Journal of the Korea Academia-Industrial cooperation Society
Vol. 25, No. 2 pp. 707-714, 2024

https://doi.org/10.5762/KAIS.2024.25.2.707
ISSN 1975-4701 / eISSN 2288-4688

학습자 유형 분석 및 성과 예측을 통한 맞춤형 이러닝 관리 시스템 제안

김민영, 윤선영, 김수현*
경북대학교 데이터사이언스대학원

A Personalized E-Learning Management System via Learner Type Analysis and Performance Prediction

Minyoung Kim, Sun-Young Yoon, Suhyeon Kim*
Graduate School of Data Science, Kyungpook National University

요 약 최근 에듀테크 산업이 활성화됨에 따라 다양한 분과에서 이러닝 학습자의 수는 꾸준히 증가하는 추세를 보이고 있다. 이러한 경향성은 이러닝에 대한 관심과 필요성이 높아지고 있다는 것을 시사하며, 이에 대응하기 위해 학습자들의 다양한 학습 성향을 이해하고 맞춤형 학습 방법을 제공하는 것이 중요하다. 본 연구에서는 실제 에듀테크 산업 데이터를 머신러닝 방법론들을 이용하여 분석하고, 분석 결과를 바탕으로 학습자 맞춤형 이러닝 학습 관리 시스템을 제안하고자 한다. 이러닝 학습자들의 학습 성향 및 패턴을 파악하기 위해 K 평균 군집화를 통한 학습자 유형 분석을 진행하였다. 또한, 다양한 머신러닝 기반 예측 모델들을 사용하여 학습 성과 예측 분석을 진행하였으며, 그 중 가장 우수한 성능을 보인 앙상블 알고리즘인 랜덤 포레스트 모델(MSE: 0.011, MAE: 0.087)을 최종 예측 모델로 선정하여 학습자별 성과를 예측하였다. 추출된 학습자 군집 유형과 학습 성과 예측치에 따라 학습자 맞춤형 학습 관리 방법을 제안함으로써, 학습 효율을 높이고 최적의 학습 환경을 제공하고자 한다.

Abstract As the EduTech industry has been gaining momentum in recent times, there has been a steady increase in the number of e-learning learners across various sectors. This trend indicates a growing interest in and the need for e-learning, so it is essential to capture the diverse patterns of e-learners and provide personalized learning methods. This study aimed to analyze actual EduTech industry data using machine learning methods and propose a personalized e-learning management system for e-learners. Initially, clustering analysis for e-learner type analysis through K-means clustering was conducted. Then e-learning performance prediction was analyzed using different machine learning models. Finally, random forest, an ensemble algorithm, was selected to identify the best prediction accuracy (MSE: 0.011, MAE: 0.087). Extracted cluster types and predictions showed that optimal learning environments that enhance learning efficiency can be provided.

Keywords : Edutech, E-learning, Machine Learning, Clustering, Ensemble Learning

This work was supported by the National Research Foundation of Korea (NRF) grant funded by the Korea government (MSIT) (RS-2023-00242528) and by the Ministry of Education (No. RS-2023-00245529).
The authors equally contributed: Minyoung Kim, Sun-Young Yoon
*Corresponding Author : Suhyeon Kim(Kyungpook National University)
email: suhyeonkim@knu.ac.kr
Received January 22, 2024 Revised February 5, 2024
Accepted February 6, 2024 Published February 29, 2024

1. 서론

4차 산업의 기술이 발달함에 따라, 머신러닝, 인공지능 등은 다양한 산업에서 필수적인 기술로 활용되고 있다. 교육 분야에서도 관련 디지털 기술들이 빠르게 확산 및 보급되어, 교육과 기술의 합성어인 '에듀테크(Edutech)', '이러닝(e-learning)' 등의 용어들이 등장하였다[1]. 에듀테크는 AR/VR, 빅데이터 및 인공지능 등을 활용하여 디지털교과서 제작, 개인 맞춤형 학습 방법 제안, 마이크로 콘텐츠 활용 등을 통해 학습자들에게 도움을 주는 역할을 한다[2].

