# 학생 중도탈락 예측 모델에 관한 연구†

이종혁*, 김대학*, 길준민**‡
*대구가톨릭대학교 빅데이터공학과
**대구가톨릭대학교 IT 공학부
e-mail : jonghyuk@cu.ac.kr

# A Study on the Prediction Model for Student Dropout

JongHyuk Lee*, DaeHak Kim*, JoonMin Gil**
*Dept. of Big Data Engineering, Daegu Catholic University
**School of Information Technology Engineering, Daegu Catholic University

**요 약**

빅데이터 산업 부상과 함께 교육 데이터 분석 분야가 새롭게 주목받고 있다. 교육 현장에서 학습 데이터의 양과 종류는 꾸준히 증가하고 있고 이를 분석하기 위한 정보기술도 계속 발전하고 있다. 한편, 학교 교육은 사회적 성취와 밀접한 관련이 있어 사회이동의 중요한 수단이 되는 만큼 학교 교육으로부터 이탈할 위험이 있는 학생들을 조기에 발견하여 이탈을 방지하는 것은 매우 중요하다. 본 논문은 대학생의 중도탈락을 예방하기 위해 로지스틱 회귀분석과 다층 퍼셉트론 기법을 이용해 학습 데이터를 분석하여 예측 모델을 생성하고 해당 모델을 평가한다. 평가 결과, 다층 퍼셉트론 모델이 로지스틱 회귀분석 모델에 비해 정확도와 재현율은 우수하였지만 정밀도는 약간 저조하였다.

## 1. 서론

빅데이터의 확산에 힘입어 교육 데이터 분석 분야가 새롭게 관심을 받고 있다. 오래 전부터 학습 데이터를 분석하여 학습자의 성과를 평가 · 관리 하였지만 최근에 학습 데이터의 양과 종류가 증가하고 고도의 데이터 분석이 가능한 정보기술이 개발됨에 따라 새로운 학습 분석이 주목받고 있다.

한편, 학교 교육은 사회적 성취와 밀접한 관련이 있어 사회이동의 중요한 수단이다[1]. 따라서 가능한 빨리 학교 교육으로부터 이탈할 위험이 있는 학생들을 파악하는 것은 물론 어떤 요인이 더 큰 영향을 미치는지 이해하는 것은 매우 중요하다. 대다수 교육자는 학생의 생활습관과 중도탈락률의 관계를 통계학적으로 명백히 밝혀냈더라도 어떤 학생이 언제 중도 탈락하는지 정확히 예측하는 일에 관심을 두기보다 어떻게 하면 중도탈락 없이 학생을 졸업을 잘 시킬 수 있는가에 관심이 있다. 이것이 교육 분야에서 데이터를 분석하는 목적이다. 본 논문은 대학생의 중도탈락을 예방하기 위해 로지스틱 회귀분석(Logistic Regression Analysis)과 다층 퍼셉트론(Multilayer Perceptron) 기법을 이용한 학습 데이터 분석을 통해 예측 모델을 생성하고 해당 모델을 평가한다.

본 논문의 2 절에서는 중도탈락 예방에 관한 연구를 살펴보고 다음 3 절에서는 중도탈락 예방을 위한 모델 생성 및 평가 시스템 아키텍처를 소개한다. 계속해서 4 절에서는 예측 모델 생성을 위해 사용한 로지스틱 회귀분석과 다층 퍼셉트론 기법을 설명하고 5 절에서는 두 기법으로 생성한 모델을 평가한다. 마지막으로 6 절에서는 결론과 향후 계획을 논한다.

