석사학위논문

기계학습 분류 기법을 활용한 대졸자 취업 예측 모델 연구

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

제 출 자 : 이 동 훈

지도교수 : 김 태 형

2019

데이터지식서비스공학과

데이터사이언스 전공

단국대학교 대학원

기계학습 분류 기법을 활용한 대졸자 취업 예측 모델 연구

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

이 논문을 석사학위논문으로 제출함.

2019년 12월

단국대학교 대학원
데이터지식서비스공학과
데이터사이언스 전공

이 동 훈

이동훈의 석사학위 논문을
합격으로 판정함

심 사 일 : 2019. 12. 10.

심사위원장 서 응 교 (인)

심 사 위 원 김 태 형 (인)

심 사 위 원 황 창 하 (인)

단국대학교 대학원

(국문초록)

기계학습 분류 기법을 활용한 대졸자 취업 예측 모델 연구

단국대학교 대학원 데이터지식서비스공학과
데이터사이언스 전공
이 동 훈
지도교수 : 김 태 형

청년실업은 우리나라에서 지속적으로 대두되고 있는 사회 문제이다. 본 연구에서는 기계학습 기법 중 의사결정나무, 랜덤포레스트, 인공신경망을 이용해 대졸자들의 취업 여부를 예측하는 모델을 생성하고 예측 결과를 통해 각 모델 간의 성능을 비교한다.

그 결과 랜덤포레스트 모델을 사용한 경우가 성능이 가장 높은 것으로 나타났으며, 인공신경망 모델이 의사결정나무 모델보다 성능이 근소한 차이로 높았다. 의사결정나무 모델에서는 가족의 경제적 지원을 받는지 여부, 전공 계열, 대학의 유형 일자리 정보를 얻는 경로, 직장을 선택할 때 근로 소득의 중요도, 졸업 대학의 소재지 등이 주요 노드로 선정되었다. 랜덤포레스트 모델에서의 변수 중요도를 파악한 결과, 중요 변수로 가족의 경제적 지원을 받는지 여부, 전공 계열, 일자리 정보를 얻는 경로, 한 달간 짜증 나는 감정을 느낀 정도, 졸업 대학의 소재지 등이 선정되었다.

독립 변수의 범주별로는 인구 통계학적 특성이 2개, 가족 및 부모의 특성이 1개, 졸업 대학의 특성이 2개, 심리적 요인이 4개, 일자리 정보를 얻는 경로 1개가 중요 변수에 포함되는 것으로 나타났다.

주요용어 : 청년실업, 기계학습, 의사결정나무, 랜덤포레스트, 인공신경망

목 차

표목차

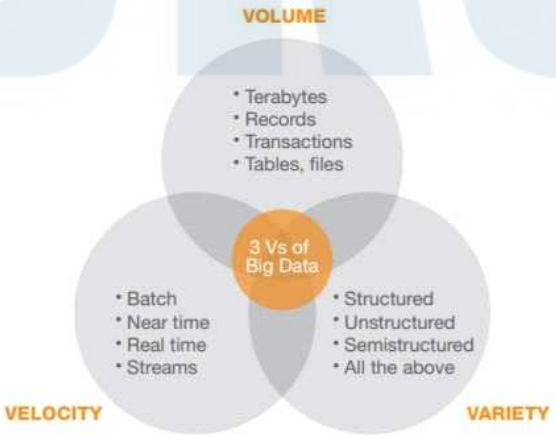
그림목차

Ⅰ. 서 론

1. 연구의 배경

1.1 빅 데이터

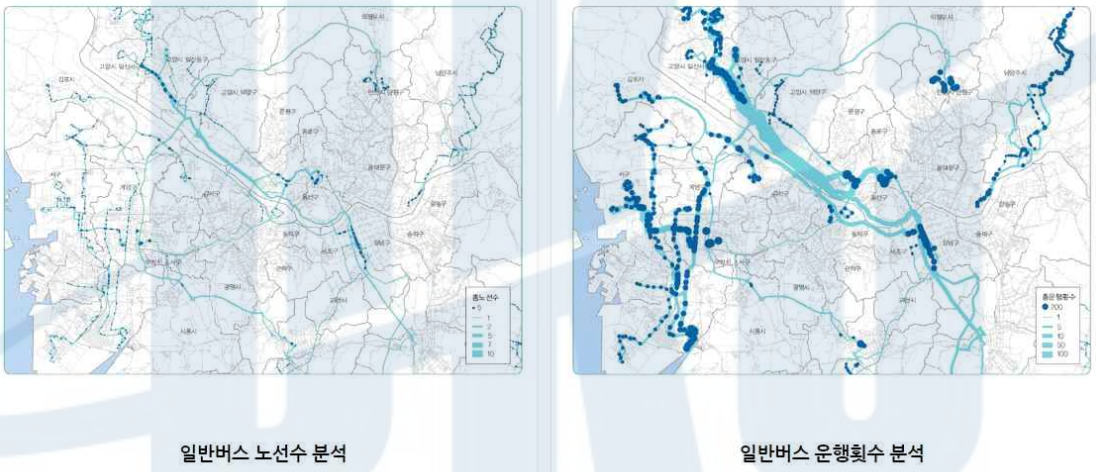
빅 데이터는 기존의 데이터베이스 방법과 도구만으로는 더 이상 효율적으로 처리할 수 없는 양의 데이터를 의미한다. 새로운 저장 매체가 발명될 때마다, 데이터에 쉽게 접근할 수 있게 되면서 데이터의 양의 폭발적으로 증가했다. 원래 빅 데이터는 구조화된 데이터에만 초점을 맞춰 정의됐지만, 많은 연구자, 실무자들은 전 세계 대부분의 정보가 텍스트나 이미지의 형태의 광범위하고 구조화되지 않은 정보로 존재한다는 것을 깨닫게 되었다. 따라서 빅 데이터는 기술의 효율적인 저장, 관리 및 처리 능력을 넘어서는 양의 데이터로 정의할 수 있다(Kaisler 등, 2016). Gartner(2011)는 기존의 데이터와 차별화되는 빅 데이터의 특징으로 '양(Volume)', '속도(Velocity)', 그리고 '다양성(Variety)'이 포함되는 3V를 꼽았다. 첫 번째, 양은 물리적인 크기뿐만 아니라 개념적인 범위까지 대규모의 데이터를 의미한다. 두 번째, 속도는 데이터가 실시간으로 생산되며 매우 빠른 속도로 유통됨을 의미한다. 세 번째, 다양성은 빅 데이터에는 기존의 구조화된 정형적인 데이터뿐만이 아니라 사진, 영상과 같은 비정형 데이터도 포함된다. 최근에는 3V와 더불어 빅 데이터를 통한 '가치(Value)' 창출도 빅 데이터의 중요한 특성으로 보고 있다(배동민 등, 2013).



[그림 1] 빅데이터의 3Vs

1.1.1 국내 빅 데이터 활용사례

국토교통부는 2013년 국가교통DB 구축사업을 통해 교통 관련 빅 데이터를 활용해 '교통 혼잡 지도'를 개발했다. 교통 혼잡 지도는 2013년 9월 한 달 동안의 차량 약 6억 개의 내비게이션의 이동궤적을 25만 개 도로 구간과 비교 분석해 도로·교차로·행정구역별로 지도에 표현하여 특정 기간의 교통 혼잡 강도를 파악하였다. 교통 혼잡 지도 시스템은 빅 데이터베이스 시스템(내비게이션 데이터, 도로 네트워크 자료), 교통 혼잡 분석시스템(혼잡 여부 판단, 각종 지표 생성), GIS 기반 표출시스템(분석 결과를 지도에 표출)으로 구성되어 있다. 교통 혼잡 지도를 활용해 분석한 결과, 광역자치단체 단위로 보면 대도시는 주중에 높은 혼잡도를 보였으며 경상남·북도, 충청남·북도, 강원도의 경우 주중보다 주말에 더 높은 교통 혼잡도를 보이는 것으로 나타났다. 교통 혼잡 지도는 향후 전국의 도로·도시별 교통망 성능 평가, 교통 수요 관리, 대중교통 활성화, 차량 이동량 측정 등의 다양한 교통정책 수립에 활용될 전망이다(김재생, 2014).



[그림 2] 국가교통DB를 활용한 대중교통 서비스 분석

신한카드는 내부 고객 데이터를 분석해 문제점과 인사이트를 발견하고 개인 맞춤형 마케팅 서비스 및 솔루션을 제공하기 위한 빅 데이터 관련 인프라를 지속적으로 강화해 나가고 있다. 2,200만 명 이상의 고객들의 개인정보, 신용정보, 거래 정보들을 다양한 채널을 통해 유입시키고 축적하고 있으며, 내부 데이터에 관련된 외부 데이터를 결합하고 이를 분석하여 마케팅 측면에서 보다 실질적인 가치 창출이 가능하다고 판단하여, 가맹점 추천, 고객 맞춤 카드 상품 추천, 다양한 카드 상품 개발 등 세 가지 영역에서 빅 데이터를 본격적으로 활용하고 있다(이유재 등, 2014).

[그림 3] 신한카드의 빅 데이터 활용 사례

보건·의료 분야에의 빅 데이터 활용의 대표적인 사례로는 보건과 의료 관련 빅 데이터에 SNS(Social Network Service) 정보를 접목한 국민건강보험공단의 국민건강주의 알람서비스가 있다. 국민건강주의 알람서비스는 국민건강정보 DB(Database)와 SNS 데이터를 연계하여 질병 관련 키워드, SNS 반응에 대한 내용을 일자 별로 집계하여 확인 가능하도록 했다. 현재 감기(인플루엔자 포함), 눈병, 식중독 등과 같은 유행성 질병의 위험도를 지역과 연령별로 구분하여 확인할 수 있도록 서비스가 제공되고 있으며, 향후 다양한 데이터 수집과 예측 정확도 향상을 통해 점진적으로 확인할 수 있는 질병의 수를 확대 제공할 전망이다. 이러한 서비스를 통해 주요 유행성 질병 등이 확산되기 전에 관련 예방 및 치료 정보를 국민들이 빠르게 제공해 선제적으로 대응할 수 있도록 한다(이지혜 등, 2014).

