KICSP
한국 융합신호처리학회
융합신호처리학회논문지 Vol. 25, No.2 : 119~126 June 2024
DOI : 10.23087/jkicsp.2024.25.2.009

머신러닝 기반 대학생 중도탈락 예측 모델 구현 방안

노윤정
동명대학교 소프트웨어학과

Implementing of a Machine Learning-based College Dropout Prediction Model

Yoon-Jung Roh
Software Department, Tong Myeong University

요 약 본 연구는 대학생의 중도탈락에 영향을 주는 주요 패턴을 기계 학습하여 대학 중도탈락에 대한 조기 경보 시스템의 타당성을 평가하고 적극적으로 예방할 수 있는 시스템의 구현 방안을 제시하고자 한다. 이를 위해 한국교육개발원에서 실시한 한국교육종단연구 2005(Korean Educational Longitudinal Study, 2005)의 데이터를 사용하여 기계학습 기반의 5종의 알고리즘을 이용하여 성능 비교 실험을 실시하였다. 실험결과, 중도탈락 의도를 가진 학생의 식별 정확률(precosion)은 랜덤 포레스트(Random Forest)를 사용할 때 최대 94.0%, 중도탈락 의도를 가진 학생의 재현율(recall)은 Logistic Regression를 사용할 때 최대 77.0%로 측정되었다. 마지막으로 가장 높은 예측 모델을 바탕으로 중도탈락 가능성이 높은 학생을 상담 관리하며 특히, 특성별로 높은 중요도를 보이는 요인을 상담법 모델에 적용하고자 한다. 본 연구는 중도탈락이 대학과 개인에게 있어 큰 비용을 초래함과 대학생들이 직면한 진로 문제를 해결하기 위해 IT 기술을 활용한 모델을 구현하고자 한다.

• 주제어 : 중도 탈락, 중도탈락 특성요인, 기계학습, 랜덤포레스트, 기술융합

Abstract This study aims to evaluate the feasibility of an early warning system for college dropout by machine learning the main patterns that affect college student dropout and to suggest ways to implement a system that can actively prevent it. For this purpose, a performance comparison experiment was conducted using five types of machine learning-based algorithms using data from the Korean Educational Longitudinal Study, 2005, conducted by the Korea Educational Development Institute. As a result of the experiment, the identification accuracy rate of students with the intention to drop out was up to 94.0% when using Random Forest, and the recall rate of students with the intention of dropping out was up to 77.0% when using Logistic Regression. It was measured. Lastly, based on the highest prediction model, we will provide counseling and management to students who are likely to drop out, and in particular, we will apply factors showing high importance by characteristic to the counseling method model. This study seeks to implement a model using IT technology to solve the career problems faced by college students, as dropout causes great costs to universities and individuals.

• Key Words : Dropout, Characteristic factors for dropping out, Machine learning, Random Forest, Technology Convergence

Received 16 April 2024, Revised 25 June 2024, Accepted 30 June 2024

* Corresponding Author Yoon-Jung Roh, Dept. of Software, Tongmyong University, 428 Sinseon-ro, Nam-gu, Busan , Korea.
E-mail: roh7299@naver.com

Ⅰ. 서론

현재 한국 대학 환경의 가장 두드러진 문제 중 하나는 학령인구 감소로 인한 정원 감소에 관한 것이다. 대학정보 통계에 따르면 매년 약 90,000명의 학생이 대학을 떠나는 놀라운 수치를 보여주고 있다[1]. 이러한 대탈출은 개인에게 취업 및 사회통합 측면에서 기회비용을 부과할 뿐만 아니라 대학에 재정적 위협을 가하여 결과적으로 교육의 질을 저하시킨다[2]. 위험에 처한 학생 그룹을 조기에 발견하면 멘토링, 상담 또는 제도적 지원과 같은 즉각적 대책을 통해 중도탈락 의도를 가진 학생들에게 실질적 도움을 줄 수 있다.

