「정보시스템연구」 제29권 제2호
한국정보시스템학회
2020년 6월, pp. 287~306

http://dx.doi.org/10.5859/KAIS.2020.29.2.287

머신러닝 기법을 활용한 대졸 구직자 취업 예측모델에 관한 연구

이동훈* · 김태형**

〈목 차〉

Ⅰ. 서 론

청년실업은 우리나라에서 항상 심각한 사회·경제적 이슈로 손꼽히고 있다. 15세~29세의 청년 실업률은 2017년에 2012년 7.5%에서 2.4%p 상승한 9.9%에 이르렀다(Hong, 2018). 학력별 실업률을 살펴보면 2010년 이후부터 대졸자의 실업률이 가장 가파르게 상승했으며, 고졸, 전문대졸의 실업률은 뚜렷하게 증가하는 추세를 보이지 않고 있다. 또한 25세~29세 연령대의 경우 과거 10여 년 동안 모든 학력 계층에서 실업률이 상승했지만, 특히 대졸자 집단의 실업률이 가장 급격하게 상승한 것을 확인할 수 있다(Hong, 2018).

최근 4차산업혁명 시대를 맞아 새로운 기술이 등장하며 사회문제 해결에 다양한 방식으로 적용되고 있다. 특히 빅데이터는 최근 많이 각광받는 기술로 사회문제 해결을 위하여 다양한 영역에서 적용되고 있다. 국토교통부는 2013년 국가교통DB 구축사업을 통해 교통 관련 빅 데이터를 활용해 교통 혼잡 지도를 개발하였다. 교통 혼잡 지도는 전국의 도로·도시별 교통망 성능 평가, 교통 수요 관리, 대중교통 활성화, 차량 이동량 측정 등의 차별화되는 데이터를 제공하여 교통정책 수립에 다양하게 활용되고 있다(Kim, 2014). 국민건강보험공단은 국민건

* 단국대학교 대학원 데이터지식서비스공학과, leedongs777@gmail.com(주저자)
** 단국대학교 대학원 데이터지식서비스공학과, kimtoja@dankook.ac.kr(교신저자)

강주의 알람서비스를 통하여 주요 유행성 질병 등이 확산되기 전에 관련 예방 및 치료 정보를 빠르게 제공해 선제적으로 대응할 수 있도록 하였다(Lee, 2014).

최근 청년취업과 관련하여 4차 산업혁명을 활용한 문제해결이 도입되고 있다. 인공지능을 활용한 모의면접 서비스가 개발돼 활용되고 있으며 맞춤형 직업소개도 진행해 주고 있지만, 빅데이터를 활용한 취업지원을 위한 서비스는 아직까지 초기단계에 머무르고 있는 실정이다(Park, 2016). 특히 대졸자 실업률에 대한 분석 연구는 부족한 실정이며 대부분 취업에 도움이 되는 활동을 요인분석차원에서 검증한 연구가 대부분이며, 대졸자 취업률 관련 빅데이터를 활용하여 진행한 연구는 없었다. 따라서 빅데이터를 접목하여 대졸 실업률에 영향을 미치는 다양한 요인을 분석해 보는 것이 반드시 필요하다.

본 연구는 18,199명을 대상으로 실시한 2016 대졸자직업이동경로조사(2016GOMS, 2016 Graduate Occupational Mobility Survey) 데이터를 기반으로 기계학습 기법 중 의사결정나무, 랜덤포레스트, 인공신경망 등을 활용하여 대졸자들의 취업 여부를 예측해 각 기법별 성능을 비교하고, 취업 여부에 영향을 미치는 중요 변수를 도출해내어 대졸자들의 고용 관련 정책 수립 및 다양한 취업 지원 프로그램 지원의 관련 자료로 활용됨을 목적으로 한다.

본 논문의 구성은 2장에서 연구에서 사용되는 방법론에 대해 소개하고, 3장에서는 분석 자료 설명과 독립변수, 종속변수 선정에 대해 설명한다. 4장에서는 각각의 모델에 대한 예측 결과를 제시하여 비교하고, 5장에서는 결론과 한계, 향후 연구 방향을 제시한다.

Ⅱ. 선행 연구

2.1 대졸자 실업률 관련 연구

국내뿐만 아니라 해외에서도 빅 데이터를 활용해 다양한 사회·경제적 문제 해결을 위한 시도를 하고 있다. 우리나라도 빅 데이터 분석을 통해 여러 사회 문제를 해결하기 위해 노력하고 있으며 앞서 언급한 대졸자 실업에 관련된 연구도 활발히 진행되고 있다. Gil and Choi (2014)는 대졸자의 취업 형태를 미취업, 비정규직, 중소기업 정규직, 대기업 정규직으로 나누어 각 취업 형태별 결정요인을 비교했다. 전체 취업 확률에 유의미한 영향을 미치는 변수로는 성별, 부모의 학력과 소득수준, 전공계열, 졸업평점, 외국어시험성적, 어학연수, 인턴십, 직업교육 및 훈련 등이며, 정규직 취업 확률에 유의미한 영향을 미치는 변수로는 전체 취업 확률에 영향을 미치는 변수 외에 성별, 졸업 대학의 국제화 수준 등이 있다고 했다(Lee et al, 2019). Kim(2018)는 대학 재학 시 참여한 진로 선택이나 취업 준비 관련 프로그램의 이수 여부가 정규직 취업 및 시기에 밀접한 관련이 있다고 했다. 또한 해외 어학연수 및 직업 능력 향상 관련 교육 훈련 참여가 고학력 청년층의 정규직 취업 시기에 일정하게 영향을 미친다고 하였다. Lee et al. (2019)은 대졸자 취업에 영향을 미치는 영향 중 취업, 봉사활동 차원의 중요한 요인을 도출하기 위하여 취업역량 강화프로그램의 측정항목을 기반으로 데이터 마이닝 기법을 활용하여 분석하였다. Oh(2003)는 대졸자의 취업 확률에 영향을 미치는 요인들로 성별, 4년제 여부, 가구소득, 연령 등을 도출하였으며, Chung

and Lee(2005)는 특정 대학의 학부 졸업생 1,500여명을 대상으로 학점이 졸업 후 취업에 유의미한 영향이 미치는 것으로 분석하였다. Choi and Shin(2016)은 학사데이터를 바탕으로 학점, 성별, 어학성적, 취업지원 프로그램이 취업성공 여부에 영향을 미친다고 분석하였다. 또한 대졸자직업이동경로조사 데이터를 바탕으로 분석된 연구들이 있다. Kim(2016)은 대졸자의 전공계열에 따라 취업여부, 임금수준, 고용안정성이 유의미하게 차이가 나는 것을 확인하였으며 Park and Chun(2016)은 일 경험, 성별, 전공에 따라 취업확률이 차이가 나는 것을 검증하였다.

Choi and Min(2018)은 지금까지의 청년층 취업 관련 연구들은 대부분이 이항・다항로짓, 패널선형회귀 등을 사용한 회귀분석 모형이라고 하며, 이와 달리 대졸자들의 개인적 특성이 배경 등 다양한 요인들이 취업에 얼마나 영향을 미치고 어떤 요인들이 중요 예측 인자인지를 기계학습 방법 중 랜덤포레스트 기법을 활용해 분석했다. 그 결과 대졸자의 취업 여부에 영향을 주는 요인으로 가구주 여부, 부모와의 동거 여부, 감정 빈도 변수 등을 꼽았다.

2.2 빅데이터 활용 사례

구글은 Google Express의 물류 배송 서비스에 자율 주행 차량 기술을 도입해 화물 운송 시장에 새로운 패러다임을 불러올 것으로 전망하고 있다. 구글은 Google Express를 통해 기존의 오프라인 유통 매장 제품을 구매해 배송하는 서비스를 제공하기 시작했고, 자율 주행 차량 개발에는 장기적인 관점에서 투자하며 물류 서비스로 영역을 확대하기 위한 기반 기술을 확보했다. 2015년 2월 무인배송플랫폼(Autonomous delivery platform)으로 미국특허를 취득했으며, 이 시스템은 화물 배송 플랫폼으로서 자율 주행 자동차와 무인 배송 플랫폼을 서로 연결했다. 이 시스템이 탑재된 구글의 수송 차량은 화물 수령자에게 도착 전 메시지를 보내고 수령자는 스마트폰이나 비밀번호를 통해 무인 차량 내 적재함을 열고 상품을 수령한 뒤 다시 닫을 수 있게 되어 있다(Oh, 2019).

