한국빅데이터학회지
제5권 제2호, 2020, pp. 155-170 https://doi.org/10.36498/kbigdt.2020.5.2.155

머신러닝을 활용한 서울시 중학생 진로성숙도 예측 요인 탐색

Exploring the Factors Influencing Students' Career Maturity in Seoul City Middle School: A Machine Learning

박 정†

흥도초등학교

요 약

본 연구의 목적은 서울시 중학생의 진로성숙도 예측 요인을 찾기 위해 머신러닝 기법(Decision Tree, Random Forest, XGBoost)을 서울교육종단연구 4~6차년도 데이터에 적용하였다. 적용에 따라 도출된 세 가지 머신러닝 모형의 변수 중요도와 각 지표별 성능을 확인하였다. 또한 XGBoostExplainer 패키지를 활용하여 모형을 해석하였으며, 데이터 전처리와 분석 모두 R과 R Studio를 활용하였다. 그 결과 각 모형별로 변수 중요도 순위는 다소 차이가 있으나 '성취목표', '창의성', '자아개념', '부모자녀와의 관계', '회복탄력성'이 높은 순위를 보였다. 또한 XGBoostExplainer를 활용하여 패널별 진로성숙도에 정적·부적 영향을 주는 요인을 탐색하였고, '성취목표'가 진로성숙도 예측 최우선 요인임을 찾을 수 있었다. 본 연구결과를 바탕으로 머신러닝 및 변수선택 방법의 비교연구와 서울교육종단연구 코호트별 비교연구가 수행되어야 함을 제언하였다.

■ 중심어 : 중학생, 진로성숙도, 머신러닝, 서울교육종단연구, SELS

Abstract

The purpose of this study was to apply machine learning techniques (Decision Tree, Random Forest, XGBoost) to data from the 4th~6th year of the Seoul Education Longitudinal Study to find the factors predicting the career maturity of middle school students in Seoul city. In order to evaluate the machine learning application result, the performance of the model according to the indicators was checked. In addition, the model was analyzed using the XGBoostExplainer package, and R and R Studio tools were used for this study. As a result, there was a slight difference in the ranking of variable importance by each model, but the rankings were high in 'Achievement goal awareness', 'Creativity', 'Self-concept', 'Relationship with parents and children', and 'Resilience'. In addition, using the XGBoostExplainer package, it was found that the factors that protect and deteriorate career maturity by panel and 'Achievement goal awareness' is the top priority factor for predicting career maturity. Based on the results of this study, it was suggested that a comparative study of machine learning and variable selection methods and a comparative study of each cohort of the Seoul Education Termination Study should be conducted.

■ Keyword : Middle school students, Career maturity, Machine learning, Seoul Education Longitudinal Study, SELS

2020년 11월 18일 접수; 2020년 12월 14일 1차 수정본 접수; 2020년 12월 20일 2차 수정본 접수; 2020년 12월 29일 게재 확정.
† 교신저자 (refree21@gmail.com)

Ⅰ. 서 론

진로발달은 유아기부터 아동기, 청소년기, 성인기, 노년기까지 아우르는 전생애적 발달과정이다[1]. 이중 청소년기는 자신의 미래를 구체적으로 계획하고 실현하기 위해 진로를 탐색하는 중요한 시기로서[2], 특히 우리나라 청소년들은 중학생이 되면서 진로에 대해 보다 실제적인 탐색을 하며 진로발달의 기초를 형성하게 된다[3].

이러한 중학생의 진로발달에서 핵심적인 개념은 진로성숙도이다[4]. 진로성숙도는 진로발달의 결과로, 자신의 특성을 인식하고 직업세계를 탐색하며 이를 바탕으로 합리적인 진로의사결정을 할 수 있는 능력을 말한다[5]. 또한 원만한 진로성숙은 성인이 된 이후 개인의 자아 실현과 삶의 질을 결정하는데 중요한 자원이 되며[6-7], 반대의 경우에는 사회부적응과 진로방황으로 이어질 수 있다[8].

학교는 청소년이 가장 많은 시간을 보내는 주요 생활공간으로 인지・사회・정서적 발달에 매우 큰 영향을 준다[5]. 특히, 중학교 시기는 학생들이 학교에서 머무는 시간이 초등학교에 비하여 많이 늘어난다. 따라서 중학교는 학교 진로교육을 통해 주도적으로 진로탐색・진학활동을 하면서 진로성숙도가 많이 발달하는 시기이다[9].

진로성숙도는 여러 선행연구를 검토해 볼 때, 다차원적 요인의 복합적인 영향을 받으며 발달한다[10-12]. 그리고 다양한 요인들은 해당 개념에 유의미한 영향을 미칠 뿐만 아니라, 요인들 간의 조합을 통해서도 영향을 주고 있으므로 각 요인들의 중요도를 비교하는 것은 쉽지 않다[13]. 또한 진로성숙도에 영향을 미치는 다양한 변인들을 포괄적으로 고려하여 수행된 연구가 부족한 실정이다[11].

이렇게 다양한 변인이 복합적으로 작용하는 개념의 경우는 회귀 분석과 같은 모수적 모형 보다는 머신러닝과 같은 비모수적 모형으로 접근하는 것이 합리적인 대안이 된다[14]. 머신러닝은 특정 가설을 검증하는 것이 아닌, 변수들 사이의 가능한 상호 작용을 최대한 탐색하는 기법으로[15], 기존 사회과학 연구방법에서 어려움을 보인 조사 대상에 대한 많은 수의 예측변수 투입을 극복하기 위한 대안으로 머신러닝이 평가되고 있다[16]. 또한 머신러닝 기법을 활용하면 인간의 심리사회학적 구인에 대한 이해를 보다 높일 수 있다[17]. 따라서 다차원적 요인들이 복합적으로 작용하는 진로성숙도를 머신러닝 방법을 활용할 경우 해당 구인에 대한 이해를 높일 수 있을 것이다.

그러므로 본 연구에서는 1) 머신러닝을 통해 중학생의 학년별 진로성숙도 예측 요인이 무엇인지 파악하고 2) 진로성숙도 예측 모형의 성능을 평가한 뒤 3) 예측 모형을 해석하여, 중등 진로교육의 내용을 명료화하고 학생의 원만한 진로성숙에 기여하고자 한다.

본 논문의 구성은 다음과 같다. 제2장에서는 본 연구에서 다룬 개념과 관련연구를 소개한다. 제3장에서는 분석방법에 대하여 설명하고, 제4장에서는 분석 결과를 정리한 뒤, 제5장에서는 결론을 맺는다.

Ⅱ. 이론적 배경

본 장에서는 진로성숙도와 머신러닝의 개념에 대하여 소개한다.

2.1 진로성숙도

진로성숙도는 청소년들이 자기 주도적으로 진로를 탐색하고 계획하기 위하여 필요한 정의적 태도, 인지적 능력 및 자신의 결정을 실행하는 정도이다[18].

진로성숙도에 영향을 미치는 요인은 연구자의 관점에 따라 다양하게 논의되었는데, 이는 다시

개인, 가정, 학교로 구분할 수 있다[12]. 개인 차원 요인으로는 학습동기, 자아존중감, 자기효능감, 자기결정성이 진로성숙도에 유의미한 영향을 미치는 것으로 나타났다[19-21]. 가정 차원 요인으로는 부모의 양육태도, 부모신뢰관계가 진로성숙도에 영향을 준다는 연구가 있었으며[22-23], 부모의 학력 및 가계소득이 높을수록 진로성숙도가 높다는 연구도 볼 수 있었다[24]. 학교 차원 요인으로는 학생들이 선생님과 긍정적 관계를 형성할수록 진로성숙도가 높게 나타나는 연구와 같이 학교의 역할을 중시하는 연구도 찾을 수 있었다[25].