교육의 학습 현장도 변화하고 있다. 전통적인 학교에서의 제도화된 교육과 교과서를 통한 표준화된 교육에서 벗어나 학습자의 역량을 고려한 교육을 중심으로 에듀테크를 활용한 학습 현장이 등장하였다[3]. 이러한 에듀테크의 확산은 개인 맞춤형 교육을 향한 미래 지향적인 방향을 강조한다. 또한 교육 현장에서 데이터의 중요성을 강조하여 학습자 개개인의 역량과 필요를 더욱 세밀하게 이해하고 반영하려는 교육의 새로운 방향성을 나타낸다. 이에 따라 교육 전문가들과 연구자들은 관련 데이터를 깊게 탐구하고 분석하여 학습자들의 역량과 그들의 필요에 최적화된 교육 방법론을 개발하고 추진하는 것이 필요하다.

빅데이터 분석 기술의 발달과 함께, 최근 교육 분야에서도 학습 관련 데이터를 분석하여 학습 특성을 파악하고 학습 성과를 개선하기 위한 학습 데이터 분석 접근이 적극적으로 시도되고 있다[4]. 학생의 학습 과정에서 생성되는 방대한 이러닝 데이터를 다양한 디지털 분석 기술을 적용하여 학생의 학습 성과 및 교육 환경을 개선하기 위한 관련 연구들이 진행되어 왔다[5]. 그러나, 기존의 교육/학습 관련 데이터 분석에서는 주로 공공 대학 교육 환경에서의 온라인 학습관리시스템(LMS: Learning Macagement System)이나 MOOC로부터 축적되는 데이터를 활용하여 분석을 진행해 왔으며[3], 실제 에듀테크 산업에서의 기업 데이터 셋을 활용하여 진행한 머신러닝 기반 데이터 분석은 극히 드물다.

이에 본 연구에서는 머신러닝 기반 에듀테크 산업 데이터 분석을 통한 학습자 맞춤형 이러닝 학습 관리 시스템을 새롭게 제안하고자 한다. 본 연구에서는 에듀테크 학회에서 제공한 에듀테크 업체의 이러닝 데이터셋을 활용하여 학습자들의 학습 데이터를 기반으로 여러 머신러닝 기법을 접목시킨 학습 관리 시스템을 제안한다. 이러닝 학습자들의 학습 성향 및 패턴을 파악하기 위해 군집 분석 알고리즘을 활용하여 학습 유형을 분석하고, 학습 포트폴리오를 기반으로 앙상블 모형을 통한 학습 성과 예측 분석을 함께 진행하는 시스템을 구축하였다. 이를 통해 학습 유형별 성향을 파악 및 학습 전략을 제안하여 교육의 효율성을 높이며 학습자들의 학습 동기를 증진해 학습 성취도 향상을 기대하고자 하였다.

2. 관련 연구

2.1 학습 데이터 분석

학습 데이터 분석은 교육 데이터 마이닝의 일종으로, 학습자 및 학습자의 맥락과 관련된 데이터를 분석 및 보고하는 것으로 정의되며[4], 주로 학습과 관련된 데이터가 대량으로 생산, 축적되는 이러닝에서 학습자 행동 모델링 및 학습 성과 예측 등을 목적으로 진행되고 있다.

기존에는 통계적인 방식을 통해 패턴이나 규칙을 찾는 연구들이 많이 진행되었으나, 최근에는 다양한 머신러닝 기반의 예측 모델을 적용하는 데에 초점을 맞추고 있다. 대표적인 분석 목적 중, 본 연구와 맥락을 같이 하는 학습자 유형 분석과 학습 성과 예측 분석 등도 최근 머신러닝 기법들을 도입하고 있다.

대부분의 대학에서 운영하는 MOOC 기반 강좌들이나 LMS 플랫폼에서는 대학교육과 관련한 에듀테크 데이터들이 축적되어 왔다. 이러한 데이터를 기반으로, 학생들의 보고서 작성에 전문가의 평가의 도움 여부를 서포트 벡터 머신(SVM: Support Vector Machine)을 활용하여 분류하는 연구[6]가 진행된 바 있다. 또한, MOOC 데이터를 활용하여 머신러닝 기반 수강생 공통 특성을 추출 연구가 진행되기도 하였다[7]. 더하여, 학습자 대상 설문 데이터를 바탕으로 k-Nearest Neighborhood 기법을 이용하여 학습자의 학습 행위 데이터의 이상치를 분석하기도 하였다[8].