이러한 중도탈락 현상을 이해하기 위한 수많은 실증 연구 수행되어 왔다. 대부분은 대학 중도탈락을 개인의 능력에 초점을 맞춘 개인적 특성, 대학 여건과 제도적 특성에 맞춘 대학 및 사회적 특성, 또는 진로 성숙과 관련된 심리적 요인에 초점을 맞춘 연구이다[1-6]. 더불어 4차 산업혁명시대에 들어서는 인공지능 기술을 접목한 빅데이터 기법을 활용한 대학생 중도탈락을 예측 논문이 증가하고 있다. 랜덤 포레스트(Random Forest), 의사결정 트리(decision tree), 로지스틱 회귀(Logistic Regression), KNN(K-Nearest Neighbors)과 같은 기계 학습 알고리즘을 사용하여 탈락에 기여하는 주요 요인을 찾아내고 있다[5-6].

이전 연구들이 주로 중도탈락 요인을 밝히거나 현상을 이해하는 것을 목표로 했다면, 본 연구에서는 머신러닝을 활용하여 주요 중도탈락 패턴을 파악하고 이를 사전에 예방함으로써 중도탈락을 적극적으로 예방할 수 있는 시스템 환경을 구축하는데 도움이 되고자 하는 것이다.

본 연구는 한국교육개발원에서 실시한 한국교육종단연구 2005(Korean Educational Longitudinal Study, 2005)의 데이터를 활용하여 머신러닝 기반 예측 모델에 대한 평가를 수행하여 더 복잡한 모델이 예측 품질을 향상시키는 정도를 평가하고자 한다. 또한, 가장 효과적인 예측 모델을 기반으로 개인적 특성, 대학 특성, 심리적 특성 중에서 중도탈락에 영향을 미치는 가장 중요한 요인을 분석하여 최적화된 상담 방법을 구축하는데 토석을 제시하고자 한다.

Ⅱ. 이론적 고찰

2.1 선행연구

대학생의 학업 지속성에 대한 이질성을 이해하고 대학생 중도탈락 발생을 예측하는 사회 및 교육 과학의 오랜 문헌에 기여한다. 이러한 선행 연구에서는 대학을 중도탈락 할 위험을 높이거나 줄이는 근본적인 요인과 결정 요인에 대해 연구한다. 일반적으로 문헌에서는 학생 중도탈락 결정에 있어 다양한 개인 및 기관 요인의 역할을 이해하는 것을 목표로 하였다.

여러 선행 연구들은 종종 확립된 이론적 틀에 기반한다. 예를 들어, Tinto(1975)는 학생 자신의 특성(학습 능력), 또래 및 사회적 환경에서 비롯된 맥락적 요인 및 제도적 배경 요인과 중도탈락 결정 사이의 인과 관계를 명시적으로 설명한다[7]. 이러한 모델의 인과적 주장과 질문(“대학생의 중도탈락 원인은 무엇입니까?”)에도 불구하고 문헌에는 학생 자기 선택 및 관찰할 수 없는 교란 변수와 같은 다른 요인과 실제 인과 관계 효과를 효과적으로 구별하기 위한 실증적 전략이 부족하다.

단일 결정 요인과 중도탈락 사이의 인과 관계를 확립하기 위해 실험적 방법을 적용한 주요 연구는 Stinebrickner(2014a), Adamopoulou 및 Tranzi(2017) 및 Horstaschraer 및 Sprietsma(2015)이다[8-10].

선행 연구의 두 번째 흐름은 그 출처를 이해하는 대신 대학생 중도탈락을 예측하는데 중점을 둔다. 학교의 관점에서 볼 때 자퇴 위험을 정확하게 평가하면 대학은 유의미한 개입을 통해 위험에 처한 학생들을 해결할 수 있다. 최근 몇 년 동안 이러한 연구는 강력한 머신러닝 방법론의 출현으로 더욱 활발히 진행되고 있다. 머신러닝 방법론의 주요 이점은 특정 교육 경로의 초기 단계에서 얻은 정보를 사용하여 정확한 예측을 제공하는 것이다. 또한 예측 분석을 위한 관리 데이터에 대한 향상된 엑세스로 인해 상당한 발전이 이루어졌다.

Sara et al.(2015)은 덴마크 학생들의 대규모 데이터를 사용하여 고등학교 중도탈락을 예측하였다. 고등학교 첫 6개월 동안 수집된 기본 인구 통계, 학교 유형, 가계 소득 및 학생들의 성과를 기반으로 학생들이 이후 3개월 동안 자퇴 여부를 예측하였다. 적용된 예측 방법(Support Vector Machine, Random Forest, CART 및 Navie Bayes) 중에서 Random Forest가 가장 정확한

예측을 제시했다[11].

