Journal of The Korean Society of Integrative Medicine, 2024, 12(2), 155~166
https://doi.org/10.15268/ksim.2024.12.2.155
ISSN 2288-1174(Print)
ISSN 2383-9651(Online)

머신러닝을 활용한 대학생 중도탈락 위험군의 예측모델 비교 연구 : N대학 사례를 중심으로

김소현[1] · 조성현[2†]

[1]연세베스트요양병원 물리치료사, [2†]남부대학교 물리치료학과 교수

A Comparative Study of Prediction Models for College Student Dropout Risk Using Machine Learning: Focusing on the case of N university

So-Hyun Kim, PT, Ph.D[1] · Sung-Hyoun Cho, PT, Ph.D[2†]

[1]Dept. of Physical Therapy, Yonsei Best Convalescent Hospital, Physical Therapist
[2†]Dept. of Physical Therapy, Nambu University, Professor

Abstract

Purpose : This study aims to identify key factors for predicting dropout risk at the university level and to provide a foundation for policy development aimed at dropout prevention. This study explores the optimal machine learning algorithm by comparing the performance of various algorithms using data on college students' dropout risks.

Methods : We collected data on factors influencing dropout risk and propensity were collected from N University. The collected data were applied to several machine learning algorithms, including random forest, decision tree, artificial neural network, logistic regression, support vector machine (SVM), k-nearest neighbor (k-NN) classification, and Naive Bayes. The performance of these models was compared and evaluated, with a focus on predictive validity and the identification of significant dropout factors through the information gain index of machine learning.

Results : The binary logistic regression analysis showed that the year of the program, department, grades, and year of entry had a statistically significant effect on the dropout risk. The performance of each machine learning algorithm showed that random forest performed the best. The results showed that the relative importance of the predictor variables was highest for department, age, grade, and residence, in the order of whether or not they matched the school location.

Conclusion : Machine learning-based prediction of dropout risk focuses on the early identification of students at risk. The types and causes of dropout crises vary significantly among students. It is important to identify the types and causes of dropout crises so that appropriate actions and support can be taken to remove risk factors and increase protective factors. The relative importance of the factors affecting dropout risk found in this study will help guide educational prescriptions for preventing college student dropout.

Key Words : college student, dropout risk, machine learning, prediction model, random forest

[†]교신저자 : 조성현, shcho@nambu.ac.kr

제출일 : 2024년 4월 16일 | 수정일 : 2024년 5월 19일 | 게재승인일 : 2024년 5월 24일

Ⅰ. 서 론

1. 연구의 배경 및 필요성

우리나라 대학생들의 학부 활동 동향을 살펴보면, 학업 지원 프로그램에 대한 관심이 매우 높은 것으로 나타났다(Shin & Song, 2022). 이러한 프로그램은 기초 교육 부족으로 인한 경고 및 잠재적 중도탈락으로 이어질 수 있는 학업 유예 및 학업 성취도 저하와 같은 문제를 해결하는 것을 목표로 한다(Bae 등, 2022). 또한, 향후 취업과 관련된 스트레스와 같은 요인이 한국 대학생들의 우울증을 유발하는 요인으로 밝혀져 정신 건강 개입의 중요성이 강조된다(Lim, 2019). 또한, 영어 학습의 어려움과 사회적 압박과 같은 동기 저하 요인도 한국 대학생의 학업 성취도와 유지율에 영향을 미친다(Mueller 등, 2021). 대학은 이러한 문제를 이해하고 목표에 맞는 개입과 지원 시스템을 구현함으로써 학생들이 장애물을 극복하고 학업 성취도를 향상시키며 중도탈락의 위험을 줄일 수 있다.

중도탈락은 학생과 대학 및 사회에 매우 부정적인 영향을 미친다(King-Dominguez 등, 2023; Silva & Diaz, 2023). 학생 개인은 학위가 없으면 소득 잠재력이 감소하고 취업 기회가 제한되며, 나중에 다른 대학에 재입학할 경우 추가 비용과 시간이 발생한다(Kang 등, 2019). 대학은 기회비용, 재정적 부담, 학생 선발과 관련된 비용 증가로 인해 재정 관리 및 정책 개발에 어려움이 있다(Nurmalitasari 등, 2023). 학생 중도탈락으로 인한 재정적 영향은 상당하며, 특히 사립대학은 1학년 중도탈락과 관련된 비용으로 인해 더 큰 영향을 받는다(Peña-Vázquez 등, 2023). 중도탈락의 다각적인 이유를 이해하는 것은 교육기관이 이 문제를 효과적으로 해결하고 학생, 대학, 사회에 미치는 악영향을 완화하는 데 매우 중요하다.

대학 기관 차원에서 중도탈락률을 관리할 때는 개선을 위한 교육 지표가 될 수 있는 거시적 변수에 초점을 맞추는 것이 중요하다(Guarda 등, 2023; Lee & Lee, 2017). 선행연구에 따르면 학생 만족도, 학업 성취도, 개인의 경제적 상황, 교수자와 학생 간의 관계, 프로그램 질과 같은 요인이 학생 중도탈락에 영향을 미친다(Llauró 등, 2023). 정성적 접근 방식과 정량적 접근 방식을 혼합하여 이러한 변수를 분석함으로써 중도탈락의 위험이 있는 학생을 조기에 파악하고 이들을 지원하고 동기를 부여하는 사전 조치를 취하여 궁극적으로 중도탈락 가능성을 줄일 수 있다(Aguirre & Carretero, 2020). 사회경제적 배경, 특별한 입학 요건, 제한된 수업 자료의 접근성, 낮은 학습 활동 참여도를 가진 학생은 학업적 어려움에 직면하고 성적이 저조할 가능성이 높다(El Ansari 등, 2013). 교육 환경에서 학생에게 효과적인 지원책을 제공하고 학생의 성공을 뒷받침하기 위해서는 이러한 다각적인 요인을 이해하고 해결하는 것이 필수적이다.

