「정보시스템연구」 제28권 제3호
한국정보시스템학회
2019년 9월, pp. 249~276

http://dx.doi.org/10.5859/KAIS.2019.28.3.249

의사결정나무 기법을 이용한 노인들의 자살생각 예측모형 및 의사결정 규칙 개발*

김덕현** · 유동희*** · 정대율****

〈목 차〉

Ⅰ. 서 론

자살은 여러 가지 죽음의 형태 중 하나로, 스스로 삶을 중단시키는 행위이다. 자살은 고의적 자해에 의한 사망이다. 자살률이란 인구 10만 명당 자살자 수를 의미한다. OECD 가입국가 자살률은 국가마다 다른 연령구조 차이를 제거하여 인구 규모를 표준화한 10만 명당 자살자 수를 백분율로 나타낸 것으로 국가별 비교를 위해 사용된다. OECD 가입국가 중 한국은 높은 자살률을 보이고 있으며, 그 중 60대 이상의 고령자 자살률은 OECD 평균보다 압도적으로 높은 수치를 기록하고 있다. 전 연령대 기준으로 대부분의 국가가 인구 10만 명당 20명 안팎으로 나타나고 있으나 한국의 경우 30명을 상회하며, 노인의 경우 60-70명 이상의 자살률을 기록하였다(중앙자살예방센터, 2018). 자살의 원인은 개인의 상황과 당시 사회 상황에 따라서 달리 나타날 수 있으나 복합적인 요인에 의한 것이라 할 수 있다. 많은 연구와 발간자료,

* 이 논문은 2019년 경상대학교 대학원 경영정보학과 김덕현의 경영학 석사 논문을 바탕으로 재수정하여 작성되었음.
** 경상대학교 경영정보학과, zoro0729@naver.com (주저자)
*** 경상대학교 경영정보학과, 경영경제연구소, dhyoo@gnu.ac.kr
**** 경상대학교 경영정보학과, 경영경제연구소, dyjeong@gnu.ac.kr (교신저자)

보고서 등을 통해 65세 이상에 해당하는 노인의 자살률이 다른 연령대에 비하여 높은 이유로써 기대수명의 연장 및 건강 악화, 가구 구성의 간소화, 그리고 노인층의 빈곤문제 등이 거론되고 있다.

자살률은 인구집단에서 큰 부분을 차지하는 연령층에서 높게 나타나는 경향이 있으므로, 점차 증대되는 노인 인구에 비례하여 노인 자살률이 증가할 것으로 전망된다. 초고령 사회로의 진입이 멀지 않은 한국의 경우 향후 노인 자살율이 급속히 증가할 것으로 예상된다(중앙자살예방센터, 2018). 이에 최근 급증하는 고령 인구의 자살의도에 영향을 미치는 요인들을 체계적은 도출하여 그 원인을 파악함으로써, 노인들의 자살을 예방하고 자살문제를 해결하기 위한 다양한 연구들이 시도되고 있다(권준동 등, 2011; 권오균, 허준수, 2013; 이금룡, 조은혜, 2013; 박민정, 2015; 김지훈, 김경호, 2018; Bonnewyn et al. 2009). 대표적으로 자살을 다차원적이고 단계적으로 구분하여 해석함으로, 이에 영향을 미치는 요인들을 규명하는 연구들을 들 수 있다(문동규, 2012; 오윤정, 김향동, 2018; Mann et al. 2005).

최근 빅데이터 기반의 의사결정 플랜폼의 개발이 활발히 진행되고 있다(장영재, 2015). 본 연구에서는 노인 자살의 예방과 해결을 위하여 빅데이터 기반의 자살생각 예측모형을 개발하였다. 예측모형의 개발을 위하여 한국복지패널에서 제공하는 패널데이터를 활용하였으며, 데이터 균형화 기법이 가미된 의사결정나무 모형 분석기법을 이용하였다. 이를 통하여 노인의 자살생각에 관한 예측모형을 개발하였고, 이를 토대로 의사결정 규칙들을 도출했다. 또한 예측모형의 정확도를 높이기 위하여 오버샘플링과 언더샘플링 기법을 동시에 사용했다.

본 연구를 통하여 노인들의 자살생각에 영향을 미치는 여러 가지의 영향변수들을 체계적으로 도출할 수 있었으며, 이들 변수들이 자살생각을 예측하는데 얼마나 정확히 영향을 미치는가를 계상하였으며, 이들 변수들이 갖는 시사점들을 찾을 수 있었다. 그리고 노인들의 자살생각을 예방하기 위한 핵심변수들의 분석을 통하여 노인자살예방을 위한 정책을 수립하는데 이용할 수 있는 기초자료를 제시하고자 한다.

Ⅱ. 이론적 배경과 선행연구

2.1 예측모형 개발의 기초

(1) 의사결정나무 기법

한국복지패널에서 제공하는 패널데이터를 활용하여 노인자살생각을 예측하기 위한모형을 개발하기 위해서는 분류예측 기법이 효과적이다. 분류예측 기법으로 로지스틱 회귀분석, 다중판별분석, 의사결정나무 등이 있다(원하림 등, 2018). 본 연구에서는 예측을 위한 분류분석 기법으로 의사결정나무 기법을 이용한다. 의사결정나무기법은 인공신경망 기법에 비하여 목표변수와 투입된 독립변수군 사이에 레이어가 히든 레이어 형태, 즉 블랙박스 형태가 아니라 화이트박스 형태로 나타나기 때문에 규칙을 발견해내거나 탐사하기 쉽다. 또한 'IF-THEN' 형식이기 때문에, 준거점에 의한 모형의 형성을 쉽게 살펴볼 수 있고 설명이 용이하다. 이러한 이유로 의사결정나무를 본 연구의 분석기법으

로 택하였다.

의사결정나무는 의사결정문제 해결에 있어 중요한 기법 중의 하나이다. 데이터마이닝 작업에서의 의사결정나무는 탐색과 모형화라는 두 가지 특성을 모두 가지고 있다. 모수적 모형을 분석하기 위해서 사전에 이상치를 검색하거나 분석에 필요한 변수 또는 모형에 포함되어야 할 상호작용의 효과를 찾아내기 위해서 사용될 수도 있고, 의사결정나무 자체가 분류 또는 예측모형으로 사용될 수도 있다(Song and Lu, 2015; 강현철 등, 2014). 의사결정나무는 의사결정 규칙을 나무구조로 도표화하여 분류와 예측을 수행하는 분석방법으로 분류 또는 예측의 과정이 나무구조에 의한 추론규칙에 의해서 표현되기 때문에, 그 과정을 쉽게 이해하고 설명할 수 있다는 장점을 가진다.

(2) 데이터 균형화 기법

데이터 균형화란 특정 클래스가 비대칭적으로 적을 경우, 예측모형에 의한 예측률이 과적합을 일으키는 것을 방지하기 위한 대안으로 제시되고 있는 샘플링 기법을 의미한다. 예를 들면 본 연구의 목표변수인 자살생각에 대한 유무는 원본 데이터를 기준으로 전체 약 200,000건 중에서 3,000여 건에 불과하다. 이를 해결하기 위한 방법으로서 특정 클래스를 구성하는 집단을 메이저 집단과 마이너 집단으로 구분하고, 마이너 집단의 수와 동일하게 메이저 집단의 수를 하향하여 균등화시키거나 메이저 집단의 수와 동일하게 마이너 집단의 수를 상향하여 균등화시키는 것이 데이터 균형화 작업이다(Burez and Poel, 2009). 데이터 균형화에서 마이너 집단의 수에 맞추어 메이저 집단의 수를 랜덤하게 뽑는 것은 언더샘플링이고, 메이저 집단의 수에 맞추어 마이너 집단의 수를 랜덤하게 증가시키는 것은 오버샘플링이다. 데이터 균형화의 목적은 예측률의 상승 혹은 하향보다는 메이저 집단으로 편향되어 예측률이 높게 나타나는 현상을 방지하기 위함에 있다. 또한 목표변수에 대한 예측모형을 개발할 때에, 메이저/마이너 구분 없이 동일 수의 인스턴스를 투입하여 50:50 확률을 인위적으로 만들어 줌으로 목표변수의 값을 예측하는 데에 투입되는 독립변수들의 작용을 보다 중립적으로 살펴볼 수 있다.

김경민 등(2014)은 불균형 데이터 처리를 위한 과표본화 기반 앙상블 학습 기법을 연구하였는데, 여기에서 과표본화는 오버샘플링을 의미한다. 일반적인 기계학습 기법들은 학습데이터가 각 범주(클래스)당 비슷한 비율로 구성되어 있다고 가정하고 학습을 진행하는데, 보통의 실세계 문제들이 불균형 데이터 문제에 속하게 된다. 이러한 경우 소수 마이너 집단에 속한 데이터들은 메이저 집단에 속한 데이터보다 잘못 분류될 가능성이 높음을 언급하고 있다. 불균형 문제를 해결하기 위하여 오버샘플링을 통한 데이터 균형화를 도모하여 분류기의 과적응 문제를 해결할 수 있다. 또 다른 샘플링 기법인 언더샘플링을 통하여 데이터 균형화를 도모한 연구들이 있다. 유동희 등(2015), 안민욱 등(2018), 김원종 등(2018)의 연구에서는 데이터 마이닝을 통하여 예측 모형을 개발하는 연구를 진행하였는데, 이들의 연구에서 데이터 균형화 기법으로 언더 샘플링을 활용하였다. 예측모형의 편향은 목표변수 내의 메이저 집단에 속하는 클래스의 수가 많을 때 일어나는데, 마찬가지로

목표변수 내의 하위집단 인스턴스 수를 동일하게 만들어 줌으로 불균형 문제를 해결할 수 있다.

본 연구에서는 연구의 자료에 대한 데이터 균형화 유무를 기준으로, 데이터 균형화 작업을 진행하지 않은 원본 데이터 셋에 대한 예측모형을 우선적으로 구축하고자 한다. 이를 통해 실제 불균형 상태의 데이터를 투입하여 예측모형을 구축할 때 발생하는 문제점을 살펴보고자 한다. 또한 데이터 균형화의 필요성에 입각하여 언더샘플링과 오버샘플링을 각각 적용하여 하위 데이터 셋을 추출한다. 이를 통해 언더샘플링의 장·단점과 오버샘플링의 장·단점을 확인하고 비교해보고자 한다.

2.2 노인 자살생각 영향변수

노인의 자살 예방과 해결을 목적으로 한 연구들에서 대표적으로 다루는 주요 변수들을 본 절에서 살펴보고자 한다. 앞서 언급한 대로 자살은 고의적 자해에 의한 죽음으로 정의되며, 자살로 인하여 생애를 마감한 경우 사인을 판명할 수는 있으나 해당 행위에 대한 직접적인 원인 규명이 어렵다. 즉, 자살로 인한 사망은 당사자가 죽음으로 인한 부재 상태에 놓이게 되기 때문에 치료 등의 후속 조치가 어려우며, 이러한 점에서 자살을 단계적으로 구분하고 접근하는 것이 중요하다.

자살을 행동 여부나 상황에 따라 구분하는데, '자살생각→자살행위→자살완수'라는 3단계로 구분하는 것이 일반적이다. 즉, 자살생각 혹은 자살의사 등으로 표현되는 자살에 대한 생각을 하는 단계, 자살시도 혹은 자살행위 등으로 표현되는 자살을 시도하는 단계, 자살행위로 인한 죽음에 이르는 자살완수의 단계로 각각을 정리할 수 있다(Mann et al., 2005; Bonnewyn et al.. 2009). 노인의 자살과 관련된 대부분의 연구에서는 자살생각을 중요한 연구변수 혹은 목표변수, 또는 종속변수로 설정하고 있다. 그 이유는 다음과 같다. 개인의 자살생각을 억제하거나 개입함으로 자살 시도를 미연에 방지하기 위해서이다. 또한, 가족과 타인 및 사회가 적극적인 개입이 가능한 사실상의 마지노선에 가깝기 때문이다. 즉, 자살생각을 하고 있거나, 했던 유경험자들에 대한 관련 요인을 탐색하고, 그 이후의 단계로 이어지지 않게 적극적인 개입을 시도하는 것이 자살 예방 및 해결책으로 제시되고 있다.

