敎育方法硏究 제33권 제1호 (2021. 2. 28):pp.29~46
http://dx.doi.org/10.17927/tkjems.2021.33.1.29

머신러닝 기반의 학업성취 예측 모형 탐색: 대학의 오프라인 수업을 중심으로*

이 현 우† (상명대학교)
이 종 문 (상명대학교)
차 윤 미 (상명대학교)

〈요 약〉

본 연구는 머신러닝 알고리즘을 활용하여 일반 대학의 오프라인 수업 맥락에서 학습관리시스템에 축적되는 학습활동 관련 데이터를 활용하여 개별 강좌에서 학습자의 학업성취 수준을 예측하는 모형 개발의 가능성을 탐색해 보고, 이를 바탕으로 선제적으로 위기의 학습자를 판별하는 시스템 개발에 대한 시사점을 제시하고자 하였다. 이를 위해 A대학의 2018학년도 2학기에 개설된 115개 강좌의 수강생인 3,500명의 학업성취도와 Moodle기반의 학습관리시스템 내의 로그 데이터, 출석데이터를 분석하였다. 예측 분석의 대상은 개별 강좌에서 받은 성적 등급을 기준으로 A+ 또는 A를 받은 '가' 집단 1,030명(29.46%), B+ 또는 B를 받은 '나' 집단 1,315명(37.57%), C+ 이하를 받은 '다' 집단 1,155명(33.00%)이었다. Gradient Boosting 모델을 적용하여 예측모형을 개발하여 예측한 결과 학기 시작 후 7주차를 기준으로 저성과 학습자라고 할 수 있는 '다' 집단에 대한 예측 성능이 재현율 72.86%, 정밀도 65.05%로 모델의 판별력이 나타났다. 본 연구를 통해 학습관리시스템의 활동 데이터를 이용한 머신러닝 모델이 개별강좌에서 학습자의 학업성취를 예측하는 유용한 모형임을 확인하였다.

주제어 : 머신러닝, 학습분석학, 학업성취, 예측모형, 행동로그, Gradient Boosting

* 본 연구는 2020학년도 상명대학교 교내연구비를 지원받아 수행하였음.
† 교신저자 : 이현우, 상명대학교, hwl@smu.ac.kr

Ⅰ. 서 론

최근 들어 고등교육에서는 학생들의 학업성취 예측을 통해 학업적으로 위기에 처한 학생들을 사전에 선별하여 이탈하지 않도록 예방하고자 하며, 특히 학령인구의 감소 추이에 따라 중도탈락률을 개선하여 재학생 유지를 위한 노력을 기울이고 있다(이은정 외, 2020). 교육학에서는 학습자의 학습과정을 모니터링하고 성취의 정도와 학습 장애를 예측하여 적시에 교육적 처방을 제공(Johnson et al., 2011)하는 빅데이터를 활용한 학습분석이 큰 보탬이 될 것으로 기대하고 있다.

학습분석 연구들의 최근 경향은 학생들을 대상으로 한 설문조사를 통한 데이터 수집방식이 아닌 학사관리시스템, 학습관리시스템 등의 자동으로 축적되는 데이터를 활용하여 분석하고 예측하고자 한다(조명희 외, 2018; 정영란, 2020).

개별 강좌 수준에서의 연구들을 살펴보면 학습관리시스템의 활용도가 높은 온라인 강좌를 중심으로 연구가 되어 왔다. 온라인학습환경에서 학습성과를 예측한 연구를 살펴보면, 학습시점 간격의 규칙성이 학업성취를 예측할 수 있는 중요한 변인으로 제시되고 있고(조일현, 김윤미, 2013), 또한 학습자–학습자/학습자–교수자 상호작용, 학습자활동 빈도, 참여의 정도가 학습성과를 예측할 수 있는 변인으로 제시되었다(Agudo–Peregrina et al., 2012).

일반 대학의 오프라인 강좌[1]를 대상으로 한 연구(Gašević et al., 2015; Kang, 2017)는 매우 제한적인데, 이미 국내 대학에 학습관리시스템의 보급이 확대되었고, 대학 내 강좌에서 이러닝, 블랜디드 러닝 교과목 외 오프라인 수업에서도 학습관리시스템을 활용하는 비율이 증가하였다. 이렇듯 학습관리시스템에 학업성취를 예측할 수 있는 학습활동 데이터의 축적이 많아졌기 때문에 오프라인 강좌에서도 축적된 데이터를 활용하여 학업성취를 예측하는 연구의 기반이 마련되었다. 또한, 일반 대학에서는 이러닝, 블랜디드 러닝의 강좌에 비해 오프라인 수업의 비중이 월등히 높다는 점을 고려했을 때, 학업성취 예측에 관한 연구가 오프라인 강좌로도 확대될 필요가 있다.

한편, 최근 중도탈락학생 및 학업성취도를 예측하는 학습분석연구에서 머신러닝(Machine Learning: 기계학습)기법을 활용하는 사례들을 찾아볼 수 있다. 예를 들면, Kang(2017)은 UCLA 대학에서 한 명의 교수자에 의해 진행된 10개 강좌의 학습관리시스템의 모든 로그데이터를 활용하여 Support Vector Machines을 통해 예측모형을 개발하였다. 비록 한 명의 교수자의 강좌만을 분석한 제한점이 있으나 위기의 학습자를 비교적 정확하게 찾아내는 모형을 만드는 성과를 냈다. 컴퓨터가 대규모의 자료를 활용하여 스스로 학습할 수 있도록 모형을 개발하는 머신러닝은

1) “일반대학”이란 『고등교육법』 제2조, 제1호부터 제4호까지 및 제6호부터 제7호의 규정에 해당하는 학교를 의미하고, 오프라인 수업이란 동법 제22조에 따라 대면수업의 보조 수단(수업자료 탑재, 질의 · 응답, 토론 등)으로서 방송 · 정보통신 매체 등을 활용하는 수업을 의미한다.