Sansone(2018)은 미국의 고등학교 종적 데이터를 사용하여 중도탈락 위험이 있는 학생들을 식별하는 알고리즘을 개발하였다. 이 예측은 고등학교 1학년 때의 학생 정보를 기반으로 중등 교육 이후의 중도탈락을 예측하였다. 기존 방법과 비교하여 Support Vector Machine, Boosted Regression 및 Post-LASSO와 같은 기술이 중도 탈락을 예측하는데 더 정확한 것으로 나타났다[12].

Berenset al.(2018)은 독일의 사립 및 공립 대학의 학부생에 대한 행정 데이터를 사용하여 학생 중도탈락을 예측하였다. 본 연구의 접근 방식과 유사하게 의사결정 트리(Decision tree), 랜덤포레스트(RF)와 신경망 같은 머신러닝 기술과 이 세 가지 기술의 예측력을 결합한 앙상블 방법을 사용하였다. 가장 정확한 예측은 여러 단일 예측 모델의 결과를 종합하는 앙상블 방법이었다. 등록 시 인구 통계학적 결정 요인만 사용하는 이 알고리즘은 단순한 분류보다 훨씬 더 나은 예측 결과를 제공하였다. 하지만, 학생들의 주관적 성과에 대한 정보를 추가해도 예측 정확도가 더 이상 향상되지 않았다[13].

이러한 예시적인 연구는 머신러닝 기술이 기존 방법을 보완하는 것으로 보이며 중도탈락 위험에 처한 학생의 감지 및 예측을 용이하게 할 수 있음을 보여준다.

Ⅲ. 데이터 및 연구방법

3.1 데이터

본 연구는 한국교육개발원에서 실시한 한국교육조단연구 2005(Korean Educational Longitudinal Study, 2005)의 자료를 분석 자료로 이용하였다. 본 자료는 2005년부터 2014년까지 조사되었으며, 2005년 때 중학교 1학년이었던 학생 대상으로 설문조사를 시작하여 대학교 1학년에 해당하는 2011년 7차년도 자료를 사용하였다. 7차년도 데이터 중 4년제 이상 국내 대학에 재학 중인 2,348명을 최종 분석 대상으로 선정하였다.

본 연구에서 4년제 일반 대학교 1학년 학생들의 중도탈락 의도에 관하여 개인적 특성, 대학 기관의 특성, 진로성숙도 기반의 개인의 심리적 특성을 하위 집합으로 분류하였다.

첫 번째 정보 세트에는 성별, 학년, 고등학교 성적, 희망하는 대학 및 학과, 부모 소득 등 기본적인 사회경제적 배경 정보가 포함된다. 두 번째 정보 세트에는 대학의 유형과 대학 교육 및 대학 생활 만족도, 교수와 학생 선후배와의 관계와 같은 대학생 중도탈락 의도에 영향을 주는 대학 특성 요인이 포함된다. 세 번째 정보 세트에는 개인의 심리적 특성 요인으로 진로성숙도에 관한 주관적인 인지 테스트와 관련하여 포함되어 있다.

핵심 변수의 하위 집합에 대한 기본 설명 통계는 Table 1.에 요약되어 있다.

Table 1. Descriptive statistics by characteristics

Div.	Var.	N	Mean	Std.	Min	Max
1st year university		2,348	0.12	0.32	0	1
personal	gender	2,348	1.5	0.5	1	2
	high-sch. grades	2,345	6.47	1.42	1	9
	college grades	2,320	3.25	0.65	1	4.5
	college preference	2,348	1.94	0.76	1	4
	department preference	2,348	3.31	0.89	1	4
	parent encourage support	2,348	3.86	0.67	1	5
	parent money support	2,348	3.14	0.72	1	5
	family income	2,348	252	1.08	1	6
	worry tutition	2,348	2.42	0.87	1	4
	father education	2,348	3.24	1.78	1	8
	mother education	2,348	2.74	1.49	1	8
University	university location	2,348	1.68	0.47	1	2
	establishment type	2,348	1.74	0.44	1	2
	Uni. satisfaction	2,348	1.94	0.76	1	3
	Curriculum Satisfaction	2,348	3.31	0.90	1	4
	Professor contact	2,348	2.69	0.57	1	6
	seniors contact	2,348	3.42	0.97	1	6
	Edu. environment	2,348	2.89	0.68	1	6
	University lecture	2,348	3.25	0.93	1	6
	Uni_life satifaction	2,348	3.31	1.11	1	5
Psycho-logical	planning	2,348	4.20	0.55	1	6
	work attitude	2,348	4.48	0.68	1	6
	self-understanding	2,348	4.46	0.84	1	6
	independence	2,348	4.00	0.73	1	6
	career behavior	2,348	4.20	0.76	1	6