노인들의 자살생각에 영향요인들을 무수히 많다. 이 중에서 우울감이 자살생각에 미치는 영향변수로 대부분의 연구에서 나타났고(이인정, 2011; 문동규, 2012; 이금룡, 조은혜, 2013; 권중돈 등, 2011; 김지훈, 김경호, 2018; Bonnewyn et al.. 2009), 이전의 자살시도 경험 또한 자살생각에 미치는 영향이 큼을 알 수 있다(김종필, 현미열, 2013). 이 외에도 개인의 인구통계학적 요인(이인정, 2011; 권중돈 등, 2011), 개인의 심리적 요인(문동규, 2012; 김지훈, 김경호, 2018), 외부 환경적 요인(오윤정, 김향동, 2018; 김지훈, 김경호, 2018), 경제적 요인(문동규, 2012; 박민정, 2015; 오욱찬 등, 2017) 등과 같은 다양한 요인들이 복합적으로 자살생각 경험 여부에 영향을 주고 있음을 알 수 있다. 선행 연구들을 바탕으로 본 연구에서는 수집 자료 내의 활용 가능한 측정항목 및 변수를 최대한 추출했다.

<표 1> 선행연구 정리

연구자	주요변수	종속변수	연구대상자	자료 분석 방법	핵심연구결과
이인정 (2011)	우울 위기사건 사회적 지지	자살생각	서울, 수도권 노인	위계적 회귀분석	위기 사건, 가족지지 조절효과 입증
김종필 · 현미열 (2013)	우울 자살시도	자살생각	J도 내 치매노인	위계적 회귀분석	치매 노인의 경우 우울 정도가 심함, 과거 자살시도 경험은 재시도를 유발함.
권중돈 등 (2011)	우울 자살시도 음주	자살생각	서울 내 독거노인	다중회귀분석	복지서비스 이용자들에 대한 자살시도경험과 문제적 음주의 조절효과 입증
이혜경 (2016)	애도수준 우울 건강상태	자살생각	사별경험 독거노인	구조방정식모형	애도수준과 자살생각 간의 관계에 대한 우울의 완전매개 효과 입증
권오균 · 허준수 (2013)	자존감 우울감 절망감	자살생각	저소득 독거노인	구조방정식모형	설정한 모든 독립·종속변수에 대한 우울감의 매개효과 입증
문동규 (2012)	개인자살력 우울 노인차별	자살생각	노인 자살생각 논문	메타분석	3가지 변인군으로 나누어 각각의 효과크기 분석
오욱찬 등 (2017)	가계부채 우울감 식생활어려움	우울, 자살생각	한국복지 패널 성인	로지스틱 회귀분석	전 연령대에 대한 연구, 가계부채의 영향력 입증, 노인층에 대한 언급 부족
박민정 (2015)	우울감 성별 가족갈등	자살생각	한국의료 패널 노인	다중로지스틱 회귀	성별에 따른 주요 변수 차이 남: 주관적 건강상태 여: 배우자 유무
김지훈 · 김경호 (2018)	우울 자아존중감 부부폭력	자살생각	한국복지 패널 성인	로지스틱 회귀분석	전 연령대에 대한 연구, 3가지 변인군으로 나누어 분석
이윤정 (2012)	정보화상태 우울	우울, 자살생각	노인 실태자료	로지스틱 회귀분석	정보화 상태는 우울 감소 역할 자살생각 증가 역할
오윤정· 김향동 (2018)	연령 소득계층 우울	자살생각	D시 내 노인	위계적 다중회귀	우울, 타인지지, 자아존중감 등이 강력한 예측요인으로 작용
김성은 · 김선아 (2013)	총 13개의 독립변수 투입	우울	농촌거주 노인	의사결정나무	운동능력, 자아존중감, 농사참여 등이 농촌노인의 우울 예측요인으로 대표됨
박명화 등 (2013)	총 49개의 독립변수 투입	우울	노인 실태자료	의사결정나무	삶의 만족도, 일상생활 수행능력, 영양상태 등이 노인의 우울 예측으로 대표됨
임성욱 · 김경희 (2018)	연령 우울	자살생각	노인 실태자료	구조방정식모형	전기 · 후기 노인에 대한 자살생각 요인 차이 입증
본 연구	총 98개의 독립변수 투입	자살생각	한국복지 패널 노인	의사결정나무	해당 절에서 서술

이들 연구 중에서 본 연구와 비슷한 기법을 활용한 선행연구들을 살펴보면 다음과 같다. 김성은과 김선아(2013)는 의사결정나무 분석을 통해 농촌거주 노인의 우울예측 모형을 구축하였는데, 연구 대상자로 경상북도와 강원도의 30개 농촌지역에 거주하는 65세 이상 노인을 선정하였다. 실험에 투입된 변수에는 목표변수로 우울상태, 독립변수로 성별, 연령, 교육수준, 월평균 소득, 결혼상태, 만성질환 수, 인지기능, 자아존중감, 운동능력, 사회활동참여 등이 포함된다. 또한 박명화 등(2013)은 의사결정나무 분석을 통해 우울 노인의 특성을 분석하였는데, 연구자료는 2008년도 노인실태조사 자료를 활용하였다. 연구대상자는 만 60세 이상의 노인으로, 목표변수에 우울을 두고 초기 독립변수 53개와 선별 작업을 거쳐 선별한 중요 독립변수 49개를 투입한 예측모형을 개발하였다. 연구 대상자의 일반적 특성 관련 변수, 가족 및 사회적 관계 관련 변수, 경제 상태 관련 변수, 건강상태 관련 변수, 건강행태 관련 변수, 기능상태 관련 변수 등이 이에 속한다.

위에서 언급한 이들 연구들의 특징을 살펴보면, 전체 데이터를 학습 데이터와 검증 데이터로 나누어 분석함으로 예측모형을 구축하고 검증하였다. 성능검증지표로 언급되는 것이 정·오분류율을 바탕으로 작성한 Confusion Matrix인데, 여기에서 분류기의 과적응 문제를 파악할 수 있다. 즉, 목표변수의 메이저 집단에 대한 분류는 더 잘 이루어지고, 마이너 집단에 대한 분류는 상대적으로 잘 이루어지지 않는 편향된 모형이 구축된 것으로 보인다. 끝으로, 투입되는 독립변수의 중요도를 산정하고, 불필요한 변수를 제거하는 래퍼 방식의 부재 또한 보완되어야 할 부분이다.

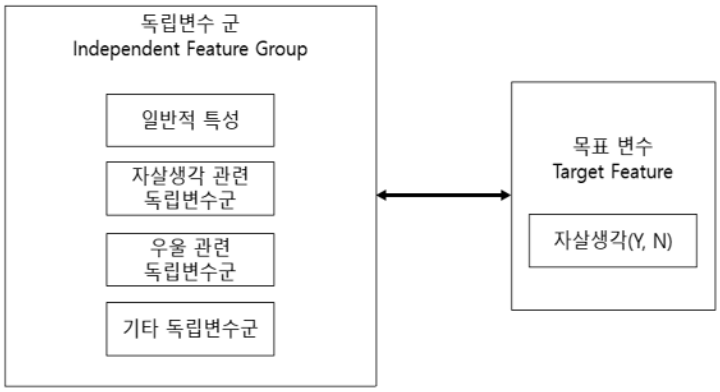
지금까지 노인 자살생각과 관련된 선행 연구들에 대하여 사용한 변수, 연구대상, 자료 분석 방법, 핵심 연구결과 등을 중심으로 정리하면 <표 1>과 같다. 이들 연구의 대부분은 인과관계 중심의 통계적 연구방법론이며, 의사결정나무 기법을 적용한 연구(김성은, 김선아, 2013; 박명화 등, 2013)는 몇 개에 불과하다. 또한 의사결정나무 분석기법을 적용한 연구의 경우 데이터 균형화 작업이나 래퍼 방식이 적용되지 않았다. 이와 달리 본 연구는, 상관관계 중심의 의사결정나무 분석을 시행하였고, 데이터 불균형 현상으로 인한 과적합 문제를 해결하기 위한 방안으로 언더샘플링과 오버샘플링을 동시에 진행하여 비교하였다.

Ⅲ. 연구방법

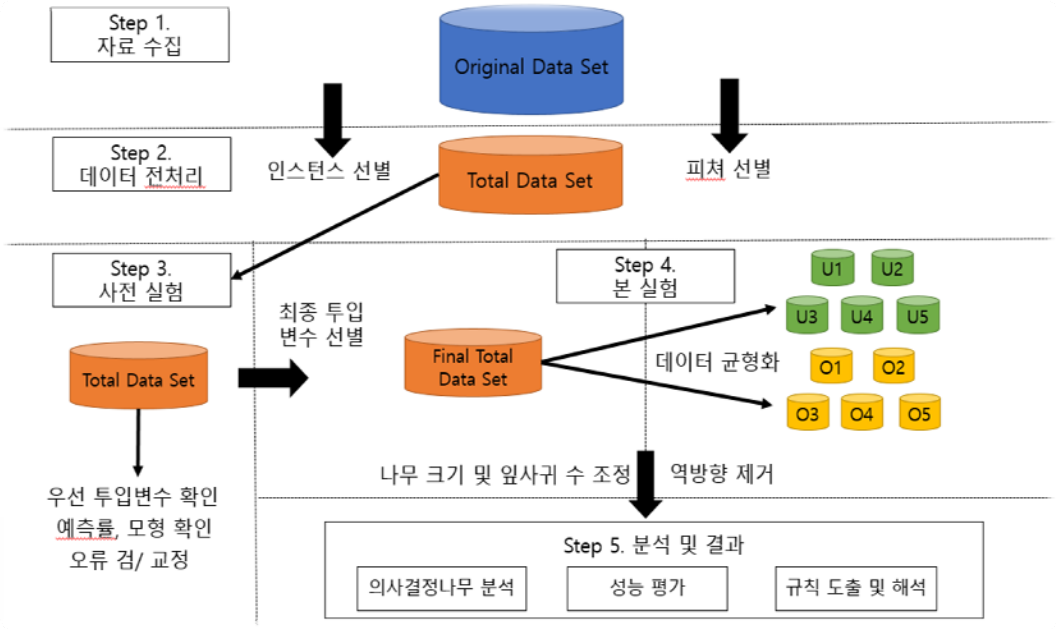
3.1 연구모형

앞의 제 2장에서 정리된 선행 연구를 바탕으로 목표변수와 영향 관계가 있을 것으로 판단되는 변수들을 추출했다. <표 1>에서 정리된 연구결과에 기초하여 연구 대상자의 일반적 특성, 자살생각과 관련된 독립변수군, 우울관련 독립변수군, 기타 독립변수군으로 대분류를 해볼 수 있다. 본 논문에서는 노인들의 자살생각에 영향을 미치는 변수를 분석하기 위하여 <그림 1>과 같이 연구 모형을 제시한다.

