

Exploring Factors Predicting Student Dropout in Online Learning: Using Random Forest Model

Hyejin Park (KAIST Global Institute For Talented Education Senior Researcher)

Seokwon Kim (KAIST Global Institute For Talented Education Researcher)

Sunghye Lee[†] (KAIST Global Institute For Talented Education Research Associate Professor)

This study was to investigate meaningful student dropout factors and to examine the differences between dropout and persistent learners in online education program for middle and high school students. Both log data(learner individual characteristics, learning activity data, learning content data) and survey data, which were recorded in the learning management system for 6 weeks out of a total of 12 weeks education period from March to June 2020, were collected from 788 learners who enrolled in the online program offered by K university. To this end, the Random Forest Model technique in Machine Learning(ML) was applied to explore factors predicting dropout, and the independent-samples t-test was carried out to examine the difference in learning behavior between dropout and persistent learners. Findings revealed that the top eight important factors predicting dropout were identified as assignment score, assignment submission, login number, number of eBook learning, number of questions, login interval, login regularity, and self-efficacy. In addition, dropout and persistent learners showed statistical differences in all the top eight factors derived as major factors for the dropout prediction. Further implications and significance of these findings were discussed.

Keywords : online learning, dropout, log data, random forest, machine learning

[†] Correspondence : Sunghye Lee, KAIST, slee45@kaist.ac.kr

온라인 학습자의 중도탈락 예측 요인 탐색: 랜덤 포레스트를 적용하여

박 혜 진 (KAIST 과학영재교육연구원 선임연구원)

김 석 원 (KAIST 과학영재교육연구원 연구원)

이 성 혜[†] (KAIST 과학영재교육연구원 연구부교수)

〈요 약〉

본 연구는 중·고등학생 대상 온라인 교육 프로그램에서 중도탈락 주요 예측 요인을 탐색하고 중도탈락 예측 요인은 중도탈락 학습자와 수강지속 학습자 간에 차이가 있는지 분석하고자 하였다. 이를 위해 본 연구에서는 2020년 3월부터 6월까지 총 12주의 교육 기간 중 6주 동안 K-대학 온라인 교육 학습관리시스템에 기록된 788명의 로그 데이터(학습자 변인, 학습활동 데이터, 학습 콘텐츠 데이터)와 콘텐츠 인식 설문 데이터를 분석에 포함하였다. 중도탈락 예측 요인 탐색을 위해 랜덤 포레스트(random forest)를 적용하여 분석하였으며, 중도탈락 예측 요인에서 중도탈락 학습자와 수강지속 학습자 간 차이를 알아보고자 독립표본 t 검증을 실시하였다. 분석 결과, 중도탈락을 예측하는 상위 주요 8개 변인은 과제 성적, 과제 제출, 로그인 횟수, eBook 학습 횟수, 학습 질문 횟수, 로그인 간격, 로그인 규칙성, 자기효능감으로 확인되었다. 또한 중도탈락 주요 예측 요인으로 도출된 상위 8개 변인 모두에서 중도탈락 학습자와 수강지속 학습자 간 차이가 있는 것으로 분석되었다. 연구결과를 토대로 본 연구의 의의와 시사점을 제시하였다.

주요어 : 온라인 학습, 중도탈락, 로그 데이터, 랜덤 포레스트, 기계학습

[†] 교신저자 : 이성혜, KAIST 과학영재교육연구원, slee45@kaist.ac.kr

I. 서 론

온라인 교육이 급격히 확산되면서 누구나 쉽게 원하는 교육에 접근할 수 있는 기회가 열렸지만, 교육을 끝까지 지속하지 못하고 그만두는 수강생의 비율도 매우 높은 편이다. 누적 학습자가 2억 명을 넘은 MOOC(Massive Open Online Course)의 경우(Shah, 2019) 이수율은 3.2% ~ 6.5% 정도로 보고되고 있으며(Impey et al., 2015; Jordan, 2014), K-MOOC은 이보다 좀 더 나은 13% 정도(조인식, 2020)로 중도탈락률은 80 ~ 90% 이상이다. 이에 비해 국내 사이버대학 평균 중도탈락률은 15.4%(대학알리미, 2021), 방송통신고등학교 10대 학생의 중도탈락률은 30% 내외(강성국 외, 2014)로 MOOC보다 높은 이수율을 보이지만 이 역시 오프라인 교육기관과 비교하면 매우 높은 수치이다. 이는 온라인 교육이 원하는 강좌에 쉽게 접근하여 시공간의 제약 없이 편리하게 학습할 수 있다는 장점이 있는 반면, 학습에 대한 책임이 전적으로 학습자에게 있고 또한 상호작용의 부족으로 쉽게 고립감을 느낄 수 있어 학습자가 동기를 유지하면서 학습을 지속하기 어려운 환경이기 때문이다(주영주 외, 2012; Carr, 2000; Rovai, 2003).

이러한 상황에서 온라인 교육을 끝까지 완료하는 학생들은 어떠한 특성이 있는가? 또는 온라인 수업을 중도에 그만두는 학습자들은 어떠한 특성이 있는가? 이러한 문제는 온라인 교육 기관과 학계의 오랜 관심 중의 하나로 이를 파악하기 위한 연구들이 다수 수행되어 왔다. 이러한 연구들은 중도탈락의 원인을 학습자 특성, 교육과정 및 교육내용, 교육환경 등에서 찾고자 하였다. 예를 들면, 성인학습자가 주된 수강생인 MOOC 및 원격대학의 맥락에서 연령, 직업 유무, 결혼 유무 등이 중도탈락과 관련 있는 학습자의 배경 특성으로 제시되기도 하였으며(James et al., 2016; Yasmin, 2013), 또한 학습자 개인의 학습동기, 자기효능감, 시간관리 역량, 자기조절학습 능력 등이 영향요인으로 나타나기도 하였다(Choi & Kim, 2017; Halawa et al., 2014; Wang & Newlin, 2002). 교육과정 및 내용과 관련해서는 수업 만족도, 교육내용, 교수설계, 학습자 상호작용 등이 중도탈락에 영향을 미치는 것으로 나타났다(주영주 외, 2007; 이현주, 2007; 정영란, 2016). 또한 중도탈락과 관련이 있는 온라인 학습 행동의 특성을 밝히고자 한 연구들도 다수 수행되었다(유지원, 2014; 이지은, 2019; 정영란, 2020). 이 밖에 온라인 교육 기관의 시설이나 인프라, 지원 등과 같은 환경 요인이 제시되기도 하였다(정선정, 2005; Botton & Gregory, 2015).

중도탈락은 초기에 발생하는 경향이 있어(Chen et al., 2019; Toledo et al., 2020), 중도탈락 위험을 조기에 예측하여 적절한 중재와 지원을 제공하는 것이 중요하다는 관점에서 최근의 중도탈락 연구는 중도탈락 학습자와 이수자 간에 특성 차이를 밝혀 중도

탈락 원인을 설명하기보다는 온라인학습시스템의 학습자 행동 데이터를 분석하여 중도탈락을 예측하는데 보다 관심이 집중되고 있으며(Borrelli et al., 2019), 머신러닝 기반의 교육 데이터 분석 기법의 발달과 함께 활발히 이루어지고 있다. 머신러닝은 데이터에서 규칙이나 패턴을 학습하고 새로운 데이터에 대해 적절한 작업을 수행하는 처리 과정을 말하며(Goodfellow et al., 2016), 컴퓨터가 축적된 데이터를 학습하여 패턴을 발견하고 이를 바탕으로 앞으로 발생할 일을 예측할 수 있도록 한다. 최근 교육 분야에서 데이터의 방대한 축적과 다양한 머신러닝 기법의 발달로 학습자 및 학습과 관련된 데이터를 분석하여 학업성취도(유진은, 2020; 이현우 외, 2021; Yoo & Rho, 2017), 학습부진 및 기초 학력 미달(박미현, 허균, 2021; 신종호, 최재원 2019; 이종현, 조규락, 2021), 진로 결정(노민정, 유진은, 2019; 박소영, 정혜원, 2021) 등을 예측하고자 하는 다양한 시도가 이루어지고 있는 가운데, 중도탈락은 머신러닝 기반 학습 데이터 분석의 주요 관심 중 하나이다. 연구자들은 온라인 교육 뿐만 아니라 오프라인 교육의 맥락에서 중도탈락 위험을 예측할 수 있는 데이터를 탐색하고 정확도가 높은 분석모델을 제시하고자 하였다(이은정 외, 2020; 황현정 외, 2021; Chung & Lee, 2019; Dass et al., 2021; Kashyap & Nayak, 2018). 머신러닝의 등장으로 연구자들은 중도탈락이 발생하기 전에 이를 탐지하여 중도탈락 가능성을 낮출 수 있는 정확한 예측모델을 생성할 수 있게 되었으며, 중도탈락률을 줄이는 것은 물론 온라인 학습을 지속하도록 하는 요인을 추출하여 다양한 처치를 제안할 수 있게 되었다.

한편, 지금까지 온라인 교육의 맥락에서 중도탈락자의 학습경험을 조사, 분석하는 것이 어려운 일이었다. 많은 온라인 교육 기관이 교육 후에 온라인 교육에 대한 경험, 만족도, 개선 사항 등에 대한 조사를 실시하지만, 이러한 교육 후 조사에 응답하는 학습자는 대부분 학습을 성공적으로 완료한 학습자들일 가능성이 높아 조사 결과가 편향될 수 있다는 것이다(Whitchill et al., 2015). Whitchill 외(2015)는 교육 프로그램에 대한 전통적인 평가 방식, 즉 모든 학생이 과정에 끝난 후에 평가하는 방식은 중도탈락률이 높은 온라인 교육의 맥락에서 잘 작동하지 않는다고 지적하며, 머신러닝 기법이 이러한 한계를 극복할 수 있는 혁신적인 대안을 제공할 수 있음을 강조한다.

본 연구는 중고등학생 학습자 대상의 온라인 수학, 과학 교육 프로그램에서 머신러닝 기법을 적용하여 중도탈락 예측 요인을 밝히고자 하였다. 해당 프로그램은 K 대학에서 전국의 초, 중, 고등학생들에게 제공하는 온라인 교육 프로그램으로 수학, 과학에 흥미가 있는 학생이면 누구나 자발적으로 신청하여 12주간 참여할 수 있는 프로그램이다. 본 연구는 해당 프로그램에서 학생들이 수강 철회, 즉 중도탈락 의사를 밝히는 전반 6주간의 온라인 학습 행태와 교육 콘텐츠에 대한 인식을 바탕으로 중도탈락 예

측력이 높은 변인을 탐색하고자 한다. 또한 중도탈락 예측 변인은 중도탈락 학습자와 수강지속 학습자 간에 차이가 나타나는지 살펴보고자 한다.

현재까지 중도탈락 연구의 대부분은 MOOC, 원격대학, 일반대학의 온라인 수업, 기업 온라인 교육 등의 맥락에서 주로 이루어져 왔으며, 따라서 대부분은 성인을 대상으로 이루어져 온 특성이 있다. 그러나 COVID-19 팬데믹은 온라인 교육을 전 교육의 영역으로 확장시켰으며, 초중고 교육에서도 예외가 아니다. 이에 본 연구의 맥락과 같이 중고등학생 학습자가 개인적인 관심과 흥미를 기반으로 선택해서 수강할 수 있는 맥락에서 이러한 요인이 파악된다면 중고등학생들의 중도탈락 예측 요인을 밝힐 수 있는 의미있는 연구가 될 것이다.

본 연구의 연구문제는 다음과 같다.

연구문제 1. 중고등학생 대상 온라인 교육에서 중도탈락을 예측하는 요인은 무엇인가?

연구문제 2. 중도탈락 예측 요인은 중도탈락 학습자와 수강지속 학습자 간에 차이가 있는가?

II. 이론적 배경

1. 온라인 교육에서 중도탈락 관련 연구

온라인 교육에서 중도탈락(Dropout)은 일반적으로 학습자가 등록한 온라인 교육 프로그램을 완료하지 못하고 중간에 학습을 포기하는 것을 의미한다(임연옥, 2007; Levy, 2007; Muse, 2005). 중도탈락 개념은 연구 대상이나 교육 형태, 교육 프로그램 등에 따라 다양하게 정의되어 왔는데 크게 세 가지로 구분하여 볼 수 있다. 온라인 교육 기관에 등록했던 수강생이 수강 철회를 통해 공식적으로 수강을 포기하거나 교육을 완료하지 못한 상태를 중도탈락의 개념으로 정의하기도 하였으며(유지원, 2014; Castles, 2004; Levy, 2007; Lim, 2016), 온라인 교육에 등록하였으나 교육 기간 중 특정 기간을 한정하여 그 기간 이후 교육 참여를 하지 않는 경우를 뜻하기도 하였다(Halawa et al., 2014). 그 외에도 재등록을 포기한 상태를 중도탈락으로 보기도 하였다(김지현, 2013; 정영란, 2020; Allen, 2017; Bettinger et al., 2017).