인공 지능의 하위 집합인 머신러닝은 알고리즘을 사용하여 방대한 양의 데이터를 분석하고 패턴을 식별하며 예측을 수행한다(Diniz, 2023). 머신러닝은 지도 학습과 비지도 학습으로 분류되며, 지도 학습은 알려진 데이터를 기반으로 분류와 예측에 중점을 둔다(Shaveta, 2023). 머신러닝의 일반적인 알고리즘으로는 결과 값 추정을 위한 선형 회귀, 의사 결정 나무, 신경망, SVM, 랜덤 포레스트 등이 있다(Rahmaty, 2023). 머신러닝의 주요 목표는 분류, 예측, 클러스터링 알고리즘을 통해 예측 모델을 생성하는 것으로, 새로운 데이터를 처리하고 결과를 예측하는 데 유용하다(Barham, 2017). 머신러닝은 선거 예측, 사회보장, 판매, 주가 예측 등 다양한 분야에 적용되고 있으며, 교육 분야에서는 학생 성적 분석부터 교육 데이터를 기반으로 한 국가 차원의 정책 결정에 이르기까지 다양한 업무에 활용되고 있다(Vyawahare, 2022). 빅 데이터를 통한 머신러닝 기법을 활용하여 학생의 경험, 목표, 과제에 대한 통찰력을 얻을 수 있으며, 정보에 기반을 둔 의사결정을 통해서 중도탈락 문제를 효과적으로 해결할 수 있다.

중도탈락 예방에 대한 연구는 중도탈락의 원인과 요인을 분석하는 데 중점을 두었지만, 최근의 선행연구는 중도탈락을 예측하고 예방 시스템을 구축하기 위한 예측 모델을 개발하는 방향으로 변화하고 있다(Krüger 등, 2023; Lee & Kang, 2019). 이러한 예측 모델링 및 예방 시스템 구축으로의 전환을 통해 중도탈락 위험의 학생을 조기에 파악하고 맞춤형 프로그램을 시행하는 연구들이 활발하게 진행되고 있다(Alladatin 등, 2023; Lee 등, 2023). 그리하여 본 연구는 머신러닝 알고리즘에 중도탈

락 위험군의 자료를 투입하여 예측모형을 개발한 후 예측 유효성을 평가하고 정보획득 지수를 통해 중도탈락에 영향을 미치는 요인을 파악하고자 한다.

2. 연구의 목적

본 연구는 중도탈락에 영향을 미치는 요인을 파악한 후 머신러닝 알고리즘을 활용하여 중도탈락 위험군 학생의 예측모형을 개발하고자 한다. 이 예측모형의 성능분석과 최적화를 수행하여 중도탈락 위험군(학사경고자, 예비학사경고자, 장기결석자)의 학사경고를 예방하고자 한다. 개발된 머신러닝 알고리즘 중에서 가장 높은 예측 유효성을 보인 머신러닝 알고리즘을 통해 중도탈락 위험군에 영향을 미치는 변인을 파악하고자 한다.

Ⅱ. 연구방법

1. 연구대상 및 데이터 정의

본 연구의 수집된 자료는 N대학교에서 운영하는 중도탈락 위험군의 자료를 바탕으로 머신러닝 알고리즘을 이용한 예측모형 개발에 투입하였다. 개발한 각 모형에 대해 예측유효성을 평가하고 머신러닝의 정보획득 지수를 통해서 파악한 중도탈락 위험군에 영향을 미치는 요인을 파악하였다. 통계적 가설을 확인하는 과정에서 데이터의 분포와 특이점에 대해서 이해하게 되고, 특정 모델에 적합하도록 추가적으로 데이터를 정제하거나 데이터의 분포에 적합한 모델을 선택하기 위해서 연구모형의 요인을 측정할 수 있는 9개의 데이터 속성자료를 추출하였다. 머신러닝 예측 모형 분석에서 예측변수는 학생이 속한 프로그램 대상년도, 학과, 학년, 성별, 연령, 입학년도, 거주지가 학교소재와 일치여부, 내외국인, 중도탈락의 유형 9개로 선정하였다. 유형완료는 중도탈락 유형으로서 가족관계부적응, 교육관계부적응, 전공부적응, 학교부적응, 학습부진 등으로 구분되어 있다. 프로그램 대상년도는 2018년 1학기부터 2022년 1학기까지 중도탈락 위험군에 포함된 학사경고, 예비학사경고, 장기결석자로 분석하였다.

본 연구에서 종속변수는 중도탈락 위험군에 의해 분류된 3개의 군을 이항변수로 변형하여 사용하였다. N대학교 중도탈락 위험군 대상자의 기준에 따라 학사경고자는 직전학기 평점평균 1.5미만, 예비학사경고자군은 직전학기 평점평균 1.5이상~1.75미만, 장기결석자군은 개강일 1~4주차 동안 3주 이상 결석에 해당되는 3개 집단으로 구분되었다. 2018년 1학기부터 2022년 1학기까지 중도탈락 위험군을 분석한 결과, 학사경고군 201명, 예비학사경고군 174명, 장기결석군 2,369명이었다. 이 중

Table 1. Data properties for factor measurement in research models

Dependent variable	Outcome	Academic Warning / Long-term absences
Predictor variable	Programme Year	Year 2018~2021 Semester 1 & 2 : Training data, Year 2022년 Semester 1 : Test data
	Departments	Categorical data
	Grade	Freshman / Sophomore / Junior / Senior
	Sex	Male / Female
	Age	Continuous data
	Entry year	Continuous data
	Residence matches the school's location	1: K city / 2: Other regions
	Nationality	Domestic / Foreigner
	Dropout type	Family relationship maladjustment, Education maladjustment, Major maladjustment, School maladjustment, Learning difficulties (other)

학사경고군의 빈도가 상대적으로 너무 적어서 학사경고군과 예비학사경고군을 합쳐서 학사경고군, 장기결석군의 이항변수로 변환하여 머신러닝 분석을 실시하였다. 본 연구의 학습 데이터는 2018년 1학기부터 2021년 2학기까지의 중도탈락 위험군 데이터 2,457건을 대상으로 구성하였다. 시험 데이터는 2022년 1학기 287건을 대상으로 구성하였다(Table 1).