서울시는 24시간 체제로 운영되는 서울의 경제 상황에 발맞추어 밤 12시 이후 시민들의 심야 이동권을 보장하고 심야 교통 수단 공급 부족, 택시 승차 거부, 영세 자영업자 및 직장인들의 경제적 부담을 해소하기 위해 심야 버스 정책을 도입했다. 서울시는 2013년 5월부터 2개의 노선을 시범적으로 운영하며 성공적으로 심야 버스를 정착시켰고, 이후 KT와 협업해 빅데이터 분석을 활용하여 심야버스 노선 최적화를 통한 버스 수요지를 보다 정확히 예측해 서비스에 반영시켰다. 심야버스 서비스는 현재도 활발히 운영되고 있으며, 이 정책은 '서울시에서 가장 잘한 교통복지 사례' 등 시민들의 긍정적인 지지를 받았다(Son, 2019).

2.3. 기계학습 분류 기법

2.3.1. 의사결정나무(Decision Tree)

분류 분석 기법 중 하나인 의사결정나무는 의사결정의 맥락에 근거하여 가능한 각 선택사항에 대한 확률을 할당하는 의사결정 방법이다. 확률$P(f/h)$에서, f는 선택의 집합이고 h는 결정의 맥락이다. 이러한 확률은 맥락적 차

원에서 q_1, q_2,…, q_3과 같은 일련의 질문에 의해 결정되며, 요청된 i번째 질문은 이전의 $i-1$번째 질문에 의해 고유하게 결정된다(Magerman, 1995). 의사결정나무를 활용하면 의사결정규칙(Decision Rule)을 도표화하여 대상 집단을 여러 개의 소집단으로 분류하거나 또는 예측할 수 있다. 또한 분석의 과정의 나무 구조로 표현되어 판별분석, 회귀분석 등의 다른 방법들에 비해 분석 과정과 결과를 쉽게 이해할 수 있는 장점이 있다(Choi and Seo, 1999; Kwak and Lee, 2019). 의사결정나무의 구성은 보통 뿌리(root), 가지(branches), 노드(nodes), 잎(leaf)으로 구성되며 노드는 집단을 상징하고, 가지는 노드를 연결하는 부분을 나타낸다. 의사결정나무는 보통 왼쪽에서 오른쪽으로, 가장 위의 뿌리부터 아래 방향으로 그려진다. 의사결정나무의 첫 번째 노드가 뿌리가 되며, <뿌리 - 가지 - 노드 - … - 노드> 순으로 구성되고 가장 마지막 노드를 잎이라고 한다(Zhao and Zhang, 2007). 의사결정나무 알고리즘의 종류에는 CHAID, CART, C4.5 등이 있다.

2.3.2. 랜덤포레스트(Random Forest)

랜덤포레스트는 각각의 트리가 독립적으로 표본 추출된 임의 벡터 값에 따라 달라지는 동시에, 모든 나무는 동일한 분포를 갖는 나무 예측 변수의 조합으로 구성되어 있다. 각각의 나무는 입력 벡터를 분류하기 위해 가장 인기 있는 클래스에 투표한다(Breiman, 1999). 랜덤포레스트의 일반화 오류는 나무의 수가 증가함에 따라 한계치로 거의 확실하게 수렴한다. 랜덤포레스트 나무 분류기의 일반화 오류는 개별 나무의 강도와 나무들 사이의 상관관계에 따라 달라지며, 각 노드를 분할하기 위해 무작위로 선택된 속성을 사용하면 Adaboost(Freund and Shapire, 1996)에 비해 오류율이 양호하지만 노이즈에 대해서는 더 견고해진다. 내부 추정치는 오류, 강도 및 상관관계를 모니터링하며, 이러한 추정치는 분할에 사용되는 속성의 수를 증가시키기 위한 반응을 나타내기 위해 사용된다. 또한 내부 추정치는 가변 중요도를 측정하는 데도 사용되며 회귀 분석에도 적용할 수 있다(Breiman, 2001).

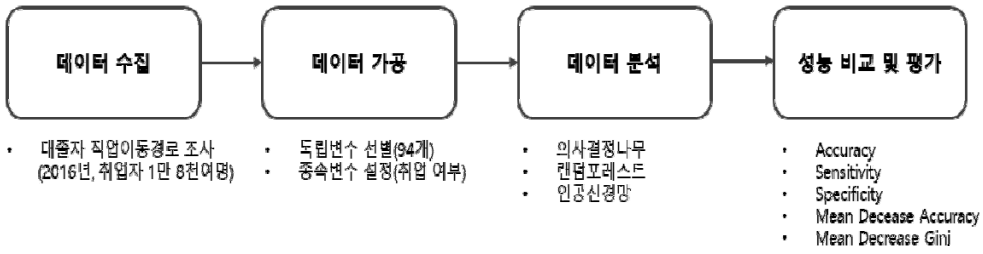
2.3.3. 인공신경망(Artificial Neural Network)

인공신경망은 신경 구조를 구성하는 가중치와 연결된 다수의 단일 단위의 인공 신경 세포로 구성된 컴퓨터 기반 모델이다. 이런 인공 신경 세포들은 정보를 처리(process)하기 때문에 처리 요소(PE, Processing Elements)라고도 한다. 각각의 PE는 가중치가 부과된 입력 값, 전달 기능, 하나의 출력 값을 가지고 있다. PE는 기본적으로 입력 값과 출력 값의 균형을 맞춰주는 방정식이다. 인공신경망에는 다양한 종류가 있지만 일반적으로 뉴런의 전달 기능, 학습 규칙, 연결 공식에 의해 설명될 수 있다(Zupan, 1992).

Ⅲ. 연구 방법

3.1 연구 개요

<그림 1>은 본 연구의 흐름을 도식화한 것이다. 본 연구는 먼저 대졸자 직업이동경로 조사 데이터를 취득한 후 독립변수를 선별하고 종속



<그림 1> 연구모형

변수를 설정하는 등의 데이터 가공을 실시했다. 그 후 R을 활용하여 의사결정나무, 랜덤포레스트, 인공신경망 모델을 생성하고 각각의 모델을 통해 대졸자들의 취업 여부를 예측하였으며, 마지막에는 각 모델의 성능을 비교하고 평가하였다.

분석 대상인 데이터는 한국고용정보원에서 실시하고 있는 GOMS 자료이다. GOMS는 매년 실시하며 전년도에 2년제 대학 이상에 해당하는 고등 교육 과정을 이수한 졸업자들을 대상으로 대학 졸업 연도의 다음 해 9월부터 3개월 동안 조사를 실시한다.

본 연구에서 사용한 2016GOMS 데이터는 총 18,199명이 대상이며 조사 항목은 총 1,295개이다. GOMS는 대졸자들이 노동 시장에 진입하는 것에 초점을 맞춘 설문 항목들로 구성되어 있으며, 졸업 후 일자리와 현재 일자리, 일자리 경험 및 구직활동 훈련, 대학 생활, 자격, 어학연수 등에 관한 문항으로 구성되어 있다. 2016GOMS의 조사 항목을 살펴보면 크게 학교생활, 경제 활동 상황, 첫 직장 일자리, 현 직장 일자리, 재학 중 경험한 일자리, 졸업 후 경험 일자리, 취업 관련 교육 및 훈련, 향후 진로, 취업 준비, 인적사항 등이 있다.