하지만 이러한 연구들은 대체로 연구자가 관심 있어 하는 요인과 진로성숙도의 관계를 본 연구가 대부분이다[7]. 그렇기 때문에 이렇게 많은 요인들이 복합적으로 작용함에 있어서 어떤 변인을 최우선적으로 고려하여야 학생의 진로성숙도를 예측할 수 있을 것인가에 대한 물음에는 답하기는 쉽지 않다. 따라서 다차원적 요인에 대하여 총체적으로 접근하여 진로성숙도와 그 우선순위를 규명하는 것이 필요한 시점으로 보인다.

2.2 머신러닝 방법

머신러닝에는 다양한 기법이 있으며 본 연구에서 활용한 세 가지 기법을 아래 기술한다.

Decision Tree(이하 DT)는 나무 형태의 시각화된 그래프가 학습 결과로 도출되기 때문에 사회 전 분야에 널리 활용되고 있다. DT는 기본적으로 데이터가 위치한 좌표평면의 재귀적 직교분할을 통해 소집단으로 분류한다. 이러한 분할을 위한 최적분기의 결정은 분할된 영역 안에 있는 데이터의 불순도(Impurity) 지표를 기준으로 하고, 그 값이 가장 낮은 지점을 분기로 선택하며 연쇄적으로 일어난다. 분할에 사용하는 불순도 지표는 데이터와 알고리즘에 따라 엔트로피, 지니계수, 카이제곱통계량 등 다양하다. 본 연구에서 활용한 C5.0 알고리즘은 불순도 지표로 정보획득량을 활용한다. DT 분할의 결과 '조건이 A이고 조건이 B이면 결과집단 C'라는 형태의 규칙으로 표현되어 직관적이며 결과를 쉽게 이해할 수 있다는 장점이 있다[26].

Random Forest(이하 RF)는 DT의 단점인 모형의 안정성과 낮은 예측력을 해소하기 위해 등장한 Ensemble 방법의 하나로 기본적으로 트리를 1,000개 이상 생성한 뒤 분류모형에서 표결에 따라 예측 모형을 생성한다. 많은 개수의 트리를 생성하면서 변수간의 상호작용을 랜덤기법을 통해 약화시켜 시킬 수 있으며, 트리를 1,000개 이상 생성할 경우 에러가 분산되어 과적합에서 자유로운 편이다[27]. 특히, 이 알고리즘은 교육학 연구자들이 현재 많이 활용하는 머신러닝 기법으로 R언어를 활용할 때 사용이 용이하고 변수중요도 그래프를 통해 직관적으로 결과를 이해할 수 있다. 하지만 RF는 블랙박스 모형으로 높은 예측력에 비하여 해석력이 낮은 점이 다소 아쉬운 점으로 여겨진다.

XGBoost(이하 XGB)는 최근 빠른 속도와 높은 성능으로 많이 활용되기 시작한 부스팅 기법이다. XGB의 목적함수는 (수식 1)과 같으며, 여기서 l은 예측 $\widehat{y_u}$과 타겟 y_i의 차이를 의미하는 손실함수이며, k는 나무의 개수, Ω은 모형 복잡도이다[28]. XGB는 손실함수를 최소화 하면서 과적합을 방지하기 위해 나무의 복잡도를 통제하는 방식으로 최적의 모형을 생성한다[29]. 트리의 깊이를 0에서 시작하여 매 분기에서 Gain 값을 계산하며 이 값이 최대가 되도록 트리는 반복적으로 분할한다[30].

$$obj = \sum_{i=1}^{n} l\left(y_i, \widehat{y_i}\right) + \sum_{k=1}^{K} \Omega(f_k) \qquad \text{(수식 1)}$$

XGBoostExplainer(이하 XGBE)는 XGB모형에 해석력을 더할 수 있게 만든 R 패키지 이다[31]. 이 패키지는 패널 별로 각각의 설명변수가 어떤

가중값으로 목표변수를 예측했는지 log-odds 값을 통해 알 수 있으며, 테스트 데이터셋 패널의 특정 변수와 목표변수 사이의 추이를 알 수 있다[29].

Ⅲ. 연구 방법

본 장에서는 연구에서 활용한 데이터와 연구 과정에 대하여 소개한다.

3.1 연구 데이터

본 연구는 서울시교육청 산하 서울교육정책연구소에서 보유하고 있는 서울교육종단연구(Seoul Education Longitudinal Study: SELS) 데이터를 활용하였다. SELS는 서울시 학생의 발달과 관련된 데이터를 장기간에 걸쳐 수집하여 정책과 학교교육이 학생에 미치는 영향 분석을 목적으로 하고 있다[32]. 이에 2010년부터 서울특별시에 소재한 국공립 및 사립학교의 초등학교 4학년 5,059명을 층화추출하여 패널로 선정한 뒤 매년 추적 수집한 데이터이다[33]. 또한 패널로 선정된 학생의 학부모, 학교에 대한 데이터도 함께 수집되며, 학생의 인지적 성취(예: 과목별 학업성취도 등)·비인지적 성취(예: 진로성숙도, 시민의식 등)와 관련된 요인 뿐만 아니라 학교생활과 관련된 여러 요인(예: 학교생활 만족도, 교우관계 등)을 함께 조사하여 데이터로 축적하고 있다.

종단연구 데이터이기 때문에 2010년 실시된 1차년도 이후 서울시교육청 관외로의 전출, 유학, 자퇴 등에 의해 더 이상 추적조사가 이루어질 수 없는 학생들이 매년 발생하여 원표본 유지율이 매년 점차 낮아지고 있다. 본 연구에서는 SELS에서 수집한 4차년도 중학교 1학년 학생 3,725명, 5차년도 중학교 2학년 학생 3,579명, 6차년도 중학교 3학년 학생 3,673명의 학생 데이터와 이와 관련된 학부모, 학교 데이터를 서울교육정책연구소로부터 제공받아 활용하였다. 제공받은 데이터는 1차년도 원표본에 비해 평균적으로 71.1%의 표본을 유지하고 있었다[34].

3.2 연구 과정

본 연구에서는 1) 데이터 전처리 2) 기초통계분석 3) 차원축소 4) 모형학습 및 평가 5) 해석 및 논의 과정을 거쳤다.

데이터 전처리에서는 [13]의 연구에서 제안한 방법에 따라 결측치 대체, 요인화, 정규화, 이진화, 데이터 유도 및 데이터 통합 과정을 통해 전처리를 실시하였다.

기초통계분석에서는 목표변수인 진로성숙도의 구체적인 문항을 살펴보고, 각 학년별 진로성숙도의 문항신뢰도, 평균과 표준편차를 확인하였다. 또한 학년별 진로성숙도 평균값의 차이를 검증하기 위해 분산분석(Analysis of Variance)을 수행하였다. 분산분석의 경우 일반적으로 세 집단 이상의 변수의 평균값에 대한 차이가 통계적으로 유의한지를 판단하는 기법이다[35]. 여기서 F값은 집단내분산에 대한 집단간분산의 비로 세 집단의 평균에 차이가 없다는 귀무가설이 참이라면 F값은 1의 값을 보이며, F값이 클수록 더욱 귀무가설을 기각할 경향이 있다[36].