온라인 교육환경은 오프라인 교육과 비교하여 보다 자기주도적인 학습을 요구하는 학습환경 특성 때문에, 학습자는 교수자 및 동료 학습자와 분리된 상황에서 쉽게 고립

감과 상호작용의 부족을 느끼며 학습동기가 저하될 수 있다. 동기의 저하, 그리고 일과 학업을 병행하는 성인학습자의 특성 등으로 인해 중도탈락률이 높은 경향이 있다(주영주 외, 2012; Carr, 2000; Rovai, 2003). 예를 들어, 2012년 MOOC(Massive Open Online Course)가 처음 등장한 이래 우수한 교육 콘텐츠에 대한 접근성을 제공하며 급성장하여 2021년 누적 학습자가 2억명을 넘었지만(Shah, 2019), 학습을 완료한 수강생의 비율은 대체로 3.2% ~ 6.5%에 그치는 것으로 보고되고 있다(Impey et al., 2015; Jordan, 2014). K-MOOC 역시 2015년 서비스를 시작한 이래 평균 이수율은 13.1%로 보고되고 있으며(조인식, 2020), 2020년 대학정보공시에 따르면 국내 사이버대학의 평균 중도탈락률은 15.4%로 일반대학 평균 7%에 비해 두 배가 넘는 수치를 보인다(대학정보공시, 2021). 청소년 대상 온라인 교육의 경우도 방송통신고등학교 10대 학생의 중도탈락률은 30% 내외로 보고된 바 있다(강성국 외, 2014). 중도탈락을 시기별로 살펴본 연구에서 방송통신대학 학습자의 중도탈락이 높은 시기는 입학 첫 학기로 나타났으며(남신동 외, 2014; 정혜령 외, 2015), Toledo 외(2020)는 온라인 기업가정신 프로그램에서 전체 탈락자의 30%가 과정 중반 이전에 이루어지고 있음을 보고하였다.

중도탈락 모형 탐구, 중도탈락 영향요인 또는 예측 요인을 탐색, 중도 탈락자와 학습 지속자를 비교한 연구는 모두 중도탈락 관련 변인을 다양한 관점에서 정리, 제시하였다. 먼저 Kember(1989)의 모형은 중도탈락을 설명하는 대표적인 중도탈락 모형으로 일반대학의 중도탈락 모형(Tinto, 1975)을 온라인 교육 상황에 맞게 재설정 하고자 하였으며 학습자의 개인적, 사회적 배경과 학습자 내적, 외적 변인에 초점을 맞추었다. 그 후 Rovai(2003)는 기존 중도탈락 모형을 요약하고 온라인 교육 중도탈락 요인을 학생 개인 특성, 학습능력, 내부 요인, 외부 요인 등으로 나누어 종합 모형을 제시하였다. 이 외에도 기존 모형을 바탕으로 영향요인을 학습자 요인, 교육과정/교육 프로그램과 내용 요인, 환경적 요인으로 중도탈락 모형을 정리하기도 하였다(주영주 외, 2008; Lee & Choi, 2011).

사이버대학, 대학 온라인 교육, 기업 교육에서 중도탈락 연구는 중도탈락 모형을 기반으로 하여 다양한 영향요인을 규명하였다. 본 연구에서는 Lee와 Choi(2011) 그리고 주영주 외(2007)가 제시한 학습자 변인, 교육과정 및 교육내용, 교육 환경 요인을 중심으로 중도탈락 요인을 살펴보고자 한다.

원격대학 재학생을 대상으로 한 중도탈락 연구에서 학습자 개인 특성인 성별, 학년, 전공과 함께 과제 제출 수, 평점, 장학금 등이 중도탈락 결정요인으로 도출되었으며(권선아 외, 2020), 연령 또한 중도탈락의 영향요인 중 하나로 제시되었다(James et al., 2016). Choi와 Kim(2017)은 사이버대학 프로그램에서 원거리 학습자의 중도탈락 결정요

인을 분석하였는데, 학습 동기, 학습자와 교수자 상호작용, 성적 등이 중요한 영향력을 미치고 있음을 확인하였다. Hart(2012)는 온라인 교육 프로그램에서 학습 지속과 관련된 변인을 연구하였는데, 교육 만족도, 가족과 동료의 지원, 시간 관리 능력이 학습 지속과 관련이 있는 것으로 나타났다. 이 외에도 정주영과 이정원(2017), Wang과 Newlin(2002), Zimmerman(2000)은 온라인 교육에서 중도탈락의 가장 중요한 변인을 자기 효능감으로 보았으며, 흥미(권혜진, 2010; Halawa et al., 2014), 학업적 통제소재(Levy, 2007; Morris et al., 2005), 자기조절학습능력(Lee et al., 2013), 온라인 교육 이수 횟수, 컴퓨터 활용 능력(Dupin-Bryant, 2004), 교수자 만족도(Moore & Kearsley, 1996) 또한 중도탈락 원인으로 밝혀졌다.

한편, 교육과정 및 교육내용을 주요 중도탈락 결정 요인으로 분석한 선행연구를 살펴보면 다음과 같다. 정영란(2020)은 학습분석학 기반으로 사이버대학 신입생의 중도탈락을 예측하고 중도탈락 양상을 비교하여 중도탈락 위험군을 분석하였는데, 그 결과 중도탈락을 예측하는 주요 변인은 성적, 학습 규칙성, 수업 만족도, 전액 장학 여부로 확인되었으며, 학습 규칙성을 확보하는 것이 학업 지속을 위해 중요하다고 보았다. 사이버대학에서의 재등록률 영향요인을 분석한 정영란(2016)의 연구에서는 시간 확보, 상호작용 횟수, 성적, 수업 만족도 등과 같은 교육과정, 내용 변인이 중도탈락의 원인으로 밝혀졌다. 주영주 외(2007)는 근거이론을 바탕으로 사이버대학 중도탈락자들의 개별 면담을 통해 중도탈락 결정요인을 심층적으로 분석하였는데, 내적 동기, 난이도, 교육 내용, 교수설계, 학습자 상호작용 등이 중도탈락 요인으로 탐색되었다. 이 외에도 출석률, 학습 시간, 강의 접속 수, 규칙적 학습, 학습 참여 빈도, 콘텐츠 접속 빈도, 게시글 읽은 빈도 등과 같은 학습활동 관련 변인이 중도탈락 영향 변인으로 밝혀졌다(유지원, 2014; 이지은, 2019; 전주성, 2010; Morris et al., 2005). 아울러 학습관리시스템(Boton & Gregory, 2015), 물리적 지원(정선정, 2005), 등록금(서선주, 2004)과 같은 교육환경적 요인 또한 중도탈락 중요 변인으로 도출되었다.

다음으로, 학습지속자와 중도탈락자를 집단으로 구분하여 차이를 비교한 연구를 살펴보면 다음과 같다. 이지은(2019)은 중도탈락 예측지수에 영향을 미치는 학습데이터를 규명하고자 일반 학습자와 중도탈락 위험군을 나누어 학습 패턴에 차이가 있는지 살펴본바, 분석 결과 총 수강 차시, 이수학점 수에서 차이가 있는 것으로 나타났다. 더불어 학습자의 학습 정보와 학사정보가 중도탈락 위험도에 미치는 영향을 중다선형 회귀분석으로 검증한 결과 학습자의 총 수강 차시, 이수학점 수, 평균 평점이 중도탈락 위험도 영향요인으로 밝혀졌다.

또한, Lee 외(2013)의 연구에서 학습지속자는 자기조절학습과 학업적 통제 소재를 높

게 지각한다고 밝혔으며, 학습지속자가 중도탈락자에 비해 학습 참여, 학습지원, 학습 만족도 등이 높은 것으로 나타났다(Morris et al., 2005; Park & Choi, 2009).

이와 같이 온라인 교육에서 중도탈락 영향요인을 규명한 선행연구에서 검증된 요인을 종합적으로 정리하면 다음과 같다.

〈표 1〉 중도탈락 영향 요인 분석 종합

구분	변인	영향 변인
학습자 변인	학습자 배경 변인	성별, 연령, 학년, 전공
	학습자 내적 변인	자기효능감, 학습 동기, 흥미, 내적 통제소재, 학업적 통제소재
	학습 능력	자기조절학습능력, 시간 관리 능력
	온라인 교육 경험	온라인 교육 경험, 온라인 교육 이수 횟수, 컴퓨터 활용 능력 부족
교육과정/교육 내용 변인	교육과정/프로그램	교수설계
	교육내용/교육 콘텐츠	교육내용, 난이도, 학습량
	학습활동	출석률, 총 학습 시간, 시간 확보, 강의 접속 수, 과제 제출 수, 규칙적 학습, 학습 콘텐츠 접속 빈도, 게시글을 읽은 빈도
	교육 결과	평점, 성적, 낮은 학점, 이수 학점 수, 교육 만족도, 장학금
	상호작용	학습자 간 상호작용, 학습자-교수자 상호작용, 상호작용 횟수
교육환경	교육시설/인프라	학습관리시스템, 물리적 지원
	심리적 지원	가족 지원, 동료 지원
	제도적 지원	등록금

2. 머신러닝 기반의 중도탈락 연구

중도탈락 연구는 머신러닝 기반의 교육 데이터 분석 기법의 발달과 함께 보다 활발히 이루어지고 있다. 머신러닝(Machine Learning)은 규칙을 일일이 프로그래밍하지 않아도 자동으로 데이터에서 규칙이나 패턴을 학습하고 새로운 데이터에 대해 적절한 작업을 수행하는 처리 과정을 말한다(Goodfellow et al., 2016). 머신러닝은 컴퓨터 과학, 통

계학, 데이터 마이닝 등과 관련이 있으며, 지도학습(Supervised Learning), 비지도 학습(Unsupervised Learning), 강화 학습(Reinforcement Learning) 등으로 분류된다. 미래 예측을 위한 모형을 완성하는 것이 핵심인 머신러닝은 최근 학습 성취 예측(유진은, 2020; 이현우 외, 2021; Yoo & Rho, 2017), 중도탈락 예측(이은정 외, 2020; 황현정 외, 2021; Chung & Lee, 2019; Dass et al., 2021; Kashyap & Nayak, 2018) 학습 부진아 또는 기초 학력 미달 예측(박미현, 허균, 2021; 신중호, 최재원 2019; 이종현, 조규락, 2021), 진로 결정 예측(박소영, 정혜원, 2021; 노민정, 유진은, 2019) 등과 같이 교육 데이터 분석 연구에 다양하게 활용되고 있다.

중도탈락은 머신러닝 기반의 교육 데이터 분석 연구의 주요 관심 중 하나였다. 앞서 언급했듯이 중도탈락은 과정 중반 이전에 발생하는 경향이 있어 중도탈락 가능성이 높은 학생들을 조기에 머신러닝 기법을 활용하여 예측함으로써 사전에 중도탈락을 방지하기 위한 지원을 제공하는 목적으로 연구가 수행되어 왔다.

이은정 외(2020)는 전국 4년제 대학의 중도탈락률 추이를 분석하고 머신러닝 기법인 랜덤 포레스트 예측 모형을 구축하여 중도탈락을 예측하는 결정요인을 탐색하였다. 연구 결과 전임교원 1인당 연구비 전임교원 1인당 논문 수, 신입생 충원률, 재적학생 수, 학생 1인당 교육비 등이 상위 결정요인으로 도출되었다. 랜덤 포레스트를 활용하여 대학의 중도탈락 요인을 예측한 또 다른 연구에서는 대학 입학 전 학교 성적, 교육 만족도, 자기평가 등이 주요한 요인으로 나타났다(Behr et al., 2020). Chung과 Lee(2019)는 2014년 국가교육정보시스템(NEIS)에서 제공되는 서울, 인천, 경북, 경상남도 지역의 고등학생을 대상으로 학업 중퇴 비율과 영향요인을 탐색하였는데, 무단결석, 무단 지각, 조퇴, 수업 결석, 자율적 활동 시간, 동아리 및 봉사 활동 시간 등이 학업 중퇴를 예측한다고 제시하였다. 그 외에도 Opazo 외(2021)에 의하면 의사결정나무(Decision Trees), 로지스틱 회귀분석(Logistic Regression), K-최근접 이웃(K-Nearest Neighbors), 서포트 벡터 머신(Support Vector Machine), 랜덤 포레스트(Random Forest) 등 다양한 머신러닝 방법을 적용하여 대학 수학 수업에서 중도탈락 영향요인을 탐색한 결과 높은 학업 성적이 중도탈락을 예측하는 요인으로 규명되었다.