[그림 4] 국민건강 알림 서비스

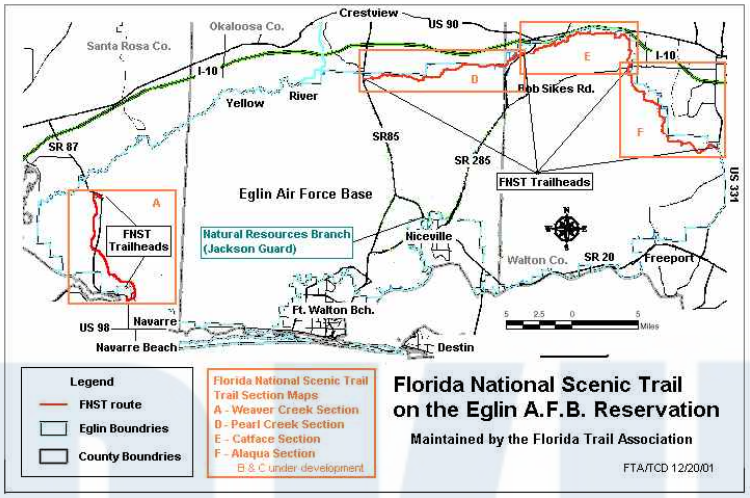
1.1.2 국외 빅 데이터 활용사례

패션브랜드인 자라(ZARA)는 빅 데이터를 분석을 생산 및 판매 전반에 걸쳐 적극적으로 활용하고 있다. 자라는 '다품종 소량생산'을 주요 마케팅 판매 전략으로 선정했다. 따라서 동종업계의 패션 브랜드들이 판매하고 있는 상품의 종류에 비해 약 2배 이상 많은 종류의 상품들을 생산한다. 또한 주문 생산, 매장 입점까지 모든 과정이 단 6주 이내에 완료된다. 이를 위해 수요를 미리 예측하고 각 매장의 재고 산출, 상품의 가격 결정, 운송 정보 등의 실시간 정보를 파악할 수 있는 빅 데이터 기반의 재고 관리 시스템을 MIT 연구팀과 연계해 개발하여 활용하고 있다(김정경, 2016).

[그림 5] ZARA의 다품종 소량생산 시스템

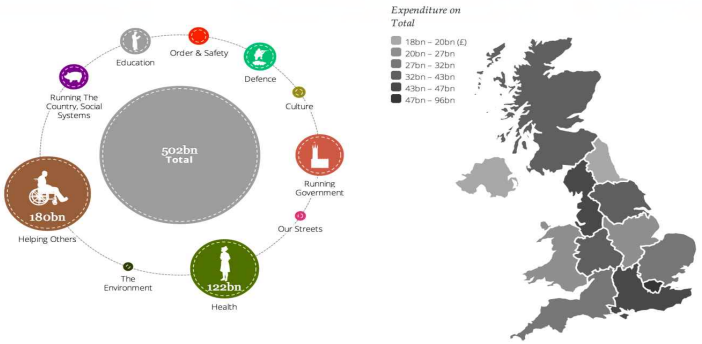
미국의 엔지니어링 회사인 맥키니(McKenney's)는 건물 자동화와 제어시스템 등 다양한 서비스를 제공하고 있다. 2012년에는 수백 개의 빌딩이 있고 약 이만 여명의 인력이 상주하며 세계에서 가장 큰 군 기지 중 하나인 에글린 공군기지의 에너지 관리 시스템을 설계 및 구축했다. 빅 데이터의 수집하고 이를 모니터링 및 분석하는 에글린 관리 시스템(Eglin Management System)은 기지 내 빌딩의 중앙 냉·난방 공조 설비에 부착된 수많은 센서에서 발생하는 데이터를 수집하고 이를 대시보드에 표현해 빌딩의 성능을 파악하고 에너지 사용의 효율성을 제고할 수 있게 했다. 이 뿐만 아니라 공군기지 내에 있는 빌딩들의 에너지 사용량이 자동으로 집계되고, 이전 에너지 사용량과 현재의 사용량을 비교 분석할 수 있는 기능도 구현되었다.

그 결과 에글린 공군기지는 효율적인 전력 사용 전략을 수립해 연간 약 10억 원의 비용 절감 효과가 있었다고 밝혔다(강희용, 2016).



[그림 6] 에글린 공군기지의 에너지 관리 시스템

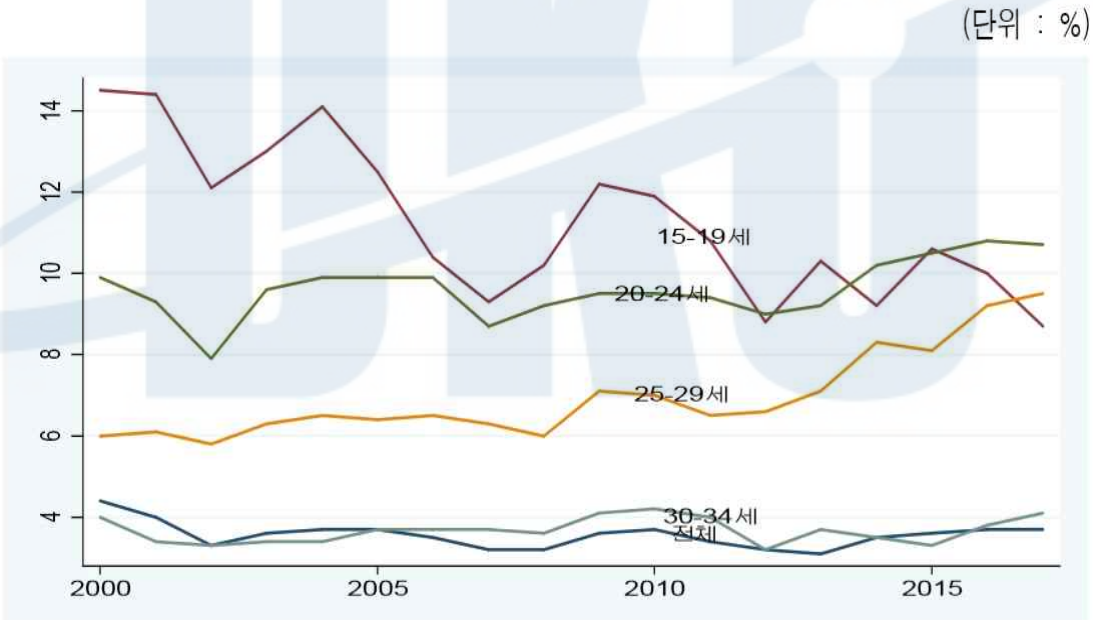
민간기관인 공개지식재단(Open Knowledge Foundation)의 웹 사이트에서는 영국 정부가 제공하고 있는 공공 데이터를 사용하여 일반인들도 이해하기 쉽도록 여러 시각화 기법을 사용하여 나타내고 있다. 그중 가장 대표적인 서비스는 '내가 납부한 세금은 어디에 쓰이는가?(Where does my money go?)'라는 상호작용이 가능한 어플리케이션이다. 개인의 수입에 상응하는 금액을 입력하면 그에 대한 세금 납부액과 해당 세금이 사용되는 항목을 의료, 복지, 국방, 교육 등의 카테고리를 2단계로 도식화해 제공하며 일반인들의 납세에 대한 이해를 돕고 있다(이만재, 2011).



[그림7] 'Where Does My Money GO' 서비스

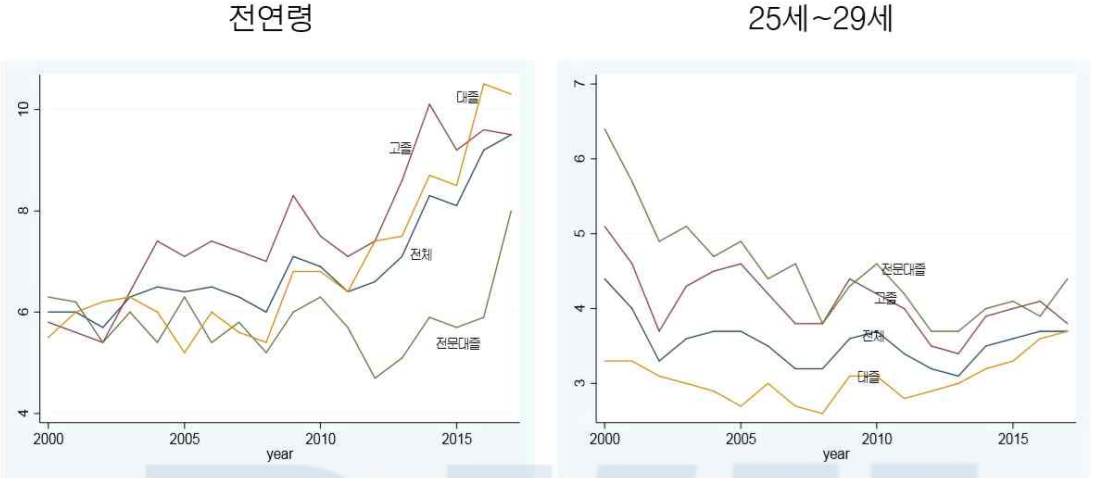
1.2 대졸자 실업률 증가

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. 이를 5세 연령별로 구분해 보면 [그림 8]과 같이 특히 25세~29세의 실업률이 지속적으로 상승하고 있는 것을 확인할 수 있다. 15세~19세의 경우 실업률이 꾸준히 하락하고 있을 뿐 아니라 다른 연령대에 비해 경제활동인구의 규모가 매우 작은 편이므로 15세~19세의 연령대가 15세~29세의 전체 청년 실업률에 미치는 효과는 크지 않다고 할 수 있다. 30세~34세 연령대의 실업률은 전체 청년 실업률과 비슷한 수준을 안정적으로 유지하고 있다. 학력별 실업률을 나타내는 [그림 9]에 의하면 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. 이에 따라 현재 청년실업 문제를 근본적으로 해결하기 위해서는 25세~29세의 대졸자 집단에 집중하여야 한다고 볼 수 있다(홍기석, 2018).



