The Journal of the Convergence on Culture Technology (JCCT)
Vol. 10, No. 4, pp.595-600, July 31, 2024. pISSN 2384-0358, eISSN 2384-0366

http://dx.doi.org/10.17703/JCCT.2024.10.4.595

JCCT 2024-7-69

Markov Chain을 응용한 학습 성과 예측 방법 개선

Improving learning outcome prediction method by applying Markov Chain

황철현*

Chul-Hyun Hwang*

요 약 학습 성과를 예측하거나 학습 경로를 최적화하는 연구 분야에서 기계학습과 같은 인공지능 기술의 사용이 점차 증가하면서 교육 분야의 인공지능 활용은 점차 많은 진전을 보이고 있다. 이러한 연구는 점차 심층학습과 강화학습과 같은 좀 더 고도화된 인공지능 방법으로 진화하고 있다. 본 연구는 학습자의 과거 학습 성과-이력 데이터를 기반으로 미래의 학습 성과를 예측하는 방법을 개선하는 것이다. 따라서 예측 성능을 높이기 위해 Markov Chain 방법을 응용한 조건부 확률을 제안한다. 이 방법은 기계학습에 의한 분류 예측에 추가하여 학습자가 학습 이력 데이터를 분류 예측에 추가함으로써 분류기의 예측 성능을 향상 시키기 위해 사용된다. 제안 방법의 효과를 확인하기 위해서 실증 데이터인 '교구 기반의 유아 교육 학습 성과 데이터'를 활용하여 기존의 분류 알고리즘과 제안 방법에 의한 분류 성능 지표를 비교하는 실험을 수행하였다. 실험 결과, 분류 알고리즘만 단독 사용한 사례보다 제안 방법에 의한 사례에서 더 높은 성능 지표를 산출한다는 것을 확인할 수 있었다.

주요어 : 기계학습, 마코브체인, 분류예측, 학습성과예측, 학습관리시스템

Abstract As the use of artificial intelligence technologies such as machine learning increases in research fields that predict learning outcomes or optimize learning pathways, the use of artificial intelligence in education is gradually making progress. This research is gradually evolving into more advanced artificial intelligence methods such as deep learning and reinforcement learning. This study aims to improve the method of predicting future learning performance based on the learner's past learning performance-history data. Therefore, to improve prediction performance, we propose conditional probability applying the Markov Chain method. This method is used to improve the prediction performance of the classifier by allowing the learner to add learning history data to the classification prediction in addition to classification prediction by machine learning. In order to confirm the effectiveness of the proposed method, a total of more than 30 experiments were conducted per algorithm and indicator using empirical data, 'Teaching aid-based early childhood education learning performance data'. As a result of the experiment, higher performance indicators were confirmed in cases using the proposed method than in cases where only the classification algorithm was used in all cases.

Key words : Machine Learning, Markov Chain, Classification Prediction, Learning Performance Prediction, Learning Management System

*정회원, 한양여자대학 빅데이터과 조교수 (단독저자)
접수일: 2024년 4월 20일, 수정완료일: 2024년 5월 20일
게재확정일: 2024년 6월 10일

Received: Apil 20, 2024 / Revised: May 20, 2024
Accepted: June 10, 2024
*Corresponding Author: chhwang@hywoman.ac.kr
Dept. of BigData, HanYang Woman Univ, Korea

Ⅰ. 서 론

교육 분야에서 인공지능 기술을 도입하는 것은 단순하게 특정 기능을 제공하는 차원을 넘어 교육 활동의 질을 높이는 본질적인 수준으로 발전하고 있다. 인공지능은 이제 커리큘럼과 콘텐츠가 학생들의 요구에 맞춰 개인화됨으로써 학습 경험과 교육 품질을 향상시켜 궁극적으로 교육 만족도를 높이는데 기여한다[1].

특히 인공지능 기술이 교육 분야에 많은 분야 가운데에도 '학습성과예측'은 교육과정 설계 및 운영 전체에 미치는 영향이 큰 핵심적인 프로세스로 매우 중요하게 취급되고 있다[2-4]. 학습 성과 예측은 각각의 교육 콘텐츠를 맞춤화·개인화하여 활용도와 유지율을 높이는데 기여하기 때문이다.

학습 성과에 대한 예측 정확도를 개선하는 것은 예측 알고리즘 자체를 개발하거나 예측에 활용되는 데이터를 개선하는 두 가지 방법이 있다.

