JPEE 실천공학교육논문지
JOURNAL OF PRACTICAL ENGINEERING EDUCATION

실천공학 교육 방법론
J. Pract. Eng. Educ. 15(1), 31-38, 2023

비전공자 대상 머신러닝 모델 학습 및 활용교육 커리큘럼

A Machine Learning Model Learning and Utilization Education Curriculum for Non-majors

허 경*

경인교육대학교 컴퓨터교육과

Kyeong Hur*

Department of Computer Education, Gyeong-In National University of Education, Anyang 13910, Korea

[요 약]

본 논문에서는 비전공자들을 위한 기초 머신러닝 모델 학습 및 활용교육 커리큘럼을 제안하고, Orange 머신러닝 모델 학습 및 분석 도구를 활용한 교육 방법을 제안하였다. Orange는 오픈 소스기반 머신러닝 및 데이터 시각화 도구로서, 복잡한 프로그래밍 없이 시각적인 위젯을 사용하여, 데이터를 학습시켜 머신러닝 모델을 만들 수 있다. Orange는 비전공자 학부생부터 전문가 그룹까지 다양하게 사용되는 플랫폼이다. 본 논문에서는 한 학기 분량의 기초 머신러닝 모델 학습 및 활용교육 커리큘럼과 주별 실습 내용을 제시하였다. 그리고, 머신러닝 모델 학습 및 활용에 대한 교육 내용 실체를 실증하기 위해, Orange 도구를 활용하여, 분류 데이터(Categorical Data) 표본과 수치 데이터(Numerical Data) 표본으로부터 머신러닝 모델을 학습시키고, 모델을 활용하여 모집단의 결과를 예측하는 활용 사례들을 제안하였다. 마지막으로 본 커리큘럼에 대한 교육 만족도를 비전공자 대상으로 조사 및 분석하였다.

[Abstract]

In this paper, a basic machine learning model learning and utilization education curriculum for non-majors is proposed, and an education method using Orange machine learning model learning and analysis tools is proposed. Orange is an open-source machine learning and data visualization tool that can create machine learning models by learning data using visual widgets without complex programming. Orange is a platform that is widely used by non-major undergraduates to expert groups. In this paper, a basic machine learning model learning and utilization education curriculum and weekly practice contents for one semester are proposed. In addition, in order to demonstrate the reality of practice contents for machine learning model learning and utilization, we used the Orange tool to learn machine learning models from categorical data samples and numerical data samples, and utilized the models. Thus, use cases for predicting the outcome of the population were proposed. Finally, the educational satisfaction of this curriculum is surveyed and analyzed for non-majors.

Key Words: AI, Curriculum, Machine learning, Model, Non-major undergraduates, Orange tool

http://dx.doi.org/10.14702/JPEE.2023.031

Received 30 March 2023; **Revised** 17 April 2023
Accepted 19 April 2023

***Corresponding Author**
E-mail: khur@ginue.ac.kr

I. 서 론

데이터 기반 사회는 빅데이터를 수집하고 가공하여 분석을 통해, 다양한 현상을 예측할 수 있는 데이터 기반 의사결정 체계를 구축하고 있다. 이러한 이유로 데이터 분석가의 역할이 부각되고 있다[1-4]. 이에 모든 분야에서 워크플로 및 시각화를 쉽게 구성할 수 있고 대량의 데이터를 분석할 수 있는 도구가 필요하다. Orange는 데이터 분석 및 시각화를 수행하고 데이터 흐름을 확인하며 생산성을 높일 수 있는 플랫폼이다[5]. 오픈 소스 플랫폼으로서 다양한 분야에 대한 최신 머신러닝 모델 및 데이터 소스를 추가할 수 있다. Orange는 직관적이며 복잡한 데이터 시각화 및 기본 기계 학습 분석을 수행할 수 있다. Orange를 사용하면 프로그래밍 방법을 몰라도 데이터를 분석할 수 있다. 또한 데이터 마이닝 전문가인 공동 작업자, 동료 및 교육생과 소통할 수 있다.