크게 지도(supervised) 학습과 비지도(unsupervised) 학습으로 구분된다(Murphy, 2012). 지도학습은 종속변인이 있는 경우 이미 수집된 과거의 자료를 학습하여 예측하는 것을 목적으로 하고, 비지도학습은 종속변인이 정해져 있지 않은 경우 입력 데이터의 숨겨진 패턴 혹은 고유한 구조를 발견하는 것을 목적으로 한다. 따라서 머신러닝은 다양하고 방대한 데이터에서 학습 알고리즘을 통해 규칙성을 찾고 예측이 가능하여(Bzdok et al., 2018) 학습관리시스템에 축적된 학습자들의 다양한 학습활동 데이터의 패턴과 학습성취와의 관계성을 예측할 수 있을 것이다.

이에 본 연구에서는 일반대학의 오프라인 수업 맥락에서 학습관리시스템에 축적되는 학습자의 학습활동 관련 데이터와 직전 학기 성취도를 활용하여 학업 성취 수준을 예측하는 모형을 머신러닝 기법을 통해 개발하고, 그 성능을 검증하고자 한다. 본 연구의 구체적인 연구 문제는 1) 학기 중 3주차, 7주차, 12주차에 학습자의 강좌별 최종 성적을 예측하는 것이 가능한가? 2) 예측 시점에서 학습자의 최종 성적에 영향을 미치는 변수는 무엇인가? 로 설정하였다. 본 연구는 개별 강좌에서 위기의 학습자를 선제적으로 판별할 수 있는 시스템 개발에 대한 시사점을 제시하고자 하였다.

Ⅱ. 이론적 배경

1. 학습분석학에서의 학업성취 예측 요인

조명희 외(2018)는 학사관리시스템에 저장된 데이터, 예를 들면 이수 학기, 이전 학사경고 회수, 도서대출 회수, 직전 학기 평점 등을 포함하여 학기 시작 후 8주차에 로지스틱 회귀분석을 통해 학사경고자를 예측하는 모형을 개발하였다. 이들의 연구가 일반대학에서의 학사경고를 예측하는 모형이었다면, 정영란(2020)은 사이버대학의 학습관리시스템과 학사관리시스템의 데이터, 예를 들면 장학 여부, 수강과목 수, 학습활동 참여, 출결, 평균 평점 등을 활용하여 로지스틱 회귀분석을 통해 신입생의 중도탈락을 예측하는 모형을 제시하였다. 두 연구 모두 추가적인 설문 조사 없이 대학이 보유하고 있는 데이터를 활용하여 중도탈락을 예측하는 모형을 개발하였다는 점에서 의의가 있다.

한편, 개별 강좌에서 학습자의 성취를 예측하는 연구로 조일현과 김윤미(2013)는 기업교육의 이러닝 맥락에서 학습시점 간격의 규칙성이 학업 성취의 예측 요인이라고 밝히고 있다. 일반대학의 교양 e-러닝 강좌의 중도탈락 예측 모형을 개발한 유지원(2014)은 중도탈락 예측 요인으로 출석과 총 학습시간을 도출하였다. 또한 Agudo-Peregrina와 동료들(2012)의 온라인 학습과정에 대한 연구에서는 학습자-학습자 상호작용, 학습자-교수자 상호작용 등이 학업 성취도를 높이는데 영향을 주는 요인으로 제시하였다. 특히 이들은 학습자의 적극적 참여, 학습자 활동(퀴

즈, 그룹/개인 과제, 설문지 등)의 빈도가 학습 성과를 예측할 수 있는 변인이라고 주장하였다.

앞의 연구들이 온라인 수업 맥락에서 학업 성취를 예측하였다면, 이현우, 이종문, 차윤미(2019)는 오프라인 수업으로 진행된 일반대학의 강좌를 대상으로 학습관리시스템에 축적된 로그 데이터를 분석하여 학업성취 수준에 따라 학습자의 학습활동에서 유의미한 차이가 있음을 확인하였고, 특히 학습관리시스템에서 학습상태 확인과 읽기 활동은 학기 시작 1주차부터 학업성취 수준에 따라 유의미한 차이가 있음을 밝혔다.

이상의 선행연구들은 학사관리시스템과 학습관리시스템에 자동으로 축적되고 있는 데이터들을 활용하고 있다는 공통점이 있고, 개별 강좌의 학업성취를 예측하는 연구들은 주로 온라인 수업 맥락에서 이루어져 왔으나 점차 오프라인 수업으로 연구범위를 확대할 수 있음을 확인하였다.

2..머신러닝 기법을 활용한 학업성취 예측 요인

학습분석 연구 중 머신러닝 기법을 활용하여 학업성취를 예측한 연구를 살펴보면, 김연희, 임수진(2020)은 학습관리시스템과 연동할 학습분석시스템을 구축하고 학습결과를 예측하기 위해 개인요인(성별, 용돈, 관심 교과과정) 학업요인(평가항목에 투여하는 노력의 정도), 행동요인(수업좌석위치, 평균 학습시간)으로 구성된 설문 조사 결과를 활용하였다. 머신러닝의 선형 회귀분석방법을 통해 학습결과를 예측하고자 하였으며, 오차율은 약 8.4%라는 결과를 보였다. 신종호, 최재원(2019)은 신입생 중 학습부진 위험 학생을 조기에 예측하고자 머신러닝 알고리즘 XGBoost를 활용하였다. 예측을 위한 주요 변수로는 성별, 수능언어/수리 백분위와 같은 입학시 제공한 입시정보데이터와 학교를 그만둘 의향, 게임 시간 등의 설문 조사 결과가 반영되었다. 앞서 두 연구는 학사시스템과 학습관리시스템의 데이터보다는 설문조사를 통해 수집된 데이터를 활용하였다. 장기적인 관점에서 설문조사 방식의 데이터 수집 및 처리 방식은 빅데이터 기반의 학습분석 및 머신러닝 기법의 활용을 제한할 수 있는 한계가 있다.

