Journal of the Korea Industrial Information Systems Research Vol. 23 No. 6, Dec. 2018 : 67-77
http://dx.doi.org./10.9723/jksiis.2018.23.6.067 ISSN:1229-3741

연관관계 규칙을 이용한 학생 유지율 관리 방안 연구

(A Study on Management of Student Retention Rate Using Association Rule Mining)

김 종 만[1], 이 동 철[2]*
(Kim Jong-Man and Lee Dong-Cheol)

요 약 최근 학령인구 감소에 따라 많은 문제점들이 나타나고 있다. 우리나라는 인구대비 가장 많은 대학을 보유하고 있기 때문에 각 대학의 생존에 필요한 최소한의 학생 유지율 관리가 점점 더 중요해 지고 있다. 따라서 본 연구는 계속되는 학력인구의 감소에 따라 각 대학들이 생존 방안으로 학생 유지율의 적절한 관리 방안을 모색한다. 이를 위하여 특정 대학에 입학한 학생들을 대상으로 성별, 출신고, 출신지역, 성적, 졸업여부 등의 데이타를 분석하여, 학생들이 입학에서 졸업에 이르기까지 지속적으로 유지될 수 있는 학생 유지율을 관리하기 위한 기본적인 방향이 어떤 것인지 알아본다. 또한, 최적의 입력 변수를 파악하고, 최적의 입력 파라메터를 기초로 apriori 알고리즘을 이용하여 연관 분석을 실행하여 유지율 관리에 가장 적합한 자료를 수집할 수 있도록 한다. 이를 바탕으로 각 대학들이 학생들을 모집하고 유지하는데 도움이 되도록 가장 효율이 높은 딥러닝(Deep Learning) 모듈을 개발하기 위한 기초 자료로 만들고자 한다.
의사결정트리를 활용하여 졸업여부를 측정한 결과는 딥러닝의 정확도 보다 낮은 75%로 나타났다. 의사결정트리에서 졸업여부를 결정하는 요인은 일반고를 졸업하고, 도시지역에 거주하면서 여성이면서 성적이 높은 학생들이 졸업확율이 높은 것으로 나타났으며 결과적으로 의사결정트리 보다는 개발된 딥러닝들이 더 효율적으로 학생들의 졸업여부를 평가할 수 있는 모델로 나타났다.

핵심주제어 : 학령인구 감소, 학생 유지율, 연관관계, 딥 러닝

Abstract Currently, there are many problems due to the decline in school-age population. Moreover, Korea has the largest number of universities compared to the population, and the university enrollment rate is also the highest in the world. As a result, the minimum student retention rate required for the survival of each university is becoming increasingly important. The purpose of this study was to examine the effects of reducing the number of graduates of

* Corresponding Author : dchlee@jejunu.ac.kr
+ 이 논문은 2018학년도 제주대학교 교원성과지원사업에 의하여 연구되었음.
Manuscript received October 17, 2018 / revised November 23, 2018 / accepted December 4, 2018
1) 제주대학교 경영정보학과, 제1저자
2) 제주대학교 경영정보학과, 교신저자

education and the social climate that prioritizes employment. And to determine what the basic direction is for students to manage the student retention rate, which can be maintained from admission to graduation, to determine the optimal input variables, Based on the input parameters, we will make associative analysis using apriori algorithm to collect training data that is most suitable for maintenance rate management and make base data for development of the most efficient Deep Learning module based on it.
The accuracy of Deep Learning was 75%, which is a measure of graduation using decision trees. In decision tree, factors that determine whether to graduate are graduated from general high school and students who are female and high in residence in urban area have high probability of graduation.
As a result, the Deep Learning module developed rather than the decision tree was identified as a model for evaluating the graduation of students more efficiently.

Key Words : School Age Population Decline, Student Retention Rate, Association Rule, Deep Learning

1. 서 론

우리나라는 인구대비 가장 많은 대학을 보유하고 있는 국가 중의 하나로서 대학 진학률 또한 세계에서 가장 높은 편이다. OECD국가와 우리나라 대학 교육 이수율을 비교해 볼 때 OECD 평균이 42%인 반면 우리나라 이수율은 69%에 육박하고 있다. 그러나 현재 국내에서는 출생율 저하에 따른 학령인구의 감소와 학업중단율의 증가로 인하여 많은 문제점들이 생겨나고 있다(교육부, 2015년 OECD 교육지표, 2015).

선행 연구에서는 설문 조사를 통하여 주로 학생들의 심리적 요인에 대한 분석 결과, 학업적 능력, 심리적 특성, 진로와 적성, 대인관계, 등과 같은 개인적 특성이 학생들의 학업 이탈에 직접적인 요인으로 제시되었다[1].

그러나 본 연구는 기존의 설문지 위주의 데이터 수집이 아닌, 현재 존재하는 실질적인 요인들의 데이터를 분석한 연구로서 Kim[2]의 학위논문을 바탕으로 작성되었다. 즉, 3~5년 간 축적된 학생의 출신지역, 출신 고등학교, 성별, 나이, 학기별 성적자료 등의 정량적 수치를 기반으로 자료를 기초로 연관분석을 진행함으로써 각 변수들 간의 연관관계를 확인하고자 한다.