3.1.1. 독립 변수 선별

Choi and Min(2018)은 랜덤포레스트 등을 포함한 기계학습 기법의 장점 중 하나로 다양한 독립 변수들의 상호 작용과 비선형성을 고려한 예측 결과를 얻을 수 있다는 것을 꼽았다. 기존의 회귀분석 기법에 비해 기계학습 기법은 독립 변수의 수가 다소 많더라도 자유도 감소의 문제를 크게 야기하지 않아 가능한 다양한 변수들을 포함시킬 수 있는 장점이 있다. 하지만 변수가 너무 많아지면 과적합이나 과도한 분석 시간 등의 문제를 야기하는 등의 문제를 일으킬 수 있다. 본 연구에서는 랜덤포레스트 기법을 활용해 취업 여부를 예측하는데 중요하게 작용하는 변수가 어떤 것들인지 알아보고자 한다. 본 연구에서는 Choi and Min(2018)이 도출한 독립 변수 96개에서 삭제되거나 추가된 변수들을 조정해 총 94개의 변수를 사용한다. 아래의 <표 1>은 연구 모델에 사용된 독립 변수를 범주별로 나타내고 있다. 먼저 채창균, 김태기(2009)는 대졸 청년층의 취업에 영향을 미치는 요인으로 출신 대학, 전공 등 한 번 정해지면 스스로의 힘으로 바꾸기 힘든 요인들의 영향이 크다고 했다. 이에 따라 가장 기본적인 변수로 대졸자들의 성별, 연령, 고등학교 거주 지역, 결혼여부, 자녀여부, 가구주 여부 등의 인구

<표 1> 범주별 독립 변수

범주	독립 변수
인구통계학적 특성 (6개)	성별, 연령, 가구주여부, 결혼여부, 자녀여부, 고등학교 거주 지역
가족 및 부모특성 (5개)	아버지 학력, 어머니 학력, 부모님 동거여부, 부모님 소득, 가정의 경제적 지원 여부
졸업대학의 특성 (6개)	학교 위치, 졸업 대학의 유형, 주/야간 여부, 본/분교 여부, 국공립 여부, 전공계열
대학생활 및 취업관련 스펙변수 (7개)	대학 입학 전형방법, 복수전공 여부, 휴학경험, 졸업 평점, 편입 여부, 어학연수 경험 여부, 학자금 대출 여부
외국어 시험 응시 여부 (9개)	toeic, toiec_speaking, opic, verbal_etc, teps, toefl_pbt, toefl_ibt, toefl_cbt, eng_etc
대학 재학 중 취업지원 프로그램 참여 (9개)	기업취업설명회, 교내 취업박람회, 취업 관련 교과목 프로그램, 직장체험 프로그램, 인적성 검사, 면접 및 이력서 작성 프로그램, 진로관련 개인 및 집단 멘토링, 취업 캠프, 기타 프로그램
대학 재학 중 취업준비 활동 (9개)	기업체 직무 적성 검사, 이력서 작성, 면접훈련, 외국어 공부(영어 포함), 봉사 활동, 공모전 수상 경력, 자격증 준비, 외모 관리, 대외 활동
교육훈련 프로그램 및 자격증 (2개)	이수한 교육 및 훈련 프로그램의 횟수, 취득한 자격증 개수
일자리 취득정보 (1개)	주로 일자리 정보를 취득하는 경로
정부지원 청년고용정책 참여 (12개)	공공기관 청년 인턴, 청년 내일 채움 공제, 재학생 직무 체험, 취업 성공 패키지, 대학 일자리 센터, 중소기업 탐방 프로그램, 내일 배움 카드제, 일학습 병행제, 청년 취업아카데미, K-MOVE, NCS 훈련, 창업 아카데미
일자리 선택 시 고려항목 (15개)	근로 소득, 근로 시간, 자신의 적성 및 흥미, 전공 분야와의 관련성, 업무 내용의 난이도, 업무량, 개인 발전 가능성, 직업 자체의 미래 전망, 직장(고용)안정성, 근무환경, 복리후생, 회사규모, 출퇴근거리, 일자리에 대한 사회적 평판, 하는 일에 대한 사회적 평판
심리적 요인(목표, 감정빈도, 만족도) (13개)	대학전공 만족도, 학교 만족도, 대학 재학 시 취업 목표 여부, 대학 재학 시 직업 목표 여부, 삶의 만족도 - 개인적 측면, 삶의 만족도 - 관계적 측면, 삶의 만족도 - 소속집단, 한 달간 감정 - 즐거운, 한 달간 감정 - 행복한, 한 달간 감정 - 편안한, 한 달간 감정 - 짜증 나는, 한 달간 감정 - 부정적인, 한 달간 감정 - 무기력한

통계학적 변수를 포함시켰다. 다음으로 길혜지, 최윤미(2014)와 정미나, 임영식(2010) 등은 대졸자의 취업 확률에 유의미한 영향을 미치는 변수 중 하나로 부모의 학력과 소득수준과 같은 가족 관련 사항을 꼽았다. 이를 근거로 부모님 동거여부・아버지 학력・어머니 학력・부모님의 소득・경제적 지원 여부 등의 가족 및 부모의 특성 변수를 포함시켰다. 또한 위의 연구 목적에서 언급했던 선행 연구들을 기반으로 졸업 대학 유형・본/분교 여부・국공립 여부・주간여부・전공계열・학교 위치 등의 졸업 대학의 특성과 대학 입학 전형방법・복수 전공 여부・졸업평점(100점 기준)・학자금 대출여부・휴학경험・편입여부・어학연수 경험 여부 등의 대학 생활 및 취업 관련 스펙 항목 또한 포함시켰다. 또한 TOEIC・TOEIC_ SPEAKING・

OPIC · VERBAL_etc(기타 영어 말하기 시험) · TOEFL_PBT · TOEFL_CBT · TOEFL_IBT · TEPS · ENG_etc(기타 영어 시험) 등의 외국어 시험 응시 여부, 취업 관련 교과목 프로그램 · 직장체험 프로그램 · 인적성 검사 · 교내 취업박람회 · 진로관련 개인 및 집단상담 · 면접 및 이력서 작성 프로그램 · 취업캠프 · 기업 취업설명회 · 기타 프로그램 등의 대학 재학 중 취업지원 프로그램 참여 여부, 기업체 직무적성 검사 · 영어 등 외국어 공부 · 봉사활동 · 공모전 수상 · 자격증 준비 · 대외활동 · 외모관리 · 이력서 작성 · 면접훈련 등의 대학 재학 중 취업 준비 활동, 이수 교육훈련 프로그램 횟수 · 자격증 개수 등의 교육 훈련 및 자격증 관련 항목, 일자리 정보 취득 루트 항목(10개 루트), 청년 내일 채움 공제 · 공공기관 청년 인턴 · 취업 성공 패키지 · 재학생 직무 체험 · 중소기업 탐방 프로그램 · 대학 일자리 센터 · 내일 배움 카드제 · 청년 취업아카데미 · 일학습 병행제 · NCS기반 훈련 · 창업 아카데미 · K-MOVE 등의 정부지원 청년 고용정책 참여 항목, 근로소득 · 근로시간 · 자신의 적성 및 흥미 · 전공 분야와의 관련성 · 업무 내용의 난이도 · 업무량 · 개인 발전 가능성 · 직업 자체의 미래 전망 · 직장(고용)안정성 · 근무환경 · 복리후생 · 회사 규모 · 출퇴근거리 · 일자리에 대한 사회적 평판 · 하는 일에 대한 사회적 평판 등과 같은 일자리 선택 시 고려 항목을 포함시켰다. 그뿐만 아니라 대학전공 만족도 · 학교 만족도 · 대학 재학 시 취업 목표 여부 · 대학 재학 시 직업 목표 여부 · 삶의 만족도 - 개인적 측면 · 삶의 만족도 - 관계적 측면 · 삶의 만족도 - 소속집단 · 한 달간 감정 - 즐거운 · 한 달간 감정 - 행복한 · 한 달간 감정 - 편안한 · 한 달간 감정 - 짜증나는 · 한 달간 감정 - 부정적인 · 한 달간 감정 - 무기력한 등의 심리적 요인도 독립 변수로 포함시켰다.

3.1.2. 종속 변수 선정

연구의 목적은 대졸자의 구직에 영향을 미치는 다양한 요인들을 통해 대졸자들의 취업 여부를 예측하는 것이다. 따라서 종속 변수는 취업 여부(YES = 취업 / NO = 미취업)의 이항변수로 설정했다. 원래 조사 자료에서는 취업 여부가 현재 직장의 종사상 지위에 따라 상용근로자 · 임시근로자 · 일용근로자 · 고용원이 있는 자영업자 · 고용원이 없는 자영업자 · 무급가족종사자 · 미취업 등 총 7개로 구분되어 있다. 따라서 본연구에서는 <표 2>와 같이 임금을 받지 않는 무급가족 종사자는 미취업자로

<표 2> 종속변수(취업/미취업) 빈도표

취업여부	현재 직장 종사상 직위	인원(명)		비율(%)	
취업	상용근로자	10,376	13,327	57%	73.2%
	임시근로자	2,243		12.3%	
	일용근로자	160		0.9%	
	고용원이있는자영업자	205		1.1%	
	고용원이없는자영업자	343		1.9%	
미취업	무급가족종사자	33	4,872	0.2%	26.8%
	미취업	4,839		26.6%	
합계		18,199	18,199	100.0%	100.0%

간주하고 종속변수를 재구성했다.