그러나 대부분의 선행 연구에서는 이러닝 학습자들의 데이터 기반으로 주로 요인 분석이나 분류 분석 등을 통해 학습자들의 현황을 파악하는 형태에 그친 바 있다. 본 연구에서는 머신러닝 기술을 통한 단순 결과 추출에 그치지 않고, 에듀테크 분야의 실산업 데이터를 기반으로 머신러닝을 활용한 이러닝 학습자 데이터 분석 프로세스를 제안한다. 제안하는 머신러닝 기반 학습자 맞춤형 이러닝 학습 관리 시스템은 기존 선행연구의 단편화된 데이터 분석 특징을 넘어 복합적인 인사이트를 제공할 수 있을 것으로 사료된다.

3. 연구 방법

3.1 분석 데이터 및 전처리

본 연구에서는 에듀테크 학회가 주최한 데이터 분석 경진대회에서 제공받은 리딩앤 에듀테크 기업의 실산업 데이터 셋을 활용하여 분석을 진행하였다. 리딩앤은 디지털 기반 영어 리딩 프로그램을 제공하는 기업으로, 본 연구에서는 리딩앤에서 제공하는 온라인 교육 과정을 수강한 학습자를 대상으로 수집된 학습 포트폴리오 데이터를 이용하였다. 리딩앤의 영어 리딩 프로그램은 Table 1과 같이 5단계로 구성되어 있다.

데이터는 2023년 1월부터 6월까지 총 6개월 동안 학습자별로 수집되었으며, 총 9가지 이러닝 학습 관련 변수들에 대한 189명의 월별 학습 통계 데이터를 최종 분석에 활용하였다. 분석에 활용된 변수들은 Table 2에 설명되어 있다.

Table 1. Multi-learning in five stages of reading program

Stage	Definition
Stage 1 (Warm Up)	Study with a book word game
Stage 2 (Listen Up)	Listen to stories with voice without text
Stage 3 (Read)	Spontaneous reading
Stage 4 (Speak Up)	Listen to and repeat five key sentences and check your score
Stage 5 (Wrap Up)	Make a final check with a game

Table 2. Variables related to monthly learning statistics

Respondents	Example
Study Time at Stage 1	1628 seconds
Study Time at Stage 2	1191 seconds
Study Time at Stage 3	1342 seconds
Study Time at Stage 4	639 seconds
Study Time at Stage 5	851 seconds
Monthly Study Days	14 days
Average Pronunciation Score	58.3
Total Number of Words (Books Completed in Stage 3)	36
Total Number of Books (Completed in Stages 1 to 5)	14

본 연구에서는 모델의 성능을 높이고 분석의 정확성을 향상시키기 위해 다음의 데이터 전처리를 진행하였다. 우선 데이터에 존재하는 이상치를 처리하였다. 이상치는 데이터 측정 과정이나 데이터를 입력할 때의 오류 등으로 인해 발생할 수 있는 특이한 값을 의미한다. 예를 들어 학습자가 학습 도중 자리를 이탈하여 학습 시간이 지속해서 누적되는 경우, 학습자의 실제 학습 시간과 관계없이 높은 학습 시간 데이터를 생성하여 분석 결과의 왜곡을 초래할 수 있다. 따라서 이상치의 발생 가능성이 있는 변수인 단계별 학습 시간 변수들을 대상으로 BoxPlot을 활용하여 극단값을 제거 후 분석에 활용하였다. 또한, 분석에 활용된 변수들이 각기 다른 범위를 가지고 있어 변수들의 수치적 일관성을 확보하고 모델의 성능 최적화를 위해 Min-Max 스케일링 방법을 통해 표준화를 실시 후 분석을 진행하였다.

3.2 분석 방법

본 연구에서는 다음의 두 가지 학습 데이터 분석 방법을 포함하는 머신러닝 기반 학습자 맞춤형 이러닝 학습 관리 시스템을 제안한다: (1) 학습자 유형 분석, (2) 학습 성과 예측 분석. Fig. 1은 본 연구에서 제안하는 시스템의 전반적인 순서를 나타낸다. 각 단계에 대한 자세한 내용은 아래에서 설명하고자 한다.