온라인 교육에서 머신러닝 기반 중도탈락 연구는 대부분 대규모의 학생들이 수강하는 MOOC을 기반으로 이루어져 왔으며, 중도탈락의 영향을 미치는 행동 패턴을 탐색하고 예측의 정확도를 높이는 분석 모형을 도출하는데 집중되어 왔다.

Kashyap와 Nayak(2018)에 의하면, HarvardX에 수집된 데이터를 활용하여 중도탈락을 예측하는 다양한 머신러닝 기법을 탐색한 연구에서 로지스틱 회귀분석(Logistic

Regression), 서포트 벡터 머신(Support Vector Machine), 랜덤 포레스트(Random Forest), 그래디언트 부스팅 의사 결정 트리(Gradient Boosting Decision Tree)를 활용하여 중도탈락 예측하는 모형을 탐색하였으며, 랜덤 포레스트가 중도탈락 예측에 최적의 결과를 보인 것으로 나타났다. 또한 중도탈락 영향요인으로서는 상호작용, 흥미, 교육과정의 질 등이 도출되었다.

Yukselturk 외(2014)는 대학 졸업 후 온라인 정보 기술 자격증 프로그램에 등록한 학생을 대상으로 중도탈락 학생을 분류하고 중도탈락 영향요인을 탐색하는 연구에서 K-최근접 이웃(K-Nearest Neighbors), 의사결정나무(Decision Trees), 나이브 베이즈(Naive Bayes), 신경망 모형(Neural Networks) 중 K-최근접 이웃과 의사결정나무가 최적의 분류를 하는 것을 확인하였으며, 자기효능감, 온라인 교육 준비성, 콘텐츠에 대한 사전 지식 등이 학생의 중도탈락을 예측하는 중요한 변인임을 보고하였다.

이대현과 조희석(2021)은 사이버대학에 재학 중인 학습자의 개인 정보, 강의 정보, 수강 정보, 과제, 퀴즈, 토론, 팀 프로젝트와 같은 학습활동 데이터를 활용하여 중도탈락자를 예측하는 정보를 제공하고자 로지스틱 회귀분석(Logistic Regression), 랜덤 포레스트(Random Forest), 서포트 벡터 머신(Support Vector Machine) Gradient Boosting, 나이브 베이즈(Naive Bayes) 분류, XGBoost를 활용하여 중도탈락 예측 연구를 수행하였다. 연구 결과 XGBoost가 높은 정밀도와 재현율을 보여 중도탈락 예측에 가장 적합한 모델로 선별되었다.

유지원(2014)은 일반대학 이러닝 강좌에서 학습관리시스템(LMS)에 기록된 데이터로부터 중도탈락 예측 모형을 구축하여 로지스틱 회귀분석으로 중도탈락 예측 요인을 검증하였다. 그 결과 출석과 총 학습 시간이 유의미하게 중도탈락을 예측하는 것으로 나타났으며, 총 접속 빈도는 유의하지 않은 것으로 나타났다. 또한 중도탈락이 결정되는 4주 시점에서 출석과 총 학습 시간은 96%의 분류 정확도로 수강 완료자와 중도 탈락자를 구분하였다. 그 외에도 Coussement 외(2020)는 Logit Leaf Model(LLM)을 활용하여 중도탈락 예측 요인을 탐색하였으며, 그 결과 인구 통계학적 특징, 교실 특성, 인지능력, 학습 참여 등이 중도탈락을 정확하게 예측할 수 있다고 확인하였다.

이와 같이 머신러닝 기반의 중도탈락 연구가 온라인 교육뿐만 아니라 오프라인 교육의 맥락에서도 활발히 이루어지고 있는 상황이다. 선행연구에서는 다양한 머신러닝 기법을 적용하여 중도탈락 예측 정확도를 높이는 분석 모형을 도출하고자 하였으며, 중도탈락 예측 변인으로 인구 통계학적 특징, 총 학습 시간, 학습 참여, 상호작용, 흥미, 교육과정의 질, 교실 특성, 인지능력 등이 탐색되었다.

III. 연구 방법

1. 연구 맥락 및 연구 대상

K 대학에서 제공하는 온라인 교육 프로그램은 수학, 과학에 흥미가 있는 학생이면 누구나 참여할 수 있어 학생들이 자발적으로 참여하고 싶은 강좌를 선택하여 온라인 교육에 참여한다. 온라인 교육 프로그램은 2020년 3월부터 6월까지 총 12주 동안 제공되었으며, 학습 참여자는 12주의 교육 일정 중 6주 이내에 수강 철회를 할 수 있는 기회가 주어진다. 온라인 교육 프로그램에서는 국가 교육과정 기반의 수학, 과학 관련 개념학습과 실생활 문제에 연계된 총 3개의 탐구과제를 eBook의 형태로 제공하고 있다. 학습자는 1개의 탐구과제를 해결하기 위해 4주 동안 eBook을 통해 수학 및 과학 개념을 학습할 수 있으며, 그 기간 동안에 학습자는 언제든지 탐구과제를 제출할 수 있도록 과제 제출 기간을 두고 있다. 본 온라인 교육 프로그램은 학습자 스스로가 자신의 학습 계획과 일정에 맞춰 자기주도적으로 학습을 수행할 수 있도록 되어있으며, 학습을 하거나 과제를 수행하는 데 도움이 필요하면 언제든지 학습관리시스템(LMS) 게시판을 통해 튜터에게 문의를 하거나 학습 동료 또는 튜터와 상호작용이 가능하다. 과제 제출 기간 종료 후, 튜터는 학습자가 제출한 과제를 채점하여 결과를 피드백과 함께 학습관리시스템(LMS)에 업로드하며 학습자는 그 결과를 바로 확인할 수 있다.

본 연구에서는 K 대학 온라인 교육 프로그램 중 중·고등 수학, 과학 과정을 수강한 학생 788명을 분석에 활용하였다. 본 연구에 참여한 학생의 성별 분포는 남학생 484명(61.4%), 여학생 304명(38.6%)이며, 과목별 분포는 수학 252명(32.0%), 과학 536명(68.0%)이었다. 학교급 및 학년별로는 중학교 1학년 261명(33.1%), 2학년 191명(24.2%), 3학년 122명(15.5%)이었고 고등학교 1학년 121명(15.4%), 2학년 88명(11.2%), 3학년이 5명(0.6%)이었다. 연구 대상자 중 중도탈락 학습자 성별 분포는 남학생 71명(60.7%), 여학생 46명(39.3%)이며, 과목별 분포는 수학 37명(31.6%), 과학 80명(68.4%)이었다. 학교급 및 학년별로는 중학교 1학년 26명(22.2%), 2학년 30명(25.6%), 3학년 16명(13.7%)이고 고등학교 1학년 20명(17.1%), 2학년 24명(20.5%), 3학년 1명(0.9%)이었다.

〈표 2〉 연구 대상자

	구분	빈도(명)	비율(%)
성별	남	484	61.4
	여	304	38.6
학년	중학교 1학년	261	33.1
	중학교 2학년	191	24.2
	중학교 3학년	122	15.5
	고등학교 1학년	121	15.4
	고등학교 2학년	88	11.2
	고등학교 3학년	5	0.6
	총계	788	100
과목	수학	252	32.0
	과학	536	68.0
총계		788	100

〈표 3〉 연구 대상자 중 중도탈락 학습자 현황

	구분	빈도(명)	비율(%)
성별	남	71	60.7
	여	46	39.3
학년	중학교 1학년	26	22.2
	중학교 2학년	30	25.6
	중학교 3학년	16	13.7
	고등학교 1학년	20	17.1
	고등학교 2학년	24	20.5
	고등학교 3학년	1	0.9
	총계	117	100
과목	수학	37	31.6
	과학	80	68.4
총계		117	100

2. 연구 변인

1) 종속변인

본 연구에서 종속변인으로 활용된 중도탈락(Dropout)은 온라인 교육이 제공되는 총 12주 중 학습자가 수강 철회를 할 수 있는 마지막 일인 6주까지 수강 철회를 한 학생을 중도탈락자로 보았으며 온라인 교육 중·고등과정 수강생 중 중도탈락 학습자는 117명(14.8%)이었다. 본 연구에서 중도탈락 학습자는 1, 수강지속 학습자는 0으로 코딩하였다.

〈표 4〉 중도 탈락자 현황

구분	빈도(명)	비율(%)
수강지속 학습자	671	85.2
중도탈락 학습자	117	14.8
총계	788	100

2) 예측 변인

온라인 교육에서 중도탈락을 예측하는 변인을 탐색하기 위하여 선행연구에서 도출된 중도탈락 예측 요인을 중심으로 학습관리시스템(LMS)에서 추출할 수 있는 학습자 변인 데이터, 학습활동 데이터 및 학습 콘텐츠 데이터 그리고 학습자가 탐구과제 수행 후 응답한 콘텐츠 인식 설문 데이터를 예측 변인으로 활용하였다. 특히, 본 연구의 주요 연구 대상인 중도탈락 학습자들은 데이터 결측이 많아 통계적 검증에 어려움이 있을 수 있어 중도탈락 학습자의 데이터 중 LMS에서 중도탈락 학습자를 모두 포함할 수 있는 데이터를 중심으로 구성하여 분석에 활용하였다. 본 연구는 콘텐츠 인식 설문 문항인 콘텐츠 흥미, 난이도, 자기효능감, 도전감 이외에 모든 변인을 LMS에 기록된 로그 데이터를 기반으로 하여 분석을 수행하였다. 예측 변인의 설명은 다음과 같으며, <표 5>에 요약 제시하였다.

선행연구에서 중도탈락 영향 요인으로 나타난 변인들은 <표 1>과 같이 학습자, 교육과정 및 내용, 교육환경 변인 등으로 구분될 수 있었으며, 이를 바탕으로 본 연구에서는 학습자 특성 데이터와 함께 특히 교육과정 및 내용 특성에 중점을 두어 온라인 학습 과정과 관련된 학습활동 및 학습 콘텐츠 데이터, 교육 내용과 관련된 콘텐츠 인식 데이터를 분석에 활용하고자 하였다. 다만, 본 연구에서 활용한 데이터는 LMS 로그 데이터를 기반으로 수집 가능한 데이터를 활용하였다.

본 연구에서 중도탈락 학습자는 수강 신청 철회가 가능한 기간인 6주 이내 수강 취소를 신청한 학습자로 보았다. 따라서 LMS 로그 데이터는 6주 시점까지를 기준으로 추출하였다. 본 연구에서 활용된 학습자 변인 데이터는 성별, 학년, 학급, 수강 과목이며, 학급은 중학교와 고등학교로 구분하였고, 과목은 수학과 과학으로 구분하였다.

LMS 로그 데이터 중 학습활동 데이터에 해당하는 로그인 횟수는 6주 동안 LMS에 로그인 한 횟수의 합을 말한다. 로그인 간격과 규칙성은 조일현과 김윤미(2013)의 산술 방법을 채택하여 산출하였는데, 로그인 간격 값은 학습자가 로그인 한 시점 기록을 바탕으로 로그인 시점 간의 간격 평균으로 산출하였고, 로그인 규칙성은 로그인 간격의 표준편차(Standard Deviation)로 산출하였다. 로그인 간격이 작을수록 접속을 자주 한 것으로, 로그인 간격의 표준편차 값이 작을수록 로그인을 규칙적으로 한 것으로 판단하였으며, 로그인 간격과 규칙성 값은 분 단위로 계산하고 해석은 부적으로 하였다. 과제 제출은 6주 기간 중 4주까지 1차시 과제 제출을 하게 되어 있어 4주까지 과제 제출을 하였는지 여부로 코딩하였으며, 학습 질문은 6주 동안 학습 질문 게시판에 학습자가 질문을 한 횟수의 합으로 산출하였다. 과제 성적은 학습자가 제출한 1차시 과제 평가 점수를 말한다.

