Journal of The Korea Convergence Society
Vol. 12. No. 10, pp. 37-43, 2021
I-SSN 2233-4890 / e-ISSN 2713-6353
DOI : https://doi.org/10.15207/JKCS.2021.12.10.037

대학생 중도탈락 예방을 위한 기계 학습 기반 추천 시스템 구현 방안

정도헌
덕성여자대학교 글로벌융합대학 조교수

Implementation of a Machine Learning-based Recommender System for Preventing the University Students' Dropout

Do-Heon Jeong
Assistant Professor, College of Global Convergence Studies, Duksung Women's University

요 약 본 연구는 대학생의 중도탈락 패턴을 식별하는 효과적인 자동 분류 기법을 제안하고, 이를 바탕으로 중도탈락을 예방하기 위한 지능형 추천 시스템의 구현 방안을 제시하는 것을 목표로 한다. 이를 위해 1) 실제 대학생의 재학/제적 데이터를 기반으로 기계 학습의 성능을 향상시킬 수 있는 데이터 처리 방안을 제안하고, 2) 5종의 기계학습 알고리즘을 이용하여 성능 비교 실험을 실시하였다. 3) 실험 결과, 제안 기법이 베이스라인에 비해 모든 알고리즘에서 우수한 성능을 보여주었다. 제적생의 식별 정확률(precision)은 랜덤 포레스트(Random Forest)를 사용할 때 최대 95.6%, 제적생의 재현율(recall)은 나이브 베이즈(Naive Bayes)를 사용할 때 최대 80.0%로 측정되었다. 4) 마지막으로, 실험 결과를 바탕으로 중도탈락 가능성이 높은 학생을 우선 상담하는 추천 시스템의 활용 방안을 제시하였다. 교육 현안 문제를 해결하기 위해 IT 분야의 기술을 활용하는 융합 연구를 통해 합리적인 의사결정을 수행할 수 있음을 확인하였으며 향후 지속적인 연구를 통해 다양한 인공지능 기술을 적용하고자 한다.

주제어 : 기술융합, 중도탈락, 자동분류, 기계 학습, 추천 시스템

Abstract This study proposed an effective automatic classification technique to identify dropout patterns of university students, and based on this, an intelligent recommender system to prevent dropouts. To this end, 1) a data processing method to improve the performance of machine learning was proposed based on actual enrollment/dropout data of university students, and 2) performance comparison experiments were conducted using five types of machine learning algorithms. 3) As a result of the experiment, the proposed method showed superior performance in all algorithms compared to the baseline method. The precision rate of discrimination of enrolled students was measured to be up to 95.6% when using a Random Forest(RF), and the recall rate of dropout students was measured to be up to 80.0% when using Naive Bayes(NB). 4) Finally, based on the experimental results, a method for using a counseling recommender system to give priority to students who are likely to drop out was suggested. It was confirmed that reasonable decision-making can be conducted through convergence research that utilizes technologies in the IT field to solve the educational issues, and we plan to apply various artificial intelligence technologies through continuous research in the future.

Key Words : Technological convergence, Dropouts, Automatic classification, Machine learning, Recommender system

*This Research was supported by Duksung Women's University Research Grants 2020 (3000005346).
*Corresponding Author : Do-Heon Jeong(doheonjeong@duksung.ac.kr)
Received August 22, 2021 Revised September 27, 2021
Accepted October 20, 2021 Published October 28, 2021

1. 서론

최근 학령인구가 급격히 감소하면서 대학의 충원율 감소와 정원미달 현상이 발생하고 있으며, 대학생이 소속 대학에서 학업을 지속하지 않고 중단하는 중도탈락의 비율 역시 증가 추세를 보이고 있다[1]. 중도탈락은 사회 진출 및 취업 과정에서 개인의 기회비용을 발생시키며, 대학의 차원에서는 대학 재정의 안정성을 위협하고 교육의 질 저하를 야기하는 문제를 안고 있다[2].

대학생의 중도탈락 현상과 관련한 분석적 연구들을 살펴보면, 우선 중도탈락에 영향을 주는 주요 요인들을 분석하는 연구들이 다수 수행되고 있다. 이러한 연구들은 주로 개인의 가정, 심리적인 문제 등과 관련된 환경적 요인, 대학생활, 대학 교육, 연구 활동 및 교육여건 등과 관련된 기관적 요인 등 원인 항목을 세분화하고 요인별 영향력을 분석한다[1-6].