[그림 8] 연령대별 실업률 추이

(단위 : %)



[그림 9] 학력별 실업률 추이

2. 연구의 목적

위 연구의 배경에서 살펴봤듯이, 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. 채창균, 김태기(2009)는 대학을 졸업한 청년층의 취업에 영향을 미치는 요인으로 출신 대학, 전공 등 한 번 정해지면 스스로의 힘으로 바꾸기 힘든 요인들의 영향이 큰 반면, 학교의 교육 지원, 일자리 경험, 어학연수, 자격증 취득과 같은 취업을 위한 노력은 크게 긍정적인 영향을 미치지 못한다고 했다. 정미나, 임영식(2010)은 대학을 졸업한 청년층이 노동시장에 진입하는데 관련된 변인으로 가구 소득(월평균), 전공학과의 향후 전망, 첫 번째 직장의 업무에 대한 전공 지식의 도움 정도, 급여, 고용 안정성 등의 구직 조건의 중요도, 전공 계열 등이 있다고 했다. 길혜지, 최윤미(2014)는 대졸자의 취업 형태를 미취업, 비정규직, 중소기업 정규직, 대기업 정규직으로 나누어 각 취업 형태별 결정요인을 비교했다. 먼저 전체 취업 확률에 유의미한 영향을 미치는 변수로는 성별, 부모의 학력과 소득수준, 전공계열, 졸업 평점, 외국어시험성적, 어학연수, 인턴십, 직업 교육 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. 그 외에 대기업 정규직 취업을 위해서는 위에서 언급한 변수들 외에 졸업 대학의 평판 및 사회 진출도와 교육과정 만족도가 있다고 했다. 노경란, 허

선주(2015)는 취업 목표 달성에 영향을 미치는 요인 중 정적 요인은 직업 관련 교육 및 훈련 참여 정도, 취득 자격증과 전공과의 관련성이 있고, 부적 요인으로는 인터넷 정보 활용성을 꼽았다. 김수혜(2018)는 대학에 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.

최필선, 민인식(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.

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.

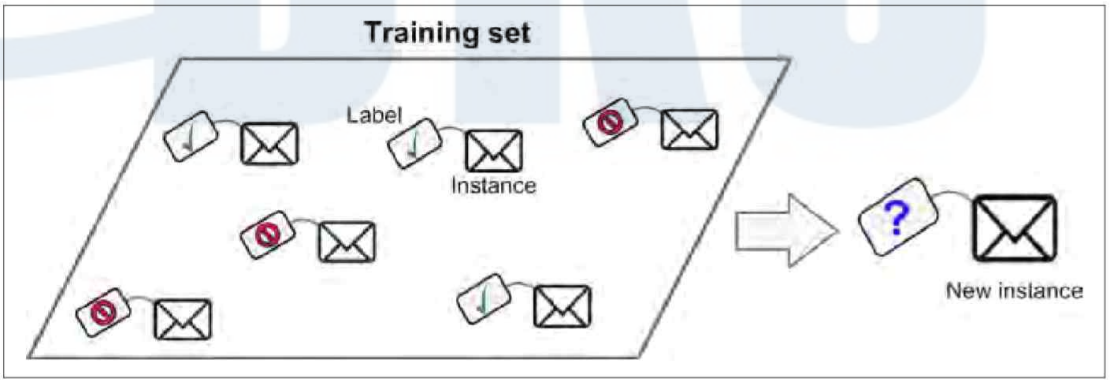
Ⅱ. 연구 방법론

1. 기계학습(Machine Learning)

1.1 지도학습(Supervised Learning)

기계 학습은 예시 데이터 또는 과거 경험을 활용해 성능 기준을 최적화시키는 컴퓨터 프로그래밍이다. 기계학습은 매개 변수까지 가기 위해 훈련 데이터 또는 과거 경험을 사용하여 학습시킨 특정한 모델에서 매개 변수의 최적화된 값을 찾기 위한 프로그램을 실행하는 것이다. 이러한 모델을 통해 우리는 미래를 예측하거나, 데이터로부터 얻은 지식을 설명할 수 있다(Alpaydin, 2010).

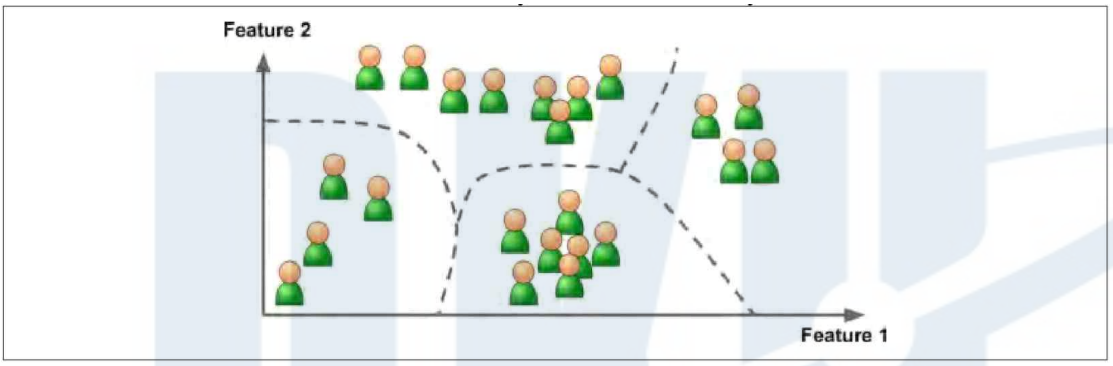
지도학습은 종속 변수의 정확한 응답 값(목표)이 있는 훈련 데이터 셋이 주어지고, 이 훈련 데이터 셋을 기반으로 한 알고리즘은 가능한 모든 독립 변수에 대한 정확한 종속 변수의 응답 값을 제공한다. 지도학습 등의 머신러닝이 보편적인 일반화보다 더 좋은 이유는 머신러닝의 알고리즘이 학습하는 동안에는 접하지 못했던 독립 변수 값이 입력되었을 때, 이에 대한 합리적인 종속 변수의 응답 값을 도출할 수 있다는 점이다. 또한 머신러닝 알고리즘은 모든 실제 현실에서 측정될 수 있으며 데이터에 내재되어 있는 부정확성을 처리할 수 있게 된다. 지도학습의 종류에는 회귀분석(Regression), 분류분석(Classification) 등이 있다(Marsland, 2015).



[그림 10] 지도학습을 활용한 스팸 메일 분류 시스템

1.2 비지도학습(Unsupervised Learning)

비지도학습은 지도학습과 반대로 훈련 데이터 셋에서 독립 변수에 대한 종속 변수, 즉 응답 값이 존재하지 않는 머신러닝 기법이다. 시각화 알고리즘(Visualization Algorithms)은 비지도학습의 좋은 예로서, 복잡하지만 종속 변수가 따로 없는 많은 양의 데이터를 제공하고, 알고리즘에 입력하면 해당 알고리즘은 입력 데이터에 대한 결과를 2D 또는 3D 그래프 등으로 출력한다. 비지도 학습에는 군집화(Clustering), 시각화 및 차원 감소(Visualization and Dimensionality Reduction), 연관 규칙 학습(Association Rule Learning) 등이 있다(Géron, 2017).



[그림 11] 블로그 방문자 클러스터링 시스템

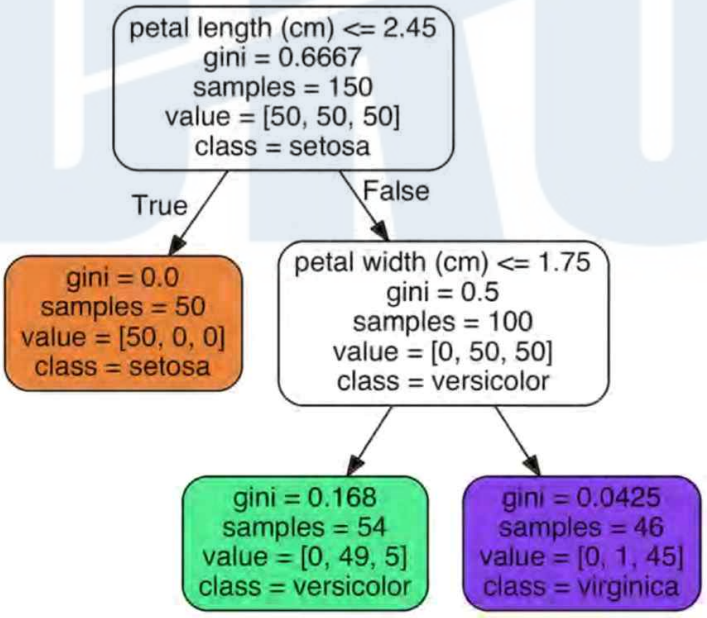
1.3 강화학습(Reinforcement Learning)

강화 학습은 목표 지향적 학습 및 의사 결정을 이해하고 자동화하는 전산적 접근 방법이다. 모범적인 또는 완전한 환경의 모델에 의존하지 않고 강화 에이전트가 환경과 직접적으로 상호작용하며 학습한다는 점에서 다른 전산 접근법과 구별된다. 강화 학습은 상태, 작업 및 보상의 측면에서 학습 에이전트와 환경 사이의 상호 작용을 정의하는 프레임 워크를 사용하며, 이 프레임 워크에는 원인과 결과, 불확실성과 비결정론, 명시적 목표의 존재 등이 포함된다. 강화학습은 로봇 공학, 게임 개발 등의 분야에서 활발히 활용되고 있다(Sutton, Barto, 2015).