개별 수업단위에서 학습자의 성취를 예측하는 연구로 이정은, 김다솜, 조일현(2020)은 동영상 기반의 학습환경에서 학습자의 행동로그를 학습기간 중 행동시간(총 일시정지시간, 총 재생시간, 코멘트 작성시간 등)과 행동빈도(북마크 및 코멘트 열람 빈도, 일시정지 빈도, 재생 빈도 등)로 구분하였고, 학업성취 지표로는 동영상 학습 사전, 사후시험 점수를 활용하였다. 머신러닝의 분석모델은 K-근접 이웃, 인공신경망, 서포트 벡터 머신, 랜덤 포레스트를 활용하였으며, 모든 모델에서 학업성취에 가장 영향을 많이 미치는 변수는 동료 학습자가 작성한 코멘트를 확인하는 행동이었고, 학업성취를 높게 예측하는 행동은 슬라이드 조정 빈도, 일시 정지 빈도, 총 일시정지 시간으로 적극적인 학습환경을 조성하고자 하는 행동이었다. Kondo, Okubo와 Hatanaka(2017)는 동영상 총 학습횟수, 야간시간 총 운영횟수, 로그인 횟수, 과제 확인 횟수, 과

제제출 완료 횟수, 학습관리시스템의 총 로그인 시간 등의 학습관리시스템 활동 데이터와 학업성적, 오프라인 출석률을 활용하였으며, 랜덤 포레스트 모델을 통해 다른 변수를 제외한 학습관리시스템 로그 데이터만으로 3주차 말에 위기학생을 약 40% 판별하는 결과를 얻었다. 조헌국(2018)은 신입생을 대상으로 운영한 이러닝 과학교양 강좌에서 학습활동 관련 로그데이터를 사용한 것이 아닌 이러닝 출결, 과제 제출현황, 중간고사 및 기말고사의 성적데이터를 활용하였다. 분석기법으로는 k-근접 이웃, 서포트 벡터 머신, 의사결정 나무, 랜덤 포레스트, 그래디언트 부스팅, 인공 신경망 등 6가지의 분석을 진행하였으나 서포트 벡터 머신 모형이 가장 예측력이 높았고, 학업성취에 중요하게 영향을 미치는 요인으로 과제, 기말고사, 중간고사, 출결 순으로 나타나, 출결정보만으로는 유의미한 학습이 이루어지지 않음을 확인하였다. 이로써, 개별 강좌 연구에서는 이러닝 강좌 중심으로 연구가 되었다는 것을 확인할 수 있다.

선행연구들을 살펴봤을 때 머싱러닝 모델을 활용한 연구들은 데이터 수집 및 활용이 용이한 이러닝 맥락에서 진행되었음을 확인할 수 있다. 일회성의 연구가 아닌 대학 내에 학생들의 학업성취를 예측하기 위해서는 별도의 설문조사나 데이터 수집이 없이 자동적으로 축적되는 데이터만으로 예측이 이루어질 필요가 있다.

3. 머신러닝 모델 구축

머신러닝 예측모형의 성능 평가는 모형이 산출한 혼동행렬(confusion matrix)의 수치를 이용하여 계산된 재현율(recall), 정밀도(precision), 정확도(accuracy)를 사용하고(Kuhn & Hohnson, 2013), 추가로 F1 score를 사용할 수 있다. <표 1>과 같이 예측치와 실제치를 행렬로 표현하고 각각의 수치를 이용하여 성능을 계산할 수 있는 혼동행렬은 다음과 같다.

〈표 1〉 혼동행렬

		예측 분류	
		양성	음성
실제 분류	양성	true positive (TP)	false negative (FN)
	음성	false positive (FP)	true negative (TN)

먼저 재현율(recall)은 실제값이 참일 때 예측값이 참인 비율로, 혼동행렬 상에서 TP/(TP+FN)로 계산한다. 반면에 정밀도(precision)는 예측값이 참일 때 실제값이 참인 비율로 TP/(TP+FP)로 계산한다. 중도탈락을 예로 들면, 재현율은 중도탈락자 중 중도탈락으로 예측된 학생의 비율이

고, 정밀도는 중도탈락할 것으로 예측된 학생 중 실제로 중도탈락한 학생의 비율이다(Davis & Goadrich, 2006).

〈표 2〉 머신러닝 예측모형의 결과

구분	결과
True Positive(TP)	관심 범주(저성취자)를 정확하게 분류함. 실제로 저성취 학생을 예측모형도 저성취 학생으로 제대로 분류함.
Tru Nagative(TN)	관심 범주(저성취자)가 아닌 것을 정확하게 분류함. 실제로 저성취자가 아닌 학생을 예측모형도 저성취자가 아닌 학생으로 제대로 분류함.
False Positive(FP)	관심 범주(저성취자)로 잘못 분류함. 실제로 저성취자가 아닌 학생을 예측모형은 저성취 학생으로 잘못 분류함
False Negative(FN)	관심 범주(저성취자)가 아닌 것으로 잘못 분류함. 실제로 저성취 학생을 예측모형은 저성취자가 아닌 학생으로 잘못 분류함

출처: 신종호, 최재원(2019), 학습분석 기반 대학 신입생 대상 학습부진 위험학생 조기예측 모델 개발 및 군집별 특성 분석. 교육공학연구 수정

재현율을 높이면 더 많은 중도탈락자를 찾을 수 있지만 실제로는 중도탈락하지 않은 학생들을 과대 표집할 가능성이 있고, 정밀도를 높이게 되면 실제로 중도탈락하는 학생들을 과소 표집하게 될 우려가 있다. 이러한 불균형문제를 해결하기 위해 재현율과 정밀도를 이용한 F1 score를 계산하여 성능을 평가한다(최환석 외, 2020). 정밀도와 재현도가 모두 높으면 F1 Scoure가 높고, 둘 중 한쪽이 낮거나 모두 낮으면 F1 Scoure가 낮아지도록 고안된 수식으로(신종호, 최재원 2019), F1 score 계산 공식은 다음과 같다.

$$\text{F1 Score} = 2 \times \frac{\text{정밀도} \times \text{재현율}}{\text{정밀도} + \text{재현율}}$$

Ⅲ. 연구방법

1. 연구대상

본 연구는 A대학의 2018학년도 2학기에 개설된 오프라인 교과목 중, 학습관리시스템에 매주 교수학습 활동이 등록된 115개 강좌를 연구대상으로 선정하였다. 해당 교과목의 수강생인

3,500명의 학업성취도와 Moodle 기반의 학습관리시스템 내의 로그 데이터와 출석데이터를 분석하였다.