연구 대상자의 일반적 특성에는 연령, 성별, 지역 구분, 경상소득, 가구 관련 변수, 학력, 장애 관련 변수 등이 있다. 자살생각과 우울관련



<그림 1> 연구 모형



<그림 2> 분석과정

독립변수군은 선행연구에 가장 많이 언급된 자살계획, 자살행위, 경제적 어려움 경험, 생활 만족도, 자아존중감, 가정생활에 대한 스트레스, 음주 관련 변수, 흡연 여부, 위기사건 관련 변수, 사회적 지지 관련 변수 등을 고려하였다.

기타 독립변수군은 측정 등의 어려움으로 선행연구에서 다루지 않았으나, 본 연구에서는 밀접한 연관이 있을 것으로 판단되는 변수들을 함께 투입하였다. 대표적으로 연금, 급여, 금융자산, 부동산 보유 정도, 주관적 최저생활비, 주관적 적정생활비, 주거 환경 관련 변수 등이 있다.

3.2 분석과정

<그림 2>는 연구자료의 분석과정을 도식화하고 있다. 분류분석을 위한 데이터마이닝은 일반적으로 데이터 수집, 데이터 전처리, 모형 구축, 성능 평가, 규칙 도출의 단계를 거쳐 진행된다. 본 연구에서도 일반적인 데이터마이닝의 기본적인 단계를 거쳤으며 일부 필요한 부분을 가미하여 진행하였다. 데이터 수집단계에서는 한국복지패널에서 제공하는 패널데이터 중 1-12차 결합 데이터를 선정하였다. 데이터 전처리 단계에서는 불필요한 인스턴스와 피처를 제거하고, 본 연구에 적합한 데이터 셋으로 가공하였다.

본 연구에서는 모형 구축, 성능 평가 단계에 앞서 가공된 데이터 셋을 바탕으로 사전 실험을 진행하였다. 사전 실험을 진행함으로 데이터 전처리 과정에서 존재할 수 있는 오류에 대해 검·교정하였고, 본 실험에 투입될 목표변수와 독립변수군을 확정하였다. 본 연구의 데이터마이닝 도구로서 Weka 3.8을 활용하였는데, 의사결정나무 기법의 알고리즘은 C4.5로 알려진 J48을 사용했다. 또한 66.6%로 전체 데이터를 학습데이터와 검증데이터로 나누어 분석을 시행했다.

본 실험 과정에서는 데이터 균형화를 위한 샘플링 기법을 적용하여, 총 데이터 셋 11개(데이터 균형화를 적용하지 않은 원본 데이터 셋 1개, 오버샘플링 데이터 셋 5개, 언더샘플링 데이터 셋 5개)에 대한 모형 구축을 시행하였다. 추가적으로 모형 구축 단계에 있어, 예측률과 나무의 크기, 잎사귀의 수와 관련이 있는 '리프 노드 내의 최소 인스턴스 수'인 minNumobj(이하 m.NO)를 조정하여 분석을 시행하였다. m.NO 값의 조정은 예측률 손실의 최소화와도 관련이 있다.

구축된 모형을 바탕으로, 각각 예측률의 상승 및 불필요한 변수 제거를 위한 역방향 제거를 수행하였으며, 이에 대한 개별 예측률을 탐색하였다. 성능평가와 규칙 도출 단계 앞서, 오버샘플링과 언더샘플링이 적용된 데이터 셋 내에서 적합한 데이터 셋을 각각 하나씩 선택하였다. 이후 원본 데이터 셋 1개, 오버샘플링 데이터 셋 1개, 언더샘플링 데이터 셋 1개에 대한 의사결정나무의 그래프를 비교·분석하고 성능을 평가하며, 의사결정규칙을 도출하였다. 최종적으로 이것을 해석하고 시사점을 찾았다.

3.3 자료 특성

연구 자료는 한국복지패널에서 제공하는 패널데이터로서 이로부터 추출한 측정항목은 가공·조작·처리하여 예측모형 개발에 적합한 변수들로 재구성하는 방식을 선택하였고, 사전 실험을 통해 본 실험에 투입될 최종 독립변수군을 한번 선별하는 과정을 거쳤다. 한국복지패널 자료는 국내에서 수행 중인 가구단위 패널조사 중 한국의료패널조사 다음으로 규모가 큰 패널조사이며, 최초 원표본 가구규모는 7,072가구로 시작하였고, 이후 조금씩 감소하였으나 2017년 기준에 조사가 완료된 원표본가구가 약 4,400가구에 이르고 있다. 한국복지패널 자료는 지역적, 가구유형 구분 등에 있어서 한국 내에 존재하는 대부분의 가구유형을 포함하기 때문에 패널조사로서는 드물게 전국적인 대표성을 지니고 있다. 끝으로, 저소득층 연구에 적합

한 패널이며 학제 간 연구가 가능한 패널조사로 소개되고 있다.

한국복지패널에서 제공하는 데이터의 형태는 원자료(source data) 형태로, 응답자(인스턴스)와 각각의 응답 항목으로 이루어져 있다. 본 연구는 선행연구에 조사된 사회적, 개인적 환경과 관련된 변수들을 가능한 한 모두 포함시키려 노력하였다. 원자료 형태의 패널데이터를 데이터 전처리 과정을 거쳐 선별한 연구 대상자는 총 6,536명으로, 이에 대한 일반적 특성은 다음과 같다.

먼저, 거주지역 구분(F1)에서 도시 거주자가 2,254명으로 가장 많았고, 권역별 지역구분(F2)에서 광주/전남/전북/제주 지역에 1,269명으로 가장 많았다. 성별(F7)에서 남성이 2,584명, 여성이 3,952명으로 여성의 비중이 조금 더 높다(<표 2> 참조). 평균 연령(F8)은 77.8±7.6세이다. 교육수준 1(F9)에서 초등학교가 2,682명으로 가장 많았고, 무학 1,499명, 중학교 1,059명으로 초등학교 이하의 저학력자가 전체의 63%로 나타났다. 교육수준 2(F10)에서 졸업이 4,046명, 무학으로 인한 비해당이 1,499명, 중퇴에 986명으로 나타났다. 장애종류(F11)와 장애등급(F12)에는 장애 없음이 5,410명(87%)이였으며, 장애가 있는 경우 지체 장애가 667명으로 가장 많았고, 장애 등급은 5급이 262명, 4급이 254명, 6급이 239명 순으로 나타났다.

연간 평균 경상소득(F3)은 2,526만원이고, 균등화 소득에 따른 가구구분(F4)에서 일반가구에 2,542명, 저소득층 가구에 4,002명이었으며, 평균 가구원 수(F5)는 2.1±1.1명이고, 가구형태(F6)는 단독이 1,928가구, 모자가 1가구, 부자가 3가구, 조손이 58가구, 기타에 4,546가구로 나타났다. 혼인상태(F13)에서 배우자 있음이 3,818명, 사별 2,392명이었다. 이외 하위 집단에 이혼, 별거, 미혼이 있다. 종교(F14)는 종교 있음에 3,755명, 종교가 없음에 2,781명으로 종교가 있는 집단의 비중이 조금 더 크다. 동거여부(F15)에는 동거가 4,574명으로 가장 많았고, 독거가 1,926명, 이외 하위 집단에 다른 지방에 근무, 회외 근무, 입원/요양 등이 있다.

끝으로 목표변수인 자살생각과 자살 관련 변수로 자살계획(F97), 자살시도(F98)가 있다. 3개 피처 모두 지금까지의 경험 유무에 대한 응답으로 경험 유(Y), 경험 무(N)로 하위 집단이 2개로 구성된다. 자살계획은 ‘경험이 있다(Y)’에 202명, ‘경험이 없다(N)’에 6,334명이 속했다. 자살시도는 ‘경험이 있다(Y)’에 81명, ‘경험

<표 2> 인구통계적 특성

투입변수	속성	표본수	투입변수	속성	표본수
거주지역 구분 (F1)	서울	824	권역별 지역 구분 (F2)	서울	824
	광역시	1,669		인천/경기 지역	1,165
	도농복합군	237		부산/울산/경남	1,156
	군지역	1,552		대구/경북	999
성별 (F7)	남성	2,584		대전/충남	572
	여성	3,952		강원/충북	551
가구구분 (F4)	일반가구	2,542		광주/전남/전북/제주	1,269
	저소득층	4,002			

이 없다(N)’에 6,455명이 속하였다. 목표변수인 자살생각은 ‘경험이 있다(Y)’에 1,514명, ‘경험이 없다(N)’에 5,022명이 속하였다.

Ⅳ. 실험결과 및 분석

4.1 실험 결과

(1) 사전 실험 결과

사전 실험(pilot test)을 실행한 이유는 두 가지로 축약할 수 있다. 본 실험에 앞서 실수나 오류를 최대한 줄이고자 함과 연구자가 처리한 변수의 활용도에 있어 최적의 조합을 찾기 위함이다. 총 2회에 걸쳐 실험을 진행했는데, 첫 번째로 데이터 전처리 과정을 거쳐 선정한 피처를 모두 투입하고 목표변수에 대한 예측 모형 구축과 역방향 제거를 반복적으로 수행하였다. 두 번째로, 수치형 데이터를 포함하는 개별 측정항목의 총합 변수 모두 제거하고, 해당 총합의 평균을 통한 범주화 과정을 거친 변수만을 투입하였다. 마찬가지로 목표변수에 대한 예측 모형 구축과 역방향 제거를 반복적으로 수행하였다. 사전 실험 1에 투입된 독립변수의 수는 123개이고, 사전 실험 2에 투입된 독립변수의 수는 98개이다.

사전 실험의 방법은 다음과 같다. 전체 데이터를 66.6%의 비율로 학습 데이터와 검증 데이터로 나누어 모형을 구축했다. 의사결정나무 알고리즘은 C4.5(J48)을, 변수 중요도 산정의 기준으로 이득비(gain ratio)를 사용했다. 역방향 제거 방식을 도입하기 전과 후로 크게 나눌 수 있으며, 리프 노드의 최소 인스턴스 수(m.NO)에 따른 예측률과 나무의 크기, 잎사귀의 수를 바탕으로 최선과 차선으로 선정하여 예측률을 평가했다.

123개와 목표변수 1개로 이루어진 사전 실험의 결과 m.NO를 100으로 설정하였을 때 예측률이 83.39%으로 나타났다. 이를 토대로 이득비를 통한 변수 중요도를 산정하였고, 최하위 독립변수부터 역방향 제거를 하였다. m.NO의 값이 20일 때 역방향 제거 방식을 통해 84.11%까지 예측률이 상승했다. 또한 수치형 데이터를 포함하는 개별 측정항목의 총합 변수 모두 제거하고, 해당 총합의 평균을 통한 범주화 과정을 거친 변수만을 투입한 결과, m.NO가 10인 경우, 역방향 제거 작업을 진행하여 예측률이 83.75%까지 상승했다.

사전 실험의 결과는 데이터 균형화가 적용되지 않은 데이터 셋에 대한 예측률이기 때문에, 자살생각의 하위 집단 중 메이저 집단에 해당되는 N에 편향된 예측 결과를 나타내었다. 대부분의 연구들에서는 편향된 예측 결과의 문제를 간과하고 있지만, 실제 데이터를 처리하고 분석하여 결과를 살펴보는 데에 있어 목표변수 내의 데이터 불균형 문제는 상당히 치명적일 수 있다. 그렇기 때문에, 본 연구의 주된 관심이자 기법인 데이터 균형화 기법의 투입으로 인한 변화 양상을 살펴볼 필요가 있다.