전 세계의 학교, 대학 및 전문 교육 과정에서 사용되는 Orange는 실습 교육 및 데이터 과학 개념의 시각적 일러스트레이션을 지원한다. Orange는 오픈 소스기반 머신러닝 및 데이터 시각화 도구로서, 다양한 도구 상자를 사용하여 시각적으로 데이터 분석 워크플로를 구축할 수 있다. Orange는 대화형 데이터 시각화로 간단한 데이터 분석을 수행한다. 수집한 데이터에 대한 통계 분포, 박스 플롯 및 산점도를 보고, 의사결정 트리, 계층적 클러스터링, 히트맵 및 선형 예측을 통해 자세히 분석할 수 있다. 특히 데이터 속성 순위 지정 및 선택을 통해 다차원 데이터도 2차원으로 분석할 수 있다. Orange는 신속한 정성 분석을 위한 대화형 데이터 탐색과 그래픽 사용자 인터페이스를 사용하여, 프로그래밍 대신 탐색적 데이터 분석에 집중할 수 있고, 기본값을 통해 데이터 분석 워크플로우의 빠른 프로토타이핑이 가능하다. 캔버스에 위젯을 배치하고, 연결하고, 데이터 세트를 로딩하고 분석한다.

데이터분석을 위한 머신러닝 모델 학습 및 활용교육에 대한 선행 연구 결과는 많지 않은 상태이다. 관련 선행 연구를 살펴보면, 참고문헌 [6]에서는 데이터분석 교육을 통해 갖추어야 할 5가지 소양을 제시하였다. 참고문헌 [7]에서는 빅데이터 가공 과정에 반드시 빅데이터 처리 프레임워크 또는 고성능 컴퓨터가 필요한 것은 아니며, 빅데이터 분석 및 응용 방법 중심으로 데이터분석 교육을 실시해야 한다고 제시하였다.

참고문헌 [8]에서는 스프레드시트를 이용한 42차시 교육 프로그램의 내용으로, 스프레드시트 기본 기능과 문제 해결 주제를 정하고 문제를 해결해 보는 데이터 수집 및 시각화 기반 교육 프로그램을 제안하였다. 참고문헌 [9]에서는 스프레드시트 기반 데이터 과학 교육 프로그램이 컴퓨팅 사고력 향상에 효과적이라고 분석하였다. 참고문헌 [10]에서는 비전공자들을 위한 기초 데이터과학 실습 커리큘럼과 스프레드시트 데이터 분석 도구를 활용한 교육 방법을 제안하고, 스프레드시트를 활용한 선형 회귀 분석 예제들을 제시하였다.

참고문헌 [10]에서는 비전공자를 대상으로 양적자료, 즉, 수치데이터(Numerical Data)의 개념과 질적자료, 즉, 분류 데이터(Categorical Data)의 개념을 교육한다. 이후, 데이터 집계, 데이터 가공 그리고 데이터 시각화를 교육한다. 이는 머신러닝 모델의 학습 데이터를 정의하고 전 처리하는 데 있어 알아야 할 선수 지식이다. 또한, 참고문헌 [10]에서는 수치데이터 항목 간 상관관계와 분류 데이터 항목 간 연관성을 분석하는 방법을 교육한다. 이 부분은 다수의 학습 데이터에서 이미 지정한 목적 변수에 영향력이 큰 설명 변수들을 추출하는 데 알아야 할 선수 지식이다. 그리고 참고문헌 [10]에서는 표본으로부터 모집단 평균값 구간을 추정하고, 가설에서 추정한 값을 검정하는 방법을 교육한다. 이 추정과 검정 방법은 기존 통계학 이론에 기초한다. 이와 같이, 참고문헌 [10]의 스프레드시트 기반 기초 데이터과학 커리큘럼을 기 이수한 학생들이 기초 머신러닝 모델 학습 및 활용교육 커리큘럼을 학습할 수 있다.