2. 자료 처리 및 분석

데이터 마이닝과 예측모형 개발에 사용할 분석 프로그램은 R version 4.3.0을 사용하였다. 구성된 데이터 셋을 기준으로 중도탈락 위험군 예방 예측모델을 개발하고자 하였다. 본 연구의 머신러닝 도구로 오픈 소프트웨어인 Orange version 3.26.0을 사용하여 중도탈락 위험군을 예측하였다(Popchev & Orozova, 2023). Orange는 데이터 시각화, 머신러닝, 데이터 마이닝 기능을 포괄하는 오픈 소스 Python 기반 툴킷으로, 탐색적 데이터 분석과 대화형 프로그래밍 프런트 엔드를 제공한다. Orange는 위젯 연결을 통한 워크플로우 생성, 데이터 처리, 시각화, 예측 모델 구축 등 다양한 기능을 제공한다(Ishak 등, 2020). ROC(receiver operating characteristic) 곡선은 진양성율(true positive rate)과 위양성율(false positive rate)의 관계를 그래프로 나타내어, 모델의 예측 성능을 시각적으로 평가하였다. 본 연구에서는 각 머신러닝 알고리즘

Table 2. General characteristics of study subjects (n= 2,744)

Characteristics	Categories	n (%)
Sex	Male	1,812 (66.03)
	Female	932 (33.97)
Grade	Freshman	961 (35.02)
	Sophomore	625 (22.78)
	Junior	603 (21.98)
	Senior	555 (20.22)
Academic disciplines	Humanities and social sciences	459 (16.73)
	Teacher series	112 (4.08)
	National sciences	730 (26.60)
	Health sciences	328 (11.95)
	Engineering	680 (24.78)
	Arts and physical education	435 (15.85)
Residence matches the school's location	K city	1,291 (47.05)
	Other regions	1,453 (52.95)
Nationality	Domestic	1,966 (71.65)
	Foreigner	778 (28.35)
Dropout type	Family relationship maladjustment	10 (.36)
	Friendship maladjustment	15 (.55)
	Other (military enlistment)	70 (2.55)
	Other (post-return maladjustment)	12 (.44)
	Other (living expenses)	193 (7.03)
	Other (foreigners)	777 (28.32)
	Other (illness. disability)	31 (1.13)
	Other (employment)	52 (1.89)
	Major inappropriateness	228 (8.30)
	Poor learning	857 (31.23)

의 성능을 비교하기 위해 ROC 곡선 아래 면적(AUC, area under the receiver operating characteristic curve)을 계산하였다. AUC 값이 1에 가까울수록 모델의 예측 성능이 우수함을 의미한다. 중도탈락 여부에 미치는 영향요인은 이항 로지스틱 회귀분석을 적용하였다.

3. 머신러닝 예측모형

머신러닝 예측 모형은 대학생의 중도탈락과 같은 실제적인 상황에서 발생하는 사건을 예측하는 메커니즘으로 예측하고자 하는 반응변수를 예측하는 특징들, 즉 설명변수를 입력값으로 투입한 뒤 주어진 과거 데이터를 기반으로 수식을 생성하여 새로운 미래 데이터에도 적용할 수 있도록 하는 모형이다. 머신러닝 알고리즘은 랜덤포레스트, 결정트리, 인공신경망, 로지스틱 회귀분석, SVM, k-NN, Naive Bayes 알고리즘을 통해 예측모델을 생성하여 7가지 알고리즘의 정확도와 예측률을 비교하였다. 개발한 중도탈락 위험군 예방예측모형의 성능평가는 분류정확도와 예측유효성에 의존한다. 예측유효성은 TP rate(진양성율), FP rate(위양성율), accuracy(정확도), recall(재현율), precision(정밀도), F-measure 지표, AUC 지표를 활용하여 알고리즘별 예측모형의 성능을 측정 및 분석하였다. 알고리즘 최적화는 랜덤 포레스트(random forest)를 활용하여 예측 모형을 구축하였다(Yang 등, 2023).

Ⅲ. 결 과

1. 연구 대상자의 일반적 특성

본 연구는 K광역시 소재 N대학교에서 2018학년도 1학기부터 2022학년도 1학기까지 학사경고자, 예비학사경고자, 장기결석자로 대학생 2,744명의 데이터를 추출하였다. 대상자의 계열의 특성을 살펴보면, 자연과학계열이 730명(26.6 %)이며, 공학계열 680명(24.8 %), 인문사회계열 459명(16.7 %), 예체능계열 435명(15.9 %) 순으로 높게 나타났다. 대상자의 사는 지역의 특성을 살펴보면, 1291명(47.1 %)이 K광역시에 거주하고 있으며, 나머지 학교소재와 일치하지 않는 인원이 1453명(52.9 %)이다. 내국인은 1,966명(71.7 %)이고, 외국인 778명(28.3 %)이다. 중도탈락 유형(유형완료)으로는 학습부진 857명(31.2 %), 외국인 777명(28.3 %), 전공부적응 228명(8.3 %), 생활고 193명(7 %) 순으로 나타났다(Table 2).