또한, 최근 학문과 기술의 융합이 활발해 지면서, 기계학습 기법 또는 텍스트 마이닝 기법 등을 적극적으로 활용한 연구도 증가하고 있다. 기계학습 알고리즘인 의사결정 나무(decision tree), 로지스틱 회귀(Logistic Regression), 랜덤 포레스트(Random Forest) 등 여러 기법을 사용하여 주요 요인을 찾아내기도 하고[5,6], 텍스트 마이닝 기술을 이용하여 대량의 데이터로부터 문맥 정보(context)를 기계적으로 추출하여 직접 해석하고자 하는 내용 분석적 연구도 수행된 바 있다[2]. 그러나 이러한 기존의 연구들은 중도탈락에 영향을 주는 요인들을 밝히거나 내용 해석을 통해 현상을 이해하는 것이 목적이므로, 학생들의 패턴을 기계 학습하여 중도탈락을 적극적으로 예방할 수 있는 시스템 환경을 구축하기에는 어려움이 있었다.

본 연구는 교육 분야의 문제 해결을 위해 IT 분야의 기술을 도입하는, 학제 간(interdisciplinary) 기술 융합의 사례로서 의미가 있다. 우선 대학생의 중도탈락 패턴을 예측하는 기계 학습 기법의 성능 향상 방안을 제안하고, 이를 활용하여 중도탈락 예방을 위한 지능형 추천 시스템 환경을 구축하는 방안[7-10]을 제시하고자 하였다.

2장에서는 데이터의 수집과 저장 과정을 설명하고 데이터의 주요 필드를 소개한다, 3장에서는 기계학습 성능을 개선하기 위한 다양한 데이터 처리 방법을 제안한다. 4장에서는 기계학습 알고리즘에 기반 한 비교 실험을 수행하고 결과를 도출한다. 실험 결과를 통해 재학생의 중도탈락을 예방할 수 있는 추천 시스템의 구현 방안을 제시한다. 마지막으로 향후 연구 계획을 언급한다.

2. 데이터 수집 및 구축

실험을 위한 데이터 수집은 학내 데이터 통합관리 시스템을 통해 2017년~2020년의 4년 치 입학생 데이터 중 일부인 2,200명을 제공받았으며, 이 중 재학생은 2,000명, 제적생은 200명이었다. 재학생과 제적생 샘플의 비율은 매년 약 10% 수준으로 나타나는 실제 중도탈락 데이터의 비율을 반영하여 설정한 것이다[2].

수집된 최종 데이터는 학적 정보, 입학 정보, 장학 정보, 학사 정보, 기타 대학활동 정보 등으로 분할된 여러 테이블로부터 학번을 통해 단일 데이터 테이블로 통합 구축되었다. 통합 테이블은 총 53개의 필드로 구성되어 있으며 주요한 항목들은 Table 1에서 보는 바와 같다. 모든 데이터는 중도탈락과 관련된 개인의 민감 정보이므로 성명과 학번, 개인 주소, 연락처 등을 모두 사전 제거하고 새로운 식별 코드를 부여하여 익명화 처리하였다.

Table 1. Main fields of database

Category	Field	Category	Field
Basic	Entrance year	Acquisition	Major transfer
	Department		Major acquisition
	Current enrollment		Liberal arts acquisition
	Number of registrations		Double major acquisition
	Current semester		Minor acquisition
	Dropout (Y/N)		Exchange student grade (domestic)
	Dropout semester		Exchange student grade (abroad)
	Dropout reason	Grades	Overall rating
Admission	Admission type		Major rating
	Entrance exam type		Academic warning
	Residence information	University Life	Service time
	High School info.		Internship
Scholarship	Income deciles		Dormitory
	Off-campus scholarship		Extracurricular program
	On-campus scholarship		Extracurricular department

3. 연구방법

3.1 연구 절차

본 연구의 목적은 대학생의 중도 탈락률을 감소시키기 위한 대학의 정책적 의사결정에 활용할 수 있는 자동화된 추천 시스템 환경을 마련하는 것이다. 이를 위해 실제 대학생의 재학/제적 데이터를 기반으로 제적생 데이터의 자동 분류 성능을 극대화할 수 있는 기계 학습

방안을 마련하고자 한다.