예를 들면 한 대학에 입학한 학생들이 입학 당시 출신 지역 및 출신 고등학교, 성별, 나이에 따라 대학에서 휴학이나, 자퇴 또는 졸업에 대한 연관성을 분석한다. 이를 통하여 학생의 학업 유지에 가장 영향을 미치는 변수를 찾아내어 학업 유지에 가장 적합한 변수들로 구성된 훈련 데이터를 수집하고, 이러한 데이터를 기초로 학생 유지 여부를 파악하는 딥러닝(Deep Learning) 모듈을 개발한다. 그 이후 개발된 딥러닝 모듈과 의사결정나무 모델을 활용한 이탈율 예측 정확도를 비교하여 개발된 딥러닝 모듈이 기존의 의사결정나무 모델보다 더 좋은 정확도를 나타낼 수 있는지 확인하고자 한다.

구체적인 연구문제는 첫째, 기계학습의 한 분야인 딥러닝 모듈을 활용하여 학생들이 중도탈락 없이 학기를 마칠 수 있는 확률을 측정할 수 있지 여부를 판단한다. 둘째, 측정 가능하다면 실제 학생 데이터에 의한 예측 정확도를 확인한다. 셋째, 개발된 모듈을 활용하여 통계적 접근 없이 다양한 형태의 학생정보를 활용 하여 실제적인 예측 모델로 사용한다. 넷째, 의사결정나무 모델과 딥러닝 모델의 예측 정확도를 비교한다.

위의 연구문제 해결을 통하여 특정 학생이 입학 이후에 학업이 중단됨이 없이 학교를 졸업할 수 있는 확률이 어느 정도인지 파악 할 수 있게 됨으로써, 학업중단 확률이 높은 학생들을 대상으로 효과적인 지도를 할 수 있는 정보를 얻을 수 있을 것이다.

기존의 추론위주의 연구와는 달리 경험을 통해 쌓인 데이터로부터 귀납적 판단을 하는 인공지능 학습방법을 이용하여 데이터를 기반으로 학생들이 학교를 중도에 포기할 가능성이 얼마나 되는지 판단할 수 있게 될 것이다.

또한 딥러닝의 가장 큰 장점인 시간을 거듭 할수록 더욱 더 많은 경험적 데이터가 축적하여 모듈을 학습시키고, 스스로 학습하는 인공지능으로 발전한다면 더욱더 효과적인 딥러닝 모듈로 발전해 나갈 수 있을 것이다.

2. 선행연구

2.1 학업유지율

학업 유지율은 전체 입학생 중에서 학업 수행을 모두 마치고 졸업을 하게 되는 학생의 비율을 말하며, 학업 중단율의 상대되는 개념이다.

학업 중단율을 줄이기 위한 노력은 그동안 중고등 학생들을 대상으로 하여 많이 연구가 되어 있으나, 고등교육기관을 대상으로 한 연구는 미흡한 실정이다. 그러나 최근의 교육부 통계에 따르면, 2015학년도 전체 고등교육기관의 학업 중단율은 7.5%로 전년대비 0.8% 증가하였다. 일반대학의 학업 중단율은 4.1%로 전년 대비 0.2% 증가하였고, 전문대학은 7.5%로 전년과 동일한 수준을 기록하였으나 높은 비율을 차지하고 있으며, 방송통신대학과 사이버 대학을 중심으로 한 기타 학교의 학업 중단율이 23.7%로 전년 대비 크게 증가(5.7% 증가)하여 전체 학업 중단율에 영향을 미친 것으로 파악된다.

학업중단율과 관련된 선행 연구는 대부분 고등학생들을 대상으로 많이 이루어지고 있다. 우리나라 고등학교의 정신건강이 학업성취도와 학업중단율에 미치는 영향에서는 학생들의 정신 건강이 학업 수행도나, 학교 이탈과 같은 교육결과와 어떤 관련성을 가지고 있는지에 대해 연구하여 낮은 학업 성취도와 잦은 결석 등 아동・청소년기에 보일 수 있는 정신건강 문제를 조기 발견할 수 있는 것에 대해 연구한 것이다[3].

또한 고등학교 학업 중단율 변화의 지역별, 학교유형별 현황 및 학교 관련 요인 탐색에서는 2010년까지 전국 고등학교의 학업중단율의 변화 실태를 파악하고, 이에 영향을 미치는 학교 특성 요인을 확인하는 연구를 진행 하였다[4]. 이는 전반적으로 학업 중단율이 증가하고 있는지를 확인하고, 지역별로, 학교 유형별로 학업 중단율 차이에 대한 원인 분석을 하는 연구 등이 있다.

Sim[5]은 자발적으로 학업을 중단하려는 의도를 가진 국립대학의 중도 이탈자들과의 개별 면담을 통한 대학 입학에서부터 학교를 중도 탈락을 결정할 때 까지 경험한 심리적, 환경적 요인들을 분석하였는데 첫째 지역여건 관련 요인, 학생의 자아실현 욕구, 셋째 일과 학습의 병행하는 세 가지 학업 이탈 모형을 제시하였다.