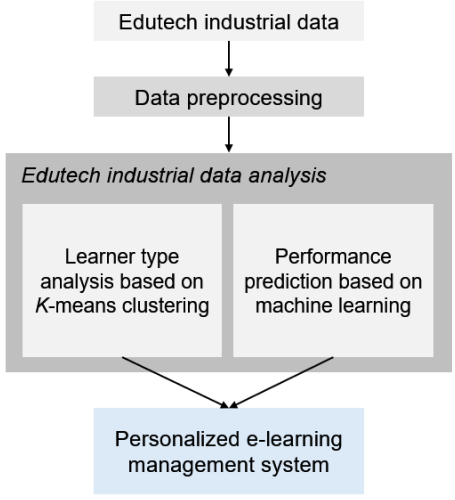

Fig. 1. Overall process of proposed method

3.2.1 Phase 1. 군집 분석 기반 학습자 유형 분석

본 연구에서는 유사한 학습 특성을 가진 학습자들을 효율적으로 관리하기 위해, 이러닝 학습 데이터를 기반으로 비슷한 특징을 가진 학습자들을 동질적인 집단으로

묶어 군집에 따른 특성을 유형화하고자 한다. 본 연구에서는 군집 분석 기법 중 가장 대표적인 알고리즘인 K-평균 군집 분석(K-means clustering)[9]을 통한 학습자 유형 분석을 진행하였다. 분석에 활용된 K-means 군집 분석의 주요 단계는 Table 3에 설명되어 있다. 최적 군집의 수 K를 설정하기 위해, 군집 간의 거리와 밀도를 감안하여 군집을 평가할 수 있는 실루엣 계수(Silhouette Factor)를 군집 수의 변화에 따라 분석하여 데이터를 잘 표현할 수 있는 적절한 군집 수를 최종적으로 선정하였다[10]. 본 연구에서는 실제 에듀테크 산업에서의 이러닝 학습 데이터로부터 새롭게 추출된 군집들의 특성을 기반으로 군집별 대표 명칭을 부여하였으며, 각 군집들의 특성을 세부적으로 분석하고 학습자 유형별 맞춤형 학습 전략을 제안함으로써 학습자의 학습 역량을 올리고자 하였다.

Table 3. Main stages of K-means clustering for proposed method

Main Stages	Contents
Initialization	To select the optimal number of clusters, we use the silhouette factor to select K.
Assignment	Each data point is assigned to the nearest cluster center, taking into account the different learning features of the learner.
Convergence Check	Repeat the process of assigning clusters and updating centers for improving the accuracy of learner clustering.
Result Retrieval	Suggest the personalized learning management for learners based on the cluster results.

3.2.2 Phase 2. 머신러닝 모형 기반 학습 성과 예측

본 연구에서는 영어 리딩 프로그램 학습 포트폴리오 데이터를 바탕으로 학습자의 지속적인 언어 능력 향상을 지원하기 위해, 머신러닝 모형 기반의 학습 성과 예측 모델링을 진행한다. 본 연구에서 사용하는 데이터의 대표적인 학습 성과로서 학습자의 발음 점수를 종속변수(i.e., 타겟 데이터)로 선정하고 나머지 변수들은 독립변수로 예측을 진행하였다. 본 연구에서는 다양한 머신러닝 모형들을 비교 분석하기 위해 회귀 분석(Linear Regression), 릿지 회귀 분석(Ridge Regression), 서포트 벡터 머신[11], 그래디언트 부스팅(Gradient Boosting)[12], XGBoost [13] 및 랜덤 포레스트(Random Forest)[14]의 6가지 머신러닝 모형들을 적용하여 예측을 진행하고 성능을 비교하였다. 분석 데이터를 훈련 및 테스트 데이터로 분할한 후 훈련 데이터를 이용하여 모델을 생성하였으며, 성능 향상을 위해 그리드 서치 기법을 이용하여 모델별 최적의 하이퍼파라미터(Hyperparameter)를 선정하였다[15]. 모델의 성능은 테스트 데이터에 대해 평균 제곱 오차(MSE: Mean Squared Error) 및 평균 절대 오차(MAE: Mean Absolute Error)를 사용하여 평가하였으며, 가장 성능이 좋은 모형을 선정하여 통해 발음 점수를 예측하는 데 영향을 미치는 변수들을 파악하고, 학습자의 다음 달 발음 점수를 최종적으로 예측하였다. 예측 결과를 바탕으로 학습자의 차별화된 학습 전략을 제안함으로써 결과적으로 학습자의 발음 능력 향상을 도모하고, 평균 발음 점수 향상을 기대할 수 있다.

4. 연구 결과

4.1 학습자 유형 분류 및 유형별 학습 특성 분석 결과

본 연구에서는 학습자별 데이터에 대해 K의 수를 2에서 10으로 변화시켜 가며 K-평균 군집 분석을 시행하였으며, K에 따라 변화하는 실루엣 계수를 측정하였다(Fig. 2). Fig. 2를 살펴보면 군집의 개수가 4개일 때 실루엣 계수가 0.4157로 가장 높은 값을 나타낸다. 이를 통해, 최적의 군집 수를 4개로 선정하였으며, 학습자 유형을 4종류로 분류하여 유형별 학습 특성에 따라 각각 '열중형(1군집)', '빠른습득형(2군집)', '표준형(3군집)', 그리고 '독서형(4군집)' 집단으로 명명하였다.