(2) 본 실험 결과

본 실험은 사전 실험의 결과를 통해 수정해야 할 부분과 보완해야 할 부분을 고치고, 독립변수의 투입을 최적화하는 작업을 진행한 후 시도되었다. 즉, 불필요한 피처를 탈락시키고, 선별된 독립변수군을 본 실험에 투입하였다. 본

실험의 실험방법도 사전 실험과 동일하다. 전체 데이터를 66.6%의 비율로 학습 데이터와 검증 데이터로 나누어 모형을 구축했다. 의사결정나무 알고리즘은 C4.5(J48)을, 변수 중요도 산정 기준으로 이득비를 사용했다. 또한 본 실험에서는 데이터 균형화를 위한 샘플링 기법이 도입되는데, 언더샘플링의 경우 상위의 실험 방법과 동일하다. 하지만 오버샘플링의 경우 오버샘플링 데이터 셋 5개 모두 m.NO가 2일 때 최고 예측률을 보였다. m.NO 값이 2인 예측 모형은 나무의 크기와 잎사귀의 수의 값이 워낙 크기 때문에, 이를 중재할 구간을 찾는 것을 목표로 3개의 m.NO 값을 선정하여 역방향 제거 작업을 진행했다. 역방향 제거 작업은 상위의 과정과 동일하며, 역방향 제거로 인한 제거 변수와 예측 모형의 예측률을 기입하였다. 데이터 균형화가 적용되지 않은 원본 데이터 셋 1개, 데이터 균형화가 적용된 언더샘플링 데이터 셋 5개, 오버샘플링 데이터 셋 5개로 총 11개 데이터 셋에 대한 본 실험 결과는 <표 3>과 같다.

원본 데이터 셋에 대한 분석결과는 다음과 같다. 자살생각(Y/N)을 목표변수로 하고, 98개의 독립변수를 투입하였으며, m.NO를 7개 단위(2, 10, 20, 50, 100, 150, 200)로 조정하여 모형을 구축했다. 원본 데이터 셋의 인스턴스 수는 6,536개이며, 목표변수의 하위 집단이 동일한 수의 인스턴스를 갖지 않는다. 즉, 해당 피처 내 특정 클래스의 샘플 수가 많은 불균형 상태

<표 3> 실험 결과 종합 및 비교 대상 선정

구분	모든 변수투입				역방향제거 완료			비교대상 선정	성능지표			
데이터 셋	m. NO	예측률	m. NO	예측률	m. NO	최고 예측률	투입 피처	선정 이유	Y TP rate	N TP rate	Y FP rate	N FP rate
Original	20	81.2781	50	81.5032	20	81.7732	21개	v	0.371	0.955	0.045	0.629
Under1	10	80.5825	20	79.8058	10	81.4470	23개	**원본 데이터 선정이유** 자살생각 N에 편향된 예측모형으로 비교대상으로 적합				
Under2	10	81.2621	50	81.1650	10	82.9126	39개	v	0.820	0.836	0.164	0.180
Under3	20	77.9612	50	77.2816	20	79.2233	43개	**2번 언더샘플링 선정이유** 언더샘플링 데이터 셋에서 최고 예측률 기록, 나무의 크기, 잎사귀의 수가 모두 적정수준, 그러므로 m.NO값 변동으로 인한 예측률 소실 고려 X				
Under4	20	78.6408	50	77.8641	20	78.6408	98개					
Under5	2	76.5192	50	78.2524	2	79.1262	11개					
Over1	2	84.1288	100	75.0805	2	84.9780	59개					
Over2	2	83.7775	150	79.9341	2	84.5388	46개	v	0.901	0.792	0.099	0.208
Over3	2	84.0117	100	75.0220	2	85.0659	70개	**2번 오버샘플링 선정이유** m.NO 값이 2일 때, 나무의 크기, 잎사귀의 수 과다, m.NO 조정 필수, m.NO 값 변동으로 인한 예측률 소실이 가장 적음 (소실값 = 0)				
Over4	2	83.5725	100	74.5242	2	85.2416	58개					
Over5	2	84.3610	100	75.4026	2	85.5051	68개					

이다. m.NO 값을 선정하는 방법으로, 나무의 크기를 기준으로 한 그룹 구분 없이 예측률 1, 2위를 기록한 경우를 선택하였다. 이때 예측률은 역방향 제거 작업을 진행하기 전에 m.NO의 값을 확정하기 위한 단계로, 모든 변수 투입된 예측률이다. 예측률 상승과 불필요한 변수 삭제라는 목표를 도모하기 위해 역방향 제거를 각각 수행하였고, 이에 해당하는 m.NO값, 예측률, 변수 제거, 변수 투입, 나무의 크기, 규칙의 수는 다음과 같다. m.NO가 20일 때, 예측률이 81.27%로 2위였으며, 나무의 크기는 93, 잎사귀의 수는 71개이다. m.NO가 50일 때, 예측률이 81.50%로 1위였으며, 나무의 크기는 19, 잎사귀의 수는 10개이다. 각각에 대한 역방향 제거 작업을 진행하였고, m.NO가 50인 경우에는 예측률의 변동이 없다가 점차 낮아졌다.

데이터 균형화를 적용한 언더샘플링 셋에 대한 분석결과는 다음과 같다. 여기서도 동일한 변수를 투입하였고, m.NO도 7개 단위(2, 10, 20, 50, 100, 150, 200)로 조정하여 모형을 구축했다. 언더샘플링은 목표변수의 마이너 집단의 수(Y=1,514)에 맞추어, 메이저 집단 내의 인스턴스를 무작위로 샘플링하였다. 각 언더샘플링 데이터 셋에 포함되는 인스턴스의 수는 3,028개였다. m.NO 값을 선정하는 방법으로, 나무의 크기가 100 이상인 그룹에서 최고 예측률을 기록한 1개, 100개 미만인 그룹에서 최고 예측률을 기록한 1개를 선정하였다. 이때 예측률은, 역방향 제거 작업을 진행하기 전에 m.NO의 값을 확정하기 위한 단계로 모든 변수 투입된 예측률이다. 예측률 상승과 불필요한 변수 삭제라는 목표를 도모하기 위해 역방향 제거를 각각 수행하였다. 그 결과 m.NO의 값이 10일 때, 최고 예측률은 82.91%, m.NO의 값이 50일 때, 최고 예측률은 81.165%로 나타났다.

데이터 균형화를 적용한 오버샘플링 셋에 대한 분석결과는 다음과 같다. 자살생각(Y/N)을 목표변수로 하고, 98개의 독립변수를 투입하였으며, m.NO를 7개 단위(2, 10, 20, 50, 100, 150, 200)로 조정하여 모형을 구축했다. 오버샘플링은 목표변수의 메이저 집단의 수(N=5,022)에 맞추어, 마이너 집단 내의 인스턴스를 중복 복제하였다.

m.NO 값을 선정하는 방법에 차이가 있는데, 모든 오버샘플링 셋에서 가장 좋은 예측률을 기록한 m.NO 값이 2이다. 이를 포함한 나무의 크기가 100 이상인 그룹에서 예측률의 1, 2위를 기록한 2개, 100개 이하인 그룹에서 최고 예측률을 기록한 1개를 더하여, 총 3개를 선정하였다. 나머지 역방향 제거 진행 방식은 동일하다. 5개의 오버샘플링 데이터 셋 중에서 나타난 최고 예측률은 85.50%로, 5번 언더샘플링의 m.NO의 값이 2일 때이다. 오버샘플링 셋 당 산출된 최고 예측률 15개 중에서 최저치는 74.87%로 4번 언더샘플링의 m.NO가 100일 때이다. 오버샘플링 데이터 셋 내에서 최고 예측률을 기록한 것은 5번 오버샘플링 셋임에도 불구하고, 2번 오버샘플링 데이터 셋을 선정한 이유는 다음과 같다. m.NO의 값이 2와 10일 때 최고 예측률 기록 시에 투입된 독립변수가 총 46개로 동일하다는 점과, m.NO가 150의 경우에도 최고 예측률(74.99%)이 나타나는 구간이 투입된 독립변수가 67개에서 37개까지 동일하다. 그렇기 때문에 m.NO의 값을 조정하는데 있어 예측률의 소실이 전혀 없다. 이러한 이유로, 오버샘플링 데이터 셋 전체에서 최고 예측률을

기록한 5번 데이터 셋 대신에 2번 데이터 셋을 선정하였다.

4.2 차이분석 및 성능평가

(1) 데이터 균형화 유무에 따른 차이 분석

데이터 균형화는 목표변수의 하위 집단의 인스턴스 수를 균등하게 하는 샘플링 기법이다. 대표적으로 언더샘플링과 오버샘플링이 있으며, 본 연구에서는 원본 데이터 셋, 언더샘플링 데이터 셋, 오버샘플링 데이터 셋을 구성하여 실험을 진행하였다. 우선, 원본데이터와 유사한 것은 오버샘플링 기법이 적용된 데이터이다. 피처의 중요도가 대부분 비슷하다. 이득비를 통한 변수의 중요도를 산정했을 때, 상위 10개에 해당하는 피처가 오버샘플링 하위 셋의 일부 순서 변동을 빼고 동일하다. 언더샘플링 기법이 적용된 데이터의 경우에는 원본데이터, 오버샘플링 데이터와는 달리 변수의 중요도가 차이를 보인다. 중요도 순위에서 하위권에 포진해있는 지역적 구분 변수 2가지(F1, F2)가 중요도 상위를 차지하고 있으며, 이사 경험(F64), 주거 위치(F37), 가족간의 다툼(F69) 등도 나타났다.

본 실험에서는 98개의 독립변수와 1개의 목표변수로 구성된 최종 투입 피처를 11개의 데이터 셋에 동일하게 투입하였다. 이득비를 통해 변수의 중요도를 산정하고, 역방향 제거 작업을 진행했다. 11개의 데이터 셋에 목표변수인 자살생각만 남을 때까지 모든 독립변수를 제거하였다. 역방향 제거 작업을 진행하면 예측률의 변동을 살펴볼 수 있다. 변동이 없거나, 하락하거나, 상승하는 등의 변동이 발생하고 어느 수준에 이르면 예측률이 점점 떨어지는 것을 살

<표 4> 최상위 3개 변수 투입 시 데이터 균형화에 따른 예측률 차이

Total Data, No Split									
원본 데이터			2번 오버샘플링				2번 언더샘플링		
m.NO	20	50	m.NO	2	10	150	m.NO	10	50
투입변수	예측률	예측률	투입변수	예측률	예측률	예측률	투입변수	예측률	예측률
역방향 제거완료	76.836	76.836	역방향 제거완료	50	50	50	역방향 제거완료	50	50
F97	79.8348	79.8348	F97	56.591	56.591	56.591	F97	56.5059	56.5039
F98, F97	79.8348	79.8348	F98, F97	56.9494	56.9494	56.591	F98, F97	56.8692	56.5059
F93, F98, F97	79.8348	79.8348	F93, F98, F97	72.3517	72.3517	72.1924	F93, F98, F97	70.9709	70.9709
Total Data: 66.6%, Training Data % Test Data									
원본 데이터			2번 오버샘플링				2번 언더샘플링		
m.NO	20	50	m.NO	2	10	150	m.NO	10	50
투입변수	예측률	예측률	투입변수	예측률	예측률	예측률	투입변수	예측률	예측률
역방향 제거완료	75.5176	75.5176	역방향 제거완료	49.019	49.019	49.019	역방향 제거완료	48.5437	48.5437
F97	78.8029	78.8029	F97	56.6618	56.6618	56.6618	F97	57.9612	57.9612
F98, F97	78.8029	78.8029	F98, F97	56.9546	56.9546	56.6618	F98, F97	57.9612	57.9612
F93, F98, F97	78.8029	78.8029	F93, F98, F97	71.9766	71.9766	71.8302	F93, F98, F97	72.7184	70.3883

펴볼 수 있었다. 특히나 변수 중요도 산정에서 상위 10개에 랭크된 피처들을 제거하기 시작하면 예측률 하락이 눈에 띄게 드러났다. 그리고 우울(Depress, F93)을 제거할 때 가장 큰 낙폭으로 예측률이 떨어졌다.

