데이터마이닝 기법을 이용한
출산자녀수 결정패턴에 관한 연구

-2000년 전국출산력 및 가족보건 실태조사를 중심으로-

연세대학교 보건대학원

국제보건학과

양 효 실

데이터마이닝 기법을 이용한 출산자녀수 결정패턴에 관한 연구

-2000년 전국출산력 및 가족보건 실태조사를 중심으로-

지도 정 우 진 교수

이 논문을 보건학 석사학위 논문으로 제출함

2005년 6월 일

연세대학교 보건대학원

국제보건학과

양 효 실

양효실의 보건학 석사학위논문을
인준함.

심사위원 _____ ㊞

심사위원 _____ ㊞

심사위원 _____ ㊞

연세대학교 보건대학원

2005년 6월 일

감사의 글

지금에서야 이렇게 논문을 마무리하게 되어 정말로 마음이 홀가분합니다. 논문이 끝날 것 같지 않았고 길게만 느껴졌던 시간들이였는데, 제게도 마무리하는 시간을 갖게 되어 얼마나 기쁜지 모릅니다. 무엇보다도 여리고 여리기만 한 나를 성장하게 하시려고 순간순간 말씀을 통해 지켜주시고, 깊은 기도를 하도록 환경을 허락해주신 하나님께 감사드립니다.

무엇보다도, 논문을 쓸 수 있도록 허락해주시고 날카로운 조언을 주셨던 정우진 교수님께 감사드립니다. 때로는 맘이 아픈 적도 많았지만 제 자신의 모습을 철저하게 숨김없이 되살펴보는 계기가 되기도 하였습니다. 학자로서의 원칙과 중심을 가지고 계셨던 교수님을 진심으로 존경합니다. 그리고 너무나 바쁘신 가운데서도 늘 편안하게 대해주시고 배려를 아끼지 않으셨던 서문희 박사님께도 감사드립니다. 또한, 방향을 못 잡고 헤매고 있을 때에 많은 부분을 도와주셨던 강대룡 박사님, 늘 격려해 주시고 같이 고민해주시고, 많은 힘이 되어 주신 것 잊지 못할 것 같습니다.

그리고 5학기 동안 함께 해준 동기 김은경에게 감사하고, 논문을 시작하기 전에 도움을 주신 황원주 선생님과 이선화 선생님에게도 감사합니다. 대학원 후배지만 동갑내기 친구인 락현이와 소양이에게도 고마움을 전하고 싶습니다. 그리고 부족하기만 한 며느리를 한결같은 사랑으로 대해주시는 시부모님께 감사드립니다. 항상 눈물로 기도해 주시고 나의 든든한 힘이 되어주시는 엄마, 중요한 결정을 앞두고 있을 때에 깊은 사랑을 아끼지 않으시는 아버지께 감사드립니다. 나에게 늘 함박웃음을 가져다주는 소중한 보배 조카 지민이에게도 감사하고 싶습니다. 바쁜

중에서도 맛있는 것도 많이 사주고 늘 자랑스럽고 멋진 오빠 양준호에게도 고마움을 전하고 싶습니다. 그리고 항상 변함없이 곁에서 함께 해주는 사랑하는 남편 이원구에게도 진심으로 감사합니다.

그리고 일일이 표현하지는 못했지만, 보이지 않는 곳에서 기도해주시고 늘 사랑과 격려를 주셨던 여러분들에게도 깊은 감사를 드립니다. 늘 가정에 평안이 가득하시길 진심으로 축복합니다.

"*너의 행사를 여호와께 맡기라 그리하면 너의 경영하는 것이 이루리라. 잠언 16: 3*"

2005년 7월

양 효 실 올림

차 례

표 차 례

그 림 차 례

국 문 요 약

본 연구는 현재 한국사회의 중요한 사회적 쟁점인 저출산에 관한 연구이다. 최근 몇 년째 저출산 현상이 지속됨으로서 앞으로 우리사회에 여러 방면에서 심각한 사회적 파장이 예측되고 있다. 본 연구의 목적은 우리 사회가 최근 지속되는 낮은 출산율에서 벗어나서 인구대체수준으로 출산율을 회복시키는 정책을 세우는데 필요한 의미 있는 기초적인 자료를 제공하고자 한다, 우리나라의 대표적인 출산관련 자료인 한국 보건사회연구원이 실시한 '2000년도 전국 출산력 및 가족보건 실태조사'를 중심으로 우리나라 유배우 여성들이 출생자녀수를 결정하는 의미 있는 패턴과 변수들 간의 상관성을 분석하기 위하여 데이터마이닝(Data Mining) 기법을 적용하고자 한다.

기존의 출생자녀수에 관련한 연구에서는 전통적 통계적 추론방법인 확증적 접근법은 유의성 검정이나 신뢰구간 추정을 통해 관측된 형태나 효과의 재현성을 평가하였다. 그래서 출산자녀수에 관한 결정요인 연구들은 기술적인 통계방법을 규명할 수는 있었으나, 저출산에 영향을 주는 각 요인을 복합적인 사회경제적 환경에서의 분석결과를 예측하고 이를 규칙화 할 수 없는 제한점이 있었다. 이와 더불어, 우리나라에서도 대규모 데이터들의 급속한 축적과 계산능력의 발전에 힘입어 탐색적 자료 분석의 중요성이 부각되고 있는 추세이다. 우리나라에서도 전국적인 국가단위의 출산력자료들이 축적되고 있지만, 최근에 국가적으로 절실하게 해결하여야 하는 쟁점중의 하나인 저출산에 관련한 부분에 대해서는 데이터 자체의 탐색적 자료 분석연구는 전무한 실정이다. 탐색적 방법인 데이터마이닝 기법을 적용하여 출산력 대규모의 데이터베이스 안에 존재하는 출산관련 변수들과 출생자녀수와의 상호 관련성이나 규칙 등을 분석 시도한 것은 향후 인구 및 보건학 분야에서의 데이터마이닝 기법의 적용가능성을 제시했다는 점에서 무엇보다 중요한 의미라고 생각되어진다.

본 연구에서는 전체 대상자 15세에서 49세까지의 유배우 부인 6,015명중 출생자

녀수가 1명 이상인 경우로 중점을 두고 분석대상자를 제한하였다. 선정된 대상자의 평균 출생자녀수가 2명을 기준으로 1명을 낳는 경우와 3명 이상을 낳는 경우에 영향을 미치는 변수들의 상관성과 규칙을 파악하기 위하여 데이터마이닝의 의사결정나무 기법을 이용하였고 CART (Classification and Regression Trees) 알고리즘 결과를 로지스틱 회귀분석 모형의 결과와 비교하여 제시하면 다음과 같다.

출생자녀수 2명에 비해 1명을 낳은 경우, 로지스틱 회귀분석 결과에서는 거주지, 부인의 교육수준, 자연유산의 경험유무, 첫째아 출생시 부인의 연령, 부인의 현 취업유무, 월평균 보육료에서 유의하게 영향을 미치는 것으로 나타났다. 반면에 데이터마이닝 기법을 통해 분석한 결과로는 첫째아 출생시 부인의 연령, 가구유형, 부인의 직업, 첫째아 출생년도, 월평균보육료, 최종임신연도에 의해서 의사결정나무 구조를 나타내었다. 이 경우에서는 첫째아 출생시 부인의 연령과 월평균 보육료가 두 가지 분석방법의 결과에서 공통되게 나타났다. 다음으로는 출생자녀수 2명에 비해 3명 이상을 낳은 경우에서 로지스틱 회귀분석을 통한 결정요인을 살펴봤을 때, 다른 요인을 모두 통제한 상태에서 거주지, 부부간 역할 분담, 남편의 교육수준, 첫째아 출생연도, 자녀의 필요성, 월평균 보육료에서 유의하게 영향을 미치는 것으로 나타났다. 그리고 데이터마이닝 기법에서는 첫째아 출생년도, 막내아 출생년도, 부인의 결혼연령, 막내아 출생시 부인의 연령에 따라 의사결정나무가 결정되었는데, 특징적인 것은 연령과 출생년도와 관계된 변수들이 다소 반복되어 가지치기를 하는 것을 볼 수 있었고, 첫째아 출생년도는 CART 알고리즘의 의사결정 패턴에서도 관련이 있는 변수임을 보였다.

I. 서론

1. 연구의 필요성

최근 유엔은 우리사회의 급격한 출산율에 대해서 한국정부에 의미심장한 조언을 했다. 한국 내에서 저출산 기조가 계속 될 경우 현재의 경제수준을 유지하기 위해서는 엄청난 규모의 국외 이민자를 받아들여야하는 권고였다. 저출산은 이와 함께 인구고령화[1]를 앞당기고 있으며 향후 노인부양부담이나 경제성장 및 국방에도 심각한 악영향을 미친다는 것이다.

국내에서도 저출산에 대한 전문가들이나 정책 당국자뿐만 아니라 일반 국민이나 언론 등의 관심이 급격하게 고조된 것은 실제로 저출산과 관련하여 제기되는 대부분의 현안들은 인구고령화의 문제와 유기적인 관계를 가지기 때문이다. 지난 2일 김근태 보사부장관은 '현재의 저출산, 고령화 사회 진전은 사전에 대비하고 개선하지 않으면 재난적 상황을 몰고 올수 있으나 구체적 실천 대책은 마련되고 있지 않고 있다'며 우려 섞인 경고와 함께 관련법의 제정을 서두르고 있다고 밝혔다(이수희, 2005 참조).

우리나라는 불과 한세대 전만하더라도 다산다사 (多産多死) 형태를 지닌 고출산국이었으나 2004년에는 소산소사 (小産小死) 형태이면서도 세계 최저 출산국으로 인구구조가 급격하게 변화되었다. 그 수준은 기존의 최저 출산율을 기록하던 일본이 1989년 출산율 "1.57쇼크"를 경험한 이후 2002년 합계출산율이 1.32명,

1) 인구고령화 (aging society)는 연령을 기준으로 한 인구분포의 중간 값 혹은 평균값들이 커지는 현상이다. 즉, 전체 인구 중에서 고령인구의 비중이 늘어남을 의미하는 것으로, 고령인구(65세 이상)의 비중이 7%를 넘으면 고령화 사회(aging society)에 진입하게 되며 고령인구비중이 14%를 넘는 경우는 고령사회(aged society)라고 하며, 고령인구비중이 20%를 넘는 경우는 초고령 사회라고 부른다.

2003년에는 1.29명으로 되었던 통계에 비해, 우리나라는 1960년대 초반 합계출산율이 6.0명에 이르던 것이 1984년에는 인국의 대체수준인 2.1명에 도달하였고, 1987년에는 1.6 수준까지 낮아졌다가, 1993년에는 1.75명으로 약간 상승하였으나 2000년에는 약 1.47명 수준으로 다시 감소하였다(통계청, 인구통계연보 2001). 그리고 2003년에는 가장 낮은 수준인 1.17명에 이르면서 일본보다 더 낮은 세계최저 수준이다.

더욱이 이러한 합계출산율의 변화는 인구대체수준(replacement level)이 합계출산율의 2.1명에 미치지 못하는 매우 낮은 수준의 출산율로의 변화가 약 40년이라는 매우 단시간 내에 이뤄져 왔다는 사실은 상당히 주목할 만한 내용이다. 산업화 이후의 우리나라의 인구구조의 변화는 과거 유럽의 인구변화 모형이 약 150여년에 걸친 변화과정임을 감안해볼 때, 매우 놀랍고 급속한 인구변화과정을 보여준다고 할 수 있다. 또한 이러한 인구변화의 추이는 위에서도 언급하였듯이 현재의 우리나라의 사회경제적 변화의 모습에 비추어 볼 때, 상당기간 지속될 가능성이 존재한다(변준한, 2004).

이러한 우리나라의 급속한 출산율 하락의 배경에는 급속한 사회경제적 변화와 60년대 초부터 실시되었던 정부의 적극적인 출산억제 정책이 있었다. 우리나라의 출산율 감소는 세계에서 유례를 찾을 수 없을 정도로 빠르게 진행되었다. 최근 저출산 수준이 지속되면서 정부는 과거 정책으로부터 선회하여 출산수준 부양에 커다란 관심을 보이고 있다. 과거 정부의 출산 억제 정책이 인구과잉으로 인한 인구부양이나 경제성장의 어려움을 개선하려는 것이었듯이 출산 수준을 높이려는 정책도 유사한 관심에서 추진되고 있다. 현재 우리사회 노령인구와 생산인구 비율은 적정한 수준을 유지하고 있지만 앞으로 30년이나 50년 후 생산인구대비 노령인구비율은 급상승하여 사회적 차원의 부양이나 경제성장과 관련하여 심각한 문제가 발생될 것이라고 지적되고 있다(장혜경 외, 2004). 다시 말하면, 노동

력은 경제성장의 가장 기초적인 동력이하는 점에서 이처럼 급격한 인구규모의 감소는 우리 경제의 성장 잠재력을 하락 시키게 될 것이다. 또한 고령인구를 위한 사회복지 관련 비용이 증가하게 될 것인데, 이 부담을 져야 할 청장년층의 경제활동인구는 저출산 현상으로 인해 감소하고 있으므로 사회 전체적으로 연금기금의 고갈 등 복지수요의 심각한 불균형이 발생할 수 있다(임일섭, 2004).

현재 우리나라가 당면하고 있는 출산율의 지속적인 하락은 우리 사회의 심각한 우려와 관심을 보이는 가운데, 최근 저출산과 관련하여 정책적 차원의 연구가 더욱 절실히 요청되고 있다. 이와 관련하여. 한사회의 출산율에 영향을 주는 지표들을 살펴보는 것은 상당히 중요한 의미가 있는 것이다. 출산율에 영향을 주는 주요 직접적 지표들을 살펴보면, 거시적인 관점에서 접근하는 방법인 합계출산율(TFR)[2]을 분석하는 것이고, 미시적인 접근방법은 출산자녀수를 분석하는 것이다. 이를 기반으로 하여, 본 연구에서는 미시적인 출산력 지표인 출산 자녀수를 이용하여, 최종적으로 우리나라 유배우 여성의 출산자녀수를 결정하는데 있어서 사회경제적으로 영향을 받는 요인들과 출산자녀수를 결정하는 특성의 패턴을 통계적 분석을 통하여 의미 있는 정보를 제공하기 위함이다. 본 연구가 실증적인 통계적 방법을 사용한 기존의 연구와 비교해서 차별되는 점은, 최근에 부각되고 있는 새로운 통계 기법인 데이터마이닝(Data Mining) 기법을 이용하여 대규모의 국가단위의 출산력 데이터를 이용하여, 저출산을 일으키는 사회경제적 영향요인 및 출산 자녀수 결정에 관한 패턴연구를 통하여 더욱 효율적인 정책적인 대책을 세우기 위한 기초적인 정보를 제공하고자 한다.

2) 합계출산율(Total fertility rate): 일반적으로 출산율의 척도로 가장 많이 이용되는 것으로 가임기간(15-49세)에 있는 여성이 평균적으로 몇 명의 자녀를 출산하는가를 나타내는 지표이다. 이는 각 년도의 연령별 출산율(age specific fertility rate)의 평균으로 계산된다.

2. 연구의 목적

본 연구는 최근 우리나라의 심각한 저출산 현상에서 벗어나서 인구대체수준으로 출산율을 회복시키기 위하여, 우리나라의 대규모의 출산력 데이터를 바탕으로 데이터마이닝 기법을 이용하여 한국사회의 출생자녀수를 결정하는 요인 및 결정패턴분석하고, 더 나아가 본 연구의 결과를 통해 향후 우리나라의 저출산에 대한 대책을 제시하고, 우리나라의 현실에 맞는 효율적이고 체계적인 출산율 장려정책 및 대책방안을 수립하는데 있어서 의미 있는 정보를 제공하는데 있다. 이 연구의 세부목적은 다음과 같다.

첫째, 우리나라 유배우 여성들의 자녀수에 영향을 주는 결정요인과 패턴을 파악하기 위하여 데이터마이닝의 기법의 의사결정나무를 이용하여 분석하였고, 어떠한 요인이 자녀수 결정에 중요하게 영향을 미치고 상호요인들 간의 관련성을 파악한다.

둘째, 데이터마이닝의 최적의 의사결정나무 모형의 안정성과 예측력을 평가하고, 기존의 선행연구의 문제점을 고려한 변수를 사용하여 실증적 분석(로지스틱 회귀분석)을 통해 비교평가를 한다.

셋째, 분석결과를 통해서 출산력 연구에 데이터마이닝 기법의 적용가능성을 타진하고 결과에 따른 효율적이고 체계적인 출산자녀수 장려를 위한 정책 및 대책방안을 수립하는데 있어서 의미 있는 정보를 제공하는데 있다.

3. 용어의 정의

1) **인구고령화 (aging society)**: 연령을 기준으로 한 인구분포의 중간 값 혹은 평균값들이 커지는 현상이다. 즉, 전체 인구 중에서 고령인구의 비중이 늘어남을 의미하는 것으로, 고령인구(65세 이상)의 비중이 7%를 넘으면 고령화 사회(aging society)에 진입하게 되며 고령인구비중이 14%를 넘는 경우는 고령사회(aged society)라고 하며, 고령인구비중이 20%를 넘는 경우는 초고령 사회라고 부른다.

2) **인구대체수준 (replacement level):** 인구를 현상 유지하는데 필요한 출산율 수준으로서, 선진국의 경우 대체로 2.1명이 이에 해당한다. 이 수치는 앞으로 인구가 늘어나거나 줄어들지 않도록 하기위해서 가임 여성 1인당 2.1명의 자녀는 낳아야 한다는 유럽경제위원회(UNECE)의 보고서에 따른 것이다. 개발도상국의 경우 대세로 3명 전후 이며, 사망률과 거의 비례한다. 인구학자들은 인구대체수준 이해로 출산율이 떨어지면 이를 저출산사회로 보는데 서유럽국가들 대부분이 이에 해당한다.

3) **합계출산율 (total fertility rate):** 일반적으로 출산율의 척도로 가장 많이 이용되는 것으로 가임기간(15-49세)에 있는 여성이 평균적으로 몇 명의 자녀를 출산하는가를 나타내는 지표이다. 이는 각 년도의 연령별 출산율(age specific fertility rate)의 평균으로 계산된다.

4) **출산억제정책:** 대한 가족보건복지협의회는 '덮어놓고 낳다 보면 거지꼴을 못 면한다'(1960년대) '딸아들 구별말고 둘만 낳아 잘 기르자'(1970년대) '잘 키운 딸 하나 열 아들 안부럽다'(1980년대)는 표어를 내세우며 출산억제 정책을 펴왔다. 그러나 지난해부터 '혼자서는 싫어요' '1,2,3 운동(결혼 후 1년내 임신하고, 2명의 자녀를, 30세 이전에 낳아 건강하게 기르자)' 등의 표어 및 포스터를 내걸고 출산장려 쪽으로 방향을 전환해 왔다.

Ⅱ. 이론적 배경

1. 저출산

가. 출산력의 통계적 상황

1) 국내 출산수준 지표의 변화

우리나라는 전통적인 저출산 지역인 유럽국가보다 출산율이 뒤늦게 떨어졌지만, 하락속도가 급속하게 빨라져서 현재는 더욱 낮은 수준을 보인다. 1960년대 초반 부인 1인당 합계출산율[3]이 6.0명에 이르던 것이, 1983년에는 인구의 대체수준(replacement level)인 2.1명에 도달하였고, 1987년에는 1.6명까지 낮아졌다가, 1993년에는 1.75명으로 약간 상승하였으나 2000년에는 약 1.47명 수준으로 다시 감소하였다(통계청 연도별자료 참조 2001). 그리고 2003년에는 세계에서 가장 낮은 수준인 1.17명에 이르는 급격한 감소현상을 보여주었다. 이는 인구대체수준인 합계출산율의 2.1명에 미치지 못하는 매우 낮은 수준의 출산율의 변화가 약 40년이라는 매우 단시간 내에 이루어져 왔음을 보여주고 있다. 물론 1997년 이후의 급격한 출산율의 감소현상은 IMF 구제 금융으로 인한 경기 침체와 대량 실업사태의 발생으로 인하여 출산의 시기를 연기하거나 결혼적령기의 남녀들이 결혼의 시기를 연기하면서 발생한 일시적인 현상이라고 파악할 수도 있으나, 이후 경기가 회복된 상태에서도 '밀레니엄 베이비'(millennium baby) 출산 현상과 같은 일시적인 출산 증가현상을 제외하고는 전체적인 출산율이 다시 회복되지 않고 계속적으로 감소하고 있다는 점을 감안할 때, 이러한 저출산 현상의 추세가 앞으로도 쉽게 변화되지 않을 것임을 보여주고 있다(변준한, 2004).

---

3) 합계출산율 (TFR: Total fertility Rate): 여자 1명이 가임 기간 동안 낳는 평균 출생아수, 한 여성의 전생애를 통한 출산 자녀수를 의미함. 합계출산율이란 출산 가능한 여성의 나이인 15세부터 49세까지를 기준으로 여성 한명이 평생의 가임기간 동안 출산하는 평균자녀수를 의미한다. (대한 가족보건 복지협회)

과거 유럽의 인구변화 모형이 약 150여년에 걸쳐 변화가 일어났지만, 산업화이후의 우리나라의 출산율 감소는 세계에서 유례를 찾아보기 힘들 정도로 불과 26년 만에 초고령 사회로 진입하는 최저의 출산율을 보여주는 것은 매우 놀랍고 심각한 사회적 파장을 예측하지 아니할 수 없다[4]. 또한 이러한 인구변화의 추이는 현재의 우리나라의 사회경제적 변화의 모습에 비추어 볼 때 상당기간 지속될 가능성이 높다. 즉, 한국의 경우에도 급속한 경제적 발전으로 인해 여성의 사회참여 비율의 증가, 소득의 증가, 만혼의 이혼의 증가, 의료 서비스의 광범위한 혜택의 보급 등의 사회, 문화적 변화를 수반 하면서 현재와 같은 세계적으로 가장 낮은 수준을 유지할 가능성이 상당히 높다.

표 1. 출산력의 통계적 상황 (단위; 여자 1인당 명)

연령	1960[1)]	1974[1)]	1983[2)]	1987[3)]	1990[4)]	1996[5)]	1999[6)]	2000[2)]	2001[2)]	2002[2)]
합계출산율	6.0	3.6	2.1	1.6	1.6	1.71	1.42	1.47	1.30	1.17

*자료출처: 한국사회의 저출산 원인과 정책적 함의, 김승권 (2004)

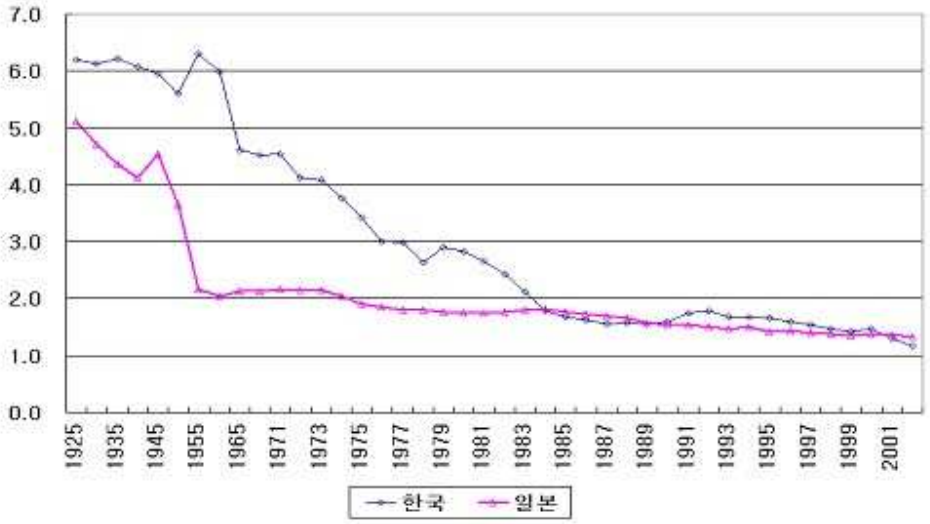
2) OECD 국가의 합계출산율 변화추이

스웨덴들 북유럽에서는 1990년 전후로 출산율이 상승하였는데(합계 출산력 2.0명 수준)하여 현재 1980년대 보다 높은 수준이다. 스페인, 이탈리아, 그리스 등 남유럽은 우리나라와 같이 1980년 이후 계속 하락하여 2000년에는 합계출산율이 1.2이하 수준을 보인다. 반면에 미국은 1980년대부터 출산율이 상승하여 서구 국가 중 예외적으로 합계출산율이 2.0이상의 수준을 보이고 있다(최경수, 2004 참조)

4) 우리나라는 고령사회(Aged Society)에서 초고령사회(Super-aged Society)로 진입하는 기간이 26년 정도에 불과할 것으로 추정되고 있다. 이는 향후 한국이 프랑스(154년), 미국(86년), 이탈리아(74년), 일본(36년)과 같은 선진국들에 비해 출산율 감소속도가 매우 빠르게 진행될 것임을 의미하는 것으로 문제의 심각성을 웅변하고 있다.[1)] 결국, 급격한 출산율 하락은 과거보다 향후에 인구고령화의 주된 원인으로 작용할 것으로 판단되며 이에 대한 원인 분석과 대책 마련의 중요성과 긴박성은 아무리 강조되어도 지나치지 않을 것이다.