학습 콘텐츠 데이터 중 eBook 학습 횟수는 6주 동안 학습자가 eBook을 열람 한 횟수의 합을 말하며, eBook 학습 간격과 규칙성은 로그인 간격과 규칙성 산출 방법과 동일하게 eBook을 열람 시점 간의 간격 평균으로 eBook 학습 간격을 산출하였다. eBook 학습 규칙성은 eBook 학습 간격의 표준편차를 통해 규칙성 값을 계산하였으며, eBook 학습 간격과 규칙성 값은 분 단위로 계산하였다. 더불어 eBook 학습 간격 값이 작을수록 자주 eBook을 열람한 것으로, eBook 학습 규칙성은 값이 작을수록 eBook을 규칙적으로 열람한 것으로 판단하여 부적으로 해석하였다. eBook 반복 열람 횟수는 동일한 eBook을 반복적으로 열람한 횟수를 계산하였으며, 최초 eBook 열람 시기는 온라인 교육 프로그램 개강 후 처음으로 eBook을 열람 한 시점을 기록하여 개강일과의 차이로 산출하였다.

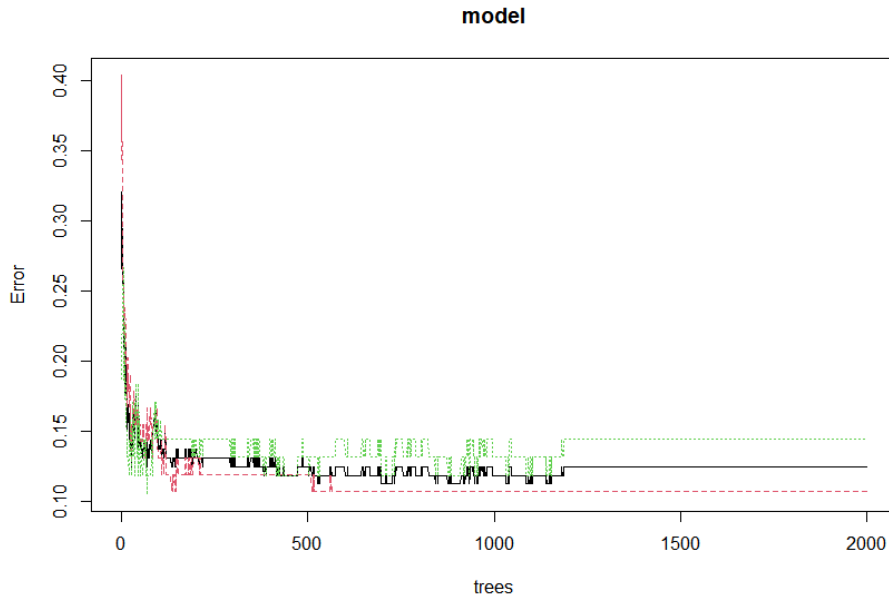
콘텐츠 인식 설문 데이터는 탐구과제와 함께 배포하여 수집하였다. 교육 콘텐츠 흥미 측정을 위해 Rotgans와 Schmidt(2011)의 상황적 흥미(Situation Interest) 척도 4문항을 사용하였으며, 자기효능감, 도점감, 난이도는 Chae와 Gentry(2007)의 SPOCQ-K 척도를 활용하였다. 설문은 ‘전혀 그렇지 않다’(1점)부터 ‘매우 그렇다’(5점)까지 5점 Likert 척도로 측정되었다.

〈표 5〉 예측 모형 투입 변인

구분	변인	척도
학습자 변인 데이터	성별	1=남, 2=여
	학년	1= 중1, 2=중2, 3=중3, 4=고1, 5=고2, 6=고3
	학급	1=중등, 2=고등
	수강 과목	1=수학, 2=과학
예측 변인	로그인 횟수	6주 동안 로그인 한 횟수
	로그인 간격	로그인 시점 간의 간격
	로그인 규칙성	로그인 간격의 표준편차
	과제 제출	1=과제 제출, 0=과제 미제출
	학습 질문	6주 동안 학습 질문을 한 횟수
	과제 성적	1차 과제 점수
	eBook 학습 횟수	6주 동안 eBook을 열람한 횟수
	eBook 학습 간격	eBook을 열람 한 시점 간의 간격
	eBook 학습 규칙성	eBook 학습 간격의 표준편차
	eBook 반복 열람 횟수	동일 eBook을 반복해서 열람한 횟수
학습콘텐츠 데이터	최초 eBook 열람 시기	개강일로부터 최초 eBook을 열람 한 시점 차이
	흥미	
	난이도	5점 Likert 척도
	자기효능감	(전혀 그렇지 않다(1) - 매우 그렇다(5))
콘텐츠 인식 설문 데이터	도전감	

3. 자료 분석 및 방법

본 연구는 온라인 교육에서 중도탈락 예측 요인을 탐색하고, 분석 결과로 도출된 중도탈락 예측 요인은 중도탈락 학습자와 수강지속 학습자 간에 차이가 있는지 탐색하고자 하였다. 이를 위해 머신러닝 분석 기법 중 하나인 랜덤 포레스트 분석을 실시하였다. 랜덤 포레스트는 붓스트랩 표본을 다수 생성하고 의사결정나무 모형을 적용하여 그 결과를 종합하는 앙상블 기법 중 하나이다. 랜덤 포레스트는 의사결정나무를 여러 개의 나무로 확장하여 정확한 예측을 하는 것을 목적으로 하고 있어 영향요인 중 최선의 결과를 도출하는 방식이다.(유진은, 2015; Breiman, 2001). 랜덤 포레스트는 잡음



(그림 1) 의사결정나무의 개수에 따른 예측 오차 비율

(Noise)이나 이상치(Outlier)로부터 크게 영향을 받지 않는 강건함(Robustness)를 가지고 있으며, 상대적으로 예측력 높고 모형이 안정적이다.(유진은 2015; Hamza & Larocque, 2005). 랜덤 포레스트 분석 시 붓스트랩 표본 수, 각 마디에서의 설명 변수 개수 등은 연구자가 선택해야 할 사항인데, Breiman(2001)은 설명 변수 개수의 경우 종속변수가 범주형이면 \sqrt{p} 개를, 종속변수가 연속형이면 $p/3$ 개를 권장하고 있다. 본 연구에서는 랜덤 포레스트를 위하여 1,500개의 붓스트랩 표본을 생성하였으며, 종속변수인 중도탈락 여부가 범주형 변수이므로 범주형 변수의 해당 기준을 적용하여 분석하였다. 이 모형의 성과를 평가하는데 붓스트랩 1,500이 충분한 숫자인지 확인하기 위해 예측 오차 비율이 어떻게 변화하는지 살펴보았다. [그림 1]에서 맨 아래쪽 그래프는 특이도, 가운데 그래프는 표본 전체에서 잘못 분류된 오차 비율(정분류율), 맨 위쪽 그래프는 민감도를 의미하는데, 본 연구에서는 의사결정나무가 1,500개를 넘어서면 예측 오차들이 안정적으로 일정한 값으로 수렴되기 때문에 1,500개로 설정하였다.

본 연구에서 투입된 중도탈락 학습자와 수강지속 학습자 데이터는 데이터 비율의 불균형으로 인해 분류 결과가 대체로 사례 수가 많은 집단 쪽으로 편향될 수 있다(김미림, 박민호, 2019). 이러한 편향 문제점을 보정하기 위한 표집 비율 조정 방법은 가중치를 적용하여 조정하거나, 사례 수가 많은 쪽을 적게 표집하는 방법(Down-Sampling) 또는 사례 수가 적은 쪽을 많이 표집하는 방법(Over-Sampling) 등을 취할 수 있다

(Breiman & Cutler, n.d.; Van Hulse et al., 2007). 본 연구에서는 사례 수가 적은 중도탈락 학습자가 실질적인 의미가 있다는 점을 참조하여 수강지속 학습자를 적게 표집하는 방식으로 붓스트랩 표본을 표집하였다. 그 후 표본을 랜덤으로 7(훈련용 자료):3(시험용 자료) 구분하여 분석과 검증 절차를 거쳤다.

랜덤 포레스트 분석 결과는 훈련자료와 시험자료에 대한 정분류율(Accuracy), 특이도(Specificity), 민감도(Sensitivity)를 기준으로 제시하였는데, 정분류율(Accuracy)은 전체 자료 중 실제와 예측이 일치하는 비율을 뜻한다. 본 연구의 종속변수는 1(중도탈락 학습자)과 0(수강지속 학습자)로 구성되어 있어, 특이도(Specificity)는 0인 자료 중 정분류된 자료의 비율을, 민감도(Sensitivity)는 1인 자료 중 정분류된 자료의 비율로 구하였다.

랜덤 포레스트 분석은 직관적인 그래프가 최종모형으로 도출되는 의사결정나무 모형과 달리 그래프가 최종모형으로 도출될 수 없어 결과에 대한 해석이 어려울 수 있다. 이 문제를 해결하기 위해 주요 변수 중요도 지수(Variable of Importance Index), 부분 의존성 도표(Partial Dependence Plots)를 제시하여 설명 변수의 중요한 영향력(중요도)을 알아볼 수 있도록 한다. 종속변수가 범주형 변수일 경우 정확도를 개선하는 정도(Mean Decrease in Accuracy 이하 MDA)와 노드 불순도를 개선하는 정도(Mean Decrease Gini 이하 MDG)로 변수의 중요도를 판단한다. MDA가 랜덤 포레스트 분석에서 가장 효율적인 중요도 척도로 널리 사용되고 있어(Strobl et al., 2007; Ishwaran, 2007; Genuer et al., 2010) 본 연구에서도 MDA를 기준으로 중요도를 판단하였다. 머신러닝 기법을 적용한 연구에서 중요도 지수에 따른 주요 변인을 보고할 때 연구자에 따라 중요도 지수 상위 10%(손윤희 외, 2020)를 보고하거나 여러 번 반복 시행 후 모두 포함된 변인(유진은 외, 2020; 정혜원 외, 2021)을 보고하는 등 중요도 변인을 상이하게 제시하고 있다. 본 연구에서는 랜덤 포레스트 분석을 10회 반복 실시하여 중요도 지수 상위 변인을 확인하였고, 반복 실시된 분석 결과 중요도 지수에서 10번 모두 포함된 변인을 확인하였다. 이렇게 중도탈락을 예측하는 중요도가 높은 설명 변수들을 선별한 뒤, 부분 의존성 도표(Partial Dependence Plot)를 활용하여 개별 독립변수와 종속변수 간의 영향력 패턴을 시각화하여 제시하였다.

랜덤 포레스트 분석을 위해 R 4.1.2 프로그램에서 randomForest(Breiman et al., 2018) 패키지를 사용하였으며, 결측치는 단순 대체를 반복 수행하여 가상적이 데이터를 생성하고 이들의 평균으로 결측값을 대체하는 방법인 다중 대체법을 사용하였으며, 이를 위해 R 4.1.2버전에서 mice 패키지를 활용하였다(van Buuren & Groothuis-Oudshoorn, 2011). 결측치를 처리하는 방법에는 결측 데이터를 완전 제거하거나 결측치를 대체하는 방법 등이 다양하게 활용되고 있다. 최근에는 결측치를 대체하는 방법으로 평균 대

체법, 중앙값 대체법, KNN, 단순 대체법, 다중 대체법 등이 주로 활용되고 있으며, 이중 다중 대체법이 결측치를 처리할 때 가장 안정적인 결과를 보여주고 있는 것으로 나타나 본 연구에서도 다중 대체법을 활용하여 결측치를 처리하였다(Mohammed et al., 2021).

중도탈락 학습자와 수강지속 학습자 간의 차이를 살펴보고자 랜덤 포레스트 분석 결과에서 중도탈락 예측 요인으로 도출된 설명 변수를 투입하여 독립표본 t 검정을 실시하였으며, 통계 분석은 SPSS 26.0을 활용하였다.

