The Journal of The Institute of Internet, Broadcasting and Communication (IIBC)
Vol. 20, No. 4, pp.187-193, Aug. 31, 2020. pISSN 2289-0238, eISSN 2289-0246

https://doi.org/10.7236/JIIBC.2020.20.4.187
JIIBC 2020-4-26

대학생의 중도탈락에 영향을 미치는 요인 다중회귀분석

A Regression Analysis of Factors Affecting Dropout of College Students

황승연[*], 신동진[**], 오재곤[***], 이용수[****], 김정준[*****]

Seung-Yeon Hwang[*], Dong-Jin Shin[**], Jae-Kon Oh[***],
Yong-Soo Lee[****], Jeong-Joon Kim[*****]

요 약 본 연구에서는 국내 대학 차원에서의 요인들을 중심으로 대학생의 중도탈락에 영향을 미치는 것이 무엇인지 분석하고자 하였다. 또한, 국립대학과 사립대학, 수도권(서울, 경기, 인천)에 소재하고 있는 대학과 수도권에 소재하고 있지 않은 대학으로 나누어 보다 대학별 특징에 맞춰 분석하였다. 대학의 학생변동사항 중 휴학과 전출을 제외하고, 자퇴를 중도탈락으로 정의하였다. 데이터는 교육부와 한국교육대학협의회에서 주관하여 운영하고 있는 "대학알리미" 포탈에서 원시자료를 받아 분석에 사용하였다. 대학 알리미에서 "대학"으로 분류된 학교 가운데 209개 대학을 최종 분석에 사용하였으며, 졸업생 취업 현황·평균 졸업 학점·재학생 1인당 장학금·평균 등록금·휴학생·재적 학생·경쟁률·전임교원 1인당 학생 수·교지확보율(%)을 독립변수로 투입하여 다중회귀분석을 통해 분석하였다. 분석결과, 첫째, 전체적으로 졸업생 평균 졸업 학점과 취업률이 높을수록 대학생 중도탈락률이 낮은 것으로 나타났다. 둘째, 국립대학과 비교하면 사립대학에서의 평균 등록금이 비쌀수록 대학생 중도탈락률에 부정적인 영향을 미치는 것으로 나타났다. 셋째, 수도권 외 소재 대학에 비해 수도권 소재 대학에서의 등록금이 비쌀수록 대학생 중도탈락률에 부정적인 영향을 미치는 것으로 나타났다.

Abstract In this study, we wanted to analyze the factors at the national university level that affect college students ' elimination. In addition, national universities, private universities, universities in Seoul and universities outside of Seoul were divided into more college-specific characteristics. Except for leave of absence and departure from school, it was defined as a middle school dropout among changes of students. The data were used for analysis by receiving raw data from "University Alerts," which are operated by the Ministry of Education and the Korean Council for Educational Universities. At the university notification, 222 universities out of the schools classified as "Universities" were utilized for final analysis, and jobs, credits, scholarships, tuition fees, students, independent students, and full-time teachers were secured through multiple education.
Overall, the higher the average graduate level and employee-rate the lower the rate of elimination from the middle of college students, the analysis showed. Second, the higher the average tuition fees at private universities, the more negatively affects the rate of elimination of university students. Third, higher tuition fees at universities outside the Seoul metropolitan area have a negative impact on the rate of elimination of students.

Key Words : Big Data, R, multiple regression analysis

[*]준회원, 안양대학교 컴퓨터공학과 석사과정
[**]준회원, 안양대학교 컴퓨터공학과 박사과정
[***]정회원, ㈜진우산전 이사
[****]정회원, 여주대학교 소프트웨어융합과 교수
[*****]정회원, 안양대학교 ICT융합학부 소프트웨어전공 교수
접수일자 2020년 4월 3일, 수정완료 2020년 6월 22일
게재확정일자 2020년 8월 7일

Received: 3 April, 2020 / Revised: 22 June, 2020 /
Accepted: 7 August, 2020
[*****]Corresponding Author: jjkim@anyang.ac.kr
Dept. ICT Convergence Engineering at Anyang University, Korea.