2. 분류분석(Classification)

2.1 의사결정나무(Decision Tree)

분류 분석 기법 중 하나인 의사결정나무는 의사결정의 맥락에 근거하여 가능한 각 선택 사항에 대한 확률을 할당하는 의사결정 방법이다. 확률 $P(f|h)$에서, f는 선택의 집합이고 h는 결정의 맥락이다. 이러한 확률은 맥락에 대한 q_1, q_2,…, q_n과 같은 일련의 질문에 의해 결정되며, 요청된 i번째 질문은 이전의 $i-1$번째 질문에 의해 고유하게 결정된다(Magerman, 1995). 의사결정나무를 활용해 의사결정규칙(Decision Rule)을 도표화하여 대상 집단을 여러 개의 소집단으로 분류하거나 또는 예측할 수 있다. 분석의 과정의 나무 구조로 표현되어 판별분석, 회귀분석 등의 다른 방법들에 비해 분석 과정과 결과를 쉽게 이해할 수 있는 장점이 있다(최종후, 서두성, 1999). 보통의 의사결정나무는 뿌리(root), 가지(branches), 노드(nodes), 잎(leaf)으로 구성되어 있다. 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, Zhang, 2007). 의사결정나무 알고리즘의 종류에는 CHAID, CART, C4.5 등이 있다.



[그림12] Iris 데이터를 활용한 의사결정나무

2.1.1 CHAID

의사결정나무 중 하나인 CHAID(Chi-Squared Automatic Interaction Detection)는 종속 변수가 이산형일 때는 카이제곱-검정을, 종속 변수가 연속형일 때는 F -검정을 이용하여 2개 이상의 다지분리를 수행하는 의사결정나무 알고리즘이다. 종속 변수가 범주형인 경우에는 Pearson의 카이제곱 통계량[1]이나 우도비 카이제곱 통계량[2]을 분리 기준으로 한다. 카이제곱 통계량은 관측 도수(f_{ij})로 구성된 $r \times c$ 분할표를 활용해 계산한다. 분할표는 [표 1[과 같다.

[표 1] 카이제곱 통계량 분할표의 구조

종속변수 / 독립변수	범주 1	범주 2	...	범주 c	합 계
범주 1	f_{11}	f_{12}	...	f_{1c}	$f_{1.}$
범주 2	f_{21}	f_{22}	...	f_{2c}	$f_{2.}$
...	...	...	...	...	...
범주 r	f_{r1}	f_{r2}	...	f_{rc}	$f_{r.}$
합 계	$f_{.1}$	$f_{.2}$	...	$f_{.c}$	$f_{..}$

Pearson의 카이제곱 통계량과 우도비 카이제곱 통계량의 자유도는 $(r-1)(c-1)$로 동일한데, 여기서 e_{ij}는 분포의 동일성, 혹은 독립적이라는 전제 하의 기대도수(expected frequency)를 뜻하며, 이 기대도수는

$$e_{ij} = \frac{f_i \times f_j}{f_{..}} \quad (1)$$

과 같이 계산한다.

자유도에 비해 카이제곱 통계량이 작다는 것은 예측 변수의 각 범주에 따라 목표 변수의 분포가 서로 같다는 것을 의미한다. 따라서 예측 변수가 목표 변수를 분류하는데 영향을 미치지 않는다고 할 수 있다. 자유도에 대해 카이제곱 통계량의 크기는 $P-$값으로 표현할 수

---

1) $x^2 = \sum_{i,j} \frac{(f_{ij} - e_{ij})^2}{e_{ij}}$ (2)

2) $x^2 = 2\sum_{i,j} f_{i,j} \times \log_e(\frac{f_{ij}}{e_{ij}})$ (3)

있는데, 자유도에 비해 카이제곱 통계량이 작으면 $P-$값이 커지게 된다. 결론적으로 분리 기준으로 카이제곱 통계량 값을 사용하면, $P-$값이 가장 작은 예측 변수와 그에 통한 최적 분리에 의해 자식마디가 형성된다는 것을 의미한다(최종후, 서두성, 1998).

2.1.2 CART

두 번째로, 1984년 Breiman에 의해 도입된 CART(Classification And Regression Trees)는 분류 나무와 회귀 나무 모두 생성이 가능하다. CART로 만든 분류 나무는 속성들의 이진 분할을 기반으로 하며, 연속적으로 실행할 수 있다(Breiman, 1984). CART는 분할 속성을 선택할 때 지니 지수(gini index) 분할 측정을 사용한다. CART에서의 가지치기는 훈련 데이터 셋의 일정 부분을 사용하여 수행된다(Podgorelec 등, 2002). CART는 의사결정나무를 구축하기 위해 연속형 속성과 범주형 속성을 모두 사용하며, 누락 속성을 처리하기 위한 내장된 속성을 가지고 있다. 회귀 분석 속성은 일련의 예측 변수들이 주어졌을 때 종속 변수를 예측하는데 사용된다. 이러한 속성으로 지니 지수, 심피니 등과 같은 많은 단일 변수 분할 기준과 하나의 다중 변수(선형 조합)가 사용되고 최적의 분할점을 결정하기 위해 모든 노드에서 데이터가 정렬된다. SALFORD SYSTEMEMS는 Breimanm의 코드를 사용해 CART®라는 버전을 구현했다. CART®는 CART의 단점을 해결할 수 있도록 그 특징과 기능이 향상되었으며, 이로 인해 보다 예측 정확도가 높은 현대적인 의사결정나무 분류기가 만들어졌다.

2.1.3 C4.5

Ross Quinlan는 자신이 고안한 ID3 알고리즘의 한계를 극복하기 위해 C4.5를 제안했다. ID3의 제한 중 하나는 입력되는 독립변수의 양이 많을 때 알고리즘이 지나치게 민감해진다는 것이다. 입력 변수의 양이 많을 때 ID3의 민감도는 사회보장 번호로 설명될 수 있다. 사회보장 번호는 개인별로 고유하기 때문에 그 가치를 테스트해보면 항상 조건부 엔트로피(entropy)[3]의 값이 낮아진다. 이러한 문제를 해결하기 위해 C4.5는 '정보 이득(Information

---

3) 엔트로피(entropy) : $Entropie(P) = -\sum_{i=1}^{n} p_i \times \log(p_i)$ (4)

일반적으로 확률 분포 $P = (p_1, p_2, \ldots, p_n)$와 표본 S가 주어졌을 때, 이러한 분포에 의해 전달되는 정보를 확률 분포 P에 의해 주어지는 엔트로피라고 한다. B, Hssina, etc (2014) A comparative

Gain)[4]'을 이용한다. 정보 이득을 계산하는 것 자체가 새로운 것을 생성하지는 않지만, 이를 통해 이득 비율을 측정할 수 있다(Hassina 등, 2014).

이득 비율은 다음과 같이 정의된다.

$$GainRaito(p,\ T) = \frac{Gain(p,\ T)}{SplitInfo(p,\ T)} \qquad (5)$$

$$SplitInfo(p, test) = -\sum_{j=1}^{n} P'\left(\frac{j}{p}\right) \times \log(P'\left(\frac{j}{p}\right)) \qquad (6)$$

$P'(j/p)$는 p위치에 존재하는 원소들의 비율이며, j번째 테스트의 값을 갖는다. 엔트로피와 달리, 앞서 말한 이 정의는 다른 클래스들의 예시 분포와 독립적이다. C4.5는 ID3과 같이 분할 속성(splitting attribute)을 결정하기 위해 데이터가 정렬된다. 이때, 정보 이득 비율 불순도 분석 방법을 사용해 분할 특성을 평가한다. C4.5는 트리의 각 노드에서 특정 클래스 또는 다른 클래스의 샘플들을 강화된 부분집합으로 분할할 수 있는 가장 효과적인 하나의 속성을 선택한다. 이 기준은 데이터를 분할하는 속성을 선택함으로써 발생하는 정규화된 정보 이득인데, 정규화 된 정보 이득이 가장 높은 속성이 결정 변수로 선택된다.

2.2 랜덤포레스트(Random Forest)

의사결정나무를 발전시킨 랜덤 포레스트는 각각의 트리가 독립적으로 표본 추출된 임의 벡터 값에 따라 달라지는 동시에, 모든 나무가 동일한 분포를 갖는 나무 예측 변수의 조합이다. 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

---

study of decision tree ID3 and C4.5

4) 정보 이득(Information Gain) : $Gain(p, T) = Entropie(p) - \sum_{j=1}^{n}(p_i \times Entropie(p_j))$ (7)

위 공식에서 p_j의 값은 속성 T에 대해 가능한 모든 값의 집합이다. 이 방법을 통해 속성들의 순위를 매기고 뿌리로부터의 경로에서 고려되지 않은 속성들 사이에서 가장 높은 정보 이득을 가진 속성들이 위치한 각 노드로 의사결정나무를 만들 수 있다. *Ibid*

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). 랜덤 포레스트 분류기는 속성을 선택하는 척도로 지니 지수를 사용하고, 이를 통해 클래스와 관련하여 속성의 불순도를 측정한다. 임의로 하나의 케이스(case)를 선택하고 그것이 특정 클래스에 속해 있는 훈련 데이터 T의 경우, 지니 지수는 다음과 같다(Pal, 2005).

$$\sum\sum_{j \neq i}(f(C_i, T)/|T|)(f(C_j, T)/|T|) \quad (8)$$

여기서 $f(C_i, T)/|T|$는 클래스 C_i에 속한 케이스가 선택될 확률이다.

하나의 나무는 매번 속성들의 조합을 사용해 새로운 데이터의 최대 깊이까지 자란다. 이렇게 완전히 자란 나무들은 가지가 쳐지지 않은 상태이다. 이것이 다른 의사결정나무 방법들과 달리 랜덤포레스트가 가지는 주요한 장점 중 하나다(Quinlan, 1993). 기존의 연구 결과들에 따르면 특성을 선택하는 방법이 아닌, 가치지치 방법을 선택하는 것이 의사결정나무의 성능에 영향을 미친다고 하였다(Mingers, 1989). 반면에 Breiman(1999)은 나무의 수가 증가하면 가지를 치지 않아도 항상 일반화 오류가 수렴하며, 대 수의 강법칙(Strong Law of Large Numbers)에 따라 과적화(Overfitting)가 발행하지 않는다고 했다. 각 노드에서 트리를 생성하는 데 사용되는 속성의 수와 나무를 자라게 하는 속성의 수는 랜덤포레스트를 형성하는데 필요한 사용자 정의 파라미터(parameter)다. 각각의 노드에서 선택된 속성들만이 최적의 분할을 위해 검색된다. 따라서 랜덤포레스트는 N개의 나무로 구성되며, 여기서 N은 사용자에 의해 정의된 나무의 수이다. 새로운 데이터를 분류하기 위해 데이터 셋의 각 케이스(case)는 각각의 N 나무로 전달된다. 랜덤포레스트는 이때 N표 중 가장 많은 표를 얻은 클래스를 선택한다(Pal, 2005).