2. 중도탈락 위험 분류에 영향을 주는 변인

1) 이항 로지스틱 회귀분석

이항 로지스틱 회귀분석을 통하여 프로그램 대상년도, 계열, 학년, 성별, 연령, 거주지가 학교소재와 일치여부, 입학연도, 내외국인 변수가 중도탈락 여부에 미치는 요인을 검증하였다. Hosmer와 Lemeshow검정 결과 χ^2=11.52, p=.174이므로 대립가설을 기각하고, 추정된 로지스틱 회귀분석 모형이 잘 적합하다는 귀무가설을 채택하는 것으로 나타났으며, 전체 분류 정확도는 86.6 %이다. 회귀계수의 유의성 검증 결과, 프로그램 대상년도, 계열, 학년, 입학년도가 중도탈락 위험에 유의한 영향이 있었다. 프로그램 대상년도가 1년 증가할수록 중도탈락 위험은 1.891배 증가하는 것으로 나타났다. 인문사회계열이 예체능계열보다 중도탈락 위험은 3.312배 증가하고, 사범계열이 예체능계열보다 중도탈락 위험은 1.986배 증가하며, 자연과학계열이 예체능계열보다 중도탈락 위험은 3.138배 증가하는 것으로 나타났다. 학년이 1년 증가할수록 중도탈락 위험에 있을 가능성은 47.8 % 감소하고, 입학년도가 1년 증가할수록 중도탈락 위험에 있을 가능성은 36.5 % 감소한다고 나타났다(Table 3).

2) 랜덤 포레스트 모델을 활용한 정보획득량 분석

머신러닝을 활용하여 대학생의 중도탈락 위험 비율이 높은 변수의 특징을 식별하기 위해 정보획득량(information gain) 분석을 실시하였고, rank는 기계학습에 반영할 핵심 속성을 추출하기 위해 점수를 산출하는 방법으로 정보획득량과 reliefF를 이용할 수 있다. 정보획득량은 어떤 속성을 선택하느냐에 따라 데이터가 더 잘 구분되는 정도를 나타낸다. 정보획득량이 클수록 분류가 더 잘 이루어진다. 본 연구에서는 학과 > 프로그램대상년도 > 학년 > 유형완료 > 연령 > 내외국인 > 거주지가 학교소재와 일치여부 > 입학년도 > 성별 순으로 중요한

Table 3. Results of binary logistic regression (N= 2,744)

Characteristics	B	SE	Wald	p	OR	95 % CI	
						LLCI	ULCI
Programme Year	.64	.07	81.31	.000	1.891	1.65	2.17
Academic disciplines							
Humanities & social sciences	1.20	.27	19.22	.000	3.312	1.94	5.66
Teacher series	.69	.19	13.26	.000	1.986	1.37	2.87
National sciences	1.14	.21	28.74	.000	3.138	2.07	4.77
Health sciences	.07	.21	.11	.737	1.072	.72	1.60
Engineering	-.88	.28	9.72	.002	.413	.24	.72
Grade	-.65	.09	50.58	.000	.522	.44	.63
Gender (male)	.11	.14	.60	.438	1.112	.85	1.45
Age	-.00	.01	.12	.731	.997	.98	1.02
Residence matches the school's location (match)	.11	.12	.78	.378	1.111	.88	1.40
Entry year	-.46	.05	72.82	.000	.635	.57	.71
Nationality (domestic)	-.05	.17	.10	.753	.949	.69	1.31
Constants	-370.69	92.57	16.04	.000	.000		

reference group*reference categories : academic disciplines*arts and physical education, gender*female, residence matches the school's location*not matched, nationality*foreigner

특성이다. 특히 학과가 유의한 분류를 할 수 있는 특성이다(Fig 1).

3. 예측모형의 성능분석

본 연구는 머신러닝을 활용하여 대학생 중도탈락 위험 비율이 높은 변수를 예측하기 위해 머신러닝의 7개 알고리즘을 통해 예측모형을 개발하였다. 시험 데이터를 통해 이 예측모형의 성능을 분석한 결과, random forest 알고리즘이 0.963의 분류정확도(CA)를 보여 가장 높게 나타났다(Fig 2)(Table 4).

이 예측모형이 시험 데이터를 얼마나 유효성 있게 결과를 예측하는지는 예측 측정 지표인 ROC curve 아래 면적(area under the ROC curve, AUC, 예측유효성)을 활용하였다(Pak & Oh, 2016). 예측성능을 비교한 결과, random forest, neural network, SVM 순으로 예측유효성이 높음을 알 수 있다(Fig 3). 본 연구의 AUC=0.996은 높은 정확도 수준이다.