차원축소 과정에서는 변수 선택 방법을 선정하였으며 구체적으로는 재귀적 변수제거(Recursive Feature Elimination: 이하 RFE) 방식을 활용하였다. 연구에 따라서는 수 백개의 변수를 모든 입력변수로 설정하여 학습시키는 방식을 취하는 방법 또한 많이 활용한다. 이러한 방법은 변수들 간의 상호작용이 모형 학습결과에 영향을 미치는 점, 결과 해석의 복잡성을 수반하는 점 때문에 본 연구에서는 차원을 축소시켜 간단하면서도 해석이 용이한 모형을 도출하고자 하였다. [37]에 따르면, 차원 축소 방법에는 변수를 제거하는 특징선택(Feature Seledtion)과 주성분 분석과 같은 특

징 추출(Feature Extracion)이 있다. 특징 선택은 중요한 특징은 남기고 그렇지 않은 특징은 제거하는 방법으로 최대의 효율을 낼 수 있는 변수의 집합을 찾는 방법이다. 반면 특징 추출은 원변수로 부터 계산되는 새로운 변수를 정의하는 주성분 분석과 같은 방법이다. 본 연구에서 활용한 RFE는 특징 선택 방법이다. 본 연구는 진로성숙도 예측 요인을 찾고 해당 요인에 대한 교육적 처방을 통해 진로성숙도가 높게 나타날 확률을 높여 학생들 모두가 원만한 진로성숙을 이루는 것에 기여함을 목적으로 하고 있다. 허나 특징 추출과 같은 방법을 사용하여 학생과 관련한 변수들의 가중치 합은 앞서 논의한 효과적인 교육처치를 하기에는 유용하지 않다고 판단되었다.

모형 학습의 과정에서는 현재까지 머신러닝 방법을 활용한 교육학 연구자들의 선행연구를 보면, 대부분 DT에 편중되어 있으며 활용한 알고리즘도 CHAID(CHi-square Automatic Interaction Detection)에 집중되어 있었다. CHAID는 사회·과학 연구분야에서 DT를 만들 때 많이 사용하는 알고리즘으로 카이제곱 통계량을 사용하여 다지분류(Multiway Splits)를 수행하는 알고리즘이다[38]. 다른 DT 알고리즘과는 다르게 가지치기를 하지 않고 가장 좋은 예측변수를 이용한 마디 분할의 검정결과가 유의적으로 향상되지 않는다면, 분할은 수행되지 않고 나무는 그대로 종료되는 방식이다[39]. 반면, 최근 몇몇의 연구에서 RF를 활용한 연구가 진행되고 있지만 사례가 많지 않으며, 소수의 연구에서는 클래스 불균형이 발생한 데이터를 그대로 활용하여 과적합된 모형을 해석하는 연구도 찾아볼 수 있었다. 이에 본 연구에서는 DT를 일차적으로 활용하여 기존 교육학 연구자와 맥을 같이 하였으며, DT의 단점으로 뽑히는 모형의 안정성과 다소 부족한 예측력 등을 보완하기 위한 대안적 방법으로 Ensemble 방법 중 RF를 활용하였다. 하지만 RF는 높은 정확도를 보이지만, 모형 학습에 너무 많은 시간이 소요되며 이를 위해 컴퓨팅 리소스를 많이 사용하는 단점이 있으며, 블랙박스 모형으로 결과 해석이 용이하지 않은 단점을 갖고 있다. 이러한 단점을 보완하기 위해 최근 빠른 속도와 정확성을 기반으로 다양한 영역에 활용되고 있는 Ensemble 방법 중 XGB를 활용하였다. 또한 XGB는 블랙박스 모형이나, XGBE 패키지를 함께 활용하여 해석가능한 XGB 모형을 구축할 수 있다는 장점을 갖고 있다. 따라서 앞서 논의한 세 가지 머신러닝 방법을 7:3의 비율로 훈련 집단:테스트 집단으로 나눈 뒤, 정확도(Accuracy)·민감도(Sensitivity)·특이도(Specificity)·Kappa계수·AUC(Area Under the Curve)지표를 활용하여 모형을 평가하고 결과를 해석하였다. 정확도와 민감도, 특이도는 혼동행렬(Confusion Matrix)를 통해 이해할 수 있는데 정확도는 학습한 머신러닝 모델이 데이터 레이블을 정확하게 분류한 비율을 말한다. 민감도는 참긍정률(True Positive Rate)이라고도 하며 긍정인 레이블을 얼마나 정확하게 긍정으로 분류하였는지에 대한 비율이다. 또한 특이도는 참부정률(True Negative Rate)이라고도 하며, 부정인 레이블을 얼마나 정확하게 부정으로 분류하였는지에 대한 비율을 말한다. 정확도, 민감도, 특이도 모두 값이 클수록 모형의 성능이 좋은 것으로 보며, 민감도와 특이도값의 편차가 크지 않는 것 또한 유의하게 살펴보아야 한다. AUC는 ROC(Receiver Operating Characteristic) 커브 그래프 하단 면적의 넓이다. ROC 커브 그래프의 x축은 1-특이도 값, y축은 민감도 값을 나타낸다. AUC값은 0.5~1 사이의 값을 보이며 가장 이상적인 ROC 커브 그래프의 AUC 값은 1이다. Kappa계수는 평가자 간의 분류에 대한 일치도를 측정하는 방법에서 가져온 개념으로 혼동행렬을 이용하여 계산한다. Kappa 계수는 0~1사이 값을 나타내며 1에 가까울수록 실제값과 예측값의 일치도가 높다는 것을 의미한다.

데이터 전처리 및 분석 툴은 R Version 3.6.0,

R studio Version 1.2.5033을 활용하였다. 전처리, 변수 선택에서는 dplyr, caret 패키지를 사용하였다. dplyr 패키지는 일반적으로 데이터 전처리에 많이 활용되는 파이프 연산자인 %>%를 통해 데이터를 손쉽게 처리할 수 있도록 돕는 패키지이다. 또한 caret(Classification And REgression Ttraining) 패키지는 복잡한 회귀와 분류 문제에 대한 모형 훈련과 튜닝 과정을 간소화하는 함수, 변수선택 및 모형평가와 같은 머신러닝 전반에 널리 활용되는 함수를 지원한다[40-41]. 학습 및 모형평가에는 C50, randomForest, xgboost, xgboostExplainer 패키지를 사용하였다.

Ⅳ. 연구의 결과

본 장에서는 DT, RF, XGB 머신러닝 기법을 활용해 중학생 진로성숙도 예측 요인을 탐색한 결과를 소개한다.

4.1 데이터 전처리 결과

전처리 과정은 결측치 대체, 요인화, 정규화, 이진화, 데이터 유도 및 데이터 통합 과정을 거쳤다[13]. 진로성숙도는 각 학년의 평균값을 기준으로 평균보다 높음(=1), 낮음(=0)으로 전처리하였다. 데이터는 요인화, 정규화, 이진화 및 데이터 유도 과정을 거치며 연속형 데이터는 모두 0~1사이의 실수 값을, 범주형 데이터는 0또는 1의 값으로 변환되었다. 이후 학생, 학부모, 학교 데이터를 학생을 중심으로 조인하여 통합된 데이터 셋을 도출하였다.