2. 분석데이터

일반적으로 대학의 시스템에 축적되는 데이터의 테이블이 일관적이지 않고 데이터 관리가 담당 업무별로 분산되어 있어 데이터 수집에 어려움이 있는 것이 사실이다. 또한 분산되어 있는 데이터를 수집하여도 수집된 데이터를 입력할 수 있도록 정리하는 전처리과정에 과도한 시간과 노력이 소요되는 문제가 있다.

이에 본 연구에서는 전처리 문제를 해결하기 위해 학습관리시스템에서 자동으로 축적되는 행동로그 데이터를 변수로 활용하였고, 전처리 과정으로 데이터를 정규화하는 과정이 포함되었으나, 매우 단순한 작업으로 전처리에 소요되는 시간과 노력을 큰 폭으로 감축할 수 있도록 하였다.

본 연구에서는 Moodle 기반의 학습관리시스템 내 모든 학습활동을 수집하고, 전처리를 통해 학습자의 동일한 행위가 중복 기록되는 것을 방지하였다. 이러한 로그 데이터는 활동의 속성에 따라 세 가지의 활동 즉, 학습상태 확인, 읽기, 쓰기로 구분하였다(<표 3> 참조). 학습상태 확인은 학습자가 본인의 학습상태를 모니터링하는 행위로 새로운 학습활동이 등록되었는지를 확인하거나 성적부, 출석부, 과제제출 상태 등을 확인하는 행동을 기록한 데이터다. 읽기활동은 학습자가 강좌에 등록된 학습자료, 동영상, 외부 학습자원, 파일 등을 접속한 로그 데이터를 말한다. 쓰기활동은 학습자가 토론, 과제, 퀴즈, 게시판 등에 글을 작성한 행동을 기록한 데이터이다.

학습관리시스템에 기록되는 학습자의 활동에 대한 로그의 양은 교과목의 특성, 교수자의 학습관리시스템 활용능력, 교수 활동 등에 영향을 받아서 로그 데이터의 절대값이 교과목별로 차이가 있으므로 주차별로 누적된 학습자 활동의 상대적인 위치값을 분석에 활용하기 위해 로그 데이터의 절대값을 정규화[2]하여 분석하였다(이현우 외, 2019). 학업성취도 데이터의 경우에는 독립변인으로 포함된 학업성취도 데이터는 직전 학기의 평점을 반영하였고, 종속변인으로서의 학업성취도는 개별 강좌에서 받은 성적 등급을 기준으로 '가' 집단(A, A+), '나' 집단(B+, B), '다' 집단(C+ 이하)의 세 집단으로 구분하였다.

2) 정규화 공식=(해당값-최소값)/(최대값-최소값)

〈표 3〉 활동 속성에 따른 세부 로그 기록 분류

속성	세부 로그
학습상태 확인	• 학습활동 목록(게시판, 과제, 동영상, 설문조사, 토론, 퀴즈, 파일, 투표 등) 확인 • 강의실 내 본인 로그 확인 • 학습진도현황 확인 • 출석부 확인 • 강의실 타임라인 확인
읽기	• 게시판, 과제, 동영상 열람 • 토론 게시글 열람 및 게시글 모아보기 • 파일, 폴더 등 자료 접속 및 열람
쓰기	• 게시글 작성 및 수정 • 댓글 작성 및 수정 • 토론 게시글 작성 및 수정, 답변 작성 및 수정 • 과제 제출 및 수정 • 설문조사, 투표, 퀴즈, 동료 평가 등 응답 제출

출처: 이현우, 이종문, 차윤미(2019), 오프라인 강좌에서 대학생의 학업성취에 따른 학습관리시스템 활동의 차이 분석. 교육정보미디어연구 재정리

3. 분석과정

머신러닝 모델은 일반적으로 데이터 세트를 훈련용 데이터 세트(Train Set)와 테스트용 데이터세트(Test Set)로 나눈다. 두 데이터 세트의 비율을 조정하고 검증하는 과정을 반복하여 최적화하였으며, 본 연구에서는 전체 3,500명의 데이터를 학습데이터 70%(2,450명), 시험데이터 30%(1,050명)로 구분하여 모델링하였다.

본 연구에서는 회귀(Regression)분석 또는 분류(Classification)분석을 수행할 수 있고, 여러 분류 모델을 생성하고 그 결과를 조합하여 분석을 진행하는 앙상블 학습(Ensemble Learning)에 해당하는 Gradient Boosting Algorithm(GBA) 기법을 사용하였다. Boosting의 장점은 여러 개의 분류기가 순차적으로 학습을 수행하며 다음 분류기의 가중치(Weight)를 부여하는 방식으로 각 분류기의 잘못된 예측을 보정하여 학습과 예측을 진행하는 방식으로 예측성능이 우수하다. 머신러닝에서 조정할 수 있는 GBA의 하이퍼 파라미터(Hyper parameter)는 max_depth(트리의 최대 깊이), max_features(최적 분할을 위해 고려할 최대 feature 개수) 등 Tree와 연관된 하이퍼 파라미터와 loss(경사하강법의 cost function), learning_rate(학습을 진행할 때 적용하는 학습률) 등 Boosting에 관한 하이퍼 파라미터가 있는데, 본 연구에서는 분석모델의 성능 향상 및 과적합 방지를 위해 learning_rate와 n_estimators(트리의 개수)를 조정하였다. 모델의 성능평가는 혼동행렬을 출력하여 정확도(Accuracy), 재현율(Recall), 정밀도(Precision)를 확인하였다. 본 연구의 모든 분석은 R Statistical software 3.5.3으로 진행하였다.

Ⅳ. 연구결과

1. 분석데이터의 기술통계

본 연구에서는 2018학년도 2학기에 개설된 오프라인 수업의 수강생 3,500명의 개인변인, 직전 학기 평점, 학습관리시스템의 활동 로그 데이터, 출석 데이터가 투입변수로 사용되었다. 수집된 예측변수의 기술통계는 <표 4>와 같다.