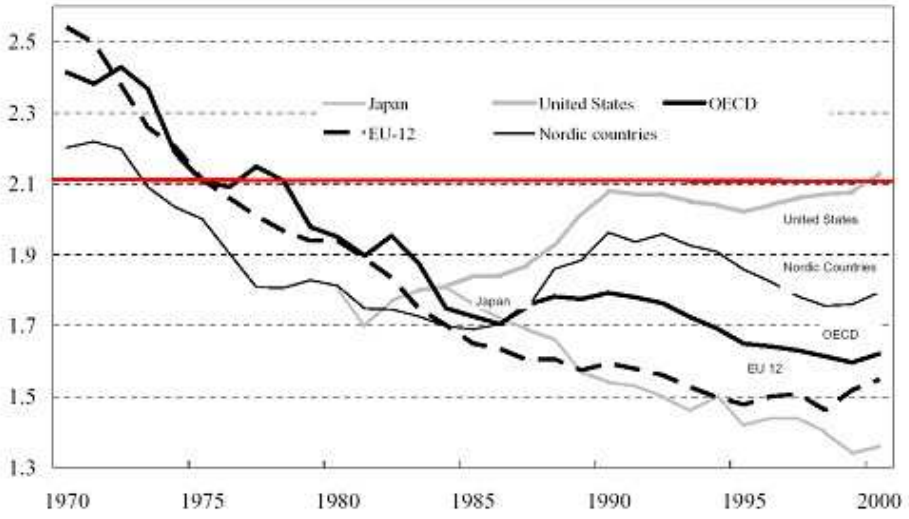
특히, 일본의 경우는 한국보다 약 15년 정도 앞서 가고 있는 인구구조 변화를 보이고 있는데, 경제활동인구는 이미 1995년 8,716만 명을 정점으로 감소세로 반전하고 있다. 이는 우리나라와 매우 유사한 합계출산율 추이를 보이고 있는데, 일본의 합계출산율은 1989년 합계출산율 1.57명으로 낮아져 사회적 '쇼크'가 발생한 이래 출산율 회복을 위한 대책마련을 고심하여 왔다. 2002년에는 합계출산율이 1.32명 2003년에는 1.29명으로 되었다. 이 수준은 인구를 유지하기 위한 적정수준인 2.09명보다 낮아 향후 총인구의 감소 및 인구 고령화가 이루어진다. 향후 총인구 통계를 보면 1980년 이전까지 크게 성장하던 인구규모가 2013년을 정점으로 감소하게 된다. 2000년 1억 2693만 명에서 2050년이 되면 1억 59만 명으로, 2100년이 되면 6천 41만4명으로 인구가 반감될 것으로 추계된다.

이와 비교할 때, 우리나라는 유례없이 일본보다 더 빠르게 출산율이 급속한 하락을 보이는 가운데, 2002년 1.17명이라는 출산율이 지속될 경우 우리나라의 인구는 2017년에는 4,925만 명으로 최대에 이른 후 , 2050년 4,610 만명, 2100년 1,621만 명으로 급격하게 감소할 것으로 전망되는 것이다(임일섭, 2004).



*자료출처: 한국 출산력하락추이에 관한 분석, 최경수(2004)

그림 1. 한국과 일본의 합계출산율



*자료출처: 한국출산력하락추이에 관한 분석, 최경수(2004)

그림 2. 최근 OECD 국가의 합계출산율 변화 추이

나. 저출산의 구체적인 사회경제적인 배경 및 원인

우리나라 출산율의 급속한 하락의 가장 큰 원인으로 정부의 출산억제정책[5]을 들 수 있다. 1960년만 하더라도 우리나라의 합계출산율은 6.0명 수준이었다. 그러나 지난 40여 년 동안 급속히 감소하여 2002년 1.17, 2003년 1.19를 기록하고 있다. 1960년대 초반부터 1990년대 중반까지 지속된 정부의 출산억제 정책이 우리사회의 급속한 출산율하락에 큰 역할을 하였다(권태환, 2002). 인구변천과정에서 의료기술발달로 사망률이 감소하였지만 출산율이 높은 수준을 유지되고 있어서 1960년대 우리나라는 높은 출산수준을 유지하고 있었다. 인구과잉을 우려한 정부는 UN의 지원을 얻어 출산수준을 낮추기 위한 다양한 정책을 실시한다. 이러한 정부의 정책을 큰 효과를 얻어서 우리나라는 짧은 시일 안에 출산주준 낮추기에 성공하였다(장혜경외, 2004). 우리사회에서 출산수준감소가 정부의 출산억제정책만으로 가능했던 것이 아니다. 정부의 출산억제정책과 동시에 사회경제발전이 동반되었기 때문에 급격한 출산율 감소가 가능했던 것이다.

그 외의 저출산의 원인으로는 결혼가치관 변화 및 초혼연령 상승, 자녀효용가지 감소, 자녀양육의 질적 강화와 양육부담증대, 자녀양육에서의 과도한 부모책임문화, '가정과 직장의 양립'을 위한 사회적 인프라부족, 여성의 자아욕구 및 사회참여 증대, 이혼 등 가족해제 증대, 불임부부 증대 등으로 정리하고 있다(김승권, 2005). 그리고 1997년 이후에는 IMF 구제 금융으로 인한 경기침체와 대량 실업사태로 미혼 남녀의 결혼 연기 및 기피현상으로, 혹은 기혼 남녀의 출산 지연 및 기피의 원인으로도 작용했다.

---

5) 출산억제정책: 대한 가족보건복지협의회는 '덮어놓고 낳다 보면 거지꼴을 못 면한다'(1960년대) '딸아들 구별말고 둘만 낳아 잘 기르자'(1970년대) '잘 키운 딸 하나 열 아들 안부럽다'(1980년대)는 표어를 내세우며 출산억제 정책을 펴왔다. 그러나 지난해부터 '혼자서는 싫어요' '1,2,3 운동(결혼 후 1년 내 임신하고,2명의 자녀를,30세 이전에 낳아 건강하게 기르자)' 등의 표어 및 포스터를 내걸고 출산장려 쪽으로 방향을 전환해 왔다.

다. 저출산으로 인한 사회경제적 영향들

현재와 같은 사회 전반의 저 출산의 기조가 지속될 경우, 사회경제적으로 엄청난 파장을 일으킨다. 앞에서 언급하였듯이, 실제로 저출산과 관련하여 제기되는 대부분의 현안들은 인구고령화와 유기적인 관계를 가진다. 예를 들면, 출산자녀수가 급격히 감소함으로서 상대적으로 노령인구비율의 급증하여 경제활동연령층의 복지비 부담의 증가를 먼저 꼽을 수 있겠다. 특별히 현재 우리나라의 연금제도는 출산율 1.8명 수준을 기초로 하여 계산된 것이기 때문에, 출산율이 1.8명 수준 이하로 저하 하게 되면 결국 연금 재정을 손실을 입을 수밖에 없다. 또 한 손실된 부분을 보충하기 위해서는 연금의 수혜에 필요한 가입기간을 연장하거나 연금요율을 상향 조정할 수밖에 없는 악순환이 반복될 수밖에 없는 실정이다. 이는 최근까지도 유럽에서 경험했던 장기적인 경기침체현상의 발생 원인과 일분 일치하는 모습이다(변준한, 2004). 우리나라의 경우 저출산으로 인한 고령화 영향이 아니더라도 사회안전망인 각종연금제도와 건강의료보험제도의 기반이 취약하기 때문에 개혁적 대책의 마련과 실시가 시급하다고 하겠다.

그리고 저출산으로 인한 사회적 문제들을 예측해 보면, 인구 중 경제활동인구의 감소, 노동투입의 감소, 저축률의 감소 등이 발생할 수 있으며, 저출산에 따른 인구고령화가 한 경제의 공급, 수요, 분배에 영향을 미쳐 경제성장이 둔화되는 것으로 정리 될 수 있다. 우선 노동공급 측면에서 고령화 진전은 생산가능 인구 감소, 생산가능 인구 자체의 고령화에 따른 노동생산성 저하로 인해 경제성장이 둔화될 것이며, 또한 자본 공급의 측면에서 볼 때 고령화 진전은 민간 저축률 하락으로 인한 가용자금 감소와 투자위축으로 이어져 경제성장을 둔화시킨다는 것이다. 수요측면에서는 주 수요자 층이 고령자 중심으로 이동하여 실버 및 보건의료산업 확대로 현대 주 소비계층이 30대~60대에서 고령자 중심으로 이동할 경우, 소비가 감소하여 경제성장이 둔화될 수밖에 없는 것이다.(이수희, 2005 참조).

2. 데이터마이닝 기법

가. 출산력 연구와 데이터 마이닝 분석방법

지금까지 출산자녀수를 결정하는 영향요인에 관한 연구는 거의 실증적인 연구방법을 사용하였다. 그러나 본 연구에서 사용할 통계기법인 데이터마이닝은 일종의 탐색적인 자료 분석 방법으로, 데이터 마이닝은 대용량의 데이터에서 통계기법이나 모델링 기법을 이용하여 기존에 발견되지 않았던 내재된 패턴을 찾아내는 과정이다. 데이터 간 상호 패턴을 찾는 것이 데이터 마이닝의 주요 목적이라는 의미이다. IT산업 전반에 걸쳐 매우 방대한 양의 데이터를 많은 기업이 축적하고 있다.(프로그램의 세계, 2003) 아울러 최근부터는 이러한 데이터를 축적하고 유지하는데 그치지 않고 이러한 대용량 데이터베이스를 실제 업무에 있어서의 국가 및 기관의 정책에 효율적 활용을 위한 방편으로, 다양한 통계기법과 정확한 정보에 근거한 전략이니 대량의 데이터를 지식으로 효과적으로 저장 관리, 활용할 수 있는 지식 탐사 방법인 데이터마이닝에 대한 중요도가 증대되고 있다.

최근 데이터 마이닝의 중요성이 강조되는 이유는 여러 가지가 있으나 첫째, 방대한 데이터베이스 속에 축적된 많은 양의 데이터를 보다 효율적으로 이용하고 둘째, 데이터마이닝 알고리즘의 발달과 컴퓨터의 용량 및 성능향상은 양적으로 증가되고 복잡한 형태를 가진 데이터의 처리과정을 보다 쉽게 처리할 수 있도록 함으로서 원하는 정보를 보다 쉽게 얻을 수 있는 환경을 제공한다. 셋째, 데이터 마이닝 기법은 기종의 전문가 시스템이 갖은 한계점인 지식획득의 병목현상을 유연하게 극복할 수 있는 대안으로 자리 잡고 있다. 데이터로부터 유용한 지식을 획득하려는 고정인 데이터 마이닝은 언급되는 분야에 따라서 여러 가지 명칭을 혼재되어 불리고 있다. 예를 들면, 지식추출(Knowledge Extraction), 정보발견(Information Discovery), 데이터 연금술(Data Archeology), 데이터 패턴처리(Data

Pattern Process), 정보수확(Information Harvesting), KDD(Knowledge Discovery in Database)등과 같다(이건창, 2001).

데이터마이닝의 이러한 중요성이 부각됨에도 불구하고, 현재까지 진행된 출산 자녀수에 관한 연구들을 살펴보면, 지금까지의 전통적 통계적 방법인 확증적(confirmatory) 접근법이였다. 이는 유의성 검정이나 신뢰구간추정을 통해 관측된 형태나 효과의 재현성(reproducibility)을 평가하였으나, 데이터의 급속한 축적과 계산 능력의 발전에 힘입어 탐색적(exploratory)자료 분석의 중요성이 부각되고 있는 추세이다(허명희외, 2000).

더불어 데이터 마이닝은 이전의 수동적인 의사결정지원 도구에 반하여 데이터 마이닝은 능동적인 의사결정 지원을 제공한다. 즉, 사용자가 문제를 정의하고, 데이터를 선택하고, 데이터 분석을 위한 도구를 결정하는 것이 아니라, 데이터 마이닝 도구는 자동적으로 데이터의 예외적인 상황과 가능성이 있는 관계를 감지하여 사용자가 발견하지 못한 문제를 파악할 수 있게 한다. 다시 말하면, 데이터 마이닝의 도구는 데이터를 분석하여 데이터 관계 속에 숨어있는 문제나 기회를 발견하고, 이러한 문제를 기초로 컴퓨터 모델을 구성하여 최소의 사용자 개입으로 사업행동을 예측한다. 따라서 사용자는 시스템에서 분석한 결과를 토대로 데이터 마이닝에 사용되는 분석 알고리즘은 인공지능, 신경망, 귀납규칙, 술어 로직, 의사결정 트리, 그리고 유전자 알고리즘 등이 있다(서길수, 2000).

나. 데이터마이닝의 정의 및 설명

데이터 마이닝이란 자동화되고 지능을 갖춘(automated and intelligent) 데이터베이스 분석기법으로 90년대 초반부터 지식발견(KDD: Knowledge Discovery in Database), 정보발견(information discovery), 정보수확(information harvesting)들의 이름으로도 소개되어 왔는데 일반적으로 “대량의 데이터로부터 새롭고 의미 있는

정보를 추출하여 의사결정에 활용하는 작업"이라 정의된다. 용어에 '채굴하다'라는 의미의 'mining'을 포함시킨 이유는 데이터로부터 정보를 찾아내는 작업이 마치 금이나 다이아몬드를 발견하기 전에 수많은 양의 흙과 잡석들을 파헤치고 제거하는 것과 유사하다는 데에 기인한다(장남식외, 1997).

다시 설명하자면, 데이터 마이닝은 대규모의 데이터베이스 내에 존재하는 숨겨져 있는 데이터간의 관계(relationships)나 새로운 규칙(rules)등을 탐색적으로 찾아내 모형화(modeling)해서 유용한 정보로 변환하는 일련의 과정이다(최연희 외, 2004).

다. 데이터마이닝의 수행과정

먼저, 데이터 마이닝의 일반적인 수행과정을 살펴보면(최국렬 외, 2001),

- 데이터샘플링 (Sampling/selection)

데이터마이닝은 수 십 메가에서 수 십 기가에 이르는 대용량의 데이터를 기반으로 한다. 그러나 방대한 양의 데이터를 살피는 것은 시간의 측변에서만보아도 많은 인내를 요하게 되는 작업이 될 수 있다. 이때 고려하여야 하는 과정이 바로 샘플링이다. 샘플링이란 방대한 양의 데이터에서 모집단을 닮은 작은 양의 데이터를 추출하는 것이다.

- 자료탐색 (Exploration)

여러 측면에서 데이터 탐색을 통해서 기본적인 정보를 검색하고 유용한 정보를 추출하는 기법들을 제공한다. 탐색과정에서는 이미 알고 있는 사실들을 확인하여 수치화하는 작업을 시작을 하여 보유하고 있는 수많은 변수들의 관계를 살펴보는 단계이다.

- 자료변화 (Modification)

탐색 단계에서 얻어진 정보를 기반으로 모형화 단계에서 모형의 성능을 향상시키기 위해, 데이터가 가지고 있는 정보를 효율적으로 사용할 수 있도록 변수변환, 수량화, 그룹화 같은 방법을 통해서 데이터를 변형하고 조정한다.

- 모형화 (Modeling)

데이터 마이닝의 핵심이라고 할 수 있는 모형화 단계는 이전단계세서의 결과들을 토대로 하여 분석목적에 따라 적절한 기법을 통해서 예측모형을 찾아내는 방법들을 제공한다.

- 평가 (Assessment)

모형화를 통해 얻어진 결과의 신뢰성, 타당성, 유용성들을 평가할 수 있다. 평가 단계에서는 리프트 도표(Lift Chart), ROC(Receiver Operating Characteristic)곡선, 이익도표(Profit Chart), ROI(Return On Investment)곡선들 다양한 평가도구가 제공된다.

라. 데이터마이닝 (Data Mining)의 기법

데이터 마이닝의 수행방법 뿐만 아니라 기법의 종류들을 살펴보면 , 데이터마이닝에는 특정문제에 적용하는 기법이 정해져있지는 않다. 또한 기법이 적용된다고 해서 모든 문제가 해결되는 것도 아니다. 얻고자 하는 결과나 데이터의 상태에 따라 적용할 수 있는 기법들은 다를 수 있다. 그러므로 기법들에 대해 어느 정도 이해가 수반되면 문제를 해결하는데 좀 더 최적의 접근으로 보다 효과적이고 적극적인 데이터마이닝을 수행할 수 있을 것이다. 데이터마이닝 기법에는 일반적으로 통계학에서 얘기되는 여러 분석 기법들을 포함하여 연관성 규칙 발견(Association Rule Discovery, Market Basket Analysis), 군집발견(Cluster Analysis), 의사결정나무(Decision Tree), 사례기반추론(Case-based Reasoning), 인공신경망

(Artificial Neural Network), OLAP(On-Line Analytic Processing), 유전자 알고리즘(Genetic Algorithm), 판별분석(Discrimination Analysis)등과 같은 기법들이 있다. (강현철, 2001)

이러한 기법들 중에 목표변수(target variable)의 유무에 따라 적용기법이 달라지는데, 일반적으로 목표변수가 정해졌을 경우에는 의사결정나무, 사례기반 추론, 인공신경망등을 사용하고, 목표변수가 정해져있지 않을 경우에는 연관성 규칙발견, 군집분석등을 사용한다. 본 연구에서처럼 '출산자녀수'라는 목표변수를 정한 경우는 위에 제시된 기법들 중 살펴보아야 할 것인데, 먼저 사례기반추론은 특정 알고리즘을 이용하여 검색된 유사 과거 사례들을 조합하여 새로운 문제에 대한 해석을 도출하는 것이다. 그리고 인공신경망은 두뇌 신경망 활동을 흉내 내어 자신이 가진 데이터로부터의 반복적인 학습과정을 거쳐 패턴을 찾아내고 이를 일반화함으로써 특히 향후를 예측하고자 하는 문제에 유용한 것으로 매우 복잡한 구조를 가진 데이터들 사이의 관계나 패턴을 찾아내는 유연한 비선형 모형의 하나이다. 그러나 결과에 대한 해석이 불가능하고 계산 속도가 느리다는 단점을 가지고 있다. 그러나 의사결정나무는 결과의 해석이 가능하고 계산속도도 빨라 대용량의 자료를 다룰 수 있으며 분석과정이 나무구조로 표현되어지기 때문에 쉽게 이해하고 설명할 수 있다는 장점이 있다. 그러므로 다음은 본 연구에 적절하다고 판단되어지는, 데이터마이닝의 모형화 단계에서 이용될 수 있는 기법 중 대표적인 기법인 의사결정나무를 택하여 분석할 것이다.

마. 의사결정나무의 구조와 표현방법

본 연구에서 선택한 데이터마이닝의 주요도구인 의사결정나무를 살펴보면,

의사결정나무는 의사결정규칙(decision rule)을 도표화하여 관심대상이 되는 집단을 몇 개의 소집단으로 분류(classification)하거나 예측(prediction)을 수행하는 방

법이다. 분석과정이 나무구조에 의해서 표현되기 때문에 판별분석 , 회귀분석, 신경망등과 같은 방법들에 비해 연구자가 분석과정을 쉽게 이해하고 설명할 수 있다는 장점을 가지고 있다(최종후외, 1999).

그리고 데이터마이닝에서의 의사결정나무는 탐색(exploration)과 모형화(modeling)의 특성을 지니며, 사전에 이상치(outlier)를 검색하거나 분석에 필요한 변수를 찾아내고 분석모형에 포함되어야 할 교호효과를 찾아내는 데 사용될 수 있고, 그 자체가 분류 또는 예측모형으로 사용될 수도 있다(호승희, 채영문외, 2000 ; 강현철외, 2001). 의사결정나무는 알고리즘에 따라서 차이는 있겠지만 흔히 그림1과 같은 형태로 그 결과가 출력된다(최종후, 1998). 그림3에서 그려진 네모들은 모두 **마디(Node)**로 불리는데, 이는 각각 데이터 베이스에서 사용되는 용어로 레코드(Record), 통계에서 사용되는 용어로는 관측치(observation value)들의 집합체라고 할 수 있다. 맨 위에 그려진 마디를 **뿌리 마디(Root Node)**라 하며 모든 레코드, 관측치 들을 의미한다. 뿌리마디를 어떤 특정한 법칙에 의하여 나누게 되면 바로 밑에 생기는 마디들은 **자식마디(Child Node)**라고 한다. 결국 자식마디는 뿌리마디의 한부분이 되는 것이다. 이러한 방법으로 특정 법칙에 따라 위의 마디를 나누는 것을 **가지치기**라 하며, 더 이상 가지가 쳐질 마디가 없으면 이 마디는 **종단마디(Leaf Node)**가 된다. 따라서 종단마디는 의사결정나무의 맨 하단에 생기게 되는데 여기서 관측치 즉, 표본들의 분류 규칙이 생기게 된다.

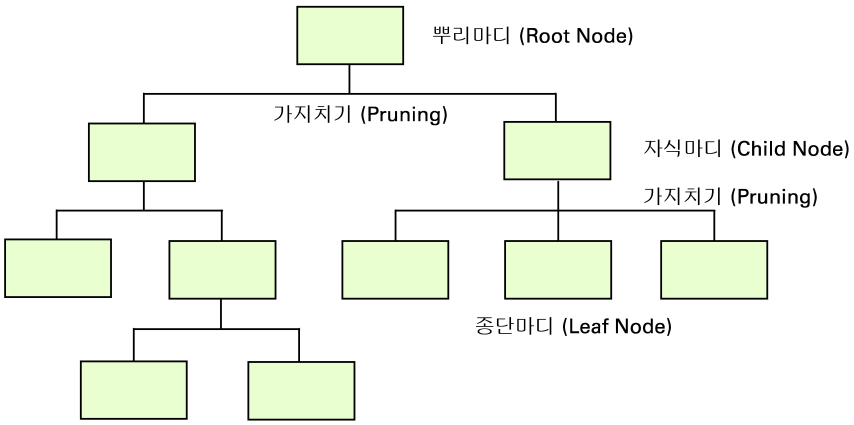


그림 3. 의사결정수의 나무구조 표현방법

위와 같은 의사결정나무 구조가 가지치기될 때 사용되는 마디분리기준(Splitting Criteria)은 이산형 결과 변수일 경우는 카이제곱 통계량, 지니지수, 엔트로피지수가 있으며, 연속형 결과변수일 경우는 F통계량, 분산의 감소량을 사용한다. 또한 마지막 분석결과가 나오는 종단마디를 어디서 멈추는가 하는 정지규칙(Stopping Rule)은 더 이상 분리가 일어나지 않고 현재의 마디가 끝마디(Terminal Node, Leaf) 되도록 하는 여러 가지 규칙을 말한다. 이러한 의사결정나무(Decision Tree Analysis)의 대표적 알고리즘은 CART, CHAID, C4.5(Quinlqn, 1993)등이 있으며, 이들은 많은 소프트웨어 회사들에 의해서 다양한 제품으로 상용화되어 있다. 이 중에서 본 연구에서는 CART(Classification and Regression Tree) 알고리즘을 이 연구에서는 이용하였다. CART 알고리즘은 의사결정나무의 대표적인 알고리즘으로써, 의사결정나무분석의 기본적인 목적은 다양한 결정요인들에 의해 나타나는 결과(종속변수)를 보다 동일한 특성(불순도의 감소)을 가진 하부그룹으로 분류하고자 할 때, 그룹간 특성은 가장 크게 하면서 그룹 내 동질성을 최대화하는 결정요인들이 무엇인지 그리고 그 요인들이 어떻게 우선적으로 관여하는지 파악하고자 하는 것이다. 분류를 단계적으로 계속 시행할수록 그룹 내의 불순도는 감소하게 된다.