하지만 데이터 균형화를 거치지 않은 원본 데이터의 경우에는 예측률 하락이 크게 일어나지 않았다. 데이터 균형화를 거친 언더샘플링과 오버샘플링의 경우에는 예측률 하락이 크게 일어났다. 우울(F94), 자살시도(F98), 자살계획(F97)만을 독립변수로 투입했을 때 의사결정나무 분석을 한 결과는 <표 4>와 같다. 표의 상단 부분은 전체 데이터를 학습 데이터와 검증 데이터로 나누지 않고 분석을 시행한 결과 값이고, 표의 하단 부분은 전체 데이터를 66.6% 비율로 학습 데이터와 검증 데이터로 나누어 분석을 시행한 결과 값이다.

분석결과를 정리하면, m.NO의 값이 다르더라도 샘플링 기법에 따라 값이 일관되게 나타났다. 또한 우울 피처를 제거할 때 오버샘플링과 언더샘플링을 적용한 경우 낙폭이 크다는 점이다. 그리고 전체 데이터를 학습 데이터와 검증 데이터로 나누지 않은 경우 각 하위 집단의 샘플 수가 그대로 나타나고, 나눈 경우 예측률의 감소가 발생한다는 점이다. 또한, 원본 데이터의 경우 여전히 70%를 상회하는 데 반해, 언더샘플링과 오버샘플링의 경우 50퍼센트 이하로 예측률이 나타났다. 결론적으로 목표변수 내에 메이저 집단의 샘플 수가 마이너 집단의 샘플 수보다 월등히 많을 경우, 샘플의 수가 많은 메이저 집단에 대한 학습이 많이 이루어져 특정 클래스(메이저 집단)만을 잘 예측하는 편향된 예측 모형이 구축된다는 점과 일치한다. 또한 데이터 균형화 작업을 수행한 오버샘플링 및 언더샘플링은 목표변수 내의 클래스 비율을 동일하게 맞추었기 때문에, 특정 클래스만을 잘 예측하는 편향된 예측 모형이 구축되지 않음을 검증하였다.

(2) 예측모형의 성능평가

예측 모형의 성능 평가의 기준으로 정・오분류율이 중요하다. 본 실험의 정・오분류표인 Confusion Matrix를 작성하면 <표 5>와 같다. 의사결정나무 분석에 있어 전체 데이터를 학습 데이터와 검증 데이터로 나누어 진행하였는데,

<표 5> 비교 대상(원본, 오버, 언더)의 Confusion Matrix

Confusion Matrix Test Data(33.3% of Total Data)											
원본 데이터				2번 오버샘플링				2번 언더샘플링			
Class	TP Rate	FP Rate	Precision	Class	TP Rate	FP Rate	Precision	Class	TP Rate	FP Rate	Precision
N	0.962	0.662	0.818	N	0.864	0.262	0.778	N	0.774	0.275	0.745
Y	0.338	0.038	0.745	Y	0.738	0.136	0.837	Y	0.725	0.226	0.755
원본 데이터				2번 오버샘플링				2번 언더샘플링			
	N	Y	Sum		N	Y	Sum		N	Y	Sum
N	1,615	63	1,678	N	458	72	530	N	1,347	394	1,741
Y	360	184	544	Y	131	369	500	Y	460	1214	1,674

Confusion Matrix는 검증 데이터에 대한 것으로, 총 인스턴스가 각각의 데이터 셋의 33.3%에 해당한다. 원본 데이터 셋은 2,222개, 언더샘플링 데이터 셋은 1,030개, 오버샘플링 데이터 셋은 3,415개가 사용되었다.

원본 데이터 셋의 경우 N에 대한 학습이 많이 일어나기 때문에, N의 TP rate가 극도로 높지만(0.962), Y의 TP rate(TN rate)는 상당히 낮다(0.338). N의 FP rate는 1-Y의 TP rate를 뺀 것으로 0.662이다. Y의 FP rate는 반대로 1-N의 TP rate를 뺀 것으로 0.038이다. 언더샘플링 셋과 오버 샘플링 셋의 경우는, Y에 대한 예측과 TP rate가 많이 상승한 것을 볼 수 있다. 즉, 실제의 값과 예측한 값이 일치하는 결과가 Y와 N 클래스 모두에게 나타난다는 것을 알 수 있다. Y의 TP rate가 각각 0.738, 0.725로 나타났다. 이는 데이터 균형화의 목적인, 목표변수의 데이터 불균형으로 인한 메이저 집단에 편향된 예측 모형 구축 문제가 해결되었음을 알 수 있다.

정밀도(precision)는 실제와 예측이 일치하는 값을 해당 클래스에 대한 예측 전체의 값으로 나누어 구할 수 있다. 즉, a/(a+c)와 d/(b+d)로 계산된다. 해당 클래스에 대한 전체 예측 중에서 실제와 일치하는 예측의 수가 차지하는 정도를 의미한다. 예를 들면, N의 TP 값은 1,615이다. 여기에서 예측을 N으로 하였으나, 실제 값이 Y인 경우에 해당하는 값이 360이다. 이를 합산하여, TP값의 비중을 구하면 정밀도가 산출된다. 비교 대상으로 선정된 세 가지의 데이터 셋 모두 0.7 이상의 양호한 정밀도를 보이고 있다.

4.3 예측률 향상과 과적합 해결 방안

본 연구의 결과로 미루어 볼 때, 과적합이 일어나지 않은 상태로의 올바른 분류분석을 시행하기 위해서는 목표변수에 대한 데이터 균형화가 필요한 것으로 사료된다. 즉, 목표변수의 하위 집단 간에 샘플 수 차이로 인한 데이터 불균형이 있는 경우, 메이저 집단에 대한 편향된 예측모형이 개발될 가능성이 크다는 점을 Confusion Matrix 등을 통해 검증하였다. 이와 더불어 데이터 균형화를 적용하였음에도 불구하고 편향된 예측모형와 비교할 때, 예측률에 있어서 성능적으로 뒤쳐지지 않음을 보이고 있다.

<표 6>은 최종 비교 대상으로 선정된 데이터 셋과 역방향 제거를 통해 탐색된 최고 예측률,

<표 6> 최종 비교 대상 종합 및 정리

최종 비교 대상 간의 예측률, 나무의 크기, 규칙 수, 성능지표										
구분	의사결정나무 모형						성능지표			
Data Set	m. NO	예측률 (%)	투입 피처수 (개)	나무 크기	규칙	특이사항	Y TP rate	N TP rate	Y 정밀도	N 정밀도
원본 Data	50	81.773	21	15	8	Y 과적합	0.338	0.962	0.465	0.818
언더 no.2	50	80.291	39	43	33	중요도 순위 역전 발생	0.738	0.864	0.837	0.778
오버 no.2	150	74.993	46	50	31	원본데이터 규칙상세화	0.725	0.774	0.755	0.745

최고 예측률을 기록했을 때의 투입 피처수, 의사결정나무 모형의 나무크기, 규칙의 수(잎사귀의 수), 특이사항, 성능평가에 대한 지표 등을 정리한 것이다. 원본 데이터의 경우, 성능 지표상에 나타난 대로 Y에 대한 정분류율이 N에 대한 정분류율에 비하여 크게 떨어지는 것을 볼 수 있다. 즉, 목표변수의 하위집단에 대한 데이터 불균형 현상이 존재함을 알 수 있다. 또한 최고 예측률 탐색을 위해 조정한 m.NO 값이 나무의 크기와 규칙의 간략화를 일으켰음을 알 수 있다.

언더샘플링의 경우, 데이터 균형화가 이루어졌음을 알 수 있고, 예측률 또한 데이터 불균형 상태인 원본 데이터에 비하여 크게 떨어지지 않았다. 그러나 목표변수의 하위집단 중 마이너 집단의 수에 메이저 집단의 수를 맞추는 식으로 샘플링을 하였기 때문에 투입된 독립변수(피처)의 중요도 순위 역전 현상이 발생했다. 이러한 점을 미루어 볼 때, 언더샘플링의 한계점인 원본 데이터 상에 존재하는 유용한 정보들을 모두 활용하지 못하는 것이 반영된 것으로 보인다.

오버샘플링의 경우, 나무의 크기와 잎사귀의 수를 해석 가능한 선에 맞추기 위하여 m.NO값을 크게 설정하였기 때문에 예측률이 비교적 떨어지지만, 투입된 피처에 대한 중요도 순위 역전이 발생하지 않은 선에서 간략히 나타난 원본 데이터 규칙의 상세화를 일으켰다. 즉, 일련의 실험과정에서 지속적으로 나타난 중요한 변수들의 중요도 순위가 역전되지 않았고, 결과적으로 언더샘플링이 기존 데이터의 유용한 정보들을 모두 활용하지 못한다는 단점을 보완하는 데이터 균형화 기법임을 알 수 있다.

예측모형의 성능평가지표가 되는 정・오분류율에 있어서도 최종 비교대상 간의 차이점을 살펴볼 수 있다. 편향된 예측모형이 개발된 원본 데이터의 경우 목표변수의 마이너 집단인 Y에 대한 정분류율이 오버샘플링 및 언더샘플링에 비하여 상당히 낮다. 그러므로 Y의 정밀도도 상당히 낮은 수치를 기록하였다. 이와 달리 데이터 균형화가 적용된 언더샘플링과 오버샘플링의 경우 목표변수 내의 메이저·마이너 집단이 존재하지 않기 때문에, Y에 대한 정분류율이 향상되었음을 알 수 있다. 즉, 목표변수에 대한 데이터 균형화로 인하여 자살생각 경험이 있음에 대한 정분류율이 높아지고, 오분류율이 낮아졌음을 알 수 있다.

본 연구에서는 최고 예측률 탐색 및 효율적 규칙 해석을 위하여, m.NO 값의 조정에 대해 지속적으로 언급하였다. m.NO 값은 리프노드당 최소 인스턴스 수에 대한 것으로 분지 및 가지치기의 기준이 된다. 이를 조정하여 나무의 크기와 잎사귀의 수, 예측률, 정・오분류율 등에 대한 차이가 발생함을 살펴보았다. 단순히 m.NO 값을 기본값인 2로 설정하여 역방향 제거 작업을 수행했을 때 관측되는 최고 예측률이 최적이 아닐 수도 있다. 나무의 크기와 잎사귀의 수가 지나치게 큰 값이라면 규칙 해석에 있어 어려움이 많다. 효과적이고 능률적인 규칙 해석을 위하여 신뢰지수의 조절이나, 본 연구에서 사용한 리프 노드 당 포함되어야 할 최소 인스턴스의 수를 조절하는 방식을 시도할 필요가 있다. 이때 발생하는 예측률의 소실을 최소한으로 줄이기 위한 방안으로, 본 연구에서는 초기 7 가지 수준으로 m.NO 값에 대한 예측률을 살펴보고 그 중 최선 및 차선을 지정하여 모두 역

방향 제거를 수행하였다. m.NO 값의 조정은 분지 및 가지치기의 기준으로 작용할 뿐 아니라, 예측률 소실의 최소화를 위한 도구로 활용 가능한 것으로 보인다.

