The Journal of the Convergence on Culture Technology (JCCT)
Vol. 8, No. 6, pp.487-492, November 30, 2022. pISSN 2384-0358, eISSN 2384-0366

http://dx.doi.org/10.17703/JCCT.2022.8.6.487

JCCT 2022-11-60

랜덤 포레스트를 활용한 만족도 사전조사에 따른 교육 역량 예측 분석

An Analysis of Educational Capacity Prediction according to Pre-survey of Satisfaction using Random Forest

남기훈*

Kihun Nam*

요 약 대학들은 급변하는 사회 환경에 적합한 교육역량 수준을 높이기 위해 다양한 방법들을 찾고 있다. 본 논문에서는 조사 항목을 수정, 보완한 만족도 사전조사를 개강 전에 실행하여 학업성취도를 높이고 전공 이탈자의 비율을 낮춰 교육 성과를 높이는 방안을 제안한다. 일반적인 만족도 조사 이후에 시행되는 교육품질 개선(CQI) 방식을 보완하고자 만족도 사전조사를 시행하였다. 학생역량을 강화하기 위해 설계가 진행 중인 인공지능형 메디치 플랫폼에 적용할 수 있는 머신러닝 기법의 랜덤 포레스트를 활용하여 중요한 데이터의 예측 및 분석을 가능하게 하였다. 만족도 사전조사 데이터들을 전처리하여 수강 신청 학생들의 정보를 설명 변수로 정의하고 분류하여 모델 생성 및 학습하였다. 실험 환경은 주피터 노트북 3.7.7, Python 3.7에서 관련 알고리즘과 사이킷런(sklearn) 라이브러리를 함께 사용하였다. 제안하는 방안의 결과를 수업에 반영하여 수업 후에 진행하는 교육 만족도 조사의 변화와 중도 탈락생 수의 동향을 비교 분석하였다.

주요어 : 만족도 사전조사, 교육품질 개선, 메디치 플랫폼, 머신러닝, 랜덤 포레스트

Abstract Universities are looking for various methods to enhance educational competence level suitable for the rapidly changing social environment. This study suggests a method to promote academic and educational achievements by reducing drop-out rate from their majors through implementation of pre-survey of satisfaction that revised and complemented survey items. To supplement the CQI method implemented after a general satisfaction survey, a pre-survey of satisfaction was carried out. To consolidate students' competences, this study made prediction and analysis of data with more importance possible using the Random Forest of the machine learning technique that can be applied to AI Medici platform, whose design is underway. By pre-processing the pre-survey of satisfaction, the students information enrolled in classes were defined as an explanatory variable, and they were classified, and a model was created and learning was conducted. For the experimental environment, the algorithms and sklearn library related in Jupyter notebook 3.7.7, Python 3.7 were used together. This study carried out a comparative analysis of change in educational satisfaction survey, carried out after classes, and trends in the drop-out students by reflecting the results of the suggested method in the classes.

Key words : Pre-survey Satisfaction, CQI(Continuous Quality Improvement), Medici Platform, Machine Learning, Random Forest

*정회원, 서경대학교 컴퓨터공학과 (단독저자)
접수일: 2022년 10월 31일, 수정완료일: 2022년 11월 6일
게재확정일: 2022년 11월 9일

Received: October 31, 2022 / Revised: November 6, 2022
Accepted: November 9, 2022
*Corresponding Author: namkh@skuniv.ac.kr
Dept. Computer Engineering, SeoKyeong Univ, Korea