본 연구에서는 샘플링 결과 클래스의 불균형 문제가 발생하고 있다. 사용근로자(Fulltime worker)와 고용원이 있는 일용근로자(Temporary worker)는 10,376대 160으로 약 65배의 차이를 보이고 있다. 두 집단의 비율이 아주 크게 차이가 나는 경우에는 분류 분석을 통해 두 집단을 정확히 분류하기가 어렵다. 이를 보완하기 위한 방법으로 랜덤 언더 샘플링(RUS, Random Under Sampling), 랜덤 오버 샘플링(ROS, Random Over Sampling), SMOTE (Synthetic Minority Over-sampling Technique) 등이 있다. RUS는 더 많은 비율의 계급에 있는 데이터가 무작위로 폐기시키고, ROS는 더 적은 비율의 계급에 있는 데이터를 무작위로 복제한다. SMOTE는 Chawla et al. (2002)에 의해 제안되었으며, 단순히 원래 데이터를 복제하는 것이 아니라 기존의 소수 계급의 데이터들 사이에서 추정할 수 있는 소수의 인위적인 데이터를 추가하는 방법이다. 따라서 본 연구에서는 RUS(Random Under Sampling)와 ROS (Random Over Sampling)의 단점을 보완할 수 있는 SMOTE(Synthetic Minority Oversampling Technique) 기법을 활용해 기존의 73(취업) : 27(미취업)의 비율을 55(취업) : 45(미취업)로 조정하여 데이터 불균형 문제를 해소했다.

Ⅳ. 결과

4.1. 예측 평가 지표 및 결과 요약

본 연구에서는 의사결정나무 모델, 랜덤포레스트 모델, 인공신경망 모델 등 3가지 모델을 활용해 대졸자들의 취업 여부를 예측하고 그 결과를 비교 분석했다. 이를 위해 총 94개의 독립 변수와 1개의 종속변수를 사용하고, 종속 변수의 계급 비율은 5.5(취업) : 4.5(미취업)로 조정했다. 학습 및 성능 평가를 위한 훈련 데이터와 검증 데이터의 비율은 7 : 3으로 무작위 설정하였다. 연구 도구로는 R을 활용하여 학습 및 평가를 수행했다.

성능 측정을 위한 지표로는 정확도(Accuracy), 민감도(Sensitivity), 특이도(Specificity)를 사용하고 랜덤포레스트의 경우 MDG(Mean Decrease Gini)를 추가로 사용했다.

<표 3>은 각 예측 모델들 간의 성능을 비교한 것이다. 그 결과 랜덤포레스트 모델을 사용한 경우가 정확도 및 민감도, 특이도가 모두 가장 높은 것으로 나타났다. 다음으로 인공신경망 모델의 경우 정확도와 특이도는 의사결정나무 모델보다 높았으나, 민감도는 의사결정나무에 비해 낮게 나타났다. 표준편차를 살펴보면 의사결정나무 모델이 정확도, 민감도, 특이도에서

<표 3> 각 예측모델들 간의 성능 비교

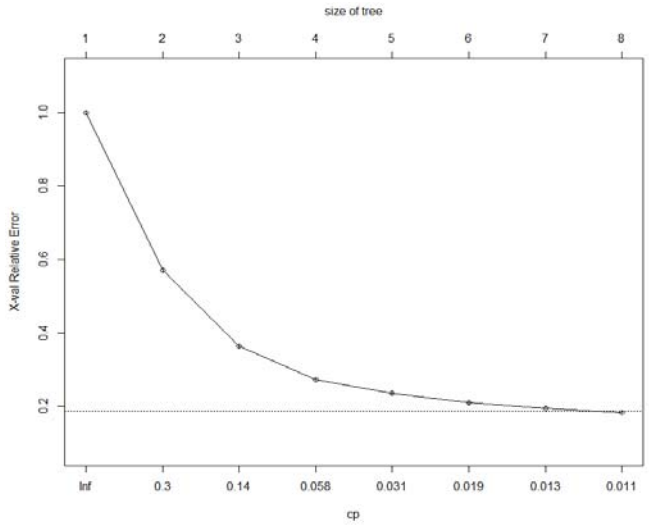
구분	의사결정나무		랜덤포레스트		인공신경망	
	평균	표준편차	평균	표준편차	평균	표준편차
정확도	0.8461	0.0022	**0.9473**	0.0038	0.8598	0.0727
민감도	0.8314	0.0059	**0.9046**	0.0083	0.7993	0.0712
특이도	0.8638	0.0045	**0.9831**	0.0045	0.9122	0.1419

모두 가장 적은 것으로 나타났다. 이는 10번의 실험을 진행할 때 각 실험의 결과 값들이 큰 차이가 없이 평균에 가깝게 산출되었다는 뜻이다. 반면에 인공신경망 모델은 의사결정나무, 랜덤포레스트보다 표준편차가 다소 큰 것으로 확인되었다. 이는 인공신경망 모델에서는 각 실험별로 결과 값들이 평균값과 차이가 있었다는 것으로 해석할 수 있다. 실제로 10번의 실험 중 의사결정나무 모델의 정확도 최댓값은 0.8486(실험 10), 최솟값은 0.8416(실험 8)으로 그 차이가 근소하지만, 인공신경망 모델의 정확도 최댓값은 0.9198(실험 9), 최솟값은 0.6671(실험 7)로 그 차이가 큰 것을 볼 수 있다. 따라서 의사결정나무, 랜덤포레스트와 같은 트리 기반(tree-based)의 방법은 훈련데이터가 달라지더라도 결과가 비교적 안정적으로 도출되며, 인공신경망은 트리 기반(tree-based)의 모델보다 훈련데이터의 변화에 민감하게 반응한다고 해석할 수 있다.

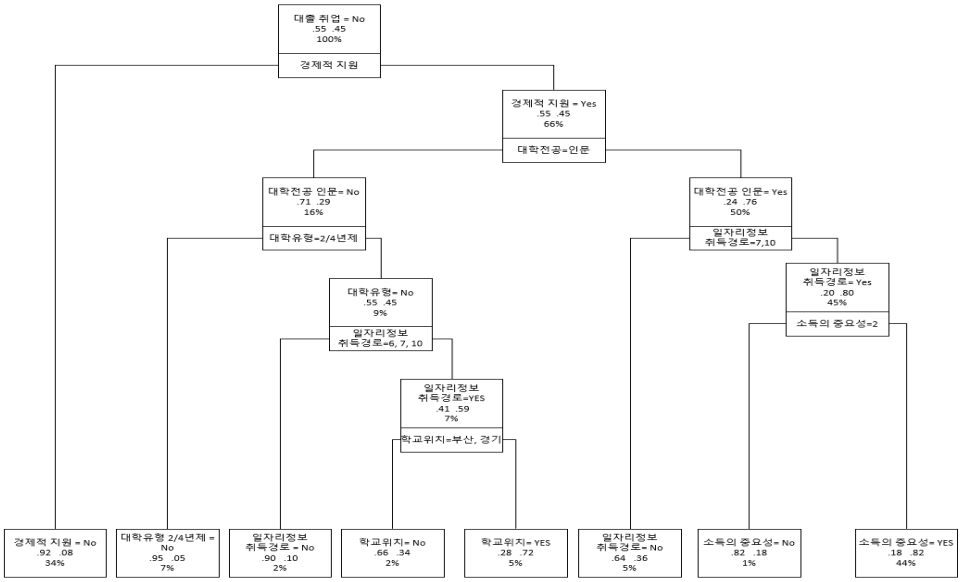
4.2. 의사결정나무 모델

본 연구에서는 R의 'rpart' 함수를 사용해 의사결정나무를 생성했다. 'rpart' 함수는 의사결정나무의 여러 기법 중 CART를 사용한다. <그림 2>는 의사결정나무 생성 후 가지 수의 증가에 따른 에러율의 변화를 나타내며, 그래프에서 가지 수가 8개일 때 에러율이 최소가 되는 것을 확인할 수 있다. 따라서 과적합(overfitting)을 막기 위해 가지 수를 8개로 하여 가지치기를 했다. <그림 3>은 가지치기를 완료한 최종 의사결정나무 모델이다.