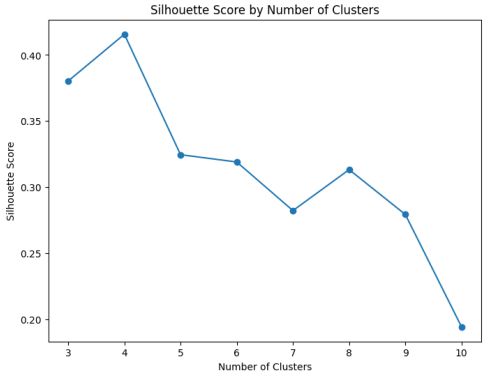

Fig. 2. Silhouette factor scores by the number of clusters

각 학습자 유형별 특성을 상세히 분석하고자, 군집별 학습 특징 차이를 나타내는 변수별 평균값과 학습 시간 분포를 각각 Table 4에 기술하였다. Table 4를 살펴보면 열중형 학습자들은 다른 유형에 비해 월간 학습 일수,

1단계 이상 완료한 도서 수, 학습 단어 수와 평균 발음 점수가 높은 값을 나타낸다.

Fig. 3에서 보는 바와 같이 열중형 학습자들은 전체 학습자의 평균 학습 시간에 비해 1~5단계 학습 시간이 월등히 많았음을 알 수 있다. 이를 토대로, 열중형 학습자들은 다른 유형 대비 많은 양의 학습을 열정적으로 수행했다는 점을 알 수 있다. 빠른 습득형 학습자들의 경우 전체 학습자의 평균 학습 시간에 비해 비 참여적인 학습 시간 분포를 보였다. 완료 도서 수, 월간 학습 일수, 학습 단어 수도 평균보다 높은 값을 나타낸다. 표준형 학습자들은 모든 단계에서 평균적인 분포를 하고 있다. 월간 학습 일수를 보아 꾸준히 학습을 수행하고, 완료 도서 수, 학습 단어 수, 평균 발음 점수도 평균적인 분포를 보인다. 마지막으로, 독서형 학습자들은 전체 학습자의 평균 학습 시간에 비해 3단계 학습 시간에만 높은 값을 가지는 것을 확인할 수 있으며, 자발적인 독서를 하는 3단계에서 많은 시간을 쏟는 것으로 보인다.

Table 4. Study data for each cluster distribution

Respondents	Cluster 1	Cluster 2	Cluster 3	Cluster 4
Study Time at Stage 1	10862	1562	4595	1005
Study Time at Stage 2	22448	2247	8531	3258
Study Time at Stage 3	27405	3409	10175	11676
Study Time at Stage 4	10292	1953	4953	452
Study Time at Stage 5	10262	1354	3783	724
Monthly Study Days	22	9	20	8
Average Pronunciation Score	47.2	46.7	41.9	4.7
Total Number of Words (Books Completed in Stage 3)	15908	2008	3684	5719
Total Number of Books (Completed in Stages 1 to 5)	105	17	47	25

학습 유형에 따른 학습자 평균 발음 점수를 상세히 비교 분석해보고자 한다. 학습 유형에 따른 학습자의 평균 발음 점수인 41.4점과 비교 분석한 결과, 열중형 학습자들의 평균 발음 점수가 47.2점으로 가장 높았고, 독서형 학습자들의 평균 발음 점수가 4.7점으로 가장 낮았다. 독서형 학습자들은 학습을 수행하고는 있지만, Table 1을 참고하면 자발적인 독서를 하는 3단계에서 많은 시간을 쏟고 발음 점수를 확인할 수 있는 4단계 학습을 상대적으로 건너뛰어 낮은 발음 점수의 값을 갖는 것으로 파악되었다. 빠른 습득형 학습자들은 평균 발음 점수보다 높은 46.7점을 갖는 것을 보아 학습 난이도가 쉽거나 짧은 시간 내에 빠른 습득으로 성적 향상이 가능한 것으로 판단하였다.