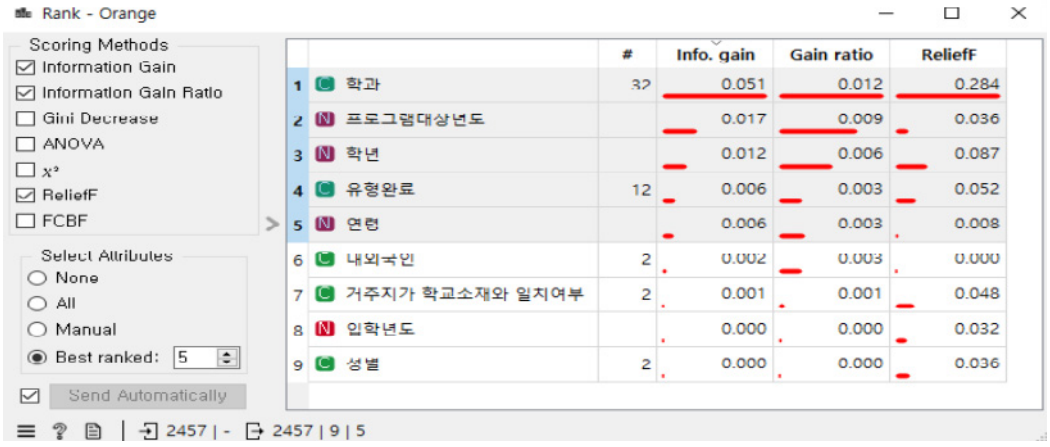

Fig 1. Rank

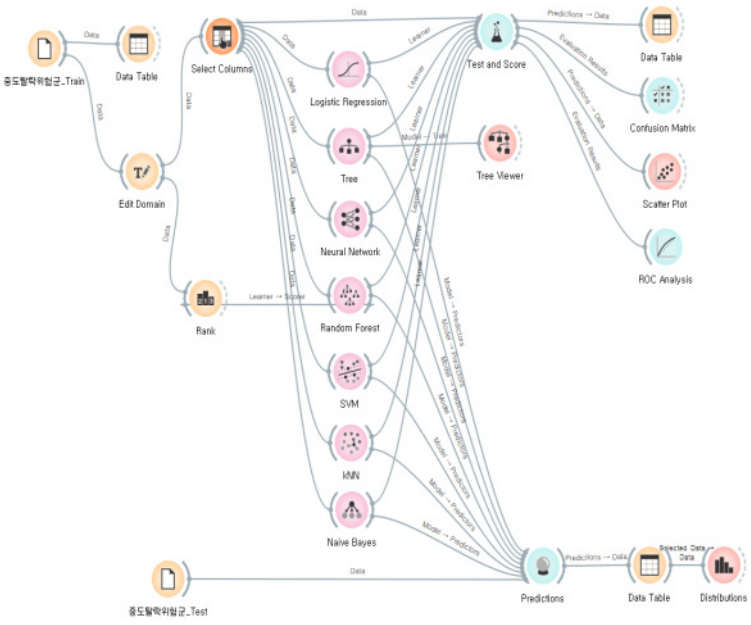

Fig 2. Orange data mining workflow

Table 4. Comparison of predictive model performance by machine learning algorithm on training data

Algorithm	AUC	CA	F1	Precision	Recall
k-NN	.684	.858	.815	.816	.858
Tree	.576	.830	.816	.807	.830
SVM	.763	.818	.828	.843	.818
Random Forest	.996	.963	.961	.965	.963
Neural Network	.928	.954	.952	.954	.954
Naive Bayes	.699	.142	.035	.020	.142
Logistic Regression	.657	.854	.793	.768	.854

4. 예측모형 알고리즘 최적화

가장 우수한 성능을 보인 랜덤 포레스트 모델에서 Explain 위젯을 이용하여 산출한 예측 변수의 상대적 중요도를 제시하였다(Fig 4). 상대적 중요도는 제공된 데이터를 사용하여 기능 값을 치환한 후 모델의 예측 오차 증가를 측정하여 예측에 대한 각 기능의 기여도를 계산한다. 예측 변수의 상대적 중요도는 학과(0.154), 연령(0.042), 학년(0.037), 거주지가 학교소재와 일치여부(0.023), 유형완료(0.011), 내외국인(0.001)의 순서로 높게 나타났다(Fig 5).

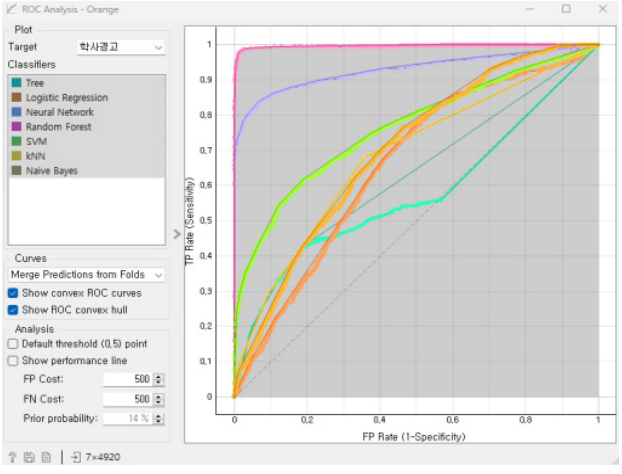

Fig 3. ROC curve

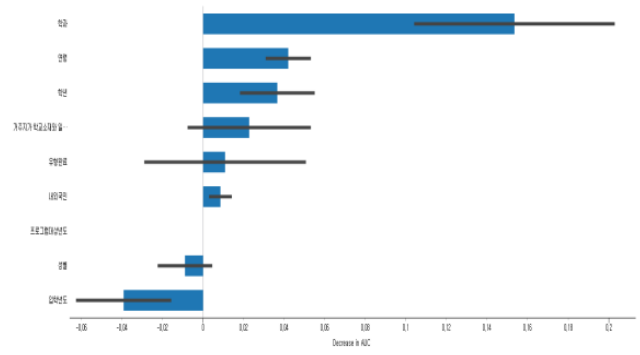

Fig 4. Graph of the relative importance of predictor variables in a random forest model

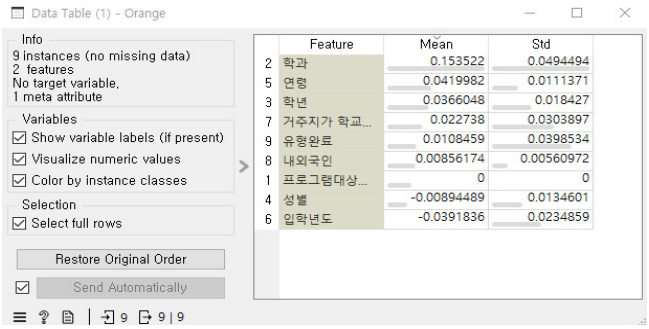

Fig 5. Relative importance of predictor variables in a random forest model

Ⅳ. 고 찰

본 연구는 중도탈락 위험군 예방을 위한 주요 요인을 도출하고, 중도탈락을 예방하기 위한 예측모형을 구축하고자 한다. 이를 위해 K광역시 N대학교의 2018~2022년 재학생을 대상으로 관찰한 데이터를 활용하여 머신러닝 알고리즘의 성능을 비교 분석하는 것이다.