전체 연구의 수행 절차는 Fig. 1과 같다. 우선 대학생의 학사 및 기타 활동 정보를 수집하고 전처리 후 데이터베이스에 구축한다(step1). 본 연구에서 제안하는 식별 성능을 향상시키기 위한 방법으로 다양한 데이터 유형별로 데이터 변환 및 추가 생성을 수행한다. (1)통계적 처리를 통해 성적 및 봉사시간과 같은 수치 데이터를 변환처리하고, (2)데이터 구조화를 통해 학과 및 계열 정보, 입학 구분, 장학금 관련 필드 등을 변환 처리하였으며, (3)자연어 처리를 통해 비교과 수행 내용과 같은 다양한 비정형 텍스트를 효과적으로 처리하는 방안을 제시한다(step2). 이를 바탕으로 기계 학습에 기반한 다양한 분류 알고리즘을 사용하여 성능을 비교 평가한다(step3). 마지막으로 재학생의 중도탈락 예방을 위한 추천 시스템 활용 방안을 제시한다(step4).

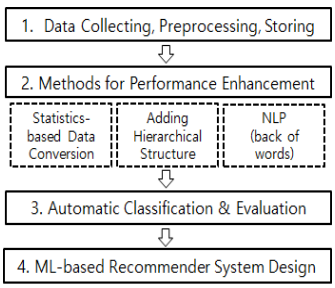

Fig. 1. Overall process

3.2 통계 기반 수치 데이터 변환

원천 데이터로부터 총 53개 필드로 구성된 1차 데이터베이스를 구축하였으며, 기계학습을 위한 다양한 방식의 데이터 추가 변환을 수행한다. 우선 연속 변수(continuous variables)인 수치 데이터 항목을 통계 기법을 이용해 범주형 변수로 변환하는 과정을 수행한다. 이는 기계학습 알고리즘 중 수치 데이터를 처리할 수 있는 트리 계열(의사결정 나무, 랜덤 포레스트 등)을 제외한 대부분의 분류기를 위해 필요한 과정이다.

본 연구에서는 종합 성적, 전공 성적, 봉사시간 등이 범주형(categorical) 데이터로의 변환이 필요한 수치형(numerical) 데이터이다. 본 연구에서는 사분위(quartile)를 활용하여 전체 구간을 25% 단위의 4개의 구간으로 범주화하여 데이터 전체의 분포를 해석하였다[11].

제적생과 재학생을 특징짓는 주요 정보인 성적 데이터를 사분위로 측정한 결과는 Table 2와 같다. 예를들어, 범주를 “하위”, “중하위”, “중상위”, “상위” 의 4단계로 구분했을 때, 전체 학생을 기준으로 종합 성적이 “3.8”이면, 해당 범주인 “종합성적_상위”로 자동 변환한다.

Table 2. Quartile conversion of continuous variables

	All students		Enrolled students		Dropout students	
	Overall Rating	Major Rating	Overall Rating	Major Rating	Overall Rating	Major Rating
Min.	0.0	0.0	0.0	0.0	0.0	0.0
1st Qt. (25%)	2.94	2.79	3.01	2.9	0.0	0.0
2st Qt. (50%)	3.32	3.33	3.36	3.38	2.53	2.13
3rd Qt. (75%)	3.71	3.81	3.73	3.83	3.22	3.25
Max	4.5	4.5	4.5	4.5	4.45	4.5

봉사 시간의 경우, 약 91.3% 이상이 ‘0’ 이었으므로 ‘0’ 값을 제외한 데이터에 대해 사분위수를 측정하여 범주화 처리를 하였다. 측정된 분위별 수치는 전체 학생 기준 1사분위: 8.0, 2사분위: 24.0, 3사분위: 54.5 이다.

마지막으로 분류 성능 향상을 위해 이진 데이터를 추가 생성하였다. ‘수행을 전혀 하지 않았다’는 의미인 ‘성적 0’, ‘봉사시간 0’과 같은 영(zero) 값은 추가로 Y/N 플래그 데이터를 생성하여 기계학습을 위한 추가 자질로 사용한다. 그 밖에 추가 생성한 이진데이터는 장학금 수혜 유/무, 행정인턴경험 유/무, 비교과수행 유/무, 기숙사 유/무 등이 있다.