그 방법적인 원리는 다음과 같다. 의사결정나무 분석은 마디 (node) 라고 불리는 구성요소들로 이루어져 있으면, 모든 관측치를 포함하는 뿌리마디 (root node) 로부터 시작하여 특정법칙에 의해 각가지가 끝마디 (terminal node) 에 이를 때까지 자식마디 (child node)를 계속적으로 형성해 나감으로 해서 완성된다. 자식마디들이 형성될 때에는 설명변수들의 선택과 병합이 여러 가지 기준에 의해 이루어진다. 여러 가지 통계량 (지니지수, 엔트로피 지수, 분산의 감소량)등을 이용하여 부모마디로부터 불순도 (impurity) 가 적은 자식마디를 형성할 때 우선적인 결정요인을 파악하는 것이다(최연희외, 2004).

바. 의사결정나무의 구성 원리

- 일정기준(나무의 최종 node수, 자료 수, 나무의 깊이)에 의해 나무 모양 결정.
- 의사결정나무의 가지분리방법 (Split), 잘라내기 (Prune)방법에 따라 CHAID, CART, C4.5, QUEST등 여러 가지 알고리즘이 사용
- 분리기준(Splitting Criteria): 순수도(Purity), 불순도(Impurity) 기준.
- 정지규칙(Stopping Rule): 더 이상 분리가 일어나지 않고 현재의 마디가 끝마디 (Terminal node, Leaf)가 되도록 하는 규칙
- 잘라내기: 지나치게 많은 마디를 가지는 의사결정나무는 새로운 자료에 적용할 때 예측 오차 (Prediction error) 가 매우 클 가능성.

사. 확증적 통계방법과 데이터 마이닝의 장단점

앞에서 언급한 바와 같이, 지금까지 인구 및 출산력 관련 분야의 대부분의 통계기법으로 확증적 통계방법을 선호해왔다. 그러나 이러한 확증적인 통계방법인 로지스틱 회귀분석을 통하여 종속변수의 위험요인을 규명하고 다른 변수를 통제한 상태에서 독립변수가 종속변수에 미치는 영향정도를 알 수 있었으나, 현실적으로 각 요인들이 상호 결합된 상황에서의 결과를 예측하고 이를 규칙화 할 수는 없는 제한점이 있었다. 그러나 데이터 마이닝의 의사결정나무 기법은 변수들간의 상호 복합적인 상황에서 상호 관련성과 패턴을 알 수 있는데, 설명변수 그룹별로 목표변수인 출산자녀수의 분류와 예측의 과정을 나무구조에 의한 추론규칙에 의해 표현하였으며, 출산자녀수 결정에 영향을 미치는 요인을 도출하고 이를 통해 대상 군별 특성과 규칙을 규명할 수 있는 것이다.

더불어, 이 방법은 분류 또는 예측의 과정이 나무구조에 의한 추론규칙에 의해서 표현되기 때문에, 다른 방법들을 예를 들면 신경망, 판별분석, 회귀분석들에 비해서 연구자가 그 과정을 쉽게 이해하고 설명할 수 있다. 그래서 본 연구에서는 로지스틱 회귀 분석과 데이터마이닝의 장단점을 비교해서 정리한 표는 다음의 <표 2>와 같다.

표 2. 로지스틱 회귀분석과 데이터마이닝의 장단점 비교

구분	확증적 통계방법 (로지스틱 회귀분석)	데이터마이닝 (의사결정나무)
장점	-현상을 선형(평면 등)으로 반영, 추정치에 대한 해석 및 변수간 영향력의 비교가 용이, 모형개발이 용이 -회귀모형의 단순성과 해석상의 편리함은 선형성을 가정하기 때문에 가능함 -회귀계수나 오즈비와 같은 회귀분석의 결과는 많은 유용한 정보를 제공하고 해석이 편리함	-분류나 예측의 근거를 알려주기 때문에 이해가 용이함 (원인설명에 대한 rule이해가 용이) -데이터를 구성하는 속성의 수가 불필요하게 많을 경우에도 모형구축 시 분류에 영향을 미치지 않는 속성들을 자동으로 제외시키기 때문에 데이터 선정이 용이함 -연속형이나 명목형 데이터 값들을 기록된 그대로 처리할 수 있기 때문에 지식발견 프로세스 중 데이터의 변환단계에서 소요되는 기간과 노력을 단축시킴. -어떠한 속성들이 각각의 부류 값에 결정적인 영향을 주는가를 쉽게 파악가능 -모형구축에 소요되는 시간이 짧다.
단점	-원천적으로 변수들 간의 관계가 복잡하여 선형성을 가정할 수 없는 경우에는 모형의 적합성(예측의 효용성)측면에서는 한계가 있음 -사전에 입력변수선택에 대한 충분한 탐색 필요. -설명변수 입력정보가 많을수록 추정이 좋아지는 의미를 갖지만, 반응변수와 관련성이 없거나 설명력이 약한 변수가 지나치게 많이 포함되면 일반화에 역기능적인 효과를 가져 올 수 있고 또한 불안정성의 원인이 될 수 있고 설명력을 떨어뜨리게 됨.	-변수별 영향정도가 민감하게 차별화되지 않으며 연속형 데이터를 처리하는 능력이 신경망이나 다른 통계기법에 비해 떨어지며, 결과적으로 예측력도 감소함 -부류가 연속형 변수의 형태를 취하며, 이것을 예측하는 모형을 구축하는 것이 목적일 경우에는 적합하지 않음 -모형을 구축하는데 사용되는 표본의 크기에 지나치게 민감함

아. 데이터마이닝을 활용하여 해결한 해외 및 국내사례들

1. 분류 (Classification)

데이터마이닝을 이용하여, 자동차보험에 가입한 피보험자를 사고 유무에 따라 분석한 외국 사례로서 여러 변수들을 통하여 분석 모형을 구해 본 결과 운전경력, 나이, 직업 등이 사고 유무를 예측하는데 큰 기여를 하였다. 분석모형에 의해 계산되는 사고발생 확률에 의해 보험료율을 산정하기로 하고 분류를 결정짓는데 공헌도가 높은 설명변수들을 중심으로 분류 산정표를 작성하였다.

2. 군집 (Clustering)

국내 한 보험 회사에서 영업효율을 극대화하기 위한 방안을 마련하고자 고객자료와 보험사 영업사원 인사자료를 연계해서 데이터 마이닝 분석을 적용한 결과 고객과 같은 나이, 같은 학력을 가진 보험모집인이 그 고객층에 대한 영업효율이 가장 높다는 사실을 발견하였다, 그래서 젊은 직장인들이 밀집해 있는 사무실에는 젊은 대졸 학력의 보험 모집인을 투입시킴으로써 영업효율을 증대시킬 수 있었다.

3. 연관 (Association)

한 수퍼마켓에서 고객들의 장바구니 분석(market basket analysis)을 실시한 결과 와인을 사는 손님 중 84%이상의 치즈와 크랙커를 구입하고, 핫케익 가루를 구입하면 87%의 확률로 시럽을 산다는 규칙을 찾아낼 수 있었다.

4. 고객 이탈방지 (Customer Churn Analysis)

한 통신회사에서는 40% 이상의 고객이 가입한지 6개월 이내에 다른 통신회사로 옮기거나 서비스를 중단하는 것으로 조사되었다. 특히 그 중 80% 정도가 대대적인 캠페인 기간에 저렴한 단말기를 구입하며 통신회사에 가입한 고객들이었다. 데이터 마이닝에 의해 이탈한 고객의 성향을 더욱 자세히 분석하여보니 캠패

인 기간 가입자 중 대부분이 20대의 일정 수입원이 없는 여성이 대부분이었다. 다시 서비스로 옮기는 경우, 서비스업체가 자신에 대한 배려가 부족하다는 다소 감정적인 면이 많았다. 영업을 목적으로 통신 서비스를 받는 사람들은 이탈율도 낮았고 서비스의 질에 크게 신경 쓰지 않는 경향이 있었다. 이과 같은 데이터 마이닝 결과를 통해 대대적 캠페인에 의한 고객 수 불리기는 자제하고 애수 고객에 대한 지속적 고객관리가 실질적 이익을 회사에 안겨준다는 사실을 알 수 있었다.

5. 교차판매 (Cross-selling)

24시간 편의점에서 고객들의 상품구매유형을 파악하기 위하여 자료를 관찰한 결과 밤 10시 이후에 남성고객들은 기저귀와 맥주를 함께 구입한다는 사실을 발견하였다. 그 이유는 30대 맞벌이 부부는 퇴근 시간 이후 쇼핑을 하게 되는데, 기저귀를 차는 정도의 아기가 있는 가정에서는 남성 혼자 편의점에 들르게 되고 맥주구입의 유혹을 느끼게 된다는 사실을 자료를 통해 그 패턴을 알 수 있었던 것이다. 그래서 기저귀 옆에 맥주를 진열함으로써 판매효율을 극대화 할 수 있었다.

자. 데이터마이닝 기법을 적용한 출산력 연구

앞서서 살펴 본 바와 같이 데이터 마이닝을 통해 우리가 얻을 수 있는 정보의 종류는 굉장히 다양하다. 이러한 방법들은 보건의료분야에서도 대규모의 환자 데이터를 이용하여 데이터 마이닝 기법을 사용한 질병패턴의 분석, 의료이용도분석, 건강증진관련 분석, 병원경영을 위한 마케팅 분석 등에 사용이 빈번해지고 그 적용방법 또한 다양화되고 있으나, 최근에 국가적으로 절실하게 해결하여야 하는 인구 및 출산력 분야의 심각한 쟁점인 저출산에 관련한 부분에 대해서는 데이터 자체의 탐색적 자료 분석연구는 전무한 실정이였다. 물론 인구 및 출산력 관련 분야에서도 대규모의 자료들이 축적되고 있는데, 거시적인 통계데이터로는 통계청에서 발간한 현재 인구동행 및 인구추계자료집과 통계청 사이트 자료 등이 있으며, 미시적 원자료로는 우리나라의 여러 연구기관에서 정기적으로 전국적인 국가단위의 대규모의 출산력자료들을 보고하고 있는데, 이는 데이터 마이닝 기법을 이용한 연구의 가능성을 제시하기도 한다. 그런 면에서 본 연구에서는 데이터마이닝 기법을 이용하여 대규모의 출산력 데이터베이스 안에 존재하는 관련성이나 규칙 등을 분석을 시도하려는 것은 상당한 의미가 있다고 본다. 새로운 통계기법인 데이터 마이닝을 사용함으로써 기존의 조사나 연구에서 발견할 수 없었던 의미 있는 새로운 상관관계, 패턴, 추세를 밝혀내고 이와 같은 유용한 정보를 집적하고 체계화하여 한국사회의 심각한 저출산의 쟁점을 해결하기위해 데이터마이닝의 기법을 출산관련 연구 분야에도 활용할 수 있는지 근거를 제시하려는 것은 중요한 의의가 있을 것이다. 더 나아가 인구 및 출산력 분야 연구에서도 데이터 마이닝의 기법을 적용가능성을 타진할 수도 있으며, 보다 다양한 연구를 통해서 의사결정 모형의 타당성과 안정성을 높여감으로써 각기 다른 특징으로 분류된 집단에 따른 정책적 전략을 세우는 것도 고려해볼 수 있을 것이다.

Ⅲ. 연구 방법

1. 연구의 틀

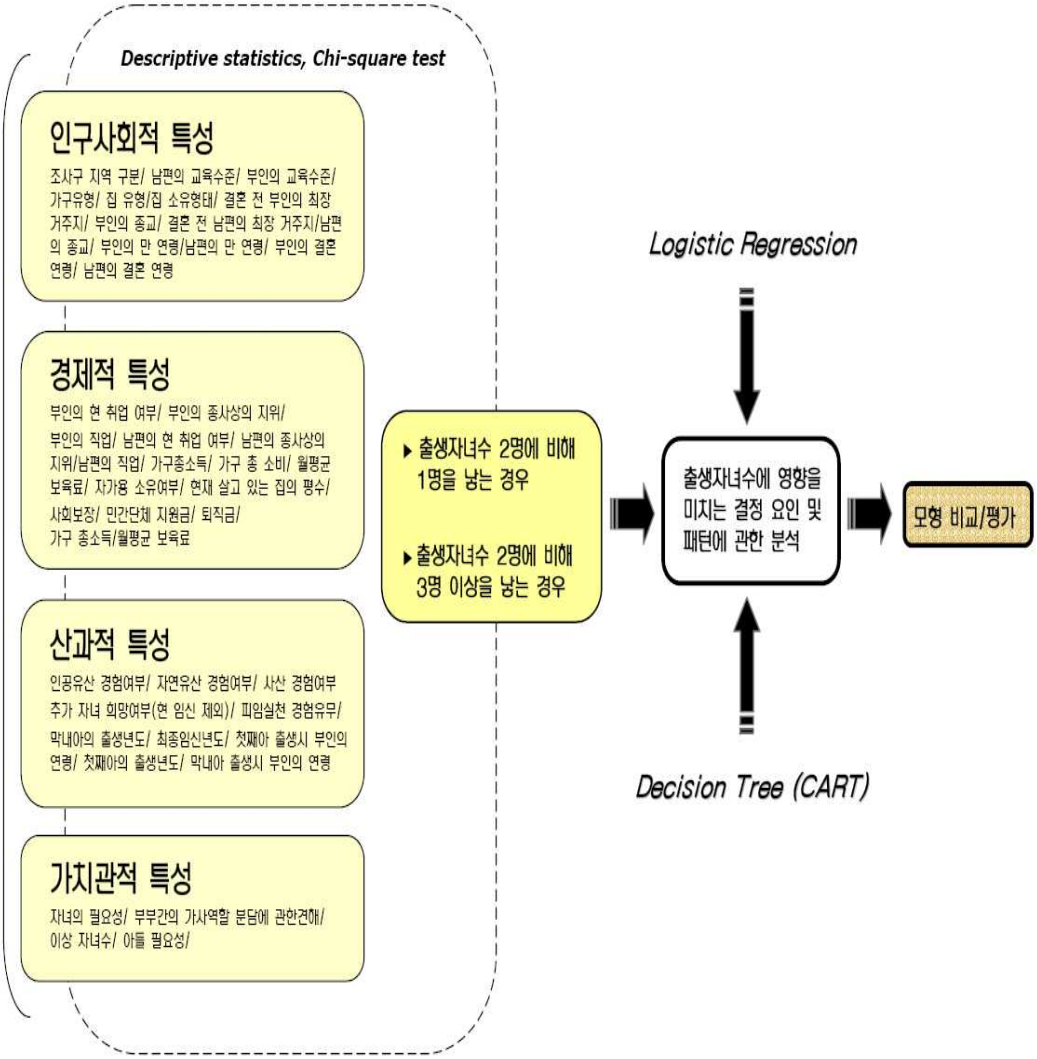


그림 4. 연구의 틀

2. 연구자료 및 연구대상의 특성

본 연구에서는 우리나라의 출산관련 대규모 데이터인 ‘2000년 전국 출산력 및 가족보건 실태조사’자료를 사용하였다. 이 설문조사는 출산력 관련 실증분석을 위한 미시적인 원 자료이며, 2000년 6월 19일부터 8월 31일 사이에 15~64세의 기혼배우 부인 11,553명을 대상으로 하여 훈련된 조사원이 직접 면접조사로 실시하였으나, 본 연구의 대상자는 그 중에서 유배우 기혼여성 19~45세의 여성 6021명의 자료를 바탕으로 통계분석을 실시하였다. 이는 우리나라에서는 다른 서구 선진국이나 모계위주의 사회체계가 존속하고 있는 국가와는 달리, 미혼여성에 의한 정상적이고 합법적인 출산이 거의 이루어지지 않으므로, 대상자를 선택하는데 있어서는 기본적으로 일부일처제로 구성된 부부이면서, 정책적인 실효성을 높이기 위하여 배우자가 있는 부인에 한하여 분석대상자를 제한하였다. 이를 위해서 형제나 자매가 가구를 이루고 있거나 편부나 편모상태에서 미혼자녀를 둔 가정이거나 조부모와 손자나 손녀로 구성된 자녀 그리고 1인 가구나 비혈연 가구의 유형은 분석에서 제외하였다. 또한, 본 연구에서는 출생자녀가 없는 경우는 제외하였으며 현재 임신 중으로 출산예정자녀를 가진 경우는 포함하여, 출생자녀수가 1명 이상인 부인들의 자료만을 선택적으로 이용하였다.

3. 통계분석을 위한 변수선정 및 정의

가. 본 연구에 투입된 최종변수들

본 연구에서는 데이터마이닝 분석을 위해서 '2000년도 전국 출산력 및 가족보건 실태조사'의 자료를 바탕으로 위에 제시된 선행연구의 변수들을 참고로 하여, 가구사항과 응답부인 및 남편의 특성에 관한 사항, 임신과 출산 그리고 피임에 대한 사항, 자녀의 보육사항과 남편 및 부인의 취업에 관련한 사항, 마지막으로 가족 가치관에 관한 사항들을 일차적으로 분석에 기초가 되는 100여개의 변수들을 최대한으로 종합하여 기술통계량을 분석하였다. 그리고 중복되는 개념들을 제외하고 의미 있는 변수들을 선택하는 작업을 몇 차례의 과정을 거친 후에 다음의 표와 같이 최종 변수 34개를 인구사회학적, 경제적, 산과적 및 가치관 특성으로 구분하여 정리하였다. 최종적으로 정리한 변수들의 내용은 다음과 같다.

표 3. 출산력과 관련된 변수정리

구분	변수목록
인구 사회학적 특성	조사구 지역 구분/ 남편의 교육수준/ 부인의 교육수준/ 가구유형/ 집 유형/집 소유형태/ 결혼 전 부인의 최장 거주지/ 부인의 종교/ 결혼 전 남편의 최장 거주지/남편의 종교/ 부인의 만 연령/ 남편의 만 연령/ 부인의 결혼 연령/ 남편의 결혼 연령
경제적 특성	부인의 현 취업 여부/ 부인의 종사상의 지위/ 부인의 직업/ 남편의 종사상의 지위/ 남편의 직업/자가용 소유여부/ 가구 총소득/월평균보육료/ 현재 살고 있는 집의 평수
산과적 특성	인공유산 경험여부/ 자연유산 경험여부/막내아의 출생년도 최종임신년도/ 첫째아 출생시 부인의 연령/ 첫째아의 출생년도/ 막내아 출생시 부인의 연령
가치관 특성	자녀의 필요성/ 부부간의 가사역할 분담에 관한 견해/ 이상 자녀수/아들 필요성

나. 목표변수(target variable)인 출생자녀수

출산자녀수에 관련한 기존의 선행연구를 살펴보면, 종속변수를 김한곤(1993)의 연구에서는 실제로 태어난 자녀의 수로 정의하였고, 보건사회연구원(2001)의 연구에서는 부인의 현재 자녀수와 출산예정인 자녀수 및 장래의 추가 자녀수를 합한 값으로 종속변수를 정의하였다. 그리고 변준한(2004)의 논문에서는 좀 더 다양하고 세부적으로 종속변수를 규정하였는데, 현존자녀수에 출산예정인 자녀수와 향후 추가자녀를 합한 자녀수와, 이상 자녀수, 그리고 현존자녀수에 출산예정자녀수를 합한 값, 마지막으로 이상 자녀수에서 현존자녀수를 뺀 값으로 다양하게 정의하였다.

그러나 본 연구에서는 선행연구와 달리 실제적으로 낳은 출생자녀수와 현재 출산예정인 경우를 합한 값을 목표변수(target variable)로 정의하였다. 현재 출산 예정인 경우는 현재 임신을 하고 출산을 계획하고 있는 경우로서 출생자녀수 값에 1명을 더 추가하였다. 이는 설문당시까지 유배우 부인이 출생자녀수를 결정하는데 영향을 미쳤던 요인들을 살펴봄으로서 그 요인들의 상호관련성을 분석하고자 한다. 다시 말하면, 자녀를 실제로 출산하기까지 어떤 요인들이 영향을 미쳤으며, 더불어 설문조사 당시에 임신을 해서 출산을 예정하기까지의 영향을 미쳤던 요인들을 알아보는 것으로, 현재 자녀의 생존 여부과 상관없이[6] 과거에 실제적으로 자녀를 출생하기로 결정하는 요인(현재 출산예정자녀포함)을 살펴보는 것이 상당히 의미 있다고 본다. 현재의 저출산의 상황에서는 현재 자녀수와 장래추가 자녀수가 포함된 개념보다 현재까지 자녀를 출산하기에 영향을 미쳤던 요인을 분석하는 것이 더욱 적합하다고 본다. 본 연구를 위해 정의한 목표 변수를 다시 정리하면 다음과 같다.

---

6) 보건사회연구원의 연구에서는 부인의 현재 자녀수와 출산예정인 자녀수 및 장래의 추가 자녀수를 합한 값으로 종속변수를 정의하였는데, 본 연구에서는 장래의 추가 자녀수를 제외하였고, 현재 자녀수에 중점을 둔 것이 아니라 실제로 자녀를 출생하는데 (현재 출산예정자녀포함) 영향을 미쳤던 요인을 파악하고자 한다.

목표변수 = 출생자녀수 + 현재출산예정

앞에서 정의한 목표변수 출생자녀수의 분포를 보면, 출생자녀수가 0인 경우는 표본 추출한 전체 6,021명에 비해 306명에 불과했다. 출생자녀수가 없는 경우는 변수에 따라 결측치가 발생하는 경우들이 빈번하게 발생하므로 적절하고 주의 깊은 변수처리가 필요하다고 생각되어진다. 출산한 자녀가 없는 이유들을 살펴보면, 결혼한 지 얼마 되지 않아서 출산자녀가 없는 경우도 포함될 것이고, 결혼한 지 충분한 시기가 지났다 할지라도 여러 가지 상황들로 인하여 자녀가 없는 경우도 발생할 수 있을 것이다. 그렇지만 본 연구에서는 더욱 중점적으로 보고자 하는 것이 출생자녀수 유무에 영향을 주는 요인을 보는 것이 아니라, 출생자녀수가 1명 이상인 연구대상자로 제한하여, 출생자녀수 2명[7]에 비해 1명을 낳은 경우와 출생자녀수 2명에 비해 3명 이상을 낳았던 경우로 목표변수를 2가지로 나누어서 각각에 영향을 미쳤던 결정요인을 세부적으로 분석하고자 하는데, 그 결과들을 중심으로 상황에 따라 효율적인 정책적인 전략을 세우는 데에 의미있는 정보를 제공하리라고 생각된다. 궁극적으로 의도하는 바는 출생자녀수를 증가시키기 위한 목적이므로, 그러한 하나의 방편으로 본 연구에서는 출생자녀수가 있는 경우로 분석대상자를 제한하여 평균 출생자녀수 2명을 기준으로 1명을 덜 낳는 경우와 1명 이상을 더 낳는 경우를 나누어서 분석하는 것은 상당한 의미가 있다고 생각되어진다.