첫 번째 방법은 예측 알고리즘 자체를 개선하기 위한 노력이다. '회귀'와 '분류' 같은 전통적인 기계학습 방법가운데 예측에 적합한 알고리즘을 '탐색'하거나 '알고리즘 자체'를 연구하는 것이다. 최근에는 심층학습(deep learning)이나 강화학습(reinforcement learning)을 도입하여 예측 정확도를 높이려는 다양한 노력이 시도되고 있다[5-6].

두 번째 방법은 예측 알고리즘에 사용될 데이터를 개선하는 방법이다. 이 방법은 도메인 지식을 활용하여 데이터를 전처리(data pre-processing)하거나 데이터 증강(data augmentation)을 통해 학습을 변화시키는 것이다. 이 방법은 '학습 부진 학생 예측'처럼 원천적으로 데이터 불균형이 발생하였을 때, 소수 분류에 대한 데이터 증강을 통해 성능을 개선하는 방법이다[2].

본 연구는 학습 성과에 대한 예측 정확도를 높이기 위해 학습 이력 데이터의 '확률 분포 값'을 활용하는 방법을 제안한다. 이 방법은 feature 변수에 정보가 충분하지 않아 원하는 예측 성능이 나오지 않을 때, 훈련 데이터 가운데 target 변수(y_test)의 확률 분포를 활용하여 최종 예측값을 산출하는 방법이다.

제안 방법의 효과를 검증하기 위해 '교구를 활용하는 유아 교육 현장의 실증 데이터'를 기반으로 다양한 case의 실험을 수행하였다.

본 논문의 구성은 다음과 같다. 2장에서는 기계학습을 활용한 학생성과 예측과 관련된 기존의 연구 결과를 살펴보고, 3장에서는 본 논문에서 제안한 방법인 Markov Process를 활용한 정확도 개선 방법에 대해 제시한다. 4장에서 제안 방법의 효과를 검증하기 위한 실험 방법과 결과를 제시하고, 5장에서는 결론을 도출한다.

Ⅱ. 관련 연구

1. 학습 분석(Learning Analysis)

학습 분석이란 학습 활동 데이터를 분석함으로써 그동안 알지 못했던 학습에 대한 통찰력을 얻기 위한 것이다[10]. 구체적으로 '학습이 이루어지는 환경과 학습'을 이해하고 최적화된 학습 환경을 제공해 주기 위해 '학습자와 학습 맥락에 대한 데이터를 수집, 측정, 분석하는 일련의 과정'이라고 정의한다[10, 13].

2. 교육 분야의 기계 학습(Machine Learning) 도입

학생의 학습 향상에 기계학습이 적용되는 중요한 방법은 학습자의 요구, 능력 및 역량에 맞춰 커리큘럼과 콘텐츠를 맞춤화하고 개인화하는 것이다. 콘텐츠 맞춤화와 개인화를 위해서는 사용자의 선호 데이터나 학습 이력 데이터를 기반으로 군집·분류·회귀 알고리즘을 활용한다[1, 7].

예를 들면, 4년제 대학에서 재학생의 중도 탈락에 영향을 미치는 결정 요인을 연구하기 위해 랜덤 포레스트(Random Forest)를 활용하였다[8]. 학습자의 학습 경로를 최적화하기 위해 유전알고리즘(Genetic Algorithm) 등을 활용하는 다양한 연구가 시도되었다[9-10].

최근에는 딥러닝을 활용하여 대학 학생들에게 전공과목을 추천하거나, 생성형 AI를 활용하여 맞춤형 교육 시스템을 제안한 연구도 진행되었다[11-13].

3. Markov Chain Process

Markov Chain Process는 20세기 초 러시아 수학자 Andrei Andreivich Markov가 제안한 방법으로 주로 확률변수의 통계값을 예측하기 위해 이용된다. Markov Chain Process에 따라 현재 상태(Xn)에서 미래 상태(Xn+1)로 전이하는 확률(Pij)에 근거하여 미래 상태를 예측하는 방법은 다음 식(1)과 같다.

$$X_{n+1} = P_{ij} \times X_n \quad (1)$$

본 연구에서는 Markov Chain Process에서 사용하는 전이 확률(Transition Probability)을 응용하여 상태가 전이된 이후 학습자의 성공 확률을 계산하는데 조건부 확률(Conditional Probability)을 사용한다.