IV. 연구 결과

1. 랜덤 포레스트 분석 결과

본 연구에서는 안정적으로 다양한 예측변수의 영향을 분석할 수 있는 랜덤 포레스트를 활용하여 중도탈락을 예측하는 요인을 분석하였다. 이를 위해 표본을 훈련자료와 시험자료로 7:3의 비율로 나누어 모형의 예측 성과를 측정하였으며, 랜덤 포레스트 분석을 통해 도출된 예측 모형의 예측 성과를 비교하기 위해서 정분류율(Accuracy), 특이도(Specificity), 민감도(Sensitivity) 값을 활용하였다. 정분류율은 분석 대상 중 중도탈락 학습자와 수강지속 학습자 모든 사례들 중 원래 집단에 속하는 것으로 정확히 예측한 비율을 의미한다. 특이도는 수강지속 학습자를 대상으로 수강지속 학습자를 정확하게 예측한 비율을 의미하며, 민감도는 중도탈락 학습자를 대상으로 중도탈락 학습자를 정확하게 예측한 비율을 나타낸다. 정분류율, 민감도, 그리고 특이도 값이 높을수록 모형의 예측력이 높아짐을 의미한다.

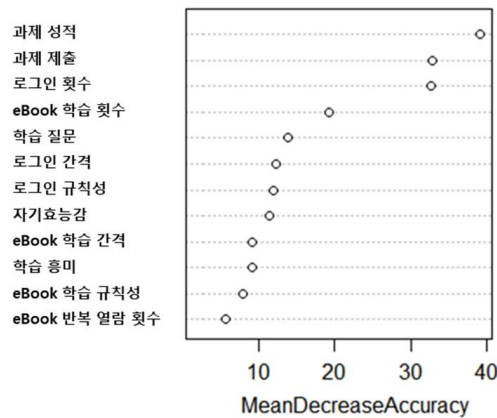
<표 6>은 훈련자료와 시험자료에 대한 랜덤 포레스트 분류 결과이다. 훈련자료의 정분류율, 민감도, 특이도는 100%로 높게 나타남을 확인할 수 있었다. 시험자료에서도 정분류율 83.8%, 민감도 72.7%, 특이도 92.7%의 높은 예측률을 보여 본 예측 모형의 성능이 신뢰할 만한 수준임을 확인하였다.

〈표 6〉 랜덤 포레스트의 정분류율, 민감도, 특이도

구분	정분류율	민감도	특이도
훈련자료	1.0000	1.0000	1.0000
시험자료	.8378	.7273	.9268

온라인 학습자의 중도탈락 예측 요인 탐색: 랜덤 포레스트를 적용하여

다음으로 중도탈락에 영향을 주는 예측변수의 상대적 중요도를 분석하기 위해 모형 정확도 개선 지수를 확인하였다. 랜덤 포레스트 분석을 통해 살펴본 온라인 학습자의 중도탈락 예측 변인의 중요도 지수는 [그림 2]와 같다. 랜덤 포레스트 분석에서 중요도 영향 변인을 산출할 때 MDA와 MDG가 사용되고 있는데, MDA가 정규분포를 따르고 랜덤 포레스트 분석에서 효율적인 지표로 활용되고 있어, 본 연구에서는 MDA 지수를

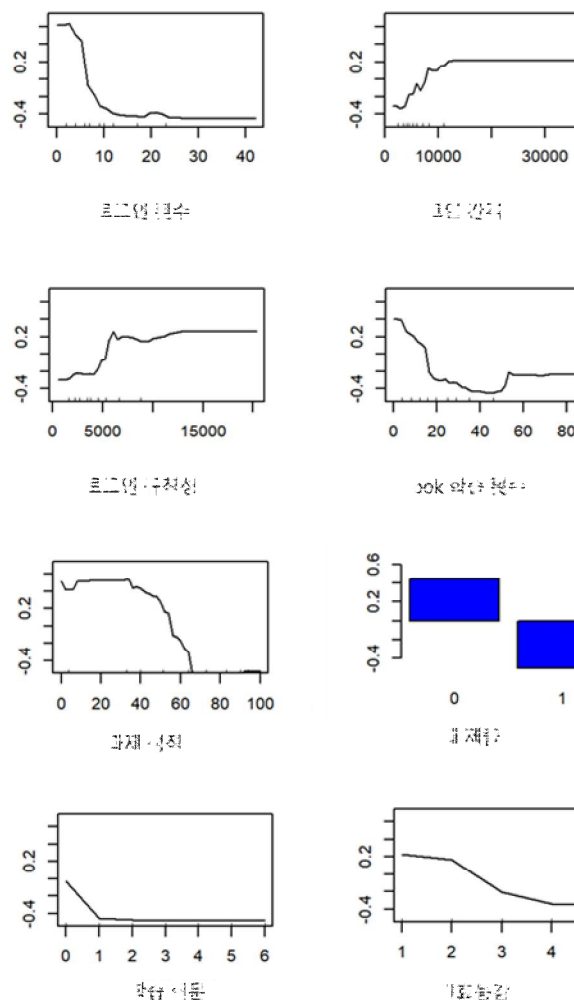


[그림 2] 중도탈락 주요 예측 변인

〈표 7〉 중도탈락 주요 예측 변인 중요도 지수

구분	예측 변수	중요도 지수
학습활동 데이터	과제 성적	30.67
	과제 제출	25.13
	로그인 횟수	23.51
학습 콘텐츠 데이터	eBook 학습 횟수	13.36
학습활동 데이터	학습 질문	9.27
	로그인 간격	8.24
	로그인 규칙성	8.12
콘텐츠 인식	자기효능감	7.59
학습 콘텐츠 데이터	eBook 학습 간격	6.65
콘텐츠 인식	학습 흥미	5.83
학습 콘텐츠 데이터	eBook 학습 규칙성	5.13
	eBook 반복 열람 횟수	4.03

기준으로 중요도를 판단하였다(Strobl et al., 2009). 선행연구(Strobl et al., 2009)에 의하면 중요도 지수 2 또는 3을 기준 값으로 중요도를 제시하고 있는데, 본 연구에서 데이터를 예측 모형에 투입한 결과 12개의 예측 변수의 중요도(MDA) 값이 3 이상인 것으로 나타났다(<표 7>). 이 중 랜덤 포레스트 분석을 10회 반복 실시한 결과 중요도 지수에서 상위에 반복적으로 나타난 변수를 확인하였는데, 분석에 투입된 예측 변수 중 로그인 횟수, 로그인 간격, 로그인 규칙성, 학습 질문, 과제 제출, 과제 성적, eBook 학습 횟수, 자기효능감이 중도탈락 예측하는 주요 상위 8개 변수로 선정되었다. 중도탈락 예측 요인의 중요도 지수가 가장 높은 변수는 과제 성적으로 나타났으며, 과제 제출이



[그림 3] 중도탈락 주요 예측 요인의 부분 의존성 도표

그 뒤를 이었다. 그 다음으로는 로그인 횟수와 eBook 학습 횟수가 중도탈락을 예측하는 주요 변인으로 나타났으며, 질문 횟수 또한 중도탈락에 중요한 역할을 하고 있음을 확인하였다. 또한 로그인 간격과 로그인 규칙성 또한 중도탈락에 상당 부분을 기여하는 것으로 나타났다. 중도탈락 예측 요인 8개의 중요도 지수 중 마지막으로 도출된 변인은 학습자 내적 변인에 해당하는 자기효능감으로 나타났다.

랜덤 포레스트 모형에서는 중요도 지수가 높은 변수를 선택하여 부분 의존성 도표를 그리는 것이 일반적이며, 이 부분 의존성 도표를 통해서 예측하고자 하는 변인과 설명 변인 간의 관계를 좀 더 상세히 알 수 있다(유진은, 2015). 중도탈락 예측 중요도 지수가 높은 8개 변인에 대해 부분 의존성 도표를 작성하였다(그림 3).

중요도 지수가 높게 나타난 중도탈락 예측 변인 중 영역별로 나누어 주요 결과를 살펴보면, 로그인 횟수가 10회 이하, 로그인 간격이 약 7일(10,000분) 이상, 로그인을 규칙적으로 하지 않을 경우 중도탈락 가능성이 높은 것으로 나타났다. 또한 과제 성적이 60점 이하이거나 과제 제출, 학습 질문을 하지 않을 경우 중도탈락 가능성이 높아지는 것을 확인할 수 있었으며, 자기효능감이 낮은 학생일수록 중도탈락 집단으로 분류될 확률이 증가하는 것으로 나타났다. 한편, eBook 학습 횟수의 경우 20회 ~ 50회에 중도탈락 가능성이 감소하는 양상을 보였으며, 50회 이상에서는 중도탈락 가능성이 조금 증가하고 20회 이하에서는 급격하게 증가하는 패턴을 관찰할 수 있었다.

2. 중도탈락 학습자와 수강지속 학습자 차이 분석 결과

온라인 교육에서 중도탈락 주요 예측 요인으로 탐색된 8개의 변인에서 중도탈락 학습자와 수강지속 학습자 간의 차이가 나타나는지 살펴보기 위해 중도탈락 주요 예측 요인으로 도출된 8개의 변인을 투입하여 두 집단 간 독립표본 t 검증을 실시하였다.

투입된 모든 예측 변인(과제 성적, 과제 제출, 로그인 횟수, eBook 학습 횟수, 학습 질문 횟수, 로그인 간격, 로그인 규칙성, 자기효능감)에서 두 집단 간 평균 차이가 통계적으로 유의한 것으로 나타났다. 과제 성적은 수강지속 학습자 71.79($SD=25.14$)가 중도탈락 학습자 40.26($SD=30.77$) 보다 높은 것으로 나타났으며($t=-10.488$, $p<.001$) 과제 제출 또한 중도탈락 학습자 0.16($SD=.37$) 보다 수강지속 학습자 0.85($SD=.35$)가 더 많이 한 것으로 나타났다($t=-19.391$ $p<.001$). 로그인 횟수 또한 중도탈락 학습자 6.03($SD=5.17$)보다 수강지속 학습자 15.65($SD=8.88$)가 더 높은 것으로 나타났다($t=-16.375$, $p<.001$). eBook 학습 횟수의 경우 중도탈락 학습자 13.05($SD=11.79$)보다 수강지속 학습자 31.34($SD=23.42$)가 더 높은 것으로 나타났으며($t=-12.918$ $p<.001$), 학습 질문

횃수도 중도탈락 학습자 0.02($SD=.13$), 수강지속 학습자 0.58($SD=1.63$)로 평균 차이가 통계적으로 유의한 것으로 나타났다($t=-8.751$, $p<.001$). 중도탈락 학습자의 로그인 간격은 평균 7898.10($SD=5322.88$)로 수강지속 학습자의 평균 로그인 간격 4845.28($SD=2784.00$)보다 높아 수강지속 학습자가 중도탈락 학습자 보다 자주 로그인을 하는 것으로 확인되었다($t=6.601$, $p<.001$). 로그인 규칙성도 중도탈락 학습자 6015.43($SD=4024.42$), 수강지속 학습자 3722.45($SD=2380.65$)로 로그인 간격과 동일하게 수강지속 학습자가 중도탈락 학습자보다 더 규칙적으로 로그인 한 것으로 확인되었다($t=5.983$, $p<.001$). 중도탈락 예측 변인으로 탐색된 자기효능감 변인은 중도탈락 학습자 2.10($SD=1.18$), 수강지속 학습자 3.59($SD=1.23$)로 나타나 수강지속 학습자의 자기효능감이 월등히 높은 것을 확인할 수 있었다($t=-12.185$, $p<.001$).

이상 중도탈락 주요 예측 요인으로 도출된 변인에서 중도탈락 학습자와 수강지속 학습자 간의 차이를 정리해 보면, 수강지속 학습자가 중도탈락 학습자보다 과제 성적

〈표 8〉 중도탈락 학습자와 수강지속 학습자의 학습 행태 차이 분석 결과

학습행태	집단	사례 수	평균	표준편차	t
과제 성적	중도탈락 학습자	117	40.26	30.77	-10.488***
	수강지속 학습자	671	71.79	25.14	
과제 제출	중도탈락 학습자	117	0.16	0.37	-19.391***
	수강지속 학습자	671	0.85	0.35	
로그인 횃수	중도탈락 학습자	117	6.03	5.17	-16.375***
	수강지속 학습자	671	15.65	8.88	
eBook 학습 횃수	중도탈락 학습자	117	13.05	11.79	-12.918***
	수강지속 학습자	671	31.34	23.42	
학습 질문 횃수	중도탈락 학습자	117	0.02	0.13	-8.751***
	수강지속 학습자	671	0.58	1.63	
로그인 간격	중도탈락 학습자	117	7898.10	5322.83	6.601***
	수강지속 학습자	671	4845.28	2784.00	
로그인 규칙성	중도탈락 학습자	117	6015.43	4024.42	5.983***
	수강지속 학습자	671	3722.45	2380.65	
자기효능감	중도탈락 학습자	117	2.10	1.18	-12.185***
	수강지속 학습자	671	3.59	1.23	

*** $p < .001$

이 높은 것으로 나타났으며, 과제 제출 또한 많이 하는 것으로 나타났다. 수강지속 학습자는 중도탈락 학습자보다 로그인 횟수, eBook 학습 횟수, 학습 질문 횟수가 더 많은 것으로 나타났으며, 수강지속 학습자가 중도탈락 학습자보다 더 자주 규칙적으로 학습에 참여하는 것을 알 수 있었다. 마지막으로, 수강지속 학습자의 자기효능감은 중도탈락 학습자보다 더 높은 것으로 확인되었다.