부가적으로 목표변수인 출생자녀수[8]가 0명인 경우는 제외하고 1명이상인 대상자의 분포를 히스토그램으로 살펴보면 그림5와 같다. 출생자녀수가 1명인 경

7) 각각의 목표변수인 출생자녀수가 1명일 경우와 3명 이상인 경우를 출생자녀수 2명을 기준으로 비교한 이유는 본 연구를 위한 자료의 설문당시 분석대상자의 평균 출생자녀수가 2.03명이며, 더불어 인구대체수준이 2.01명이라는 점을 볼 때 더욱 의미가 있다고 생각되어진다.

8) 출생자녀수: 15~49세 유배우 부인의 출생아수와 현 임신 출산예정자녀수 (1명)를 합한 값으로 정의한다.

우는 1,182명 (20.71%), 2명인 경우는 3,422명 (59.96%), 3명 이상인 경우를 모두 합하면 1,103명(19.33%)을 나타내었다. 본 연구를 위한 대상자는 5,715명이였고, 이들의 평균 출생자녀수는 2.03±0.75명이였다.

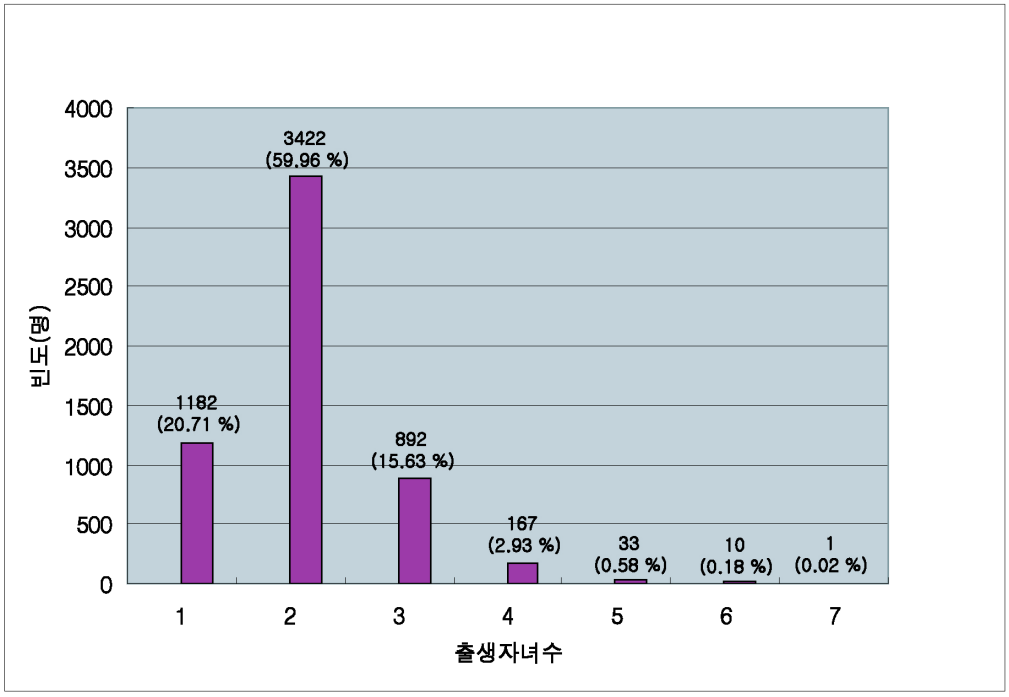


그림 5. 목표변수(target variable)인 출생자녀수의 분포

4. 분석방법 및 도구

본 연구의 분석대상자인 6021명을 중심으로 위에서 정의한 목표변수인 출생자녀수와 선정된 34개의 변수를 가지고 데이터를 세부적으로 파악하기 위하여 연속형의 변수형태로 ANOVA, t-test를 시행해 보았으며, 거의 모든 변수에서 P-value에서는 0.05이하로서의 대부분 유의하다는 결과를 나타내었으나 본 연구를 위한 의미있는 정보를 찾아내지 못했다. 그러나 범주형 변수로 변환하여 카이제곱 검정(Chi-square test)를 실시한 출생자녀수에 따른 유의미한 차이를 보이는 변수들을 볼 수 있었다.

이러한 결과를 기반으로 하여 본 연구의 분석방법은 출산력 자료에서 유배우 부인의 자녀수를 결정 요인을 파악하기 위해서 정리한 변수들을 중심으로 로지스틱 회귀분석을 실시하였다. 그리고 연구를 위한 분석자료 자체가 보여주는 자녀수에 영향을 미치는 요인을 파악하기위해 데이터마이닝 기법중의 대표적인 방법으로 의사결정나무를 도출하였다. 분류기준은 CART 알고리즘의 Gini index를 이용하였고 유의수준은 0.05로 지정하였다. 이 때 모형구축을 위해 그림 3에 제시한 바와 같이 분석용(Train) 데이터 셋을 75%, 테스트용(Test) 데이터 셋은 25%로 할당하였다. 통계분석 프로그램은 SAS 8.1 Enterprise Miner를 사용하였다.

Ⅳ. 연구 결과

1. 연구대상자의 일반적인 특성

본 연구를 위하여 '2000년도 전국 출산력 및 가족보건 실태조사 자료'를 통해 15~49세 유배우 부인 6,021명을 대상자 중에서 앞서 언급한 바와 같이 목표변수인 출생자녀수가 0인 경우를 제외하여 5715명을 대상으로 하였다. 몇 차례의 통계적 과정을 거친 후에 <표3>과 같이 최종변수들을 중심으로 정리하였고, 인구사회학적, 경제적, 산과적 및 가치관 특성별로 분류하여 범주형 변수와 연속형 변수를 구분하여 정리하였다. 연구대상자들의 특성을 살펴보면 다음의 <표4> 및 <표5>와 같다.

가. 인구사회학적, 경제적, 산과적 및 가치관 특성의 범주형 변수들

먼저 인구사회학적 특성을 살펴보면, 조사구 지역별 결과에서 전체 분석대상자의 대부분인 84.29%가 우리나라의 도시(광역시, 중소도시)에 살고 있었으며, 읍면에 거주하는 대상자는 15.71%를 차지하고 있었다. 분석대상자의 남편의 교육수준은 중졸이하가 22.20%였으며, 고졸이 43.28%, 대졸이상이 34.51%이였다. 부인의 교육수준도 남편의 교육수준의 경우처럼 고졸의 분포가 가장 많은 상태로, 중졸이하가 29.72%, 고졸이 47.93%, 대졸이상은 22.36%였다. 가구유형[9]을 보면, 2세대가 같이 사는 가구가 80.55%로 가장 많았으며, 그 다음은 3세대가 같이 사는 경우가 13.24%, 1세대만 사는 경우가 6.22%를 차지하고 있었다. 그리고 집 유형을 보면, 단독주택에 사는 대상자가 44.97%, 아파트, 연립, 기타에

9) 가구유형에서 1세대인 경우에는 부부만 같이 살고 있다거나, 미혼자녀나 형제, 자매가 같이 사는 경우가 해당된다, 2세대인 경우는 부부와 기혼자녀 혹은 미혼자녀가 같이 살고 있다거나, 편부나 편모인 경우도 여기에 해당되며, 부부와 양친, 편부모, 혹은 자녀와 부부의 형제자매가 같이 사는 경우도 여기에 해당하며, 조부모와 손자녀가 같이 사는 경우도 2세대에 포함시켰다. 그리고 3세대인 경우는 부부와 자녀(기혼, 미혼), 양친이나 편부모가 같이 산다거나 손자녀가 같이 사는 경우 등을 포함하여 분류하였다.

거주하는 대상자는 55.03%이였다. 집 소유 형태에서는 자가인 경우가 56.76%로 가장 많았으며, 전세, 월세가 각각 31.50%, 11.74%를 차지하고 있었다. 결혼 전 부인의 최장거주지를 살펴보면, 도시가 71.64%, 읍면이 28.37%를 보였다. 부인의 종교는 불교가 29.52%, 개신교와 천주교가 31.30%, 기타가 39.19%였다. 결혼 전 남편의 최장거주지는 도시가 69.36%, 읍면이 30.63%를 차지하고 있었다. 남편의 종교를 살펴보면 불교가 22.87%를 보였고, 개신교와 천주교가 27.23%, 기타가 49.90%였다.

그 다음으로는 경제적 특성을 살펴보면, 부인의 현 취업여부에서 취업을 하고 있는 부인이 45.39%이고, 비취업 부인이 54.61%이였다. 부인의 종사상의 지위는 임시고용직, 일용고용직, 자영업자인 경우가 41.53%로 가장 높았으며, 고용주, 상용고용자인 경우가 30.61%, 무급가족종사자가 27.87%이였다. 부인의 직업을 살펴보면. 고위전문직에 종사하는 대상자가 고위전문직, 사무기술직인 경우는 21.03%, 서비스판매직이 35.78%, 기타가 43.19%이였다. 기타에는 단순노무직, 1차산업종사자, 기술공 등이 이에 포함된다. 남편의 종사상의 지위를 보면, 고용주, 상용고용직에 있는 남편들이 60.58%를 보였으며, 임시고, 일용고, 자영업자 및 무급가족 종사자의 해당자가 39.42%였다. 남편의 직업은 고위전문직 및 사무기술직에는 34.95%, 서비스판매직은 17.55%, 기타인 경우는 47.50%로 나타났다. 자가용을 가지고 있는 응답자는 소유하고 있는 경우가 72.94%를 보였으며, 나머지 27.06%은 자동차를 소유하고 있지 않았다. 가구 총소득[10]을 보면, 저소득층인 경우가 24.55%, 중간소득층의 경우는 43.80%, 고소득층인 경우는 31.65%이였다. 월평균보육료[11]를 살펴보면, 100만원 미만인 경우가 24.83%,

10) 가구 총소득을 저소득층, 중간소득층, 고소득층으로 분류하였는데, 여기서 저소득층은 가구 총소득이 100만원 이하인 경우이고, 중간 소득층은 100만원 초과 200만원 이하이며, 고소득층은 200만원 이상인 경우라고 정의하였다.

11) 월평균보육료 변수는 대상자의 아이들을 돌보아주는 시설의 월평균 보육료를 말하는 것으로 자녀가 2명이상인 경우는 첫째아를 비롯하여 둘째아 및 셋째아 까지 해당하는 월평균 보육료를 모두 합한 값이다.

100만원 이상 150만원 미만인 경우가 22.18%, 150만원 이상 250만원 미만인 경우는 25.35%, 250만원 이상인 경우는 27.63%임을 보였다.

응답대상 부인의 산과적 특성을 살펴보면, 인공유산을 경험여부에서 경험을 한 경우가 77.21%로 많았고, 그 외 22.79%는 인공유산의 경험이 없었다. 반면에 자연유산은 경험하지 않은 경우가 65.57%이고, 경험자가 34.43%를 보였다. 그리고 막내아의 출생년도는 1990년대가 56.88%임을 보였고, 1980년대가 36.98%, 1950년대에서 1970년대인 경우는 6.14%의 순으로 나타났다. 최종임신 년도는 1980년대가 48.13%로 가장 많았고, 1990년대에 태어난 대상자가 41.37%, 1950년대~1970년대가 6.14% 순으로 나타났다. 첫째아 출생시 부인의 연령은 25세 미만인 경우가 49.30%로 가장 많았으며, 25세 이상 30세미만이 44.48%, 30세 이상이 6.21%를 보였다. 그리고 첫째아의 출생년도를 보면, 1990년대인 경우가 42.35%였고, 1980년대는 39.92%, 1950년대에서 1970년대인 경우가 17.73%임을 알 수 있었다. 막내아 출생시 부인의 연령은 25세 이상 30세 미만인 경우가 56.41%로 가장 많았고, 30세 이상이 25.67%, 그리고 25세 미만이 17.92%임을 나타내었다.

마지막으로 설문 대상자의 가치관 특성을 살펴봤을 때, 자녀가 필요하다에 찬성한 응답자가 60.37%였으며, 그 외가 39.63%를 보였다. 부부간의 가사역할분담에 관한 견해로 '아내가 주로 가사행위를 한다'라고 대답한 경우가 33.13%, '반반씩 나눠서 한다'고 응답한 경우가 66.87%을 차지하였다. 이상 자녀수에서는 2명이 이상적이라고 대답한 분석대상자가 62.63%로 가장 많았고, 3명인 경우는 21.44%, 1명인 경우가 8.07%, 그리고 4명 이상이 7.86%의 순으로 나타났다. 아들이 필요한가에 대한 견해에는 필요하다고 대답한 군이 64.53%였으며, '필요없다'의 경우는 35.47%이였다.

표 4. 인구사회학적, 경제적, 산과적 및 가치관 요인에 따른 범주형 변수들

구분	변수	분류	빈도수	백분율
인구사회학적 특성	조사구 지역 구분	도시	4817	84.29
		읍면	898	15.71
	남편의 교육수준	중졸이하	1268	22.20
		고졸	2472	43.28
		대졸이상	1971	34.51
	부인의 교육수준	중졸이하	1691	29.72
		고졸	2727	47.93
		대졸이상	1272	22.36
	가구유형	1세대	355	6.22
		2세대	4600	80.55
		3세대	756	13.24
	집유형	단독	2570	44.97
		아파트, 연립, 다세대, 기타	3145	55.03
	집 소유형태	자가	3133	56.76
		전세	1739	31.50
		월세	648	11.74
	결혼전 부인의 최장거주지	도시	4094	71.64
		읍면	1621	28.37
	부인의 종교	기타	2225	39.19
		개신교+천주교	1676	29.52
		불교	1777	31.30
	결혼전 남편의 최장거주지	도시	3964	69.36
		읍면	1751	30.64
	남편의 종교	기타	2826	49.90
		개신교+천주교	1542	27.23
		불교	1295	22.87
경제적 특성	부인의 현 취업 여부	예(취업)	2593	45.39
		아니오(비취업)	3120	54.61
	부인의 종사상의 지위	고용주, 상용고	793	30.61
		임시고, 일용고, 자영업자	1076	41.53
		무급가족종사자	722	27.87
	부인의 직업	고위전문직/ 사무기술직	545	21.03
		서비스판매직	927	35.78
		기타	1119	43.19
	남편의 종사상의 지위	고용주, 상용고	3240	60.58
		임시고, 일용고, 자영업자	2108	39.42
		무급가족종사자		
	남편의 직업	고위전문직 / 사무기술직	1864	34.95
		서비스판매직	936	17.55
		기타	2533	47.50
	자가용 소유여부	소유하고 있지 않음	1545	27.06
		소유하고 있음	4165	72.94
	가구 총소득	저소득층	1403	24.55
		중간소득층	2503	43.80
		고소득층	1809	31.65
	월평균보육료	100만원 미만	674	24.83
		100만원 이상 150만원 미만	602	22.18
		150만원 이상 250만원 미만	688	25.35
		250만원 이상	750	27.63

(계속)

구분	변수	분류	빈도	백분율
산과적 특성	인공유산 경험여부	무	718	22.79
		유	2433	77.21
	자연유산 경험여부	무	2066	65.57
		유	1085	34.43
	막내아의 출생년도	1950년대~1970년대	338	6.14
		1980년대	2035	36.98
		1990년대	3130	56.88
	최종임신년도	70년대	440	10.50
		80년대	2016	48.13
		90년대	1733	41.37
	첫째아 출생 시 부인의 연령	~25세미만	2713	49.30
		25세이상~30세미만	2448	44.48
		30세이상~	342	6.21
	첫째아의 출생년도	1950년대~1970년대	976	17.73
		1980년대	2197	39.92
		1990년대	2331	42.35
	막내아 출생 시 부인의 연령	~25세미만	986	17.92
		25이상~30세미만	3103	56.41
		30세 이상~	1412	25.67
가치관 특성	자녀의 필요성	찬성	3424	60.37
		그 외 (반대, 모르겠다)	2248	39.63
	부부간의 가사역할 분담에 관한 견해	아내가 주로 한다	1876	33.13
		반반씩 나눈다.	3786	66.87
	이상 자녀수	1명	457	8.07
		2명	3546	62.63
		3명	1214	21.44
		4명이상	445	7.86
	아들 필요성	무	1988	35.47
		유	3616	64.53

* 자료출처: '2000년 전국 출산력 및 가족보건 실태조사', 한국보건사회연구원

나. 인구사회학적 및 경제적 특성의 연속형 변수들

각종 일반적 특성의 변수들중 앞에서는 범주형으로 처리한 변수의 결과에 대하여 살펴보았다. 여기서는 연속형 변수들로서 부인과 남편의 만연령, 부인과 남편의 결혼 연령, 현재살고 있는 집의 평수 그리고 가구의 총소득에 대한 것이다.

본 연구를 위해 응답한 부인의 평균 만 연령은 37.03±6.78세, 남편의 평균 만 연령 보다 40.46±7.47세보다 3살가량 더 젊었다. 그리고 부인의 결혼연령이 23.61±3.40세인데 비해, 남편의 결혼연령은 29.04±3.68세로 평균 5살가량 더 높았다. 그러나 설문당시인 2000년도의 결과와 최근 통계청의 인구동태통계연보의 결과를 평균 결혼연령에 비교한다면 2002년에는 남자 29.8세, 여자 27.0세였으며 2003년에는 남자가 30.1세 여자가 27.3세인 것을 보면 평균 초혼 연령이 현저하게 더 증가한 것을 알 수 있다.

표 5. 인구사회학적 및 경제적 특성의 연속형 변수들

구분	변수목록	N	평균±표준편차
인구 사회학적 및 경제학적 특성	부인의 만 연령	5715	37.03 ± 6.78
	남편의 만 연령	5714	40.46 ± 7.47
	부인의 결혼 연령	5711	23.61 ± 3.40
	남편의 결혼 연령	5710	27.04 ± 3.68
	현재 살고 있는 집의 평수	5688	23.37 ± 19.67

다. 인구사회학적, 경제적, 산과적 및 가치관 특성에 따른 출생자녀수

여러 가지 특성에 따른 출생자녀수를 알아보기 위하며 종속변수를 범주화하였는데, 앞에서도 언급하였지만, 출생자녀수가 0명인 경우는 제외한 표본 5715명으로 chi-square test를 시행하였다. 최근의 저출산의 기조로 볼 때에 출생자녀를 1명을 낳는 경우와 2명을 낳는 경우, 그리고 3명 이상을 낳는 경우를 구분하여 파악하는 것을 상당히 의미가 있다고 본다. 각 독립변수에 대해서도 범주화하였고, 인구사회학적, 경제적, 산과적 및 가치관 특징에 따라 구분하여 정리하였다. <표6, 표7, 표8 참조>

첫 번째로 인구사회학적 특성에 따른 9개의 변수에 따른 출생자녀수를 나타내는 결과는 표6과 같다. 모든 변수의 경우에서 출생자녀수가 2명인 경우가 상당한 비중을 차지하고 있었으며, Chi-square test 결과에서 모든 변수들이 유의하게 나타났다. 조사구 지역별로 보면 도시, 읍면 모두 출생자녀수 2명을 낳는 경우가 각각 61.28%, 52.90% 으로 가장 많았으며, 도시에서는 출생자녀수가 1명인 경우가 두 번째로 21.94%였으며, 읍면에서는 3명 이상을 낳은 경우가 32.96% 으로 많은 것을 볼 수 있었다. 가구유형에서는 1세대에서는 1명을 낳는 경우가 44.07%로 가장 높았으며, 2세대와 3세대에서는 2명을 낳은 경우가 각각 62.87%, 55.50%로 가장 많았다. 그 다음으로는 1세대에서는 2명을 낳은 경우가 33.33%로 많았으며, 윗세대와 함께 거주하는 3세대에서는 3명 이상을 낳은 경우가 두 번째로 26.09%를 차지하였다. 남편의 교육수준에서도, 학력에 상관없이 출생자녀수 2명을 낳은 경우가 출생자녀수 1명과 3명 이상에 비해 가장 많았는데, 중졸이하에서는 50.95%, 고졸에서는 62.51%, 대졸이상에서는 62.68%의 비중을 차지하였다. 그 다음의 순으로는 중졸이하에서는 3명 이상이 37.60%, 대졸이상에서는 1명을 낳은 경우가 26.89%로 많았다. 부인의 교육수준에서도 2명을 낳은 경우가 가장 높으며, 중졸이하 52.90%, 고졸 64.74%, 그리고 대졸이상이 59.40%로 나타났다. 중졸이하에서는 출생자녀수가 3명 이상인 경우가 두번째로 36.02%로 높았고, 대졸이상에서는 출생자녀수 1명인 경우가 32.65%로 그 다음으로 높았다.

표 6. 인구사회학적 변수에 따른 출생자녀수 단위: %

변수	분류	출생자녀수			χ^2값	P값
		1명	2명	3명 이상		
조사구 지역 구분	도시	1055 (21.94)	2947 (61.28)	807 (16.78)	133.59	<.0001
	군읍명	127 (14.14)	475 (52.90)	296 (32.96)		
가구유형	1세대	156 (44.07)	118 (33.33)	80 (22.60)	174.58	<.0001
	2세대	886 (19.29)	2884 (62.78)	824 (17.94)		
	3세대	139 (18.41)	419 (55.50)	197 (26.09)		
남편의 교육수준	중졸이하	145 (11.45)	645 (50.95)	476 (37.60)	417.29	<.0001
	고졸	505 (20.45)	1544 (62.51)	421 (17.04)		
	대졸이상	529 (26.89)	1233 (62.68)	205 (10.42)		
부인의 교육수준	중졸이하	187 (11.08)	893 (52.90)	608 (36.02)	552.55	<.0001
	고졸	571 (20.97)	1763 (64.74)	389 (14.29)		
	대졸이상	415 (32.65)	755 (59.40)	101 (7.95)		
부인의 종교	기타	520 (23.40)	1340 (60.31)	362 (16.29)	53.39	<.0001
	불교	286 (17.10)	979 (58.52)	408 (24.39)		
	개신교/천주교	367 (20.68)	1085 (61.13)	323 (18.20)		
남편의 종교	기타	612 (21.68)	1702 (60.29)	509 (18.03)	23.00	<.0001
	불교	274 (17.80)	911 (59.19)	354 (23.00)		
	개신교/천주교	280 (21.66)	782 (60.48)	231 (17.87)		
집유형	단독	481 (18.76)	1413 (55.11)	670 (26.13)	138.36	<.0001
	.아파트/연립/					
	다세대/기타	701 (22.30)	2009 (63.92)	433 (13.78)		
결혼전 부인의 최장 거주지	도시	967 (23.66)	2551 (62.42)	569 (13.92)	292.57	<.0001
	읍면	215 (13.27)	871 (53.77)	534 (32.96)		
결혼전 남편의 최장 거주지	도시	944 (23.85)	2496 (63.06)	518 (13.09)	342.32	<.0001
	읍면	238 (13.61)	926 (52.94)	585 (33.45)		

부인의 종교에서도 출생자녀수 2명인 경우가 가장 높은 결과를 보였는데 기타(무교, 기타종교)가 60.31%, 불교에서는 58.52%, 개신교와 천주교에서는 61.13%를 보였으며, 그다음 순으로 기타에서는 출생자녀수 1명의 경우가 23.40%를 나타냈고, 불교에서는 3명 이상이 24.39%로 높은 것으로 나타났다. 남편의 종교에서도 비슷한 결과를 보였는데, 기타에서 60.29%, 불교에서는 59.19%, 개신교와 천주교에서는 60.48%가 출생자녀수 2명인 경우로 가장 많았고, 두 번째로는 불교에서 출생자녀수가 3명 이상인 경우가 23.00%를 차지하는 것을 볼 수 있었다. 집 유형에서도 역시 출생자녀수 2명인 경우가 가장 많이 나타났는데, 단독주택인 경우 55.11%, 아파트·연립·다세대등에서는 63.92%이였으며, 단독에서는 3명 이상인 경우도 26.13%, 아파트·연립·다세대등에서는 출생자녀수 1명인 경우가 22.30%로 그 다음으로 높은 것을 볼 수 있었다. 결혼 전 부인의 최장거주지가 도시와 읍면이 각각 62.42%, 53.77%로 출생자녀수 2명을 가지는 경우가 가장 많았으며, 도시에서는 출생자녀수 1명인 경우는 23.66%, 읍면에서 출생자녀수 3명 이상인 경우가 32.96%로 그 다음 순이였다. 결혼 전 부인의 최장거주지에서는 도시와 읍면에서 각각 62.42%, 53.77%로 출생자녀수 2명인 경우가 가장 많았고, 도시에서는 출생자녀수가 1명인 경우는 23.66%, 읍면에서 출생자녀수 3명 이상인 경우가 32.96%로 두 번째로 비중을 차지하였다. 결혼 전 남편의 최장거주지에서도 도시와 읍면에서 각각 63.06%, 52.94%로 출생자녀수 2명인 경우가 가장 많았고, 도시에서는 출생자녀수 1명인 경우는 23.85%, 읍면에서 출생자녀수 3명 이상인 경우가 33.45%였다.