취업대상자에 대한 경제적 지원은 취업을 하지 못하는 원인에 절대적인 영향을 미치는 것으로 확인이 되었는데, 전체 대졸 취업대상자 중에서 취업을 하지 못한 55%의 사람들 중 92%의 대상자가 경제적 지원을 받지 못한 것으로 나타났다. 또한 취업을 한 사람들 45%의 사람들 중에서 경제적 지원을 받은 사람들이 65%에 달했으며, 이 중 인문학 전공자가 76%에 달



<그림 2> 의사결정나무 가지 수에 따른 에러율 변화



<그림 3> 가지치기를 완료한 최종 의사결정나무 모델

<표 4> 의사결정나무 모델 예측 성능

구분	정확도	민감도	특이도
Experiment 1	0.8444	0.8325	0.8588
Experiment 2	0.8461	0.8313	0.8638
Experiment 3	0.8453	0.8318	0.8615
Experiment 4	0.8481	0.8365	0.8621
Experiment 5	0.8437	0.8283	0.8622
Experiment 6	0.8482	0.8402	0.8578
Experiment 7	0.8475	0.8358	0.8617
Experiment 8	0.8416	0.8173	0.8707
Experiment 9	0.8477	0.8276	0.8718
Experiment 10	0.8486	0.8328	0.8675
평균	**0.8461**	**0.8314**	**0.8638**
표준편차	**0.0022**	**0.0059**	**0.0045**

하였다. 특히 인문학 전공자 중에서 일자리 정보를 체계적으로 얻을 수 있었던 사람들이 80%에 달하였다. 따라서 변수의 중요성 중에서 소득의 중요성, 경제적 지원, 대학유형, 학교위치의 순으로 중요성이 나타났다.

<표 4>는 최종적으로 생성된 의사결정나무 모델의 예측 성능을 나타낸다. 평균 정확도는 84.6%이며, 평균 민감도는 84.1%로 이는 실제 미취업자를 의사결정나무 모델이 미취업자로 올바르게 예측한 정도를 나타낸다. 평균 특이도는 86.4%로 이는 의사결정나무 모델이 미취업자라고 예측한 값 중 실제 미취업자의 정도를 나타낸다.

4.3. 랜덤포레스트 모델

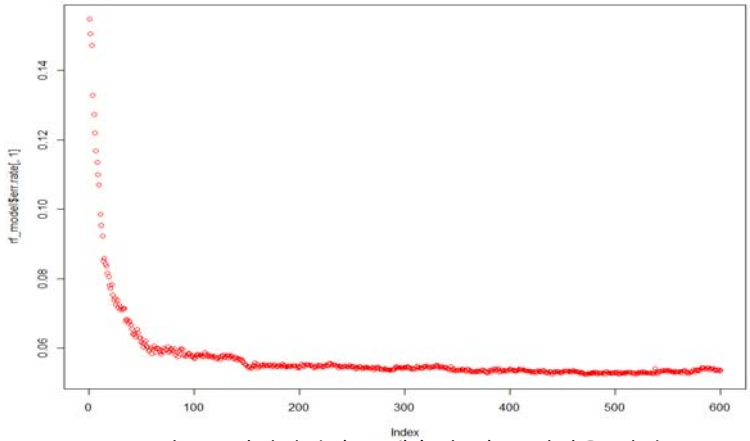
랜덤포레스트 모델은 R의 ‘randomForest’ 함수를 사용해 생성했다. <표 5>는 랜덤포레스트 모델을 사용해 대졸자들의 취업 여부를 예측한 결과이다.

랜덤포레스트로 예측한 결과, 평균 정확도는 94.7%로 의사결정나무 모델에 비해 10.1% 증가했다. 평균 민감도는 90.5%로 실제 미취업자를 랜덤포레스트가 미취업자로 올바르게 예측한 정도이며, 의사결정나무 모델보다 7.3% 상승했다. 평균 특이도는 98.3%로 랜덤포레스트 모델이 미취업자라고 예측한 값 중 실제 미취업자의 정도를 나타내며, 의사결정나무 모델보다 11.9% 상승했다.

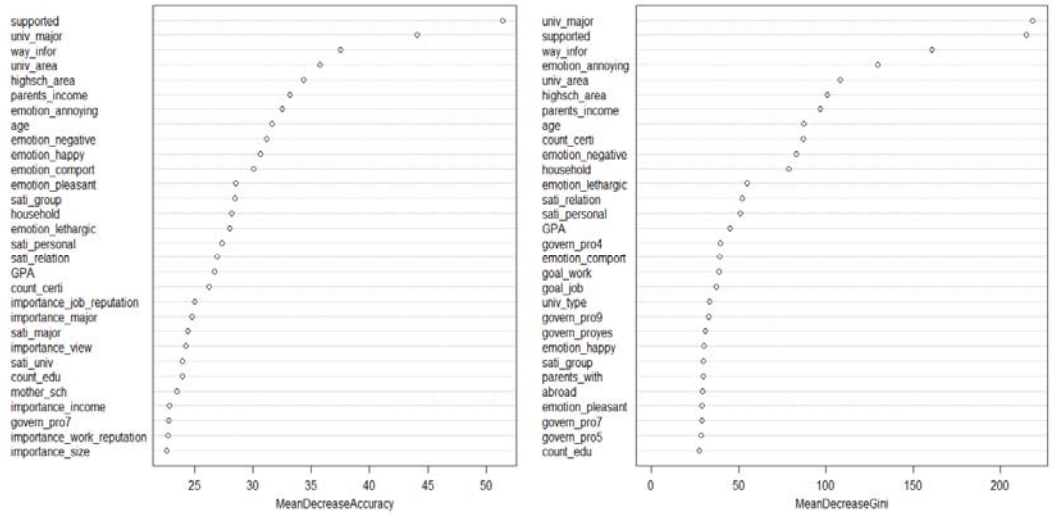
<그림 4>는 의사결정나무의 개수에 따른 에러율의 변화를 나타낸다. 랜덤포레스트 모델에서는 생성할 의사결정나무의 개수를 결정해야 하는데, 본 연구에서는 최대 600개의 의사결정나무를 생성하여 관찰한 결과, 의사결정나무의 개수가 471개 일 때 에러율이 5.23%로 가장 작았다.

<표 5> 랜덤포레스트 모델 예측 성능

구분	정확도	민감도	특이도
Experiment 1	0.9496	0.9091	0.9852
Experiment 2	0.9418	0.8956	0.9812
Experiment 3	0.9490	0.9040	0.9857
Experiment 4	0.9503	0.9194	0.9778
Experiment 5	0.9478	0.9001	0.9871
Experiment 6	0.9550	0.9158	0.9902
Experiment 7	0.9461	0.9104	0.9747
Experiment 8	0.9456	0.9008	0.9827
Experiment 9	0.9462	0.8966	0.9860
Experiment 10	0.9417	0.8944	0.9799
평균	**0.9473**	**0.9046**	**0.9831**
표준편차	**0.0038**	**0.0083**	**0.0045**