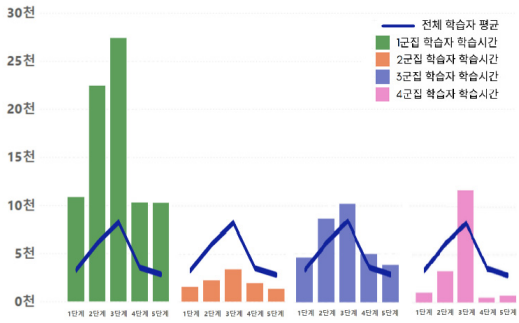

Fig. 3. Learning time distribution by each cluster

4.2 학습 성과 예측 모델 결과

본 연구에서는 학습자가 언어 능력을 지속해서 향상할 수 있도록 학습 데이터를 기반으로 학습자들의 개인별 다음 달의 발음 점수를 예측하는 다양한 모델의 성능을 비교하였다. 총 6가지 모델들의 성능 비교 결과는 Table 5에 나타나 있다.

Table 5. Performance prediction results of six machine learning models

Model	MSE	MAE
Random Forest	**0.01167**	**0.08759**
Linear Regression	0.01581	0.10156
SVM	0.01685	0.10339
Gradient Boosting	0.01406	0.08983
Ridge	0.01604	0.10248
XGBoost	0.01573	0.10091

성능 메트릭인 MSE와 MAE의 수치를 살펴본 결과, 랜덤 포레스트 모델의 MSE 값이 0.01167, MAE 값은 0.08759로 다른 모델들에 비해 가장 낮은 값을 보여 예측 성능이 좋다고 판단하였다. 따라서, 랜덤 포레스트 모델을 최종 예측 모델로 선정 후 예측 결과를 분석하였다. Table 6은 열중형과 독서형 군집에서 무작위로 선발된 두 사용자의 학습 특성을 바탕으로 다음 달 발음 점수를

예측한 결과이다. Table 6을 살펴보면, 본 연구의 최종 예측 모델은 User 1의 경우 이전 달 대비 약 3점, User 2의 경우 약 7점 상승한 발음 점수를 예측하였다. User 1은 일관된 학습 빈도와 충분한 학습량을 바탕으로 더 높은 발음 점수 향상이 예측되었으며, 이는 지속적인 학습량 증가에 따른 발음 능력 향상을 나타낸다. 꾸준히 학습량을 늘려간다면 더욱 높은 발음 점수를 예측해 볼 수 있을 것으로 기대된다.

Table 6. An example of comparisons between two users

Respondents	User 1	User 2
Study Time at Stage 1	11129	745
Study Time at Stage 2	16347	459
Study Time at Stage 3	15463	28213
Study Time at Stage 4	7234	491
Study Time at Stage 5	13105	684
Monthly Study Days	23	9
Average Pronunciation Score	40.9	24.6
Total Number of Words (Books Completed in Stage 3)	21854	32823
Total Number of Books (Completed in Stages 1 to 5)	88	85
Performance Prediction		
Next Month's Pronunciation Score	**43.9**	**31.9**

User 2는 주로 3단계에서의 학습 시간이 상대적으로 높은 비율을 차지하며, 발음 점수를 확인할 수 있는 4단계의 학습은 거의 이루어지지 않아 학습의 균형의 필요함을 나타낸다. 학습 일수를 증가시키고 모든 단계에서의 학습을 균등하게 수행할 경우, 발음 점수의 더 큰 상승을 예측해 볼 수 있을 것으로 사료된다.

5. 결론 및 제언

본 연구에서는 4차 산업의 기술이 발달함에 따른 교육의 변화와 에듀테크 등장에 주목하여 학습자들의 학습 특성을 파악하여 그에 맞는 학습 전략을 제안하는 것이다. 에듀테크 업체의 데이터를 활용하여 교육을 수강한 학습자들의 학습 특성을 분석하여 군집으로 분류하였고 4가지의 학습 유형을 도출하였다. 또한 학습 유형에 따른 학습자의 발음 점수를 예측하였다.