4.4 의사결정규칙 도출 및 특성 분석

의사결정나무는 목표변수와 독립변수 사이에 과정이 화이트박스 형태로 나타나기 때문에 규칙이나 준거점을 파악하는 것이 용이하다. 이를 통한 규칙과 시사점, 결론 등을 도출해내는 데에 있어서도 용이하고 해석에 큰 이점을 준다. 본 실험에서 다루었던 세 가지 데이터 셋을 바탕으로 형성된 의사결정나무를 바탕으로 규칙을 해석하고 설명하는 것이 중요하며, 이는 이론적, 실무적 도움을 줄 수 있을 것으로 파악된다. 목표변수인 자살생각의 유무에 있어서 자살생각이 있는 경우, 그들이 갖고 있는 특성들이나 상태를 파악할 수 있는 독립변수들로 이루어진 예측모형이기 때문에 이를 해석함으로 노인 자살 예방에 반드시 기여할 수 있을 것이라 판단된다.

(1) 원본 데이터의 의사결정 규칙 도출

먼저 원본 데이터를 이용하여 의사결정나무 모형과 의사결정 규칙을 도출하면 <표 7>과 같다. 표에서 각 규칙별 예측결과와 예측률을 보여주고 있다. 원본 데이터 셋에 대한 의사결정나무 모형의 규칙 분석 및 해석은 다음과 같다.

의사결정 나무의 최상단에 '자살계획 노드'를 중심으로 분지가 이루어진다. 1번 규칙은 자살계획에 관한 경험이 있다면, 해당 노드의 인스턴스 중 98.51%가 자살생각을 경험한 것으로 나타났다. 또한 '우울 노드'를 중심으로 크게 분지가 일어났는데, 이것은 자살계획에 관한 경험이 없고 우울하지 않은 경우, 해당 노드의 인스턴스 중 90.75%가 자살생각을 경험하지

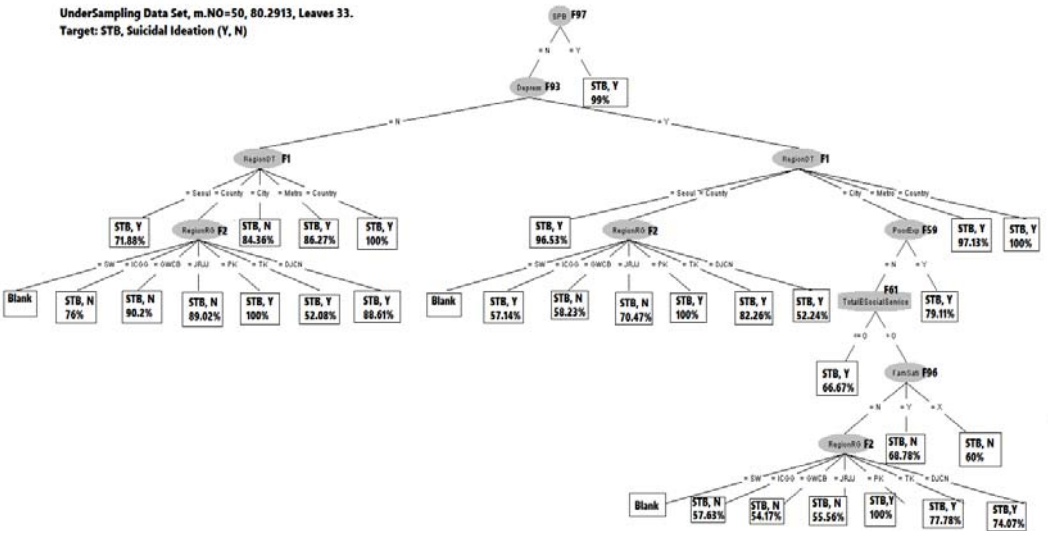
<표 7> 원본 데이터의 의사결정 규칙 및 예측률

규칙	내용	예측결과	예측률(%)	예측성공
1	자살계획 Y	자살생각 Y	98.51	199
2	자살계획 N & 우울 N	자살생각 N	90.75	3366
3	자살계획 N & 우울 Y & 경제적 어려움 경험 Y & 가정생활에 대한 스트레스 Y	자살생각 Y	68.70	259
4	자살계획 N & 우울 Y & 경제적 어려움 경험 Y & 가정생활에 대한 스트레스 N	자살생각 N	54.59	101
5	자살계획 N & 우울 Y & 경제적 어려움 경험 N & 사회적 급여 N	자살생각 N	71.60	1306
6	자살계획 N & 우울 Y & 경제적 어려움 경험 N & 사회적 급여 Y & 가정생활에 대한 스트레스 N	자살생각 N	61.68	66
7	자살계획 N & 우울 Y & 경제적 어려움 경험 N & 사회적 급여 Y & 가정생활에 대한 스트레스 Y & 연령 <=81세	자살생각 Y	63.64	49
8	자살계획 N & 우울 Y & 경제적 어려움 경험 N & 사회적 급여 Y & 가정생활에 대한 스트레스 Y & 연령 >81세	자살생각 N	61.82	34

않은 것으로 나타났다. 3번 규칙에서 8번 규칙까지는 우울한 집단 Y에 대한 규칙들이 도출되었다. 3번 규칙은 자살계획에 관한 경험이 없고, 우울하며, 경제적 어려움 경험이 있고, 가정생활에 대한 스트레스가 있을 때에 68.7%가 자살생각을 경험한 것으로 나타났다. 4번 규칙은 다른 것은 동일하나 3번 규칙와 달리 가정생활에 대한 스트레스가 없는 경우이며, 해당 노드의 인스턴스 중 54.59%가 자살생각을 경험하지 않은 것으로 나타났다.

5번-8번 규칙에서는 경제적 어려움으로 인한 곤궁 경험이 없는 집단 N에 대한 규칙들이며, 5번 규칙에서 사회적 급여를 지급받지 못하는 경우 자살생각을 경험하지 않은 것으로 나타났다. 6번 규칙에서는 사회적 급여를 지급받는 경우일지라도, 가정생활에 대한 스트레스가 없다면 자살생각을 경험하지 않았다. 7번과 8번 규칙에서는 가정생활에 대한 스트레스가 있을지라도, 81세를 기준으로 81세 이하일 경우엔 자살생각을 경험했고, 81세 초과인 경우엔 자살생각을 경험하지 않았다.

원본 데이터에서 나타나는 특이사항은 다음과 같다. 경제적 어려움 경험(F59)는 경제적인 어려움으로 인해 국민건강 보험료를 미수납했거나, 식사 끼니 등을 걸렀던 경험에 관한 것이다. 사회적 급여(F62)는 생계 급여나, 주거 급여, 교육 급여, 국민기초생활 급여 등으로 생계유지비와 관련되어 있다. 생계유지비를 지급받지 않는 집단(규칙 5번)의 경우엔 자살생각을 경험하지 않은 것으로, 생계유지비를 지급받은 집단 중에서도 연령이 81세 이하라면(규칙 7번)의 경우엔 자살생각을 경험하는 것으로 나타났다. 생계유지비를 지급받는다는 의미는 가구의 경제적 가난을 의미하는 것으로 해석된다. 하지만, 원본 데이터는 데이터 균형화를 거치지 않았기 때문에, 자살생각의 하위 집단인 N에 편향된 예측모형임을 간과해서는 안 된다.



<그림 3> 언더샘플링의 의사결정나무 모형과 예측성공률

<표 8> 언더샘플링 데이터의 의사결정 규칙 및 예측률

규칙	내용	예측결과	예측률(%)	예측성공
1	자살계획 Y	자살생각 Y	99.00	199
2	자살계획 N & 우울 N & 도시규모-서울특별시	자살생각 Y	71.88	46
3	자살계획 N & 우울 N & 도시규모-도시	자살생각 N	84.36	588
13	자살계획 N & 우울 Y & 도시규모-서울특별시	자살생각 Y	96.53	139
14	자살계획 N & 우울 Y & 도시규모-광역시	자살생각 Y	97.13	237
25	자살계획N&우울Y&도시규모-도시&경제적어려움경험N&노인가구 의사회복지서비스경험횟수>0회 & 가족에 대한 만족도 비해당	자살생각 N	60.00	3
26	자살계획N&우울Y&도시규모-도시&경제적어려움경험N&노인가구 의사회복지서비스경험횟수>0회 & 가족에 대한 만족도 Y	자살생각 N	68.78	141

(2) 언더 샘플링의 의사결정 규칙 도출

다음으로 언더샘플링의 경우 의사결정나무 모형과 각 노드별 예측 성공률을 도식화 하면 <그림 3>과 같으며, 주요 규칙을 제시하면 <표 8>과 같다. 언더샘플링 데이터의 경우에도, 위와 마찬가지로 최상단에 위치한 '자살계획 노드'에 의해 크게 분지된다. 언더샘플링의 그래프 특징은 앞서 언급한 변수 중요도를 매겼을 때, 상위 10개에 지역 구분과 관련된 변수가 포함된다는 것이다. 5개 권역 구분(F1)은 도시 규모에 따른 차이고, 7개 권역 구분(F2)은 지역 분포에 따른 차이이다. 지역 구분에 관련된 규칙들이 도출되었기 때문에 규칙의 수(잎사귀의 수)가 많다.

규칙 1은 자살계획에 대한 경험이 있는 경우, 99%가 자살생각을 경험한 것으로 나타났다. 우울 노드를 기준으로, 우울하지 않은 집단 N과 관련된 규칙 2번-12번 규칙이 도출되었고 우울한 집단 Y와 관련된 규칙 13번-33번 규칙이 도출되었다. 먼저 우울하지 않은 집단의 경우, 도시 규모에 따라 서울특별시, 광역시, 도시, 도농복합군과 직접적인 규칙이 도출되었다. 자살계획 경험이 없고, 우울하지 않기 때문에 도시 거주자들의 84.36%가 자살생각을 경험하지 않은 것으로 나타났다. 그러나 이외에는 모두 '자살생각을 경험했다'에 대한 규칙이 도출되었다.

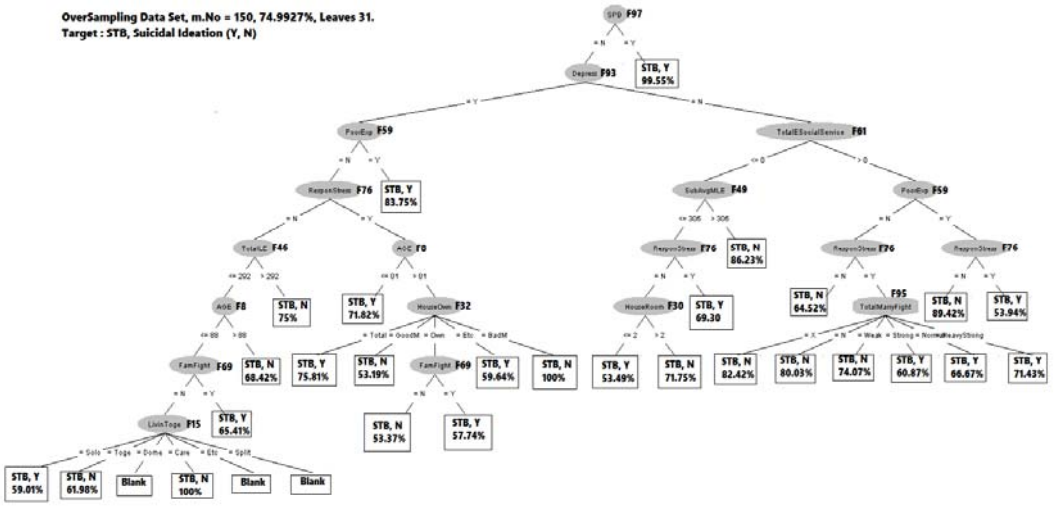
규칙 25를 보면 자살계획 경험이 없고, 우울하며, 도시 규모에 거주하는 노인 중에서 경제적 어려움을 경험한 경우에, 해당 노드의 79.11%가 자살생각 경험이 있음으로 나타났다. 규칙 26에서는 경제적 어려움을 경험하지 않고, 사회복지서비스를 한 번도 경험하지 않은 노인 가구의 경우에, 해당 노드의 66.67%가 자살생각을 경험한 것으로 나타났다. 특히, 가족이 없는 경우인 X집단의 경우, 해당 노드의 60%가 자살생각을 하지 않는 것으로 나타났다. 가족관계에 만족하는 경우인 Y 집단의 경우, 해당 노드의 68.78%가 자살생각을 하지 않는 것으로 나타났다.