<그림 4> 의사결정나무 개수에 따른 에러율 변화



<그림 5> 랜덤포레스트 모델 변수들의 중요성 지수

<표 6>은 랜덤포레스트 모델을 활용한 10번의 실험에서 도출한 변수 중요도의 평균을 산출한 결과이다. 의사결정나무에서 변수 중요도를 산출하기 위해서 특정 변수를 제거했을 때 감소하는 정확도 차이의 평균값인 MDA(Mean Decrease Accuracy)와 랜덤포레스트 각 나무 가지별로 선택되는 변수의 불순도 감소량을 측정하는 평균치 값이 MDG(Mean Decrease Gini)가 사용된다. 본 연구에서는 MDA 값 상위 10개의 중요 변수를 산출하여 대졸자의 취업에 영향을 미치는 요인을 분석하였다. 먼저 가족의 경제적 지원을 받는지 여부가 MDA 값 0.0501로 가장 높게 나타났고, 두번째로 전공계열의 인문계열 여부가 취업에 높게 영향을 미치는 것으로 분석되었다. 그 다음으로는 한 달간 짜증 나는 감정을 느낀 정도, 일자리 정보를 얻는 경로, 취업 목표 여부, 졸업 대학의 소재지, 직업 목표 여부, 고등학교 거주 지역, 한 달간 부정적인 감정을 느낀 정도, 가구주 여부 순으로 선정되었다. 이를 범주별로 구분해보면 가족 및 부모의 특성이 1개(가족의 경제적 지원을 받는지 여부)이지만 가장 중요한 범주로 나타났고, 졸업 대학의 특성이 2개(전공 계열, 졸업 대학의 소재지)로 그 다음으로 중요한 범주로 나타났다. 또한 일자리 정보를 얻는 경로 1개의 범주도 중요하게 나타났으며 심리적 요인이 2개(한 달간 짜증 나는 감정을 느낀 정도, 한 달간 부정적인 감정을 느낀 정도) 범주로 나타났으며, 마지막으로 인구 통계학적 특성이 2개(가구주 여부, 고등학교 거주 지역)요인을 가진 범주의 순으로 나누어졌다. 따라서 MDA 결과에서는 대학 재학 시 취업에 대한 목표가 있는지 여부, 대학 재학 시 직업에 대한 목표가 있는지 여부, 짜증나거나 부정적인 감정을 느끼는 정도와 같은 심리적 요인들이 대졸자들의 취업을 예측하는데 중요한 변수라고 파악되었다.

반면에 랜덤포레스트의 나무들이 가지를 뻗을 때마다 선택되는 변수의 불순도 감소량의

<표 6> 랜덤포레스트 모델의 변수 중요도

순위	변수명	MDA	순위	변수명	MDG
1	가정의 경제적 지원여부	0.0501	1	가정의 경제적 지원여부	220.0185
2	전공계열	0.0412	2	전공계열	210.0841
3	한달간 감정-짜증나는	0.0267	3	일자리 정보를 취득하는 경로	152.2371
4	일자리 정보를 취득하는 경로	0.0264	4	한달간 감정-짜증나는	126.5622
5	대학 재학 시 취업목표 여부	0.0234	5	학교 위치	112.0212
6	학교 위치	0.0228	6	전공분야와의 관련성	106.3344
7	대학 재학 시 직업목표 여부	0.0227	7	부모님 소득	97.9603
8	전공분야와의 관련성	0.0216	8	취득한 자격증 개수	91.3499
9	한달간 감정-부정적인	0.0206	9	한달간 감정-부정적인	88.9131
10	세대주 여부	0.0201	10	연령	84.7870
11	부모님 소득	0.0196	11	세대주 여부	70.4667
12	취득한 자격증 개수	0.0196	12	삶의 만족도-개인적 측면	55.0032
13	연령	0.0174	13	한달간 감정-무기력한	54.8269
14	삶의 만족도-관계적 측면	0.0141	14	삶의 만족도-관계적 측면	52.7240
15	삶의 만족도-개인적 측면	0.0136	15	졸업평점	44.9612
16	한달간 감정-무기력한	0.0127	16	정부지원 청년고용정책-취업성공패키지 참여	39.7371
17	대학의 유형	0.0119	17	한달간 감정-편안한	39.6415
18	정부지원 청년고용정책 참여여부	0.0118	18	대학 재학 시 취업목표 여부	39.4536
19	부모님 동거여부	0.0118	19	대학 재학 시 직업목표 여부	38.2272
20	정부지원 청년고용정책-내일 배움 카드제 참여	0.0110	20	대학의 유형	35.7860

평균인 MDG(Mean Decrease Gini) 결과의 상위 10개 변수에는 가족의 경제적 지원을 받는지 여부, 전공 계열, 일자리 정보를 얻는 경로, 한 달간 짜증 나는 감정을 느낀 정도, 졸업 대학의 소재지, 고등학교 거주 지역, 부모님의 소득 수준, 자격증의 개수, 한 달간 부정적인 감정을 느낀 정도, 연령 등이 선택되었다. 이를 범주별로 구분해보면 인구통계학적 특성이 2개(고등학교 거주 지역, 연령), 가족 및 부모의 특성이 2개(가족의 경제적 지원을 받는지 여부, 부모님의 소득 수준), 졸업 대학의 특성이 2개(전공 계열, 졸업 대학의 소재지), 일자리 정보를 얻는 경로가 1개, 교육 훈련 프로그램 및 자격증이 1개(취득한 자격증의 개수), 심리적 요인이 2개(한 달간 짜증 나는 감정을 느낀 정도, 한 달간 부정적인 감정을 느낀 정도)가 있다.

분석 결과 MDA와 MDG에서 모두 중요하다고 선정된 변수에는 가족의 경제적 지원을 받는지 여부, 전공 계열, 일자리 정보를 얻는 경로, 한 달간 짜증나는 감정을 느낀 정도, 졸업 대학의 소재지, 고등학교 거주 지역, 한 달간 부정적인 감정을 느낀 정도로 파악되었다. 따라서 이러한 변수들이 대졸자들의 취업을 예측하기 위한 중요한 영향을 미친다고 할 수 있다.

4.4. 인공신경망 모델

인공신경망 모델은 R의 ‘nnet’ 함수를 사용해 생성했다. <표 7>은 은닉층(Hidden layer)이

<표 7> 인공신경망 예측모델 결과

구분	정확도	민감도	특이도
Experiment 1	0.8363	0.6617	0.9893
Experiment 2	0.8946	0.8285	0.9509
Experiment 3	0.9154	0.8351	0.9810
Experiment 4	0.8436	0.8072	0.8759
Experiment 5	0.8921	0.7640	0.9976
Experiment 6	0.8205	0.6767	0.9492
Experiment 7	0.6671	0.8760	0.4996
Experiment 8	0.9198	0.8578	0.9709
Experiment 9	0.9187	0.8348	0.9860
Experiment 10	0.8903	0.8512	0.9218
평균	0.8598	0.7993	0.9122
표준편차	0.0727	0.0712	0.1419

1개인 인경신경망 모델을 사용해 대졸자들의 취업 여부를 예측한 결과이다.

인공신경망 모델을 생성하여 예측한 결과, 평균 정확도는 86.0%로 의사결정나무 모델에 비해 1.4% 높았으며, 랜덤포레스트 모델의 비해서는 8.8% 낮았다. 평균 민감도는 79.9%로 실제 미취업자를 인공신경망 모델이 미취업자로 올바르게 예측한 정도를 나타내며, 의사결정나무 모델보다는 3.2% 낮았고, 랜덤포레스트 모델보다는 10.5% 낮았다. 평균 특이도는 91.2%로 인공신경망 모델이 미취업자라고 예측한 값 중 실제 미취업자의 정도를 나타내며, 의사결정나무 모델보다는 4.8% 높았고, 랜덤포레스트 모델 보다는 7.1% 낮았다. 인공신경망 모델은 결과가 도출되는 사용된 변수 등의 파악이 불가능해 해석이 용이하지 않았다. 따라서 본 연구에서는 의사결정나무, 랜덤포레스트 모델와 같은 트리 기반의 분석 기법이 인공신경망을 활용한 모델보다 성능, 안정성, 해석의 용이함 측면에서 모두 더 높았음을 확인할 수 있었다.

Ⅴ. 결 론

5.1. 시사점

본 연구의 분석 결과를 정리하면 다음과 같다. 첫 번째, 의사결정나무 모델에서는 10번의 실험에서 모두 가족의 경제적 지원을 받는지 여부(supported) 변수가 뿌리 노드로 선정되었다. 이 노드에서 가족에게 경제적 지원을 받고 있다고 응답한 대졸자들의 92%가 취업을 하지 못했다고 분류되었다. 실험에 따라 근소하게 차이는 있지만 경제적 지원 여부 다음의 주요 노드들로는 전공 계열, 대학의 유형(전문대학, 4년제 대학, 교육대), 일자리 정보를 얻는 경로, 직장을 선택할 때 근로 소득의 중요도, 졸업 대학의 소재지 등이 선택되었다.