III. 학습성과예측방법 제안

1. 제안 방법의 개요

본 연구의 제안은 학습 데이터의 종속변수가 가지는 조건부 확률값을 활용하여 최종 예측값을 생성하는 방식으로 학습성과 예측의 정확도를 높이는 방법이다.

2. 제안 방법의 수행 절차

제안 방법은 그림 1과 같은 세 단계로 수행된다.

첫째, 기존 접근 방법대로 분류 알고리즘을 이용하여 학습 성과를 예측하기 위한 기계학습을 수행하고, 성공 여부와 여부에 대한 각각의 확률값을 산출한다.

둘째, 직전 학습 이력 정보를 활용하여 t-1의 상태와 학습 성공 확률값을 구한다. 이 값은 상태 t-1에서 상태 t로 상태가 전이될 때 학습이 성공할 조건부 확률을 나타낸다. 이 값은 Markov Chain의 전이 확률(state transition probability)을 응용한 것이다.

셋째, 기계학습 과정에서 산출된 성공 확률값에 보정 값을 반영한 후 조건부 확률값을 곱하여 최종 예측 확률값을 산출한다. 보정 값은 예측 확률(predicted probability)이 가지는 값의 특징을 보정해주는 역할을 담당한다. 산출값이 0.5 이상일 경우 '성공'으로 분류한다.

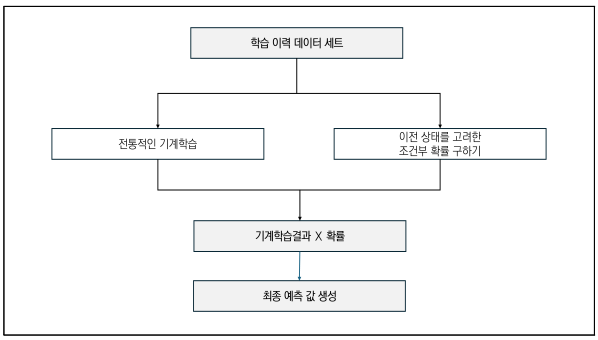


그림 1. 제안 방법의 개념적 절차
Figure 1. Conceptual procedure of the proposed method

1) 기계학습에 의한 분류 예측 수행

제안 방법의 첫 번째 단계인 분류 예측은 학습자의 학습성과를 예측하기 위해 분류 알고리즘을 적용하는 과정이다. 이 단계에서는 학습자의 이전 학습 이력과 학습 프로그램 정보를 활용하여 학습의 성공 여부를 예측한다. 이를 위해 기존 기계학습에서 사용되는 'XGB classification', 'Random Forest'와 같은 기계학습 알고리즘을 사용하여 모델을 구축한다.

2) 직전 상태를 고려한 조건부 확률 구하기

제안 방법의 두 번째 단계는 예측하고자 하는 학습의 직전 프로그램 정보를 활용하여 성공 확률을 구하는 과정이다. 보정계수 k는 다음 식(2)와 같이 전체 학습 이력 데이터 수에서 성공 데이터 수의 비율을 말한다.

$$k = \frac{count(All)}{count(Suc)} \qquad (2)$$

- count(Suc) : 학습에 성공한 이력 수
- count(All) : 학습에 참여한 전체 이력 수

다음은 t-1에서 프로그램 P_a를 이수하였고, t 시점에서 P_b 프로그램의 성공 확률을 산출하는 과정이다. 다음 식 (3)은 직전 상태를 고려한 성공 확률을 구하는 방법을 설명하고 있다.

$$\Pr_{ab} = P(A_{t-1} \cap B_t) \qquad (3)$$

- $\Pr_{ab}$: i번째 학습생의 보정된 성공 확률

3) 최종 분류 예측값 산출

학습 프로그램에 대한 최종 분류 예측값을 산출하기 위해 다음 식(4)와 같은 절차를 수행한다.