Ⅰ. 서 론

우리나라의 고등교육은 양적으로 매우 급격하게 팽창해 왔다. 고등교육의 양적인 증가는 연도별 대학 수의 변화를 통해 확인할 수 있는데, 1980년 85개교에서 꾸준히 증가하여 2000년 161개교, 2010년 179개교로 증가하였으며, 2017년에 189개의 대학이 운영되고 있다(교육통계서비스). 1980년 27.2%에 불과했던 대학 진학률도 지속적으로 증가하여 2000년 62%, 2010년 75.4%에 달하였으며, 2017년에는 68.9%로 조금 감소하였지만 여전히 높은 수준을 유지하고 있다. (교육통계서비스)

최근 전 세계적으로 지식과 정보의 경제적 가치가 중요하게 고려되면서 대학 경쟁력 제고 및 교육의 질 향상에 대한 요구가 높아지고 있다. 더구나 출산율 저하에 따른 지속적인 학령인구의 감소로 인해 앞으로 각 대학은 입학자원의 부족으로 대학 운영과 존립에 위기를 맞을 수 있다(김수연, 2006). 안정적인 대학 운영을 위해 대학 경쟁력에서 우위를 점하기 위한 많은 노력이 필요한 시점에 직면하였다.

학령인구의 감소에 따라 2018년부터 대학 입학정원이 고교 졸업자 수를 초과하게 될 것으로 예상되며, 이는 대학의 충원율 감소로 이어질 것으로 전망하고 있다(교육부, 2014). 이뿐만 아니라 대학생의 중도탈락률은 해마다 증가하는 추세를 보이고 있어 문제가 심각한 상황이다. 대학 알리미 포탈에서의 공시자료를 통해 대학생의 중도탈락률을 살펴보면, 전문대학의 경우 2000년 4.6%에서 2013년 7.6%로 중도탈락학생 비율이 증가하였으며, 4년제 대학 또한 2000년 3.6%에서 2013년 6.4%로 증가하였다 실제로, 대학의 중도탈락 학생수가 2010년부터 지속적으로 매년 14만명 정도로 나타나고 있고, 이에 대한 사회적 비용이 3조원이 넘는 것으로 추정되고 있다(한국일보, 2015). 또한 교육부의 '중도탈락 대학생의 경제·사회적 비용' 자료에서는 학업을 중도에 포기하는 학생들의 등록금, 입학금, 교재비 등을 추산하면 1인당 800만원 상당에 달하며, 중도탈락 학생들이 취업했을 때의 경제적 이익은 1인당 1,729만원으로 이를 종합하여 계산하면 총 2조 5,187억원의 기회비용이 발생하는 것으로 나타난다(중앙일보, 2013).

국내에서 이루어진 선행연구들을 살펴보면, 대학생의 중도탈락에 영향을 미치는 요인으로 설정한 변수들이 대부분 개인적인 요인에 집중되어 있으며, 대학 수준의 요인을 중점적으로 분석한 연구는 부족한 상황이다. 따라서 본 연구에서는 대학의 중도탈락 현황을 살펴보고, 중도탈락에 영향을 미치는 대학 요인이 무엇인지 분석하여 대학생 중도탈락률을 낮추기 위해 대학 차원에서 고려해야 할 사항에 대한 시사점을 도출하고자 한다.

Ⅱ. 관련기술

1. 빅데이터(Big Data)

빅데이터란 기존 데이터베이스 관리도구로 대량의 정형 또는 비정형 데이터를 포함한 데이터로부터 가치를 추출하고 수집, 저장, 관리, 분석할 수 있는 기술이다.

가트너는 빅데이터를 세 개의 차원으로 정의하였는데 이는 데이터의 양(Volume), 데이터 입출력의 속도(Velocity), 데이터 종류의 다양성(Variety)을 뜻한다. 2012년에는 가트너의 기존 정의를 개정하였는데 정확성(Value)이나 복잡성(Complexity)을 덧붙이기도 한다[1].

대용량 데이터를 가지고 자료 관리기술과 자료 분석기술을 이용할 수 있다. 자료관리 기술은 Hadoop등을 이용하고 자료 분석 기술로는 통계학, 기계학습, 인공신경망, 데이터 마이닝 등을 이용할 수 있다[2,3].

2. 하둡(Hadoop)

하둡은 대용량 데이터를 분산 처리할 수 있는 자바 기반의 오픈 소스 프레임워크이다. 하둡은 분산시스템인 HDFS(Hadoop Distributed File System)에 데이터를 저장하고, 맵리듀스(MapReduce)를 이용해 데이터를 처리한다. 일반적인 파일 시스템처럼 불륨 기반으로 파일을 적재하거나 파일 단위로 저장하지 않고, 파일을 블록 단위로 쪼개 여러 서버에 분산 저장한다. 따라서 서버의 디스크 용량보다 큰 수십 테라바이트 또는 페타바이트 이상의 대용량 파일도 저장하고 처리할 수 있다[4,5].