Ⅰ. 서 론

최근 대다수의 교육 기관들은 산업체 요구역량에 맞는 인재 양성을 목표로 다양한 교육 편재 및 새로운 프로그램 발굴에 노력하고 있다[1]. 그러한 노력에도 불구하고 산업체에서는 전문 인재 부족을 호소하고 있는 실정이다[2]. 이는 교육제도와 학생들 사이에 문제가 있음을 보이고 있다. 다양한 원인 중 전공 수업을 학생들이 제대로 소화해내지 못하고 있다는 점이 가장 큰 요인이라 말할 수 있다. 멘토링, 컨설팅, 카운슬링 등을 적용하여 학생들과 소통을 통한 문제점 인식 및 해결점을 찾으려는 방안들이 많지만 기대 이상 큰 실적은 나오지 않고 있다[3][4]. 요즘 학생들이 아는 것에 대한 소신 발언을 주저하지 않고 주장하기도 하지만 모르거나 불확실한 상황에 대한 자신의 속내를 밝히기를 꺼려하는 측면도 많기 때문이다[5]. 익명성 보장 및 비접촉 상황에서의 만족도 조사에서는 다른 양상을 보인다. 초등학교부터 만족도 조사에 익숙해서인지 냉철하게 판단하고 신중하게 설문에 응답하는 경향을 나타낸다[6]. 그러한 특이점을 감안하여 만족도 조사의 내용을 세분화하고 시행단계를 개선하여 학생들의 수업에 대한 다양한 소견을 정확하게 파악하고 분석하여 교육의 질을 높이는 방안을 제안하려 한다.

본 논문에서는 수업 만족도 사전조사를 랜덤 포레스트를 활용하여 데이터를 학습하고 예측하여 수업 후 만족도 조사와 비교 분석하여 주요 변수의 결과에 따라 교육의 질이 향상되었는가를 보여주려 한다. 2장에서 관련 선행 연구, 3장에서는 연구 방법을 설명하고 4장과 5장은 연구 결과 및 결론 순으로 구성된다.

Ⅱ. 관련 선행 연구

일반적인 만족도 지수는 만족도를 구성하는 요인들의 만족도와 중요도를 가중 평균하여 산출하는 경우가 대부분이다. CSI(Customer satisfaction Index) 또는 IPA(Important-Performance Analysis)를 이용하여 만족도를 구한다. 교육 기관마다 자체적인 설문 문항을 선정하고 5점 또는 7점 척도를 사용하여 교육 만족도를 구하며 그 결과 CQI로 지속적인 교육과정 등을 개선할 수 있도록 주기적인 평가와 개선사항을 도출하는 과정을 수행한다[7]. 하지만 CQI는 수업의 특성상 보통 6개월 후에나 반영되어야 하기에 CQI 사항들을 기억하고 수행하기에 적절한 대응이 쉽지 않다는 문제가 있다[8]. 또 다른 문제로는 강의를 듣는 학생들의 수준을 고려해야 한다. 같은 내용의 수업이라 할지라도 해마다 수강하는 학생들의 교육 자세 및 기본 지식의 깊이에 따라 관심도 및 이해력에 있어서 교육 결과의 차이가 생기고 있기에 수업 전 정확한 수강생들의 상태를 파악하는 것이 중요하다고 판단된다.

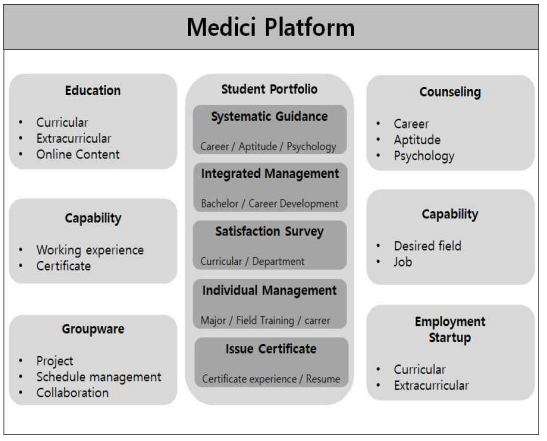

그림 1. 메디치 플랫폼 구성요소
Figure 1. Medici Platform components

그림1은 대학에서 자체 개발 중인 인공지능형 메디치 플랫폼이다. AI(Artificial Intelligence) 기반 학생역량 관리 프로그램으로 학생 취업 역량 강화 및 교육역량 관리를 위한 목적으로 구축되어있다. 지속적 진로상담 및 멘토링과 만족도 조사를 통해 급변하는 사회 환경에서 기업이 요구하는 역량을 개발하기 위해서는 학생의 역량과 산업체 요구역량을 실시간으로 확인하고 모니터링할 수 있는 플랫폼이 필요했기 때문이다. 본 논문에서 교육역량 강화를 위한 만족도 사전조사를 인공지능화하여 역량분석 및 제안 솔루션을 제공하려 한다.