Table 1.를 살펴보면 개인적 특성 요인에서 선호하는 대학의 평균이 1.94로 낮은 수치를 보이고 있다. 반면 선호하는 학과의 평균은 3.31로 높은 수치를 보이고 있어 분석 대상자들이 재학하고 있는 대학에 대해서는 기대가 낮으며 학과에 대해서는 만족하고 있는 것으로 보인다. 또한, 대학 특성 요인에서 수업 외 교수와의 교류가 다른 변수들 보다 낮은 수치를 보이며 만족도가 낮게 나타나고 있다. 진로성숙도에 기반한 심리적 특성 요인에서는 5개 변인 모두 4.0 이상으로 평균 이상 높은 결과를 보여주고 있다.

3.2 데이터 전처리

KELS 2005 데이터 세트에는 원시 데이터가 포함되어 있으며 누락 된 값은 대체 프로세스를 거쳤다. 특히, 랜덤포레스트(RF), 로지스틱 회귀(LR), KNN과 같은 기술의 경우, 결측값이 평균과 중앙값을 모두 사용하여 대체된다. 반대로 나이브 베이즈(NB)와 LightGBM의 경우 누락 된 데이터는 처리되지 않은 상태로 누락 된 값을 그대로 두고 분석하였다. 구조화된 데이터 프레임을 생성하기 위해 Python 기반 Pandas 라이브러리를 사용하여 데이터 조작을 수행하였다.

3.3 연구 방법

본 연구의 목표는 중도 탈락 위험이 있는 그룹을 예측하도록 설계된 머신러닝 기반의 예측 알고리즘의 타당성과 예상 성능을 평가하는 것이다. 또한, 평가된 알고리즘 중 최고 예측 성능의 모델을 기반으로 대학 중도 탈락에 영향을 주는 주요 변인들의 중요도를 분석한다. 이를 위해, 구체적으로 사용 가능한 데이터가 정확한 예측에 적합한지 확인하고 예측 품질 향상을 위해 여러 선행 연구들을 바탕으로 한 머신러닝 분야에서 널리 활용되는 Random Forest(RF), Naive Bayes(NB), Logistic Regression(LR), K-Nearest Neighbors(KNN)과 LightGBM의 5종 알고리즘 모델과 방법론을 사용한다[8-12].

모델 간 예측 및 변수 중요도의 구조적 차이를 도출하기 위해 사용되는 정보와 적용된 방법이 다른 예측 모델을 분석하고 이를 통해 불균형을 일반화 할 수 있는 패턴을 찾아내고 이러한 결과를 기반으로 예측 모델을 구현하기 위한 지침을 세우는 것을 목표로 한다.

실험적 평가 방법은 무작위 샘플링을 10회 실시하여 평균값을 도출하였으며, 학습 세트 비율은 전체 데이터 세트의 95%로 구성했다. 기계학습에서 일반적으로 사용되는 성능 평가 지표인 정밀도와 재현율을 Table 2.와 같이 활용했으며, 다음 [식 1.]과 같다.

Table 2. Exemplary Confusion Matrix[1)]

Div.		Prediction	
		Dropout-intention	Non Dropout-intention
Truth	Dropout-intention	10 (true positives)	10 (false negatives)
	Non Dropout-intention	20 (false positives)	60 (true negatives)

$$\mathrm{Pr}ecision = \frac{TP}{TP+FP},\ Recall = \frac{TP}{TP+FN} \quad (1)$$

$$F1\,Score = 2 \times \frac{\mathrm{Pr}ecision \times Recall}{\mathrm{Pr}ecision + Recall}$$

Ⅳ. 연구결과

4.1 실험 및 성능 비교

우선 실험을 통해 본 연구의 제안 방법을 비교하였다. 전체 실험 결과는 Table 3.과 같다. F1 값은 정밀도(P)와 재현율(R) 값의 균형을 제공하는 척도이다. 최고값 1(완벽한)에서 최저값 0으로 도달한다.