V. 논의 및 결론

본 연구에서는 중고등학생이 자발적으로 참여하여 학습을 진행하는 온라인 교육의 맥락에서 학생들이 수강 철회, 즉 중도탈락 의사를 밝히는 전반 6주간의 온라인 학습 행태와 교육 콘텐츠에 대한 인식을 바탕으로 중도탈락 예측력이 높은 변인을 탐색하였다. 또한 분석 결과 도출된 중도탈락 주요 예측 변인이 중도탈락 학습자와 수강지속 학습자 간에 차이가 있는지 살펴보았다. 이를 위해 온라인 교육을 수강한 학생 788명을 중도탈락 여부에 따라 집단을 구분하였으며 학습관리시스템(LMS)에서 추출할 수 있는 학습자 변인 데이터, 학습활동 데이터와 학습 콘텐츠 데이터, 학습자가 과제수행 후 콘텐츠 인식 설문에 응답한 데이터를 중심으로 예측 변인을 구성하여 분석하였다. 온라인 교육에서 중도탈락을 예측하는 요인을 탐색하기 위하여 랜덤 포레스트 분석을 실시하였으며, 중도탈락 주요 예측 요인으로 도출된 변인을 가지고 중도탈락 학습자와 수강지속 학습자의 차이를 독립표본 t 검정으로 분석하였다.

랜덤 포레스트 분석 결과 본 예측 모형의 성능이 신뢰할 만한 수준임을 확인하였으며, 예측 변인의 중요도를 고려하여 중도탈락을 예측하는 상위 주요 8개 변인을 도출했을 때 과제 성적, 과제 제출 여부, 로그인 횟수, eBook 학습 횟수, 학습 질문 횟수, 로그인 간격, 로그인 규칙성, 자기효능감 순으로 포함되었다. 독립표본 t 검정 결과 예측 요인으로 나타난 8개 변인 모두 중도탈락 학습자와 수강지속 학습자 간에 유의미한 차이가 있는 것으로 나타나 예측 모형이 타당함을 확인할 수 있었다.

이와 같은 결과를 선행연구를 바탕으로 논의하면 다음과 같다.

첫째, 학습자의 성적, 과제 제출 수 및 여부는 선행연구에서 중도탈락에 영향을 미치는 요인으로 확인된 바 있다(권선아 외, 2020; 정영란 2016, 2020; Choi & Kim, 2017). 다만 본 연구가 기존의 연구와 차별되는 점은 기존의 연구에서는 성적이 중도탈락에 영향을 미치는 변인이라는 점을 밝혀 중도탈락의 원인을 설명하는 방식이라면, 머신러닝을 적용한 본 연구의 결과는 다양한 데이터 중 중도탈락에 속할 가능성을 가장 잘

예측하는 변인이 중도탈락 전에 학생에게 부여된 과제의 성적이라는 것이다. 기존의 연구에서 성적이 학업성취도가 누적된 결과라면 중도탈락 학생들의 성적이 낮은 것은 당연한 결과일 수 있으며, 이에 본 연구에서는 초기 과제 제출 여부 및 점수가 중도탈락에 의미 있게 예측함을 확인하였다.

둘째, 로그인 횟수, eBook 학습 횟수, 학습 질문 횟수, 로그인 간격, 로그인 규칙성 등과 같이 학습 행태 변인 역시 중도탈락을 예측하는 것으로 나타났다. 이러한 변인은 학습분석 연구에서 일관되게 학업성취도, 이수 및 미이수, 중도탈락 등 학습의 결과에 영향을 미치는 것으로 나타난 변인이었다(이성혜 외, 2021; 조일현, 김윤미, 2013; 한정윤, 이성혜, 2019; You, 2016). 학습자가 보다 자주, 규칙적으로 온라인 학습 시스템에 로그인하여 학습 콘텐츠에 빈번하게 접근하고, 적극적으로 질문을 올릴수록 성적이 높아지거나 중도탈락 가능성이 낮아질 수 있다는 결과는 학습자 데이터 기반의 연구에서 일관성 있게 제시되고 있다. 이는 학습분석 연구에서 이들을 타당한 행동 특성 변인으로 고려할 수 있으며, 이러한 결과를 바탕으로 학습자에게 개별화된 의미 있는 피드백을 제공할 수 있음을 의미한다. 예컨대, 개별 학습자의 데이터를 분석하여 학습자가 보다 자주, 규칙적으로 로그인하거나 학습에 참여하도록 촉진하는 메시지를 제공하는 것이 가능하며, 이를 통해 보다 긍정적인 학습 결과를 기대할 수 있을 것이다.

셋째, 자기효능감 역시 선행연구에서 중도탈락에 영향을 미치는 요인 중 하나로 제시되고 있다(Yukselturk et al., 2014). 본 연구 역시 학습자가 중도탈락 이전 제시된 과제에 대해 느끼는 자기효능감이 중도탈락을 예측하는 것으로 나타났다. 자기효능감은 학업 수행에 요구되는 자신의 능력에 대한 판단을 의미한다(Bandura, 1997). 자기효능감은 학습 참여 및 학업지속과 관련이 높은 변인이며(Pajares, 2002; Urdan & Parajes, 2006), 온라인 교육에서 학습자의 성공적인 학습에도 영향을 미치는 것으로 보고되고 있다(Wang & Newlin, 2002). 자기효능감이 낮은 학생들은 과제를 미루는 경향이 나타나며(Haycock et al., 1998), 온라인 학습 환경에서 이러한 지연행동은 결과적으로 중도탈락에 영향을 주는 것이다(Klingsieck et al., 2012).

본 연구는 머신러닝 기법인 랜덤 포레스트 모델을 활용하여 온라인 교육 시스템에서 추출 가능한 예측 변인을 투입하였으며, 이를 통해 상대적으로 중요한 예측 변인을 도출하고자 시도하였다. 기존의 통계적 분석 방법이 이론을 기반으로 연구모형을 제시하고 가설을 이를 검증하는 것이라면, 머신러닝 기반의 예측 분석은 명확한 연구모형을 기반으로 하여 통계적 검증을 하기보다는 데이터로부터 패턴을 발견하여 앞으로의 양상을 예측하는 데 중점을 두는 방법이라고 할 수 있다(유진은, 2019). 따라서 중도탈락 연구의 경우 기존의 통계분석 방법이 중도탈락을 설명하기 위한 모델을 찾는 것이

라면, 머신러닝은 학습자가 중도탈락 또는 완료 중 어디에 속할 것인지는 예측할 수 있도록 하는 것이다. 이러한 머신러닝의 특징을 고려할 때, 본 연구는 중도탈락이 확정된 시점에서의 데이터를 통해 예측 모형을 구축했다는 점에서 한계가 있으며, 향후 이 모형을 새로운 학기의 데이터에 적용하여 모형의 정확도를 검증하고 중도탈락 위험군을 예측해 볼 필요가 있다. 또한, 머신러닝 기법 중 정확성과 예측도가 높은 랜덤 포레스트를 적용하여 분석하였는데, 다양한 머신러닝 분석 방법을 적용, 비교한다면 예측 정확도가 더 높은 모형을 구축할 수 있을 것이다.

중도탈락 연구의 공통적인 어려움 중의 하나는 중도탈락 학습자와 학습지속 또는 완료자 간 샘플 수의 불균형이며, 중도탈락이 발생하면 이후 해당 학습자에 대한 데이터를 확보하기 어렵기 때문에 이러한 차이가 결과적으로 예측 모델의 정확도와 일반화에 영향을 미친다는 점이다(Dalipi et al., 2018). 본 연구 역시 6주 시점을 기준으로 중도탈락 학습자 117명, 수강지속 학습자 671명으로 분류되어 이러한 한계를 극복하지 못하였으며, 또한 중도탈락 학습자의 데이터 특성상 결측률이 높은 변인들을 분석에서 제외하였다. 따라서 중도탈락을 정확하게 예측할 수 있는 시점 규명 및 다양한 초기 데이터가 중도탈락 예측에 있어 매우 중요한 이슈이며, 이를 위한 후속 연구가 지속적으로 이루어져야 할 것이다.

서론에서 언급했듯이 중도탈락자의 학습경험을 조사하여 분석하는 것은 한계가 있었다. 대부분의 온라인 교육기관에서 교육이 끝난 후 온라인 교육에 대한 경험, 만족도, 개선 사항 등에 대한 조사를 실시하지만 응답률이 매우 낮으며, 응답자 또한 교육을 끝까지 이수한 학습자에 편중되어 있기 때문이다. 또한 많은 중도탈락 연구들이 교육 프로그램이 끝난 시점에서 로그인 횟수, 게시글 수 등과 같이 누적된 학습행동 데이터를 분석에 활용해 왔는데, 이는 앞서 성과와 마찬가지로 중도에 탈락한 학생들과 수강완료 학생들 간에 누적 활동 수가 차이가 나는 것은 당연한 결과일 것이다. 이에 머신러닝은 중도탈락 이전 학습자의 데이터를 기반으로 중도탈락을 예측하고 조기에 원인을 파악하거나 혹은 중도탈락이 예측되는 시점에서 조사를 실시하는 등 전통적인 온라인 교육 프로그램 평가에 획기적인 대안을 제시한다(Whitchill et al., 2015). 현재 다양한 온라인 교육의 맥락에서 중도탈락을 정확하게 예측하기 위한 다양한 알고리즘을 탐색하는 연구들이 주를 이루고 있으며, 또한 이를 바탕으로 중도탈락 위험 학습자들에게 적절한 시점에 개별화된 처치를 제공하기 위한 방법들이 탐색되고 있다. 특히, 중도탈락이 과정이 진행되는 동안 어느 시점에서도 발생할 수 있어 데이터를 기반으로 이를 실시간으로 탐지하는 것이 매우 중요한 과제이다.

앞서 제시한 바와 같이 본 연구는 여러 가지 한계를 지니지만, 기존 머신러닝 기반

의 중도탈락 연구가 오프라인 대학, MOOC 등 대규모 온라인 강의와 같은 맥락에서 이루어졌던 것과 달리 본 연구는 중고등학생 대상의 온라인 교육에서 수행한 점에서 의미가 있다고 할 수 있다. 향후, 본 연구에서 제시된 모델을 지속적으로 개선하고 성인이 아닌 중고등학생 대상 온라인 교육의 맥락에 적응적으로 활용할 수 있도록 지속적인 연구가 이루어져야 할 것이다. 마지막으로 본 연구에서는 중도탈락과 관련된 선행연구를 기반으로 해당 프로그램의 맥락에서 수집 가능한 데이터를 분석에 활용했다는 점에서 한계가 있다. 특히, 선행연구에서 온라인 학습 과정에서 상호작용과 관련된 행동은 중도탈락을 예측하는 주요 변인 중 하나였으나(정영란, 2016; Kashyap & Nayak, 2018), 본 온라인 교육 프로그램의 맥락에서 상호작용 행동이 의미있는 수준으로 발생하지 않아 이를 분석에서 제외하였다. 이에 향후 보다 타당한 중도탈락 연구를 위해서는 상호작용 데이터를 포함하여 온라인 학습에서 중도탈락에 영향을 미칠 수 있는 보다 다양한 학습자 및 학습행동 데이터의 탐색과 분석이 요구된다.