3.3 데이터 구조화를 통한 계층정보 활용

데이터 분류 성능 개선을 위한 두 번째 방안은 데이터 구조화, 계층화를 통한 추가 정보 생성이다. 본 연구에서는 학사 정보와 비교과, 대내활동 데이터의 특징에 따라 구조화하여 중분류, 대분류 등의 추가 항목을 생성하는 계층화 과정을 수행하였다.

예를 들어, Table 3과 같이 각 학과들은 계열정보와 단과대학으로 계층화 할 수 있다. 여기에 대학의 고유한 비즈니스 규칙(business rule)이 반영될 수 있다. 본교의 경우, 2020학년도부터 대단위 학부제가 시행됨에 따라 이 시기 이후의 입학생들은 1학년 시기에 학과 정보가 존재하지 않는다. 또한 2020년 이전 학과정보에는 “융합 대학”과 같은 학부 정보가 존재하지 않는다. 이 경우, Table 3과 같이 데이터 구조화를 통해 기계학습 성능을 개선할 수 있는 데이터의 연관정보를 제공할 수 있다[12,13]. 즉, 데이터 구조화는 비즈니스 모델을 이해하여 데이터의 특징을 반영하는 과정이라 할 수 있다. 이

밖에 입학데이터 “수시-논술”, “수시-학생부” 등과 같은 입시 유형은 상위 범주인 “수시” 항목을 추가 생성하고, “교내장학-성적우수”, “교내장학-저소득” 등의 경우, 중분류인 “교내장학금”, 대분류인 “장학금 수혜” 등의 값을 추가 생성한다. 계층화 수준은 모델링 방식에 따라 2단계 또는 3단계로 구분할 수도 있고, 더 많은 단계를 부여할 수도 있다.

Table 3. Example of hierarchical data structure

category	raw data	Hierarchy level 1	Hierarchy level 2
Major data	Korean language and literature	College of humanities	College of convergence studies
	History		
	Economics	College of social science	
	Psychology		
	College of convergence studies *(2020-)*		
	Computer science	College of engineering	College of science and technology
	Biotechnology		
	College of science and technology *(2020-)*		
Admission data	Regular: general	-	Regular decision
	Regular: etc.		
	Early: essay	-	Early decision
	Early: Student Record		
Scholarship data	On-camp: high grades	On-campus scholarship	Scholarship
	On-camp: low income		
	Off-camp: ministry of edu.	Off-campus scholarship	
	Off-camp: etc.		

3.4 자연어 처리를 통한 자질 생성

마지막은 자연어 처리(NLP; natural language processing) 기법의 형태소 분석(morphological analysis)을 통해 기계 학습의 성능 향상을 위한 자질(features)을 추가 생성하는 과정이다[14,15]. 이를 위해 파이썬 기반의 KoNLPy (https://konlpy-ko.readthedocs.io/) 라이브러리를 사용하여 명사 상당 어구(noun phrases)를 추출하였다. KoNLPy는 품사(POS; part of speech) 태깅을 위해 Kkma, Komoran, Hannanum 등 다양한 클래스를 임포트하여 사용할 수 있다. 본 연구에서는 사전 테스트를 통해 결과가 적합하다고 판단된 Kkma를 POS 태거로 결정하였다.

본 과정은 BOW(back of words) 개념을 기반으로 추가 자질 생성을 통해 데이터 희소 환경에서 안정적으로 분류기(classifiers)가 작동할 수 있도록 한다. 추가 자질 생성을 위한 방법은 다음과 같다. “NN* + NN* + ...”와 같이 연속된 명사 패턴인 “〈NN*〉*〈NN*〉”이 나타나면 가능한 명사구의 조합을 생성 처리한다[15].

Table 4와 같이 자연어 처리 기법은 주로 다양하게 내용 서술이 되어있는 비정형 데이터 항목에 적용을 하며 장학금 수혜 내용, 입학 정보, 비교과 수행 부서와 비교과 프로그램 등이 이에 해당된다.