## 2. 관련연구

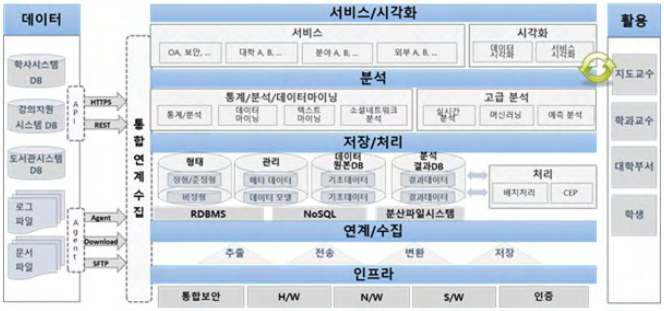
중도탈락 예방에 대한 연구는 중도탈락의 원인 및 요인 분석이 주로 연구되어 왔지만 최근 중도탈락 예측 모델 개발과 이를 이용한 예방 시스템 구축에 관한 연구가 늘어나고 있는 추세이다. 지방대 학생들의 학교생활 만족도를 분석하여 학생들의 학교생활 유지에 긍정적 또는 부정적 영향을 미치는 요소들을 이해하기 위해 재학생의 심층면담을 통해 자료를 수집하고 영역 분석 방법으로 분석한 연구[2]가 진행되었다. 그리고 대학생의 중도탈락에 영향을 미치는 요인이 무엇인지 분석하기 위해 교육부와 한국교육학술정보원에서 운영하고 있는 에듀데이터 서비스 시스템의 알리미 원시자료를 분석한 연구[3]가 있다. 분석 결과, 입학 전형 최종 등록률과 정원내 신입생 경쟁률이 높을수록, 학생 1 인당 교육비가 높을수록, 기숙사 수용률이 높을수록, 전임교원 1 인당 학생수가 적을수록 낮은 중도탈락률을 보였다. 중도탈락 학생을 사전에 예측하는 모델을 개발하여 이를 이용한 웹 기반 중도탈락 예방 시스템을 구축하는 연구[3]도 진행되었다. 이 연구는 예측 모델을 생성하기 위해 진단 속성에

† 이 논문은 2016 년도 정부(교육부)의 재원으로 한국연구재단의 지원을 받아 수행된 기초연구사업임(No. NRF-2016R1D1A3B03933370).
‡ 교신저자

대한 통계적 방법과 나이브 베이즈(Naïve Bayes)를 사용하였다. 이 밖에 중도 탈락 가능성이 높은 학생을 예측하기 위해 다층 퍼셉트론, 서포트 벡터 머신(Support Vector Machine), 의사결정 테이블(Decision Table) 등의 데이터 마이닝 기법을 사용한 여러 연구[6-8]가 진행되었다. 본 논문은 회귀분석 방법과 다층 퍼셉트론 기법을 이용해 예측 모델을 생성하고자 한다.

## 3. 중도탈락 예측 시스템 아키텍처

중도탈락 예방을 위한 모델 생성 및 평가 시스템 아키텍처는 다음 그림과 같이 역할별로 수집, 저장, 처리, 분석 및 시각화 계층으로 나뉜다.




(그림 1) 중도탈락 예측 시스템 아키텍처

- 수집 계층: 본 계층에서는 조직의 내외부에 있는 다양한 시스템으로부터 데이터를 수집한다. 본 수집기는 학사 시스템, 강의지원시스템, 도서관 시스템 등의 데이터베이스, 로그 파일 등 정형 또는 비정형 데이터를 다양한 인터페이스(REST, HTTPS, SFTP 등)로 연결하여 수집한다.
- 저장 계층: 본 계층에서는 수집한 데이터를 분산 스토리지에 영구 또는 임시로 저장한다. 본 저장소는 로그 파일 등 대용량 파일을 영구적으로 저장하기 위해 HDFS 를 사용하고 대규모 메시징 데이터를 영구 저장하기 위해 NoSQL 을 사용한다. 개인정보 데이터는 비식별화 처리하여 저장된다.
- 처리 계층: 본 계층에서는 저장소에 저장된 데이터를 분석에 활용하기 위해 데이터를 정형화 및 정규화 한다. 예를 들면, 강의지원시스템 비정형 로그 파일을 처리하여 학생별 로그인 횟수 및 시간과 같은 정형 데이터를 생성한다.
- 분석 계층: 본 계층에서는 대규모 데이터로부터 새로운 패턴을 찾고 해석하여 새로운 통찰력을 확보한다. 예를 들면, 중도탈락 가능성이 있는 학생을 조기 발견하기 위한 예측 모델을 생성한다.
- 시각화 계층: 본 계층에서는 빅데이터 분석 결과를 쉽게 이해할 수 있도록 시각화 한다. 예를 들면, 새로운 학생 정보를 예측 모델에 입력하여 학생의 중도 탈락 여부를 시각화 한다.