Lee[6]은 도시 학교에 제학중인 중학교 2학년부터 고등학교 1학년 사이에서 학생들이 학교 적응이 시간에 따라 어떻게 변화하는 지 학교적응을 예측하는 변수는 무엇인지를 잠재성장모형을 통하여 검증하였는데 도시지역의 학교에 다니는 청소년의 학교 적응이 증가하는 것으로 나타났고, 도시지역에서 다니는 청소년의 학교적응의 초기값을 측정하는 요인은 부모의 교육수준, 우울, 주의집중, 부모의 방임과 학대 , 수업수행 효과와 만족도, 학교폭력 가해행위와 공동체 의식 변수였고, 변화율을 예측하는 요인은 주의집중, 학업수행 효과와 만족도, 학교폭력 가해 행동과 공통체 의식 변수로 나타났다.

Cho and Han[7]은 부모와 학생의 배경이 학교적응에 미치는 영향을 분석하였는데 이 연구에서 학교적응 개념을 학습활동, 학교규칙, 교유관계, 교사관계로 구분하였으며 연구결과, 학교에 잘 적응하는 상위 20%집단은 학교적응에 정적 영향을 주는 요인이 높게 나타났으며, 하위 20% 집단은 부적 요인이 높은 것으로 나타났고, 부모의 방임적 양육태도는 자녀의 학교 적응과 부정적 상관관계가 있었고 학생의 학습습관 요인들은 학교적응에 정적인 관계가 높게 나타난다.

Lee and Kang[8]는 학생들의 학교생활 부적을 요인을 파악하기 위하여 학교 부적응 학생들을 대상으로 평균과 표준편차와 위계적 회귀분석을 통하여 결과를 분석하였다. 수업에 적응하지 못하는 행동은 첫째, 부적응 행동 경험과 교사의 기

대, 학업성적과 부모자녀 관계 요인이 영향을 미치는 것으로 나타났다. 둘째, 가정폭력 셋째, 학교 생활에 흥미가 떨어지는 학생일수록 학교생활에 어려움을 격고 있는 것으로 나타났다.

이와 같이 중 고등학생들을 대상으로 중도탈락과 같은 학업 중단율에 영향을 미치는 연구는 비교적 많이 진행 되어져 왔으나, 대학과 같은 고등교육기관을 대상으로 학업 중단율에 영향을 미치는 요인을 분석하는 연구에서는 많이 부족하다.

또한 학업 중단율의 영향을 연구하는 방법도 대부분 학생들의 심리적 요인 등을 대상으로 연구되어 있으며, 실제 학생들의 실제 정량적 데이터로 접근한 연구는 아직 없는 실정이다.

따라서, 본 연구에서는 기존 연구에서의 심리적인 요인에서 중도 탈락 관련 원인을 찾는 형태가 아닌 대학교 학생들을 대상으로 한 심리적 설문조사나, 사전 조사 없이 현재의 실질적인 학생의 정량적 데이터만을 활용하여 학생들의 학업 유지율, 즉, 중도 탈락 가능성을 예측하는데 활용 가능한 유용한 모델을 개발하고자 한다.

3. 연구설계

3.1 데이터의 구성

학생들의 실제 데이터를 Training data와 Testing Data로 하고, Training data 및 Testing Data의 구성은 출신학교(School), 성별(Sex), 연령(Age), 출신지역(Address), 통학시간(Travel time), 성적(G1, G2, G3)등으로 구분하고, 졸업여부(Grad)를 판정하기 위한 데이터로 구성하였다.

Table 1에서와 같이 출신학교는 일반고, 특성화고, 검정고시 그리고 대졸 및 외국인 학교로 구분하였으며 출신지역은 도시출신과 시골지역으로 구분하였고, 통학시간은 시간대별로 범위로 구분하였으며, 졸업여부는 정상 졸업과 자퇴, 장기 휴학상태, 제적 등의 경우로 구분하였다.

Table 1 Composition of Data

seq	Item	Description
1	School	Graduate School (Binary : GP, MS, SF, ET)
2	Sex	Student's Sex (binary : F, M)
3	Age	Student's age (numeric : 15 ~ 22)
4	Address	Student's Hometown (binary : U, C)
5	Traveltime	Travel Time (numeric : 1, 15, 30, 60)
6	G1	First period grade (numeric : 0 to 20)
7	G1	Second period grade (numeric : 0 to 20)
8	G3	Final period grade (numeric : 0 to 20)
9	Grad	Graduate status (binary : G, H, D, C P)

3.2 데이터의 특성

2011학년도부터 2014학년도까지의 A대학교 신입생 및 재학생 대상 8,000명을 대상으로 하여 이름, 학번 등의 개인정보 데이터를 제외하고, 학생들의 출신 지역, 출신 고등학교, 성별, 나이, 주소, 성적 등의 데이터를 활용하였다. 이 데이터는 학생이 대학 입학 시 기록한 것으로 성별, 나이, 출신고 등을 나타내는 변수들과 통학거리를 나타내는 변수들, 그리고 학기별 성적 데이터를 나타내는 변수로 구성되어 있으며 학생이 정상적으로 학교를 졸업 했을 경우 와 그렇지 않을 경우를 고려하여 데이터를 구성하였다.