Table 4. Korean NLP-based noun phrases generation

Original form	Morphological analysis (NLP)
교외장학금	교외, 장학금, 교외장학금
수시-농어촌학생	수시, 농어촌, 학생, 수시농어촌, 농어촌학생, 수시농어촌학생
커리어개발센터	커리어, 개발, 센터, 커리어개발, 개발센터, 커리어개발센터

4. 실험 및 논의

4.1 실험 환경 및 평가 방법

우선 기계학습을 위한 입력용 데이터 프레임(data frame)을 생성하였다. 학습 시 바이어스(bias)가 될 수 있는 ‘제적’ 관련 정보는 사전 제거하였으며, 분류기가 행렬 연산을 수행할 수 있도록 가중치로 표현된 고차원 벡터 형식으로 변환하였다. 원천데이터를 그대로 사용한 베이스라인(Baseline) 방식과 제안 방안(Proposed Method)을 적용한 데이터프레임을 각각 생성하였다. 데이터프레임 생성에는 파이썬 기반 판다스(Pandas) 라이브러리를 활용하였다.

실험을 위해 사용한 분류기(classifiers)는 기계 학습 분야에서 널리 활용되는 k-최근접 이웃(kNN; k-Nearest Neighbor), 서포트 벡터 머신(SVM; Support Vector Machines), 랜덤 포레스트(RF; Random Forest), 나이브 베이즈(NB; Naive Bayes), 로지스틱 회귀(LR; Logistic Regression)의 5종 알고리즘이다. 분류 실험을 위해 기계 학습 및 데이터 분석 소프트웨어인 오렌지(Orange)를 활용하였다.

Table 5. 2x2 contingency table

		Observation	
		YES	NO
Prediction	YES	TP	FP
	NO	FN	TN

실험 평가 방식은 랜덤 샘플링(random sampling) 10회를 실시하여 평균값을 구하였으며, 학습 셋 비율을 전체의 95%로 설정하였다. 기계 학습 분야의 보편적인 성능 평가 지표인 정확률(precision)과 재현율(recall)을 이용하였다.

Table 6. Performance comparisons of overall experiment

		KNN			SVM			RF			NB			LR		
		F1	P	R	F1	P	R	F1	P	R	F1	P	R	F1	P	R
Overall	Baseline	0.932	0.929	0.935	0.724	0.85	0.631	0.944	0.944	0.945	0.858	0.906	0.815	0.939	0.936	0.942
	ProposedMethod	0.940	0.939	0.942	0.808	0.863	0.759	0.950	0.95	0.95	0.890	0.912	0.869	0.940	0.939	0.942
Dropouts	Baseline	0.611	0.81	0.49	0.501	0.463	0.545	0.662	0.928	0.515	0.530	0.44	0.665	0.648	0.887	0.51
	ProposedMethod	0.650	0.88	0.515	0.536	0.499	0.58	**0.694**	**0.956**	0.545	0.551	0.42	**0.80**	0.665	0.852	0.545

Table 5와 같이 분류기가 긍정으로 예측하여 정답을 맞힌 경우를 TP(true positive), 긍정으로 예측했으나 오답인 경우를 FP(false positive), 부정으로 옳게 예측한 경우를 TN(true negative), 부정으로 예측했으나 오답(긍정)인 경우를 FN(false negative)로 구분하여 정확률과 재현율을 산출한다(공식 1). 또한 두 성능 지표를 종합하여 단일 지표로 나타내기 위해 F1 점수를 사용하였다(공식 2). 실험에 사용한 각 분류기의 주요 파라미터는 kNN은 neighbors=5, euclidean distance 유사도, SVM은 linear 방식, c(cost)=1.0, RF는 생성트리수=10, ε(엡실론)=0.1, iteration=100, LR은 정규화 형=Lidge, c(strength)=1 이다.

$$Precision = \frac{TP}{TP+FP}, \quad Recall = \frac{TP}{TP+FN} \qquad (1)$$

$$F1\,Score = 2 \times \frac{Precision \times Recall}{Precision + Recall} \qquad (2)$$

4.2 실험 및 성능 비교

4.2.1 베이스라인 vs. 제안 방법 성능 비교

우선 실험을 통해 베이스라인 성능과 본 연구의 제안 방법을 비교하였다. 전체 실험 결과는 Table 6과 같다. Fig. 2는 Table 6을 요약하여 재학생을 포함하는 전체 학생(Overall)을 분류기 별로 비교한 결과이다. 성능 편차는 있으나 제안 방법을 적용한 결과가 모든 알고리즘에 대해 우수한 것으로 나타나고 있다. KNN, RF, LR이 90% 중반 대의 우수한 성능을 보이며 반면, SVM이 가장 저조한 F1 점수를 나타내고 있다. 이 경향은 이후 실험 결과에서도 유사하게 나타난다.