두 번째로 경제적 변수에 따른 출생자녀수를 살펴보면<표7>, 집 소유형태가 자가, 전세 그리고 월세인 경우에도 출생자녀수가 2명인 경우가 가장 많았는데 각각 62.8%, 58.55%, 54.56%를 나타내었다. 그 다음 순으로는 자가인 경우는 출생자녀수 3명 이상이 24.26%로 많았고, 전세와 월세인 경우 29.99%, 30.29%로를 차지하였다. 가구 총소득에서도 100만원 미만, 100만원 이상 200만원 미만, 그리고 200만원 이상의 모든 범주에서도 마찬가지로 출생자녀수가 2명인 경우가 가장 많았다. 각각 52.86%, 60.95%, 64.08%를 나타내었으며, 그 다음으로는

표 7. 경제적 변수에 따른 출생 자녀수 단위: %

변수	분류	출생자녀수			χ^2값	P값
		1명	2명	3명 이상		
집 소유 형태	자가	427 (13.65)	1942 (62.08)	759 (24.26)	283.71	<.0001
	전세	521 (29.99)	1017 (58.55)	199 (11.46)		
	월세	196 (30.29)	353 (54.56)	98 (15.15)		
가구 총소득 (만원)	100 만원미만	315 (22.53)	739 (52.86)	344 (24.61)	49.82	<.0001
	100이상 200만원미만	526 (21.02)	1525 (60.95)	451 (18.03)		
	200만원 이상	341 (18.87)	1158 (64.08)	308 (17.04)		
남편의 취업여부	예(취업)	1101 (20.59)	3240 (60.61)	1005(18.80)	19.58	<.0001
	아니오(비취업)	80 (22.60)	177 (50.00)	97 (27.40)		
부인의 취업여부	예(취업)	428 (16.54)	1529 (59.08)	631 (24.38)	103.50	<.0001
	아니오(비취업)	754 (24.19)	1892 (60.70)	471 (15.11)		
부인의 직업지위	고용주.상용고	196 (24.75)	488 (61.62)	108 (13.64)	140.95	<.0001
	임시고,일용고.자영업자	157 (14.63)	658 (61.32)	258 (24.04)		
	무급가족종사자	74 (10.26)	382 (52.98)	265 (36.75)		
부인직업	고위전문직,사무기술직	168 (30.88)	327 (60.11)	49 (9.01)	220.99	<.0001
	서비스판매직	147 (15.89)	593 (64.11)	185 (20.00)		
	기타	113 (10.12)	607 (54.34)	397 (35.54)		
남편의 직업지위	고용주.상용고	785 (24.25)	2020 (62.40)	432 (13.35)	188.38	<.0001
	임시고,일용고.자영업자					
	또는 무급가족종사자	315 (14.96)	1217 (57.81)	573 (27.22)		
남편직업	고위전문직,사무기술직	409 (24.64)	1180 (63.34)	224 (12.02)	119.38	<.0001
	서비스판매직	205 (21.93)	562 (60.11)	168 (17.97)		
	기타	432 (17.08)	1486 (58.74)	612 (24.19)		
월평균 보육료 (만원)	100만원 미만	157 (23.33)	391 (58.10)	125 (18.57)	59.64	<.0001
	100이상 150만원 미만	100 (16.61)	413 (68.60)	89 (14.78)		
	150이상 250만원미만	93 (13.52)	487 (70.78)	108 (15.70)		
	250만원 이상	76 (10.13)	554 (73.87)	120 (16.00)		

100만원 미만에서 출생자녀수가 3명 이상인 경우가 24.61%, 100만원 이상 200만원 미만인 경우 출생자녀수가 1명인 경우가 21.02%를 나타내었다. 남편의 취업여부에서는 취업여무와 상관없이 출생자녀수가 2명인 경우가 가장 많았는데, 각각 60.61%, 50.00%를 보였다. 그 다음 순으로는 남편이 취업했을 때 출생자녀수가 1명인 경우는 20.59%, 비취업 상태인 경우 출생자녀수가 3명 이상이 27.40%를 나타내었다. 부인의 취업여부를 살펴보면, 마찬가지로 취업여부와 상관없이 출생자녀수 2명이 각각 59.08%, 60.70%로 가장 많았다. 부인이 취업한 경우 출생자녀수가 3명 이상인 경우가 24.38%, 비취업인 경우 출생자녀수 1명인 경우가 24.19%를 보였다. 부인직업의 지위를 보면 고용주나 상용고용직인 경우는 출생자녀수가 2명인 경우 61.62%로 가장 많았으며, 출생자녀수 1명인 경우가 24.75%를 보였다. 임시고용직, 일용고용직 및 자영업자일 때 출생자녀수가 2명인 경우 61.32%, 3명 이상인 경우 24.04%를 나타내었으며, 무급가족종사자일 때는 출생자녀수 2명이 52.98%, 3명 이상이 36.75%를 차지하였다. 부인의 직업에서는 고위전문직 혹은 사무기술직에 종사하는 경우 출생자녀수 2명이 60.11%, 1명이 30.88%를 나타내었다. 그리고 서비스 판매직에서는 출생자녀수가 2명인 경우가 64.11%로 가장 많았고, 3명 이상인 경우가 20.00%로 그 다음 순 이였다. 그러나 기타의 직업에서는 출생자녀수 2명이 54.34%, 3명 이상에서 35.54%의 순으로 나타났다. 남편직업의 지위를 보면 고용주나 상용고용직인 경우는 출생자녀수가 2명인 경우 역시 62.40%로 가장 많았으며, 출생자녀수 1명인 경우가 24.25%를 보였다. 그리고 임시고용직, 일용고용직. 또는 자영업자등에서는 출생자녀수가 2명인 경우가 57.81%로 가장 많았고, 3명 이상인 경우가 27.22%로 그 다음 순 이였다. 남편의 직업에서도 직업구분에 상관없이 출생자녀수 2명이 가장 많았으며, 고위전문직 혹은 사무기술직이 63.34%, 서비스 판매직이 60.11%, 기타인 경우 58.74%를 보였다. 그리고 그 다음 순으로는 고위전문직, 사문기술직에서는 출생자녀수가 1명인 경우 24.64%를 보였고, 기타에서는 출생자녀수 3명 이상일 때 24.19%를 나타내었다. 또 월평균 보육료가 100만원 미만인 경우는 출생자녀수 2인 경우가 58.10%, 출생자녀수가 1명인 경우 23.33%이였고, 100만원이상 150만원미만인 경우 출생자녀수 2명이 68.60%였

으며, 150만원에서 250만원인 경우는 출생자녀수2명이 70.78%, 250만원 이상에서는 출생자녀수가 2명인 경우는 73.87%, 출생자녀수가 3명 이상인 경우는 16%임을 보였다.

산과적 특성 및 가치관 변수에 따른 출생자녀수를 살펴보면<표8>, 먼저 최종임신연도에서도 70년대, 80년대, 그리고 90년대와 상관없이 출생자녀수가 2명인 경우가 가장 많았는데, 각각 53.86%, 64.73%, 그리고 67.17%를 보였다. 그 다음으로는 70년대와 80년대에서는 출생자녀수가 3명 이상인 경우가 많았는데 각각 38.64%, 23.26%를 나타내었다. 반면에 90년대에서는 출생자녀수 1명인 경우가 16.21%로 3명 이상인 16.22%와 비슷하였다. 첫째아 출생시 부인의 연령을 보면, 25세 미만일 때 첫째아를 출생했을 경우는 출생자녀수가 2명인 경우가 59.87%, 그 다음으로는 3명 이상인 경우가 28.77%로 많았으며, 25세 이상 30세 미만일 때 첫째아를 낳은 경우는 출생자녀수가 2명일 때 66.03%, 1명일 때 22.20%를 나타내었다. 그리고 첫째아를 30세 이상일 때 낳은 경우는 출생자녀수가 1명인 경우가 53.51%로 가장 많았고, 그 다음으로 출생자녀수가 2명이 42.98%임을 보였다. 막내아 출생시 부인의 연령에서는, 25세 미만일 때 막내아를 출생했을 경우는 출생자녀수가 2명인 경우가 58.07%, 그 다음으로는 출생자녀수가 1명인 경우가 31.07%로 많았으며, 25세 이상 30세 미만일 때 막내아를 낳은 경우는 출생자녀수가 2명일 때 65.99%, 1명일 때 17.54%를 나타내었다. 그리고 막내아를 30세 이상일 때 낳은 경우는 출생자녀수가 2명인 경우가 54.26%로 가장 많았고, 그 다음으로 출생자녀수가 3명 이상이 32.77%임을 보였다. 첫째아 출생연도가 1950년대에서 1970년대인 경우에는 출생자녀수가 3명 이상일 때가 50.56%로 가장 많았고 그 다음으로는 출생자녀수가 2명이 47.08%였으며, 1980년대에는 출생자녀수가 2명일 때 69.98%, 그 다음으로는 출생자녀수 3명 이상에서 18.22%를 보였고, 1990년대는 출생자녀수 2명이 59.70%, 1명이 32.27%였다. 막내아가 1950년대에서 1970년대에 출생한 경우에는 출생자녀수가 2명일 때가 61.42%로 가장 많았고 그 다음으로는 출생자녀수 3명 이상이 31.75%였으며, 1980년대에는 출생자녀수가 2명일 때 62.52%, 그 다음으로는

출생자녀수 3명 이상에서 24.84%를 보였고, 1990년대는 출생자녀수 2명이 60.98%, 1명인 경우가 24.07%였다. 또한 인공유산 경험여부에서도 인공유산 경험여부와 상관없이 출생자녀수가 2명인 경우가 가장 많았으며, 인공유산 경험이 있었던 경우는 66.42%로 가장 많았으며, 그 다음으로는 출생자녀수 3명이상인 경우에서 21.74%를 보였다. 자연유산의 경험여부에서도 자연유산 경험이 있든지, 없든지 각각 출생자녀수 2명이 58.99%, 67.13%로 가장 많았으며, 그 다음으로는 자연유산의 경험이 있는 경우에는 출생자녀수가 3명 이상인 경우 21.94%, 자연유산의 경험이 없는 경우도 출생자녀수가 3명 이상인 경우 21.25%, 이였다. 가치관 특성을 나타내는 질문중 자녀의 필요성에서는 자녀가 꼭 필요하다고 대답한 경우 출생자녀수가 2명인 경우가 58.59%로 가장 많았으며, 출생자녀수가 3명 이상인 경우 21.67%로 그 다음 순 이였다. 그 외(반대, 모름 등)인 경우는 출생자녀수 2명이 62.06%로 가장 많았고, 그 다음으로는 출생자녀수 1명이 22.11%로 많았다. 아들의 필요성에서는 '아들이 필요하다'고 대답한 경우 출생자녀수가 2명이 60.43%, 출생자녀수가 3명 이상이 23.23%였다. 그리고 '아들이 필요하지 않다'고 대답한 경우에는 출생자녀수가 2명인 경우 59.21%, 출생자녀수가 1명인 경우 28.27%를 보였다. 이상 자녀수를 물어보는 질문에서는 이상 자녀수가 1명이라고 대답한 대상자들은 출생자녀수가 2명인 경우가 48.40%, 1명인 경우 37.54%였으며, 이상 자녀수가 2명인 경우는 출생자녀수가 2명인 경우는 62.15%, 출생자녀수가 1명일 때는 22.76%였다. 이상 자녀수가 3명일 때는 출생자녀수 2명이 58.40%, 출생자녀수 3명 이상이 27.59%를 나타내었고, 이상 자녀수가 4명 이상일 경우는 출생자녀수가 2명이 58.20%, 출생자녀수가 3명 이상이 36.85%를 나타내었다. 지금까지 설명한 인구사회학적, 경제적, 산과적 및 가치관 특징을 가진 모든 변수에서 유의한 결과를 나타난 것을 알 수 있다.

표 8. 산과적 특성 및 가치관 변수에 따른 출생자녀수 단위: %

변수	분류	출생자녀수			X^2값	P값
		1명	2명	3명 이상		
최종 임신연도	70년대	33 (7.50)	237 (53.86)	170 (38.64)	11.55	<.0001
	80년대	242 (12.00)	1305 (64.73)	469 (23.26)		
	90년대	281 (16.21)	1164 (67.17)	288 (16.62)		
첫째아 출생시 부인의 연령	~25세미만	308 (11.36)	1623 (59.87)	780 (28.77)	579.76	<.0001
	25세이상~30세미만	543 (22.20)	1615 (66.03)	288 (11.77)		
	30세이상~	183 (53.51)	147 (42.98)	12 (3.51)		
막내아 출생시 부인의 연령	~25세미만	306 (31.07)	572 (58.07)	107 (10.86)	309.08	<.0001
	25이상~30세미만	544 (17.54)	2047 (65.99)	511 (16.47)		
	30세 이상~	183 (12.98)	765 (54.26)	462 (32.77)		
첫째아 출생연도	1950년대~1970년대	23 (2.36)	459 (47.08)	493 (50.56)	1119.35	<.0001
	1980년대	259 (11.80)	1536 (69.98)	400 (18.22)		
	1990년대	752 (32.27)	1391 (59.70)	187 (8.03)		
막내아 출생연도	1950년대~1970년대	23 (6.82)	207 (61.42)	107 (31.75)	201.52	<.0001
	1980년대	257 (12.64)	1271 (62.52)	505 (24.84)		
	1990년대	753 (24.07)	1908 (60.98)	468 (14.96)		
인공유산 경험여부	무	159 (22.14)	411 (57.24)	148 (20.61)	49.11	<.0001
	유	288 (11.84)	1616 (66.42)	529 (21.74)		
자연유산 경험여부	무	240 (11.62)	1387 (67.13)	439 (21.25)	34.42	<.0001
	유	207 (19.08)	640 (58.99)	238 (21.94)		
자녀의 필요성	꼭 필요하다	676 (19.74)	2006 (58.59)	742 (21.67)	30.26	<.0001
	그 외(반대, 모름,)	497 (22.11)	1395 (62.06)	356 (15.84)		
아들 필요성	무	562 (28.27)	1177 (59.21)	249 (12.53)	164.63	<.0001
	유	591 (16.34)	2185 (60.43)	840 (23.23)		
이상 자녀수	1명	172 (37.64)	223 (48.80)	62 (13.57)	317.70	<.0001
	2명	807 (22.76)	2204 (62.15)	535 (15.09)		
	3명	170 (14.00)	709 (58.40)	335 (27.59)		
	4명이상	22 (4.94)	259 (58.20)	164 (36.85)		

2. 로지스틱 회귀모형을 통한 출산관련변수와 출생자녀수

출생자녀수 결정에 영향을 미치는 요인을 알아보기 위하여 로지스틱 회귀분석을 이용하였다. 연구대상자는 자녀를 1명이상 출생한 총 5715명의 샘플을 가지고 평균 출생자녀수는 2명이므로 2명을 기준으로 1명을 덜 낳게 되는 요인과 2명을 기준으로 3명 이상을 더 낳게 되는 요인을 로지스틱 회귀분석을 이용하여 살펴보았다. 출생자녀수가 2명에 비해 1명을 덜 낳게 되는 결정요인과 2명에 비해 3명 이상을 더 낳게 되는 결정요인이 다르므로 각각의 로지스틱 분석을 통하여 사용될 변수를 선정하기 위해서 단계별 선정(Stepwise Selection)을 실시하였고, 그 변수들을 기준으로 본 연구자가 판단하기에 더욱 적합하다고 생각되는 변수를 추가하였고, 단계별 회귀분석 결과에서 통계학적으로 유의한 변수라고 선별되었다 할지라도 논리에 맞지 않는 변수들은 제외시켰다.

가. 출생자녀수 2명에 비해 1명을 낳게 한 결정요인

먼저 출생자녀수 2명에 비해 1명을 낳게 한 결정요인을 살펴보기 위하여, 다른 요인을 모두 통제한 상태에서 거주지, 부인의 교육수준, 자연유산의 경험유무, 첫째아 출생시 부인의 연령, 부인의 현 취업유무, 월평균 보육료에서 유의하게 영향을 미치는 것으로 나타났다. 도시에 비해 읍면에 거주하는 사람들이 출생자녀수 2명보다 1명을 낳을 확률이 0.53배로 통계적으로 유의하게 감소하였다(95% CI: 0.290~0.965). 이는 비도시에 사는 사람들이 자녀를 1명 출산하는 것 보다는 2명 출산할 확률이 1.89(=0.53^{-1})배로 유의하게 높음을 의미한다. 부인의 교육수준에서는 출생자녀수 2명보다 1명을 낳을 확률이 중졸이하인 경우에 비해 고졸인 경우가 2.122배이고, 대졸이상인 경우가 2.078배였다. 그리고 95% 신뢰구간이 각각 (1.19~3.79), (1.05~4.13)으로 통계적으로 유의한 것을 알 수 있다. 즉, 부인의 교육수준이 높을수록 저출산할 가능성이 높다. 자연유산 경험유무를 보면 자연유산의 경험이 있는 경우 경험이 없는 경우보다 출생자녀수 1명을 낳을 확률이 1.84배로 유의한 결과를 보였다.(95% CI: 1.29~2.62). 첫째아 출생시 부인의 연령에서는 25세 미만인 경우보다 25세 이상 30세 미만, 30세 이상인 경우가 출생자

녀수 2명에 비해 1명을 낳는 경우가 각각 1.42배, 6.94배 높았다. 그러나 95% 신뢰구간에서는 각각 (0.94~2.16), (3.90~12.34)으로 30세 이상인 경우만 통계적으로 유의한 결과를 보였다. 부인의 현 취업유무에서는 취업을 하지 않은 경우가 취업을 한 경우보다 출생자녀수 2명보다 1명을 낳을 확률이 0.41배로 유의한 결과를 나타내었다(95% CI:0.282~0.598). 이는 부인이 취업을 하지 않은 경우가 자녀를 1명 출산하기 보다는 2명을 출산할 확률이 2.44(0.41^{-1})배로 유의하게 높음을 의미했다. 월평균 보육료를 살펴보면, 월평균보육료를 250만원 이상을 소비하는 경우보다 150만원 이상 250만원미만, 100만원 이상 150만원 미만 및 100만원 미만인 경우가 출생자녀수 2명보다 1명을 낳을 확률이 각각 4.38배, 3.00배, 1.95배 더 높았다. 그리고 95%의 신뢰구간이 각각 (2.526~7.592), (1.72~5.212), (1.11~3.40)으로 통계적으로 유의하게 높았다. 다시 말하면, 월평균보육료가 적은 구간으로 갈수록 저출산할 가능성이 높아짐을 보였다. 동시에 월평균보육료가 많은 집단에서는 대체로 출생자녀수도 많다고 결과가 나온 것을 볼 수 있는데, 이는 소득과도 밀접한 관련된 변수이며 소득이 높을수록 보육료도 또한 비례한다는 것을 나타낸다. 소득이 높을수록 자녀교육의 질을 중요시하기 때문에 월평균보육료가 상승하는 현상이 생긴다.

그리고 그 외의 다른 특성인 부부간 역할분담, 집유형, 남편의 직업 및 자녀의 필요성의 변수는 95%의 신뢰구간에서 출생자녀수 2명에 비해 1명을 낳는 확률에 의미적으로 영향을 주지 못했다.

나. 출생자녀수 2명에 비해 3명 이상을 낳게 한 결정요인

다음으로는 출생자녀수 2명에 비해 3명 이상을 낳게 한 결정요인을 살펴보도록 하겠다. 다른 요인을 모두 통제한 상태에서 거주지, 부부간 역할 분담, 남편의 교육수준, 첫째아 출생연도, 자녀의 필요성, 월평균 보육료에서 유의하게 영향을 미치는 것으로 나타났다. 앞에서 분석한 출생자녀수 2명에 비해 1명을 낳게 한 결정요인과 공통되게 유의한 영향을 나타낸 변수는 거주지 변수와 월평균 보육료에 관한 변수 2가지였다.

거주지별로 볼 때 읍면에 거주하는 경우 도시에 거주하는 것에 비해 출생자녀수 2명에 비해 3명 이상을 낳을 확률이 2.47배로 통계적으로 유의하게 증가했다(95% CI: 1.69~3.62). 부부간 역할분담에서는 '아내가 주로 한다'에 비해서 '반반씩 나눈다'인 경우가 출생자녀수를 3명 이상 낳은 확률이 0.69배로 통계적으로 유의하게 낮았다(95% CI: 0.51~0.92). 즉 가사분담을 하는 가구는 그렇지 않은 가구에 비해 자녀를 3명이상 출산할 확률이 낮음을 알 수 있다. 남편의 교육수준에서는 출생자녀수 2명보다 3명 이상을 낳을 확률이 중졸이하인 경우에 비해 고졸인 경우가 0.74배, 대졸 이상인 경우가 0.53배로 낮은 결과를 보였다. 그리고 95%의 신뢰구간이 각각 (0.47~1.16), (0.31~0.92)으로 남편의 교육수준이 대졸이상인 경우에서 통계적으로 유의함을 알 수 있다 즉, 남편의 교육수준이 높을수록 고출산할 가능성이 낮아진다고 볼 수 있다. 첫째아 출생연도는 1950년대에서 1970년대인 경우에 비해 1980년대, 1990년대인 경우가 출생자녀수 2명보다 3명 이상을 낳은 확률이 각각 0.09배, 0.04배 정도나 두드러지게 낮았다. 그리고 95%의 신뢰구간이 각각 (0.03~0.34), (0.01~0.14)으로 통계적으로 유의함을 알 수 있다. 즉, 첫째아 출생연도가 1980년대 이후에는 자녀를 3명 이상을 낳을 가능성은 그 전에 비해 현저히 낮음을 알 수 있다. 자녀의 필요성에서는 찬성한 경우에 비해 그 외(반대, 모르겠음)인 경우가 출생자녀수 2명보다 3명 이상을 낳은 확률이 0.62배로 유의하게 낮은 것을 알 수 있다(95% CI: 0.45~0.84). 월평균 보육료를 살펴보면, 월평균보육료를 250만원 이상을 소비하는 경우보다 150만원 이상 250만원 미만, 100만원 이상 150만원 미만 및 100만원 미만인 경우가 출생자녀수 2명에 비해 3명 이상을 낳을 확률이 각각 0.93배, 0.57배, 0.90배 정도 낮게 나타났다. 여기서는 95%신뢰구간이 각각 (0.63~1.38), (0.36~0.89), (0.60~1.36)으로 100만원 이상 150만원미만인 경우에서 통계적으로 유의한 것을 볼 수 있다.

그리고 그 외의 변수들인 부인과 남편의 최장거주지, 자연유산 경험유무, 남편의 직업 및 자동차 소유여부의 변수들에서는 95%의 신뢰구간에서 출생자녀수 2명에 비해 3명 이상을 낳는 확률에 유의미하게 영향을 주지 못한 것으로 나타났다.