정밀율(P)은 모델이 수행한 모든 긍정적인 예측 중에서 실제 긍정적인 예측의 비율을 측정한 값이다. 이는 양성 예측의 정확성을 나타내며 참양성(올바르게 예측된 양성값)과 참양성 및 거짓양성(양성의 잘못 예측된 양성값)의 비율을 의미한다.

재현율(R)은 데이터 세트의 모든 실제 양성 사례 중에서 참양성 예측의 비율을 측정한 값이다.

1) 기본 모집단 100명 중에서 20명의 중도탈락 의도를 가진 자와 80명의 재학 의지를 가진 학생이다. 예제 알고리즘은 10개의 탈락(참 양성)을 올바르게 식별하지만 10개의 탈락을 잘못 분류한다(거짓 음성). 반면에 80명 중 60명의 재학의지 학생은 올바르게 분류하지만(참 음성), 20명의 학생을 중도탈락 의도(거짓 양성)로 잘못 분류한다.

Table 3. Performancecomparison of overall experiment

Div.	KNN			NB			RF			Logit			LightGBM		
	F1	P	R	F1	P	R	F1	P	R	F1	P	R	F1	P	R
Overall students	0.87	0.93	0.83	0.83	0.91	0.77	0.89	0.94	0.86	0.90	0.87	0.94	0.72	0.85	0.63
Dropout - intention	0.61	0.83	0.48	0.59	0.56	0.62	0.85	0.94	0.73	0.68	0.61	0.77	0.60	0.62	0.58

전체 학생과 중도탈락 의도를 가진 학생을 분류기별로 비교한 결과 전체 학생 대상으로 편차는 있으나 전체적으로 모든 알고리즘에 대해 우수하게 나타났다. KNN, RF, LR의 F1 값이 80% 후반에서 90% 초반 대의 우수한 성능을 보이며, LightGBM이 가장 저조한 F1 값이 나타났다.

본 연구의 목적은 중도탈락 의도를 가진 학생 대상 중도탈락 패턴을 찾고 예방하기 위한 조치를 하기 위함으로 Table 3.에서 보여주는 것과 같이 5종의 성능 측정 결과를 비교하였다. 분석 결과, 중도탈락 의도를 가진 학생들의 성능은 전체 학생 그룹에 비해 낮게 나타나고 있다. 전체 학생 수 대비 중도탈락 의도를 가진 학생 수가 작아서 나타나는 현상이라 생각할 수 있다. 중도탈락 의도를 가진 학생들의 실험 결과, 랜덤포레스트(RF)가 94.0%로 가장 높은 정밀율을 보이고 있으며, 재현율(R) 역시 73.0%로 높은 수치를 보이고 있다. 따라서, 랜덤포레스트(RF)가 1에 가까운 0.85 F1 값으로 가장 우수한 모델 성능을 보이고 있다. 재현율(R)은 Logistic Regression(LR)이 77%로 가장 높은 수치로 측정되었다.

분석 결과를 요약하면, 랜덤포레스트(RF)가 전체 정보를 사용한 비교를 기반으로 가장 선호되는 모델로 나타났다. 이는 거짓 부정을 최소화하면서 실제 중도탈락 의도를 가진 학생을 정확하게 분류하는 정밀도가 높은 것으로 해석할 수 있다. 또한 Sara et al.(2015) 연구 결과와 일치한다[11]. 따라서 랜덤포레스트(RF)가 더 높은 수준의 해석 가능성을 유지하고 개별 예측 변수의 중요성에 대한 더 나은 통찰력을 제공한다는 점으로 고려하여 랜덤포레스트(RF)를 앞으로 선택한 모델로 정하였다.

4.2 중도탈락 특성별 중요 변수 예측

이전 장에서는 특정 예측 변수 세트를 사용할 때 랜덤포레스트(RF)가 일반적으로 다양한 결정 임계값에서 추가 예측 모델보다 성능이 뛰어난 것으로 관찰되었다. 데이터 세트 크기 변화와 관련된 이러한 결과의 견고성을 평가하기 위해 사용 가능한 변수를 특성에 따라 개인, 대학 및 심리적 특성을 기반으로 하위 집으로 분류하였다. 이러한 기본 특성의 제한으로 인해 예측 품질이 저하 될 수 있음에도 불구하고 기계학습 기반 예측은 OLS와 같은 간단한 방법에 비해 지속적으로 우수한 분류를 보여준다. 이는 상대적으로 작은 데이터 세트, 특히 대학에서 얻은 정보로 제한된 데이터 세트를 사용하는 경우에도 기계학습 기반의 랜덤포레스트(RF) 모델을 사용하면 대학 중도탈락 의도를 가진 학생을 식별해내는 예측 경보 시스템을 상대적으로 간단하고 비용 면에서도 효율적으로 제공될 수 있다[14].