대학생의 중도탈락은 개인적, 학업적, 경제적, 제도적 측면과 같은 요인이 중요한 역할을 한다(Santos-Villalba 등, 2023). 선행연구에 따르면 심리적 요인, 사회적 적응, 진로 결정 미숙, 학습 동기, 학교 지원 서비스와 같은 요인이 대학 중도탈락에 중요한 영향을 미친다고 보고하였다(Lee 등, 2023). 또한 중도탈락에 영향을 미치는 요인으로는 개인의 특성, 가정환경, 교수와의 관계 및 인프라 같은 학교 관련 측면이 있다(Alladatin 등, 2023). 이러한 다각적인 요인을 이해하는 것은 대학생의 중도탈락을 효과적으로 예방하기 위한 프로그램을 개발하는 데 필수적이다.

본 연구에서는 이항 로지스틱 회귀분석을 실시하여 프로그램 대상년도, 계열, 학년, 성별, 연령, 거주지가 학교소재와 일치여부, 입학연도, 내외국인 변수가 중도탈락 여부에 미치는 요인을 검증하였다. 회귀계수의 유의성 검증 결과, 프로그램 대상년도, 계열, 학년, 입학년도가 중도탈락 위험에 유의한 영향을 미치는 것으로 나타났다. Kemper 등(2020)은 logistic regressions와 decision trees의 중도탈락 예측률을 비교한 결과에서 decision trees가 보다 높은 예측률을 보였다. 특히 3학기 이상의 학생들을 대상으로 예측하였을 때 두 방법 모두 95 % 이상의 예측의 정확성을 보였다. 머신러닝을 적용한 중도탈락 위험군 예측에 이용된 선행연구에서의 머신러닝 알고리즘은 artificial neural network, Naive Bayes, decision tree, support vector machine, random forest, k-NN이었다(Shynarbek 등, 2022). Moreira da Silva 등(2022)은 중도탈락 위험군 학생을 예측하기 위해 random forest 앙상블 기법 중 Gradient Boosting의 XGBoost에 의해 분석하였다. 학생들의 중도탈락은 생활 조건과 관련되어 있으며, 나이로 추정된 학생의 성숙도(maturity)가 도출되었다. Rodríguez-Muñiz 등(2019)은 머신러닝 기반으로 분석을 수행하여 다른 연구들과 유사한 결론인 개인 및 교육맥락 변인들, 첫해의 학업성과가 영향을 미치고 있으며, 나이가 예측에 기여하고 있음을 확인하였다. 다른 선행연구에서는 수학 점수가 높은 학생이 첫 해에 대학을 그만둘 가능성이 낮은 이유는 첫 해에 언어능력을 많이 요구하지 않는 과목들이 많기 때문이다. 따라서 첫 해에 수학 과목에서 어려움을 겪는 학생들을 위한 지원 프로그램을 강화하는 것이 중요하다. 각 대학의 특성과 환경에 맞춘 개별화된 중도탈락 예측 모델을 개발하여 적용하는 것이 중요하다. 통합된 모델보다는 개별 모델이 더 정확한 예측을 제공하기 때문에 중도탈락을 효과적으로 예측하고 예방할 수 있다(Opazo 등, 2021).

본 연구에서는 학과, 프로그램 대상년도, 학년, 유형완료, 연령, 내외국인, 거주지가 학교소재와 일치여부, 입학년도, 성별, 총 9개 변수를 예측 변수로 하고, 중도탈락 위험에 대한 학사경고, 예비학사경고, 장기결석자군

의 이항변수를 종속변수로 하여 k-NN, decision tree, SVM, random forest, neural network, Naive Bayes, logistic regression에 적용하여 산출된 머신러닝 모델의 성능을 비교 평가하였다. 본 연구를 통해 분석한 결과, 중도탈락을 예측하는 중요도는 학과, 연령, 학년, 거주지가 학교 소재와 일치여부, 중도탈락 유형, 내외국인이 상위 6개 변수인 것으로 확인되었고, 그 중 랜덤포레스트의 성능이 가장 좋은 것으로 나타났다. 예측 변수의 상대적 중요도는 학과(0.154), 연령(0.042), 학년(0.037), 거주지가 학교소재와 일치여부(0.023)의 순서로 높게 나타났다. 중도탈락의 여러 변인들 중에서 '학과'의 상대적 중요도가 가장 높은 결과가 나타났다. 이는 중도탈락 위험이 높을 것으로 예측되었으나 실제로는 중도탈락 위험 비율이 낮은 학과의 경우에도 중도탈락 위험 예방에 관하여 재점검해 볼 수 있는 기회를 준다. 또한 중도탈락의 문제를 개선하기 위한 방안과 관련하여 정책 결정이나 지원책에 대한 필요성도 생각해 볼 수 있다.

Kabathova와 Drlik(2021)의 연구에서는 이러닝에서 중도탈락 예측에 머신러닝 알고리즘 (random forest, logistic regression, support vector machine, descision tree, Naive Bayes, neural network model 등)을 3년간 학생 데이터에 적용하였다. 모두 77 % 이상의 예측률을 보였으나 random forest 알고리즘의 예측률이 가장 높았고 Naive Bayes와 neural network model이 가장 낮았다. 다른 연구에서의 제안된 이중 레이어 앙상블 모델의 예측 정확도는 테스트 데이터 셋에서 92.18%로, 단일 모델에 비해 높은 성능을 보였고 대학 수업에서의 학생 중도탈락을 효과적으로 예측할 수 있다고 하였다(Niyogisubizo 등, 2022). 대학은 제안된 모델을 지속적으로 개선하고 업데이트하여 예측 성능을 높이는 것이 중요하다고 하였다. 선행연구에서는 중도탈락의 예측률을 높이기 위해서는 충분한 데이터가 필요함을 강조하였다. 향후 다른 예측 연구에서도 많은 대상자를 통한 장기간의 시계열 데이터를 통한 체계적인 분석이 필요하다고 생각된다.