랜덤포레스트는 CART와 같은 의사결정나무와 다르게 붓스트래핑(Bootstrapping)을 통해 반복적으로 의사결정나무를 생성하기 때문에 분류 예측에 있어서 안정성을 확보할 수 있는 장점이 있는 반면에, 최종으로 분류된 결과를 트리 형태로 확인할 수 없기 때문에 예측 모형을 직관적으로 이해하기 어렵다는 단점이 있다(최필선, 민인식, 2018).

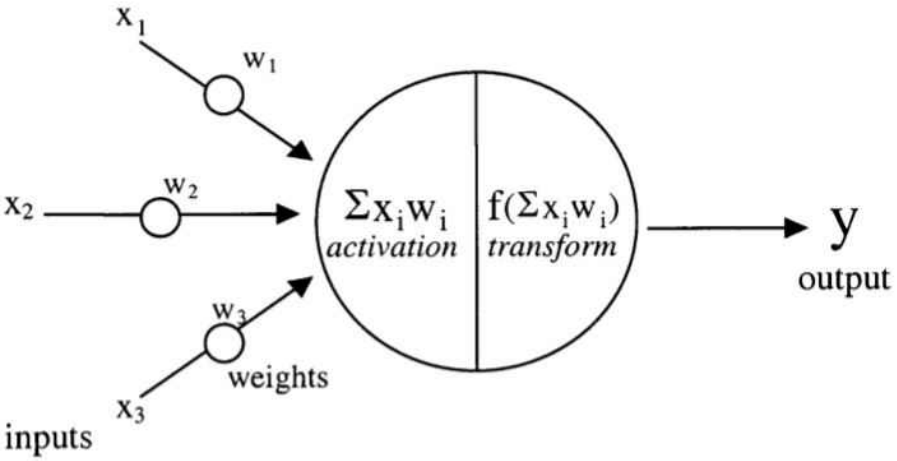
2.3 인공신경망(Artificial Neural Network)

인공신경망(ANN)은 인간의 두뇌를 디지털화한 모델로서, 인간의 뇌가 정보를 처리하는 방식을 시뮬레이션하기 위해 고안된 컴퓨터 프로그램이다. ANN은 컴퓨터 프로그래밍으로부터가 아닌 사람들이 배우는 방법처럼 적절한 학습 예시들을 가진 경험을 통해 학습하거나 훈련받는다. 신경망은 데이터의 패턴과 관계를 탐지함으로써 그것들에 대한 지식을 모은다. 뇌는 훌륭한 패턴 인식 도구인데, 예를 들어 우리가 펜을 볼 때, 우리는 그것이 펜이라는 것을 안다. 그 이유는 우리 뇌의 특정 영역에 있는 생물학적 뉴런들이 이전에도 비슷한 입력 패턴을 접했고 그 특정한 패턴을 '펜'과 연결하는 방법을 학습했기 때문이다. 우리의 뇌는 완전하게 연결되어 있는 수십억 개의 뉴런을 포함하고 있기 때문에, 수없이 다양한 입력 패턴을 배우고 인식할 수 있다(Agatonovic, Beresford, 1999).

인공신경망은 신경 구조를 구성하는 계수(가중치)와 연결된 수백 개의 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 뉴런(neurons)의 전달 기능(transfer function), 학습 규칙(learning rule) 그리고 연결 공식(connection formula)에 의해 설명될 수 있다(Zupan 등, 1992).

2.3.1 뉴런(Neurons)

인공 뉴런은 생물학적 뉴런의 기능을 시뮬레이션하기 위해 고안된 ANN의 구성 요소이다. 입력 값인 도달 신호에 조정된 연결 가중치를 곱한 후 전달 함수(transfer function)를 통해 해당 뉴런의 출력 값을 생성한다. 활성화 함수(activation function)는 뉴런 입력 값의 가중치의 합이며, 가장 일반적으로 사용되는 전달 함수로는 시그모이드 함수(sigmoid function)가 있다(Zurada, 1992).

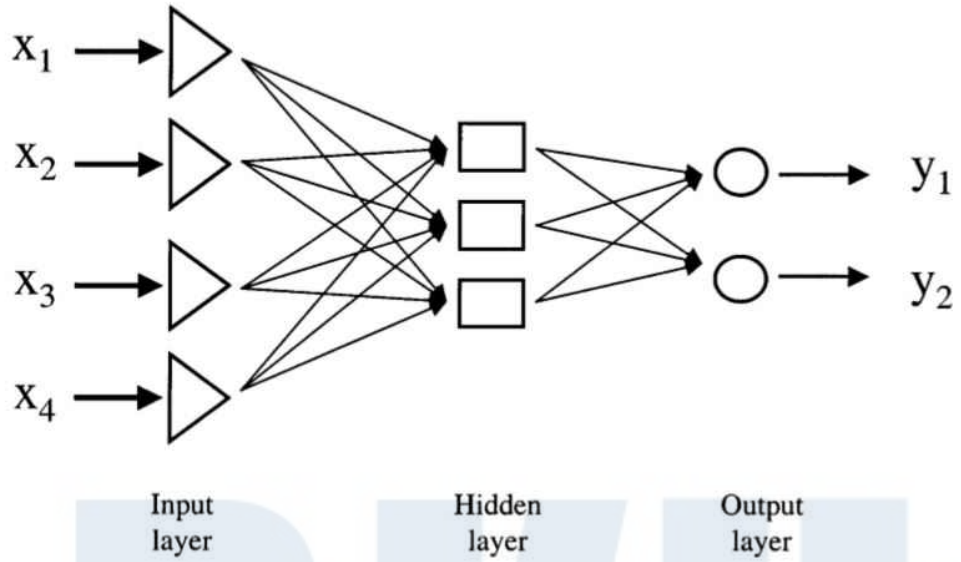


[그림 13] 인공 뉴런 모델

2.3.2 연결 함수(Connection Formula)

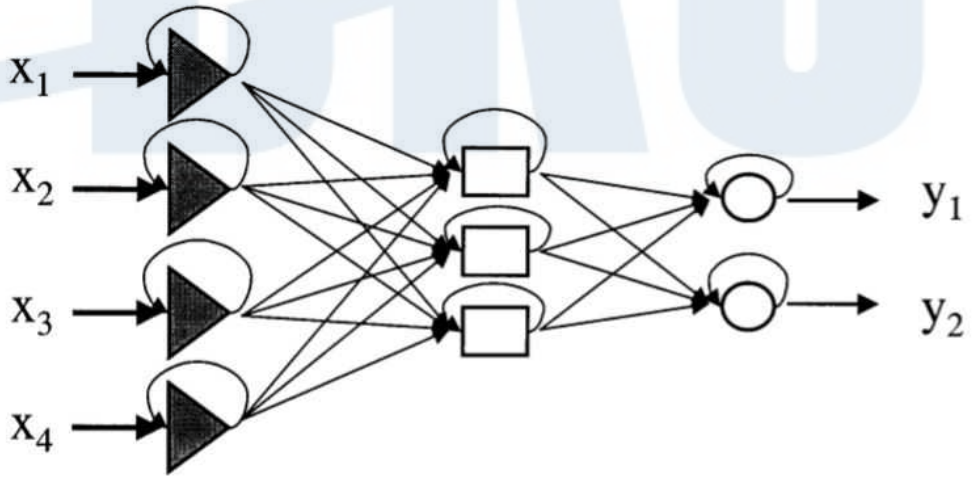
뉴런들이 서로 연결되는 방식은 인공신경망의 작동에 상당히 큰 영향을 미친다. 실제 뉴런처럼 인공 뉴런은 흥분성 또는 억제성 입력을 받을 수 있다. 흥분성 입력 값은 다음 뉴런의 메커니즘(mechanism)을 더하는 반면 억제 입력은 메커니즘의 감소를 유발한다. 뉴런은 또한 같은 층(layer)의 다른 뉴런들을 억제할 수도 있다. 이것을 '측방향 억제'라고 한다. 인공신경망은 가장 높은 확률을 선택하고 그 외에는 모두 억제시키는데, 이런 개념을 경쟁(competition)이라고도 한다(Agatonovic, Beresford, 1999).

피드백(feedback)은 한 층의 출력이 바로 이전 층이나 동일한 층의 입력 값으로 되돌아가는 또 다른 유형의 연결이다. 인공신경망의 피드백 연결 존재 유무에 따라 이를 두 가지 유형의 아키텍처(architecture)로 구분할 수 있다. 먼저, Feedforward 아키텍처는 [그림 14]와 같이 출력에서 입력 뉴런으로 되돌아가는 연결이 없으므로 이전 출력 값의 기록을 유지하지 않는다.



[그림 14] Feedforward network

Feedback 아키텍처는 출력 뉴런에서 입력 뉴런까지 서로 연결되어 있다. 훈련 시 오류를 최소화하려고 할 때, 각 뉴런은 [그림 15]와 같이 추가 자유도를 허용하는 입력 값으로서 하나의 추가 가중치를 가진다. 이러한 인공신경망은 입력 신호뿐만 아니라 네트워크의 이전 상태에 다음 상태가 의존할 수 있도록 이전 상태의 메모리를 유지한다(Zurada, 1992).