기초 데이터 과학의 이해를 위해 제안된 참고문헌 [10]의 커리큘럼과 연계하여, 머신러닝 모델을 학습하고 활용하는 교육 방법을 제안하기 위해, 본 논문에서는 비전공자들을 위한 기초 머신러닝 모델 학습 및 활용교육 커리큘럼을 제안하고, Orange 도구를 활용한 교육 방법을 제안하였다. Orange는 오픈 소스기반 머신러닝 및 데이터 시각화 도구로서, 복잡한 프로그래밍 없이 시각적인 위젯을 사용하여, 데이터를 학습시켜 머신러닝 모델을 만들 수 있다. 본 논문에서는 한 학기 분량의 기초 머신러닝 모델 학습 및 활용교육 커리큘럼과 주별 실습 내용을 제시하였다. 그리고, 머신러닝 모델 학습 및 활용에 대한 교육 내용 실체를 실증하기위해, Orange 도구를 활용하여, 분류 데이터 표본과 수치 데이터 표본으로부터 머신러닝 모델을 학습시키고, 모델을 활용하여 모집단의 결과를 예측하는 활용 사례들을 제안하였다. 마지막으로 본 커리큘럼에 대한 교육 만족도를 비전공자 대상으로 조사 및 분석하였다.

II. 기초 머신러닝 모델 학습 및 활용 교육 커리큘럼

본 논문에서는 기초 머신러닝 모델 학습 및 활용 교육 커리큘럼을 개발하기 위해, 그림 1의 ADDIE 교육과정 개발 모형을 적용하였다[11]. 분석(Analysis) 단계의 환경분석에서는

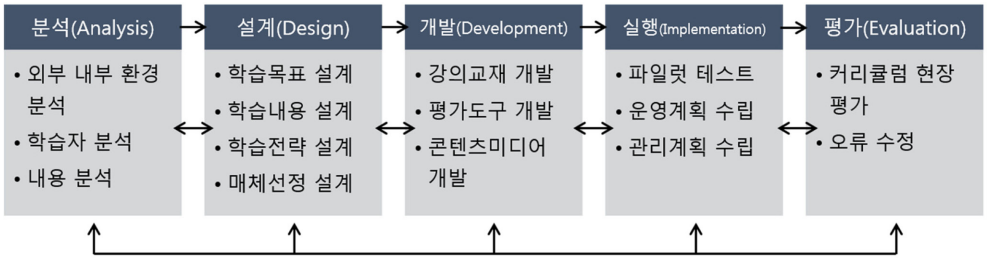

그림 1. '기초 머신러닝 모델 학습 및 활용 교육 커리큘럼' 개발에 적용된 ADDIE 모형의 절차

Fig. 1. Procedure of an ADDIE model applied to the basic machine learning model learning and utilization education curriculum development.

표 1. Orange를 활용한 기초 머신러닝 모델 학습 및 활용 교육 커리큘럼

Table 1. A basic machine learning model learning and application curriculum using Orange