3. 하이브(Hive)

HIve는 하둡에서 동작하는 데이터 웨어하우스(Data Warehouse) 인프라 구조로서 데이터 요약, 질의 및 분석 기능을 제공한다. HiveQL은 SQL언어와 유사하지만 기능은 다소 부족하다. 하지만 고급 조인이 필요하지 않은 경우 SQL로 할 수 있는 모든 작업을 SQL과 동일하게 처리가 가능하다. Hive 엔진을 사용하여 mapreduece를 작성하지 않고 쿼리 언어만으로 hadoop의 비정형 데이터 분석이 가능하다[6].

4. R 기반 통계분석

R은 무료 오픈소스 통계 프로그래밍 환경이며, 다양한 통계적 기법과 수치 해석 기법을 제공할 뿐만 아니라 우수하고 다양한 그래픽 방법이 있어 이용자가 새로운 함수를 작성하여 확장 및 추가할 수 있다. 또한, 한 번에 하나의 프로세스를 수행하는 대화식이기 때문에 분석하는 동안 보이는 것에 기초하여 변경이 가능하다. R은 주로 연구 및 산업별 응용 프로그램으로 많이 사용되고 있으며 최근에는 기업에서도 많이 사용하기 시작했다. 특히 빅데이터 분석을 목적으로 주목받고 있으며 5000개가 넘는 패키지들이 다양한 기능을 지원하고 있으며 수시로 업데이트 되고 있다[7].

자바(Java), C, C++, 파이썬(Python) 등 다른 프로그램 언어와도 쉽게 연동할 수 있으며, 윈도, 리눅스·유닉스, 맥(Mac) OS 등 대부분의 개발 환경을 지원한다. 또한, 하둡 분산처리 환경을 지원하는 라이브러리가 제공되기 때문에 구글, 페이스북, 아마존 등도 빅데이터 분석에 R을 사용하고 있다. R은 빅데이터 분석에 가장 강력하면서 유용한 도구로 점차 자리를 잡아가고 있다[8].

5. 다중 회귀 분석(multiple regression analysis)

회귀분석은 종속변수(Dependent Variable)와 독립변수(Independent Variable)간의 상관관계를 검증하여 독립변수가 종속변수에 어떠한 영향력을 미치는지 파악하거나, 독립변수의 변화에 따라 종속변수의 변화를 예측하기 위하여 사용되는 통계학적 분석방법이다. 회귀분석은 독립변수의 개수에 따라 독립변수가 둘 이상인 경우는 다중회귀분석, 하나인 경우는 단순회귀분석이라 한다[9,10].

회귀분석이 사용되는 이유는 결과(종속변수)의 일부 원인(독립 변수)을 한 번에 분석이 가능하기 때문이다. 또한 회귀분석에서는 종속변수에 대한 각각의 독립변수들이 어떠한 영향을 미치는지 개별적으로 분석이 가능하기 때문에 특정 변수를 통제할 시 다른 독립변수가 종속변수의 변화에 어떠한 상호관련성이 있는지 쉽게 판단이 가능하다. 하지만, 독립변수간의 상호연관성을 배제하고, 단방향의 관계만을 취급하는 특징을 갖고 있다. 또한 측정 오차를 허용하지 않는 특징을 가지고 있다. 따라서 독립변수내의 관련성 문제 및 다중공선성 문제를 극복 가능한 간단한 인과모형을 대상으로 할 시, 종속변수에 대한 독립변수간의 상호영향력의 크기를 비교 가능한 뛰어난 통계기법 중 하나이다[11].

Ⅲ. 연구방법

1. 분석대상

본 연구에서는 대학생의 중도탈락에 영향을 미치는 대학 요인을 분석하기 위해 교육부와 한국교육대한협의회에서 주관하여 운영하고 있는 '대학 알리미' 포탈을 통해 원시자료를 받아 분석에 사용하였다. 가장 최근 자료인 2018년도 자료에는 일부 누락된 데이터가 있어 분석 자료의 기준년도는 2017년으로 설정하였다.

대학 유형에 따라 대학의 특성이 다양하며, 학교 유형에 따라 중도탈락률에 차이가 나타날 수 있기 때문에 대학 알리미에서 "대학"으로 분류된 학교 가운데 분교와 전문대학 및 사이버대학, 방송통신대학을 제외하고 최종적으로 209개 대학을 분석에 사용하였다. 또한 지역별로 수도권(서울, 경기, 인천)에 소재한 대학과 수도권에 소재하지 않은 대학, 국립대학과 사립대학으로 나누어 나타난 결과를 분석해 보았다. 분석에 사용한 주요 변수는 표 1과 같다.