4.2 기초 통계 분석

이번 단계에서는 목표변수의 문항 구성을 살펴보고, 해당 문항의 기초통계 및 학년별 차이를 검증하였다. SELS의 진로성숙도는 4~6년차 중학교 1~3학년 동일하게 8개 문항 5점 Likert 척도인 자가설문 방식을 이용하였으며 그 구체적인 문항은 <표 1>과 같다. 중학생의 학년별 진로성숙도 평균(Mean)은 중1: 4.01, 중2: 3.89, 중3: 3.91로, 표준편차(Standard Deviation)는 중1: 0.75, 중2: 0.73, 중3: 0.74로 나타났다. 진로성숙도 평균값은 중학교 1학년에 비해 중학교 2학년

〈표 1〉 목표 변수의 문항구성, 기초통계 및 차이 검증

목표 변수	설문 문항 내용	학년	N	Cronbach's α	Mean	SD	F
진로 성숙도	q1. 내가 좋아하는 일이 무엇인지 알고 있다 q2. 내 성격에서 좋은 점이 무엇인지 알고 있다 q3. 내가 관심을 가지고 있는 진로(전공 혹은 직업)에 대한 구체적인 정보를 알아본 적이 있다	중1	3,725	0.92	4.01	0.75	28.49***
	q4. 장래 희망을 이루기 위해 지금 무엇을 해야 하는지 생각하고 있다. q5. 나의 진로(전공 혹은 직업)를 스스로 결정한다.	중2	3,579	0.92	3.89	0.73	
	q6. 내 직업에서 사람들의 인정을 받는 최고 전문가가 되고 싶다. q7. 내 일과 관련된 결정을 내릴 때, 누구보다 중요한 역할을 하는 사람이 되고 싶다. q8. 희망하는 직업을 갖기 위한 어떤 어려움이 있어도 이겨낼 것이다.	중3	3,673	0.93	3.91	0.74	

***p<.001

때 낮아졌으며 중학교 3학년에 다소 증가한 것을 찾을 수 있었다. 또 각 학년별 문항 신뢰도(Cronbach's α)는 0.92, 0.92, 0.93으로 상당히 높은 값을 보이고 있는 것으로 나타났다. 또한 각 학년별 진로성숙도의 차이가 있는지를 확인하기 위해 분산분석으로 각 학년별 진로성숙도 평균값의 차이를 분석하였다. 그 결과 F값은 28.49을 보였으며, 이 값의 유의확률(p-value)이 0.001 미만의 값으로 나와 학년별 진로성숙도가 통계적으로 유의한 차이를 보이고 있는 것을 확인할 수 있었다.

4.3 차원 축소

변수들 간의 상호작용을 방지하고 모형 해석의 용이함과 같은 이점을 취하기 위해 RFE 방식을 활용하여 10개 내외로 선택하였다. 이에 더하여 선행연구에서 유의미한 요인으로 작용한 변수까지 종합적으로 고려하여 차원 축소과정을 수행하였다. 각 학년별 진로성숙도를 목표변수로 하여 변수를 선택한 결과는 <표 2>와 같다. 이를 통해 11개 변수로 중학교 1~3학년 모두 동일하게 추출하였다.

선택된 변수들간의 상관 계수(Correlation Coefficient)를 확인한 결과 변수간의 상관이 크지 않아(<0.75) 선택한 모든 변수를 모형학습에 활용할 수 있었다[27].

4.4 모형 학습 및 평가

모형의 학습 및 평가를 위해 데이터 분할 방법을 활용하였다. 이를 위해 각 학년별 차원 축소된 데이터를 훈련용 데이터세트와 테스트 데이터세트로 7:3 비율 분할하였다. 훈련용 데이터세트는 모형을 적합할 때 활용하였으며, 모형의 성능을 평가할 때에는 테스트 데이터 세트를 활용하였다. 모형학습 결과로 목표변수인 진로성숙도가 평균보다 높음과 낮음으로 분류하는 과정에서 중요하게 활용된 변수의 순위를 확인할 수 있다.

모형평가에서 활용한 지표 중 Kappa계수는 실제값과 예측값의 일치도를 말하며 본 연구에서는 [42]에서 제안한 0~1사이의 값을 6등급으로 구분한 기준을 활용하였다. 이 기준에 따르면 등급의 숫자가 낮을수록 높은 성능(1/6등급이 가장 높은 성능을 나타내며 1에 가까운 값을 보임)으로 평가된다. 추가적으로 AUC 지표는 5등급으로 구분한 [43]의 해석 기준을 활용하였다. 등급의 숫자가 낮을수록 AUC값이 1에 가까운 것으로 높은 성능을 보이는 것이다.

4.4.1 중학교 1학년 진로성숙도 모형

중학교 1학년 데이터를 DT, RF, XGB를 통해 학습한 결과를 살펴보면 <표 3>과 같다. 각 모형별로 진로성숙도가 평균보다 높음과 낮음으로 분류하는 과정에서 가장 중요하게 다루어진 변수는 세 모형 모두 '성취목표'였다. 뒤를 이어 활용된 변수를 각각 살펴보면 DT는 '부모자녀와의 관계', '자기주도학습능력: 학습태도', '자기주도학습능력: 학습노력', '회복탄력성'이었다. RF는 '창의성', '자아개념', '부모자녀와의 관계', '자기통제력', '수업태도', '교사의 수업능력' 순으로 변수 중요도를 보였다. XGB는 RF와 거의 동일하였으나 '회복탄력성', '교사의 수업능력',

〈표 2〉 차원 축소 결과

	선택된 변수(가나다 순)
중학교 1~3학년	(학생이 인식하는) 교사의 수업능력, 부모자녀와의 관계, 성취목표, 수업태도, 자기주도학습능력: 학습노력, 자기주도학습능력: 학습태도, 자기주도학습능력: 학습방법, 자기통제력, 자아개념, 창의성, 회복탄력성

‘수업태도’, ‘자기통제력’에서 차이를 보였다.

모형의 성능 평가 결과인 <표 4>를 보았을 때, 정확도는 앙상블 모형인 RF와 XGB가 0.83으로 나타났으며 민감도와 특이도에서 큰 차이가 없어 모형이 과적합되지 않았다는 것을 확인할 수 있었다. 또한 Kappa계수는 [42]의 기준에 따라 DT는 3/6등급, RF와 XGB는 2/6등급의 기준에 도달하였다. 그리고 AUC 값은 [43]의 기준에 따라 DT는 3/5등급, RF와 XGB는 최상위 등급인 1/5 값을 보였다.

4.4.2 중학교 2학년 진로성숙도 모형

중학교 2학년 데이터를 DT, RF, XGB를 통해 학습한 결과를 살펴보면 <표 5>와 같다. 가장 높은 중요도를 보인 요인은 세 모형 모두 중학교 1학년 모형과 동일한 ‘성취목표’였다. 다음으로 중요한 요인은 DT의 경우 ‘수업태도’, ‘교사의 수업능력’, ‘자기통제력’, ‘회복탄력성’, ‘자아개념’이었다. RF는 ‘창의성’, ‘회복탄력성’, ‘자아개념’, ‘부모자녀와의 관계’, ‘수업태도’, ‘교사의 수업능력’, ‘자기통제력’ 순으로 나타났다. XGB는 ‘창의성’, ‘자아개념’, ‘수업태도’, ‘회복탄력성’,’ 교사의 수업능력’, ‘자기통제력’ 순을 보였다.

모형의 성능 평가 결과인 <표 6>을 통해 모형의 정확도를 보면 0.77~0.80의 값을 보였다. 민감도와 특이도는 모든 모형에서 상호간 편차가 크게 나타나지 않아 모형이 과적합되지 않았다. Kappa와 AUC 또한 [42], [43]에 따라 DT는 3/6, 3/5 등급을 RF는 2/6, 2/5 등급을 XGB는 3/6, 2/5 등급을 보였다.