〈표 4〉 변수의 기술통계 (n=3,500)

분류	변수	*M*	*SD*	*min*	*max*	frequency(%)
개인변인	외국인(Y)	–	–	–	–	802 (22.91%)
	외국인(N)	–	–	–	–	2,698 (77.09%)
	성별(M)	–	–	–	–	1,742 (49.77%)
	성별(W)	–	–	–	–	1,758 (50.23%)
이전 학업성취	직전 학기 평점	3.21	0.841	0.0	4.5	
활동 로그[a]	3주차 쓰기 활동	14.306	24.098	0	100	–
	3주차 읽기 활동	26.534	25.028	0	100	–
	3주차 학습상태확인 활동	25.931	23.341	0	100	–
	3주차 출석 데이터(결석)	0.179	0.428	0	3	–
	7주차 쓰기 활동	20.543	25.621	0	100	–
	7주차 읽기 활동	30.416	24.628	0	100	–
	7주차 학습상태확인 활동	30.559	23.404	0	100	–
	7주차 출석 데이터(결석)	0.397	0.821	0	7	–
	12주차 쓰기 활동	23.484	25.664	0	100	–
	12주차 읽기 활동	32.557	24.438	0	100	–
	12주차 학습상태확인 활동	32.513	23.11	0	100	–
	12주차 출석 데이터(결석)	0.687	1.339	0	12	–
학업 성취도	성적집단 (가)	–	–	–	–	1,030 (29.46%)
	성적집단 (나)	–	–	–	–	1,315 (37.57%)
	성적집단 (다)	–	–	–	–	1,155 (33.00%)

a. 활동로그 중 대표적으로 3, 7, 12주차의 기술통계만 제시함.

2. 머신러닝 모델의 성능

개별 강좌에서 학습자의 학업성취도를 조기에 예측하기 위해 학습데이터에 Gradient Boosting 모델을 적용하여 예측모형을 개발하고, 이 모델에 시험데이터를 반영하여 모델 성능을 도출하

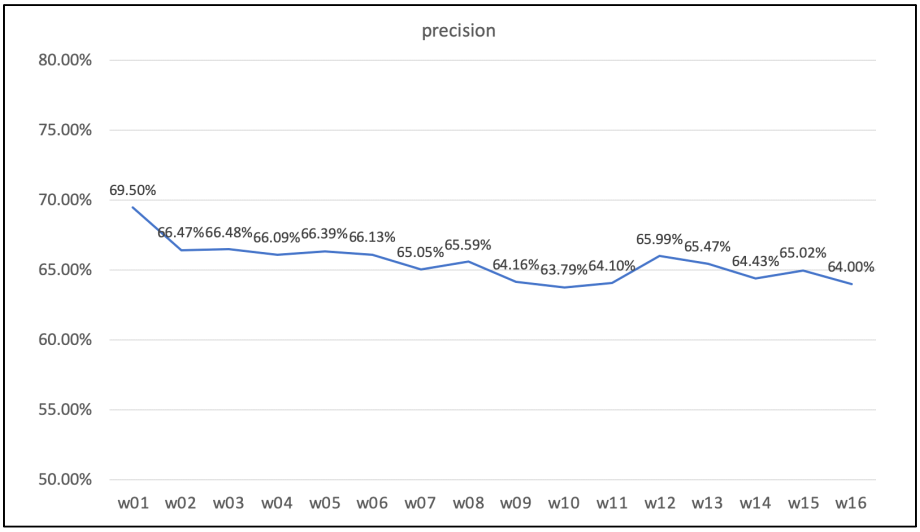
였다. 각 주차별 재현율과 정밀도는 <표 3>에 제시하였다. 주차별 재현율과 정밀도를 살펴보았을 때, 학습위험군이라고 할 수 있는 C 그룹에 대한 예측이 중간고사 이전인 7주차에서 재현율 72.86%, 정밀도 65.05%로 양호한 수준을 나타내고 있다.

〈표 3〉 주차별 학업성취 예측의 재현율과 정밀도

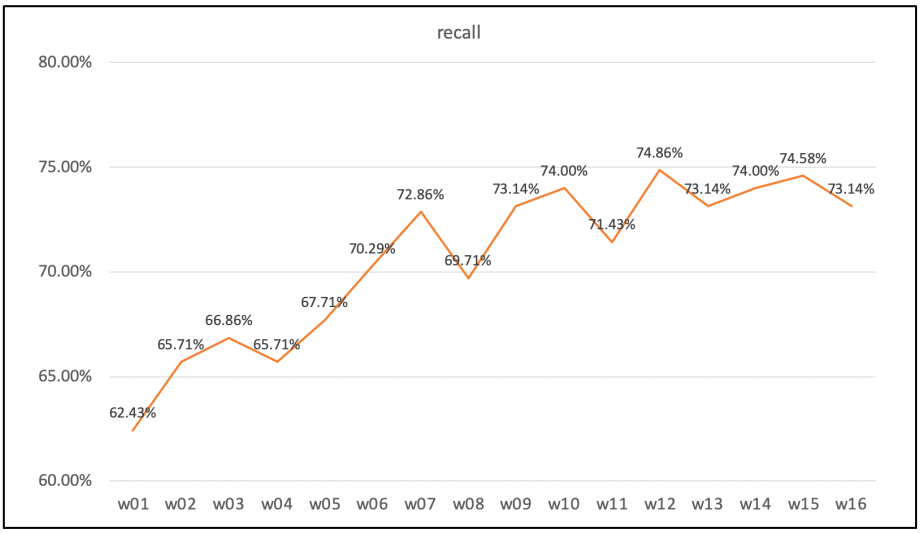
예측 성취도		실제 성취도			정밀도
		A를 받은 학습자	B를 받은 학습자	C를 받은 학습자	
1주차	A로 예측된 학습자	141	110	23	51.46%
	B로 예측된 학습자	142	206	110	44.98%
	C로 예측된 학습자	25	72	221	69.50%
	재현율	45.78%	53.09%	62.43%	
2주차	A로 예측된 학습자	113	52	8	65.32%
	B로 예측된 학습자	164	255	112	48.02%
	C로 예측된 학습자	32	84	230	66.47%
	재현율	36.57%	65.22%	65.71%	
3주차	A로 예측된 학습자	113	52	8	65.32%
	B로 예측된 학습자	164	253	108	48.19%
	C로 예측된 학습자	32	86	234	66.48%
	재현율	36.57%	64.71%	66.86%	
4주차	A로 예측된 학습자	122	68	15	59.51%
	B로 예측된 학습자	155	237	105	47.69%
	C로 예측된 학습자	32	86	230	66.09%
	재현율	39.48%	60.61%	65.71%	
5주차	A로 예측된 학습자	113	51	8	65.70%
	B로 예측된 학습자	164	252	105	48.37%
	C로 예측된 학습자	32	88	237	66.39%
	재현율	36.57%	64.45%	67.71%	
6주차	A로 예측된 학습자	113	52	8	65.32%
	B로 예측된 학습자	162	247	96	48.91%
	C로 예측된 학습자	34	92	246	66.13%
	재현율	36.57%	63.17%	70.29%	
7주차	A로 예측된 학습자	138	81	17	58.47%
	B로 예측된 학습자	135	209	78	49.53%
	C로 예측된 학습자	36	101	255	65.05%
	재현율	44.66%	53.45%	72.86%	
8주차	A로 예측된 학습자	118	58	13	62.43%
	B로 예측된 학습자	156	240	93	49.08%
	C로 예측된 학습자	35	93	244	65.59%
	재현율	38.19%	61.38%	69.71%	