4.2.1 개인 특성 분류 중요 변수

대학 중도탈락 요인 분석에서 가장 기본적인 개인적 특성 요인 중 하위 요인들의 중요도 분석 결과는 Figure 1.와 같다.

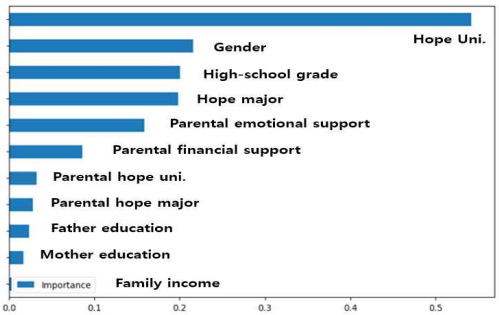

Fig. 1. Feature importance : Personal characteristic

선호하는 대학이 가장 중요한 예측 변수로 나타났다. 이는 개인이 대학 입학 전부터 희망하는 대학의 여부가 대학교 1학년 학생들에게 중도 탈락의 중요한 변수임을 설명하고 있다. 또한, 기존의 노윤정(2022) 자료의 연구 결과와도 일치한다[15]. 우리나라에 만연해

있는 대학 서열화를 비추어 볼 때 개인적 특성 중 희망하는 대학의 여부가 대학을 포기할지 재학할지에 관한 중요 예측 변수임을 다시 한번 보여주고 있다.

4.2.2 대학 특성 분류 중요 변수

대학 중도탈락 요인 분석에서 대학 설립유형 및 대학 교육과정 등 대학 생활에 관련한 대학 특성 요인 중 하위 요인들의 중요 요인 분석 결과는 Figure 2.와 같다.

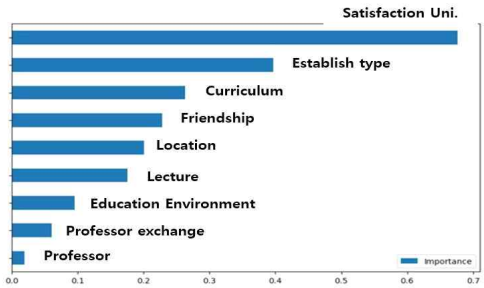

Fig. 2. Feature importance : University characteristic

대학 관련 요인 중 가장 영향력 있는 예측 변수로 재학하고 있는 대학에 대한 만족도로 나타났다. 이러한 결과는 대학에 대한 사회적 인식이 낮을수록 중도탈락 가능성이 높다는 선행연구와 일치한다(김수연, 2012; 한송이, 2019; 노윤정, 2022)[15-17]. 이전 연구에서 대학의 평판이 포함된 전반적인 대학에 대한 만족도가 중도 탈락 가능성의 중요한 결정 요인으로 일관되게 강조해 오고 있는 부분과 일치하고 있다. 하지만 선행 연구 결과와 다른 결과를 보이고 있는 변수가 교수와의 교류이다[15]. 노윤정(2022) 연구에서는 수업 외 교수와의 상호작용이 증가할수록 중도탈락 확률이 높아지는 것으로 중도탈락에서 중요 요인으로 보였다. 하지만 본 연구인 중요성 분석에서는 교수와의 상호작용이 두 번째로 낮은 예측 변수로 나타나 교수와의 상호작용이 중도탈락에 크게 영향을 미치지 않는 것으로 해석할 수 있다.

4.2.3 심리적 특성 분류 중요 변수

대학 중도탈락 요인 분석에서 진로성숙도를 나타내는 심리적 특성 요인 중 하위 요인들의 중요도 분석 결과는 Figure 3.와 같다.