(그림 2)는 모델 생성 및 배포 과정을 설명한다. 평가 결과가 사용자가 원하는 임계값 이상일 경우 모델을 배포하고 새로운 학생 정보 데이터를 입력하여 결과를 예측한다.


(그림 2) 모델 생성 및 배포

## 4. 모델 생성 기법

학생의 중도탈락 여부를 예상하기 위해서는 특정 데이터를 사전에 정해진 기준에 따라 몇 가지 카테고리로 분류하는 기법을 사용해야 한다. 본 논문은 로지스틱 회귀분석과 다층 퍼셉트론 기법을 사용한다.

로지스틱 회귀분석 기법은 지도학습에 속하는 대표적인 분류 알고리즘으로 종속 변수와 독립 변수 간의 관계를 유추하는 것을 목적으로 한다. 로지스틱 회귀분석 결과 값 중 하나인 오즈비를 이용하여 어느 독립 변수가 종속 변수에 얼마만큼 영향을 주는지 파악할 수 있다. 반면 다층 퍼셉트론 기법은 뉴런을 통한 뇌 학습 기능을 모델링한 알고리즘으로 단층 신경망 모델의 단점을 보완하기 위해 입력층과 출력층 사이에 여러 개의 은닉층을 사용하는 모델이다. 최근 다층 퍼셉트론 기법을 사용한 인공 신경망을 심층 신경망(Deep Neural Network)이라고 하며 심층 신경망을 학습하기 위해 고안된 특별한 알고리즘을 딥러닝(Deep Learning)이라고 한다. 보통 다층 퍼셉트론 기법은 로지스틱 회귀 분석 기법에 비해 예측 성능이 우수하지만 어느 입력 값이 출력 값에 얼마만큼 영향을 주는지 해석하기 어렵다. 따라서 본 논문에서는 실제 예측 시에는 다층 퍼셉트론 기법을 사용하고 어느 데이터를 집중 관리할지를 판단 시에는 로지스틱 회귀 기법을 사용 한다.

## 5. 모델 생성 및 평가

본 논문은 중도탈락 예측 모델 생성을 위한 데이터를 수집한 후 익명화 처리하였다. 그리고 중도탈락에 영향을 줄만 한 항목을 구분하여 <표 1>과 같이 종속변수와 독립변수만을 모델 생성 데이터로 제한하였고 데이터를 정제하였다. 그리고 정제 데이터는 트레이닝 데이터와 테스트 데이터로 구분하여 사용하기 위해 7:3 의 비율로 분리되었다.

<표 1> 변수 정의 및 데이터 정제 방법

| 변수 종류 | 테이블 | 항목 | 데이터 정제 방법 |
|---|---|---|---|
| 종속 변수 | 학적 기본 정보 | 학적상태 | 학적상태가 총 4 가지 상태(제적생, 수료(미졸), 휴학생, 재학생)가 존재하나 제적생과 재학생인 것만을 대상으로 함 |