참고문헌

- 강성국, 금지현, 손찬희, 이쌍철, 이은철, 안성훈, 이경상 (2014). 교육소의 청소년을 위한 방송통신 중·고등학교 운영방안 연구. 한국교육개발원.
- (Translated in English) Kang, S., Keum, J., Son, C., Lee, S., Lee, E., Ahn, S., & Lee, K. (2014). *A study on the operation of broadcasting and communication middle and high schools for educational underprivileged youth*. Korean Educational Development Institute.
- 권선아, 김명진, 서희정, 김민정 (2020). 원격고등교육에서 성인학습자의 중도탈락에 관한 로지스틱 회귀분석. 평생학습사회, 16(4), 149-169.
- (Translated in English) Kyun, S., Kim, M., Seo, H., & Kim, M. (2020). Logistic regressions analysis of the dropout of adult-learners in higher distance education. *Journal of lifelong learning society*, 16(4), 149-169.
- 권혜진 (2010). 개인, 교육기관, 사회적 변인이 사이버대 재학생의 중도탈락의도 결정에 미치는 영향. 한국콘텐츠학회논문지, 10(3), 404-412.
- (Translated in English) Kwon, H. (2010). The effects of personal, institutional, social variables on determination of the cyber university students' dropout intention. *The Journal of the Korea Contents Association*, 10(3), 404-412.
- 김미림, 박민호 (2019). 랜덤 포레스트를 활용한 대학생의 최초 취업 사교육 참여 시점 별 특성 분석. 교육연구논총, 40(1), 1-33.
- (Translated in English) Kim, M., & Park, M. (2019). An analysis of the characteristics of college students according to first-time participation in private tutoring using a random forest. *CNU Journal of Educational Studies*, 40(1), 1-33.
- 김지현 (2013). 사이버대학생의 중도탈락 결정요인에 관한 연구. 사이버교육연구, 7(2), 1-16.
- (Translated in English) Kim, J. (2013). Study on the determinants of student drop-out in online universities. *Journal of Cyber Education*, 7(2), 1-16.
- 남신동, 정영숙, 황지원, 정연희 (2014). 방송대생 학업지속률 제고를 위한 학업중단경험의 제유형 및 결정요인의 분석. 한국방송통신대학교 원격교육연구소.
- (Translated in English) Nam, S., Jung, Y., Hwang, J., & Jung, Y. (2014). *Analysis of types and determinants of academic dropout experience for improving the academic continuity of KNOU students*. Institute of Distance Education, Korea Open National University.
- 노민정, 유진은 (2019). Adaptive LASSO를 통한 진로결정 관련 변수 탐색. 열린교육연구,

27(4), 133-155.

(Translated in English) Rho, M., & Yoo, J. E. (2019). Exploration of variables relating to career decisions via Adaptive LASSO. *The Journal of Yeolin Education*, 27(4), 133-155.

대학알리미 (2021). 중도탈락 학생 현황(대학).

<https://www.academyinfo.go.kr/uipnh/unt/unmcom/RdViewer.dodptj>에서 검색

(Translated in English) Higher Education in KOREA. (2021). *Status of dropout students(University)*.

<https://www.academyinfo.go.kr/uipnh/unt/unmcom/RdViewer.do>

박미현, 허 균 (2021). 머신러닝 분류기법을 적용한 중학생의 학습 부진 예측모형 개발연구: 대구교육종단자료를 중심으로. *교육공학연구*, 37(3), 627-648.

(Translated in English) Park, M., & Heo, G. (2021). Development of prediction model for middle school students' underachievement using machine learning classification techniques: A focus on DELS. *Journal of educational technology*, 37(3), 627-648.

박소영, 정혜원 (2021). 랜덤 포레스트를 활용한 고등학생의 진로개발역량 예측변수 탐색. *열린교육연구*, 29(1), 239-265.

(Translated in English) Park, S., & Chung, H. (2021). Exploring predictors affecting career development competence of high school students using random forest. *The Journal of Yeolin Education*, 29(1), 239-265.

서선주 (2004). 사이버대학생들의 중도탈락 요인에 관한 연구 [석사학위논문]. 중앙대학교. (Translated in English) Seo, S. J. (2004). *A study on the cause of dropout by the students of Cyber University* [Unpublished master's thesis]. ChungAng University.

손윤희, 박현정, 박민호 (2020). 랜덤 포레스트를 활용한 읽기소양 수준에 따른 집단 결정요인 분석 PISA 2018 자료를 중심으로. *아시아교육연구*, 21(1), 191-215.

(Translated in English) Son, Y., Park, H., & Park, M. (2020). Random forest analysis of factor influencing the students's reading literacy levels: using PISA 2018 korea data. *Asian Journal of Education*, 21(1), 191-215.

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

(Translated in English) Shin, J., & Choi, J. (2019). Developing the prediction model of at-risk freshmen students and analyzing characteristics of cluster based on learning analytics. *Journal of educational technology*, 35(2), 425-454.

유지원 (2014). 일반대학에서 교양 e-러닝 강좌의 중도탈락 예측모형 개발과 조기 판별 가능성 탐색. *컴퓨터교육학회 논문지*, 17(1), 1-12.

- (Translated in English) You, J. (2014). Dropout prediction modeling and investigating the feasibility of early detection in e-Learning courses. *The Journal of Korean association of computer education*, 17(1), 1-12.
- 유진은 (2015). 랜덤 포레스트: 의사결정나무의 대안으로서의 데이터 마이닝 기법. *교육평가연구*, 28(2), 427-448.
- (Translated in English) Yoo, J. E. (2015). Random forests, an alternative data mining technique to decision tree. *Journal of Educational Evaluation*, 28(2), 427-448.
- 유진은 (2019). 기계학습: 대용량/패널자료와 학습분석학 자료 분석으로의 활용. *교육공학연구*, 35(2), 313-338.
- (Translated in English) Yoo, J. E. (2019). Machine learning for large-scale/panel data and learning analytics data analysis. *The Journal of educational technology*, 35(2), 313-338.
- 유진은 (2020). 기계학습 기법을 활용한 플립러닝 강좌의 LMS 로그파일 분석 사례 연구. *열린교육연구*, 28(5), 79-102.
- (Translated in English) Yoo, J. E. (2020). Learning management system Log Data Analysis via machine Learning: A case study from flipped learning. *The Journal of Yeolin Education*, 28(5), 79-102.
- 유진은, 김형관, 노민정 (2020). Group Mnet 기계학습 기법을 통한 중학생의 끈기 (grit) 관련 변수 탐색. *한국청소년연구*, 31(1), 157-182.
- (Translated in English) Yoo, J. E., Kim, H., & Rho, M. (2020). An Exploration of the variables relating to middle school students' grit via a machine learning technique, Group Mnet. *The Studies on Korean Youth*, 31(1), 157-182.
- 이대현, 조희석 (2021). 학습분석 기반 중도탈락 예측 모델. *한국컴퓨터교육학회 학술대회논문집*, 25(1), 211-213.
- (Translated in English) Lee, D., & Cho, H. (2021). A dropout prediction model based on learning analytics. *The Korean Association of Computer Education* 25(1), 211-213.
- 이성혜, 박혜진, 성은모 (2021). 온라인학습환경에서 학업성취도에 영향을 미치는 자기조절학습 변인 및 학습행동 데이터 특성 탐색. *교육정보미디어연구*, 27(2), 723-748.
- (Translated in English) Lee, S., Park, H., & Sung, E. (2021). Exploration of self-regulated learning variables and learning behavior data affecting academic achievement in an online learning environment. *The Journal of Educational Information and Media*, 27(2), 723-748.
- 이은정, 송영수, 김지하, 오수현 (2020). 랜덤 포레스트를 활용한 4년제 대학 중도탈락률 예측 요인 탐색: 대학 수준 결정요인을 중심으로. *교육공학연구* 36(1), 191-219.

- (Translated in English) Lee, E., Song, Y., Kim, J., & Oh, S. (2020). An exploratory study on determinants predicting the dropout rate of 4-year universities using random forest: focusing on the institutional level factors. *Journal of educational technology*, 36(1), 191-219.
- 이종현, 조규락 (2021). 머신러닝을 활용한 중학교 수학 기초학력 미달 비율 예측모형 탐구. *교육공학연구*, 37(1), 95-129
- (Translated in English) Lee, J., & Cho, K. (2021). A study on the prediction model for the ratio of mathematics low-performing students in middle school using machine learning. *Journal of educational technology*, 37(1), 95-129.
- 이지은 (2019). 학생 중도탈락 예측지수에 관한 사후검증 연구. *한국빅데이터학회지*, 4(2), 175-183.
- (Translated in English) Lee, J. (2019). Post-examination analysis on the student dropout prediction index. *The journal of Bigdata*, 4(2), 175-183.
- 이현우, 이종문, 차윤미 (2021). 머신러닝 기반의 학업성취 예측 모형 탐색. *교육방법연구*, 33(1), 29-26.
- (Translated in English) Lee, H. W., Lee, J. M., & Cha, Y. M. (2021). Exploring a model for predicting academic achievement with machine learning for off-line courses in higher education. *Journal of Educational Metodology*, 33(1), 29-26.
- 임연옥 (2007). 사이버대학 학습자관련 변인과 중도탈락 간의 관계 규명을 위한 실증적 연구. *정보교육학회논문지*, 11(2), 205-219.
- (Translated in English) Im, Y. (2007). A subatantial study on the relationship between students' variables and dropout in cyber university. *Journal Of The Korean Association of information Education*, 11(2), 205-219.
- 전주성 (2010). 사이버대학의 잠재적 중도탈락자 예측에 관한 연구. *한국성인교육학회*, 13(1), 121-139.
- (Translated in English) Jun, J. (2010). Andragogy today: Interdisciplinary journal of adult & continuing education. *Interdisciplinary Journal of adult & continuing education*, 13(1), 121-139.
- 정선정 (2005). 직업교육 이러닝(e-Learning)의 중도탈락 원인 분석. [석사학위논문]. 이화여자대학교.
- (Translated in English) Jeong, S. J. (2005). *The analysis of the cause behind attrition in e-learning vocational education* [Unpublished master's dissertation]. Ewha Womans University.
- 정영란 (2016). 사이버대학에서의 재등록률 영향 요인 분석. *교육방법연구*, 28(4), 791-814.

- (Translated in English) Joung, Y. (2016). Analysis on factors affecting the retention rate in cyber university. *The Korean Journal Of Educational Methodology Studies*, 28(4), 791-814.
- 정영란 (2020). 학습분석학 기반의 사이버대학의 중도탈락 예측 분석. *교육방법연구*, 32(2), 205-232.
- (Translated in English) Joung, Y. (2020). A prediction analysis on the dropout of cyber university based on learning analytics. *The Korean Journal Of Educational Methodology Studies*, 32(2), 205-232.
- 정주영, 이정원 (2017). 사이버대학생의 중도탈락 의도에 영향을 미치는 요인 탐색연구. *한국교육문제연구*, 35(4), 149-168.
- (Translated in English) Jung, J., & Lee, J. (2017). An exploratory study on dropout intention of cyber university students. *Journal of the research institute of Korean education*, 35(4), 149-168.
- 정혜령, 윤창국, 우영희 (2015). 방송대 학습자의 학업지속 장애요인 극복 방안. 한국방송통신대학교 원격교육연구소.
- (Translated in English) Jung, H., Yoon, C., & Woo, Y. (2015). *A method for overcoming obstacles to academic continuity of KNOU learners*. Institute of Distance Education, Korea Open National University.
- 정혜원, 박소영, 김정인, 김아름 (2021). 청소년 읽기 소양과 삶의 만족도의 영향변인 탐색: PISA 2018 한국 필란드 국제비교. *교육과정평가연구*, 24(1), 123-152.
- (Translated in English) Chung, H., Park, S., Kim, J., & Kim, A. (2021). Exploring variables affecting adolescents' reading literacy and life satisfaction: PISA 2018 international comparison of Korea and Finland. *The Journal of Curriculum and Evaluation*, 24(1), 123-152.
- 조인식 (2020). K-MOOC(한국형 공개 온라인 강좌)의 현황과 개선과제. 국회입법조사처.
- (Translated in English) Cho, I. (2020). *Current status and improvement tasks of K-MOOC (Korean open online course)*. National Assembly Research Service.
- 조일현, 김윤미 (2013). 이러닝에서 학습자의 시간관리 전략이 학업성취도에 미치는 영향: 학습분석학적 접근. *교육정보미디어연구*, 19(1), 83-107.
- (Translated in English) Jo, I., & Kim, Y. (2013). Impact of learner's time management strategies on achievement in an e-learning environment: A learning analytics approach. *The Journal of Educational Information and Media*, 19(1), 83-107.
- 주영주, 심우진, 김수미 (2008). 기업 사이버교육에서 학습자의 중도탈락에 대한 결정요인 분석. *교육정보미디어연구*, 14(4), 5-25.
- (Translated in English) Joo, Y., Shim, W., & Kim, S. (2008). A study on the factors affecting