표 1. 변수설명
Table 1. Variable Description

변수구분	변수설명
종속변수 재학생 중도탈락률	(중도탈락 학생 수/재학생수) *100
독립변수 졸업생 취업 현황	(취업자 수/{졸업자 수-(진학자 수+입대자 수+취업불가능자 학생 수+외국인유학생 수+건강보험 직장가입 제외대상 수)} *100
평균 졸업 학점	(졸업자 평점의 합/졸업자수)
재학생 1인당 장학금	(교내장학금+교외장학금)/재학생수
평균 등록금	전체학과 등록금의 합/정원내 입학정원의 합
휴학생	정원내 휴학생
재적학생	정원내 휴학생+정원내 재학생
경쟁률	정원내 지원자/정원내 모집인원
전임교원 1인당 학생수	재학생학생현황/ 재학생기준전임교원
교지확보율(%)	보유면적(m^2)/재학생 기준면적(m^2)*100

2. 데이터 정제

대학의 학교 요인이 대학생의 중도탈락에 미치는 영향력을 분석하기 위하여 먼저 대학의 학교 요인들을 추출하는 작업이 필요하다. 원하는 변수가 존재하는 파일들을 Hadoop에 저장하여 Hive를 이용해 테이블 형태로 저장하였다. 저장된 테이블에서 필요한 컬럼들을 HiveQL을 사용하여 분석에 필요한 변수들로 이루어진 새로운 테이블을 생성하여 하둡에 저장하였고, 이를 다시 로컬에 저장하여 그림 1과 같은 결과를 얻을 수 있다.

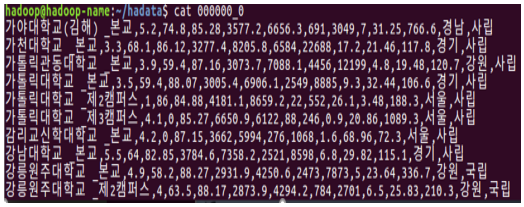

그림 1. 필요한 변수들을 추출한 최종 형태
Fig. 1. Required parameter extraction method

3. 분석방법

대학의 학교 요인이 대학생의 중도탈락에 미치는 영향력을 분석하기 위하여 다중회귀분석을 사용하였다. 다중회귀분석은 독립변수가 2개 이상인 추정식을 이용하는 회귀분석으로 여러 개의 독립변수 중 종속변수에 가장 큰 영향을 미치는 변수가 무엇인지, 종속변수를 설명해 줄 수 있는 가장 적합한 모형이 무엇인지 밝히는 통계적 방법이다. 알고자하는 독립변수들을 순차적으로 투입하면서 대학의 요인들의 설명력을 확인할 수 있기 때문에 분석방법으로 사용하였다.

본 연구에서는 대학생 중도탈락률을 종속변수로 설정하고, 독립변수로는 졸업생 취업 현황·평균 졸업 학점·재학생 1인당 장학금·평균 등록금·휴학생·재적 학생·경쟁률·전임교원 1인당 학생 수·교지확보율(%) 변인을 투입하여 분석하였다. 모형에 투입하여 대학의 변인들이 중토탈락률에 미치는 영향력 정도를 살펴보고자 하였다. 다중회귀분석에 사용된 수식은 그림 2와 같고, Y_i는 종속변수인 표 1의 재학생 중도탈락률을 의미하며, 각 X_{ni}는 표 1의 독립변수를 의미한다.

$$Y_i = B_1 + B_2X_{2i} + B_3X_{3i} + \ldots + B_{10\Sigma}X_{10i} + u_i (i = 1,2,\ldots,n)$$

그림 2. 다중회귀분석 수식
Fig. 2. Multiple Regression Analysis Formula

Ⅳ. 분석 결과

본 연구에서는 대학생의 중도탈락에 영향을 미치는 대학 요인을 분석하기 위해 분교와 사이버대학과, 전문대학, 교육대학을 제외한 209개 학교를 분석대상으로 사용하였다. 209개 4년제 대학의 대학생 중도탈락률 평균은 4.83%로 나타났으며, 가장 중도탈락이 높은 대학은 19.2%에 달하는 것으로 나타났다. 대학 평판도와 관련된 정원 내 신입생 경쟁률 평균은 1:8로 나타났으며 졸업생 취업 현황의 평균은 66%, 평균 졸업 학점의 평균은 86.8점, 재학생 1인당 장학금 평균은 3,423만원, 평균 등록금의 평균은 6,530만원, 휴학생의 평균은 2496명, 재적학생의 평균은 8,646명, 전임교원 1인당 학생 수의 평균은 27명, 교지확보율을 평균은 318%로 나타났다.