머신러닝은 지도 학습(Supervised Learning), 비지도 학습(Unsupervised Learning), 강화 학습(Reinforcement Learning)로 구분된다[9]. 랜덤 포레스트는 지도 학습에 속한다. 분류 및 회귀 작업에 사용될 수 있으면서 비선형 특성을 고려하면 다양한 데이터를 얻어내는 장점을 가지고 있다[10]. 사전에 프로그램되어 있지 않은 데이터로부터 패턴을 학습하는 일반화 수행 중 예측 결과를 도출하는 과정과 결과 데이터의 해석이 어려울 때도 있다[11].

표 1. 요인별 사전 설문 문항(이공대)
Table 1. Pre-questionnaire by Factor(Institute of Technology)

요인	사전 설문 문항		척도
이론	IT1	대면, 비대면, 혼용(대면/비대면), 실시간, 비실시간 중 방식은?	5점 리커트 척도 (1점: 전혀 그렇지 않다~ 5점: 매우 그렇다)
	IT2	수강을 신청한 이유나 동기를 선택하시오.	
	IT3	귀하의 이론적 배경지식이 있나요?	
	IT4	배경지식이 있다면 어느 정도 인지를 선택하세요.	
	IT5	강의, 토론, 발표, 강의+토론, 강의+발표 중 선택하세요.	
	IT6	귀하가 원하는 과제 횟수는 몇 회 정도인가요?	
	IT7	귀하가 원하는 퀴즈 횟수는 몇 회 정도인가요?	
	IT8	교재 내용 중 호기심 및 관심이 있는 장을 선택하세요.	
	IT9	본 수업에 귀하는 어느 정도 노력하실 계획입니까?	
	IT10	귀하가 희망하는 성적 처리 반영 비율을 선택하시오.	
이론 및 실기	IT11	귀하가 다룰 수 있는 프로그램을 선택하세요.	
	IT12	선택한 프로그램 수준은 어느 정도입니까?	
	IT13	이론과 실기 비중을 몇 %로 희망하는지를 선택하세요.	
	IT14	귀하가 새롭게 배우고 싶은 프로그램은 무엇입니까?.	
	~	이하 생략	
실기 및 산업체	IT21	귀하는 산학활동 경험이 있습니까?	
	IT22	귀하는 산업 전문 기술을 어는 정도 습득하고 있습니까?	
	IT23	교수님과 산업체 초빙 강사님의 수업 비율을 선택하세요.	
	IT24	귀하는 산학협력업체에 취업을 희망하십니까?	
	~	이하 생략	

III. 연구 방법

강의 시작 전 수강 신청(장보기)때에 만족도 사전조사를 실시하여 수업에 관한 여러 데이터를 분석 판단한다. 수강생들의 요구 사항들을 최대한 수업에 반영되도록 만족도 사전조사 설문 양식을 새롭게 구성하였다[12][13][14].

표1은 사전 조사용 설문 문항들이며 내용을 축소 표기한 것이다. 사전 설문 문항은 5점 리커트 척도로 측정하였다.

표 2. 교육 분야의 범주의 설명 변수
Table 2. Explanatory variables for categories in education

범주	설명 변수
수강생 기본정보 (7개)	단과대학, 학과, 학번, 학년, 성적, 나이, 성별
전공 수업 (76개)	사전 설문 문항(30), 전공기초(8), 전공필수(38), 전공 자격증(7), 부전공
교양 수업 (47개)	교양필수(6), 교양 및 자유 선택(41)
비교과 (13개)	상담, 진로탐색, 취업, 창업, 자격증(8)

조사에 참여한 이공대학 A학과의 교과목 구성은 이론 과목들이 6개, 이론 및 실기는 34개, 실기 및 산업체는 6개로 전공 교과목이 구성되어있다. 조사에 참여한 학생은 1학년 80명, 2학년 75명, 3학년 84명, 4학년 57명, 총 296명이다. 표1의 변수만을 가지고는 학습효과가 미흡하기에 사전 설문 문항들을 표2에서 제시하는 설명변수에 포함하여 함께 처리하였다. 143개의 설명 변수를 이용하여 수업에 영향을 미치는 중요 변수를 예측하기 위해 랜덤 포레스트를 활용하였다.