[그림 15] Feedback network

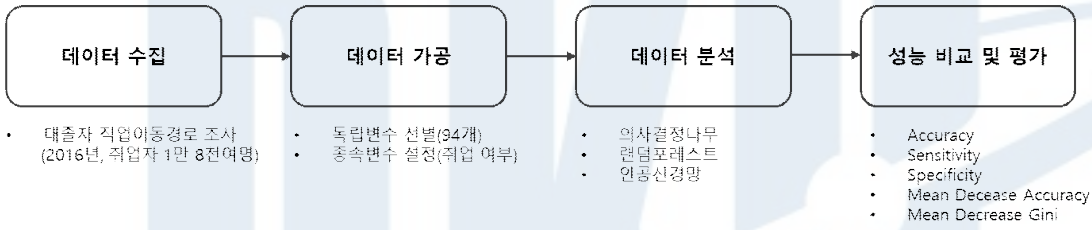
2.3.3 학습 규칙(Learning Rule)

ANN을 학습시킬 수 있는 다양한 학습 규칙이 있지만, 그 중 가장 자주 사용되는 것은 델타 규칙(Delta rule) 또는 역전파(Back-propagation) 규칙이다. 신경망은 가중치를 반복적으로 조정하여 입력 데이터 셋을 연결하도록 훈련을 받는다. 가중치가 부여된 링크(link)의 사용은 ANN의 인식 능력에 있어 필수적이다. 입력 값으로부터의 정보는 뉴런 사이의 가중치를 최적화하기 위해 네트워크를 통해서 앞으로 전달된다. 가중치의 최적화는 훈련이나 학습 단계에서 오차의 역전파를 통해 이루어진다. ANN은 훈련 데이터 셋의 입력 및 출력 값을 읽고 가중치가 부여된 링크의 값을 변경하여 예측 값과 목표 값의 차이를 줄인다. 예측 오류는 네트워크가 지정된 정확도 수준에 도달할 때까지 많은 훈련을 반복하며 최소화된다. 그러나 네트워크가 너무 오랫동안 훈련하도록 방치되면, 네트워크가 훈련 데이터에 과적합 될 수 있다 (Agatonovic, Beresford, 1999).

Ⅲ. 연구 설계

1. 연구 모형

[그림 16]은 본 연구의 흐름을 도식화 한 것이다. 본 연구는 먼저 대졸자 직업이동경로 조사 데이터를 취득한 후 독립변수를 선별하고 종속변수를 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.



[그림 16] 연구 모형도

2. 분석 데이터

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 조사의 시작연도인 2006년에 실시된 2005GOMS는 2004년 8월 및 2005년 2월에 2~3년제 대학 이상의 졸업자 총 502,764명의 모집단 중 약 5%에 해당하는 25,000여 명을 대상으로 1차 년도 조사를 실시했다. 1차 년도에 조사를 실시한 결과 총 26,544명의 표본 구축을 완료했으며, 2008년에는 3차 년도의 조사를 마치며 2005년 졸업자 조사를 종료했다.

2007년 졸업자를 대상으로 한 2007GOMS 부터는 졸업 다음 해에 1회 조사 이후 2년이 지난 뒤 1회에 한하며 추적 조사하는 단기 패널 조사로 2007년에 조사 설계를 변경했다. 그 후 2013년에 추가 조사 설계 변경을 통해 2011년 졸업자 조사부터는 횡단면 조사를 실시하고 있

다.

고학력화로 인해 청년층의 실업 문제가 지속적으로 심화되면서 학교를 졸업하고 노동시장으로의 이동 경로 현황을 분석하고 이를 지원하기 위한 다양한 방면의 정책적 수요가 증가하는 추세에서, GOMS의 조사 목적은 대학을 졸업한 청년층의 경력 개발과 직장 및 직업 이동 경로를 조사해 교육과 노동 시장 간의 인력 미스매칭(miss matching)을 완화하기 위한 정책 수립 과정에서 기초 자료로 활용되는 것에 있다.

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, 어학 연수 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. 독립 변수 선별

최필선, 민인식(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 기법들과 달리 기계학습 기법은 독립 변수의 수가 많더라도 자유도 감소의 문제를 크게 야기하지 않기 때문에 가능한 다양한 변수들을 포함시켜 취업 여부를 예측하는 데 중요하게 작용하는 변수가 어떤 것들인지 알아보고자 한다. 본 연구에서는 최필선, 민인식(2018)이 도출한 독립 변수 96개에서 삭제되거나 추가된 변수들을 조정해 총 94개의 변수[5]를 사용한다. 아래의 [표 2]는 연구 모델에 사용된 독립 변수를 범주별로 나타내고 있다.

[표 2] 범주 별 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

5) 변수 중 JPT · JLPT · HSK · CPT · BCT · LANG_etc(기타 외국어 시험) 등 6개의 제 2외국어 응시여부 변수들은 결측치가 약 90%에 달하기 때문에 변수에서 제외했다.

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

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 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 전형 방법 · 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 · 기업취업설명회 · 기타 프로그램 등의 대학 재학 중 취업지원 프로그램 참여 여부, 기업체 직무적성검사 · 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 교육훈련 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 · 회사규모 · 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.

4. 종속 변수 선정

본 연구의 목적은 대졸자의 구직에 영향을 미치는 다양한 요인들을 통해 대졸자들의 취업 여부를 예측하는 것이다. 따라서 종속 변수는 취업 여부(YES = 취업 / NO = 미취업)의 이항 변수로 설정했다. 원래 조사 자료에서는 취업 여부가 현재 직장의 종사상 지위에 따라 상용근로자·임시근로자·일용근로자·고용원이 있는 자영업자·고용원이 없는 자영업자·무급가족종사자[6]·미취업 등 총 7개로 구분되어 있다. 따라서 임금을 받지 않는 무급가족 종사자는 미취업자로 간주하고 종속변수를 재구성했다.

[표 3] 종속변수(취업/미취업) 빈도표

취업여부	현재 직장의 종사상 지위	인원(명)		비율(%)	
취업 (YES)	상용근로자	10,376	13,327	57%	73.2%
	임시근로자	2,243		12.3%	
	일용근로자	160		0.9%	
	고용원이 있는 자영업자	205		1.1%	
	고용원이 없는 자영업자	343		1.9%	
미취업 (NO)	무급가족종사자	33	4,872	0.2%	26.8%
	미취업	4,839		26.6%	
	합계	**18,199**	**18,199**	**100.0%**	**100.0%**

[표 3]과 같은 종속 변수의 비율로 기계학습을 실행하면 데이터의 계급 불균형이 발생하게 된다. 다양한 종류의 분류 문제에서는 스팸 메일 여부, 기업의 파산 여부 등 데이터의 균형이 맞지 않는 경우가 있다. 특히, 두 집단의 비율이 아주 크게 차이가 나는 경우에는 분류 분석을 통해 두 집단을 정확히 분류하기가 어렵다(김동아 등, 2015).

이런 불균형을 해소하기 위한 방법으로는 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 등(2002)에 의해 제안되었으며, 단순히 원래 데이터를 복제하는 것이 아니라 기존의 소수 계급의 데이터들 사이에서 추정할 수 있는 소수의 인위적인 데이터를 추가하는 방법이다. 이 기법은 먼저 소수 계급의 데이터들에서 k 근접 이웃(k nearest neighbors)을 찾아낸다.

6) 무급가족종사자 : 자영업자의 직계 가족, 친인척(동일 가구가 아니어도 가능)으로서 무임금으로 사업체 정규 근로 시간의 1/3 이상을 종사하는 사람을 일컫는 말이다. 통계표준용어

이러한 인위적인 데이터들은 원하는 오버 샘플링의 양에 따라 일부 또는 모든 방향으로 생성될 수 있다(Hulse 등, 2007).

본 연구에서는 RUS와 ROS의 단점을 보완할 수 있는 SMOTE 기법을 활용해 기존의 73(취업) : 27(미취업)의 비율을 55(취업) : 45(미취업)로 조정하여 데이터 불균형 문제를 해소했다.

Ⅳ. 분석 결과

1. 결과

1.1 예측 및 평가 지표

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 4] 정확도, 민감도, 특이도에 대한 용어 정의표

진단 테스트 결과	실제 질병의 유무		
	질병 있음(True)	질병 없음(False)	행 합계
질병 있음	TP(True Positive)	FP(False Positive)	TP+FP
질병 없음	FN(False Negative)	TN(True Negative)	FN+TN
열 합계	TP+FN	FP+TN	N = TP+TN+FP+FN

[표 4]에서 만약 질병이 환자에게 존재하는 것으로 밝혀질 때, 진단 검사의 결과 또한 질병이 존재한다고 하는 경우를 TP라고 한다. 이와 유사하게, 만약 질병이 환자에게서 없을 때, 검사에서 질병이 없다는 진단이 나오는 경우를 TN이라고 한다. 반면에 진단 검사에서 실제로 그런 질병이 없는 환자에게 질병이 존재한다고 하는 경우에는 이를 FP이다. 반대로 진단 검사의 결과 질병이 확실히 없는 것으로 나타난 환자가 질병을 가지고 있는 경우를 FN이라고 한다. 이를 기반으로 정확도, 민감도, 그리고 특이도는 아래와 같이 정의할 수 있다.

$$\text{민감도(Sensitivity)} = TP/(TP+FN) \quad (9)$$

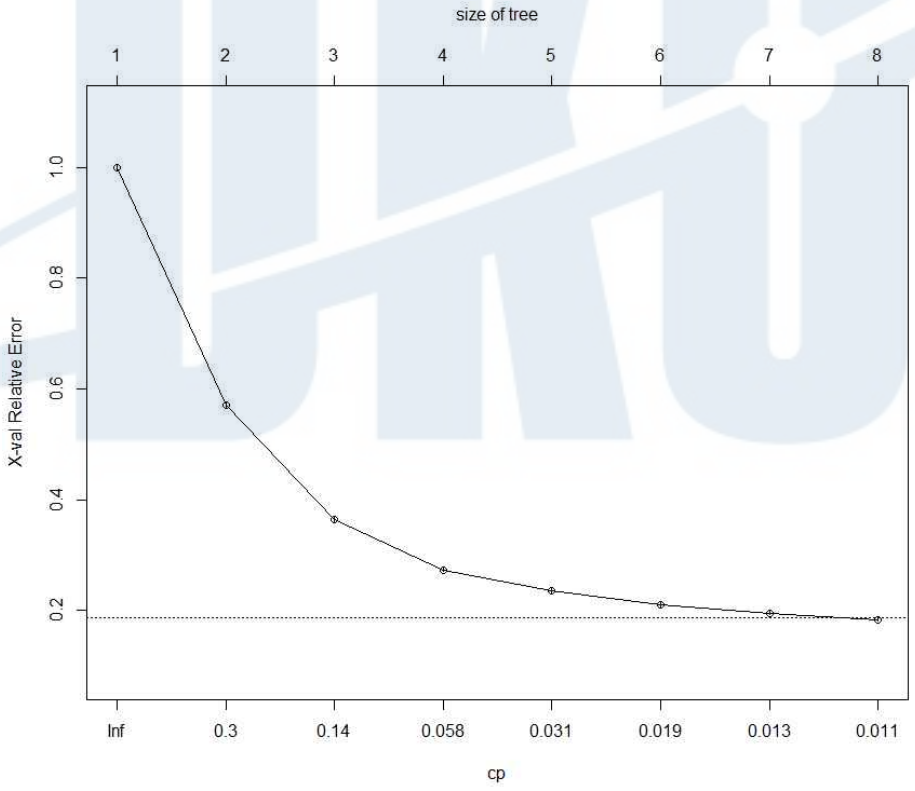
$$\text{특이도(Specificity)} = TN/(TN+FP) \quad (10)$$