| | | | |
|---|---|---|---|
| 독립변수 | 학기별 성적 | 평점평균 | 학생별 학차가 다르므로 평균 학점으로 변환하여 사용 |
| | 학적 기본 정보 | 연령 | |
| | | 학차 | |
| | | 성별 | |
| | | 국적 | 한국인과 외국인 구분 |
| | | 보호자주소 | 대구, 경북, 경남 지역과 그 외 지역으로 구분 |
| | 기타 활동 정보 | 비교과 취득점수 | |
| | | 봉사활동 참여횟수 | |
| | | 비교과 만족도 참여횟수 | |
| | | 학과만족도 참여횟수 | |
| | | 학생상담센터 상담횟수 | |
| | | 신입생캠프이수여부 | |
| | | 기숙사입사이력 | |
| | | 동아리활동이력 | |

본 논문은 중도탈락 예측 모델 생성 시 사용되는 독립변수의 특성 및 개수에 따라 평가 결과의 변화 정도를 알아보고자 다음과 같이 다섯 가지 케이스로 독립변수를 사용하였다. Case #5 는 로지스틱 회귀분석 모델 해석 결과 중도탈락 여부에 크게 영향을 주지 않는 독립변수를 Case #4 에서 제외한 케이스이다.

- Case #1: 평균학점
- Case #2: 평균학점, 연령, 학차, 성별
- Case #3: 평균학점, 연령, 학차, 성별, 기숙사입사이력, 동아리활동이력
- Case #4: 평균학점, 연령, 학차, 성별, 기숙사입사이력, 동아리활동이력, 국적, 보호자주소, 비교과 취득점수, 봉사활동 참여횟수, 비교과 만족도 참여횟수, 학과만족도 참여횟수, 학생상담센터 상담횟수, 신입생캠프이수여부, 인성캠프이수여부
- Case #5: 평균학점, 연령, 학차, 성별, 동아리활동이력, 비교과 취득점수, 봉사활동 참여횟수, 비교과 만족도 참여횟수, 학과만족도 참여횟수, 신입생캠프이수여부

본 논문은 Spark[5] 환경에서 로지스틱 회귀분석과 다층 퍼셉트론 기법을 프로그래밍하여 각각의 모델을 생성하였다. 그리고 두 모델을 평가하기 위해 다음과 같은 정확도(accuracy), 재현율(recall), 정밀도(precision)에 관한 식을 사용하였다.

$$\text{Accuracy} = \frac{tp + tn}{tp + tn + fp + fn}$$

$$\text{Recall} = \frac{tp}{tp + fn}$$

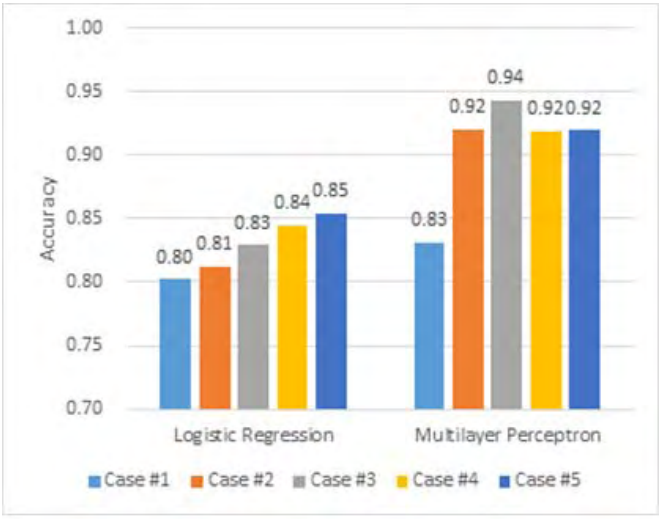
$$\text{Precision} = \frac{tp}{tp + fp}$$