본 연구에 사용된 데이터의 특성은 Table 2에서 제시된 바와 같이 성별로는 남자가 52%, 여자가 48%이며 출신학교는 일반 고등학교가 67% 특성화고가 24.6% 검정고시등 기타 고등학교가 8.4%의 비율로 나타났다. 정상졸업인 경우 62.6%이며 제적, 중퇴, 장기휴학등 졸업하지 않은 비율

이 37.4%이며, 이중 남성인 경우 정상 졸업이 44.7%, 중도탈락이 55.3%, 여성인 경우 졸업이 79%, 중도탈락이 21%로서 남성 보다는 여성이 중도탈락 없이 학교를 정상적으로 마치는 비율이 높았다.

출신학교, 성별, 지역, 연령 등의 구분자를 구성하여 각각의 변수에 대입한 후 연속형 값을 이산화하기 위해 정렬한 후 R의 data frame 함수를 사용하여 이산화형으로 나누기 위한 분할점을 구성한 후 이렇게 나뉘어진 데이터를 거래데이터(이산화형)로 변환한다.

- 연속된 변수의 값을 이산화하기 위해 정렬한다.
- 범주형을 이산화형으로 나누기 위한 분할점을 구성한다.
- 분할점으로 구성된 데이터를 이산화형(Transaction Data)로 변환한다.

Table 2 Main Characteristics of Data

Main characteristics		Frequency	Remarks
Sex	Male	52%	
	Female	48%	
School	General High School	67%	
	specialized vocational high schools	24.6%	
	Etc	8.4%	
Graduate	Graduated	62.6%	* Male - Graduated : 44.7% - Drop out : 55.3% * Female - Graduated : 79% - Drop out : 21.0%
	Drop out	37.4%	

3.3 연관분석

본 연구는 이론적 배경을 기반으로 실제적으로 데이터에서 포함하는 변수들이 실제 학생이 학업 중단에 영향을 미치는지에 대한 연관성을 분석하고, 주요 연구변수들이 실제 학업 중단, 즉 이탈율에 영향을 미친다면 이것을 바탕으로 좀 더 많은 데이터를 수집하여 실제 학생 개개인이 입학과 동시에 학기일정을 정상적으로 마칠 수 있는 확률값을 확보하기 위한 것이다.

학생들의 출신학교는 일반고등학교 출신일 경우 'GP', 취업을 목적으로 하는 특성화고 등은 'MS', 그리고 검정고시, 대졸출신, 외국고 출신 등은 'ET' 로 표시하였다. 성별은 남성일 경우 'M', 여성일 경우 'F'로 구분 하였으며 주소는 직할시를 포함한 시 지역일 경우 'U' 그 외 지역 출신일 경우 'C'로 표기 하였으며 성적은 평점이 3.3 이상인 경우 'high' 평점 3.3 미만인 경우 'low' 로 표기하였으며, 졸업여부는 정상적으로 졸업한 학생은 'G' 로 표기하고 제적, 휴학생태, 자퇴, 수료인 경우는 모두 'H' 로 표기한 후 출신고등학교, 출신지역, 성별, 성적 등이 졸업여부와 어떤 연관성이 있는지 분석 하였다.

Apriori 알고리즘은 구현이 간단하고 성능 또한 만족할 만한 수준을 보여주는 알고리즘으로 패턴 분석을 위해 자주 이용되는 알고리즘이다. 연관 관계를 계산하기 위해서 아이템들의 출현 빈도를 이용하여 계산하게 되는데 여기서는 간단히 아이템들이 동시에 출현하게 될 경우의 확률에 대해서 신뢰도 s라고 하고, 아이템 X가 출현할 때 또 다른 아이템 Y 역시 포함되어 있을 조건부 확률을 신뢰도 c라고 할 때 우선 도시, 출신고, 졸업의 연관관계를 살펴보면, 도시 출신일 때 졸업할 확률은 대략 66% 정도가 나오고, 전체 데이터에서 도시와, 졸업이 일어날 확률이 50%가 된다. 따라서 지지도는 50%, 신뢰도는 66.6%가 된다.

Apriori 알고리즘은 구현하기 쉽고, 이해하기 쉬우며 어느 정도 만족할 만한 결과를 주기 때문에 자주 쓰이고 있으나, 이 알고리즘은 후보 집합 생성 시 아이템의 개수가 많아지면 계산 복잡도가 급격히 증가하게 되는 단점이 있다.

Fig. 1 How to utilize data

3.4 데이터의 활용 방안

본 연구의 목적은 Fig. 1에서 보는 것과 같이 학생들이 입학한 이후에 중도에 학업을 중단한 확률인 학생들의 이탈율을 효과적으로 계산 할 수 있는 딥러닝 모듈을 개발하고 개발된 모듈에 효과적으로 학습할 수 있는 기초 데이터를 확보하는 것에 있다.

딥러닝 모듈을 개발한 후 개발된 모듈을 학사정보와 연계하여 학생이 입학한 경우 출신 고등학교, 연령, 출신지역 등의 데이터를 모듈에 대입하면 통계데이터의 확보 없이 이탈 확률이 높은 학생들을 선별하여 효과적인 지도학습이 가능하다.