$$\text{정확도(Accuracy)} = (TN+TP)/(TN+TP+FN+FP) \quad (11)$$

위의 공식에서 제시하는 바와 같이 민감도는 진단 테스트에 의해 정확하게 식별되는 실제 양성의 비율이다. 민감도는 진단을 수행한 테스트가 질병을 발견하는데 얼마나 좋은 테스트인지를 나타낸다. 특이도는 진단 테스트에 의해 정확하게 식별되는 실제 음성의 비율이다. 특이도는 진단을 수행한 테스트가 정상 상태를 식별하는데 얼마나 좋은 테스트인지를 보여준다. 정확도는 모집단에서의 TP와 TN의 합의 비율이다. 따라서 정확도는 특정 조건에서 진단 테스트의 진실성을 측정한다(Zhu 등, 2010).

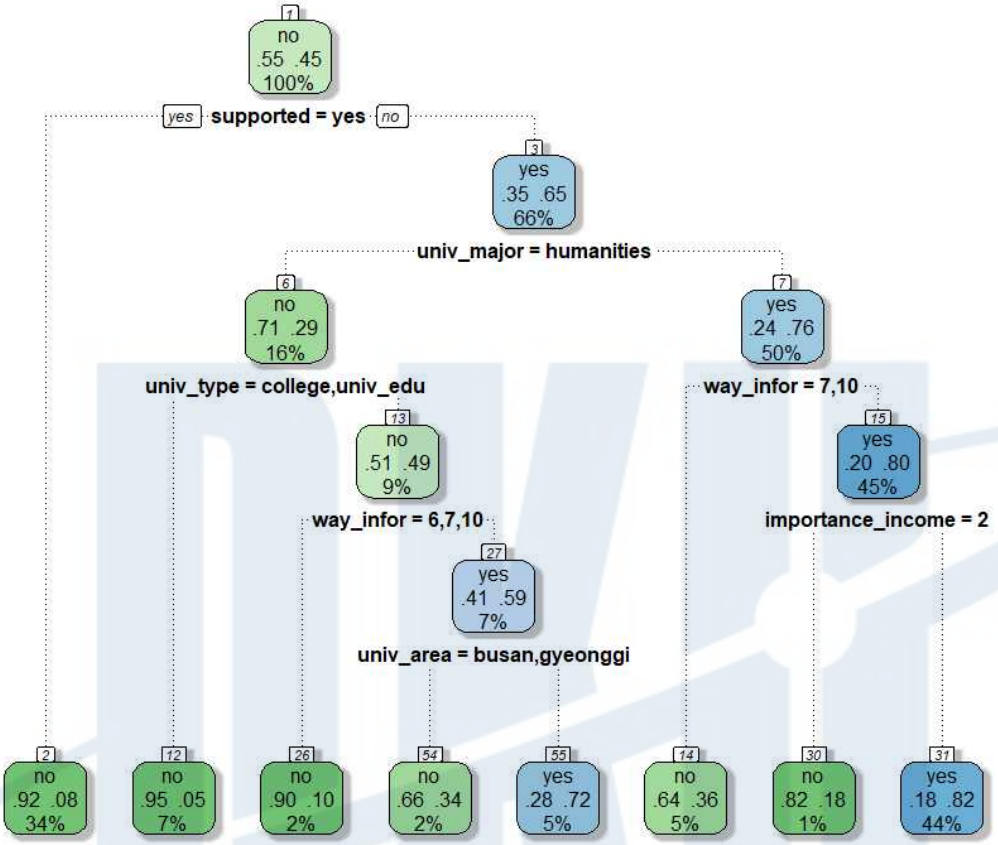
1.2 의사결정나무 모델

본 연구에서는 R의 'rpart' 함수를 사용해 의사결정나무를 생성했다. 'rpart' 함수는 의사결정나무의 여러 기법 중 CART를 사용한다. [그림 17]은 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 17] 의사결정나무 가지 수에 따른 에러율 변화

아래의 [그림 18]은 가치치기를 완료한 실험 1의 최종 의사결정나무 모델을 'rattle' 패키지의 'fancyRpartPlot' 함수로 시각화한 것이다.



[그림18] 가지치기를 완료한 최종 의사결정나무 모델(실험1)

[표 5] 의사결정나무 모델 예측 성능

구분	정확도	민감도	특이도
실험1	0.8444	0.8325	0.8588
실험2	0.8461	0.8313	0.8638
실험3	0.8453	0.8318	0.8615
실험4	0.8481	0.8365	0.8621
실험5	0.8437	0.8283	0.8622
실험6	0.8482	0.8402	0.8578
실험7	0.8475	0.8358	0.8617
실험8	0.8416	0.8173	0.8707
실험9	0.8477	0.8276	0.8718
실험10	0.8486	0.8328	0.8675
평균	**0.8461**	**0.8314**	**0.8638**
표준편차	**0.0022**	**0.0059**	**0.0045**

[표 5]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.

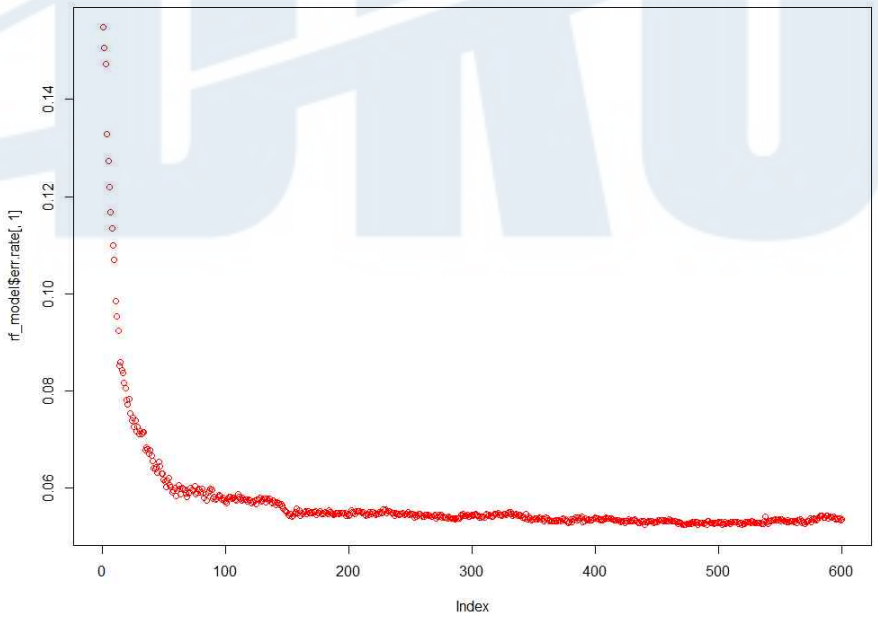
1.3 랜덤포레스트 모델

랜덤포레스트 모델은 R의 ‘randomForest’ 함수를 사용해 생성했다. [표 6]은 랜덤포레스트 모델을 사용해 대졸자들의 취업 여부를 예측한 결과이다.

[표 6] 랜덤포레스트 모델 예측 성능

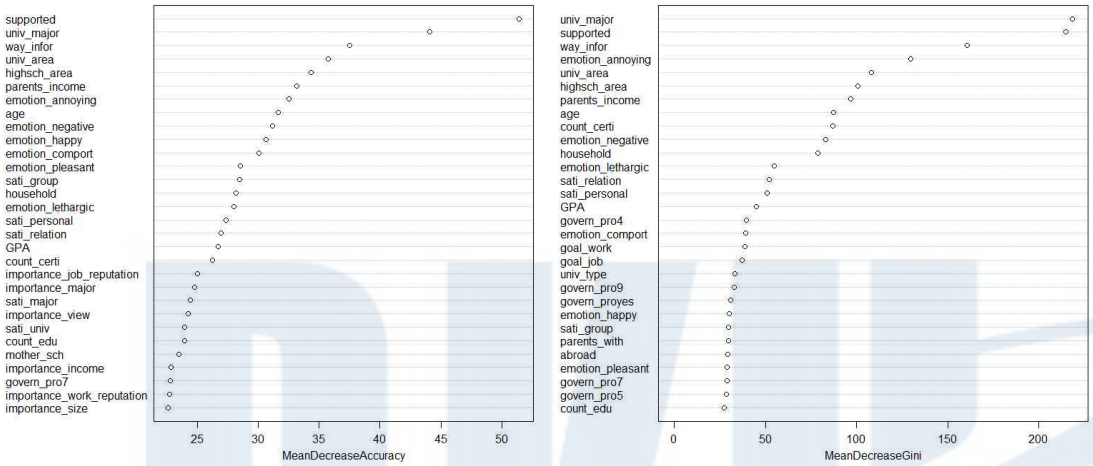
구분	정확도	민감도	특이도
실험1	0.9496	0.9091	0.9852
실험2	0.9418	0.8956	0.9812
실험3	0.9490	0.9040	0.9857
실험4	0.9503	0.9194	0.9778
실험5	0.9478	0.9001	0.9871
실험6	0.9550	0.9158	0.9902
실험7	0.9461	0.9104	0.9747
실험8	0.9456	0.9008	0.9827
실험9	0.9462	0.8966	0.9860
실험10	0.9417	0.8944	0.9799
평균	**0.9473**	**0.9046**	**0.9831**
표준편차	**0.0038**	**0.0083**	**0.0045**

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 19] 의사결정나무 개수에 따른 에러율의 변화

[그림 19]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.