언더샘플링에서 나타나는 특이사항은 다음과 같다. 사전 실험과 원본 데이터 및 오버샘플링 데이터에서 중요도 순위가 낮은 순위를 기록한 지역적 구분 변수에 의한 규칙이 많이 도출되었다. 언더샘플링 기법의 단점으로 전체 데이터에 포함된 유용한 정보를 모두 활용할 수 없음이 인식된다. 하지만, 지역적 구분에 대한 설명과 도시규모 구분에 대한 설명을 반드시 짚고 넘어가야 하는 것으로 판단된다.

(3) 오버 샘플링의 의사결정 규칙 도출

<그림 4>는 오버샘플링 데이터의 의사결정나무를 근거로 규칙을 도출한 결과이며, 주요 의사결정 규칙을 제시하면 <표 9>와 같다. 오버 샘플링의 경우에도 언더 샘플링과 마찬가지로, 자살계획 경험 유무에 의해 분지되며 자살계획 경험이 있는 경우에 99.55%가 자살생각 경험이 있는 것으로 나타났다. 그래프 상에 언더샘플링 데이터에서 분지 기준으로 자주 나타난 지역 구분 독립변수들은 전혀 나타나지 않았다. '우울 노드'를 기준으로 우울 집단과 비우울 집단에 대한 규칙이 도출되었다.

규칙 2-14번은 우울하지 않은 집단에 대한 규칙이며, 15-31번은 우울한 집단에 대한 규칙이다. 2번 규칙은 자살계획 경험이 없고, 우울하지 않으며, 노인복지서비스 경험이 없고, 주관적 월평균최적생활비가 305만원 이상일 때, 해당 노드의 인스턴스 중 86.23%가 자살생각을 경험하지 않은 것으로 나타났다. 3번 규칙은 주관적 월평균최적생활비가 305만원 이하일 때, 가정생활에 대한 스트레스가 있다면 자살생각을 경험하는 것으로 규칙이 도출되었다(69.3%). 4-5번 규칙은 가정생활에 대한 스트레스가 없다면, 주거지의 방 개수(F30)가 2개 이하라면 자살생각을 경험하는 것으로, 3개 이상이라면 자살생각을 경험하지 않는 것으로 각각 규칙이 도출되었다. 6번 규칙은 자살계획 경험이 없고, 우울하지 않으며, 노인복지서비스와 경제적 어려움 경험이 있고, 가정생활에 대한



<그림 4> 오버샘플링의 의사결정나무 모형과 예측성공률

<표 9> 오버샘플링 데이터의 의사결정 규칙 및 예측률

규칙	내용	예측결과	예측률(%)	예측성공
1	자살계획 Y	자살생각 Y	99.55	665
2	자살계획 N & 우울 N & 노인가구의 사회복지서비스 경험 횟수 <= 0회 & 주관적 평균최적생활비(月) > 305 만원	자살생각 N	86.23	144
3	자살계획N&우울N&노인가구의사회복지서비스경험횟수<=0회&주관적평균최적생활비(月)<=305만원 & 가정생활에 대한 스트레스 Y	자살생각 Y	69.30	316
4	자살계획N&우울N&노인가구의사회복지서비스경험횟수<=0회&주관적평균최적생활비(月)<=305만원 & 가정생활에 대한 스트레스 N & 방 수 <= 2개	자살생각 Y	53.49	92
5	자살계획N&우울N&노인가구의사회복지서비스경험횟수<=0회&주관적평균최적생활비(月)<=305만원 & 가정생활에 대한 스트레스 N & 방 수 > 2개	자살생각 N	71.75	193
6	자살계획 N & 우울 N & 노인가구의 사회복지서비스 경험 횟수 > 0회 & 경제적 어려움 경험 Y & 가정생활에 대한 스트레스 N	자살생각 N	64.52	100
7	자살계획 N & 우울 N & 노인가구의 사회복지서비스 경험 횟수 > 0회 & 경제적 어려움 경험 Y & 가정생활에 대한 스트레스 Y	자살생각 Y	53.94	89
8	자살계획 N & 우울 N & 노인가구의 사회복지서비스 경험 횟수 > 0회 & 경제적 어려움 경험 N & 가정생활에 대한 스트레스 N	자살생각 N	89.42	1750
15	자살계획 N & 우울 Y & 경제적 어려움 경험 Y	자살생각 Y	83.75	1129
16	자살계획 N & 우울 Y & 경제적 어려움 경험 N & 가정 생활에 대한 스트레스 N & 총 생활비(月) > 292 만원	자살생각 N	75.00	129

스트레스가 없을 때, 해당 노드의 인스턴스 중 64.52%가 자살생각을 경험하지 않은 것으로 나타났다. 7번 규칙은 가정생활에 대한 스트레스가 있을 때, 해당 노드의 인스턴스 중 53.94%가 자살생각을 경험한 것으로 나타났다. 8번 규칙은 자살계획 경험이 없고, 우울하지 않으며, 노인복지서비스와 경제적 어려움 경험이 없고, 가정생활에 대한 스트레스가 없을 때, 해당 노드의 인스턴스 중 89.42%가 자살생각을 경험하지 않은 것으로 나타났다.

오버샘플링에서 나타나는 특이사항은 다음과 같다. 앞서 언급했던 우울과 자살생각 간의 관계에 대한 선행연구들에 나타난 변수들을 바탕으로 추출했기 때문에, 풍부한 규칙 도출 및 해석이 가능했다. 이는 오버샘플링 데이터 셋의 인스턴스 수가 약 1만개인데, 많은 양의 인스턴스를 분류하는 과정에서 m.NO의 값이 가장 큼에도 불구하고 그래프 상에 중요 노드들이 분지의 준거점으로 나타났다. 우울을 중심으로 대분지가 일어났고, 우울한 집단의 경우 자살생각 경험이 있는 것으로 나타난 규칙의 수가 7개, 비우울 집단의 경우 6개가 도출되었다. 우울이 자살생각에 미치는 영향이 큼은 여러 연구를 통해 입증되었으나, 우울한 경우라도 어떠한 작용에 의해 자살생각을 경험하는지와 우울하지 않은 경우라도 어떠한 작용에 의해 자살생각을

경험하는지를 자세히 살펴볼 수 있었다.

경제적 요인과 관련된 월평균 총 생활비, 주관적 최저-적정 생활비, 경제적 어려움 경험을 기준으로 분지가 일어나고 규칙이 도출되었음을 알 수 있다. 특히, 월평균 총 생활비가 292만원 초과인 경우, 자살생각 경험이 없는 것으로 나타났으며 생활비가 많을수록 자살생각이 억제된다고 볼 수 있다. 주관적 생활비에 관련해서 해당 값이 클수록 저소득층이 아닌 일반가구에 속할 것으로 유추할 수 있는데, 월평균 최적의 생활비가 305만원 이상이면 자살생각 경험이 없는 것으로 나타났다. 최적이라는 의미는 최저생활비와 적정생활비로부터 가공된 평균값이기 때문에, 의미 있는 변수를 발견한 것으로 보인다.

가족 관계 요인과 관련된 가정생활에 대한 스트레스, 가족간의 다툼, 부부싸움 등도 분지의 기준이 됨을 볼 수 있다. 특히나 우울하지 않은 집단의 경우, 부부싸움의 정도에서 자살생각 경험에 관련된 규칙이 3개가 도출되었다. 일반적인 언쟁 정도의 다툼이라면 자살생각을 하지 않지만, 신체적인 폭력이나 심각한 언어폭력이 자행된 경우에는 자살생각을 경험한 것으로 나타났다.

본 절에서 각 데이터 셋별 의사결정나무 분석 결과로 나타난 규칙들 종합적으로 정리하면 다음과 같다. 3개의 그래프 상에 모두 나타난 변수는 자살계획(F97), 우울(F93), 경제적 어려움 경험(F59)이 있으며, 이외에는 약간의 차이가 존재한다. 전체적으로 볼 때, 원본 데이터와 오버샘플링 데이터의 의사결정나무가 유사한 형태를 띄며 언더샘플링 데이터의 경우 지역구분 변수가 그래프 상에 나타났다. 변수의 중요도 순위에서 높은 위치를 차지한 자아존중감, 전반적 생활 만족도 등의 변수는 분지의 기준으로 등장하지 않았다. 오버샘플링의 그래프를 통해, 비교적 간결했던 원본 데이터의 그래프를 면밀히 살펴볼 수 있었다.

Ⅴ. 결론 및 시사점

본 연구는 노인의 자살생각을 예측하는 데에 있어 선행 연구를 바탕으로 투입가능한 모든 변수들을 고려하여 연구를 수행하였다. 여러 단계를 거쳐 최종적으로 목표변수 포함 총 99개의 변수를 활용하였다. 역방향제거 방식과 데이터 균형화 작업을 통한 샘플링 기법을 도입하여 분류분석을 시행하였다. 방의 개수나 주거점유 형태, 주관적 생활비 등과 같은 변수에 대한 선행연구가 부족한 점이 단점으로 지적될 수 있다. 그러나 한국복지패널에서 제공하는 패널데이터의 자료들을 충분히 활용함으로 실제 노인들의 세계를 보다 구체적으로 반영했다는 점에서 기여하는 바가 있다고 여겨진다. 또한 노인의 자살생각에 대한 기존 선행 연구들은 대부분 통계적 방법론으로 시도되었으며, 자살생각의 가장 주요한 요인으로 꼽히는 우울에 대한 의사결정모형 분석을 시도한 연구들이 많았다. 그러나 본 연구에서는 우울 외에도 보다 더 다양한 변수들을 이용하였다. 앞서 언급한 연구들과 본 연구의 차이점과 본 연구의 시사점은 다음과 같다.

노인의 자살생각을 예측하기 위하여, 예측모형의 개발방법론으로 의사결정나무 분석을 사용했다. 데이터 마이닝의 분류분석 기법인 의사

결정나무 모형을 통해 목표변수인 자살생각에 대한 예측모형을 개발함으로 여러 규칙들을 도출하였다. 대부분의 연구에서 우울이 자살생각에 미치는 영향이 크기 때문에 우울한 노인에 대한 시사점이나 제언 등은 상당히 많이 도출되어 있다. 그러나 본 연구에서는 우울한 집단의 자살생각 경험에 대한 규칙뿐만 아니라, 우울하지 않은 집단의 자살생각 경험에 대한 규칙을 도출했기 때문에 노인 전체에 대한 실무적 시사점이나 유의점을 얻어 낼 수 있다. 특히, 우울하지 않은 노인이 자살생각을 경험한 규칙들(오버샘플링 규칙3, 4, 7, 12, 13, 14)을 살펴보면 노인복지, 경제적 여건, 가족 관계 등과 관련이 있다. 이에 해당하는 규칙과 같은 노드를 공유하는 규칙들의 차이가 무엇인지(분지의 기준점)를 살펴보고, 자살생각을 경험케 하는 것이 구체적으로 어떤 점에서 비롯되는지와 이를 해결하기 위한 맞춤형 전략과 방안을 모색해야 한다.