표 9. 로지스틱 회귀모형을 통한 출산관련변수와 출생자녀수

변수	분류	모형I †		모형II §	
		OR	CI 95%	OR	CI 95%
조사구 지역구분	도시	1.00		1.00	
	읍면	0.53	0.29-0.97	2.47	1.69-3.62
부부간 역할분담	아내가 주로 한다	1.00		1.00	
	반반씩 나눈다	1.00	0.68-1.47	0.69	0.51-0.92
부인 최장거주지	도시	-	-	1.00	
	읍면	-	-	1.39	0.95-2.00
남편 최장거주지	도시	-	-	1.00	
	읍면	-	-	1.33	0.90-2.96
부인의 교육수준	중졸이하	1		-	-
	고졸	2.12	1.19-3.79	-	-
	대졸이상	2.08	1.05-4.13	-	-
남편의 교육수준	중졸이하	-	-	1.00	
	고졸	-	-	0.74	0.47-1.16
	대졸이상	-	-	0.53	0.31-0.92
짐유형	단독	1.00		-	-
	아파트,연립,다세대,기타	1.16	0.78-1.72	-	-
자연유산경험유무	무	1.00		1.00	
	유	1.84	1.29-2.62	1.33	0.99-1.79
첫째아 출생시	25세미만	1.00		-	-
부인의 연령	25세 이상 30세미만	1.42	0.94-2.16	-	-
	30세 이상	6.94	3.90-12.34	-	-
첫째아 출생연도	1950-1970년대	-	-	1.00	
	1980년대	-	-	0.09	0.03-0.34
	1990년대	-	-	0.04	0.01-0.14
남편의 직업	고위전문직 ,사무기술직	1.00		1.00	
	서비스판매직	1.05	0.63-1.77	1.12	0.55-1.25
	기타	1.14	0.73-1.81	0.83	0.59-1.36
부인 현 취업유무	예(취업)	1.00		-	-
	아니오(비취업)	0.41	0.28-0.60	-	-
자동차 소유여부	소유하고 있지 않음	-	-	1.00	
	소유하고 있음	-	-	0.98	0.67-1.44
자녀의 필요성	찬성	1.00		1.00	
	그외	0.84	0.59-1.21	0.62	0.45-0.84
월평균보육료	250만원이상	1.00		1	
	150만원이상 250만원미만	4.38	2.53-7.59	0.92	0.63-1.38
	100만원이상 150만원미만	3.00	1.72-5.21	0.57	0.36-0.89
	100만원미만	1.95	1.11-3.41	0.90	0.59-1.36

† 출생자녀수 2명에 비해 1명을 적게 낳게 되는 결정요인을 분석하기 위한 로지스틱 회귀분석 모형

§ 출생자녀수 2명에 비해 1명이상을 더 낳게 되는 결정요인을 분석하기 위한 로지스틱 회귀분석 모형

3. 출생자녀수에 따른 의사결정나무 모형

본 연구에서 중점적으로 보고자 하는 것은 출생자녀수 2명을 중심으로 하여 1명을 덜 낳는 경우(1명)와 출생자녀수 2명을 중심으로 하여 1명이상을 더 낳는 경우(3명 이상)로 로지스틱 회귀분석에서 결과를 분석한 것처럼, 본 장에서도 위의 두 가지 경우로 나누어서 의사결정나무 분석을 실행하였다. 어떤 특징을 가진 집단에 따라서 출생자녀수에 따른 확률을 나타내고 있다. 의사결정 나무 모형구축을 위하여 CART 알고리즘을 이용하였으며, 효율적으로 모형구축을 하기 위해서 분석용 데이터 셋(train data set) 75%, 테스트 데이터 셋(test data set) 25% 으로 자료를 분할하였다. CART 분석을 이용한 의사결정나무의 분석 흐름도는 다음 그림과 같다.

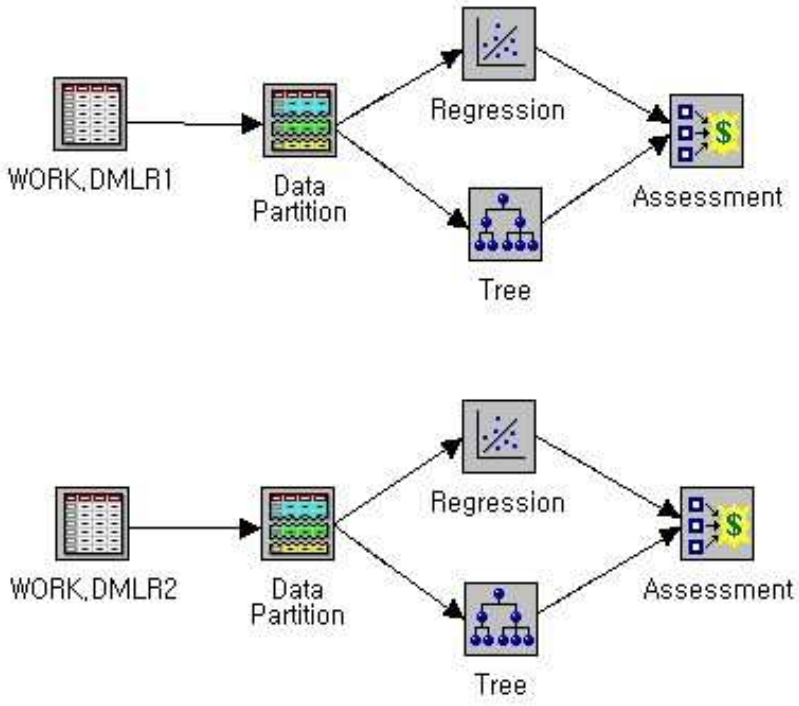


그림 6. CART 분석을 이용한 의사결정나무의 분석흐름도

가. 출생자녀수 2명에 비해 1명을 낳을 경우의 의사결정나무

출생자녀수 1명과 2명을 결정하는 변수는 첫째아 출생시 부인의 연령, 가구유형, 부인의 직업, 첫째아 출생년도, 월평균보육료, 최종임신연도에 의해서 의사결정나무 구조를 나타내었다. 처음 시작할 때 노드 (Root node)의 분석용 데이터를 살펴보면, 출생자녀수가 1명인 경우와 2명인 경우의 분석대상자 전체인원은 3452명이였다. 그중 자녀수가 1명인 경우는 25.9%이고, 자녀수 2명인 경우는 74.1%를 이루었다. 그림7의 의사결정 나무에서는 최종 종단마디 (leaf node)는 모두 7개이고, 이들 중에는 2번째 노드에서 가지 친 결과가 종단마디를 이룬 것이 하나 포함된다. 이들 최종 종단마디에 대하여 왼쪽부터 번호를 붙여서 각각 이름을 붙여 살펴보면 다음과 같다.

Node1은 첫째아 출생시 부인의 연령이 30세 이하이고, 가구유형이 1세대인 경우이며, 부인의 직업이 고위전문직이거나 사무기술직인 특성을 가진 집단인데, 이는 출생자녀수가 1명이 경우가 96.6%, 2명인 경우가 3.4%임을 알 수가 있다. 종단마디의 결과에서 출생자녀수 2명을 낳는 것보다 1명을 가장 많이 낳은 집단이였다. 즉, 이는 자녀수를 1명을 출산하는 것보다 2명을 낳을 가능성이 가장 적은 경우였다. Node2는 node1과 같은 특성의 패턴을 보이다가 마지막 노드인 부인의 직업에서 서비스판매직에 종사하는 군에서는 출산자녀수를 1명을 낳은 확률이 56.2%, 2명을 낳을 확률은 43.8%의 결과를 보인다. Node3은 출생시 부인의 연령이 30세 이하이고, 가구유형이 2세대 혹은 3세대가 같이 거주하는 경우이고, 첫째아 출생년도가 1950년대에서 1980년대인 특성을 가진 집단으로 출생자녀수가 1명인 경우가 11.4%, 2명인 경우는 88.6%였고, 이는 출생자녀수가 2명인 경우보다 1명을 낳은 가능성이 전체에서 가장 낮았으며, 달리 말하면 출생자녀수를 1명을 낳은 경우보다 2명을 낳는 가능성이 가장 큰 집단이였음을 의미했다. Node4는 node3과 같은 특성을 보이다가 마지막 노드인 첫째아 출생년도가 1990년대인 군으로 출생자녀수가 1명인 경우는 31.4%, 2명인 경우는 68.6%의 결과를 보였다. Node5는 첫째아 출생시 부인의 연령이 30세 이상인 경우이며, 월평균보육료가 100만원 미만인 분류군으로 출생자녀수가 1명인 경우가 66.0%, 2명인 경

우는 34.0%임을 알 수가 있다. Node6은 Node5과 같은 특성을 보이다가 월평균보육료가 100만원 이상인 경우이며, 최종 임신년도가 70년대, 80년대인 특성의 집단으로 출생자녀수가 1명인 경우는 66.7%, 2명인 경우는 33.3%였다. Node7은 첫째아 출생시 부인의 연령이 30세 이상이면서, 월평균보육료가 100만원 이상이면서 최종 임신년도가 1990년대인 특성의 집단으로 출생자녀수 1명인 경우, 2명인 경우 모두 50.0%임을 알 수가 있었다.

최종 7개의 terminal node 중에서 출생자녀수 1명을 가장 많이 낳은 순서대로 정리해보면, node 1이 96.6%, node5에서 66.7%, node2가 56.2%, node6이 50.0%, node4가 31.4%, 그리고 node3이 11.4%인 순으로 나타났다. 이러한 결과에서 의미있는 규칙을 생성한 node를 살펴보면, node3에서는 출생자녀수가 1명이였던 경우가 Root node에서 25.9%였던 것이 첫째아 출생시 부인의 연령이 1980년대 이전으로 분류되면서 출생자녀수를 1명을 낳은 확률이 23.5%였고, 또 다시 여기서 가구유형이 2세대 혹은 3세대가 같이 사는 경우에서는 출생자녀수 1명을 낳은 경우가 20.9%를 차지하였고, 마지막으로 첫째아 출생년도가 1990년대 이전인 경우에서는 11.4%가 출생자녀수 1명을 차지하고 있었다. 이는 node3의 종단마디로 갈수록 출생자녀수 1명을 가진 확률이 음의 방향으로 낮아지는 패턴을 보였다. 즉 이는 출생자녀수 2명을 낳은 경우가 증가하는 패턴이기도 한 것이다.

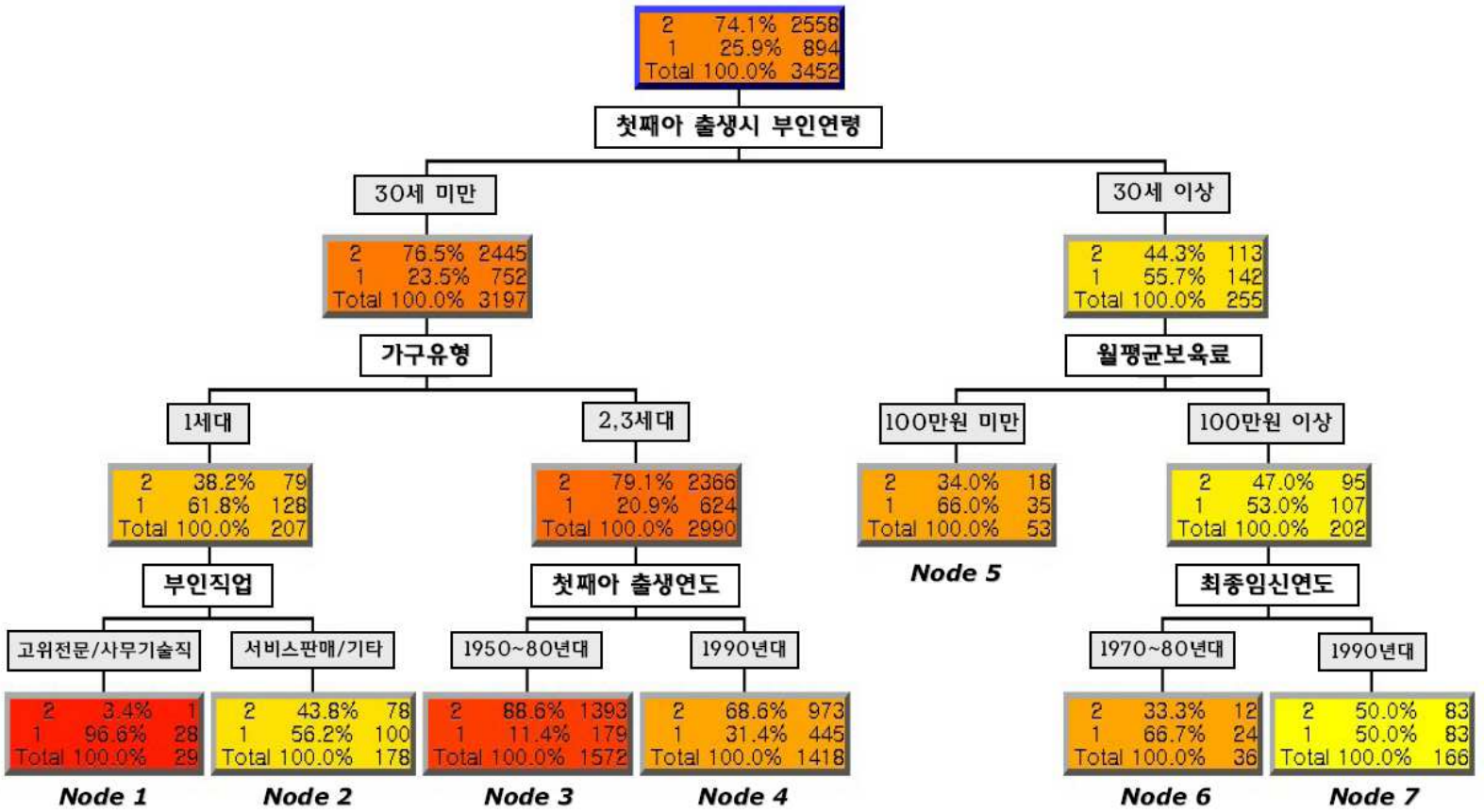


그림 7. CART 알고리즘에 의한 의사결정나무 (출산자녀수 2명에 비해 1명인 경우)

나. 출생자녀수 2명에 비해 3명 이상을 낳을 경우의 의사결정나무

출생자녀수 2명과 3명 이상을 결정하는 변수는 첫째아 출생년도, 막내아 출생년도, 부인의 결혼연령, 막내아 출생시 부인의 연령에 따라 의사결정나무가 결정되었고, 일부 연령과 관계된 변수들이 다소 반복되면서 가지치기를 하는 것을 볼 수 있었다. 그림 8의 의사결정나무에서는 최종 종단마디(Leaf node)가 모두 8개인 것을 볼 수 있다. 처음 시작할 때 노드(Root node)인 분석용 데이터를 살펴보면, 분석대상자가 총 3,393명인 경우로 출생자녀수가 2명인 경우는 전체의 76.1%이고, 출생자녀수 3명 이상인 경우는 23.9%이었다. 이번 결과에서 나온 각각의 node에 대한 내용은 다음과 같다.

Node1은 첫 번째 node에서 첫째아의 출생년도가 1950년대~1970년대인 경우이며, 막내아의 출생년도 역시 1950~1970년대이면서, 부인의 결혼연령이 21.5세 이하의 특징으로 가지치기되어지는 군으로서, 출생자녀수 2명인 경우는 56.8%, 3명 이상인 경우는 43.2%임을 보였다. Node2는 node1과 동일한 특성으로 가치치기 되어다가 마지막 노드인 부인의 결혼연령에서 21.5세 이상인 경우의 특성을 지닌 군에서는 출생자녀수가 2명인 경우는 86.7%, 출생자녀수가 3명 이상인 경우는 13.3%로 나타났다. Node3은 첫째아의 출생년도가 1950년대~1970년대인 경우이며, 막내아의 출생년도가 1980년대, 1990년대이면서, 막내아 출생시 부인의 연령이 25세 미만인 특징으로 분류되는 군으로서, 출생자녀수 2명인 경우는 73.7%, 3명 이상인 경우는 26.3%임을 보였다. Node4는 node3 과 같은 특성으로 분류되다가 마지막 노드인 막내아 출생시 부인의 연령이 25세 이상인 군으로서, 출생자녀수가 2명인 경우는 33.3%, 3명 이상인 경우는 66.7%으로 나타났다. Node5에서는 첫째아의 출생년도가 1980년대, 1990년대인 경우이며, 막내아 출생시 부인의 연령이 30세이하인 경우이며, 부인의 결혼연령이 20.5세 이하인 특징을 가지니 집단에서는 출생자녀수가 2명인 경우가 74.5%, 3명 이상인 경우가 25.5%임을 보였다. Node6은 node5와 같은 가지치기가 되어오다가 마지막 노드인 부인의 결혼연령에서 20.5세 이상인 특성을 가진 군으로 출생자녀수가 2명이 91.5%, 3명 이상인 경우는 8.5%이였다. Node7은 첫째아의 출생년도가 1980년대, 1990년대인 경우

이며, 막내아 출생시 부인의 연령이 30세 이상인 경우이며, 부인의 결혼 연령이 24.5세 이하인 특성을 가진 집단으로서 출생자녀수가 2명인 경우는 44.7%, 3명 이상인 경우는 55.3%을 나타내었다. Node8은 node7과 같은 특성으로 분류되다가 마지막 노드 부인의 결혼 연령에서 24.5세 이상인 특성의 군으로서 출생자녀수가 2명인 경우가 82.6%, 3명 이상인 경우가 17.4%임을 알 수 있었디.

최종 8개의 종단 노드들 중에서는 출생자녀수 2명을 낳는 경우에 비해 출생자녀수 3명을 가장 많이 낳은 경우를 다시 정리를 하면 node4가 66.3%, node7이 55.3%, node1이 43.2%, node3가 26.3%, node5이 25.5%, node8은 17.4%, node2이 13.3%, 그리고 node6에서는 8.5%인 순으로 나타났다. 이러한 결과의 패턴을 살펴볼 때, node4인 경우는 출생자녀수가 3명 이상이였던 확률이 Root node에서 23.9%였던 것이 첫째아 출생년도가 1950년대에서 1970년대로 분류되면서 출생자녀수를 3명이상 낳은 확률이 52.4%였고, 또 다시 여기서 막내아 출생년도가 1980년대 이후인 경우에서는 출생자녀수 3명이상이 61.4%, 마지막으로 막내아 출생시 부인의 연령이 25세 이상인 경우는 출생자녀수 3명이상인 경우가 66.3%로 양의 방향으로 증가하는 것을 볼 수 있다.

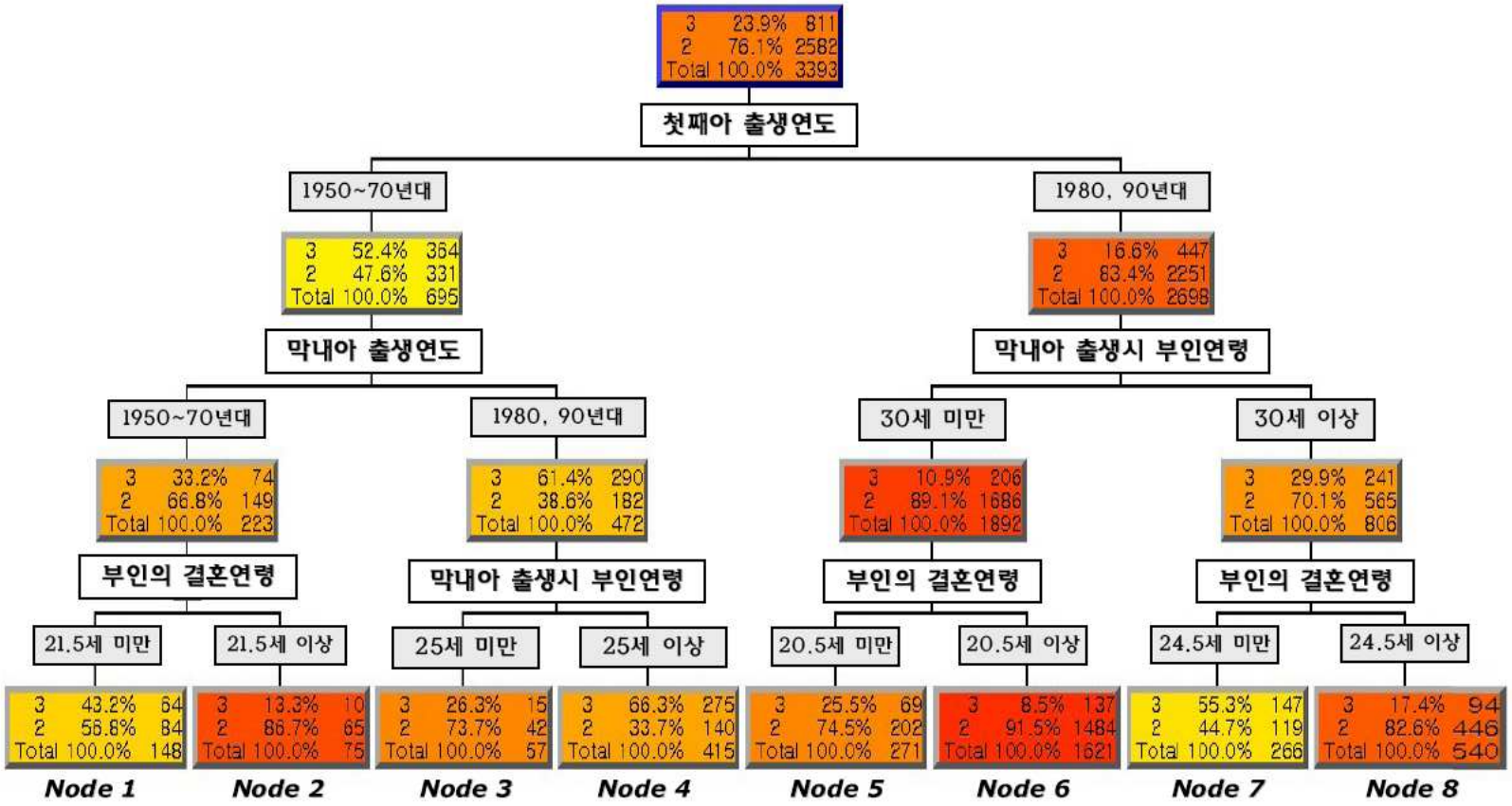


그림 8. CART 알고리즘에 의한 의사결정나무 (출산자녀수 2명에 비해 3명 이상인 경우)

4. 로지스틱 회귀분석 모형과의 비교평가

출생자녀수에 따른 출산관련 요인들을 제시한 두 가지 모형, 즉 로지스틱과 데이터마이닝의 CART 알고리즘에서 제시한 모형을 비교평가하기 위해서 각 모형으로부터 추정된 사후확률을 통한 예측력을 비교하기 위해 각각의 연구결과에서 제시한 오분류율(Misclassification Rate)과 'Lift value 및 'ROC curve'를 계산하였다.

가. 출생자녀수 2명에 비해 1명을 낳을 경우의 의사결정모형 평가

먼저, 출생자녀수 2명에 비해 1명을 낳을 경우에 대해 실행시킨 의사결정나무를 살펴보면, 의사결정나무 기법의 CART 알고리즘을 통하여 분리기준을 Gini index를 사용하였으며, 유의수준을 0.05로 지정하여 Tree node를 실행시켰다. 의사결정나무의 분석결과 모형의 오분류율은 표14에서도 표현하고 있는데, 분석용(training) 데이터에 대하여 16.13%의 오분류율를 보였으며, 테스트용 데이터에서는 17.10% 오분류율을 보였다. Root ASE (Asymptotic Standard Error)는 33.65%, 테스트용 Root ASE는 35.61%를 나타내었다. 이 값들은 모두 0에 가까이 갈수록 좋은 모형이라고 할 수 있는데, 이 모형에서도 대체적으로 좋은 모형이라고 평가할 수 있다. 앞서 시행한 로지스틱 회귀분석과 비교했을 때 의사결정모형의 오분류율과 ASE기 더욱 낮은 값을 나타내는 것을 알 수 있는데, 이는 로지스틱 회귀모형보다 의사결정모형이 더욱 좋은 모형이라고 할 수 있다.