대학 설립유형별로 살펴보면 국공립학교 46개교(22%), 사립학교 163개교(78%)로 사립대학의 비중이 높게 나타났으며 수도권(서울, 경기, 인천)에 소재한 대학이 75개교(35.8%), 수도권에 소재하지 않은 대학이 134개교(64.1%)로 나타났다.

표 2. 대학별 데이터
Table 2. University Data

fail	employ ment	grade	scholar ship	tuition	absenc e	enrollm ent	compet ition	teacher	grand
5.2	74.8	85.28	3,577.20	6,656.30	691	3,049	7	31.25	766.6
3.3	68.1	86.12	3,277.40	8,205.80	6,584	22,688	17.2	21.46	117.8
3.9	59.4	87.16	3,073.70	7,088.10	4,456	12,199	4.8	19.48	120.7
3.5	59.4	88.07	3,005.40	6,906.10	2,549	8,885	9.3	32.44	106.6
1	86	84.88	4,181.10	8,659.20	22	552	26.1	3.48	188.3
4.1	0	85.27	6,650.90	6,122	88	246	0.9	20.86	1,089.30
4.2	0	87.15	3,662	5,994	276	1,068	1.6	68.96	72.3
5.5	64	82.85	3,784.60	7,358.20	2,521	8,598	6.8	29.82	115.1
4.9	58.2	88.27	2,931.90	4,250.60	2,473	7,873	5	23.64	336.7
4	63.5	88.17	2,873.90	4,294.20	784	2,701	6.5	25.83	210.3
2.6	57.4	87.57	2,459.60	4,050.50	7,015	19,950	7.9	22.36	175.6
3.6	66.5	86.9	3,430.60	4,242.70	4,011	9,657	5	32.97	277.4
2.6	61.7	88.01	3,156.60	8,167.60	6,192	19,110	16	27.01	82.2
4.2	64.2	87.55	3,680.60	7,820	3,582	10,394	10.2	16.19	137
5.4	76.2	86.21	3,607	6,321.30	1,324	5,269	6.7	27.1	168.3
2.9	86.4	86.95	3,444.40	7,179.70	587	3,734	6.9	16.46	134
3.4	62.6	86.1	3,621.40	7,263.10	4,658	15,137	10.8	30.44	133.2

표 2의 fail은 퇴학률을, employment은 졸업생 취업 현황, grade은 평균 졸업 학점, scholarship은 재학생 1인당 장학금, tuition은 평균 등록금, absence은 휴학생, enrollment은 재적 학생, competition은 경쟁률, teacher은 전임교원 1인당 학생 수, grand은 교지확보율(%)을 지칭하며 “대학알리미” 포탈에서 제공하는 데이터를 위 그림과 같이 정제하였다. 또한 분석에 앞서 변수간 충분한 변량을 확보하였는지 확인하였다.

전체 데이터에서 다중회귀분석을 한 결과에서 취업률 변수인 employee_rate, 학점 변수인 grade, 경쟁률 변수인 competition, 전임교원 1인당 학생 수 변수인 teacher의 p-value 값이 0.05보다 작으므로 중도탈락률과 연관성이 있음을 아래 그림 3에서 확인할 수 있다. 또한, Estimate 값은 다른 변수가 고정되어있고, 각 X가 한 단위 변화했을 때 중도탈락률의 평균적인 변화량을 나타낸다.

```
> summary(out1)

Call:
lm(formula = drop_rate ~ ., data = dropout1)

Residuals:
    Min      1Q  Median      3Q     Max
-6.1206 -1.4172 -0.2727  0.7672 12.1202

Coefficients:
                Estimate Std. Error t value Pr(>|t|)
(Intercept)    4.164e+01  7.162e+00   5.815 2.39e-08 ***
employee_rate -9.124e-02  1.907e-02  -4.783 3.35e-06 ***
grade         -3.898e-01  7.719e-02  -5.050 9.97e-07 ***
scholarship    2.501e-04  2.624e-04   0.953  0.34164
tuition        1.808e-04  1.325e-04   1.365  0.17384
absence       -1.370e-04  4.171e-04  -0.328  0.74294
enrollment     4.186e-06  1.313e-04   0.032  0.97460
competition   -1.122e-01  3.750e-02  -2.993  0.00311 **
teacher        8.215e-02  2.829e-02   2.903  0.00411 **
grand          5.358e-05  4.399e-04   0.122  0.90319
---
signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 2.5 on 199 degrees of freedom
Multiple R-squared:  0.3241,    Adjusted R-squared:  0.2936
F-statistic:  10.6 on 9 and 199 DF,  p-value: 2.245e-13
```