예측 성취도		실제 성취도			정밀도
		A를 받은 학습자	B를 받은 학습자	C를 받은 학습자	
9주차	A로 예측된 학습자	142	89	21	56.35%
	B로 예측된 학습자	127	199	73	49.87%
	C로 예측된 학습자	40	103	256	64.16%
	재현율	45.95%	50.90%	73.14%	
10주차	A로 예측된 학습자	135	82	15	58.19%
	B로 예측된 학습자	134	202	76	49.03%
	C로 예측된 학습자	40	107	259	63.79%
	재현율	43.69%	51.66%	74.00%	
11주차	A로 예측된 학습자	134	75	15	59.82%
	B로 예측된 학습자	137	214	85	49.08%
	C로 예측된 학습자	38	102	250	64.10%
	재현율	43.37%	54.73%	71.43%	
12주차	A로 예측된 학습자	133	90	16	55.65%
	B로 예측된 학습자	140	202	72	48.79%
	C로 예측된 학습자	36	99	262	65.99%
	재현율	43.04%	51.66%	74.86%	
13주차	A로 예측된 학습자	117	60	9	62.90%
	B로 예측된 학습자	156	232	85	49.05%
	C로 예측된 학습자	36	99	256	65.47%
	재현율	37.86%	59.34%	73.14%	
14주차	A로 예측된 학습자	129	72	12	60.56%
	B로 예측된 학습자	140	216	79	49.66%
	C로 예측된 학습자	40	103	259	64.43%
	재현율	41.75%	55.24%	74.00%	
15주차	A로 예측된 학습자	135	81	14	58.70%
	B로 예측된 학습자	134	204	76	49.28%
	C로 예측된 학습자	39	103	264	65.02%
	재현율	43.83%	52.58%	74.58%	
16주차	A로 예측된 학습자	149	89	14	59.13%
	B로 예측된 학습자	126	192	80	48.24%
	C로 예측된 학습자	38	106	256	64.00%
	재현율	47.60%	49.61%	73.14%	

[그림 1]과 [그림 2]에서 C 그룹에 대한 예측의 재현율과 정밀도의 추이를 살펴보면 재현율은 7주차까지 상승하다가 16주차까지 유지되는 것을 알 수 있다. 정밀도는 69.0%에서 64%까지로 유지되고 있다.

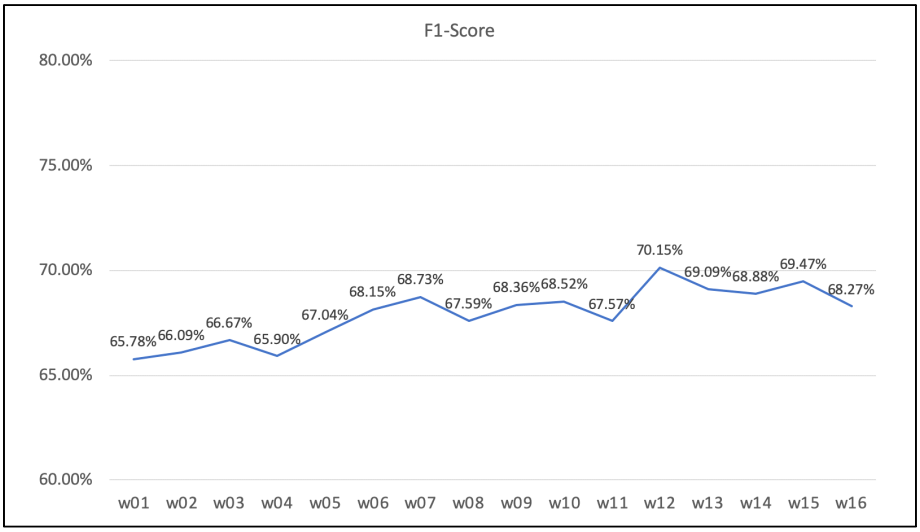


〔그림 1〕 C그룹에 대한 주차별 정밀도 추이



〔그림 2〕 C그룹에 대한 주차별 재현율 추이

<표 2>의 정밀도와 재현율을 활용하여 계산한 F1 Score는 [그림 3]과 같다. 7주차에서 68.73%로 양호한 성능을 보였으며, 12주차에서 70.15%로 가장 높은 수치를 보였다.