- 절차1에서 산출한 확률값이 최대 1이 될 수 있도록 보정계수 k를 곱해준다. p(n)은 전체 교육 이력 대비 성공 이력의 비율이다.
- 위 결과에 절차 2에서 산출한 조건부 확률을 반영하여 최종 발생확률을 산출한다.
- 임계치를 0.5로 하여 최종 발생확률의 값이 0.5 이상일 경우 '성공'이라 분류하고 0.5 미만일 경우 '실패'로 분류한다.

$$Pc_{ab} = f(\Pr_{ab} \times k) \qquad (4)$$

$$f = \begin{cases} 1 \text{ if } x > 0.5 \\ 0 \; else \end{cases}$$

- Pc_{ab} : 최종 산출된 분류 예측 결과

Ⅳ. 실험 및 결과

제안 방법의 효과를 검증하기 위해 교구를 활용한 학습자들의 학습 이력을 기록한 실증 데이터를 사용하여 실험하였다. 실험은 기존의 기계학습 알고리즘을 활용한 분류 예측의 정확도와 제안 방법의 분류 예측 정확도를 비교하는 방법으로 진행하였다.

1. 실험 데이터 세트

본 연구에서는 학습자들이 유아 교육용 교구를 활용하여 학습한 결과를 기록한 826개의 실증 데이터를 활용하였다. 이 데이터는 학습자와 프로그램 식별자, 교육 시간, 시도·성공 횟수, 최종 성공 여부 판단 등의 데이터를 포함하고 있다. 실험에 사용된 데이터의 구성은 다음 표 1과 같다.

표 1. 실험 데이터의 구조
Table 1. Structure of experimental data

데이터 항목		유형
L	교육 프로그램 식별자	Char(10)
2	학습자 식별자	Char(12)
3	교육 프로그램 분류	Char(1)
4	교육 프로그램 수준 분류	Number(1)
5	교육 시작 시간	DateTime
6	교육 종료 시간	DateTime
7	오류(실패) 횟수	Number(3)
8	교육 프로그램 성공 유무	Char(1)

2. 기계학습용 데이터 생성

확보된 데이터를 기계학습에 활용하기 위해서 표2와 같이 파생 데이터의 생성, one-hot encoding, scaling 절차를 거쳐 최종 기계학습용 데이터를 생성한다.

먼저 실험에서 학습 이력 데이터의 특성을 더 잘 설명하고, 궁극적으로 기계학습의 성능을 향상하고자 다음 표2와 같은 파생 변수를 생성하여 활용한다.

각 파생 변수는 예측 대상이 되는 학습 프로그램과 동일 교육 분류(program category) 내의 이력과 전체 교육 분류 이력을 대상으로 생성한다.

예를 들면 '총 학습 시간'의 경우 예측 대상이 되는 학습 프로그램 분류와 동일한 분류 내의 학습 이력만을 대상으로 '총 학습 시간'을 산출하고, 동시에 전체 학습 이력 데이터를 대상으로 모두 2건의 '총 학습 시간'을 산출한다.

표 2. 파생변수 생성
Table 2. Create derived variables

데이터 항목		유형
1	교육(학습) 횟수	Number(5)
2	총 학습 시간	Number(5)
3	회당 평균 학습 시간	DateTime
4	총 실패 횟수	Number(5)
5	회당 평균 실패 횟수	Number(5.2)
6	교육 프로그램 총 성공 횟수	Number(5)
7	교육 프로그램 성공 확률	Number(5.2)

원-핫 인코딩(One-hot Encoding)은 범주형 변수인 '교육 프로그램 분류' 1개 항목을 대상으로 수행하였고, Scaler는 수치형 데이터에 대해 MinMaxScaler를 적용하였다. 또한 기계학습을 위해 random split을 수행하여 660개의 훈련 데이터 세트와 166개의 시험 데이터 세트를 분리하여 기계학습용 데이터 세트를 구성하였다.