그림 3. 전체 다중회귀 분석 결과
Fig. 3. multiple regression analysis results

수도권을 기준으로 한 다중회귀 분석 결과에서는 취업률 변수인 employee_rate, 학점 변수인 grade, 장학금 변수인 scholarship의 p-value 값이 0.05보다 작으므로 중도탈락률과 연관성이 있음을 아래 그림 4에서 확인할 수 있다.

```
> summary(out2)

Call:
lm(formula = drop_rate ~ ., data = dropout2)

Residuals:
    Min      1Q  Median      3Q     Max
-2.9952 -1.0299 -0.0445  0.7965  5.1236

Coefficients:
                Estimate Std. Error t value Pr(>|t|)
(Intercept)    2.453e+01  8.299e+00   2.956  0.00434 **
employee_rate -5.415e-02  2.303e-02  -2.351  0.02178 *
grade         -2.213e-01  9.116e-02  -2.428  0.01796 *
scholarship    9.470e-04  3.458e-04   2.739  0.00795 **
tuition       -9.884e-05  1.645e-04  -0.601  0.55001
absence        2.115e-05  5.278e-04   0.040  0.96816
enrollment    -4.349e-05  1.587e-04  -0.274  0.78490
competition   -4.824e-02  3.182e-02  -1.516  0.13440
teacher        1.939e-02  3.156e-02   0.615  0.54102
grand          3.290e-05  3.905e-04   0.084  0.93312
---
signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 1.52 on 65 degrees of freedom
Multiple R-squared:  0.4081,    Adjusted R-squared:  0.3262
F-statistic:  4.98 on 9 and 65 DF,  p-value: 4.264e-05
```

그림 4. 수도권을 기준으로 한 다중회귀분석 결과
Fig. 4. multiple regression analysis of Capital Region

```
> summary(out3)

Call:
lm(formula = drop_rate ~ ., data = dropout3)

Residuals:
    Min      1Q  Median      3Q     Max
-5.9491 -1.6213 -0.4716  1.1669 10.9637

Coefficients:
                Estimate Std. Error t value Pr(>|t|)
(Intercept)    3.510e+01  1.084e+01   3.237 0.001551 **
employee_rate -1.168e-01  2.514e-02  -4.647 8.49e-06 ***
grade         -3.200e-01  1.153e-01  -2.776 0.006362 **
scholarship   -2.673e-04  3.445e-04  -0.776 0.439185
tuition        6.237e-04  1.997e-04   3.123 0.002230 **
absence       -3.460e-04  5.373e-04  -0.644 0.520775
enrollment     4.352e-05  1.743e-04   0.250 0.803302
competition   -8.853e-02  7.369e-02  -1.201 0.231863
teacher        1.455e-01  3.858e-02   3.770 0.000251 ***
grand         -3.727e-06  6.745e-04  -0.006 0.995600
---
signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 2.721 on 124 degrees of freedom
Multiple R-squared:  0.3759,    Adjusted R-squared:  0.3306
F-statistic: 8.298 on 9 and 124 DF,  p-value: 1.321e-09
```

그림 5. 비수도권을 기준으로 한 다중회귀분석 결과
Fig. 5. multiple regression analysis of NonCapital Region

비수도권을 기준으로 한 다중회귀분석 결과에서는 취업률 변수인 employee_rate, 학점 변수인 grade, 등록금 변수인 tuition, 전임교원 1인당 학생 수 변수인 teacher의 p-value 값이 0.05보다 작으므로 중도탈락률과 연관성이 있음을 아래 그림 5에서 확인할 수 있다.