두 번째, 랜덤포레스트의 나무들이 가지를 뻗을 때마다 선택되는 변수의 불순도 감소량의 평균인 MDG(Mean Decrease Gini) 결과의 상위 10개 변수에는 가족의 경제적 지원을 받는지 여부, 전공 계열, 일자리 정보를 얻는 경로,

한 달간 짜증 나는 감정을 느낀 정도, 졸업 대학의 소재지, 고등학교 거주 지역, 부모님의 소득 수준, 자격증의 개수, 한 달간 부정적인 감정을 느낀 정도, 연령 등이 선택되었다. 이를 범주별로 구분해보면 인구통계학적 특성이 2개(고등학교 거주 지역, 연령), 가족 및 부모의 특성이 2개(가족의 경제적 지원을 받는지 여부, 부모님의 소득 수준), 졸업 대학의 특성이 2개(전공 계열, 졸업 대학의 소재지), 일자리 정보를 얻는 경로가 1개, 교육 훈련 프로그램 및 자격증이 1개(취득한 자격증의 개수), 심리적 요인이 2개(한 달간 짜증 나는 감정을 느낀 정도, 한 달간 부정적인 감정을 느낀 정도)가 있다.

분석 결과 MDA와 MDG에서 모두 중요하다고 선정된 변수에는 가족의 경제적 지원을 받는지 여부, 전공 계열, 일자리 정보를 얻는 경로, 한 달간 짜증나는 감정을 느낀 정도, 졸업 대학의 소재지, 고등학교 거주 지역, 한 달간 부정적인 감정을 느낀 정도로 파악되었다. 따라서 이러한 변수들이 대졸자들의 취업을 예측하기 위한 중요한 영향을 미친다고 할 수 있다.

지금까지 대졸자 취업과 관련된 대부분의 연구들은 주로 회귀분석을 활용해 독립 변수들 간의 상관관계나 독립 변수와 종속 변수 간의 인과관계를 파악하는 것이 연구의 주된 목적이었다. 그러나 실업 문제와 같이 수요보다 공급이 더 많아 발생하는 사회 문제의 경우에는 종속 변수를 직접적으로 예측하여 종속 변수 값을 도출할 수 있는 중요 독립 변수들을 파악하는 것이 중요한 과제일 수 있다. 예를 들어 지금까지 대졸자들의 취업 관련 정책을 수립할 때는 기존의 객관적 지표들만 활용을 했다면, 향후에는 본 연구의 결과와 같이 대졸자들의 취업 여부에 영향을 미치는 중요한 변수 중 하나인 심리적 요인을 분석해 취업 관련 정책 수립을 고려해볼 수 있을 것이다. 또한 Chae and Kim(2009)가 말한 것처럼 인구통계학적 특성이나 가족 및 부모의 특성, 졸업 대학의 특성 등 스스로의 노력으로 바꾸기 쉽지 않은 요소들이 대학 생활이나, 취업 준비, 기타 스펙 등과 같은 노력보다 취업을 결정하는데 더 중요한 영향을 미치는 요소로 나타나는 점을 미루어 보아 여전히 취업 시장에서 대졸자들이 자신의 노력 여하만으로 취업을 하기에는 다소 어려운 점이 있다고 할 수 있다. 이는 우리나라의 많은 기업이나 기관 등이 채용 과정에서 아직도 인구통계학적, 졸업 대학의 특성 등 바꾸기 힘든 특성들이 주요 평가 요소로 활용하고 있음을 시사하며, 이 문제를 해결하기 위해서는 채용 과정에서 기존의 지표들이 아닌 보다 더 다양하고 객관적인 지표들을 도입하여 대졸자들이 자신의 노력 여부에 따라 취업을 할 수 있도록 해야 할 것이다. 특히 취업을 준비하는 청년들에게 더 직접적인 도움을 줄 수 있는 정책이 필요하다. 앞서도 중요도 분석결과에도 나타났듯 가정의 경제적 지원이 가장 중요한 부분인데 경제적 지원에 있어 어려움이 있는 소득분위가 낮은 청년들에게 직접적인 경제적 지원 정책이 필요하다. 또한 대학 차원에 있어서 새로운 정책이 필요하다. 일자리 정보를 얻는 경로도 매우 중요한 요소로 분류된 바, 대학별 기 존재하는 취업지원부서의 역할을 대폭 강화하여 학생들이 원하는 맞춤형 정보를 즉시적으로 제공할 수 있는 기반을 구축해야 한다. 최근 몇 대학들이 학생 빅데이터를 기반으로 인공지능 분석을 통하여 맞춤형 취업데이터를 제공해주는 사례

가 나타나고 있는데, 이러한 4차산업혁명 기술을 활용한 적극적 지원이 더욱 필요하다. 또한 학생들의 심리적 지원이 중요한바, 대학 상담센터의 기능을 강화해 취업을 전담으로 담당하는 상담의 기능도 구축할 필요가 있다.

특히 최근 개인 특성, 가정적 특성, 대학 특성을 배제할 수 있는 자기소개서 작성과 블라인드 면접을 도입하는 회사가 많아지고 있다. 이는 우리나라 채용시장에서 불고 있는 혁신이며 이에 따라 기존의 인구통계학적 특성기반의 채용시장은 점차 바뀌어 갈 수 있으리라 생각된다.

5.2. 한계 및 향후 연구 방향

본 연구에서는 취업자에 상용 근로자뿐만 아니라 임시 근로자나 일용 근로자 등 비정규직도 취업자에 포함시켜 대졸자들의 취업의 질을 구분하여 파악하는 것에 한계가 있었다. 또한 교차검증을 실시하지 못하였다. 향후 연구에서 종속변수를 종사상 지위 등으로 보다 세분화하여 정규직, 비정규직, 창업 등으로 나누고 각 계급을 예측하는데 필요한 더 많은 독립 변수들을 입력시켜 중요도를 측정하면, 이를 통해 정규직, 비정규직, 창업을 하는데 중요한 변수들을 각각 도출해낼 수 있을 것이다. 또한 본 연구에서는 랜덤포레스트 기법의 성능이 가장 높게 산출되었지만, 향후에는 보다 다양한 분류 기법을 활용하여 성능을 비교하고 더 나은 성능과 안정성을 갖춘 분류 기법을 제시하고자 한다. 나아가 GOMS 조사의 이전 자료들을 분석하여 시간의 흐름에 따른 대졸자 취업 예측을 위한 중요 변인의 변화 등을 살펴보고자 한다.

참고문헌

Agatonovic, S. and Beresford, R. “Basic concepts of artificial neural network (ANN) modeling and its application in pharmaceutical research,” *Journal of Pharmaceutical and Biomedical Analysis*, Vol. 22. 1999, pp. 718-720.

Alpaydin, E. “Introduction to Machine Learning. second edition,” *The MIT Press*, Cambridge, Massachusetts, USA. 2010. pp. 20-24.

Anyanwu. M. and Shiva. S., “Comparative Analysis of Serial Decision Tree Classification Algorithms,” *International Journal of Computer Science and Security*, Vol. 3, No. 3, 2009, pp. 233.

Breiman, L., Friedman, J., Olshen, L., and Stone, J., “Classification and Regression trees,” CHAPMAN and HALL/CRC, USA. 1989. pp. 55-58.

Breiman, L. “RANDOM FORESTS,” 1999. pp. 2-3.

Breiman, L. “RANDOM FORESTS,” 2001. pp. 7-8.

Chae, C. G. and Kim, T. G., “Analysis of Determinants of Employment Performance of Young College Students,” *The Journal of Vocational Education Research*, Vol. 28, No. 2., 2009. pp. 96-99.

Chawla, N. V., Bower, K. W., Hall, L. O., and Kegelmeyer, W. P., “SMOTE: Synthetic

Minority Over-sampling Technique," *Journal of Artificial Intelligence Research,* Vol. 16, 2002, pp. 321-357.

Choi, I. S. and Shin, E. J., "An Empirical Study of the Determinants of Successful Job Seeking of College Students - Focusing on the Impacts of Job Education Programs," *Economic Education Research,* Vol. 23, No. 1. 2016, pp. 23-49.