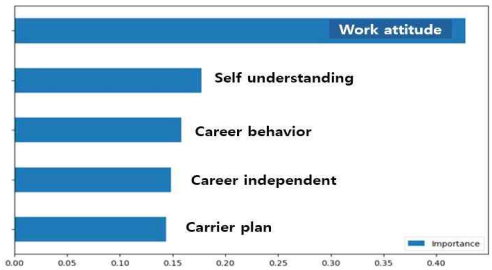

Fig. 3. Feature importance : Psychological characteristic

진로성숙도 기반의 심리적 특성 요인 중에서 일에 대한 태도가 중도탈락 의도를 가장 강력하게 예측하는 변수로 나타났다. 이러한 결과는 노윤정(2022)의 연구와 일관되게 보여진다[15]. 일과 진로에 대한 독립성에 관해 긍정적인 태도가 중도탈락 의도 감소와 관련이 있다는 선행연구 결과와 일치한다. 일에 대한 태도는 일의 중요성에 대한 인식 뿐만 아니라 자신의 직업적 진로에 대한 긍정적인 관점을 키우는 것도 포함한다. 이러한 결과는 대학 입학 전과 후의 진로와 직업 세계에 대한 적극적인 교육의 필요함을 시사해준다.

Ⅴ. 결론 및 시사점

대학생의 중도탈락은 개인과 대학에 큰 비용을 초래하며, 이를 예방하기 위한 선제적 연구가 필요하다. 또한, 인공지능 기술을 활용한 대학생 중도탈락 예측 연구가 증가하고 있으며, 머신러닝을 활용하여 주요 중도탈락 패턴을 파악하고 예방하는 시스템 환경을 구축하는 것이 무엇보다 필요한 상황이다.

이를 위해 본 연구는 한국교육개발원에서 실시한 한국교육종단연구(KELS) 2005의 데이터를 활용하여 머신러닝 기반 예측 모델에 대한 평가를 수행했으며, 대학 중도탈락의 중요 요인 분석에 관한 선행 연구를 기반하여 특성별 중요도 요인 예측을 실시하였다.

5종의 알고리즘 모델과 방법론을 사용하여 중도탈락 의도에 관한 예측 모델 성능을 평가한 결과, 랜덤포레스트(RF)가 94.0%로 가장 높은 정밀율(P)을 보이고 있으며, 재현율(R) 역시 73.0%로 높게 나타났다. 또한, 가장 높은 예측 성능의 모델인 랜덤포레스트(RF)로 대학생 중도탈락의 중요도 요인 분석을 실시한 결과 1) 개인적 특성 요인 중에서 학생이 선호하는 대학이 중도탈락 의도를 결정하는 가장 중요한 예측 변수로 나

타났다. 2) 대학 관련 요인 중 가장 영향력 있는 예측 변수로는 현재 재학하고 있는 대학의 만족도로 나타났다. 3) 진로성숙도 기반의 심리적 특성 요인 중에서는 일에 대한 태도가 중도탈락 의도에 중요한 변수로 나타났다.

이러한 결과는 대학생 중도탈락 예방을 위한 연구에 중요한 시사점을 제공한다.

우선 대학 중도탈락자를 조기에 예측하기 위해 가장 우수한 성능을 보였던 랜덤포레스트(RF)로 접근하는 것을 기본 모델로 구현할 수 있다. 또한, 대학 중도탈락은 대학의 외부적 특성 뿐만 아니라 개인이 가지는 주관적인 요인에 영향을 받는다는 선행 연구에 기반하면[18], 랜덤포레스트(RF)가 학생의 대학의 전반적인 생활에 대한 만족도나 심리적 특성 요인에 기반한 주관적인 데이터를 사용했을 경우 예측 품질을 높일 수 있다는 것도 확인했다.

두 번째로 대학 중도탈락에 대한 조기 예측 모델을 통해서 대학과 학생의 중도탈락에 대한 비용 효율성에 대해서도 고려할 수 있다. 중도탈락 의도를 가진 학생 그룹을 조기에 발견하면 대학에서 멘토링, 상담 또는 제도적 지원과 같은 목표 조치를 즉시 개입함으로써, 학생의 교육 경력 중단과 관련된 상당한 비용을 완화할 수 있으며, 대학 역시 중도탈락 의도를 가진 학생을 올바르게 예측하여 대학 재정을 중도탈락 학생에 대한 대책 예상 비용을 효율적으로 분배할 수 있다.