4.4.3 중학교 3학년 진로성숙도 모형

중학교 3학년 기계학습 결과를 보면 <표 7>과 같다. 가장 중요한 변수는 앞서 살펴본 모형과 마찬가지로 세 모형 모두 ‘성취목표’였으며 2~4번째 중요한 변수 또한 모두 ‘창의성’, ‘자아개념’,

〈표 3〉 중학교 1학년 모형별 변수 중요도 순위

Decision Tree: C5.0	Random Forest	XGBoost
성취목표	성취목표	성취목표
부모자녀와의 관계	창의성	창의성
자기주도학습능력: 학습태도	자아개념	자아개념
자기주도학습능력: 학습노력	부모자녀와의 관계	부모자녀와의 관계
회복탄력성	자기통제력	회복탄력성
-	수업태도	교사의 수업능력
-	교사의 수업능력	수업태도
-	회복탄력성	자기통제력
-	자기주도학습능력: 학습노력	자기주도학습능력: 학습노력
-	자기주도학습능력: 학습방법	자기주도학습능력: 학습방법
-	자기주도학습능력: 학습태도	자기주도학습능력: 학습태도

〈표 4〉 중학교 1학년 모형 평가 결과

	정확도	민감도	특이도	Kappa	AUC
Decision Tree: C5.0	0.77	0.77	0.78	0.55	0.78
Random Forest	0.83	0.85	0.81	0.66	0.90
XGBoost	0.83	0.85	0.79	0.65	0.90

〈표 5〉 중학교 2학년 모형별 변수 중요도 순위

Decision Tree: C5.0	Random Forest	XGBoost
성취목표	성취목표	성취목표
수업태도	창의성	창의성
교사의 수업능력	회복탄력성	자아개념
자기통제력	자아개념	수업태도
회복탄력성	부모자녀와의 관계	회복탄력성
자아개념	수업태도	교사의 수업능력
-	교사의 수업능력	자기통제력
-	자기통제력	자기주도학습능력: 학습태도
-	자기주도학습능력: 학습태도	부모자녀와의 관계
-	자기주도학습능력: 학습방법	자기주도학습능력: 학습방법
-	자기주도학습능력: 학습노력	자기주도학습능력: 학습노력

〈표 6〉 중학교 2학년 모형 평가 결과

	정확도	민감도	특이도	Kappa	AUC
Decision Tree: C5.0	0.78	0.77	0.79	0.57	0.78
Random Forest	0.80	0.80	0.81	0.61	0.88
XGBoost	0.77	0.78	0.77	0.55	0.87

〈표 7〉 중학교 3학년 모형별 변수 중요도 순위

Decision Tree: C5.0	Random Forest	XGBoost
성취목표	성취목표	성취목표
창의성	창의성	창의성
자아개념	자아개념	자아개념
회복탄력성	회복탄력성	회복탄력성
자기주도학습능력: 학습태도	부모자녀와의 관계	부모자녀와의 관계
부모자녀와의 관계	교사의 수업능력	자기주도학습능력: 학습태도
-	수업태도	수업태도
-	자기통제력	교사의 수업능력
-	자기주도학습능력: 학습태도	자기통제력
-	자기주도학습능력: 학습방법	자기주도학습능력: 학습방법
-	자기주도학습능력: 학습노력	자기주도학습능력: 학습노력

〈표 8〉 중학교 3학년 모형 평가 결과

	정확도	민감도	특이도	Kappa	AUC
Decision Tree: C5.0	0.81	0.82	0.80	0.63	0.78
Random Forest	0.80	0.78	0.81	0.60	0.89
XGBoost	0.81	0.85	0.77	0.62	0.89

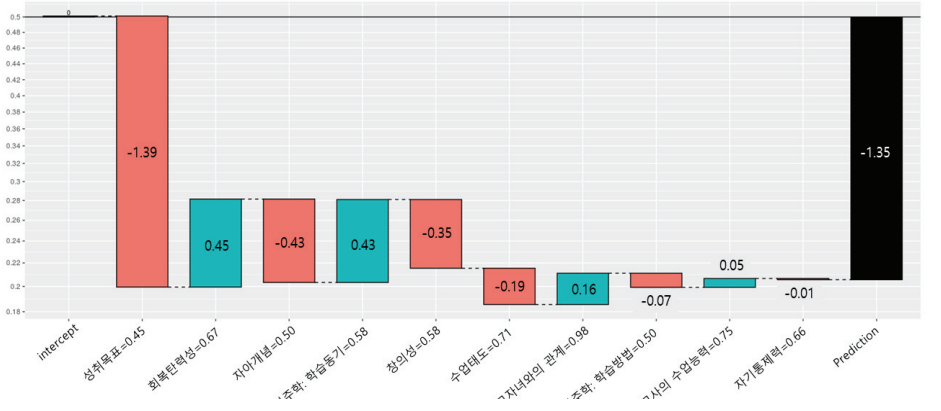
‘회복탄력성’ 순으로 나타났다. 이보다 낮은 중요도를 보인 변수는 DT에서는 ‘자기주도학습능력: 학습태도’, ‘부모자녀와의 관계’ 순으로 나타났다. RF는 ‘부모자녀와의 관계’, ‘교사의 수업능력’, ‘수업태도’, ‘자기통제력’ 과 같은 중요도를 보였다. XGB는 ‘부모자녀와의 관계’, ‘자기주도학습능력: 학습태도’, ‘수업태도’, ‘교사의 수업능력’ 순을 보였다.

모형의 성능 평가 결과는 <표 8>과 같았다. 정확도에 있어서는 DT, RF, XGB 각각 0.81, 0.80, 0.81을 보였다. 민감도와 특이도는 모든 모형에서 상호간 편차가 크게 나타나지 않았다. Kappa와 AUC 또한 [42], [43]에 따라 DT는 2/6, 3/5, RF는 3/6, 2/5, XGB는 2/6, 2/5 등급을 보였다.

4.5 해석 및 논의

중학교 1~3학년까지의 진로성숙도 분류 모형의 정확도, 민감도, 특이도, Kappa 계수 및 AUC를 종합적으로 고려한 결과 단일 트리인 DT보다는 앙상블 모형인 RF와 XGB가 우수한 것으로 볼 수 있었다. 여기에 XGB는 XGBE 패키지의 2가지 그래프를 활용하여 설명력을 더할 수 있다.

첫 번째, 폭포수 그래프를 통해 학생 각자의 요인별 목표변수에 대한 영향력을 확인할 수 있었다. 3번 패널 학생의 중학교 2학년 요인별



〈그림 1〉 3번 패널 중학교 2학년 예측 log-odds 그래프

〈표 9〉 3번 패널 중2 진로성숙도 예측 확률

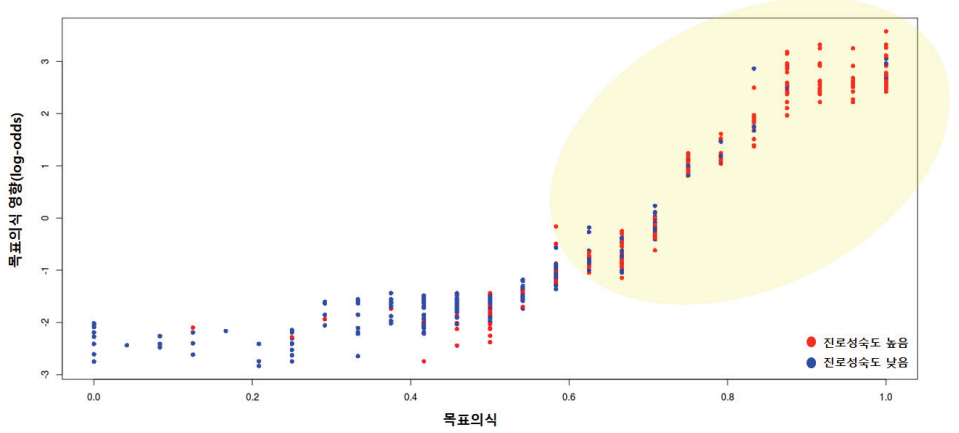
변수별 log-odds	변수	누적 log-odds
0	Baseline (Intercept)	0
-1.39	성취목표	-1.39
+0.45	회복탄력성	-0.94
-0.43	자아개념	-1.37
+0.43	자기주도학습능력: 학습동기	-0.94
-0.35	창의성	-1.29
-0.19	수업태도	-1.48
+0.16	부모자녀와의 관계	-1.32
-0.07	자기주도학습능력: 학습방법	-1.39
+0.05	교사의 수업능력	-1.34
-0.01	자기통제력	-1.35