분석 결과, 학습자들은 열중형, 빠른습득형, 표준형, 독서형의 4가지 유형으로 분류되었다. 군집에 따른 학습 전략은 다음과 같다. 열중형 학습자들에게는 발음 점수 예측을 제공함으로써 꾸준한 학습을 장려한다. 또한, 발음 점수 예측을 통해 학습자들에게 발음 향상의 가능성을 보여줌으로써 학습 동기를 높인다. 빠른습득형 학습자들은 난이도 조절이 필요한 학습자가 있을 것으로 판단되며, 우수한 학습 능력을 갖춘 학습자에게는 학습 일수, 시간을 늘리는 것을 권장함으로써 학습 성과를 극대화하도록 독려한다. 또한, 표준형 학습자들에게는 더 좋은 발음 점수를 얻기 위해서 학습 시간을 추가로 투자할 수 있도록 권장한다. 또한 학습 시간의 증진을 통해 어휘력을 향상시킬 수 있도록 한다. 마지막으로, 독서형 학습자들은 3단계에서 많은 시간을 투자하고 있기 때문에, 더 많은 단어를 습득할 수 있는 맞춤형 도서를 추천한다. 학습 이해도를 확인하기 위해 학습의 모든 단계를 완수하도록 적극 독려한다.

발음 점수 예측을 통해서는 다음과 같은 기대효과를 불러올 수 있다. 첫째, 학습자들은 발음 향상을 위해 예측된 발음 점수를 고려하여 다음 달의 학습 계획을 세울 수 있다. 개인의 발음 능력 향상을 위해 집중적으로 계획을 수립하는 데 도움을 줄 수 있다. 둘째, 예측된 발음 점수는 학습자에게 목표 달성에 대한 동기부여를 제공한다. 목표를 달성하면 예측 점수와 비교하여 얼마나 성장했는지 확인함으로써 학습 성취감을 주기도 한다. 셋째, 발음 점수 예측을 기반으로 개별화된 학습 자료나 연습 방법을 제공하여 학습자가 개인적으로 발음을 향상시킬 수 있도록 돕는다. 이와 같이 학습자들의 학습 최적화에 도움을 줄 수 있을 것이다.

학습자들의 측면에서는 데이터 수집 후 군집 별 맞춤형 학습 방법이 제안되어 이로 인해 학습 동기 유발이 기대된다. 또한, 발음 점수 예측의 기능을 통해 학습자들은 자신의 발음 능력을 개선하기 위한 구체적인 목표와 계획을 세울 수 있게 된다. 에듀테크 기업 측면에서는 학습자들의 학습 동기의 상승으로 인한 학습률 향상으로 수집되는 데이터의 양도 증가할 것으로 기대된다. 증가된 데이터는 기업들이 더욱 정밀한 데이터 분석을 가능하게 하며, 이를 기반으로 학습자들에게 최적화된 서비스와 학습 경험을 제공하는 데 도움을 줄 것이다.

본 연구의 시사점은 다음과 같다. 첫째, 군집별 맞춤형 학습 방법이 제안됨에 따라 학습자들의 공통으로 선호하는 분야, 학습 방법 등을 파악하고 이에 맞는 맞춤형 교육이나 서비스를 제공하는 것의 중요성이 강조된다. 이는 학습자들의 참여도와 학습 효과를 극대화하고 학습의 질을 높이게 된다. 둘째, 학습 유형을 더욱 정확하게

파악하기 위해서는 광범위하고 다양한 데이터의 수집이 필수이다. 이렇게 얻어진 분석 결과를 통해 교육과 서비스를 지속적으로 개선할 수 있으며, 학습자에게 학습에 필요한 맞춤형 서비스를 제공하고 보다 효과적인 교육 환경을 구축할 수 있다는 점에서 중요하다.

학습자 중심의 학습 최적화와 교육 방향성을 강조한 본 연구는 에듀테크 산업 전반에 걸쳐 참고하고 활용될 가치가 크다. 본 연구에서는 영어 리딩과 관련된 에듀테크 산업 데이터를 활용하였으나, 제안하는 머신러닝 기반 학습 관리 시스템은 학습자의 정량적 수치 정보 데이터 및 학습 성과 데이터가 존재하는 다양한 에듀테크 산업에 범용적 활용이 가능하다. 향후에는 다양한 학습 데이터와 신기술의 통합을 통해 개인별 학습 경험과 효과를 더욱 세밀하게 분석하고 학습 최적화가 필요하다. 연구의 지속적인 추진을 통해 에듀테크 산업의 끊임없는 발전과 혁신을 끌어내고 교육자와 학습자 모두에게 더욱 향상된 교육 환경을 제공하기를 기대한다.

References

[1] H. M. Alakrash, N. A. Razak, "Education and the Fourth Industrial Revolution: Lessons from COVID-19", *Computers, Materials & Continua*, Vol.70, No.1, 2022. DOI: https://doi.org/10.32604/cmc.2022.014288

[2] J. S. Park, J. M. Gil, "Edutech in the Era of the 4th Industrial Revolution", *KIPS Transactions on Software and Data Engineering*, Vol.9, No.11, pp.329-331, 2020. DOI: https://doi.org/10.3745/KTSDE.2020.9.11.329

[3] S. H. Jin, "Exploring learning data for supporting self-directed learning in the perspective of learning analytics", *Journal of Educational Technology*, Vol.32, No.3, pp.487-533, 2016.