국립대학을 기준으로 한 다중회귀 분석 결과에서는 취업률 변수인 employee_rate, 학점 변수인 grade, 휴학생 수 변수인 absence, 교지확보율 변수인 grand의 p-value 값이 0.05보다 작으므로 중도탈락률과 연관성이 있음을 아래 그림 6에서 확인할 수 있다.

```
> summary(out4)

Call:
lm(formula = drop_rate ~ ., data = dropout4)

Residuals:
    Min      1Q  Median      3Q     Max
-4.3307 -1.1269 -0.1413  1.0242  6.4719

Coefficients:
                Estimate Std. Error t value Pr(>|t|)
(Intercept)   53.2122897 21.0637770   2.526   0.0161 *
employee_rate -0.1811629  0.0377349  -4.801 2.76e-05 ***
grade         -0.4792354  0.2356731  -2.033   0.0494 *
scholarship   -0.0006466  0.0005883  -1.099   0.2790
tuition        0.0013097  0.0006505   2.013   0.0516 .
absence       -0.0015784  0.0007522  -2.098   0.0430 *
enrollment     0.0004034  0.0002351   1.716   0.0948 .
competition   -0.1771498  0.1255931  -1.411   0.1670
teacher        0.1284935  0.0686471   1.872   0.0694 .
grand         -0.0033530  0.0015859  -2.114   0.0415 *
---
signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 2.292 on 36 degrees of freedom
Multiple R-squared:  0.5444,    Adjusted R-squared:  0.4305
F-statistic:  4.78 on 9 and 36 DF,  p-value: 0.0003188
```

그림 6. 국립대학을 기준으로 한 다중회귀분석 결과
Fig. 6. multiple regression analysis of national university

마지막으로 사립 대학을 기준으로 한 다중회귀분석 결과에서는 학점 변수인 grade, 등록금 변수인 tuition , 경쟁률 변수인 competition, 전임교원 1인당 학생 수 변수인 teacher의 p-value 값이 0.05보다 작으므로 중도탈락률과 연관성이 있음을 아래 그림 7에서 확인할 수 있다.

```
> summary(out5)

Call:
lm(formula = drop_rate ~ ., data = dropout5)

Residuals:
    Min      1Q  Median      3Q     Max
-6.3535 -1.2605 -0.2336  0.9655  8.4208

Coefficients:
                Estimate Std. Error t value Pr(>|t|)
(Intercept)   42.6312503  7.1487017   5.963 1.65e-08 ***
employee_rate -0.0384429  0.0221208  -1.738  0.08425 .
grade         -0.3619574  0.0767900  -4.714 5.43e-06 ***
scholarship   -0.0001938  0.0003096  -0.626  0.53223
tuition       -0.0005744  0.0002201  -2.610  0.00997 **
absence        0.0004165  0.0004547   0.916  0.36115
enrollment    -0.0001660  0.0001441  -1.151  0.25135
competition   -0.1023512  0.0366545  -2.792  0.00590 **
teacher        0.0904303  0.0306839   2.947  0.00371 **
grand          0.0001751  0.0004291   0.408  0.68377
---
signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 2.313 on 153 degrees of freedom
Multiple R-squared:  0.3701,    Adjusted R-squared:  0.3331
F-statistic: 9.989 on 9 and 153 DF,  p-value: 5.47e-12
```

그림 7. 사립대학을 기준으로 한 다중회귀분석 결과
Fig. 7. multiple regression analysis of private university

V. 결 론

본 논문에서는 대학알리미에서 제공하는 데이터의 대학생을 중심으로 중도탈락에 영향을 미치는 변수에 대해 전체, 수도권, 비수도권, 국립, 사립 조건에서 측정 하였다. 측정된 데이터를 Hive 솔루션을 통해 정제 후, 다중 회귀분석을 사용하여 분석하고, 어떠한 상관관계가 있는지를 확인하였다.

학점 변수인 grade와 취업률 변수인 employee_rate가 중도탈락률 변수인 drop_rate에 가장 유의한 관계임을 확인할 수 있었다. 또 상대적으로 수도권(서울, 경기, 인천)대학이 비수도권대학 보다 등록금이 낮게 측정되고, 국립대학 보다는 사립대학이 낮게 측정되었다. 수도권대학의 분석에서 장학금 변수인 scholarship이 연관성이 있음을 확인하고, 사립대학의 분석에서 등록금 변수인 tuituion이 연관성이 있음을 확인하여 경제적으로 등록금의 부담이 큰 학생들이 자퇴를 할 확률이 높다고 할 수 있다. 마지막으로 관련성이 적은 휴학생, 재적학생, 교지확보율 변수인 absence, enrollment, grand는 대학 차원에서 중도탈락률과 관련하여 고려할 필요성이 없음을 확인할 수 있다.

향후 연구에서는 대학생의 중도탈락에 영향을 미치는 대학 수준의 요인을 보다 정확하게 분석하기 위해서 개인 요인과 결합된 자료를 수집하여 추가 분석할 예정이다.

References

[1] Man-Mo Kang, Sang-Rak Kim, Sang-Mu Park, "Analysis and utilization of big data", Journal of Information Science and Technology, Vol. 30, No. 6, pp. 25-32, 2012.3.