전체적인 성능이 우수하지만, 본 연구의 목적은 제적생의 중도탈락 패턴을 찾고 예방하기 위한 것이므로 실제로 주목해야 하는 대상은 제적생 그룹이다. 이에 따라 Table 6을 바탕으로 Fig. 3과 같이 dropouts 데이터에 대한 성능 측정 결과를 비교하였다. 실험 결과, 재학생을 배제한 제적생 그룹의 전체적인 식별 성능은 앞선 실험에 비해 저조하게 나타나고 있다. 이는 관련 연구[2]에 따르면, 중도탈락 현상은 대학생활의 초기인 1학년 1학기, 2학기에 주로 나타나기 때문에 식별을 위한 자질의 수가 부족한 현상에 의한 것이라 볼 수 있다. 전체적인 성능은 KNN, RF, LR이 우수하며 RF가 최고 69.4%의 F1점수를 보이고 있다. 또한, 모든 알고리즘에서 제안방법이 베이스라인보다 우수한 것으로 나타나고 있다.

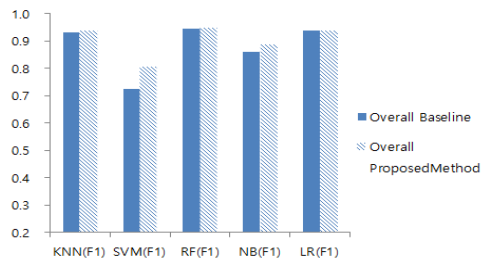

Fig. 2. F1 scores of "overall" data (comparison between baseline and proposed method)

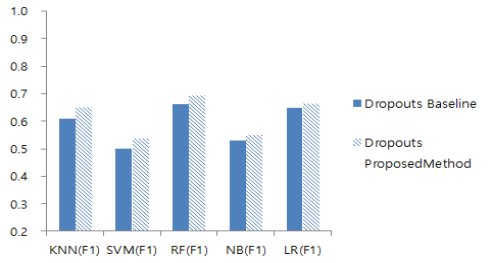

Fig. 3. F1 scores of "dropouts" data (comparison between baseline and proposed method)

4.2.2 알고리즘별 제적생의 정확률과 재현율 비교

베이스라인과의 비교 평가를 통해 제안방법을 적용할 때 식별성능의 향상이 있음을 확인하였다. 최적의 결과를 도출하기 위해, 알고리즘별로 제적생 그룹에 대

한 정확률과 재현율을 비교한 결과는 Fig, 4와 같다. 정확률은 기계(분류기)가 제적생이라 판단한 결과 중에서 제적생으로 확인된 건수의 비율로 RF가 95.6%로 가장 우수한 수치를 보여주었다. 재현율은 기계가 실제 제적생 전체 중에서 제적생이라 판단한(즉, 실제 제적생 중에서 몇 명이나 찾아냈는지를 측정한) 건수를 의미하며, NB가 대체적으로 정확률이 낮은 데 비해 재현율 성능이 우수하게 나타나 최고 80.0%로 측정되었다. 다음 장에서는 정확률과 재현율의 측정 결과를 바탕으로 수립한 시스템 활용 방안을 제시하고자 한다.