[4] S. Kim, "In the digital big data classroom reality and application of smart education: Learner-centered education using edutech", *Journal of the Korea Entertainment Industry Association*, Vol.15, No.4, pp.279-286, 2021. DOI: https://doi.org/10.21184/jkeia.2021.6.15.4.279

[5] J. E. Lee, "Post-Examination Analysis on the Student Dropout Prediction Index", *The Journal of Bigdata*, Vol.4, No.2, pp.175-183, 2019.

[6] H. N. Ocharo, S. Hasegawa, "Using machine learning to classify reviewer comments in research article drafts to enable students to focus on global vision", *Education and Information Technologies*, Vol.34, No.5, pp.2093-2110, 2018. DOI: https://doi.org/10.1007/s10639-018-9705-7

[7] K. F. Hew, Y. Tang, "Understanding student engagement in large-scale open online courses: a machine learning facilitated analysis of students' reflections in 18 highly rated MOOCs.", *International Review of Research in Open and Distributed Learning*, Vol.19, No.3, pp.69-93, 2018. DOI: https://doi.org/10.19173/irrodl.v19i3.3596

[8] T. B. Yoon, Y. M. Jung, J. H. Lee, H. J. Cha, S. H. Park, Y. S. Kim, "Outlier Analysis of Learner's Learning Behaviors Data using k-NN Method." *Proceedings of HCI KOREA*, pp.524-529. 2007.

[9] K. P. Sinaga, M. S. Yang, "Unsupervised K-means clustering algorithm." *IEEE access*, Vol.8, pp.80716-80727, 2020. DOI: https://doi.org/10.1109/ACCESS.2020.2988796

[10] T. M. Kodinariya, P. R. Makwana, "Review on determining number of Cluster in K-Means Clustering." *International Journal*, Vol.1, No.6, pp.90-95, 2013.

[11] C. Cortes, V. Vapnik, "Support-vector networks." *Machine learning*, Vol.20, No.3, pp.273-297, 1995. DOI: https://doi.org/10.1007/BF00994018

[12] A. Natekin, A. Knoll, "Gradient boosting machines, a tutorial", *Frontiers in neurorobotics*, Vol.7, No.21, 2013. DOI: https://doi.org/10.3389/fnbot.2013.00021

[13] M. Nalluri, M. Pentela, N. R. Eluri, "A Scalable Tree Boosting System: XG Boost", *International Journal of Research Studies in Science, Engineering and Technology*, Vol.7, No.12, pp.36-51. 2020. DOI: https://doi.org/10.48550/arXiv.1603.02754

[14] S. J. Rigatti, "Random forest." *Journal of Insurance Medicine*, Vol.47, No.1, pp.31-39, 2017. DOI: https://doi.org/10.17849/insm-47-01-31-39.1

[15] R. Andonie, "Hyperparameter optimization in learning systems." *Journal of Membrane Computing*, Vol.1, No.4, pp.279-291, 2019. DOI: https://doi.org/10.1007/s41965-019-00023-0

김 민 영(Minyoung Kim) [준회원]

- 2023년 2월 : 영남대학교 정보통신공학과 (공학사)
- 2023년 3월 ~ 현재 : 경북대학교 데이터사이언스학과 석사과정

〈관심분야〉
자연어 처리, 그래프 데이터 분석

윤 선 영(Sun-Young Yoon) [준회원]

- 2022년 2월 : 인제대학교 통계학과 (통계학사)
- 2023년 3월 ~ 현재 : 경북대학교 데이터사이언스학과 석사과정

〈관심분야〉
텍스트 마이닝, 인공지능 응용

김 수 현(Suhyeon Kim) [정회원]

- 2020년 2월 : 울산과학기술원 융합경영대학원 비즈니스분석 (이학석사)
- 2023년 2월 : 울산과학기술원 산업공학과 (공학박사)
- 2023년 3월 ~ 2023년 8월 : 서울대학교 공학연구원 박사후연구원
- 2023년 9월 ~ 현재 : 경북대학교 데이터사이언스대학원 조교수

〈관심분야〉
데이터 마이닝, 인공지능 응용