위 식을 이해하기 위해 다음 <표 2>를 참고 하면 tp 는 실제 중도탈락 하는 학생을 중도탈락 할 것으로 예측한 경우인 반면 fp 는 중도탈락 하지 않은 학생을 중도탈락 할 것으로 예측한 경우이다. 여기서 fp 를 1 종 오류라고 한다. fn 는 실제 중도탈락 하는 학생을 중도탈락 하지 않을 것으로 예측한 경우인 반면 tn 는 중도탈락 하지 않은 학생을 중도탈락 하지 않을 것으로 예측한 경우이다. 여기서 fn 를 2 종 오류라고 한다. 예측 모델이 실제 진리를 모두 맞추려면 tp 와 tn 횟수만 존재해야 한다. 즉, 이 때 정확도가 1 이 된다. 그런 의미에서 재현율은 실제 중도탈락 하는 학생 중에서 2 종 오류의 정도를 알 수 있는 측정치이고 정밀도는 중도탈락 할 것으로 예측한 학생 중에 1 종 오류의 정도를 알 수 있는 측정치이다.

<표 2> Confusion Matrix

| | | 실제 진리 | |
|---|---|---|---|
| | | 참 (중도탈락 함) | 거짓 (중도탈락 안함) |
| 예측 결과 | 참 (중도탈락 함) | tp (true positive) | fp (false positive) (1 종 오류) |
| | 거짓 (중도탈락 안함) | fn (false negative) (2 종 오류) | tn (true negative) |

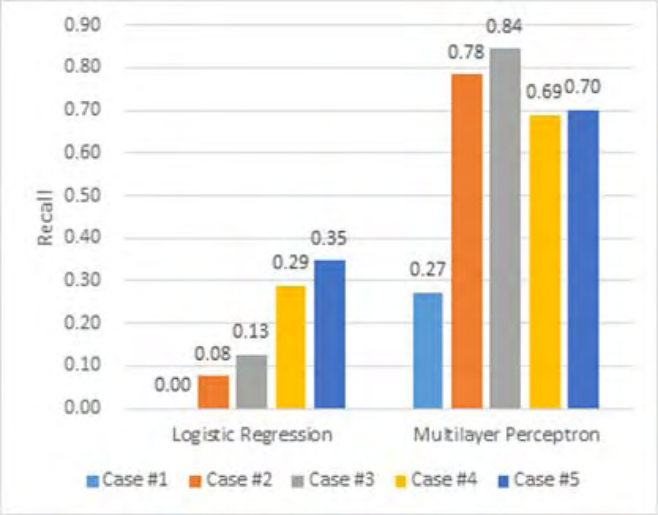
(그림 3)은 로지스틱 회귀 모델과 다층 퍼셉트론 모델을 정확도 관점에서 비교하기 위한 그림이다. 다층 퍼셉트론 모델의 정확도가 로지스틱 회귀 모델의 것에 비해 상대적으로 높으며 특히 Case #3 에서 다층 퍼셉트론 모델을 사용하여 예측할 경우 0.94 의 정확도를 보인다. 로지스틱 회귀분석 모델은 사용한 독립변수를 늘릴수록 정확도는 증가하는데 반해 다층 퍼셉트론 모델은 그렇지 않다.




(그림 3) 정확도 관점 비교

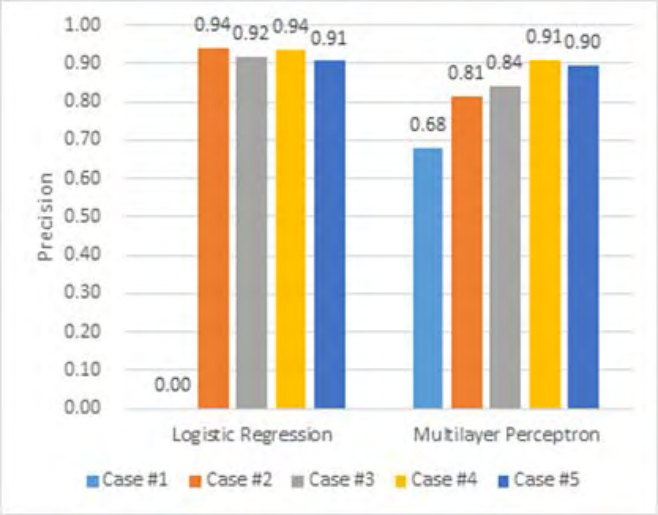
(그림 4)는 로지스틱 회귀 모델과 다층 퍼셉트론 모델을 재현율 관점에서 비교하기 위한 그림이다. 다층 퍼셉트론 모델의 재현율이 로지스틱 회귀 모델의 것에 비해 상대적으로 높으며 특히 Case #3 에서 다층