log-odds 값을 나타낸 그래프는 <그림 1>이다. XGBE모형에서 예측확률이 0.5 미만으로 나올 경우 진로성숙도 낮음, 0.5 이상일 경우 진로성숙도 높음으로 분류한다[29]. 즉, 0에 가까울수록 진로성숙도가 낮을 확률이 높으며, 1에 가까울수록 진로성숙도가 높을 가능성이 커진다. 일예를 들어보면, XGBE 모형은 3번 패널의 중학교 2학년 진로성숙도 예측 확률을 0.205로 예측했다. 이 과정을 <그림 1>과 <표 9>의 통계값을 통해 살펴보면 다음과 같다. <그림 1> 변수별 폭포수 그래프에 나타난 log-odds 값을 누적하여 계산하면 가장 오른쪽 검은색 바에 나타난 -1.35이다. 이를 XGBE 패키지의 로지스틱 함수에 적용하면 (수식 2)와 같고, 이를 통해 진로성숙도 예측 확률을 구할 수 있다.

$$\frac{1}{1+e^{-(-1.35)}}=0.205 \qquad \text{(수식 2)}$$

3번 패널 학생이 중학교 2학년 시기에 진로성숙도가 낮을 것으로 예측된 이유는 (+)값을 보인 '회복탄력성', '자기주도학습능력: 학습동기', '부모자녀와의 관계', '교사의 수업능력'의 비중보다 (-)값을 보인 '목표의식', '자아개념', '창의성', '수업태도', '자기주도학습능력: 학습방법', '자기통제력'의 비중이 더 컸기 때문이다. 이상의 내용을 동일 패널 중1, 중3 그래프와 통계값을 확인해 보았을 때, 같은 변수라 하더라도 때로는 진로성숙도에 긍정적 영향을 끼치는 요인으로, 때로는 부정적 영향을 끼치는 요인이 됨을 찾을 수 있었다. 이를 통해 사회과학적 구인에서 주로 다루어지고 있는 요인 중심적 접근 방식의 대안인 사람 중심적 연구방식[6]의 하나로 XGBE 폭포수 그래프를 활용할 수 있을 것이다. 요인 중심적 접근 방식은 학생 각자가 갖고 있는 사회과학적 구인의 형성 정도와 과정이 모두 다르지만, 이를 하나의 동일한 모집단으로 규정하고 특정한 변수들 끼리의 관계가 모든 학생들에게 일반화됨을 가정하는 분석방법을 말한다. 따라서 여기에서 얻은 것은 학생 각각의 차이를 살피고 학생에게 적합한 교육처치를 할 수 있는 근거자료로 본 그래프가 활용될 수 있을 것이다.

두 번째, 개별 변수와 목표변수와의 관계를 그래프로 확인할 수 있다. 개별 변수를 x값, 개별변수 log-odds 비중을 y값으로 하면 테스트 패널을 대상으로 진로성숙도에 개별 변수의 영향 추이를 확인할 수 있다. <그림 2>는 진로성숙도에 목표의식이 어떤 영향을 미치는지를 중학교 3학년 테스트 데이터셋 패널을 대상으로 플로팅한 것이다. 노란색으로 음영처리된 부분을 중심으



〈그림 2〉 중3 테스트 패널 목표의식 변수의 영향

로 보면 ‘목표의식’이 평균값(0.67) 이상일 때에 비례하여 진로성숙도가 높은 학생들의 비율이 많아짐을 보여준다. 상대적으로 목표의식이 평균값 미만일 때는 특정한 관계를 발견하기 어려웠다. 이러한 목표의식 변수의 영향은 중학교 1학년, 2학년 모두 <그림 2>와 동일한 양상을 보임을 확인할 수 있었다. 나머지 다른 변수들 또한 모두 확인해 보았으나 <그림 2>와 같이 두드러진 특징을 찾을 수는 없었다. 이 그래프를 통해 중1~3학년 XGB모형에서 가장 중요하게 다루는 변수가 ‘목표의식’인 이유를 설명할 수 있었다.

위에서 논의한 내용을 종합해 볼 때, XGB모형은 머신러닝 기법 중 예측력이 좋으며 XGBE를 활용하여 모형의 해석력을 높일 수 있었다. 따라서 사회과학적 구인을 머신러닝을 통해 분석함에 있어서 XGB모형이 본 연구를 통해서 DT, RF, XGB 세 가지 기법 중 예측력과 설명력을 모두 갖춘 적합한 모형이라 할 수 있다.

이에 선택한 XGB 모형 학습 결과를 바탕으로 논의를 진행하면 다음과 같다. 첫째, 청소년기 진로성숙도 관련 선행연구에서 주로 다루지 않은 ‘성취목표’, ‘회복탄력성’이 높은 변수 중요도를 보이는 것으로 나타났다. 이 두 요인은 주로 대학생 진로성숙도에 영향을 미치는 요인으로 주로 다루어진 것으로 파악된다[44-47]. 하지만 본 연구 결과를 통해 ‘성취목표’, ‘회복탄력성’이 성인기 뿐만 아니라 청소년기 진로성숙도에도 큰 영향을 주는 요인임을 확인하였다. 이는 [48]과 [49]의 연구에 의해 지지되는 결과였다. 따라서 성인기 뿐만 아니라 청소년기 또한 ‘성취목표’가 진로성숙도에 다른 어떠한 요인들 보다 중요하다는 사실을 찾을 수 있었다.

둘째, ‘창의성’, ‘자아개념’, ‘교사의 수업능력’, ‘부모자녀와의 관계’, ‘자기주도학습’과 같은 변수는 여러 선행연구에서 나타난 바, 본 연구의 타당성을 확인할 수 있었다. ‘창의성’은 [50]의 연구와 맥을 같이 하며, ‘자아개념’은 다수의 연구에서 진로성숙도에 정적(+)인 영향을 미치는 요인임을 찾을 수 있었다[10, 51-52]. 그리고 교사의 수업능력은 [11]과 [53]의 연구에서 말한 교사의 열의와 같은 맥락 안에서 지지되는 결과였다. 또한 ‘부모자녀와의 관계’는 [23], [53-54]의 선행연구와 유사함을 보이고 있음을 확인하였다. ‘자기주도학습’ 관련 변인은 [55]의 연구에 의해 지지됨을 찾을 수 있었다.