대학 정책에 활용성 높은 연구 결과를 기술하기 위해서는 기본적으로 예측률이 높은 머신러닝 알고리즘 기반으로 도출된 중도탈락을 감소하는데 기여하는 변인을 나열할 필요가 있다. 분류기법에 해당되는 분석들을 통해 학과별로 구체적인 접근 시나리오를 제공하는 것도 도움이 될 것이다. 예를 들어, 1학년 학생 중 대학적응도 및 성숙도 등과 숙소와 학교 간 이동거리가 높은 학생들이 특히 높은 중도탈락률을 보인다는 내용이 학과별로 의사결정에 도움이 될 것이다.

본 연구의 제한점은 단일 대학의 데이터를 사용하였기 때문에 결과를 일반화하는 데 한계가 있다. 본 연구에서는 중도탈락의 요인을 학생의 개인의 외적인 요인으로만 국한되어 예측 변수만을 사용하였기 때문에 다양한 변수들을 추가적으로 고려할 필요가 있다. 향후 연구에서는 더 많은 대학의 데이터를 수집하여 분석하고, 보다 다양한 예측 변수를 포함한 연구를 통해 예측모형의 정확도를 높일 필요가 있다.

본 연구를 통해 다양한 머신러닝 알고리즘을 활용한 결과를 바탕으로 예비학사경고자, 학사경고자 등을 지도 및 상담할 때 기본 대상자 선정 및 전략 수립의 근거로 활용할 수 있다. 또한, 학생 개별의 다양한 배경과 학업 상황을 고려하여 이들의 필요에 맞는 지원을 계획하는 데 도움을 줄 수 있다. 이러한 정보를 바탕으로 향후 대학은 중도탈락률을 줄이고, 학생들의 학업 성취도를 향상시키기 위한 전략을 수립할 수 있다. 그리고 본 연구는 향후 교육 정책 입안자와 대학 관리자들에게 중도탈락 예방과 학업 지원에 관한 의사결정 과정에서 실질적인 가이드라인을 제공할 것으로 기대된다.

Ⅴ. 결 론

대학생의 중도탈락 위험군 관련 데이터를 활용하여 머신러닝 알고리즘인 K-최근접 이웃, 로지스틱 회귀, 서포트 벡터 머신, 의사결정트리, 랜덤 포레스트에 적용하여 산출된 머신러닝 모델의 성능을 비교 평가하였다. 그 결과 예측 변수의 상대적 중요도는 학과, 프로그램 대상자의 순서로 높게 나타났다. 이는 학과별 특성과 프로그램 진행 과정이 중도탈락에 큰 영향을 미친다는 것을 의미한다. 각 머신러닝 알고리즘의 성능 결과, 랜덤 포레스트의 성능이 가장 좋은 것으로 나타났다. 본 연구의 예측모형은 데이터 기반의 의사결정을 가능하게 하여, 보다 효과적이고 체계적인 중도탈락 예방 전략을 수립하

는 데 기여한다.

머신러닝을 활용한 중도탈락 위험군 예방 예측은 학업중단 위기학생을 조기에 발굴하는데 초점을 맞추고 있다. 중도탈락의 조기 예방 및 맞춤형 개입 프로그램 개발에 있어 실질적인 방향을 제시한다. 이러한 접근은 학생들의 학업 중단을 방지하고, 궁극적으로 학업 성취도를 향상시키는 데 기여할 것이다. 학업중단 위기의 유형과 원인은 학생에 따라 매우 다양하게 나타날 수 있다. 대학생의 학년별, 연령별 특성을 고려한 지원 정책을 수립하여 학생들의 학업 성취도를 높일 수 있다. 중도탈락 위험군 예방예측 모델 개발을 통하여 학사경고를 예방하고, 이에 맞는 학생 맞춤형 프로그램을 개발하는데 중요한 기초 자료가 될 것으로 생각된다. 향후 연구에서는 더 많은 대학의 데이터를 수집하여 분석하고, 보다 다양한 예측 변수를 포함한 연구를 통해 예측모형의 정확도를 높일 필요가 있다. 본 연구를 통하여 향후 중도탈락 예방에 대한 방향을 적절히 안내해 주고, 중도탈락 위험군 영역에 대한 연구가 활발하게 이뤄질 것이다.

참고문헌

Aguirre CE, Carretero J(2020). Predictive data analysis techniques applied to dropping out of university studies. 2020 XLVI Latin American Computing Conference (CLEI), 512-521. https://doi.org/10.1109/CLEI52000.2020.00066.

Alladatin J, Gnanguenon MA, Goza A, et al(2023). Research on school attendance and dropout: synthesis of the scientific literature. J Soc Sciences, 6(2), 89-98. https://doi.org/10.52326/jss.utm.2023.6(2).08.

Bae SH, Hwang SJ, Byun BK(2022). Pattern of out-of-class activities of Korean university students: latent profile analysis. Int J Res Ext Educ, 10(1), 59-74.

Barham H(2017). Achieving competitive advantage through big data: a literature review. 2017 Portland International Conference on Management of Engineering and Technology (PICMET), 1-7. https://doi.org/10.23919/PICMET.2017.8125459.

Diniz PS(2023). Signal processing and machine learning theory. 1st ed, Massachusetts, Academic Press, pp.869-959.

El Ansari W, Sebena R, Stock C(2013). Socio-demographic correlates of six indicators of alcohol consumption: survey findings of students across seven universities in England, Wales and Northern Ireland. Arch Public Health, 71(29), Printed Online. https://doi.org/10.1186/2049-3258-71-29.

Guarda T, Barrionuevo O, Victor JA(2023). Higher education students dropout prediction. In developments and advances in defense and security: proceedings of MICRADS 2022. Singapore, Springer Nature Singapore, pp.121-128.

Ishak A, Siregar K, Ginting R, et al(2020). Orange software usage in data mining classification method on the dataset lenses. IOP Conference Series: Materials Science and Engineering, 1003, Printed Online. https://doi.org/10.1088/1757-899X/1003/1/012113.

Kabathova J, Drlik M(2021). Towards predicting student's dropout in university courses using different machine learning techniques. Appl Sci, 11(7), Printed Online. https://doi.org/10.3390/app11073130.