주	강의내용 (주별 3시간)
1주차	• Orange3 소개, 설치, 플랫폼, 기본 사용법 • 머신러닝 모델 종류
2주차	• 실생활문제 적용 사례 1 • 정형데이터와 선형회귀모델을 사용한 데이터 수치 예측
3주차	• 실생활문제 적용 사례 2 • 정형데이터와 로지스틱회귀 모델을 사용한 데이터 분류
4주차	• 실생활문제 적용 사례 3 • 정형데이터와 나이브베이즈 모델을 사용한 데이터 분류
5주차	• 실생활문제 적용 사례 4 • 정형데이터와 K-Means 모델을 사용한 데이터 군집화 분류
6주차	• 실생활문제 적용 사례 5 • 비정형데이터(이미지)와 로지스틱회귀 모델을 사용한 이미지 데이터 분류
7주차	• 실생활문제 적용 사례 6 • 비정형데이터(이미지)와 인공신경망 모델을 사용한 이미지 데이터 식별 분류
8주차	• 실생활문제 적용 사례 7 • 정형데이터와 랜덤포레스트 모델을 사용한 데이터 수치 예측
9주차	• 실생활문제 적용 사례 8 • 정형데이터와 K-NN 모델을 사용한 데이터 분류
10주차	• 실생활문제 적용 사례 9 • 비정형데이터(이미지)와 로지스틱회귀 모델을 사용한 이미지 데이터 분류
11주차	• 실생활문제 적용 사례 10 • 비정형데이터(이미지)와 서포트벡터머신 모델을 사용한 이미지 데이터 분류
12주차	• Orange 도구를 활용하여, 분류 데이터(Categorical Data) 표본으로부터 머신러닝 모델을 학습시키는 사례 • 분류 데이터 모델을 활용하여 모집단의 결과를 예측하는 활용 사례
13주차	• Orange 도구를 활용하여, 수치 데이터(Numerical Data) 표본으로부터 머신러닝 모델을 학습시키는 사례 • 수치 데이터 모델을 활용하여 모집단의 결과를 예측하는 활용 사례
14주차	• 과제 1 수행 – 분류 데이터 표본 수집, 머신러닝 모델 학습 및 활용 과제 1. 분석하고자 하는 분류 데이터 표본 수집 및 데이터 분석목표 설정 2. 수집한 표본 분류 데이터로부터 머신러닝 모델 학습결과 작성 3. 생성한 모델을 활용하여 모집단의 결과를 예측하는 활용 결과 작성
15주차	• 과제 2 수행 – 수치 데이터 표본 수집, 머신러닝 모델 학습 및 활용 과제 1. 분석하고자 하는 수치 데이터 표본 수집 및 데이터 분석목표 설정 2. 수집한 표본 수치 데이터로부터 머신러닝 모델 학습결과 작성 3. 생성한 모델을 활용하여 모집단의 결과를 예측하는 활용 결과 작성

외부 교육 트렌드를 분석하여 머신러닝 모델 학습 및 활용 교육의 필요성을 분석하고, 교육 장소 및 기자재 등 내부 환경을 분석하였다. 분석 단계의 학습자 분석에서는 비전공자들의 사전지식, 학습동기 및 교육 목표의 수준을 분석하였고, 내용 분석에서는 기초 머신러닝 모델 학습 및 활용 교육과정과 가용 여건을 종합적으로 분석하였다.

설계(Design) 단계의 학습목표 설계에서는 기초 머신러닝 모델 학습 및 활용 교육과정이 비전공자 대상으로 실시되는 상황을 정의하고, 본 과목을 이수한 학습자들이 실제적으로 데이터분석 활동을 할 수 있는 구체적인 목표 달성 기준들을 설정하였다. 여기서는 Orange를 이용한 구체적인 데이터분석결과를 제시하도록 목표를 구체화하였다. 학습내용 설계에서는 주 단위 강의 및 평가 내용을 세분화하였다. 학습전략 및 매체선정 설계에서는 프로그래밍 부담을 없애고, 머신러닝 모델을 학습시키고 학습시킨 모델들을 활용하는 워크플로우에 대한 이해도를 높이기 위해, Orange라는 매체도구를 선정하고 공개된 데이터를 활용한 실습과정을 설계하였다.

개발(Development) 단계에서는 교재, 평가도구 및 동영상강의 콘텐츠를 개발하였다. 이에 본 기초 머신러닝 모델 학습 및 활용 커리큘럼에 사용될 실습교재를 선정하였다. 평가도구 부분에서는 머신러닝 모델 학습 및 활용 과제 및 과제 평가 기준을 설계하였고, 주별 동영상강의 콘텐츠를 개발하였다. 실행(Implementation) 단계에서는 교육대상자, 교육일정, 교육과정 및 교육장소 등 운영계획과 강의평가를 통한 개선 관리 계획을 수립하였다. 그리고 기초 머신러닝 모델 학습 및 활용 커리큘럼의 적절성을 검토하기 위해, 한학기 강의를 실행하였다.

평가(Evaluation) 단계에서는 기초 머신러닝 모델 학습 및 활용 커리큘럼에 대해 비전공자 대상 만족도 조사 계획을 수립하였고, 본 논문에 조사결과를 제시하였다. 이러한 현장평가 결과를 바탕으로, 커리큘럼의 효과성을 평가하고 보완하였다. 표 1은 ADDIE 개발 과정을 통해 도출된 15주 기초 머

신러닝 모델 학습 및 활용 커리큘럼 결과를 나타낸다. 표 1의 커리큘럼은 머신러닝 모델 학습 및 활용을 위한 실습 교재를 선택하여 활용한다[12].