데이터 전처리를 통하여 불필요한 변수 제거 및 학습에 사용할 Training 데이터와 최종적으로 예측에 활용될 Testing 데이터로 나눈다.

Algorithm 1 Random forest algorithm for classification

1. For $b = 1$ to B:
   a. Draw a bootstrap sample Z^* of size N from the training data.
   b. Grow a random-forest tree $T_b(\Theta_b)$ with the bootstrapped data, by recursively repeating the following steps for each terminal node of the tree, until the minimum node size is reached.
      i. Select m variables at random from the p variables.
      ii. Determine the best variable and split-point among the m variables using the Gini index.
      iii. Split the node into two daughter nodes.
2. Output the ensemble of trees $\{T_b(\Theta_b)_1^B\}$.

To classify a new feature vector f_n:

Classification: Let $\hat{C}_b(f_n, \Theta_b)$ be the class prediction of the bth random-forest tree.
Then $\hat{C}_{rf}^B(f_n) = \text{majority vote}\{C_b(f_n, \Theta_b)\}_1^B$.

그림 2. 랜덤 포레스트 알고리즘
Figure 2. Random Forest algorithm

랜덤 포레스트 알고리즘은 부스트랩 사이즈 T만큼의 Training 데이터를 만들어내고, b = 1부터 B까지가 부스트랩을 만든다. 앙상블(Ensemble)은 각각의 leaner의 개수이며, 의사결정나무는 p개의 변수 중에서 m개의 변수만을 선택$(m < p)$하여 생성한다. Bagging 복원추출을 통해서 Original 데이터의 숫자만큼을 샘플링하여 Training 데이터 집합과 OOB(Out-of-Bag) 데이터 집합을 만든다[15].

변수의 중요도 결정은 Original 데이터 집합에 대해서 OOB Error(e_i)를 구하고 특정 변수의 값을 임의로 뒤섞은 데이터 집합에 대한 OOB Error(p_i)를 구해 두 개의 $p_i - e_i$ 평균과 분산을 고려하여 차이가 최소가 되는 변수를 선택한다.

표 3. 최적의 하이퍼 파라미터
Table 3. Optimal Hyper-Parameter Set

파라미터 명	설정값(Value)
n_estimators (트리의 개수)	42
min_samples_splt	14
min_samples_leaf	6
max_features	None
max_depth	10

표3의 최적화된 하이퍼 파라미터 셋으로 다시 모델을 학습하여 Testing 데이터에서 예측 정확성을 측정하였다.

표 4. 상위 10개 변수 중요도와 정확성
Table 4. Top 10 variables importance and accuracy

상위 변수 (10개)	중요도 (Importance)	정확성 (Accuracy)
1. BIT001	0.410714	0.958046
2. BIT013	0.325392	
3. BIT024	0.251742	
4. CES003	0.144253	
5. BIT005	0.138536	
6. BIT012	0.121204	
7. CFR012	0.061204	
8. DBI002	0.047812	
9. BIT024	0.034061	
10. DBI003	0.025605	