3.5 의사결정트리

연관 분석을 통한 데이터를 이용한 의사결정모델을 평가하기 위하여 R의 caret패키지를 사용해서 데이터를 70% 의 train set과 30%의 test set으로 구분한 후 의사결정트리의 성능을 평가하였다. train set으로 의사결정트리 모델을 구성하고, test set을 구성된 모델이 입력하여 졸업여부를 얼마나 잘 예측하는지 확인하였다.

의사결정트리 분석은 회귀분석이나, 랜덤포레스트 등의 알고리즘에 비해 직관적으로 이해할 수 있으며, 설명 또한 쉽게 할 수 있는 장점이 있는 지도학습 방법이다. 분석방법에는 party 패키지를 사용하여 분석하는 것이 가장 높은 정확도를 나타내었다.

3.6 딥러닝

연관 분석을 통한 데이터를 딥러닝 모듈을 만들기 위한 데이터로 활용하기 위하여 총 8891건의 학생 데이터를 모델을 만들기 위한 Training Data, 만들어진 모델을 평가하기에 필요한 Testing 데이터로 분리하였다. Training Data로 모델을 훈련시키고 만들어진 딥러닝 모델이 새로 들어오는 입력에 대해 결과 예측치를 제시 하고자 한다.

연관분석을 바탕으로 한 데이터는 Fig. 2에서와 같이 각각의 분류 데이터를 바이너리 구조로 변환한 후 4,000개의 Training Data 2set, 800개의 Testing Data 로 구분하여 사용하였다.

Training 및 Testing 결과로 학생들의 중도이탈여부 판단에 사용될 딥러닝 모듈을 구성하고 향후에 실제 학생 데이터를 만들어진 모듈에 적용하여 학생 지도에 활용할 수 있을지에 대한 연구를 하고자 한다. Machine Learning 라이브러리는 구글에서 사용하고 있는 오픈소스 소프트웨어 라이브러리인 TensorFlow을 사용하였다. 딥러닝 모듈의 구성은 Fig. 3에서와 같이 1개의

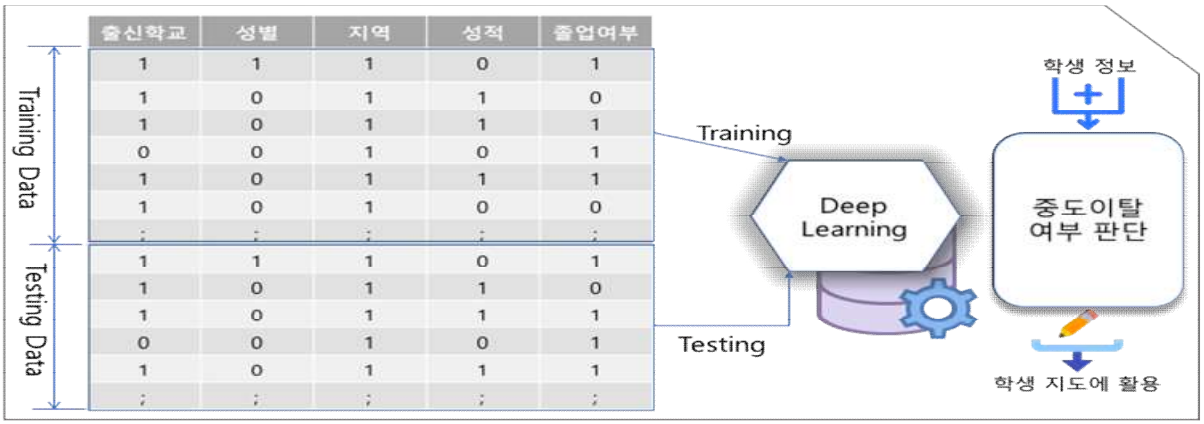

Fig. 2 Using Training Data, Testing Data

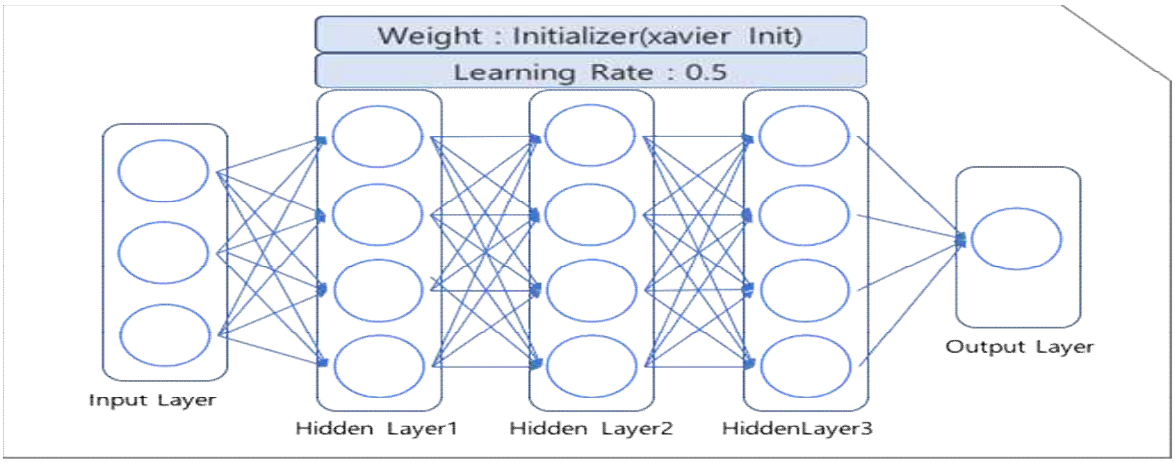

Fig. 3 Nueral Network(NN)

입력계층, 3개의 은닉계층, 1개의 출력계층으로 구성 하였다.