- the drop-out in corporate cyber learning. *Journal of Korean Association for Educational Information and Media*, 14(4), 5-25.
- 주영주, 장미진, 이현주 (2007). 사이버대학 학생의 중도탈락 경험에 근거한 중도탈락 요인에 관한 질적 연구. *교육정보미디어연구*, 13(3), 209-233.
- (Translated in English) Joo, Y., Jang, M., & Lee, H. (2007). An in-depth analysis of dropout factors based on cyber university student's dropout experiences. *Journal of Korean Association for Educational Information and Media*, 13(3), 209-233.
- 주영주, 정애경, 유나연, 이상희 (2012). 사이버대학에서 인지된 고립감, 조직의 지원, 만족도, 학습지속의향간 구조적 관계 규명. *전자공학회논문지*, 49(10), 240-250.
- (Translated in English) Joo, Y., Chung, A., Yoo, N., & Yi, S. (2012). Investigating the structural relationship among perceived isolation, organizational support, satisfaction and consistency in cyber university. *Journal of the Institute of Electronics and Information Engineers*, 49(10), 240-250.
- 한정윤, 이성혜 (2019). 온라인 소프트웨어 교육에서 학습자의 자기조절학습 관련 특성에 기반한 온라인학습 유형 분석: 계층적 군집 분석 기법을 활용하여. *컴퓨터교육학회논문지*, 22(5), 51-65.
- (Translated in English) Han, J., & Lee, S. (2019). Investigating online learning types based on self-regulated learning in online software education: Applying hierarchical cluster analysis. *The Journal of Korean Association of Computer Education*, 22(5), 51-65.
- 황현정, 박솔잎, 박형용 (2021). 학습결과 분석을 통한 원격대학 중도탈락 예측 시스템 AI 알고리즘 적용방안. *컴퓨터교육학회 논문지*, 24(5), 63-73.
- (Translated in English) Hwang, H., Park, S., & Park, H. (2021). Application of AI algorithm in distance learners' dropout prediction system by analyzing learning results. *The Korean Association of Computer Education*, 24(5), 63-73.
- Allen, J. S. (2017). *Online faculty behaviors that impact student persistence* [Unpublished doctoral dissertation]. San Diego State University, San Diego, USA.
- Bandura, A. (1997). *Self-Efficacy: The exercise of control*. W.H. Freeman and Company, New York.
- Behr, A., Giese, M., Tegum K, H., & Theune, K. (2020). Early Prediction of University Dropouts - A Random Forest Approach. *Jahrbücher für Nationalökonomie und Statistik*, 240(6), 743-789.
- Bettinger, E., Doss, C., Loeb, S., Rogers, A., & Taylor, E. (2017). The effects of class size in online college courses: Experimental evidence. *Economics of Education Review*, 58, 68-85.

- Boton, E. C., & Gregory, S. (2015). Minimizing attrition in online degree courses. *Journal of Educators Online*, 12(1), 62-90.
- Borrella, I., Caballero, S., & Ponce-Cueto, E. (2019). Predict and Intervene: Addressing the Dropout Problem in a MOOC-based Program. In *Proceedings of Sixth ACM Conference*, 1-9.
- Breiman, L. (2001). Random forests. *Machine Learning*, 45, 5-32.
- Breiman, L., & Cutler, A. (n.d.). *Random forests*.
http://www.math.usu.edu/adele/forests/cc_home.htm
- Breiman, L., Cutler, A., Liaw, A., & Wiener, M. (2018). *RandomForest: Breiman and cutler's random forests for classification and regression*. R package version, 4.6-14.
- Carr, S. (2000). As distance education comes of age, the challenge is keeping the students. *Chronicle of Higher Education*, 46(23), A39-A41.
- Castles, J. (2004). Persistence and the adult learner: Factors affecting persistence in Open University students. *Active Learning in Higher Education*, 5(2), 166-179.
- Chae, Y., & Gentry, M. (2007). Korean high school student perceptions of classroom quality: Validation research. *Gifted and Talented International*, 22(2), 68-76.
- Chen, J., Feng, J., Sun, X., Wu, N., Yang, Z., & Chen, S. (2019). MOOC dropout prediction using a hybrid algorithm based on decision tree and extreme learning machine. *Mathematical Problems in Engineering*, 1-11.
- Choi, H. J., & Kim, B. U. (2017). Factors affecting adult student dropout rates in the korean cyber-university degree programs. *Journal of Continuing Higher Education*, 66(1), 1-12
- Chung, A. Y., & Lee, S. (2019). Dropout early warning systems for high school students using machine learning. *Children and Youth Services Review*, 96, 346-353.
- Coussement, K., Phan, M., De Caigny, A., Benoit, D. F., & Raes, A. (2020). Predicting student dropout in subscription-based online learning environments: The beneficial impact of the logit leaf model, *Decis. Support Syst.* 135, 113325.
- Dalipi, K., Imran, A., & Kastrati, Z. (2018, April 17-20). MOOC dropout prediction using machine learning techniques: Review and research challenge. In *2018 IEEE Global Engineering Education Conference(EUCON)*. Santa Cruz de Tenerife, Canary Islands, Spain.
- Dass, S., Gary, K., & Cunningham, J. (2021). Predicting student dropout in self-paced MOOC course using random forest model. *Information*. 12(11), 476.
- Dupin-Bryant, P. (2004). Pre-entry variables related to retention in online distance education. *American Journal of Distance Education*, 18(4), 199-206.

- Genuer, R., Poggi, J. M., & Tuleau-Malot, C. (2010). Variable selection using random forests. *Pattern Recognition Letters*, 31, 2225-2236.
- Goodfellow, I., Bengio, Y., & Courville, A. (2016). *Deep learning*. Cambridge: The MIT Press.
- Halawa, S., Greene, D., & Mitchell, J. (2014). Dropout prediction in MOOCs using learner activity features. *Proceedings of the second European MOOC stakeholder summit*, 37(1), 58-65.
- Hamza, M., & Larocque, D. (2005). An empirical comparison of ensemble methods based on classification trees. *Journal of Statistical Computation and Simulation*, 75(8), 629-643.
- Hart, C. (2012). Factors associated with student persistence in an online program of study: A review of the literature. *Journal of Interactive Online Learning*, 11(1), 19-42.
- Haycock, L. A., McCarthy, P., & Skay, C. L. (1998). Procrastination in college students: The role of self-efficacy and anxiety. *Journal of Counseling & Development*, 76(3), 317-324.
- Impey, C. D., Wenger, M. C., & Austin, C. L. (2015). Astronomy for astronomical numbers: A worldwide massive open online class. *The International Review of Research in Open and Distributed Learning*, 16(1), 57-79.
- Ishwaran, H. (2007). Variable importance in binary regression trees and forests. *Electronic Journal of Statistics*, 1, 519-537.
- James, S., Swan, K., & Daston, C. (2016). Retention, progression and the taking of online courses. *Journal of Asynchronous Learning Network*, 20(2), 75-96.
- Jordan, K. (2014). Initial trends in enrollment and completion of massive open online courses. *International Review of Research in Open and Distance Learning*, 15(1), 133-169.
- Kashyap, A., & Nayak, A. (2018). Different machine learning models to predict dropouts in MOOCs. *2018 International Conference on Advances in Computing, Communications and Informatics (ICACCI)* (pp. 80-85). IEEE.
- Kember, D. (1989). A longitudinal-process model of drop-out from distance education. *The Journal of Higher Education*, 60(3), 278-301.
- Klingsieck, K. B., Fries, S., Horz, C., & Hofer, M. (2012). Procrastination in a distance university setting. *Distance Education*, 33(3), 295-310.
- Lee, Y., & Choi, J. (2011). A review of online course dropout research: Implications for practice and future research. *Educational Technology Research and Development*, 59(5), 593-618.
- Lee, Y., Choi, J., & Kim, T. (2013). Discriminating factors between completers of and dropouts from online learning courses. *British Journal of Educational Technology*, 44(2), 328-337.
- Levy, Y. (2007). Comparing dropouts and persistence in e-learning courses. *Computers and*

- Education*, 48(2), 185-204.
- Lim, J. M. (2016). Predicting successful completion using student delay indicators in undergraduate self-paced online courses. *Distance Education*, 37(3), 317-332.
- Mohammed, M. B., Zulkafli, H. S., Adam, M., Ali, N., & Baba, I. (2021). Comparison of five imputation methods in handling missing data in a continuous frequency table. *Computer Science, AIP Conference Proceedings* 2355.
- Moore, M., & Kearsley, G. (1996). *Distance education: A systems view*. Belmont, CA: Wadsworth.
- Morris, L. V., Wu, S., & Finnegan, C. L. (2005). Predicting retention in online general education courses. *American Journal of Distance Education*, 19(1), 23-36.
- Muse, H. E. (2005). *At-risk factors for the community college web-based student*. 20th Annual Conference on Distance Teaching and Learning.
- Opazo, D., Moreno, S., Álvarez-Miranda, E., & Pereira, J. (2021). Analysis of first-year university student dropout through machine learning models: A comparison between universities. *Mathematics*, 9(20), 2599.
- Pajares, F. (2002). *Overview of social cognitive theory and of self-efficacy*. Retrieved May 29, 2009, from <http://www.des.emory.edu/mfp/eff.html>
- Park, J., & Choi, H. (2009). Factors influencing adult learners' decision to drop out or persist in online learning. *Educational Technology & Society*, 12(4), 207-217.
- Rotgans, J. I., & Schmidt, H. G. (2011). Situational interest, task engagement, and achievement in an active-learning environment. *Learning and Instruction*, 21(1), 58-67.
- Rovai, A. P. (2003). In search of higher persistence rates in distance education online programs, *The Internet and Higher Education*, 6(1), 1-16.
- Shah, D. (2019). *By the numbers: MOOCs in 2021*. Retrieved December 1, 2021, from <https://www.classcentral.com/report/mooc-stats-2021/>
- Strobl, C., Boulesteix, A.-L., Zeileis, A., & Hothorn, T. (2007). Bias in random forest variable importance measures: Illustrations, sources and a solution. *BMC Bioinformatics*, 8, 25.
- Strobl, C., Malley, J., & Tutz, G. (2009). An introduction to recursive partitioning: Rationale, application, and characteristics of classification and regression trees, bagging, and random forests. *Psychological Methods*, 14, 323-348.
- Tinto, V. (1975). Dropout from higher education: A theoretical synthesis of recent research. *Review of Educational Research*, 45(1), 89-125.
- Toledo, I., Alborno, C., & Schneider, K. (2020). Learning analytics to explore dropout in

- online entrepreneurship education. *Psychology*, 11, 268-284.
- Urdan, T., & Parajes, F. (2006). *Self-efficacy beliefs of adolescents*. Charlotte: Information Age Publishing.
- van Buuren, S., & Groothuis-Oudshoorn, K. (2011). Mice: Multivariate imputation by chained equations in R. *Journal of Statistical Software*, 45(3), 1-67.
- Van Hulse, J., Khoshgoftaar, T. M., & Napolitano, A. (2007). Experimental perspectives on learning from imbalanced data. *Proceedings of the International Conference on Machine Learning (ICML 2007)* (pp. 935-942).
- Wang, A. Y., & Newlin, M. H. (2002). Predictors of web-student performance: The role of self-efficacy and reasons for taking an on-line class. *Computers in Human Behavior*, 18, 151-163.
- Whitehill, J., Williams, J., Lopez, G., Coleman, C., & Reich, J. (2015). *Beyond prediction: First steps toward automatic intervention in MOOC student stopout*. <https://ssrn.com/abstract=2611750> or <http://dx.doi.org/10.2139/ssrn.2611750>
- Yasmin, D. (2013). Application of the classification tree model in predicting learner dropout behaviour in open and distance learning. *Distance Education*, 34(2), 218-231.
- Yoo, J. E., & Rho, M. (2017). TIMSS 2015 korean student, teacher, and school predictor exploration and identification via random forests. *The SNU Journal of Education Research*, 26(4), 43-61.
- You, J. W. (2016). Identifying significant indicators using lms data to predict course achievement in online learning. *Internet and Higher Education*, 29, 23-30.
- Yukselturk, E., Ozekes, S., & Türel, Y. K. (2014). Predicting dropout student: An application of data mining methods in an online education program. *European Journal of Open, Distance and E-Learning*, 17, 118-133.
- Zimmerman, B. J. (2000). Self-efficacy: An essential motive to learn. *Contemporary Educational Psychology*, 25(1), 82-91.