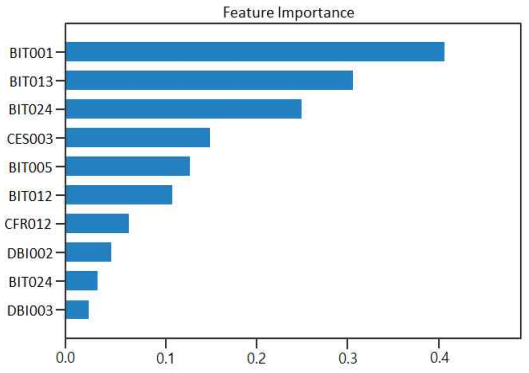

그림 3. 중요 변수 산출
Figure 3. Calculation of importance variables

Ⅳ. 연구 결과

최상위 변수 BIT001은 대면, 비대면, 혼용(대면/비대면), 실시간, 비실시간 중 방식을 묻는 항목이다. 팬데믹으로 인하여 비대면 수업으로 전환되는 과정에서 비대면 수업을 실시간 또는 비실시간 수업 운영이 최대 관심사로 나타났으며 변수의 결과값에 따라 수업 운영 방식을 정하였다. 상위의 변수 항목들을 최대한 수업환경에 적용되게 하였다. 그 결과 수강생들의 성적 확인 시 시행하는 교육 만족도 조사 결과는 그림4와 같으며 사전조사를 시행한 학과 교과목들의 만족도 결과가 조사를 미시행한 학과의 교과목보다 평균적으로 개선되는 것이 확인되었다.

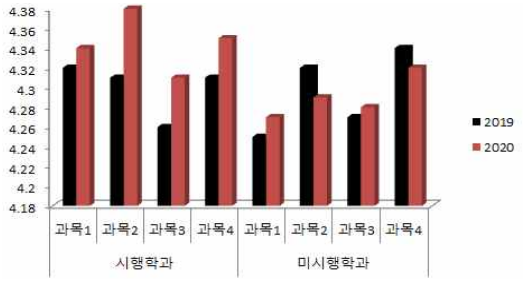

그림 4. 2019~2020년 교육 만족도 조사 결과
Figure 4. Results of Education Satisfaction in 2019~2020

그림5와 그림6은 학교 포탈 BI 시스템에 공개된 데이터 결과이며 연도별 중도 탈락자 수를 그래프로 나타낸 것이다. 중도 탈락자 수는 영역별로 3가지(개인 사유 및 기타, 교육(제적) 및 미적성, 편입학 및 진학) 영역

으로 함축시켰으며 교육(제적) 및 미적성 관한 지표는 실선으로 구분하였다. 2022년 데이터는 완전하게 취합되지 않은 상태이다. 그림5와 그림6의 결과를 통해 교육(제적) 및 미적성 영역의 변화에 따라 중도 탈락자의 수의 상승 및 하락추세를 비교할 수 있었다.

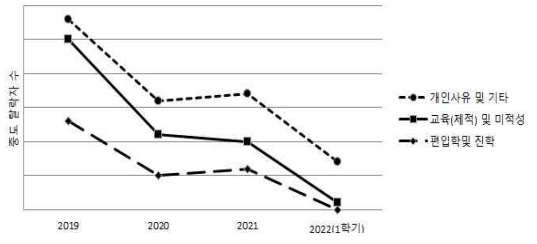

그림 5. 조사를 시행한 학과의 결과
Figure 5. Results of the department that conducted the survey

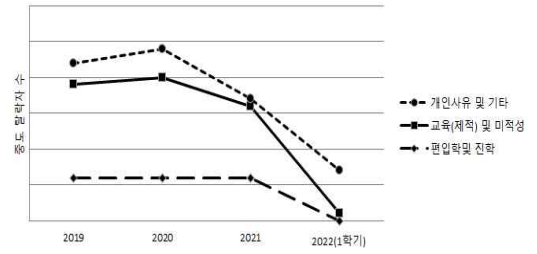

그림 6. 조사를 시행하지 않은 학과의 결과
Figure 6. Results of the department that has not conducted the survey

V. 결 론

본 논문에서 제안한 랜덤 포레스트를 활용한 만족도 사전조사를 실시한 결과 수강생들이 수업에 대한 기대감이 향상되었으며 수업 참여도 역시 높아졌음을 확인하였다. 수업에 영향을 주는 변수들의 미세한 변화를 줌으로써 전공을 포기하지 않고 중도 탈락자 수를 소폭 줄이는 긍정적인 결과를 얻게 되었다. 앞으로 개선되어야 할 사항은 세밀한 분석이 가능한 설문 문항들의 지속적인 발굴과 더 나은 예측 및 평가 알고리즘 개발하는 과제가 남아있다. 그 과정을 마치면 전체 학과에 만족도 사전조사를 확대 적용하려 한다. 시행 시점이 팬데믹으로 인한 비대면 수업으로 진행되었다는 점에서 결과의 신뢰성에 대한 의심의 여지가 있을 수 있기에 차년도를 기점으로 실험을 재운영할 계획이다.