은닉계층을 3개보다 더 많이 구성하면 과대적합 문제가 발생하여 Fig. 4에서와 같이 정확도가 더 낮게 나타났다. Error loss(cost)를 위한 Learning Rate는 몇 번의 Training 결과를 반영하여 가장 큰 인식율을 나타낼 수 있는 0.5 로 결정하였다. 각각의 계층에 입력되는 Input Data에 각각의 Weight를 부여해야 되는데 각 계층별 Weight 초기화 또한 중요한데 Weight에 대한 초기값을 0으로 설정하면 딥러닝 알고리즘이 전혀 동작하지 않기 때문에 본 연구에서는 지금까지 가장 단순하면서도 가장 좋은 결과를 나타내는 Xavier 초기화 모듈을 사용하였다. Xavier 초기화는 입력값과 출력값 사이의 난수를 선택해서 입력값의 제곱근으로 나누는 초기화 방법이다.

```
       [ True],
       [ True],
       ...,
       [False],
       [False],
       [False]], dtype=bool), 0.74825001]
Accuracy: 0.74825
```

Fig. 4 Accuracy when Configuring 5 Hidden Layers

4. 분석 결과

본 연구에서는 Table 3에서와 같이 분석에 사용된 데이터는 8891레코드에 12컬럼의 크기를 가지는데 밀도가 0.416667정도에 불과하다. 즉 8891 * 12 * 0.4166667 = 49,789의 항목만이 분석에 포함되었다. 학업 중단율은 26.88%로서 전문대학 평균보다 월등히 높아 반드시 관리해야 될 항목으로 고려되었다.

Table 3 Top 10 Items with 10% Frequency of Use and Support of Data

transactions as itemMatrix in sparse format with 8891 rows (elements/itemsets/transactions) and 12 columns (items) and a density of 0.4166667 most frequent items:					
address =U	grad =P	school =GP	g1 =low	sex =F	Other
8255	6501	5988	4927	4622	14162

Table 4 Analysis Result Summary

rulelengthdistribution (lhs + rhs):sizes		
1 2 3 4 5		
9 74 206 214 74		
Min. 1st Qu. Median Mean 3rd Qu. Max.		
1.000 3.000 3.000 3.468 4.000 5.000		
support	confidence	lift
Min.: 0.01001	Min.: 0.2000	Min.: 0.4623
1st Qu.: 0.03284	1st Qu.: 0.4295	1st Qu.: 0.9438
Median: 0.09538	Median: 0.6357	Median: 1.0685
Mean**: 0.13228**	Mean: 0.6237	Mean: **1.6519**
3rd Qu.: 0.18018	3rd Qu.: 0.8218	3rd Qu.: 1.2255
Max.: 0.92847	Max.: 0.9981	Max.: 13.1853

Apriori 알고리즘을 사용해서 minimum support = 0.01, minimum confidence = 0.2로 하여 분석한 결과 총 577개의 rule 중에서 rule이 3개의 아이템으로 이루어져 있는 rule이 대략 206개, 4개의 아이템으로 이루어져 있는 rule이 대략 214개로 구성되어 있다.

의미있는 지지도(Support)는 0.13, 신뢰도(Confidence)는 0.62, 향상도(Lift)는 1.65로서 도시지역에 거주하면서 일반계고등학교를 졸업하고 여성으로 구성된 데이터가 최소 지지도와 최소 신뢰도를 넘어서는 것으로 나타나는 것으로 볼 때 전반적으로 이러한 항목들이 졸업여부에 영향을 미치는 것으로 나타났다(Table 4).

연관분석이 완료된 데이터를 의사결정 나무모델에 적용하기 위해 각각 Training Data를 구성하고 Testing Data로 의사결정 나무모델을 활용하여 평가한 결과 Table 5와 같이 예측 정확도는 약 75%로 측정되었다.

딥러닝의 예측 정확도 80% 보다 조금 낮은 예측을 나타내었다. 비교적 직관적이고 판정기준을 명확하게 확인할 수 있는 점에서는 많은 장점이 있는 예측모델이나, 통계적인 데이터가 미리 준비되어 있어야 하는 점과, 현재의 데이터를 바로 대입하여 이탈율을 예측하지 못한다는 점에서 딥러닝 모델 보다는 제약이 많은 것으로 나타났다.