퍼셉트론 모델을 사용하여 예측할 경우 0.84 의 재현율을 보인다. 낮은 재현율은 실제 중도탈락 하는 학생을 중도탈락 하지 않는 것으로 예측할 가능성이 높은 것을 의미하므로 재현율이 최대 0.35 인 로지스틱 회귀분석 모델은 높은 재현율을 요구하는 경우에는 사용하기 어렵다.




(그림 4) 재현율 관점 비교

(그림 5)는 로지스틱 회귀 모델과 다층 퍼셉트론 모델을 정밀도 관점에서 비교하기 위한 그림이다. 로지스틱 회귀 모델의 정밀도는 다층 퍼셉트론 모델의 것에 비해 상대적으로 높다. 이는 다층 퍼셉트론 모델이 로지스틱 회귀 모델에 비해 상대적으로 1 종 오류가 많음을 의미한다.




(그림 5) 정밀도 관점 비교

## 6. 결론 및 향후 계획

본 논문은 학생 중도탈락을 예방하기 위해 교육 데이터 분석을 통해 중도탈락 예측 모델을 생성하였고 이를 평가하였다. 중도탈락 예측 모델 생성 기법으로 로지스틱 회귀분석과 다층 퍼셉트론 기법을 사용하였다. 평가 결과, 중도탈락 예측 모델을 위한 모델 생성 기법으로 다층 퍼셉트론 기법을 사용하는 것이 로지스틱 회귀분석 기법에 비해 적합하며 다층 퍼셉트론 모델의 최대 정확도는 0.94 이다.

계속해서 다층 퍼셉트론 모델의 성능 향상을 위한 연구와 생성된 모델을 실제 예방 시스템에 적용 구축하는 계획이 있다.

## 참고문헌

[1] 여유진. “한국에서의 교육을 통한 사회이동 경향에 대한 연구, 보건사회연구.” 보건사회논집 28.2 (2008): 53-80.

[2] 김선영. “지방대학 학생들의 학교생활유지에 영향을 미치는 요소 연구.” 교육학연구 51.4 (2013): 27-55.

[3] 정제영, 선미숙, 정민지. “대학생의 중도탈락에 영향을 미치는 대학수준 요인 분석.” 아시아교육연구 16 (2015): 57-76.

[4] 송미영. “지능형 학생중도탈락 예방시스템을 위한 중도탈락자 예측모형 개발.” 한국컴퓨터정보학회논문지 22.10 (2017): 9-17.

[5] Apache Spark, https://spark.apache.org/

[6] Manhães, Laci Mary Barbosa, Sérgio Manuel Serra da Cruz, and Geraldo Zimbrão. “WAVE: an architecture for predicting dropout in undergraduate courses using EDM.” Proceedings of the 29th Annual ACM Symposium on Applied Computing. ACM, 2014.

[7] Guarín, Camilo Ernesto López, Elizabeth León Guzmán, and Fabio A. González. “A model to predict low academic performance at a specific enrollment using data mining.” IEEE Revista Iberoamericana de tecnologias del Aprendizaje 10.3 (2015): 119-125.

[8] Costa, Evandro B., et al. “Evaluating the effectiveness of educational data mining techniques for early prediction of students' academic failure in introductory programming courses.” Computers in Human Behavior 73 (2017): 247-256.