세 번째로 대학 중도탈락에 관한 개인, 대학, 심리적 특성 요인의 하위 변인 중에서 중요 변수를 고려하여 대학은 맞춤형 지원을 제공할 수 있다. 또한, 대학에서는 학생들의 만족도를 높이기 위해 교육 및 지원 시스템을 개선하고 학생들과의 소통을 강화하는 것이 무엇보다 중요함을 본 연구 결과로 다시 한번 확인할 수 있었다.

ACKNOWLEDGMENTS

이 논문은 2022학년도 동명대학교 교내학술연구비 지원에 의하여 연구되었음(2022B005)

REFERENCES

[1] Korea Educational Development Institute, Korean Education Statistics Yearbook, 2023

[2] D. H. Jeong & J. Y. Park, “Data Analysis of Dropouts of College Students Using Topic Modeling”, Journal of the Korea Institute of Information and Communication Engineering, vol.25, no.1, pp.88-95, 2021

[3] P. Perchinunno, M. Bilancia, & D. Vitale, “A Statistical Analysis of Factors Affecting Higher Education Dropouts”, Social Indicators Research, vol.156, pp.341-362, 2021

[4] M. Kang, E. Lee & E. Lee, “Trends and influencing factors of college student’s dropout intention “, In Forum for Youth Culture, vol.58, pp.5-30, 2019

[5] C. Park, “Development of Prediction Model to Improve Dropout of Cyber University “Journal of the Korea Academia-Industrial Cooperation Society, vol.21, no.7, pp.380-390, 2020

[6] E. J. Lee, Y. Song, J. H. Kim & S. H. Oh, “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, vol.36, no.1, pp.191-219, 2020

[7] Tinto, V, “Dropout from higher education: A theoretical synthesis of recent research”, Review of educational research, vol.45, no.1, pp.89-125, 1975

[8] Stinebrickner, Todd R. Stinebrickner, R, “A Major in Science? Initial Beliefs and Final Outcomes for College Major and Dropout,” The Review of Economic Studies, vol.8, no.1, pp.426-472, 2014

[9] Adamopoulou, E., Tanzi, G. M, “Academic Dropout and the Great Recession,“ Journal of Human Capital, vol.11, no.1, pp.35-71, 2017

[10] Horstschräer, J., Sprietsma, M, “The Effects of the Introduction of the Bachelor Degree on College Enrollment and Dropout Rates”, Education Economics, vol.23, no.3. 2015

[11] Sara, N. B., Halland, R., Igel, C., Alstrup, S, “ High-school dropout prediction using machine learning: A danish large-scale study”, In ESANN 2015 proceedings, European Symposium on Artificial Neural Networks, Computational Intelligence, pp.319-24, 2015

[12] Sansone, D, “Beyond Early Warning Indicators: High School Dropout and Machine Learning”, Oxford Bulletin of Economics and Statistics, 2018

[13] Berens, J., Schneider, K., Goertz, S., Oster, S., and Burghoff, J, "Early Detection of Students at Risk-Predicting Student Dropouts Using Administrative Student Data and Machine Learning Methods", CESifo Working Paper, No.7256, 2018
[14] I. E. Isphording, T. Raabe, "Early Identification of College Dropouts Using Machine-Learning: Conceptual Considerations and an Empirical Example", Institute of Labor Economics, no.89, 2019
[15] Y. J. Roh, "Empirical analysis of factors influencing failure to drop out; Focusing on career adaptability", International Next-generation Convergence technology Association, vol.6, no.5, pp.876-889, 2022
[16] S. Y. Kim, "Analysis of the movement path structure of college dropouts", Educational Science Research, vol.43, no.3, pp.131-163, 2012
[17] S. I. Han, "Analysis of Elements of Retention Corresponding to College Dropout Factors", Korean Association For Learner-Centered Curriculum And Instruction, vol.19, no.4, pp.1239-1257, 2019
[18] D. F. Williams, "The impact of career workshops on freshman college students at risk for dropout: An action research study", Journal of College Student Retention, vo.13, no.1, pp.37-62, 2011

저자소개

노 윤 정 : (Yoon-Jung Roh)

1994년 2월 : 부경대학교 미생물학과(학사)
2016년 8월 : 부산대학교 경영학과 대학원(석사)
2021년 8월 : 부경대학교 경제학과 대학원(박사)
2014년 3월~현재 : 동명대학교 소프트웨어학과 조교수

관심분야 : 공학교육, 머신러닝, 교육경제