Table 5 Measure Accuracy Using Decision Tree Model

Confusion Matrix and Statistics
Reference
Prediction H P
H 408 358
P 311 1590
Accuracy : 0.7492
95% CI : (0.7322, 0.7655)
No Information Rate : 0.7304
P-Value [Acc > NIR] : 0.01490
Kappa : 0.3759
Mcnemar's Test P-Value : 0.07533
Sensitivity : 0.5675
Specificity : 0.8162
Pos Pred Value : 0.5326
Neg Pred Value : 0.8364
Prevalence : 0.2696
Detection Rate : 0.1530
Detection Prevalence : 0.2872
Balanced Accuracy : 0.6918

딥러닝은 TensorFlow Library를 활용하였으며, Weight 초기화는 xavier, Loss를 결정하는 Learning Rate는 0.5로 하여 개발된 딥러닝 모듈을 Training한 후 별도의 Testing 데이터로 Testing한 결과 Fig. 5에서와 같이 정확도가 80% 정도로서 의사결정나무 모델보다는 비교적 양호한 결과를 나타내었다.

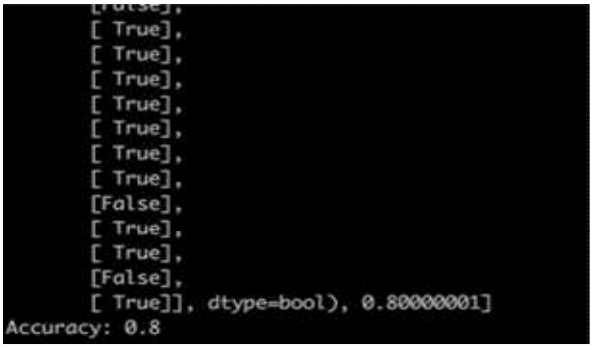

Fig. 5 Deep Learning Accuracy

또한 Fig. 6에서와 같이 Loss를 나타내는 Cost 역시 감소하는 것으로 나타났다.

의사결정트리를 활용하여 졸업여부를 측정한 결과는 Table 6에서와 같이 딥러닝의 정확도 보다 낮은 75%로 나타났다. 의사결정트리에서 졸업여부를 결정하는 요인은 일반고를 졸업하고, 도시지역에 거주하면서 여성이면서 성적이 높은 학생들이 졸업확율이 높은 것으로 나타났다.

결과적으로 의사결정트리보다는 개발된 딥러닝 모듈이 더 효율적으로 학생들의 졸업여부를 평가할 수 있는 모델로 규명되었다.

Table 6 Decision Tree and Deep Learing Prediction Accuracy Comparison

Devision	Constitution	Accuraty
Decision Tree	predict confusionMatrix	75%
Deep Learning	Hidden Layer : 3 Learning Rate : 0.5 초기화 : xavier	80%

딥러닝 모듈의 구성은 3개의 은닉계층으로 구성될 경우, Learning Rate를 0.5, Weight 초기화에는 xavier초기화 모듈 사용, 각 계층의 출력 보정에는 렐루함수(ReLU, Rectified Leanear Unit)를 사용한 결과 정확도는 80%정도로서 현재 까지 가장 좋은 정확도를 보여 주었으며 은닉계층을 3개 이상 구성 시 Fig. 7에서와 같이 과대적합의 문제로 인하여 정확도가 점점 감소하였다.

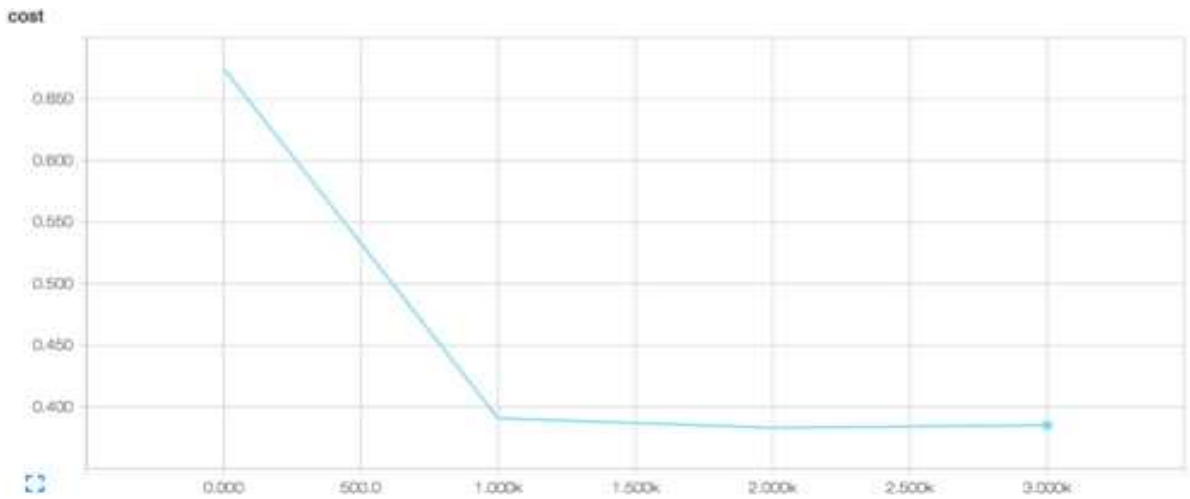

Fig. 6 Cost Change Status

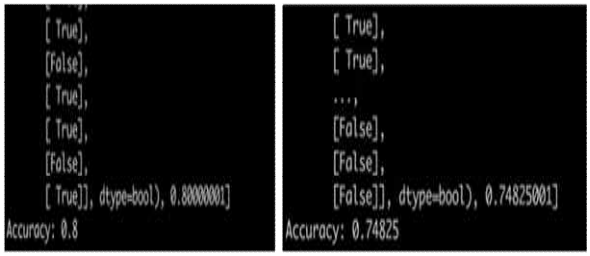

Fig. 7 Comparison of Accuracy Difference according to Hidden Layer Configuration (Layer 3, Layer 5)