셋째, 본 연구 결과를 통해 중등 진로교육 내용 명료화를 위해 ‘성취목표’, ‘자아개념’, ‘회복탄력성’, ‘자기통제력’과 같은 내용을 강화하여야 할 것이다. 또한 중등교육 전반적으로 ‘자기주도학습능력’을 신장시키기 위한 노력을 기울여야하며, ‘부모자녀와의 관계’를 긍정적으로 유지하기 위해 학교는 가정에서 부모와 자녀간의 상호작용을 늘릴 수 있는 교육활동을 기획하여 운영함이 학생의 진로성숙도를 높이는 데 효과적일 것이다. 이에 더하여 학생의 ‘창의성’을 신장시키고 ‘교사의 수업능력’을 높이기 위해서 단위 학교에서는 허용적인 수업 분위기와 창의성을 신장시킬 수 있는 교수・학습 방법을 적용하며, 이를 장려하기 위해 교사의 열의를 높일 수 있는 환경을 조성하기 위해 노력해야 할 것이다. 즉, 학교 진로교육계획을 수립할 때, 앞서 논의한 내용을 충실히 반영하고 뿐만 아니라 교육현장에서 명료화된 내용을 교육할 수 있도록 지속적인 노력이 필요할 것이다.

Ⅴ. 결론 및 제언

본 연구는 SELS 4차~6차년도 데이터를 머신러닝 기법 중 DT(C5.0), RF, XGB를 활용하여 중학생의 진로성숙도에 영향을 주는 요인을 탐색하였다. 그 결과로 XGB 기법이 본 연구에서 가장 적합한 알고리즘임을 규명하였으며 ‘성취목표’, ‘창의성’, ‘자아개념’, ‘부모자녀와의 관계’, ‘회복탄력성’과 같은 요인이 중학생의 진로성숙

도를 예측함에 있어 가장 높은 변수 중요도를 보임을 찾을 수 있었다. 또한 XGBE를 통해 개별 학생의 진로성숙도에 긍정적·부정적 영향을 주는 요인을 확인하였으며, 학생이 성장해 감에 따라 각 요인의 변화하는 추이도 함께 발견하였다. 마지막으로 각각의 변수와 진로성숙도의 전반적인 상관관계를 확인하여, 특히 '성취목표' 요인의 경우 평균 이상의 집단에서 진로성숙도가 높은 학생들이 나타날 확률이 높아짐을 발견해 중학생의 진로성숙도 예측에 가장 중요한 요인임을 규명하였다. 또한 성인기 진로성숙도에서 주로 다루었던 '성취목표', '회복탄력성' 요인이 청소년기의 진로성숙도에서도 중요한 변수임을 발견한 점이 주목할만하다. 이에 더하여 XGB 및 XGBE의 결과가 진로성숙도 영향요인과 관련된 선행연구를 지지해 그 설명력을 높여 줄 수 있었다는 점에서 그 의의를 찾을 수 있었다. 또한 사회과학적 구인에 대한 머신러닝을 활용하여 인간의 심리에 대한 이해를 확장시켰다는 점에서도 의의를 들 수 있다. 나아가 머신러닝 결과를 활용하여 중학교 진로교육 내용의 명료화를 위한 방향성을 찾는 중요한 기초자료를 제공하였다. 하지만 XGB의 경우 트리를 형성할 때 수치형 변수에 비하여 범주형 변수가 저평가될 수 있다는 점, 트리의 복잡도가 높을수록 독립변수와 목표변수의 방향성을 일부 가지에 의존한다는 점이 한계점으로 지적된다[29].

이와 같은 결과를 바탕으로 후속 연구를 제언하면 다음과 같다. 첫째, XGBE 이외의 예측력과 설명력을 함께 활용할 수 있는 머신러닝 기법을 사회과학 구인에 적용하여 각각의 결과를 비교·설명하는 연구가 이루어져야 할 것으로 보인다. 둘째, 차원을 축소할 때 활용하는 특징 선택과 특징 추출 중 어느 것이 더 모형의 예측력과 설명력을 높일 수 있는지에 대한 비교·검토 연구 또한 이루어져야 할 것으로 보인다. 셋째, 현재 SELS 1차 코호트 조사는 종료되고 2차 코호트 조사가 이루어지고 있다. 이에 1차와 2차 코호트 자료를 비교하여 약 10년간의 시차를 두고 진로성숙도에 영향을 주는 요인은 어떻게 변하였는지를 규명하는 연구 또한 필요할 것으로 보인다.

참 고 문 헌

[1] 이종범, 정철영, "초등학생 진로발달 검사도구 타당화 연구", 진로교육연구, 제18권, 2호, pp.79-105, 2005.

[2] 이시연, "중학생의 진로성숙도가 학교생활적응에 미치는 영향: 공동체의식의 매개효과를 중심으로", 한국콘텐츠학회논문지, 제17권, 6호, pp.622-631, 2017.

[3] 김보람, 김봉환, "진로탐색집단상담프로그램이 학업중단청소년의 진로결정수준과 진로정체감에 미치는 효과", 진로교육연구, 제28권, 2호, pp.1-22, 2015.

[4] 하아영, 조한익, "중학생의 지능에 대한 암묵적 신념, 공동체의식 및 진로성숙도의 관계:국어, 영어, 수학교과 자기효능감의 조전된 매개효과", 한국인간발달학회, 제27권, 3호, pp.109-135, 2020.

[5] 백승원, 윤채영, "중학생의 학업적 자기효능감, 진로성숙도, 학교적응의 종단적 관계", 교육혁신연구, 제30권, 3호, pp.175-199, 2020.

[6] 이종범, I. H. Lee, "초등학생 진로발달 유형에 관한 잠재프로파일분석", 한국실과교육학회지, 제33권, 1호, pp.65-81, 2020.

[7] 임현정, "초등학생의 진로성숙도에 대한 개인, 가정, 학교의 영향", 한국교육문제연구, 제34권, 4호, pp.265-285, 2016.

[8] 박지현, "중학생의 진로성숙도 영향요인 연구-자아존중감 조절효과를 중심으로-", 인문사회21, 제11권, 3호, pp.329-344, 2020.

[9] 허균, "잠재성장모형을 활용한 진로성숙도의

변화궤적과 성별, 자아존중감 및 부모애착 시간 효과의 구조관계", 직업교육연구, 제31권, 2호, pp.193-209, 2012.

[10] 이현미, 정제영, "중학생의 진로성숙도에 영향을 미치는 요인 분석: 경기교육종단연구 (GEPS)를 중심으로", 청소년학연구, 제24권, 2호, pp.117-139, 2017.

[11] 전현정, 정혜원, "중학생의 진로성숙도와 학교 특성 변인 및 학생 특성 변인과의 관계 분석", 한국청소년연구, 제29권, 3호, pp. 213-240, 2018.

[12] 전화숙, 임혜정, 이기혜, "중학생의 진로성숙도 영향요인 분석: 학교 효과를 중심으로",한국교육사회학회 학술대회자료집, pp.1-28, 2016.

[13] 박정, 조완섭, "교육종단연구 분석을 위한 빅데이터 플랫폼 개발 및 적용", 한국빅데이터학회지, 제5권, 1호, pp.11-27, 2020.

[14] Cupples, L. A., Bailey, J. N., Cartier, K. C., Falk, C. T., Liu, K.-Y., Ye, Y., Yu, R., Zhang, H., and Zhao, H., "Data mining", Genetic Epidemiology, Vol.29, No.S1, pp.103-109, 2005.

[15] 박정, "교육종단연구 분석을 위한 빅데이터 플랫폼 연구 개발 및 응용", 충북대학교 박사학위논문, 2020.

[16] 김영식, 김훈호, "머신러닝 기법을 활용한 사교육 참여 예측 모형 탐색", 교육재정경제연구, 제28권, 3호, pp.29-52, 2019.