〔그림 3〕 주차별 F1 Score 추이

F1 score 추이에서 유의미한 변화가 있는 시점인 3주차, 7주차, 12주차에서 학습자의 학업성취를 예측하는 예측변수의 가중치는 <표 4>와 같다. 학업성취를 예측하는 예측변수 중 상대적으로 가중치가 높은 변수는 직전 학기 평점 그룹이고, 활동로그 중에서는 쓰기활동이 높게 나타났다.

〈표 4〉 주차별 각 예측변수의 가중치(Weight)

	3주차	7주차	12주차
직전 학기 평점 그룹	0.386	0.321	0.282
출석데이터(결석)	0.081	0.139	0.226
개인변인(외국인)	0.072	0.093	0.064
개인변인(성별)	0.046	0.021	0.017
활동로그(쓰기 활동)	0.106	0.122	0.149
활동로그(읽기 활동)	0.089	0.108	0.143
활동로그 (학습활동 확인)	0.088	0.086	0.112

Ⅳ. 결론 및 논의

본 연구는 일반대학의 오프라인 수업에서 학습관리시스템에 자동으로 저장되고 있는 로그데이터와 이전 학기의 성적 데이터를 활용하여 개별 강좌에서의 학습자의 학업성취수준을 예측하는 모형의 가능성을 탐색하였다. 연구결과를 바탕으로 세 가지의 이슈에 대해 논의하고자 한다.

먼저 본 연구를 통해 개별 강좌에서 학습자의 성취를 예측할 수 있는 변인으로 직전 학기까지의 누적 평점이 가중치가 가장 높게 나타났다. 학사경고자를 예측하기 위한 학습분석학적 접근을 했던 조명희 외(2018)의 연구에서는 직전 학기의 총 평점이 학사경고를 예측하는 중요한 변수임을 밝혔고, 사이버대학의 중도탈락을 예측한 정영란(2020)의 연구에서도 평균 평점이 중요한 변인으로 검증되었다. 즉 이전 학기까지의 학업 성취도가 이후의 학습성과를 예측할 수 있는 변인임을 확인할 수 있었고, 이러한 결과를 바탕으로 이전 학기까지의 저성과자들의 학업성과를 높일 수 있는 방안이 처방되어질 필요가 있다.

둘째, 학습관리시스템에 축적된 활동 데이터를 분석에 반영한 결과 학습자 참여 활동이 중요한 예측 변인임을 확인하였다. 학습자-학습자 간 상호작용과 학습자-교수자 간 상호작용을 포함하는 다양한 학습자 활동의 빈도가 학습성과를 예측하는 중요한 변인임을 보고한 Agudo-Peregrina 외(2012)의 연구결과와 유사한 맥락에서 이해할 수 있다. 본 연구를 통해 오프라인 수업에서도 학습자의 활동 빈도가 학습 성과를 예측할 수 있는 중요한 변인임이 밝혀졌고, 특히 쓰기 활동이 중요한 변인이라는 점을 고려하여 교수자들이 학습자의 적극적인 참여를 촉진하려는 노력이 필요함을 강조하고 있다.

셋째, 머신러닝을 활용한 연구가 의학, 건축, 교통, 기계, 기상, 금융 등 자연과학과 공학뿐만 아니라 사회과학에서도 활발하게 진행되고 있다. 반면에 교육분야에서 머신러닝을 활용한 연구는 이제 시작하는 단계에 있는 것으로 판단된다. 예를 들어, KCI 등재지 기준으로 2021년 1월 기준 머신러닝을 키워드로 검색되는 논문의 총수는 678건인데 반해 교육관련 학술지에 실린 논문은 단 9건에 불과하다. 사전에 학습에 어려움을 겪고 있는 학습자를 찾아내고 이들에게 적절한 도움을 제공하기 위해서는 머신러닝을 활용한 연구들이 활발하게 이루어질 필요가 있어 보인다.

머신러닝을 활용한 예측모형의 성능평가와 관련하여 아직은 절대적인 기준을 찾아볼 수는 없다는 한계점이 존재한다. 본 연구의 결과에서 7주차를 기준으로 C그룹을 예측하는 재현율은 72.86%였고, 정밀도는 65.05%였다. 본 연구의 목적이 학습에 어려움을 겪는 학습자를 사전에 찾아내서 도움을 제공하는 것이라면 재현율, 즉 실제로 C 이하를 받은 학생 중 C 그룹에 속할 것으로 예측된 비율이 72%라는 의미는 C 이하를 받은 학생 10명 중 7.2명 정도를 7주차에 찾아낼 수 있다는 의미이다. 반면에 정밀도, 즉 C 그룹으로 예측된 학생 중 실제로 C 이하를 받

은 학생의 비율이 65%라는 의미는 C 그룹으로 예측된 10명 중 6.5명이 실제로 C 이하를 받았다는 의미이고, 이는 곧 3.5명은 실제로는 B 이상의 성적을 받았지만, C 그룹으로 판별되었음을 의미한다. 예측모형의 성능과 관련해서는 일정 수준의 가치 판단이 요구되며 일부 해당하지 않은 학습자들이 포함되더라도 더 많은 위기의 학습자를 찾아내야 한다면 재현율을 높일 필요가 있고, 오히려 판별의 오류로 인해 학습동기에 부정적인 영향을 미칠 수 있다면 위기의 학생을 일부 포함하지 못하더라도 정밀도를 높여야 할 필요가 있다(최환석 외, 2020).

본 연구는 연구의 대상이 된 강좌의 특성, 즉 전공과 교양과목의 구분, 강의 규모, 전공계열 등을 고려하지 않고 분석을 시도하였다는 한계가 있다. 그럼에도 불구하고 머신러닝 기법을 활용하여 일반대학의 개별강좌에서 C 이하의 성적을 받을 것으로 예상하는 학습자를 학기가 시작되고 7주차에 판별해낼 수 있는 모형을 개발하였다는 점에서 의의가 있다. 학습관리시스템에 저장되는 학습자들의 학습 과정을 보여주는 데이터들을 활용하여 학습자들에게 사전에 적절한 처방을 내리는 것이 필요해 보인다. 이러한 맥락에서 본 연구를 기초로 개별 강좌에 대한 저성과자 판별 머신을 개발해나갈 필요가 있고, 특히 개별 강좌에서 판별한 결과를 학습자별로 정리하면 위기의 학습자를 예측하는 것도 가능할 것으로 판단되어 위기의 학습자를 조기에 찾아낼 수 있을 것이다.