Choi, J. H. and Seo, D. S., "Application of data mining decision tree," *Statistical Analysis Research.* Vol. 4, No. 1. 1999, pp. 62, 63-67.

Choi, P. S. and Min, I. S., "Employment prediction model for college graduates using machine learning techniques," *Vocational competency development research.* Vol. 21, No. 1, 2018, pp. 32-38.

Chung, T. Y., and Lee, K. Y., "Determinants of Job Finding among College Graduates - with Emphasis on the Effects of GPA -," *Korea Business Review,* Vol. 8, No. 2, 2006, pp. 159-184

Géron, A. "Hands-On Machine Learning with Scikit-Learn and TensorFlow," O'Reilly, USA. 2017, pp. 10-12.

Gil, H. Y. and Choi, Y. M., "Analysis of Determinants of Employment Types of Graduates," *The Journal of Vocational Education Research,* Vol. 33, No. 6. 2014, pp. 13-19.

Hong, G. S., "Study on determinants of youth unemployment," *Analysis of the Korean economy.* Vol. 24, No. 2, 2018, pp. 91-93.

Hssina, B., Merbouha, A., Ezzikouri, H., and Erritali, M., "A comparative study of decision tree ID3 and C4.5," International Journal of Advanced Computer Science and Applications, *Special Issue on Advances in Vehicular Ad Hoc Networking and Applications:* 2014, pp. 13-18.

Jung, M. N. and Yim, Y. S., "Path analysis of variables related to entering the labor market among college graduates," *The Journal of Career Education Research,* Vol. 23, No. 2, 2010, pp. 143-146.

Kaisler, S., Armour, F., Espinosa, J., and Monet, W., "Big Data: Issues and Challenges Moving Forward," *Hawaii International Conference on System Sciences,* 46, 2016, pp. 996-997

Kang, H. Y., "Big data application examples and utilization strategies," *Magazine of the SAREK,* Vol. 45, No. 1, 2016, pp. 32-33.

Kim, D. A., Kang, D. A., and Song, J. W., "Classification Analysis for Unbalanced Data," *The Korean Journal of Applied Statistics,* Vol. 28, No. 3, 2015, pp. 495.

Kim, J. K., "Domestic and foreign big data trends and success stories," *Industrial*

Engineering Magazine, Vol. 23, No. 1, 2016. pp. 48-49.

Kim, J. S., "Big data analysis technology and use cases," *Journal of Contents*, Vol. 12, No. 1. 2014, pp. 18-19.

Kim, S. H., "The effect of job preparation activities of young college graduates on entering the labor market: focusing on whether and when to get a full-time job," *Education Culture Research*, Vol. 24, 2014, pp. 313-318.

Kwak, H. and Lee, S. W., "Competitiveness Analysis for Artificial Intelligence Technology through Patent Analysis," Journal of Information Systems, Vol 28, No. 3., 2019, pp. 141-158.

Lee, J. H., Lee, C. K. and Lee, H. H., "An Analysis of College Graduates Employment factors using Data Mining: The Importance of Volunteer Services," *Logos Management Review*, Vol. 17, No. 2., 2019, pp. 143-156.

Lee, H. W., Lee, S. R., and Chung, K, K., "The impact of Artificial Intelligence Adoption in Canadidates Screening and Joub Interview on Intentions to Apply," *Journal of Information Systems*, Vol. 28, No. 2, 2019, pp. 25-52.

Lee, Y. J., Lee, S. H., and Lee, J. S., "KB Card's marketing activities and bigdata utilization," *Korea Business Review*, Vol. 18, No. 1, 2014, pp. 162-163.

Magerman, D., "Statistical Decision-Tree Models for Parsing," ACL '95 Proceedings of the 33rd annual meeting on Association for Computational Linguistics. 1995, pp. 276-279.

Marsland, S., "Machine Learning Algorithmic Perspective," CRC Press, USA. 2015. pp. 6-9.

Mingers, J., "An empirical comparison of pruning methods for decision tree induction," *Machine Learning*, 4, 1989, pp. 241-242.

Noh, K. R. and Heo, S. J., "Analysis of Factors Influencing Achievement of Employment Goals," *The Journal of Vocational Education Research*, Vol. 1, No. 22, 2015, pp. 10-14.

Oh, J. Y, Lee, W. G., Lee, J. M. and Park, M. S., "Big Data Use Cases of Last Mile Logistics," The Korea *Institute of Information and Communication Engineering.*, 2019, pp. 121-123

Oh, S. G., "An analysis of the determinants of youth employment probability in Korea," Master's thesis, 2003, Yonsei University.

Park, G. H., "To foster leaders of the 4th Industrial Revolution Training market research," *Ministry of Employment and Labor*. 2018, pp. 167-171

Pal, M., "Random forest classifier for remote sensing classification," *International Journal of Remote Sensing*, Vol. 26, No. 1, 2005, pp. 2-3.

Park, K. Y. and Cheon, Y. M., “A Factor analysis of employment for college graduates,” *Employment survey Conference 2016*, 2016.

Podgorelec, V., Kokol, P., Stiglic, B. and Rozman, I., “Decision trees: an overview and their use in medicine,” *Journal of Medical Systems*, Kluwer Academic/ Plenum Press, Vol. 26, No. 5, 2002, pp. 9-10.

Son, J. S., “A study on transportation policy using big data: Focusing Seoul city late night bus case,” *Korean Policy Studies Review*. 2019. pp. 19-34

Sutton, R. and Barto, A., “Reinforcement Learning: An Introduction,” A Bradford Book, USA, England. 2015, pp. 2-4

Zhao, Y. and Zhang, Y., “Comparison of decision tree methods for finding active objects,” *Advances in Space Research*, Vol. 41, No. 12, 2007, pp. 3-4.

Zhu, W., Zeng, W. and Wang, N., “Sensitivity, Specificity, Accuracy, Associated Confidence Interval and ROC Analysis with Practical SAS®,” *Health Care and Life Sciences, Implementations*, 2010, pp. 1-2.

Zurada, J., “Introduction to Arificial Neural Systems,” West Publishing Company, USA. 1992, pp. 37-38

이 동 훈 (Lee, Dong Hun)

단국대학교 경영학사와 단국대학교 석사학위를 취득하였다. 현재 엑셀 및 데이터 분석 강사로 활동하고 있으며, 주요 관심 분야는 엑셀을 활용한 데이터 분석, 머신러닝, 기업 데이터 분석 등이다.

김 태 형 (Kim, Tae Hyung)

성균관대학교 언론학석사와 동국대학교 멀티미디어 디자인 박사학위와 성균관대학교 미래도시융합공학 박사학위를 취득하였다. 현재 단국대학교 데이터지식서비스공학과 교수로 재직하고 있으며, 주요 관심분야는 디자인 싱킹, 리빙랩, 스마트시티, 머신러닝 등이다.

<Abstract>

Study on the Prediction Model for Employment of University Graduates Using Machine Learning Classification

Lee, Dong Hun · Kim, Tae Hyung

Purpose

Youth unemployment is a social problem that continues to emerge in Korea. In this study, we create a model that predicts the employment of college graduates using decision tree, random forest and artificial neural network among machine learning techniques and compare the performance between each model through prediction results.

Design/methodology/approach

In this study, the data processing was performed, including the acquisition of the college graduates' vocational path survey data first, then the selection of independent variables and setting up dependent variables. We use R to create decision tree, random forest, and artificial neural network models and predicted whether college graduates were employed through each model. And at the end, the performance of each model was compared and evaluated.

Findings

The results showed that the random forest model had the highest performance, and the artificial neural network model had a narrow difference in performance than the decision tree model. In the decision-making tree model, key nodes were selected as to whether they receive economic support from their families, major affiliates, the route of obtaining information for jobs at universities, the importance of working income when choosing jobs and the location of graduation universities. Identifying the importance of variables in the random forest model, whether they receive economic support from their families as important variables, majors, the route to obtaining job information, the degree of irritating feelings for a month, and the location of the graduating university were selected.

Keyword: Machine Learning, Youth unemployment, Decision-Tree, Randome Forest, Artificial neural network.

* 이 논문은 2020년 5월 18일 접수, 2020년 5월 26일 1차 심사, 2020년 6월 19일 2차 심사, 2020년 6월 24일 게재 확정되었습니다.