3. 실험 계획 및 결과

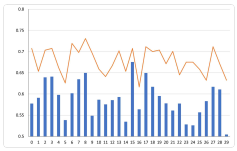
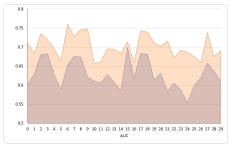
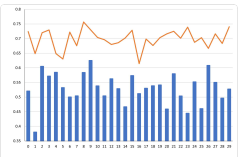
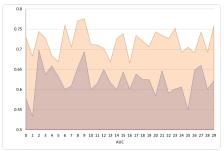
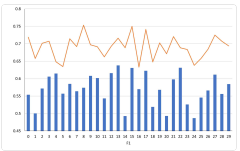
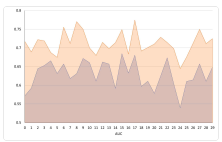
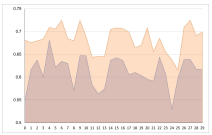
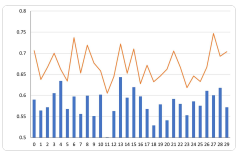
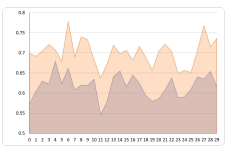
실험에 사용된 분류 알고리즘은 XGBoost(eXtreme Gradient Boosting), Random Forest, KNN, LGBM, Logistic Regressor 등 총 5개의 알고리즘을 사용하여 학습 예측 모델을 구축하였다. 분류 예측 성능 평가를 위한 지표로 이진 분류(Binary Classification)의 성능 측정 지표인 F1 Score와 AUC(Area Under the Curve)를 사용하여 예측 성능을 상호 비교하였고, 실험의 객관성을 확보하기 위해 데이터를 random하게 추출하여 알고리즘당 30회의 실험을 수행하였다.

이와 같은 다수의 실험 case에서 제안된 방법은 동일한 알고리즘을 사용하는 기존의 기계학습 방법에 비해 F-Score와 AUC가 일관되게 상승했음을 확인할 수 있었다. 이를 통해 직전 교육 프로그램의 성과를 조건부 확률로 기계학습에 반영하면 성능 지표 상승에 도움을 주는 것을 확인할 수 있다.

다음 표 3은 앞서 제시한 실험 결과를 도식화하여

제시한 자료이다. 각 알고리즘 별로 제안 방법의 효과를 측정하기 위해 데이터 수와 교육 프로그램 유형을 고려하여 총 30회의 실험을 수행하였다. 실험 결과에서 F1 Score와 AUC의 두 가지 성능 지표 항목에서 기존의 기계학습 방법에 비해 성능이 향상된 결과를 볼 수 있다.

표 3. 실험 결과 시각화
Table 3. Summary of experiment results

알고리즘	F1 Score	AUC
KNN		
Logi-stic Reg		
Ran-dom Forest		
XGB		
Light GBM		

다음 표 4는 앞서 제시한 실험 방법의 결과를 대푯값으로 요약하여 제시하였다. Random하게 추출된 30개의 데이터 세트를 활용하여 실험 결과로 추출된 각 지표의 최대, 최소, 평균 값을 제시하고, 기존 방법과 비교하였다.

표 4. 실험 결과 요약
Table 4. Summary of experiment results

알고리즘	성과 지표	접근 방법	Min	Max	Mean	GT. As-Is
KNN	F1	AsIs	0.50	0.67	0.58	30/30
		ToBe	0.61	0.73	0.67	
	AUC	AsIs	0.55	0.69	0.62	30/30
		ToBe	0.65	0.76	0.70	
Logi-stic Reg	F1	AsIs	0.38	0.62	0.52	30/30
		ToBe	0.61	0.75	0.69	
	AUC	AsIs	0.53	0.69	0.62	30/30
		ToBe	0.66	0.77	0.71	
Ran-dom Forest	F1	AsIs	0.48	0.63	0.57	30/30
		ToBe	0.63	0.75	0.69	
	AUC	AsIs	0.53	0.68	0.63	30/30
		ToBe	0.64	0.77	0.71	
XGB	F1	AsIs	0.52	0.63	0.57	30/30
		ToBe	0.58	0.70	0.65	
	AUC	AsIs	0.53	0.68	0.61	30/30
		ToBe	0.61	0.72	0.68	
Light GBM	F1	AsIs	0.48	0.64	0.58	30/30
		ToBe	0.60	0.74	0.67	
	AUC	AsIs	0.54	0.67	0.61	30/30
		ToBe	0.63	0.77	0.70	

V. 결 론

본 연구에서는 교구를 활용한 유아 교육 환경에서 학습 결과를 예측하기 위한 정확도 향상 방법을 제안하였다. 제안 방법은 기존의 분류 예측결과에 직전에 수행된 선행 학습 결과를 조건부 확률로 반영하는 방법이다. 제안 방법의 효과를 검증하기 위해 교육 현장의 실증 데이터를 수집하여 5개의 기계학습 알고리즘 별로 30회의 실험을 수행하였다.