III. 분류 데이터 머신러닝 모델 학습 및 활용 교육 사례

표 2는 분류된 표본 데이터를 갖고 머신러닝 모델을 학습시키고, 모집단의 특정 분류 데이터를 예측하는 데 사용된 데이터 속성들을 설명하고 있다. 표본으로 수집된 학습데이터 세트는 A1부터 A15까지 원인에 해당하는 설명변수들을 갖고, 결과에 해당하는 목적변수 B1을 포함한다. A1~A15와 B1, 즉, 16개 값으로 구성된 표본 데이터 인스턴스 수는 50개이며, 모두 1, 2, 3, 4, 5 중 하나의 값을 갖는다. 여기서 1, 2, 3, 4, 5가 의미하는 것은 5개 보기 중 하나를 고른 값 또는 만족도와 같이 각 숫자가 사전에 분류한 특정한 카테고리에 해당한다는 것을 나타낸다.

학습 데이터 세트 중 A1~A15 값들로부터 B1 분류 데이터를 예측하는 머신러닝 모델을 만든다. 그리고 학습된 모델에 모집단의 수에 해당하는 A1~A15 테스트 데이터 인스턴스들을 랜덤하게 발생시킨다. 여기서 가정한 모집단의 수는 400이며, 이에 따라 모집단테스트 랜덤데이터 세트는 모든 A1~A15 값들이 1, 2, 3, 4, 5 중 하나의 값을 균등하게 갖는다. 이러한 모집단테스트 랜덤데이터 세트(A1~A15, 400개)를 학습된 모델에 입력하여, 400개 B1 값들을 예측한다. 여기서 400개 B1 값들을 모아 통계적으로 분석한 것이 모집단의 B1 분류 데이터 특성이다.

그림 2는 Orange로 작성한 분류 데이터 머신러닝 모델 학습 및 예측 워크플로우를 나타낸다. Training Data 위젯을 통해, A1~A15와 B1으로 구성된 50개 표본 학습데이터가 입력

표 2. 분류 데이터 머신러닝 모델 학습 및 예측에 사용된 데이터 속성

Table 2. Data features used to train and predict categorical data machine learning models

데이터 종류	데이터명	속성	데이터 수	할당된 분류 값
학습데이터(표본)	A1~A15	설명변수	50	1, 2, 3, 4, 5
학습데이터(표본)	B1	목적변수	50	1, 2, 3, 4, 5
모집단테스트 랜덤데이터	A1~A15	설명변수	400	1, 2, 3, 4, 5