[17] Yarkoni, T., and Westfall, J., "Choosing prediction over explanation in psychology: Lessons from machine learning", Perspectives on Psychological Science, Vol.12, No.6, pp.1100-1122, 2017.

[18] 임언, 정윤경, 상경아, "진로성숙도 검사개발 보고서", 한국직업능력개발원, 2001.

[19] 강혜정, 강성현, 임은미, "일반계 고등학생의 학업동기와 진로동기 수준에 따른 집단분류 가능성 탐색", 아시아교육연구, 제17권, 2호, pp.151-175, 2016.

[20] 길혜지, 윤지윤, "진로성숙도가 높은 학생과 학교 특성 분석", 제1회 경기교육종단연구 학술대회 논문집, pp.143-162, 2014.

[21] 오석영, "중학생의 관계형성 및 자기효능감이 진로경험 및 진로성숙도에 미치는 영향-진학희망계열 집단 비교를 중심으로", 진로교육연구, 제25권, 3호, pp.77-94, 2012.

[22] 윤현, "초등학생의 진로성숙도에 영향을 미치는 요인에 관한 연구 - 성별에 따른 차이를 중심으로-", 사회복지 실천과 연구, 제8권, pp.97-124, 2011.

[23] 이형실, "부모의 양육행동이 청소년의 진로성숙도에 미치는 영향: 또래관계의 매개효과", 한국가정과교육학회지, 제27권, 4호, pp.109-119, 2015.

[24] 진성미, "서울시 초·중·고교생의 진로성숙도 관련 변인별 집단 비교", 한국교육문제연구, 제29권, 2호, pp.133-156, 2011.

[25] 정윤경, 이상은, "우리나라 청소년의 진로성숙도와 관련 변인: 개인, 가정, 학교 특성을 중심으로", 한국사회학회 사회학대회 논문집, pp.259-305, 2005.

[26] 박나영, 김장일, 정용규, "Naive Bayes 분석기법을 이용한 유방암 진단", 서비스연구, 제3권, 1호, pp.87-93, 2013.

[27] Kuhn, M., and Johnson, K.(권정민 옮김), "실천 예측 분석 모델링", 서울:에이콘, 2018.

[28] 이유나, "Cost-Sensitive Learning을 활용한 심뇌혈관 질환 발생 예측 모형 개발", 충북대학교 석사학위 논문, 2019.

[29] 황혜진, 김수현, 송규원, "XGBoost 모델 해석을 통한 노인의 인지능력 개선·악화 요인 탐구", 한국차세대컴퓨팅학회 논문지, 제14권, 3호, pp.16-24, 2018.

[30] Chen, T., and Guestrin, C., "Xgboost: A scalable tree boosting system", Proceedings of the 22nd ACM Sigkdd International Conference on Knowledge Discovery and Data Mining, ACM,

pp.785-794, 2016.

[31] David, F., “NEW R Package: The XGBoost Explainer –Applied Data Science-”, Medium, 2017.

[32] 서울교육연구정보원, “데이터 기반 서울교육 정책 설계: 서울교육종단연구”, 서울특별시교육청, 2010.

[33] 유진은, 노민정, “Elastic net 을 통한 학생의 창의성 예측 모형 연구”, The SNU Journal of Education Research, 제27권, 3호, pp.185-205, 2018.

[34] 김성식, 김준엽, 곽한석, 최유림, 박미희, 차미선, 곽나람, “서울교육종단연구 8차년도 기초 분석 보고서”, 서울특별시교육연구정보원, 2017.

[35] 조철호, “SPSS/AMOS 활용 구조방정식모형 논문 통계분석”, 서울: 청람, 2015.

[36] 정충영, 최이규, “SPSSWIN 을 이용한 통계분석”, 서울:무역경영사, 2011.

[37] Alpaydin, E., “Introduction to machine learning”, MIT press, 2020.

[38] Han, J., Kamber, M., and Pei, J.(정사범, 송용근 옮김), “데이터 마이닝: 개념과 기법”, 서울:에이콘, 2016.

[39] Kass, G., “An Exploratory technique for investigating large quantities of categorical data”, Applied Statistics, Vol.29, No.2, pp.119-129, 2018.

[40] 나종화, “R 데이터마이닝”, 경기:자유아카데미, 2017.

[41] 서민구, “R을 이용한 데이터 처리&분석 실무”, 서울:길벗, 2014.

[42] Viera, A.J., and Garrett, J. M., “Understanding inter observer agreement: the kappa statistic”, Fam med, Vol.37, No.5, pp.360-363, 2005.

[43] Trifonova, O. P., Lokhov, P. G., and Archakov, A. I., “Metabolic profiling of human blood”, Biochemistry Supplement Series B: Biomedical Chemistry, Vol.7, No.3, pp.179-186, 2013.

[44] 박상희, 이지민, “대학생이 지각한 부모의 심리적 통제가 학습된 무기력과 진로성숙도에 미치는 영향 -회복탄력성의 조절된 매개효과-”, 한국가족관계학회지, 제24권, 4호, pp.129-150, 2020.

[45] 성낙훈, “체육전공대학생들의 성취목표지향성과 진로의식성숙도 관계”, 한국체육과학회지, 제23권, 4호, pp.69-80, 2014.

[46] 이호현, “캡스톤디자인 항공모의비행 교과목 수강 후 진로결정 자기효능감, 진로태도 성숙도, 진로결정수준의 변화연구 : 서울 소재 2년제 대학 항공과 재학생들을 대상으로”, 취업진로연구, 제10권, 2호, pp.1-23, 2020.

[47] 이희주, 이해영, 강경자, “간호대학생의 임상수행능력, 교수-학생상호작용, 진로성숙도가 회복탄력성에 미치는 영향”, 한국웰니스학회지, 제12권, 1호, pp.425-437, 2017.

[48] 성춘향, 박용한, “청소년기 생애목표지향의 종단적 변화 탐색”, 한국교육문제연구, 제36권, 4호, pp.97-122, 2018.

[49] 이병임, “중학생의 회복탄력성 및 도덕지능과 진로성숙도의 관계”, 인격교육, 제13권, 3호, pp.143-155, 2019.

[50] 어윤경, “진로성숙도를 매개로 한 진로체험 활동의 창의성 함양 효과 연구”, 한국교육학연구, 제21권, 2호, pp.197-219, 2015.

[51] 박혜숙, 전명남, “서울시 교육종단자료에 나타난 초・중・고등학생의 자아존중감과 진로성숙도의 관계 탐색”, 한국교육문제연구, 제32권, 2호, pp.59-83, 2014.

[52] 차정원, 이형실, “청소년의 가족환경 및 또래환경과 진로성숙도의 관계에서 자아존중감의 매개효과”, 한국가정과교육학회지, 제26권, 3호, pp.53-67, 2014.

[53] 최인희, “중학생의 진로성숙도의 변화와 영향요인 탐색”, 한국교육, 제46권, 1호, pp. 161-186, 2019.

[54] 정경화, 김기승, “중학생들의 부모애착과 진로성숙도의 관계분석”, 한국산학기술학회 논문지, 제19권, 10호, pp.475-482, 2018.
[55] 김지선, 원새연, 이원재, “경기도 중·고등학생의 자기주도적 학습, 건강및 의식, 교사사기, 진로의식 간의 관계 분석”, 제1회 경기교육종단연구학술대회논문집, pp.77-102, 2014.

저 자 소 개

박 정(Jung Park)
·2020년: 충북대학교 빅데이터협동과정(박사)
·2010년~현재: 대전광역시교육청(교사)
·관심분야: 빅데이터, 머신러닝, 교육데이터마이닝