실험 결과 제안 방법은 모든 실험 Case에서 기존의 기계학습 알고리즘만을 적용하는 전통적인 방법에 비해 높은 성능 지표를 보이는 것을 확인할 수 있다.

본 연구는 학습성과예측 플랫폼을 위한 핵심 프로세

스를 제안하였으며 실증 데이터를 활용한 실험 결과에서 제안 방법이 성능 향상에 기여할 수 있다는 것을 확인 하였다.

References

[1] L. Chen, P. Chen and Z. Lin, "Artificial Intelligence in Education: A Review," *in IEEE Access*, vol. 8, pp. 75264-75278, 2020, doi: 10.1109/ACCESS.2020.2988510.

[2] Chul-Hyun Hwang, "Improvement of early prediction performance of under-performing students using anomaly data," *Journal of the Korea Institute of Information and Communication Engineering(JKIICE)*, Vol. 26, No. 11, pp. 1608-1614, Dec 2022, DOI : 10.6109/jkiice.2022.26.11.1608

[3] H. Lakkaraju, E. Aguiar, C. Shan, D. Miller, N. Bhanpuri, R. Ghani, and K. L. Addison, "A Machine Learning Framework to Identify Students at Risk of Adverse Academic Outcome," *in Proceedings of the 21st ACM SIGKDD, International Conference on Knowledge Discovery and Data, Sydney, Australia*, pp. 1909-1918, 2015.

[4] B. Albreiki, N. Zaki, and H. Alashwal, "A Systematic Literature Review of Student' Performance Prediction Using Machine Learning Techniques," *Education Science*, vol. 11, no. 9, pp. 1-27, Sep. 2020.

[5] E. Alyahyan and D. Dustegor, "Predicting academic success in higher education: Literature review and best practices," *International Journal of Educational Technology in Higher Education,* vol. 17, no. 3, Feb. 2020.

[6] W. Xing and D. Du, "Dropout Prediction in MOOCs: Using Deep Learning for Personalized Intervention", *Journal of Educational Computing Research*, vol. 57, no. 3, pp. 547-570, Mar. 2019.

[7] T. A. Mikropoulos and A. Natsis, "Educational virtual environments: A ten-year review of empirical research (1999 - 2009)", *Comput. Edu.*, vol. 56, no. 3, pp. 769-780, Apr. 2011.

[8] Chunhong Liu, Haoyang Zhang, Jieyu Zhang, Zhengling Zhang, Peiyan Yuan, "Design of a Learning Path Recommendation System Based on a Knowledge Graph", *International Journal of Information and Communication Technology Education (IJICTE), Vol. 19, Issue. 1, pp. 1-18, Dec 2023,* , DOI : 10.4018/IJICTE.319962

[8] Eunjung Lee, Youngsoo Song, Jiha Kim, Suhyun 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, , DOI: 10.17232/KSET.36.1.191

[9] Chunhong Liu, Haoyang Zhang, Jieyu Zhang, Zhengling Zhang, Peiyan Yuan, "Design of a Learning Path Recommendation System Based on a Knowledge Graph", *International Journal of Information and Communication Technology Education (IJICTE)*, Vol. 19, Issue. 1, pp. 1-18, Dec 2023, , DOI : 10.4018/IJICTE.319962

[10]Yeon-Hee Kim, Soo-Jin Lim, "A Study on the Prediction of Learning Results Using Machine Learning", *Jounal of Educational Technology*, vol 36, No 1, pp. 191-219, July 2020

[11]Lee Jae Kyu , PARK HEESUNG , Wooju Kim, "Major Class Recommendation System based on Deep learning using Network Analysis", *Journal of Intelligence and Information Systems(JIIS)*, Vol. 27, No. 1, pp. 95-112, Dec 2023

[12]Oakyoung Han, "A Study on Components for Designing Personalized Education Systems Based on Generative AI", *The Journal of Korean association of computer education,* vol 26, No 6, pp. 127-141, Oct 2023

[13]Hyeon-Seong Kim, Jin-Seok Kim, "A Study on Regional-customized education program selection model using big data analysis", *The Journal of the Convergence on Culture Technology (JCCT)*, Vol. 9 No. 2, Mar, 2023

[14]Siemens. G, Long. P, "Penerrating the foganalytics in learning and educations." *Education Review,* Vol.46, No.5, pp.30-32, 2011.