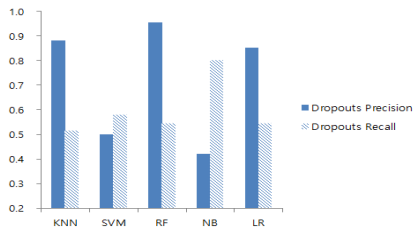

Fig. 4. Comparison between precision and recall rates of “dropouts” data (under the proposed method)

4.3 기계학습 기반 추천 시스템 구현 방안

본 장에서는 중도탈락 학생의 패턴을 예측하는 기계학습 시스템의 대학 내 활용 방안에 대해 살펴보고자 한다. 많은 대학이 학년별 지도교수 제도를 통한 학생 상담 과정을 매학기 운영하며 모든 학생의 100% 상담 달성을 목표로 하고 있다. 이때 기계학습을 활용한 예측 결과를 이용한다면 매 학기 실시하는 상담의 우선순위를 데이터에 근거하여 설정할 수 있을 것이다. 요약하자면, Fig. 5와 같이 정확률 우선 1차 상담, 재현율 우선 2차 상담, 미상담 잔여 학생에 대한 3차 상담으로 체계적인 학생 상담 서비스 전략을 수립할 수 있다. 정확률 우선 전략은 최소의 노력으로 신속하고 효율적으로 대상을 찾는 과정이며, 재현율 우선 전략은 중도탈락 예상 학생을 최대한 많이 찾아내는 과정이라 볼 수 있다.

본 연구의 실험 결과를 적용한다면, 랜덤 포레스트(RF) 기반으로 찾아낸 학생들을 대상으로 1차 상담을 조속히 진행(중간고사 직후)하고 나이브 베이즈(NB)를 기반으로 광범위하게 찾아낸 학생들을 대상으로 2차 상담을 진행(기말고사 전후)한 후, 학기 마감 시점까지 나머지 재학생을 상담한다. 이와 같이 학생 상담 서비스의 100% 목표를 달성하면서 동시에 중도탈락 예방 정책을 수행할 수 있다.

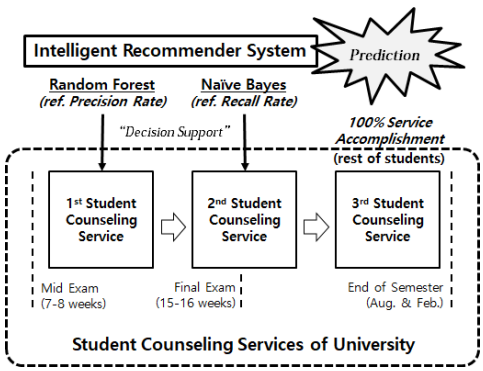

Fig. 5. Strategic utilization of machine learning-based data analytics system

5. 결 론

본 연구는 중도탈락의 가능성이 있는 대학생의 패턴을 찾아내는 효율적인 기계학습 방안을 도출함으로써 대학이 중도탈락률을 감소시키기 위해 기계학습 기반의 지능형 추천 시스템을 어떻게 활용할 수 있는가를 제시하고자 하였다.

이를 위해 1) 실제 대학생의 재학/제적 데이터를 기반으로 기계 학습의 성능을 향상시킬 수 있는 데이터 처리 방안을 제안하고, 2) 5종의 기계학습 알고리즘을 이용해 비교 실험을 실시하였다. 3) 실험 결과, 베이스라인에 비해 제안 방법이 모든 알고리즘에 대해 우수한 성능을 보여 주었다. 제적생의 식별 정확률은 랜덤 포레스트(RF) 알고리즘을 사용했을 때 최대 95.6%, 제적생의 재현율은 나이브 베이즈(NB)를 사용할 때 최대 80.0% 성능을 보여주었다. 4) 마지막으로, 실험 결과를 바탕으로 중도탈락 가능성이 있는 학생을 우선 상담하는 목표 전략을 수립할 수 있도록 지원함으로써, 효율적인 의사결정을 지원하는 기계학습 기반의 추천 시스템의 구축방안을 제시하였다.

본 연구는 교육 분야에서의 당면 문제를 해결하기 위해 텍스트 마이닝 기법을 적극적으로 도입한 실험적 연구로써 대표적인 자동분류 알고리즘들을 적절히 활용하여 기술 융합을 통한 사회적 문제해결의 가능성을 제시하고자 하였다. 향후 지속적인 연구를 통해 보다 다양한 인공지능 기법을 적용하여 성능을 향상시키고자 한다.