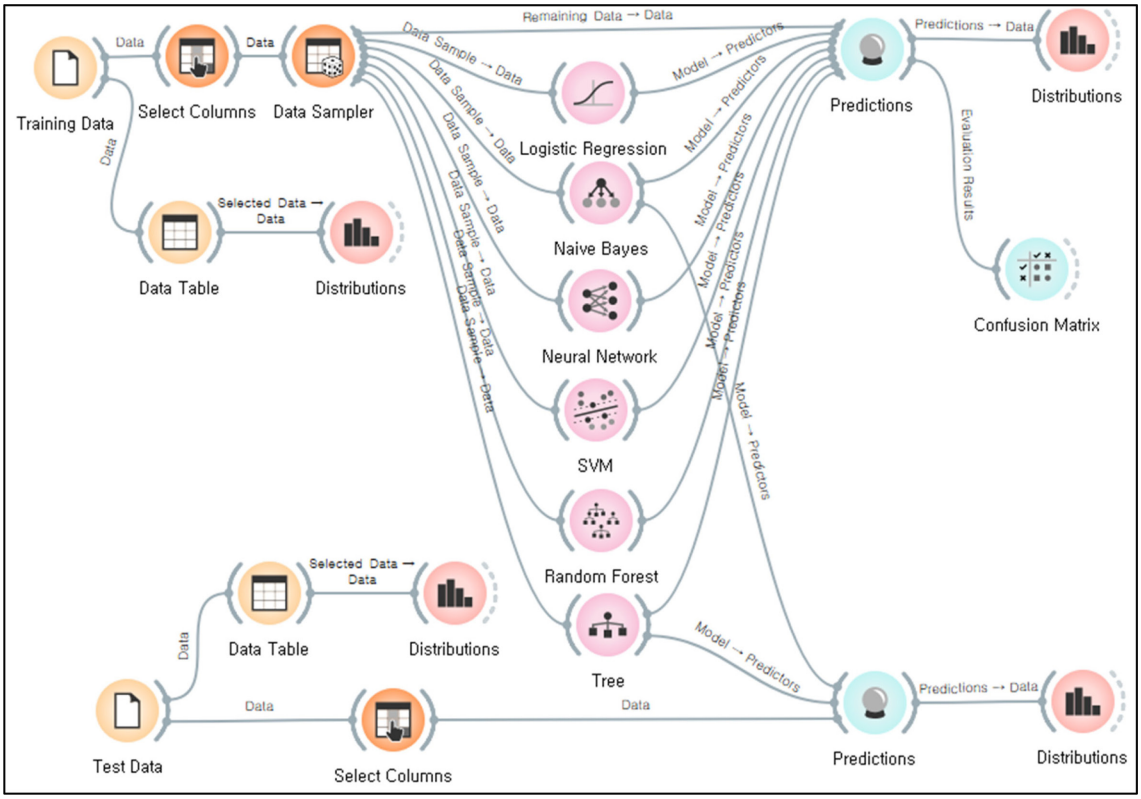

그림 2. 분류 데이터 머신러닝 모델 학습 및 예측 워크플로우

Fig. 2. Categorical data machine learning model training and prediction workflow.

된다. Data Table 위젯을 통해 입력된 학습데이터 세트를 확인할 수 있다. 그리고 이후 연결된 Distribution 위젯을 통해, 위 16개 각 속성에 대한 도수분포 그래프를 확인한다. 학습데이터와 연결된 Select Columns 위젯은 입력된 학습데이터 중 설명변수 세트와 목적변수 1개를 지정한다. Data Sampler 위젯은 50개 학습데이터들 중 머신러닝 모델들에 입력되는 학습용 데이터와 테스트용 데이터의 비율을 결정하고, 실제 학습용 및 테스트용 데이터들을 추출한다. 그림 2에서 설정한 비율은 70%이며, 50개 표본 학습데이터 중 실제 모델 학습에 사용된 데이터 수는 35개, 테스트용 데이터 수는 15개이다.

연속적인 수치 데이터가 아닌 특정 값들로 분류된 데이터들을 분석하는 머신러닝 모델로, 그림 2에서는 Logistic Regression, Naive Bayes, Neural Network, SVM, Random Forest와 Decision Tree 모델들을 사용하였다. 그림 3은 Predictions 위젯에서 6개 모델을 테스트한 CA(Classification Accuracy) 결과를 나타낸다. 그리고 이중 가장 높은 CA를 나타낸 Naive Bayes 모델과 Decision Tree 모델에 대해, Confusion Matix 위젯에서 실제 분류 데이터와 예측한 분류 데이터를 비교한 결과를 나타낸다. 15개 데이터 중 1개를 틀려 CA는 14/15, 즉, 0.933을 나타내었다.

그림 2의 Test Data 위젯을 통해, A1~A15으로 구성된 모집단테스트 랜덤데이터 400개가 입력된다. 이 위젯과 연결된 Select Columns 위젯은 입력된 데이터 중 설명변수 세트 A1~A15를 지정한다. 모집단테스트 랜덤데이터 400개가 Predictions 위젯으로 연결되고, 그림 3에서 가장 높은 CA 성능을 나타낸 Naive Bayes 모델과 Decision Tree 모델도 Predictions 위젯과 연결하였다. 이를 통해, 이 두 모델에 모집단테스트 랜덤데이터 400개를 입력하여 예측한 결과를 표 3에 나타내었다.