또한, 본 연구는 머신러닝 기법을 적용하기 위한 데이터 전처리과정을 최소화하기 위해 학습관리시스템에 자동으로 축적되는 데이터만을 활용하는 유용성도 보여주었다. 이러한 유용성이 향후 머신러닝을 활용한 예측모형 개발을 촉진할 수 있을 것으로 판단되며, 향후 이 모형을 통해 판별된 학습자들에게 제공할 수 있는 교수-학습적 처방에 관한 연구가 이어질 필요가 있다.

참고문헌

김연희, 임수진. (2020). 기계학습을 활용한 대학생 학습결과 예측 연구. **한국콘텐츠학회논문지**, 20(6), 695-704.

신종호, 최재원. (2019). 학습분석 기반 대학 신입생 대상 학습부진 위험학생 조기예측 모델 개발 및 군집별 특성 분석. **교육공학연구**, 35(2), 425-454.

유지원. (2014). 컴퓨터교과교육: 일반대학에서 교양 e-러닝 강좌의 중도 탈락 예측모형 개발과 조기 판별 가능성 탐색. **컴퓨터교육학회논문지**, 17(1), 1-12.

이은정, 송영수, 김지하, 오수현. (2020). 랜덤 포레스트를 활용한 4년제 대학 중도탈락률 예측 요인 탐색: 대학 수준 결정요인을 중심으로. **교육공학연구**, 36(1), 191-219.

이정은, 김다솜, 조일현. (2020). 동영상 기반 학습 환경에서 머신러닝을 활용한 행동로그의 학업성취 예측 모형 탐색. **컴퓨터교육학회 논문지**, 23(2), 53-64.

이현우, 이종문, 차윤미. (2019). 오프라인 강좌에서 대학생의 학업성취에 따른 학습관리시스템 활동의 차이 분석. **교육정보미디어연구**, 25(1), 201-222.

정영란. (2020). 학습분석학 기반의 사이버대학의 중도탈락 예측 분석. **교육방법연구**, 32(2), 205-232.

조명희, 김은진, 이현우. (2018). 학사경고자 예측을 위한 학습분석학적 모형 탐색. **교육공학연구**, 34(4), 877-900.

조일현, 김윤미. (2013). 이러닝에서 학습자의 시간관리 전략이 학업성취도에 미치는 영향: 학습분석학적 접근. **교육정보미디어연구**, 19(1), 83-107.

조헌국. (2018). 머신 러닝을 활용한 이러닝 학습 환경에서의 학습자 성취 예측 모형 탐색. **학습자중심교과교육연구**, 18, 553-572.

최환석, 팽전, 이우섭. (2020). 머신러닝 기반 음식점 추천시스템 설계 및 개발. **한국디지털콘텐츠학회논문지**, 21(2), 259-268.

Agudo-Peregrina, Á. F., Hernández-García, Á. & Iglesias-Pradas, S. (2012). Predicting academic performance with learning analytics in virtual learning environments: A comparative study of three interaction classifications. *2012 International Symposium on Computers in Education*, 1-6.

Bzdok, D., Altman, N., & Krzywinski, M. (2018). Points of significance: Statistics versus machine learning. *Nature Methods, 15*(4): 233-234.

Davis, J., & Goadrich, M. (2006). The relationship between precision-recall and roc curves. *Proceeding Icml '06 Proceedings of the 23rd International Conference on Machine Learning*, 233-240.

Gašević, D., Dawson, S., & Siemens, G. (2015). Let's not forget: Learning analytics are about learning. *TechTrends*, *59*(1), 64-71.

Johnson, L., Levine, A., Smith, R., & Stone, S. (2011). *The 2010 Horizon Report*. Austin, TX: The New Media Consortium. In International Conference on Educational Data Mining (pp. 38–70). Montréal, Québec, Canada.

Kang, E. (2017). *A Pilot Study of Predicting Failing Grades Using Data from UCLA's Learning Management System*. Unpublished doctoral dissertation, UCLA.

Kondo, N., Okubo, M., & Hatanaka, T. (2017, July). *Early detection of at¬risk students using machine learning based on LMS log data*. In 2017 6th IIAI International Congress on Advanced Applied Informatics (IIAI–AAI) (pp. 198–201). IEEE.

Murphy, K. P. (2012). *Machine learning: a probabilistic perspective*. MIT press.

Reigeluth, C. M. (1983). *Instructional design theories and models: An overview of their current status*. Hillsdale, NJ: Lawrence Erlbaum Associates.

◎ 논문접수: 2020. 10. 12 / 수정본 접수: 2021. 01. 11 / 게재승인: 2021. 02. 15

〈Abstract〉

Exploring a Model for Predicting Academic Achievement with Machine Learning for Off-line Courses in Higher Education

Hyeon Woo Lee (Sangmyung University)
Jong Moon Lee (Sangmyung University)
Yoon Mi Cha (Sangmyung University)

The purpose of this study is to explores the possibility of developing a model that predicts the level of academic achievement of college students in individual courses by using data related to learning activities accumulated in the learning management system in the context of face-to-face classes in general colleges using machine learning algorithms. Also, the study aims to present implications for discriminating learners in crisis. Academic achievement, log data in the Moodle-based learning management system, and attendance data of 3,500 students of 115 courses in the fall semester of 2018 at University A were analyzed. In terms of academic achievements, there were 1,030 students in group A (29.46%), 1,315 students in group B (37.57%), and 1,155 students in group C (33.00%). As a result of developing a prediction model by applying the Gradient Boost model, the predictive performance for the C group, which can be said to be a learner in crisis, was good at 72.86% recall and 65.05% precision, based on the 7th week after the start of the semester. Through this study, we confirmed the usefulness of the machine learning model using the activity data of the learning management system to predict learners' academic achievement in individual courses.

Key words : Machine Learning, Learning Analytics, Learning Achievement, Prediction Model, Behavioral Log, Gradient Boosting