5. 결 론

본 연구에서는 일반 대학교에서 정규학기를 마치지 않고 학교를 중도에 포기하는 학생들의 연관성이 있다는 것을 알아보았다. 특히 특성화고를 졸업하고 대학에 진학한 학생들이 학교를 중도에 포기하는 확률이 높다는 것을 확인 하였다. 이러한 결과는 특성화 고교가 학업 성취도 측면에서나 학업 지속적인 면에서 더 어려움을 겪고 있다는 것을 의미하며 이들에 대한 우선적인 관리 및 관심이 필요할 것으로 보인다.

딥러닝의 정확도는 80% 정도로 기대했던 만큼의 결과를 보여 주긴 하였으나, 정확도 90% 이상의 결과를 나타내기 위해서는 지역별, 대학별로 좀 더 다양한 Training 데이터를 확보하고 딥러닝 모듈의 계층을 좀더 넓게 적용할 필요가 있다.

현재의 정확도만으로도 학생 이탈율 방지에 적용하여 지표의 보조 데이터로 활용할 수 있을 만큼 유의미한 정보를 제공하는 것으로 판단되나, 지방의 A대학만을 대상으로 분석하였으므로 향후 추가 연구를 통해서 이들을 체계적으로 관리할 수 있는 일반화 된 지표 및 항목을 개발할 수 있을 것이다. 향후 개발된 지표를 활용하여 학교 현장에서 입학 단계부터 학생유지율 확보를 위한 방안 마련을 위한 좀 더 효과적인 자료로 활용이 가능할 것이다.

References

[1] Kim, S. S., “A Exploratory Study on Withdrawal and Transfer of Korean College Students: The Influence of College-Choice Reason and Satisfaction Afterwards,” The Journal of Korean Education, Vol. 35, No. 1, pp. 227-249, 2008.

[2] Kim, J. M., “ Study on Prevention of Student Drop out Rate Using Deep Learning,” Jeju National University, KOREA, 2017.

[3] Bang, E. H. and Others 7, “Effect of Korean High School Student’s Mental Health on Academic Achievement and School Dropout Rate,” J Korean Acad Child Adolesc Psychiatry, Vol. 27, No. 3, pp. 173-180, 2016.

[4] Lee, H. J. and Kim, Y. N., “The School Related Factors Affecting High School Dropout Rates,” Asian Journal of Education Vol. 13, No. 1, pp. 149-185, 2012.

[5] Sim, H., “A Grounded Theory-Based Ana lysis on the Factors that Causing Dropout of Students in Korean National Universitie s,” Journal of Education and Culture, Vol. 23, No. 2, pp. 105-128, 2017.

[6] Lee, H. J., “Individual, Family, School and Community Related Variables Predicting Longitudinal School Adjustment in Urban Adolescents,” Journal of Education and Culture, Vol. 21, No. 2, pp. 27-56, 2015.

[7] Cho, Y. J. and Han, S. G., “An Analysis of Student and Parent Factors Influencing School Adjustment,” The Education Assignment Institute of Chonbuk National University, KOREA, pp. 117-144, 2015.

[8] Lee, B. H. and Kang, D. K., “A Study of School Maladjustment Action Factors in Secondary School Students,” Journal of Education and Culture, Vol. 20, No. 3, pp.

125-148, 2014.
[9] Lee, J. H. and Lee, H. K., "A Study on Unstructured Text Mining Algorithm through R Programming Based on Data Dictionary," Journal of the Korea Industrial Information Systems Research, Vol. 20, No. 2, pp. 113-123, 2015.
[10] Lee, Y. H. and Jeon, H. J., "Students' Information Communication Skill Affecting Relationship among Technology Acceptance, Education Service Quality, Relationship Quality, and Education Service Satisfaction," Journal of the Korea Industrial Information Systems Research, Vol. 16, No. 5, pp. 73-81, 2011.
[11] Lee H. G. and Shin, Y. H., "Protein Disorder/Order Region Classification Using EPs-TFP Mining Method," Journal of the Korea Industrial Information Systems Research, Vol. 17, No. 6, pp. 59-72, 2012.

김 종 만 (Kim Jong-Man)

- 정회원
- 강원대학교 지질학과 학사
- 제주대학교 경영정보학과 석사
- 관심분야 : 인공지능, MIS, Database

이 동 철 (Lee Dong-Cheol)

- 정회원
- 충남대학교 전기공학교육과 학사
- 국민대학교 MIS학과 석사
- 성균관대학교 산업공학과 박사
- 제주대학교 경상대학 경영정보학과 교수
- 관심분야 : MIS, Agent, 정보시스템, 콘텐츠 비즈니스