표 10. 자녀수 2명에 비해 1명을 낳은 경우 의사결정나무 모형의 오분류율

		Asymptotic Standard error	Misclassification Rate
Logsistic Regression	Train data Set	0.4427	0.2505
	Test Data Set	0.4434	0.2508
Decision Tree	Train Data Set	0.3365	0.1613
	Test Data Set	0.3561	0.1710

다음에 제시된 그림9(좌)는 자녀수 2명에 비해 1명을 낳은 경우의 lift 이익 도표[12]이다. lift 이익도표는 분류모형의 성과를 비교하기 위한 것이며, 범주 1인 각 집단 내에서 평균적으로 가지는 빈도와 해당 집단내의 범주1에 대한 비율을 말하는데, 만약 lift가 1보다 크면 특정범주가 전체 평균보다 크다는 것을 의미한다. 그림 9(좌)에서 보는 바와 같이 의사결정나무 모형에 의한 기대 반응율(Expected Response)이 로지스틱 회귀모형에 비해 의사결정나무 모형이 Lift value가 더 큰 것을 알 수 있다. 그래서 Lift 이익도표에서도 로지스틱 회귀모형에 비해 의사결정나무 모형은 분류모형의 성과가 더 큰 것을 알 수 있다.

더불어 다음으로 살펴볼 것은 그림 9(우)에 위치한 ROC 도표인데, 이는 구축한 모형의 성능을 민감도와 특이도에 의해 판단하고자 사용하는 것으로, 민감도(sensitivity)는 (출생자녀수를 예측한 수/ 실제로 낳은 출생 자녀수)*100을 말하는 것으로(최국렬외, 2001), 그림9(우)의 ROC chart 에서는 각각의 1-특이도를 나타내는 값에 대하여 민감도가 높은 경우 좋은 모형이라고 할 수 있는데, 이 경우도 의사결정 모형이 로지스틱 회귀분석 모형에 비해 더욱 높은 값을 보이며 모형구축의 경우 상당히 좋은 모형이라고 말할 수 있다.

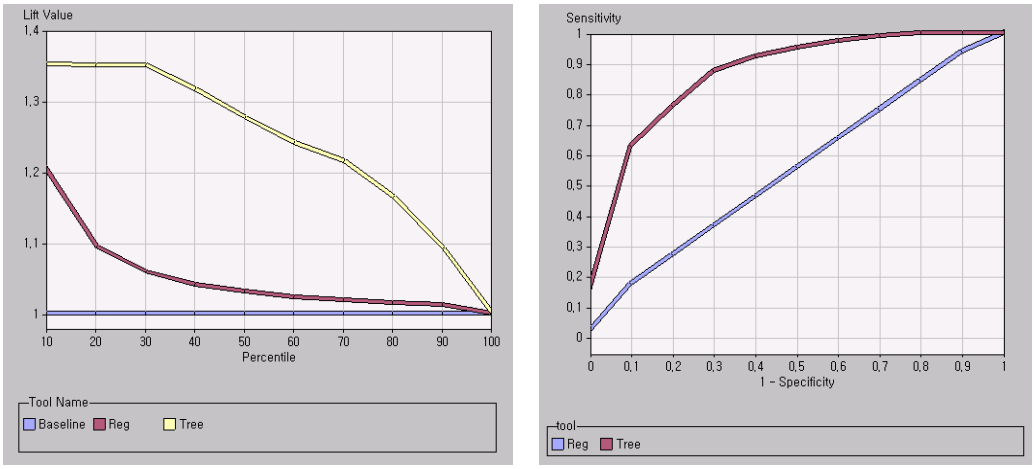


그림 9. Lift 이익도표(좌)와 ROC 도표(우) (자녀수2명에 비해 1명일 경우).

12) lift= 해당집단에서 목표변수의 특정범주빈도/전체에서 목표변수의 특정범주빈도*0.1

나. 출생자녀수 2명에 비해 3명 이상을 낳을 경우의 의사결정나무 평가

여기서는 출생자녀수 2명에 비해 3명 이상을 낳을 경우에 대한 의사결정나무를 평가하였다. CART를 이용한 의사결정나무의 분석결과 모형의 오분류율은 분석용(training) 데이터에 대하여 15.64%를 보였으며, 테스트용 데이터에서는 19.25% 오분류율을 보였다. 그리고 Root ASE는 분석용은 35.10%, 테스트용은 Root ASE는 38.15%를 나타내었다. 반면에 로지스틱 모형의 오분류율은 분석용(training) 데이터에 대하여 22.78%를 보였으며, 테스트용 데이터에서는 22.32% 오분류율을 보였다. 그리고 Root ASE는 41.87%, test용 Root ASE는 44.35%를 나타내는 것을 파악할 수 있는데 각각의 데이터 셋 간의 오분류율과 Root ASE가 거의 일치하는 것을 알 수 있다. 그리고 출생자녀수 2명에 비해 3명 이상을 낳을 경우의 통계적 모형에서도 역시 로지스틱 회귀모형보다 의사결정나무 모형이 더욱 ASE와 오분류율이 낮은 좋은 모형이라고 할 수 있겠다.

표11. 자녀수 2명에 비해 3명 이상을 낳을 의사결정나무 모형의 오분류율

		Asymptotic Standard error	Misclassification Rate
Logsistic Regression	Train data Set	0.4187	0.2278
	Test Data Set	0.4435	0.2232
Decision Tree	Train Data Set	0.3510	0.1564
	Test Data Set	0.3813	0.1925

그림10(좌)의 Lift curve 에서는 자녀수 2명에 비해 1명을 낳은 경우의 lift 이익도표[13]이다. 이는 그림 9(좌)와 같이 분류모형의 성과를 비교하기 위한 것이며, 그림10(좌)에서도 의사결정나무 모형에 의한 기대 반응율이 로지스틱 회귀모형

13) lift= 해당집단에서 목표변수의 특정범주빈도/전체에서 목표변수의 특정범주빈도*0.1

에 비해 의사결정나무 모형이 Lift value가 더 큰 것을 알 수 있다. 그래서 로지스틱 회귀모형에 비해 의사결정나무 모형은 분류모형의 성과가 더 큰 것을 알 수 있었다.

ROC chart 에서도 민감도가 높으면 높을수록 오분류율이 가장 낮음을 뜻하는데, 그림10(우)에서도 로지스틱 회귀모형에 비해서 의사결정나무 모형이 각각의 1-Specificity에 대하여 민감도가 현저하게 큰 것을 알 수 있다.

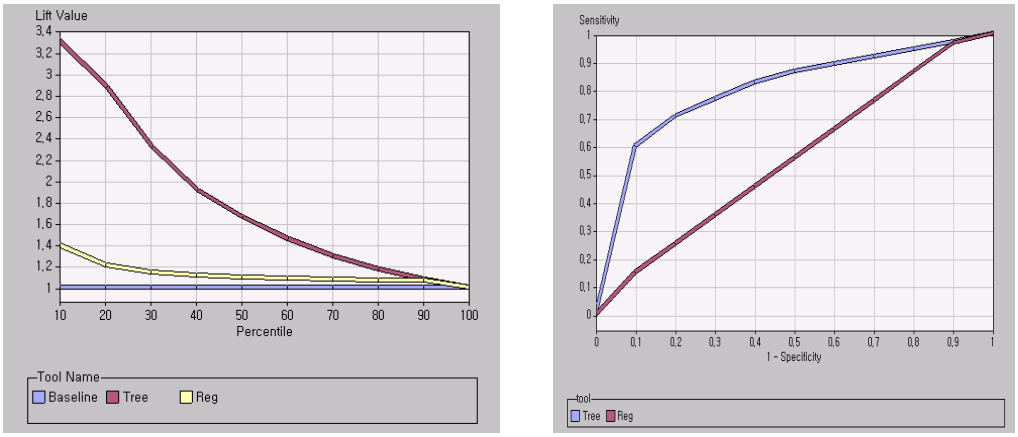


그림 10. Lift 이익도표(좌)와 ROC 도표(우) (자녀수2명에 비해 3명 이상일 경우).

그러므로 종합적으로 로지스틱 모형과 의사결정나무 모형을 비교해볼 때, 출생자녀수가 2명에 비해 1명을 낳을 경우와 3명 이상을 낳을 경우에서 모두 오분류율과 ASE가 낮게 나타내는 값을 보였고, Lift 이익도표와 ROC 도표에서도 평가기준이 로지스틱 회귀모형 보다 의사결정나무 모형이 훨씬 더 높은 것을 알 수 있다.

V. 고찰

1. 연구방법에 대한 고찰

본 연구가 지금까지 출산자녀수 결정요인과 관련된 기존의 연구와의 두드러진 것은 새로운 탐색적 통계기법인 데이터마이닝을 통해 분석하였다는 것이다. 물론 보건 의료분야에서도 이미 데이터마이닝 기법의 활용을 시도하고 있는데, 인구보건학 분야에서는 이러한 기법을 사용한 것은 전무한 실정이였기 때문이다. 그래서 데이터마이닝 기법을 통해 원 자료 자체의 관련성과 규칙을 파악하기 위한 것으로, 인구 보건학 분야에서는 시행되지 않았던 데이터 마이닝 기법을 통해서 출산력 연구의 제시 가능성과 타당성을 제시하고자 한 것에 새로운 의미가 있다고 생각된다. 아직까지는 대용량 자료 자체가 보여주는 패턴을 출생자녀수를 결정하는 완전한 패턴으로는 해석하기에는 제한점이 있지만, 출생자녀수 결정에 관여하는 다양한 특성의 요인들이 복합적으로 일정한 패턴을 가지고 있다는 점과 기존의 확증적인 통계기법에서 보지 못했던 변수들의 상호관련성을 의사결정나무라는 그림을 통해서 쉽게 알 수 있다는 점과 기존의 실증적 통계방법중의 하나인 로지스틱 회귀분석에서 보지 못했던 결과들을 보완적으로 살펴보는 데에도 의의가 있다고 생각된다. 앞에서도 언급했지만, 이 연구는 단면적인 연구이기에 연구 설계 자체가 가지는 제한점, 즉 원인 결과관계를 제시하는 잘 고안된 연구가 뒤따라야 하는 문제가 있으나 이 연구주제와 관련하여 우리나라의 출산력자료 자체가 보여주는 관련성을 분석하기 위해 시도하였다는 것은 상당한 의의가 있다고 생각된다.

그 중 데이터마이닝의 의사결정나무 분석은 연구결과를 쉽게 이해할 수 있고, 시각적 효과로 해석이 용이하다는 점, 일반 통계에서 의의가 없는 정보를 버리지 않고 가급적으로 많이 사용한 분류를 보여준다는 장점이 있는 반면, 분석도구가

민감하기 때문에 분석을 위한 분리기준, 정지규칙, 평가기준을 어떻게 지정하느냐에 따라서 서로 다른 결과를 나타낼 수 있으므로 연구자의 각별한 주의가 요구되었다. 따라서 실제로 단 한번 만에 분석에 끝나지 않고, 연구자는 모형을 만드는 과정을 수차례의 반복수행을 통해서 실제적용에 따르는 문제점을 비교검토하여야 하는 노력이 불가피하다.

그리고 본 연구를 위해 사용되었던 분석 자료인 '2000년도 전국 출산력 및 가족보건 실태조사'를 살펴보면, 조사원이 표본으로 선정된 조사구를 방문하여 면접 조사한 것을 통해 분석을 실시한 것이다. 이는 인구사회적, 경제적, 산과적, 및 가치관적 특성에 관한 내용 등을 설문할 때 피면접자의 기술에 근거하고 있기 때문에 자료의 신뢰성에는 한계가 있을 수 있다. 그리고 면접자들의 경우 출생자녀를 최근에 낳거나 임신하고 있는 경우도 있지만, 자녀를 오래전에 출생한 경우도 존재하며 대상자의 연령, 교육수준에 따라, 또는 결혼이나 임신, 출산시기가 오래 지난 사람들이 정확하게 기억하여 조사에 응답하지 못한 경우 기억편의(recall bias)가 개입될 가능성이 높다. 그리고 본 설문조사의 많은 독립변수들이 자녀를 출생하던 그 당시의 특성이 아니라, 조사를 실시한 당시의 요인들이기 때문에 출산 당시의 시점과는 차이가 발생할 수 있다는 한계가 있다. 그러므로 이와 같은 시간 의존적(time-dependent) 변수들의 영향은 통제되지 않았기 때문에 본 연구에서의 결과와 분석에 있어 제한점이 있을 수 밖에 없다. 다시 말하면, 목표변수인 출생자녀수의 정의상 실제적으로 과거의 특성이 출생자녀수에 영향을 미치지만, 현재의 특성을 가진 변수들이 결과에 나타나는 경우 결과와 분석에 한계를 가져올 수 있다는 것이다. 물론 의사결정나무 결과에서 현재 변수나 시간과 관련된 변수가 나오는 것은 한계를 가져다 주지만, 그 변수들을 포함해서도 서로간의 순수한 관련성을 보는 데에도 의의가 있다고 생각한다. 만약 현재 특성인 변수들을 제외하고 난 나머지 변수들과의 결과를 도출할 수도 있겠지만, 이러한 변수들이 출생자녀수에 결정적으로 영향을 미친다면 그 외의 변수들로만 잠

재적 전략을 세운다할지라도 제외된 변수의 영향을 고려하지 않았기 때문에 정책의 효과성에 있어서도 한계에 부딪힐 수밖에 없을 것이다. 정책적으로 실행시키기 위한 정보를 얻어내기 위해서는 보다 다양하고 한계성을 보완한 모형을 점진적으로 개발해나가야 한다고 본다.

그리고 본 연구는 전체 대상자 15세에서 49세까지의 유배우 부인 6,015명중 출생자녀수가 1명 이상인 경우로 중점을 두고 분석대상자를 제한하였고, 선정된 대상자의 평균 출생자녀수가 2명을 기준으로 2명보다 1명을 덜 낳는 경우에 영향을 미치는 결정요인과 출생자녀수 2명을 기준으로 3명 이상을 낳는 경우에 영향을 미치는 결정요인을 파악하기 위하여 로지스틱 회귀분석과 데이터마이닝 기법을 통해 분석을 실시한 것이다. 다시 말하면, 출생자녀수 2명을 낳은 경우에 비해 1명을 덜 낳게 하는 결정요인과 출생자녀수 2명을 낳은 경우에 비해 1명 이상을 더 낳게 하는 결정요인을 두 가지 경우로 구별하여 분석한 것으로 출생자녀수 증감에 따른 결정요인을 분석을 통해서 세부적인 정보를 제공한 것도 의미가 있다고 본다. 그러나 앞서 언급한 데로 본 연구에서는 출생자녀수가 없는 경우 306명은 제외하였고 출생자녀수가 1명 이상인 경우만으로 분석대상자를 제한하였으므로, 이는 출생자녀수가 없는 경우와 있는 경우에 대해서도 어떠한 요인들이 영향을 미치는지 살펴보는 것은 차후에 별도의 의미있는 연구주제가 되리라고 본다.

2. 연구결과에 따른 고찰

자녀수와 관련된 국외의 선행연구에서는 출산율에 영향을 주는 개인적인 요인에 대해서는 학자마다 다소 공통된 의견을 제시하고 있다. 기존의 연구에 의하면 부인의 개인적 배경인 연령, 학력, 인종, 경제활동 참여 여부, 수입, 종교, 자녀관련 요인인 자녀수 및 막내의 연령, 남편 관련 요인인 학력, 수입, 주거지역 등이 출

산력에 영향을 주고 있다고 말하고 있다(Lehrer and Kawasaki,1985; Mason and Kuhlthau, 1992; Del Boca, 2002). 그리고 국내의 연구 중 김한곤(1993)의 출산율 변화원인에 관한 결과를 보면, 거시적 접근에서는 출산율 감소는 사회경제발전, 여성의 지위 , 불임시술 등과 복합적으로 작용하고, 미시적 접근에 의하면 부인의 초혼 연령, 교육수분, 효과적인 피임 사용 등이 출산력 저하에 기여하였으나 기혼여성의 현재 고용상태에는 별다른 영향을 주지 못한 것으로 밝혀졌다. 보건사회연구원(2001)의 연구에서는 남편의 혼전 최장거주지, 현재거주지, 남편연령, 부인의 교육수준, 부인의 근로소득, 현 거주 집평수, 가구 소득원수, 자녀의 필요성, 이상 자녀수, 아들의 필요성, 자녀의 부도 대리성취라는 변수가 자녀수 결정에 영향을 주고 있다고 한 적 있다.

위와 같은 선행연구의 결과에 비해 본 연구를 통한 우리나라 유배우 부인의 출생자녀수 결정요인 및 결정패턴의 주요 연구결과를 정리하면 다음과 같다.

출생자녀수 2명에 비해 1명을 낳은 경우, 로지스틱 회귀분석을 통한 결정요인을 살펴보면 다른 요인을 모두 통제한 상태에서 거주지, 부인의 교육수준, 자연유산의 경험유무, 첫째아 출생시 부인의 연령, 부인의 현 취업유무, 월평균 보육료에서 유의하게 영향을 미치는 것으로 나타났다. 이 경우에 데이터마이닝 기법을 통해 분석한 결과로는 첫째아 출생시 부인의 연령, 가구유형, 부인의 직업, 첫째아 출생년도, 월평균보육료, 최종임신연도에 의해서 의사결정나무 구조를 나타내었다. 이 경우에서는 첫째아 출생시 부인의 연령와 월평균 보육료가 두 가지 분석방법의 결과에서 공통되게 나타나는 것을 볼 수 있다.

다음으로는 출생자녀수 2명에 비해 3명 이상을 낳은 경우에서 로지스틱 회귀분석을 통한 결정요인을 살펴봤을 때, 다른 요인을 모두 통제한 상태에서 거주지, 부부간 역할 분담, 남편의 교육수준, 첫째아 출생연도, 자녀의 필요성, 월평균 보육료에서 유의하게 영향을 미치는 것으로 나타났다. 그리고 데이터마이닝의 의사

결정나무 구조에서는 첫째아 출생년도, 막내아 출생년도, 부인의 결혼연령, 막내아 출생시 부인의 연령에 따라 의사결정나무가 결정되었고, 대체적으로 연령과 출생년도와 관계된 변수들이 다소 반복되어 가지치기를 하는 것을 볼 수 있었고, 첫째아 출생년도는 CART 알고리즘의 의사결정 패턴에서도 관련이 있는 변수임을 알 수 있다.

더불어서 본 연구에서는 데이터마이닝을 통해 생성된 의미있는 패턴들을 살펴본다면, 출생자녀수가 2명인 것에 비해 1명을 낳게 된 경우와, 출생자녀수가 2명인 것에 비해 3명 이상을 낳게 된 경우에서 결정패턴을 살펴보면 각각 다음과 같다.

먼저, 출생자녀수가 2명인 것에 비해 1명을 낳은 경우를 나타내는 <그림7>에서 최종 7개의 terminal node들을 출생자녀수 1명을 가장 많이 낳은 순서대로 정리해 보면, node1이 96.6%, node5에서 66.7%, node2가 56.2%, node6이 50.0%, node4가 31.4%, 그리고 node3이 11.4%인 순으로 나타났다. 그러나 이러한 결과의 패턴을 살펴볼 때, 본 연구의 목적과 부합하여 출생자녀수가 2명인 것에 비해 1명을 낳게 된 경우에서는 출생자녀수 1명을 적게 낳게 되는 생성규칙을 살펴보는 것이 더욱 의미가 있다고 생각되어진다. 출생자녀수 1명을 가장 적게 낳은 node3에서는 출생자녀수가 1명이였던 경우가 Root node에서 25.9%였던 것이 첫째아 출생시 부인의 연령이 1980년대 이전으로 분류되면서 출생자녀수를 1명을 낳은 확률이 23.5%였고, 또 다시 여기서 가구유형이 2세대 혹은 3세대가 같이 사는 경우에서는 출생자녀수 1명을 낳은 경우가 20.9%를 차지하였고, 마지막으로 첫째아 출생년도가 1990년대 이전인 경우에서는 11.4%가 출생자녀수 1명을 가지고 있으며, node3의 종단마디로 갈수록 출생자녀수 1명을 가진 확률이 음의 방향으로 낮아지는 패턴을 보였다. 즉 이는 상대적으로 출생자녀수 2명을 낳은 경우가 증가하는 패턴이기도 한 것이다.

출생자녀수가 2명인 것에 비해 3명 이상을 낳은 경우를 나타내는 <그림8>에서는 최종 8개의 종단 노드의 의사결정나무가 구성되어 있다. 출생자녀수 2명을 낳는 경우에 비해 출생자녀수 3명 이상 낳은 경우의 가능성이 큰 순서대로 정리를 하면 node4가 66.3%, node7이 55.3%, node1이 43.2%, node3가 26.3%, node5이 25.5%, node8은 17.4%, node2이 13.3%, 그리고 node6에서는 8.5%인 순으로 나타났다. 이러한 결과의 의미있는 패턴을 살펴볼 때, node4의 Root node에서 출생자녀수가 3명 이상이였던 분률이 Root node에서 23.9%였던 것이 첫째아 출생년도가 1950년대에서 1970년대로 분류되면서 출생자녀수를 3명 이상 낳은 확률이 52.4%였고, 또 다시 여기서 막내아 출생년도가 1980년대 이후인 경우에서는 출생자녀수 3명이상이 61.4%, 마지막으로 막내아 출생시 부인의 연령이 25세 이상인 경우는 출생자녀수 3명 이상인 경우가 66.3%로 양의 방향으로 증가하는 패턴이 생성된 것을 볼 수 있었다.

3. 정책적 시사점

정책적 차원이 출산율에 미치는 영향을 살펴보면, 정책을 크게 직접적 정책과 간접적 정책으로 나누어 볼 수 있다 직접적인 정책에는 금전적인 지원이 있는데 여기에는 현금급여, 특별대부, 세금공제, 특별지원금 등이 포함된다. 비금전적인 간접적 정책으로는 보육시설, 자녀양육휴가, 근무시간, 사회복지제도, 조세제도 등이 있다. 우선, 양질이면서 적절한 비용의 보육시설이 존재한다면 출산율을 높일 수 있고, 자녀를 돌보기 위한 휴가제도, 자녀와 가사 일을 돌볼 수 있도록 조정된 근무시간, 특성에 대해서 특혜나 불이익을 주지 않는 복지제도, 자녀 양육을 지원하는 조세제도의 유무 등이 출산율에 긍정적 영향을 준다고 할 수 있다 (서문희, 2004)

그러면 본 연구의 결과를 가지고 현재의 저출산의 심각한 문제를 해결하기위해 어떻게 정책적 방향을 설명할 수 있을 것인가? 데이터마이닝의 의사결정나무의 결과를 보면, 복합적인 여러 요인과 관련하여 출생자녀수가 증가하는 방향과 출생자녀수가 감소하는 방향으로 생성되는 패턴들을 볼 수 있다. 앞서서 설명한 바와 같이 연구결과들을 통해 출생자녀수를 결정하는 데 영향을 미치는 결정요인과 결정패턴을 파악할 수 있었다. 이는 세부적으로 출생자녀수 2명에 비해 1명을 낳는 경우와 출생자녀수 2명에 비해 3명 이상을 낳는 경우를 나누어서 분석하였는데, 출생자녀수가 증가하도록 의사 결정하는 패턴을 가지는 경우에 정책적인 전략을 세워볼 수 있을 것이다. 각각의 집단의 특성에 맞게 출생자녀수를 증가하는 방향으로 정책을 세우는 것이 필요하다. 위의 제시된 결과에서 첫째아 출생시 부인의 연령이 30세미만과 30세 이상인 군에서 출생자녀수를 낳는 확률이 현저하게 달라지는 것을 볼수 있다. 이런 경우 출생자녀수를 많이 낳는 패턴으로 결정한 대상자에게 재정적인 혜택을 주는 것인데, 앞서 언급하였지만 직접적인 정책의 현금급여, 특별대부, 세금공제, 특별지원금 등과 같은 재정적 유인책을

출생자녀수가 증가하는 방향으로 사용한다면 좋은 방법일 것이다. OECD 국가들이 미흡하나마 현금급여 정도와 세제혜택 등이 출산율 증가에 긍정적인 영향을 미치고 있음을 보고하고 있다는 점에서 (Sleebo, 2003) 우리나라의 정책방향에서도 현재 이러한 방법들이 고려되고 있는 중이다.