[그림 20] 랜덤포레스트 모델 변수들의 중요성 지수(실험1)

위에서 언급했듯이 랜덤포레스트는 결과를 예측하는데 결정적인 역할을 하는 중요 변수를 파악할 수 있다는 장점이 있다. [그림 20]은 실험 1의 랜덤포레스트 모델을 구성하는 변수들의 중요성 지수를 나타낸 것이다. MDG(MeanDecreaseGini)를 통해 확인한 결과, supported(가족에게 경제적 지원을 받는 여부), univ_major(졸업한 전공 계열), way_info(일자리 정보를 얻는 루트), emotion_annoying(지난 한 달간 감정 - 짜증나는), univ_area(졸업 대학의 소재지) 순으로 중요도가 산출되었다.

1.4 인공신경망 모델

인공신경망 모델은 R의 'nnet' 함수를 사용해 생성했다. [표 7]은 은닉층(hidden layer)이 1개인 인공신경망 모델을 사용해 대졸자들의 취업 여부를 예측한 결과이다.

[표 7] 인공신경망 모델 예측 성능

구분	정확도	민감도	특이도
실험1	0.8363	0.6617	0.9893
실험2	0.8946	0.8285	0.9509
실험3	0.9154	0.8351	0.9810
실험4	0.8436	0.8072	0.8759
실험5	0.8921	0.7640	0.9976
실험6	0.8205	0.6767	0.9492
실험7	0.6671	0.8760	0.4996
실험8	0.9198	0.8578	0.9709
실험9	0.9187	0.8348	0.9860
실험10	0.8903	0.8512	0.9218
평균	**0.8598**	**0.7993**	**0.9122**
표준편차	**0.0727**	**0.0712**	**0.1419**

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.

Ⅴ. 결론

본 연구에서는 기계학습 기법 중 의사결정나무, 랜덤포레스트, 인공신경망을 이용해 대졸자들의 취업 여부를 예측하는 모델을 생성하고 예측 결과를 통해 각 모델 간의 성능을 비교하였다. [표 8]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 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.

[표 8] 각 예측 모델들 간의 성능 비교

구 분	의사결정나무		랜덤포레스트		인공신경망	
	평균	표준편차	평균	표준편차	평균	표준편차
정확도	0.8461	**0.0022**	**0.9473**	0.0038	0.8598	0.7027
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

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 노드

들로는 전공 계열, 대학의 유형(전문대학, 4년제 대학, 교육대), 일자리 정보를 얻는 경로, 직장을 선택할 때 근로 소득의 중요도, 졸업 대학의 소재지 등이 선택되었다.

[표 9] 랜덤포레스트 모델의 변수 중요도

순위	구분	MDA	순위	구분	MDG
1	**supported**	**0.0501**	**1**	**supported**	**220.0185**
2	**univ_major**	**0.0412**	**2**	**univ_major**	**210.0841**
3	**emotion_annoying**	**0.0267**	**3**	**way_infor**	**152.2371**
4	**way_infor**	**0.0264**	**4**	**emotion_annoying**	**126.5622**
5	**goal_work**	**0.0234**	**5**	**univ_area**	**112.0212**
6	univ_area	0.0228	6	highsch_area	106.3344
7	goal_job	0.0227	7	parents_income	97.9603
8	highsch_area	0.0216	8	count_certi	91.3499
9	emotion_negative	0.0206	9	emotion_negative	88.9131
10	household	0.0201	10	age	84.7870
11	parents_income	0.0196	11	household	70.4667
12	count_certi	0.0196	12	sati_personal	55.0032
13	age	0.0174	13	emotion_lethargic	54.8269
14	sati_relation	0.0141	14	sati_relation	52.7240
15	sati_personal	0.0136	15	GPA	44.9612
16	emotion_lethargic	0.0127	16	govern_pro4	39.7371
17	univ_type	0.0119	17	emotion_comport	39.6415
18	govern_proyes	0.0118	18	goal_work	39.4536
19	parents_with	0.0118	19	goal_job	38.2272
20	govern_pro7	0.0110	20	univ_type	35.7860

두 번째, [표 9]는 랜덤포레스트 모델을 활용한 10번의 실험에서 도출한 변수 중요도의 평균을 산출한 결과이다. 의사결정나무에서 특정 변수를 제거했을 때 감소하는 정확도 차이의 평균값인 MDA(Mean Decrease Accuracy)에서 상위 10개의 중요 변수는 가족의 경제적 지원을 받는지 여부, 전공 계열, 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 인구 통계학적 특성이 2개(가구주 여부, 고등학교 거주 지역), 가족 및 부모의 특성이 1개(가족의 경제적 지원을 받는지 여부), 졸업 대학의 특성이 2개(전공 계열, 졸업 대학의 소재지), 심리적 요인이 4개(한 달간 짜증 나는 감정을 느낀 정도, 한 달간 부정적인 감정을 느낀 정도), 일자리 정보

를 얻는 경로 1개로 나누어진다. 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 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.

세 번째, 인공신경망 모델은 정확도 측면에서 의사결정나무보다는 높은 성능을 보였지만 랜덤포레스트 모델보다는 성능이 낮았다. 또한 트리 기반의 다른 두 방법보다 훈련 데이터의 선정에 따라 결과가 민감하게 반응하는 것으로 나타났다. 또한 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.

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 채창균, 김태기(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 영향

을 미치는 요소로 나타나는 점을 미루어보아 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.

본 연구에서는 취업자에 상용 근로자뿐만 아니라 임시 근로자나 일용 근로자 등 비정규직도 취업자에 포함시켜 대졸자들의 취업의 질을 구분하여 파악하는 것에 한계가 있었다. 향후 연구에서 종속 변수를 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

강희용 (2016) 빅데이터 적용 사례 및 활용 전략. Magazine of the SAREK, 45(1): 32-33.

길혜지, 최윤미 (2014) 대졸자의 고용형태 결정요인 분석 연구. The Journal of Vocational Education Research, 33(6): 13-19.

김동아, 강수연, 송종우 (2015) Classification Analysis for Unbalanced Data. The Korean Journal of Applied Statistics, 28(3): 495.

김수혜 (2018) 대졸 청년층의 취업준비 활동이 노동시장 진입에 미치는 영향 : 정규직 취업 여부와 시점을 중심으로. 교육문화연구, 24: 313-318.

김정경 (2016) 국내외 빅데이터 동향 및 성공사례. Industrial Engineering Magazine, 23(1): 48-49.

김재생 (2014) 빅데이터 분석 기술과 활용 사례. 한국콘텐츠학회지, 12(1): 18-19.

노경란, 허선주 (2015) 대졸 청년층의 취업목표 달성 영향요인 분석. The Journal of Vocational Education Research, 1(22): 10-14.

이유재, 이신형, 이종세 (2014) KB국민카드의 마케팅 활동과 빅데이터 활용. Korea Business Review, 18(1): 162-163.

정미나, 임영식 (2010) 대졸 청년층의 노동시장 진입관련 변인에 대한 경로분석. The Journal of Career Education Research, 23(2): 143-146.

채창균, 김태기 (2009) 대졸 청년층의 취업 성과 결정 요인 분석. The Journal of Vocational Education Research, 28(2): 96-99.

홍기석 (2018) 청년실업의 결정요인 연구, 한국경제의 분석. 24(2): 91-93.

최종후, 서두성 (1999) 데이터마이닝 의사결정나무의 응용, 통계분석연구. 4(1): 62, 63-67.

최필선, 민인식 (2018) 머신러닝 기법을 이용한 대졸자 취업예측 모형. 직업능력개발연구. 21(1): 32-34, 38.

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

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

Anyanwu. M., Shiva. S. (2009) Comparative Analysis of Serial Decision Tree Classification Algorithms. International Journal of Computer Science and Security, 3(3): 233.

Breiman, L., Friedman, J., Olshen, L. and Stone, J. (1984) Claasification and Regression trees. CHAPMAN & HALL/CRC, USA. pp. 55-58.

Breiman, L. (1999) RANDOM FORESTS: 2-3.

Breiman, L. (2001) RANDOM FORESTS: 7-8.

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

Hssina, B., Merbouha, A., Ezzikouri, H. and Erritali, M. (2014) 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: 13-18.

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

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

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

Mingers, J. (1989) An empirical comparison of pruning methods for decision tree induction. Machine Learning, 4: 241-242.

Pal, M. (2005) Random forest classifier for remote sensing classification. International Journal of Remote Sensing, 26(1): 2-3.

Podgorelec, V., Kokol, P., Stiglic, B. and Rozman, I. (2002) Decision trees: an overview and their use in medicine. Journal of Medical Systems, Kluwer Academic/ Plenum Press, 26(5): 9-10.

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

Zhao, Y., Zhang, Y. (2007) Comparison of decision tree methods for finding active objects. Advances in Space Research, 41(12): 3-4.

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

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

(Abstract)

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

Donghun Lee

Department of Data Knowledge Service Engineering

Graduate School

Dankook University

Advisor : Professor Taehyung Kim

Youth unemployment is continually becoming a social problem in Korea. In this study, By using the decision tree, random forest, and artificial neural network, we generate models to predict college graduates employment, and compare the performance among each model through forecast results.

The results showed that the use of random forest models had the highest performance, and that the artificial neural network models had a close performance gap over the decision tree models. In the decision tree model, whether to receive financial support from family members, the

major line, the path to obtaining information on type jobs at universities, the importance of working income when choosing a job, and the location of college and university were selected as the main nodes. In the Random Forest model, important variables was selected, including whether they receive financial support from family members with important variables, their major line of study, the route of obtaining job information, the degree of irritating feelings for a month, and the location of graduation colleges.

By category of independent variables, two demographic characteristics, one family and parent characteristics, two graduation characteristics, four psychological factors, and one path to obtaining job information are included in the important variables.

Keywords : Youth Unemployment, Mechanical Learning, Decision Tree, Random Forest, Artificial Neural Network(ANN)