정부의 정책적 관점에서 본다면 노인들의 자살예방을 위해서는 노인들의 우울을 예방하기 위한 다양한 프로그램의 개발이 필요하다. 현재 대부분의 지자체에서 노인들을 대상으로 노인대학을 운영하고 있다. 노인들의 우울증을 해소하고 건강증진을 위해서는 활동중심의 놀이프로그램의 개발을 통하여 건장증진과 우울증 해소를 동시에 추구해야 한다. 또한 노인들의 경제력을 향상시키기 위한 다양한 노인 일자리 정책을 개발해야 한다. 그리고 노인들을 위한 전문상담인력을 대폭 늘이기 위한 자격증 제도의 도입이 필요하다.

끝으로 본 연구에서는 예측모형 개발에 있어서도 데이터 균형화 적용의 유무를 통해, 직접적으로 데이터 불균형 현상에 따른 과적합 문제를 규명하였다. 또한 단순히 높은 예측률을 지향하기보다, 예측률 손실의 최소화 및 예측모형의 최적화 등을 도모하였다. 또한 한국보건사회연구원에서 제공하는 한국복지패널 데이터를 사용하여 분석하였고, 의사결정나무의 그래프를 바탕으로 규칙들을 도출하였기 때문에 자료에 대한 신빙성과 대표성, 규칙의 정확성이 담보되었다고 생각한다.

마지막으로 본 연구의 제한점이나 향후 연구방향에 대해 언급하고자 한다. 본 연구의 제한점은 크게 세 가지를 들 수 있다. 우선적으로 패널데이터의 특성을 완벽하게 살리지 못하였다. 패널데이터의 특성은 횡단면 분석과 종단면 분석의 한계를 모두 넘기 위한 것으로, 연구 대상의 연속성이나 그 자료를 바탕으로 진행하는 연구의 풍성함을 더한다. 본 연구에서는 목표변수를 중심으로 인스턴스의 중복제거 및 시간적 통제 등과 같이 연구 대상자 선정과정에 있어 기준점이 되는 부분들이 결과적으로 패널데이터의 특성을 100% 활용하지 못하게 되었다. 이는 향후의 연구에서 보완되어야 할 부분으로 생각된다.

두 번째로, 변수를 조작하는 과정과 추출해내는 과정에 있어서 분야의 이질적인 차이로 인한 한계점을 들 수 있다. 예를 들면, 자산은 자본+부채의 개념을 가지고 있는데 이를 순자산 등으로 조작화 하지 않고 각각의 변수로 한꺼번에 투입하였거나 응답자들의 큰 편차로 인한 연속척도 값들을 일부 명목척도로 변경하여 변수를 조작화한 점 등이 있다. 이러한 부분에 대해서는 다양한 분야의 연구자들의 참여나 연구를 통해 보완할 수 있을 것이다.

끝으로 향후 연구방향에 대한 언급을 하고자 한다. 본 연구는 여러 기준에 의한 실험 방법의 다양화를 추구하였으며, 매우 많은 인스턴스와 피처의 투입으로 예측모형을 개발하였기 때문에 이와 비슷한 연구를 수행함으로 해당 연구 방법론의 실효성에 대한 검증과 보완이 필요하다. 앞서 언급한 한계점과 미비점들을 보완해나감으로 연구의 질을 높이고, 보다 최적화된 연구방법론을 정립하고자 한다.

참고문헌

강현철 외 5명, 빅데이터 분석을 위한 데이터마이닝 방법론, 자유아카데미, 2014.

권오균, 허준수, “저소득 독거노인의 자살생각 인과모형에 관한 연구-자아존중감, 우울감, 절망감의 매개효과를 중심으로,” 정신보건과 사회사업, 제41권, 제4호, 2013, pp.65-93.

권중돈, 김유진, 엄태영, “노인돌봄서비스 이용 독거노인의 자살생각에 영향을 미치는 요인에 관한 연구-자살시도경험과 음주행위와의 관계를 중심으로”, 노인복지연구, 제51권, 2011, pp.297-320.

김경민, 장하영, 장병탁, “불균형 데이터 처리를 위한 과표본화 기반 앙상블 학습 기법”, 한국정보과학회지, 제20권, 제10호, 2014, pp.549-554.

김성은, 김선아, “의사결정나무 분석기법을 이용한 농촌거주 노인의 우울예측모형 구축”, Journal of Korean academy of nursing, 제43권. 제3호, 2013, pp.442-451.

김원종, 최연식, 유동희, “데이터 마이닝을 활용한 한국 프로야구 구단의 승패예측과 승률 향상을 위한 전략 도출 연구”, 한국스포츠산업경영학회지, 제23권, 제3호, 2018, pp.88-104.

김종필, 현미열, “치매노인의 우울과 자살의도”, Journal of Korean academy of nursing, 제43권, 제2호, 2013, pp.296-303.

김지훈, 김경호, “자살생각, 자살계획 및 자살시도와 관련된 유발요인의 영향력 분석: 6차년도 한국복지패널 참여자를 대상으로”, 한국콘텐츠학회논문지, 제18권, 제2호, 2018, pp.344-360.

문동규, “노인의 자살생각과 관련된 유발변인의 메타회귀분석”, 노인복지연구, 제55권, 2012, pp.133-157.

박명화, 최소라, 신아미, 구철회, “의사결정나무 분석법을 활용한 우울 노인의 특성 분석”, Journal of Korean academy of nursing, 제43권, 제1호, 2013, pp.1-10.

박민정, “노인의 성별에 따른 자살생각과 영향요인-2010년도 한국의료패널자료를 이용하여”, Journal of the Korean Data Analysis Society, 제12권, 제2호, 2015, pp.1087-1099.

장영재, “빅데이터, 비즈니스 애널리틱스, IoT: 경영의 새로운 도전과 기회,” 정보시스템연구, 제24권, 제4호, 2015, pp.139-152.

안민욱, 김태운, 유동희, “데이터 마이닝 기법을 활용한 근로자의 고용유지 강화 방안

개발", 한국콘텐츠학회논문지, 제18권, 제5호, 2018, pp.265-279.

오욱찬, 박정민, 구서정, "가계부채가 정신건강에 미치는 영향-우울감과 자살생각을 중심으로", 한국사회복지학, 제69권, 제2호, 2017, pp.171-190.

오윤정, 김향동, "노인의 자살생각 예측요인", 융합정보논문지, 제8권, 제2호, 2018, pp.1-9.

원하림, 김무전, 안현철, "온라인 무료 샘플 판촉의 효과적 활용을 위한 기계학습 기반 고객분류예측 모형," 정보시스템연구, 제27권, 제3호, 2018, pp.63-80.

유동희, 최근호, 서용무, "산재근로자들의 고용안정과 건강한 삶을 위한 데이터마이닝 기반의 규칙 도출 연구", 한국직업재활학회, 25(3), 2015, pp.5-24.

이금룡, 조은혜, "독거노인의 자살생각에 영향을 미치는 주요 변인에 관한 연구: 사회적 지지의 직접 및 간접효과를 중심으로," 보건사회연구, 제33권, 제1호, 2013, pp.162-189.

이윤정, "노인가구의 정보화 상태가 우울과 자살생각에 미치는 영향", 한국가정관리학회 제52차 추계학술대회, 2012, pp.215-229.

이인정, "노인의 우울과 자살생각의 관계에 대한 위기사건, 사회적 지지의 조절효과", 보건사회연구, 제31권, 제4호, 2011, pp.34-62.

이혜경, "배우자 사별을 경험한 독거노인의 애도수준과 자살생각 간의 관계-우울의 매개효과를 중심으로," 정신보건과 사회사업, 제44권, 제1호, 2016, pp.24-47.

임성옥, 김경희, "전기・후기노인의 자살생각에 대한 영향요인," 사회복지정책, 제45권, 제3호, 2018, pp.5-29.

중앙자살예방센터, 2018자살예방백서, 2018.

한국복지패널, 1-12차 결합데이터 코드북(가구데이터), 2018.

한국복지패널, 1-12차 결합데이터 코드북(머지데이터: 가구+가구원+부가), 2018.

한국복지패널, 2016년 한국복지패널 기초분석 보고서, 2016.

한국복지패널, 12차년도 조사자료 유저가이드, 2018.

Bonnewyn, A., Shah, A., and Demyttenaere, K. "Suicidality and Suicide in Older People," *Reviews in Clinical Gerontology*, Vol. 19, 2009, pp. 271-294.

Burez, J. and Van den Poel, D., "Handling Class Imbalance in Customer Churn Prediction," *Expert Systems with Applications*, Vol. 36, No. 3, 2009, April, pp. 4626-4636.

Longade, R., Dongre, S., and Malik, L., "Class Imbalance Problem in Data Mining: Review," *International Journal of Computer Science and Network*, Vol. 2, Iss. 1, 2013, February, pp. 1-6.

Mann, J., Apter, A., Bertolote, J., Beautrais, A., Currier, D., and Haas, A. "Suicide Prevention Strategies: A Systematic Review." *JAMA*, Vol. 294, No. 16, 2005, pp. 2064–2074.

Song, Y. and Lu, Y., “Decision Tee Methods: Applications for Classification and Prediction,” *Shanghai Arch Psychiatry*, Vol. 27, No. 2, 2015, pp. 130-135.

김 덕 현 (Kim, Deok Hyun)

경상대학교 경영정보학과에서 경영학사와 석사를 취득하였다. 현재 경상대학교 경영정보학과 박사과정에 재학중이다. 주요 관심분야는 빅데이터 분석, 노인복지 등이다.

유 동 희 (Yoo, Dong Hee)

고려대학교에서 경영학사와 경영학 박사학위를 취득하였다. 현재 경상대학교 경영정보학과에서 부교수로 재직하고 있으며, 주요 관심분야 는 데이터마이닝, 빅데이터 분석, 지능형시스템 등 이다.

정 대 율 (Jeong, Dae Yul)

부산대학교 경영학과에서 경영학사, 석사, 박사학위를 취득하였다. 현재 경상대학교 경영정보학과 교수로 재직하고 있으며, 주요 관심분야는 시스템분석 및 설계, 데이터마이닝, 의사결정지원시스템 등이다.

<Abstract>

A Development of Suicidal Ideation Prediction Model and Decision Rules for the Elderly: Decision Tree Approach

Kim, Deok Hyun · Yoo, Dong Hee · Jeong, Dae Yul

Purpose

The purpose of this study is to develop a prediction model and decision rules for the elderly's suicidal ideation based on the Korean Welfare Panel survey data. By utilizing this data, we obtained many decision rules to predict the elderly's suicide ideation.

Design/methodology/approach

This study used classification analysis to derive decision rules to predict on the basis of decision tree technique. Weka 3.8 is used as the data mining tool in this study. The decision tree algorithm uses J48, also known as C4.5. In addition, 66.6% of the total data was divided into learning data and verification data. We considered all possible variables based on previous studies in predicting suicidal ideation of the elderly. Finally, 99 variables including the target variable were used. Classification analysis was performed by introducing sampling technique through backward elimination and data balancing.

Findings

As a result, there were significant differences between the data sets. The selected data sets have different, various decision tree and several rules. Based on the decision tree method, we derived the rules for suicide prevention. The decision tree derives not only the rules for the suicidal ideation of the depressed group, but also the rules for the suicidal ideation of the non-depressed group. In addition, in developing the predictive model, the problem of over-fitting due to the data imbalance phenomenon was directly identified through the application of data balancing. We could conclude that it is necessary to balance the data on the target variables in order to perform the correct classification analysis without over-fitting. In addition, although data balancing is applied, it is shown that performance is not inferior in prediction rate when compared with a biased

prediction model.

Keyword: Decision Tree, Data Mining, Balanced Data, Elderly Suicidal Ideation, Prediction Model, Decision Rules

* 이 논문은 2019년 9월 6일 접수, 2019년 9월 21일 1차 심사, 2019년 9월 28일 게재 확정되었습니다.