[2] Man-Jai Lee, "Big Data and the Utilization of Public Data", Internet and Information Security, Vol. 2, No. 2, pp. 47-64, 2011.11.

[3] Seung-Yeol Bang, Hyo-Dong Ha, and Chang-Jae Kim, “A Study on BigData-based Software Architecture Design for Utilizing Public Open Data”, The Journal of KIIT, Vol. 13, No. 10, pp. 99-107, 2015.10.
DOI : https://dx.doi.org/10.14801/jkiit.2015.13.10.99

[4] Hyun-Jong Lee, "Big Data Leverages the Hadoop Platform", The Korean Institute of Conmunications and Information Sciences, Vol. 29, No. 11, pp. 43-47, 2012.10.

[5] Hyun-Joo Kim, “Design and Implementation of an Efficient Web Services Data Processing Using Hadoop-Based Big Data Processing Technique”, The Journal of the Korea Academia-Industrial cooperation Society, Vol. 16, No. 1, pp. 726-734, 2015.
DOI: https://doi.org/10.5762/KAIS.2015.16.1.726

[6] Ki-Chan Park, Hong-Kuen Yoon, Seok-Ju Jang, Jung-Jun Lee, "Development of web based hive management tool using open source", Proceedings of the Korea Information Science Society Conference, 1671-1673, 2014.6.

[7] Jong-gi Lee, "A Case Study of R Programming for Big Data Analysis", Computational Accounting Research, Vol. 13, No. 1, pp. 1-22, 2015.6.

[8] Ji-Hee Lee, Joon-Sung Lee, Jung-Wook Son, "R programming based unstructured construction data analysis", Journal of the Architectural Institute of Korea - Structural Systems, Vol. 32, No. 5, pp. 37-44, 2016.5.

[9] Yoo-jae Lee, "A Study on the Verification of the Main Effect in Multiple Regression Analysis including Interaction Effect", Management Research, Vol. 23, No. 4, pp. 183-210, 1994.7.

[10] Uh-Soo Kyun, Sung-Hoon Cho, Jeong-Joon Kim, “A Study on Perception for Public Safety of Seoul Citizens using Multiple Regression Analysis”, The Journal of The Institute of Internet, Broadcasting and Communication, Vol. 18, No. 1, pp. 195-201, Feb 2018.
DOI: https://doi.org/10.7236/JIIBC.2018.18.1.195

[11] Bong-Woo Nam, Kyung-Bin Kim, Kyu-Ho Kim, Jun-Min Cha, "Regional Power Demand Forecasting Algorithm Using Multiple Regression Analysis", Journal of the Korean Institute of Illuminating and Electrical Installation, Vol. 22, No. 2, pp. 63-70, 2008.2.

저 자 소 개

황 승 연(준회원)

• Seung-Yeon Hwang received his BS in Department of Computer Science at Korea Polytechnic University in 2019. He is currently studying MS in Department of Computer Science at Anyang University. His research interests include Database System, Big Data, Data Analysis, Machine Learning, etc.

신 동 진(준회원)

• Dong-Jin Shin received his BS in Enginerring at Korea Polytechnic University in 2018. He is currently a Master's course in the department of Smart Manufacturing Engineering at Korea Polytechnic University. His research interests include Big Data, Internet of Things(IoT), Network Security, etc.

오 재 곤(정회원)

• Jae-Kon Oh received his BS and MS at Kwangwoon University in 1994 and Ajou University in 2005, respectively. In 2017, he received his PhD in at Chonbuk University. He is currently a CEO at SEINSystems. His research interests include Database Systems, BigData, Semantic Web, Geographic Information Systems (GIS) and Ubiquitous Sensor Network (USN), etc.

이 용 수(정회원)

• Yong-soo Lee received his MS in Computer Science at Konkuk University in 1989. In 2015, he received his PhD in Information & Control Engineering at Kwangwoon University. He is currently a professor at the Department of software convergence at Yeoju Institute of Technology. He is the Member of the Korea Institute of Internet, Broadcasting & Communication (IIBC). His research interests include Database Systems, Data Mining, BigData, Wireless Sensor Networks and Ubiquitous Sensor Network (USN), etc.

김 정 준(정회원)

•Jeong Joon Kim received his BS and MS in Computer Science at Konkuk University in 2003 and 2005, respectively. In 2010, he received his PhD in at Konkuk University. He is currently a professor at the department of Computer Science at Anyang University. His research interests include Database Systems, Big Data, Semantic Web, Geographic Information Systems (GIS) and Ubiquitous Sensor Network (USN), etc.