REFERENCES

[1] J. Y. Chung, M. Sun & M. J. Jeong. (2015). An Analysis of Institutional Factors Affecting on College Dropout Rates. *Asian Journal of Education, 16(4),* 57-76.
URI : https://hdl.handle.net/10371/95751

[2] D. H. Jeong & J. Y. Park. (2021). Data Analysis of Dropouts of College Students Using Topic Modeling. *Journal of the Korea Institute of Information and Communication Engineering, 25(1),* 88-95.
DOI : 10.6109/jkiice.2021.25.1.88

[3] P. Perchinunno, M. Bilancia, & D. Vitale. (2021). A Statistical Analysis of Factors Affecting Higher Education Dropouts. *Social Indicators Research, 156,* 341-362.
DOI : 10.1007/s11205-019-02249-y

[4] M. Kang, E. Lee & E. Lee. (2019). Trends and influencing factors of college student's dropout intention. *In Forum for Youth Culture, 58,* 5-30.
DOI : 10.17854/ffyc.2019.04.58.5

[5] C. Park. (2020). Development of Prediction Model to Improve Dropout of Cyber University. *Journal of the Korea Academia-Industrial Cooperation Society, 21(7),* 380-390.
DOI : 10.5762/KAIS.2020.21.7.380

[6] E. J. Lee, Y. Song, J. H. Kim & S. H. Oh. (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.

[7] H. J. Kim, H. S. Lee, B. J. Choi, & Y. H. Kim. (2019). Machine Learning-based Quality Control and Error Correction Using Homogeneous Temporal Data Collected by IoT Sensors. *Journal of the Korea Convergence Society, 10(4),* 17-23.
DOI : 10.15207/JKCS.2019.10.4.017

[8] D. H. Jeong. (2017). Prescriptive analytics system design fusing automatic classification method and intellectual structure analysis method. *Journal of the Korean Society for information Management, 34(4),* 33-57.
DOI : 10.3743/KOSIM.2017.34.4.033

[9] S. Zhang, L. Yao, A. Sun, & Y. Tay. (2019). Deep Learning Based Recommender System: A Survey and New Perspectives. *ACM Computing Surveys, 52(1),* 1-38.
DOI : 10.1145/3285029

[10] K. Lepenioti, A. Bousdekis, D. Apostolou, & G. Mentzas. (2020). Prescriptive analytics: Literature review and research challengesInternational. *Journal of Information Management, 50,* 57-70.
DOI : 10.1016/j.ijinfomgt.2019.04.003.

[11] S. Goswami & A. Chakrabarti. (2012). Quartile Clustering: A quartile based technique for Generating Meaningful Clusters. *Journal of Computing, 4(2),* 48-55. arXiv: 1203.4157

[12] R. Bai, X. Wang, & J. Liao. (2010). Extract Semantic Information from WordNet to Improve Text Classification Performance. *AST 2010, ACN 2010: Advances in Computer Science and Information Technology.* 409-420
DOI : 10.1007/978-3-642-13577-4_36

[13] A. Kehagias, V. Petridis, V. G. Kaburlasos, & P. Fragkou. (2003). A Comparison of Word- and Sense-Based Text Categorization Using Several Classification Algorithms. *Journal of Intelligent Information Systems, 21,* 227-247.
DOI : 10.1023/A:1025554732352

[14] A. Stavrianou, C. Brun, T. Silander, & C. Roux. (2014). NLP-based feature extraction for automated tweet classification. *Proceedings of the 1st International Conference on Interactions between Data Mining and Natural Language Processing, 1202,* 145-146.
https://aclanthology.org/2020.nlpcovid19-acl.17.pdf

[15] D. H. Jeong. (2019). Enhancing Classification Performance of Temporal Keyword Data by Using Moving Average-based Dynamic Time Warping Method. *Journal of the Korean Society for information Management, 36(4),* 83-105.
DOI : 10.3743/KOSIM.2019.36.4.083

정 도 헌(Do-Heon Jeong) [정회원]

· 1997년 8월 : 연세대학교 문헌정보학과(학사)
· 2003년 8월 : 연세대학교 문헌정보학과 대학원(석사)
· 2014년 2월 : 연세대학교 문헌정보학과 대학원(박사, 정보공학)

· 2017년 3월 ~ 현재 : 덕성여자대학교 글로벌융합대학 조교수
· 관심분야 : 자동분류, 토픽모델링, 텍스트마이닝, SNA
· E-Mail : doheonjeong@duksung.ac.kr