References

[1] Kim, Pil-Seong, “A Poetic Exploration on How to Enhance Various Capacities Based on “OECD Education 2030” in Higher Education”, Journal of Character Education and Research, Vol. 4, No. 2, pp.1-14, 2019, DOI: 10.46227/JCER.4.2.1.
[2] https://it.donga.com/102288/
[3] Kim Sinae, “Qualitative Case Study on College Students’s Experience of Academic Counseling: Focused on Experience with Different Grades”, Journal of Learner-Centered Curriculum and Instruction, Vol. 20, No. 3, pp.549-577, 2020, DOI: 10.22251/jlcci.2020.20.3.549.
[4] Park Hyun-Jeong, Shin Junghwi, “Trend of Academic Resilience and Random Forest Analysis on Factors related to Academic Resilience in Korea”, Asia Journal of Education, Vol. 23, No. 3, pp.501-527, 2022.
[5] Jo So Young, Kyoo-Lak Cho, “A study on the needs of university education service and campus life adjustment for college freshmen”, Journal of Learner-Centered Curriculum and Instruction, Vol. 19, No. 18, pp.1073-1097, 2019, DOI: 10.22251/jlcci.2019.19.18.1073.
[6] Cho, WonKee, Lee, Soojeong, “Relationship among College Admission System, College Students’ College Life and Satisfaction on College Major, and their Academic Achievements”, Journal of Learner-Centered Curriculum and Instruction, Vol. 16, No. 7, pp.673-700, 2016.
[7] Minjung Lee, Kim Soo Dong, “A study on the CQI System for the Quality Management of Competency-Base Education - Focusing on D University in Korea”, Culture and Convergence, Vol. 41, No. 3, pp.35-48, 2019.
[8] Bo-Ram Cho, “A Study on CQI Form for Quality Management and Improvement of Class”, Journal of Digital Convergence, Vol. 18, No. 5, pp.115-125, 2020, DOI: 10.14400/JDC.2020.18.5.115.
[9] Mitchell, T. M. “Evaluating Hypotheses”, Machine Learning, 1997.
[10]Jin Eun Yoo, “Machine Learning for Large-scale/Panel Data and Learning Analytics Data Analysis”, Journal of Educational Technology, Vol. 35, No. 2, pp.313-338, 2019, DOI: 10.17232/KSET.35.2.313.
[11]Oh Miae, Choi Hyeonsu, Kim Suhyun, Chang Joonhyuk, Jin Jaehyeon, Cheon Mikyeong, “A Study on Social security Big Data Analysis and

Prediction Model based on Machine Learning", Korea Institute for Health and Affairs, 2017.
[12]Yoon Eugene, "A Study on Status Analysis and the Development Strategies of University General Education for the 4th Industrial Revolution Era", Korea Journal of General Education, Vol. 14, No. 2, pp.311-325, 2020, DOI: 10.46392/kjks.2020.14.2.311.
[13]Park Joo-Ho, Ryu Kiung, "The Analyses of the Concepts, Contents, Satisfaction and Future Needs for College General Education", Korea Journal of General Education, Vol. 8, No. 2, pp.43-82, 2014.
[14]Shin Soyoung, Kwon Soungyoun, "A Study on the Development and Validity Verification of a Measurement Tool for Educational Satisfaction in University", Journal of Education Science, Vol. 12, pp.107-132, 2013.
[15]Alexandru Gegiuc, Markku Simila, Juha Karvonen, Mikko Lensu, Marko Makynen, Jouni Vainio, "Estimation of degree of sea ice ridging based on dual-polarized C-band SAR data" The Cryosphere, Vol. 12, pp.343-364, 2018, DOI: 10.5194/tc.12.343.2018.

※ This work was supported by Seokyeong University in 2021.