그러나 무엇보다도 이러한 연구결과를 실제적으로 정책화 활용을 위한 가능성 갖추기 위해서는 데이터마이닝을 이용한 출생자녀수 결정요인 및 결정패턴에 관한 보다 안정되고 타당한 모형을 개발하여야 한다고 생각되어 진다. 출산력 연구에서도 탐색적 방법인 데이터마이닝 기법을 적용한 다양한 연구를 통해 연구의 한계점을 보완하고 집단의 특성에 따라 출생자녀수를 보다 정확하게 예측하는 신뢰도 있는 의사결정 모형을 개발하여 점진적으로 개선해 나가야 할 것이다. 그리하여 출생자녀수가 증가하는 방향에는 긍정적인 유인책을, 출생자녀수가 감소하는 방향에는 억제책을 시행함으로서 집단군의 특성에 맞게 전략적으로 적용한다면 보다 효율적인 결과를 이룰 수 있을 것이다.

그리고 본 연구결과에서 두드러진 변수들을 살펴보면, 자녀 출생시 부인의 연령, 부인의 결혼연령과 출생년도에 관한 변수인 것 같다. 결혼이라는 제도가 출산을 규제한다고 할 때, 그 중 가장 중요한 방법 중의 하나가 바로 혼인 연령에 대한 부분이다. 혼인 연령은 대부분의 사회에서 사람들의 자녀 출산의 가능기간을 결정하는 가장 중요한 요인으로 작용한다(Braun, 1978; Chambers, 1965; 이희연, 2004) 최근의 우리나라의 현저한 혼인연령상승은 여성의 교육에 대한 열망 및 미혼 여성의 취업기회 확대 그리고 자아성취 욕구의 증대 등에 기인될 수밖에 없는 것이다(김승권, 2001). 더불어 흥미로운 것은, 첫째아 출생시 부인의 연령이 낮을수록, 그리고 막내아 출생시 부인의 연령이 늦어질수록 출생자녀수가 증가하였다. 그러면 이러한 연구결과와 관련하여 출산시기의 폭을 늘려주는 정책적 방향을 생각해봐야 할 것이다. 한국사회에서 출산력은 지난 30~40년 사이에 엄청나게 빠른 속도로 그리고 매우 낮은 수준으로 떨어졌기 때문

에 한 여성이 출산을 완료하는데 걸리는 시간이 해를 거듭할수록 매우 짧아졌을 것임은 누구나 쉽게 짐작할 수 있다(은기수, 2001) 그러므로 혼인연령이 어릴수록 출생자녀수를 증가시키는 패턴을 나타내므로 일찍 결혼할 때 혜택을 준다든가, 부인의 첫째아 출산시기를 앞당겨주고, 막내아 출산에 있어서도 이른 나이에 완료하지 않도록 하는 직접적, 간접적인 정책을 세워야 할 것이다. Ermisch(1988)의 연구를 보면 가족현금을 더 많이 지원할 때 여성들이 어머니가 되는 시기가 더 빨라지는 경향이 있다고 한다. Barmby & Cigno(1990) 역시 영국에 대한 연구에서 아동수단의 지급 수준을 상향조정하는 것이 가족의 수보다는 어머니가 되는 시기에 영향을 미친다고 보고하고 있다. 그런데 무엇보다도 정책의 효과성이 일시적이지 않도록 하게 하기 위해서는 위의 선행연구와 같은 직접적인 요인들에 근거한 효율적인 정책도 중요하지만, 그러한 결과를 가져오게 된 복합적인 사회경제적 변화 및 여러 복합적인 부분이 관련된 문제들을 바라보는 것 또한 중요하다. 직접적인 출산 유인정책보다는 장기적인 시각에서 우리 사회가 자녀를 출산하기 위한, 그리고 출생한 자녀가 잘 양육되어지도록 안정적이고 바람직한 환경정비를 목표로 하는 것이 더욱 중요할 것이다. 이러한 사회적 현상은 더욱이 1997년 이후에는 IMF 구제 금융으로 인한 경기침체와 대량 실업사태로 미혼 남녀의 결혼 연기 및 기피현상으로, 혹은 기혼 남녀의 출산 지연 및 기피의 원인으로도 작용하였다는 점도 감안하여야 할 것이다. 장기적으로는 전반적인 사회 경제 발전뿐만 아니라 무엇보다 극심한 청년실업사태를 해결하기위한 전략도 간과되어서는 안 될 것이다.

그리고 저출산 현상에 대응한 정책적 전략의 장기적인 기본 방향은 아동 양육과 사회적 책임의 확대와 양성평등 환경조성이 되어야 할 것이며 이를 위한 구체적인 정책 목표로는 첫째, 가족의 자녀양육에 대한 경제적 부담의 완화, 둘째, 여성의 경제활동 지속보장의 성평등한 노동환경 구축, 셋째, 여성 및 남성의 성평등한 양육책임을 확산시키는 것이다.

최근 들어서 출산력이 급격하게 감소하고 여성 취업률이 50%로 넘지 못하면서 보육, 육아지원 등 가정에서 아이를 낳아서 기르는 부모의 역할을 국가와 사회가 지원하여야 한다는 공감대도 크게 강화되고 있는데, 기혼 여성의 노동시장 참가율을 높이기 위해서는 다양한 육아지원정책을 강화하여야 한다. 이용률이 낮은 출산휴가와 육아휴직의 실효성을 확보하여 이용률을 증가시키고, 공보육시설의 설치, 비용지원 등 질 높은 공보육을 확보할 수 있는 방안을 구체화하며, 이외에도 가족의 자녀양육비용 직접 지원 등 다양하고 다각적인 가족지원정책의 통합적 추진이 필요하다(서문희, 2004)

이러한 실질적으로 자녀를 낳아서 양육하기에 좋은 환경을 갖추기 위한 사회 전반적인 노력이 일어나야 할 것이다. 즉 직접적으로 출생자녀수에 영향을 미치는 요인들을 바탕으로 정책을 세워서 일시적인 출산수준을 늘리는 것에 급급할 것이 아니라, 혼인 연령이나 첫출산 연령이 증가하게 된 것도 복합적인 요인들이 간접적으로 영향을 미쳤을 것이다. 특히 직접적으로 미친 여러 요인들에 간접적으로 영향을 미치는 환경들을 개선함으로써 저출산 현상에서 근본적으로 벗어나는 장기적인 전략을 세워야 할 것이다.

VI. 결론

이 연구에서는 2000년 전국 출산력 및 가족보건 실태조사 자료를 중심으로 유배우 부인의 출산력 관련 변수 중에 출생자녀수정도를 결정하는데 미치는 영향요인 및 결정패턴을 파악하고자 하였다. 이를 위해 데이터 마이닝 기법 중 의사결정나무분석인 CART 알고리즘을 이용하여 자료자체가 보여주는 구조적인 패턴을 알아보고 향후 인구 및 보건학 분야에서의 적용가능성을 제시 하고자 하였다.

분석방법은 출산관련 변수에 따른 출생자녀수정도를 파악하기위해 위험요인들의 기술통계량을 비교 분석한 후 로지스틱 회귀분석을 통해 출생자녀수에 영향을 주는 요인를 파악하였다. 대규모 출산력 자료자체에서 보이는 출생자녀수에 영향을 미치는 특성을 파악하기위해 CART 알고리즘에 의한 의사결정나무를 도출하였고, 분석용 데이터 셋을 75%, 테스트용 데이터 셋을 25%로 할당하였다. 연구결과는 다음과 같았다.

데이터마이닝의 CART 알고리즘의 결과에서는 목표변수에 따라서 다른 결과를 나타내었는데, 출생자녀수가 2명인 것에 비해 1명을 낳은 경우에 데이터마이닝 기법을 통해 분석한 결과로는 첫째아 출생시 부인의 연령, 가구유형, 부인의 직업, 첫째아 출생년도, 월평균보육료, 최종임신연도에 의해서 의사결정나무 구조를 나타내었다. 이 경우에서는 첫째아 출생시 부인의 연령와 월평균 보육료는 로지스틱 회귀방법의 결과와 공통되게 나타나는 것을 볼 수 있다. 그리고 출생자녀수가 2명인 것에 비해 출생자녀수 1명을 낳게 된 경우에서는 적게 낳게 되는 생성규칙을 살펴보는 것이 더욱 의미가 있다고 생각되어진다. 결과로 나온 node들 가운데 출생자녀수 1명을 가장 적게 낳은 node3을 살펴보면, 출생자녀수가 1명이였던 경우가 Root node에서 25.9%였던 것이 첫째아 출생시 부인의 연령이 1980년대 이전으로 분류되면서 출생자녀수를 1명을 낳은 확률이 23.5%였고, 또 다시 여기서 가구유형이 2세대 혹은 3세대가 같이 사는 경우에서는 출생자녀수 1명을

낳은 경우가 20.9%를 차지하였고, 마지막으로 첫째아 출생년도가 1990년대 이전인 경우에서는 11.4%가 출생자녀수 1명을 가지고 있으며, node3의 종단마디로 갈수록 출생자녀수 1명을 가진 확률이 음의 방향으로 낮아지는 패턴을 보였다.

다음으로는 출생자녀수 2명에 비해 3명 이상을 낳은 경우의 데이터마이닝의 의사결정나무 구조에서는 첫째아 출생년도, 막내아 출생년도, 부인의 결혼연령, 막내아 출생시 부인의 연령에 따라 의사결정나무가 결정되었고, 대체적으로 연령과 출생년도와 관계된 변수들이 다소 반복되어 가지치기를 하는 것을 볼 수 있었고, 첫째아 출생년도는 로지스틱 회귀분석에서도 유의한 변수임을 알 수 있었다. 출생자녀수가 2명인 것에 비해 3명 이상을 낳은 경우 최종 8개의 종단 노드의 의사결정나무가 구성되어 있다. 다음의 결과에서 의미있는 패턴을 살펴볼 때, node4의 Root node에서 출생자녀수가 3명 이상이였던 분율이 Root node에서 23.9%였던 것이 첫째아 출생년도가 1950년대에서 1970년대로 분류되면서 출생자녀수를 3명 이상 낳은 확률이 52.4%였고, 또 다시 여기서 막내아 출생년도가 1980년대 이후인 경우에서는 출생자녀수 3명이상이 61.4%, 마지막으로 막내아 출생시 부인의 연령이 25세 이상인 경우는 출생자녀수 3명 이상인 경우가 66.3%로 양의 방향으로 증가하는 패턴이 생성된 것을 볼 수 있다.

본 연구결과를 통해 출생자녀수를 결정하는 데 영향을 미치는 결정요인과 결정패턴을 파악할 수 있었다. 이는 세부적으로 출생자녀수 2명에 비해 1명을 낳는 경우와 출생자녀수 2명에 비해 3명 이상을 낳는 경우를 나누어서 분석하였는데, 복합적인 여러 요인과 관련하여 출생자녀수가 증가하는 방향과 출생자녀수가 감소하는 방향으로 생성되는 패턴들을 볼 수 있다. 이에 따라 정부에서는 각각의 집단의 특성에 맞게 출생자녀수를 증가하는 방향으로 정책을 세우는 것이 필요하다. 이런 경우 출생자녀수를 많이 낳는 패턴으로 결정한 대상자에게 재정적인 혜택을 주는 것인데, 앞서 언급하였지만 직접적인 정책의 현금급여,

특별대부, 세금공제, 특별지원금 등과 같은 재정적 유인책을 출생자녀수가 증가하는 방향으로 사용한다면 좋은 방법일 것이다.

이러한 연구결과를 정책적으로 활용하기 위한 가능성 갖추기 위해서는 무엇보다도 데이터마이닝을 이용한 출생자녀수 결정요인 및 결정패턴에 관한 보다 안정되고 타당한 모형을 개발하여야 한다고 생각되어 진다. 집단의 특성에 따라 출생자녀수를 보다 정확하게 예측하는 신뢰도 있는 의사결정 모형을 개발하여 점진적으로 개선해 나가야 할 것이다. 출생자녀수가 증가하는 방향에는 긍정적인 유인책을, 출생자녀수가 감소하는 방향에는 억제책을 시행함으로서 집단군의 특성에 맞게 전략적으로 적용한다면 보다 효율적인 결과를 이룰 수 있을 것이다.

그런데 무엇보다도 정책의 효과성이 일시적이지 않도록 하게 하기 위해서 위의 선행연구와 같은 직접적인 요인들에 근거한 효율적인 정책도 중요하지만, 그러한 결과를 가져오게 된 복합적인 사회경제적 변화 및 여러 복합적인 부분이 관련된 문제들을 바라보는 것 또한 중요하다. 직접적인 출산 유인정책보다는 장기적인 시각에서 우리 사회가 자녀를 출산하기 위한, 그리고 출생한 자녀가 잘 양육되어지도록 안정적이고 바람직한 환경정비를 목표로 하는 것이 더욱 중요할 것이다. 실질적으로 자녀를 낳아서 양육하기에 좋은 환경을 갖추기 위한 사회전반적인 노력이 일어나야 할 것이다. 즉 직접적으로 출생자녀수에 영향을 미치는 요인들을 바탕으로 정책을 세워서 일시적인 출산수준을 늘리는 것에 급급할 것이 아니라, 혼인 연령이나 첫출산 연령이 증가하게 된 것도 복합적인 요인들이 간접적으로 영향을 미쳤을 것이다. 특히 직접적으로 미친 여러 요인들에 간접적으로 영향을 미치는 환경들을 개선함으로써 저출산 현상에서 근본적으로 벗어나는 장기적인 전략을 세워야 할 것이다.

참고문헌

1) 단행본

김승권외 (2001), 2000년 전국 출산력 및 가족 보건실태조사, 한국보건사회연구원

김승권외 (2001), 출산력 및 가족보건 실태의 변화양상과 대응방안에 관한 연구, 한국사회연구원

김한곤 (1993), 한국 출산력 변화의 원인과 전망, 영남대 출판부

김승권 (2004), 한국사회의 저출산 원인과 정책적 함의, 보건사회연구원

장혜경외 (2004), 저출산 시대 여성과 국가대응전략, 한국여성개발원

최경수 (2004), 한국출산력 하락 추이에 관한 분석, 한국개발연구원 통계청, 인구통계연보, 연도별자료

2) 정책토론회

김승권 (2005), 한국사회 저출산의 원인과 대응방안, 한국보건사회연구원

이수희 (2005), 저출산이 사회경제적으로 미치는 파급효과, 한국경제연구원

임일섭 (2004), 세계 최저출산율, 어떻게 대응해야 하나, LG주간경제

3) 국내 학회지

김성혜, 김초강(1991), 결혼관련 요인의 출산력에 미치는 영향연구, 한국보건교육학회지

김승권, 한국사회의 저출산의 원인과 정책적 함의(2004), 한국인구학회 전기 학술대회.

은기수 (2001), 첫 출산부터 마지막출산까지 출산기간의 차별성, 한국보건통계학회지.

이건창, 정남호, 신경식(2001), 신용카드 시장에서 데이터 마이닝을 이용한 이탈고객 분석, 한국지능정보시스템 학회지, 춘계정기 학술대회.

전광희 (2002), 한국의 저출산: 추이와 전망, 충남대학교 사회과학연구소

최경수, 한국 출산력 하락 추이에 관한 분석, 2004, 한국인구학회 전기 학술대회.

최연희, 강대룡외 (2004), CART알고리즘을 이용한 구강건강과 전신건강과의 관련성, 대한구강보건학회지

최종후, 서두성 (1999), 데이터 마이닝의 의사결정나무의 응용, 통계청 통계분석연구 제4권 제1호.

최재훈, 이상훈 (2003), 데이터마이닝 기법을 적용한 안개 예보칙 장성방안 연구, 데이터베이스연구, 제19권 제4호.

호승희, 채영문, 조경일외(1995), 의료분야의 지식경영을 위한 데이터마이닝 응용, 대한의료정보 학회지.

4) 국내 학위논문

김혜숙(2001), 데이터마이닝을 이용한 의료의 질 측정지표 분석 및 의사결정 지원 시스템 개발, 연세대학교 학위논문.

변준한(2004), 기혼유배우 여성의 자녀수 결정요인 분석, 고려대 학위논문

용왕식(2000), 의료보험 지식경영구축 방안에 관한 연구, 연세대학교 학위논문.

신선미(2002), 저체중 청소년 건강증진을 휘한 데이터 마이닝 응용, 연세대학교 학위논문.

전용진(2003), 약물의 유효성 평가에서의 변수선택법 연구, 전용진, 연세대학교

학위논문.

조윤정(2001), 데이터 마이닝을 이용한 종합건강진단텐터의 데이터베이스 마케팅에 관한 연구. 서울대 학위논문.

최명애(2001), 데이터 마이닝 기법을 활용한 DRG 분류체계 및 재원일수 관련 요인 연구, 연세대학교 학위논문.

5) 국외문헌

Breimen L, Friedman JH, Olshen RA, Stone CJ.(1983) Classification and Regression Trees. Belmont, CA; Wadsworth statistical Press; 1984:1-58

Buja A, and Lee Y-S(1999), Data mining criteria for tree-based regression and classification, http//research.att.com/~andreas/paers/trees.p.,gz.

Sleebo, Joelle. 2003."Low fertility rates in OECD countries: Facts and policy respons." OECD Social employment and Migration working papers.

國立社會保障人口問題研究所, population of Japan, 2003

Keith B. Hermiz, Critical success factors for Data mining projects, SourceMedia and DM Review, 2005

Hans-peter Kohler, Jere R Behrman, and Susan C. Watkins, 2001, the density of social networks and fertility decisions: evidence from south Nyanza district, Kenya, Demograghy, Volume 38, Number 1.

NBER, “The Impact of Population Aging on Financial Markets”, NBER Working Paper No. 10851, October 2004.

OECD,『OECD Health Data』, 2004.

The American Economic Review, “Declining Population and Sustained Economic growth: Can they Coexist?”, *The American Economic Review*, Vol.88, No 2, 1998.

UN,『The Sex and Age Distribution of World Population』, Annual report

Kass, G. (1980). An exploratory technique for investigating large quantities of categorical data. *Applied Statistics.* 29:2, 119-129.

Chae Y.M. Data-mining approach to policy analysis in a health insurance domain, Int J Med Inf. 2001; July 62(2-3) 103-1

Sleebos, Joelle E., Low Fertility Rate in OECD Countries: Fact and Policy Response, OECD, 2003

Del Boca, Daniela, "Low Fertiliy and labour Force Participation of Italian Women; Evidence and Interpretation,' OECD labour market and Social Plicy-Occasional Papers, N61, 2002

Lehrer, Evelyn L., and Seiichi Kawasaki, ' Child care Arrangements and Fertility: An Analysis of Two-Earner Households' Demography, vol.22, No.4, 1985

6) 도서

김현승, 김현진,(2003) 늙어가는 대한민국-저출산 고령화의 시한폭탄, 삼성경제연구소

서길수(2000), 데이터베이스의 관리, 박영사, 2000

장남식, 홍성완, 장재호(2000), 데이터마이닝, 대청미디어.

최국렬, 조대현, 이상열, 석경하, 박일수, 김유미, 기옥남, 김병수, 강성홍 공저 (2001), 데이터마이닝 이론과 실습, 청구문화사.

허명희, 문승호(2000), 탐색적 자료분석(EDA), 개정판, 서울; 자유아카데미.

홍순영, 황인성(2004), SERI 전망 2004, 삼성경제연구소

강현철, 서두성, 최종후(2002), Enterprise Miner의 의사결정나무분석 알고리즘

이희연(2003), 인구학; 인구의 지리학적 이해, 법문사

7) 연구보고서

서문희(2004)외, 여성 사회활동 증진을 위한 보육환경 개선방안 연구, 한국보건사회연구원.

7) 인터넷 사이트

대한민국통계청 (http://nso.go.kr)

일본통계청 (http://ipss.go.jp)

대한 가족보건복지협회(www.ppfk.or.kr) javax.datamining.algorithm.tree

Abstract

This study is related to a low birth, one of the key social issues of Korean society. In recent years, it has been unfavorably predicted that the low birth rate in Korea was going up and would be worst over the world. This is the important reason why we should study this social phenomenon such as a low birth. Also, it is strongly expected to provide the fundamental information for a national birth policy, which can overcome the present phenomenon of low birth and make it a desirable population replacement level in Korea. Based on the data from Year 2000 Korea National Fertility Survey carried by Korea Institute for Health and Social Affairs (KIHASA), we analyzed its inter-relationship among several variables about fertility, and several decision-making patterns for the number of children in Korea, using data-mining technique (i.e., decision tree).

In most studies about the number of children, some confirmatory approaches, one of the traditional statistical reasoning methods, have been used to assess the determinants which cause a low fertility in Korea until now. In these a few studies of them, some factors about the number of children were examined by technical statistical methods. Therefore, there were some limitations in that each factor of low birth was not completely predicted and considered in the complicated socioeconomic environments. However, as an importance of data mining and analysis has recently been emerged since more powerful data processing was fully developed with a rapid computation and a huge accumulation of massive data, the data mining and analysis is a widely known method in various practical analyses: There have been few studies using this exploratory technique such as data-mining in population and public health

field. Although a national concern of low birth is drastically increased in recent years, this technique have not applied in order to analyze decision-making patterns for the number of children. This study is the first approach to analyze an interrelation and rule among all birth-related variables which affect the number of children in nationwide fertility database by using data-mining technique. Furthermore, these results would be contributed to provide applicability of data-mining technique for future studies of population and public health.

In this study, we focused the respondents into only married women, which have over one children, among 6,015 respondents in a range from 15 years to 49 years. We achieved a CART (Classification And Regression Trees) algorithm, one of decision-making trees, to exactly know the decision-making patterns which strongly influence on determining the number of children. And the results from these trees were compared with those of logistic regression model for two cases; one is that a chosen respondent has one child, the other is that she has over three children. All these results were compared with the case that the average number of children is two.

In the first case, the number of children has been influenced by the region of residence, wife's education level, spontaneous abortion, wife's age at first childbirth, the present employment state of wife, the average monthly nurturing cost from the logistic regression analysis adjusting the effects of the other factors. But the decision-making trees were formed by wife's age at first childbirth, type of house, wife's occupation, the year of first childbirth, the average monthly nurturing cost, the last year of pregnancy from the data-mining technique. Particularly wife's age at first childbirth and the average monthly nurturing cost were used as the key factors of deciding the number of children in both analyses.

In the second case, the number of children has been influenced by the region of residence, a conjugal household assignment, husband's education level, the

year of first childbirth, the need of children, the average monthly nurturing cost from the logistic regression analysis adjusted the effects of the other factors. And the decision-making trees were formed by the year of first childbirth, the year of last childbirth, wife's age at marriage, wife's age at last childbirth from the data-mining technique. In other words, the year of childbirth and the age of husband and wife made many complicated branches of decision-making patterns. Obviously, it can be shown that the year of first childbirth is an important factor of decsion-making patterns of CART algorithms.

In a brief summary, we achieved the CART analysis with the decision tree to understand several decision-making patterns for the number of children, and also compared these results with those of logistic regression model to assess the stability and predictability of the model for the optimized decision tree. Furthermore, those results would be contributed to provide more reliable information on a prospective birth promotion policy and give more farsighted measures for a potential solution to the lowest birth rate and ongoing superaging society in future Korea.