Kang MH, Lee EK, Lee ET(2019). Trends and influencing factors of college students' dropout intention. Forum For Youth Culture, 58, 5-30. https://doi.org/10.17854/ffyc.2019.04.58.5.

Kemper L, Vorhoff G, Wigger BU(2020). Predicting student dropout: a machine learning approach. European J High Educ, 10(1), 28-47. https://doi.org/10.1080/21568235.2020.1718520.

King-Dominguez AA, Amestica-Rivas L, Gonzalez VR, et al(2023). Student dropout, the economic cost for Chilean universities. Universidad Ciencia y Tecnología, 27(118), Printed Online. https://doi.org/10.47460/uct.v27i118.683.

Krüger JGC, de Souza Britto Jr A, Barddal JP(2023). An explainable machine learning approach for student

dropout prediction. Expert Syst Appl, 233(1), Printed Online. https://doi.org/10.1016/j.eswa.2023.120933.

Lee EH, Kang SH(2019). The research trends and implications of college dropouts in Korea. J Korean Assn Learn, 19(10), 169-199. http://doi.org/10.22251/jlcci.2019.19.10.169.

Lee K, Lee H(2017). Korean graduate students' perceptions of guidance and professional development. High Educ, 73, 725-740. https://doi.org/10.1007/s10734-016-9988-9.

Lee SH, Lee MJ, Baek ES(2023). Analysis of university dropout research trends using structural topic modeling. J Korean Assn Learn, 23(11), 293-308. https://doi.org/10.22251/jlcci.2023.23.11.293.

Lim IR(2019). A meta-analysis on the effects of support program for underachieving college students in Korea. Journal of Educational Innovation Research, 29(3), 77-95. https://doi.org/10.21024/pnuedi.29.3.201909.77.

Llauró A, Fonseca D, Amo-Filva D, et al(2022). Academic analytics applied in the study of the relationship between the initial profile of undergraduate students and early drop-out rates. defining the variables of a predictor instrument. In: International conference on technological ecosystems for enhancing multiculturality. Singapore, Springer Nature Singapore, pp.982-990.

Moreira da Silva DE, Solteiro Pires EJ, Reis A, et al(2022). Forecasting students dropout: a UTAD university study. Future Internet, 14(3), Printed Online. https://doi.org/10.3390/fi14030076.

Mueller AS, Abrutyn S, Pescosolido B, et al(2021). The social roots of suicide: theorizing how the external social world matters to suicide and suicide prevention. Front Psychol, 12, Printed Online. https://doi.org/10.3389/fpsyg.2021.621569.

Niyogisubizo J, Liao L, Nziyumva E, et al(2022). Predicting student's dropout in university classes using two-layer ensemble machine learning approach: a novel stacked generalization. Computers and Education: Artificial Intelligence, 3, 100066. https://doi.org/10.1016/j.caeai.2022.100066.

Nurmalitasari, Long ZA, Noor MFM(2023). Factors influencing dropout students in higher education. Educ Res Int, 2023, Printed Online. https://doi.org/10.1155/2023/7704142.

Opazo D, Moreno S, Álvarez-Miranda E, et al(2021). Pereira J. Analysis of first-year university student dropout through machine learning models: a comparison between Universities. Mathematics, 9(20), 2599. https://doi.org/10.3390/math9202599.

Pak SI, Oh TH(2016). Application of receiver operating characteristic (ROC) curve for evaluation of diagnostic test performance. J Vet Clin, 33(2), 97-101. https://doi.org/10.17555/jvc.2016.04.33.2.97.

Peña-Vázquez R, González Morales O, Álvarez-Pérez PR, et al(2023). Building the profile of students with the intention of dropping out of university studies. Revista Española de Pedagogía, 81(285), 291-316. https://doi.org/10.22550/rep81-2-2023-03.

Popchev I, Orozova D(2023). Algorithms for machine learning with orange system. Int J Online Biomed Eng, 19(4), 109-123. https://doi.org/10.3991/ijoe.v19i04.36897.

Rahmaty M(2023). Machine learning with big data to solve real-world problems. J Data Anal, 2(1), 9-16. https://doi.org/10.59615/jda.2.1.9.

Rodriguez-Muniz LJ, Bernardo AB, Esteban M, et al(2019). Dropout and transfer paths: what are the risky profiles when analyzing university persistence with machine learning techniques?. PloS One, 14(6), Printed Online. https://doi.org/10.1371/journal.pone.0218796.

Santos-Villalba MJ, Alcalá del Olmo Fernández MJ, Montenegro Rueda M, et al(2023). Incident factors in Andalusian university dropout: a qualitative approach from the perspective of higher education students. Front Educ, 7, Printed Online. https://doi.org/10.3389/feduc.2022.1083773.

Shaveta(2023). A review on machine learning. Int J Sci Res, 9(1), 281–285. https://doi.org/10.30574/ijsra.2023.9.1.0410.

Shin MC, Song MJ(2022). A study on the effects of a

supporting program for low achieving students at university-centered on learning-men in K university. Korean J General Educ, 16(4), 247-264.

Shynarbek N, Saparzhanov A, Sagyndyk N, et al(2022). Forecasting dropout in university based on students' background profile data through automated machine learning approach. 2022 International Conference on Smart Information Systems and Technologies (SIST), 1-5.

Silva CAG, Diaz JP(2023). Dropout among students in higher education: a case study. Universidad Ciencia y Tecnología, 27(119), 18-28.

Vyawahare HR(2022). Machine learning: a solution approach for complex problems. Indian Sci J Res Eng Manag, 6(4), Printed Online. https://doi.org/10.55041/ijsrem16123.

Yang P, Li C, Qiu Y, et al(2023). Metaheuristic optimization of random forest for predicting punch shear strength of FRP-reinforced concrete beams. Materials, 16(11), Printed Online. https://doi.org/10.3390/ma16114034.