표 3에서 두 모델 모두 모집단 400명이 나타내는 목적변수 B1 분류 데이터는 3, 4, 5로 예측되었고, 각 값을 몇 명이 나타낼 지도 예측하였다. 이를 통해, Naive Bayes 모델과 Decision Tree 모델은 모집단 400명이 B1 값으로 평균 4.115와 4.34를 나타낼 것으로 각각 예측하였다. 이 두 모델이 모집단에 대해 예측한 값들 중 어느 모델의 값을 사용할 것인가에 대해, 그림 3에서 15개 데이터에 대해 예측한 결과의 분

표 3. 분류 데이터 모집단 예측 결과

Table 3. Prediction results of population group categorical data

머신러닝 모델	분류데이터 B1 예측값과 도수			기대값 합계	모집단예측 평균
Naive Bayes	3 (135)	4 (84)	5 (181)	1646(400)	4.115
Decision Tree	3 (51)	4 (162)	5 (187)	1736(400)	4.34

Model	AUC	CA
Logistic Regression	0.958	0.867
Naive Bayes	0.958	0.933
Neural Network	0.896	0.867
SVM	0.896	0.800
Random Forest	0.931	0.733
Tree	0.972	0.933

Actual \ Predicted	3	4	5	Σ
3	4	0	0	4
4	1	3	0	4
5	0	0	7	7
Σ	5	3	7	15

그림 3. 분류 데이터 머신러닝 모델 테스트 결과

Fig. 3. Results of categorical data machine learning model test.

포 특성과 유사한 Naive Bayes 모델 분석결과를 모집단 분류 데이터 B1 예측 결과로 선정할 수 있다.

IV. 수치 데이터 머신러닝 모델 학습 및 활용 교육 사례와 커리큘럼 만족도 분석

표 4는 연속성을 갖는 표본 수치 데이터를 갖고 머신러닝 모델을 학습시키고, 모집단의 특정 수치 데이터를 예측하는데 사용된 데이터 속성들을 설명하고 있다. 표본으로 수집된 학습데이터 세트는 A1부터 A6까지 원인에 해당하는 설명변수들을 갖고, 결과에 해당하는 목적변수 B1을 포함한다. A1~A6과 B1, 즉, 7개 값으로 구성된 표본 데이터 인스턴스 수는 120개이며, 모두 0~100 중 하나의 값을 갖는다.

학습 데이터 세트 중 A1~A6 값들로부터 B1 수치 데이터를 예측하는 머신러닝 모델을 만든다. 그리고 학습된 모델에 모집단의 수에 해당하는 A1~A6 테스트 데이터 인스턴스들을 랜덤하게 발생시킨다. 여기서 가정한 모집단의 수는 400이며, 이에 따라 모집단테스트 랜덤데이터 세트는 모든 A1~A6 값들이 0~100 중 하나의 값을 갖고 균등한 분포를 나타내었다. 이러한 모집단테스트 랜덤데이터 세트(A1~A6, 400개)를 학습된 모델에 입력하여, 400개 B1 값들을 예측한다. 여기서 400개 B1 값들을 모아 통계적으로 분석한 것이 모집단의 B1 수치 데이터 특성이다.

그림 4는 Orange로 작성한 수치 데이터 머신러닝 모델 학습 및 예측 워크플로우를 나타낸다. Training Data 위젯을 통해, A1~A6와 B1으로 구성된 120개 표본 학습데이터가 입력된다. Data Table 위젯을 통해 입력된 학습데이터 세트를 확인할 수 있고 Distribution 위젯을 통해, 위 7개 각 속성에 대한 도수분포 그래프를 확인한다. 학습데이터와 연결된 Select Columns 위젯은 입력된 학습데이터 중 설명변수 6개와 목적변수 1개를 지정한다. Data Sampler 위젯은 120개 학습데이터들 중 머신러닝 모델들에 입력되는 학습용 데이터와 테스트용 데이터의 비율을 결정하고, 실제 학습용 및 테스트용 데이터들을 추출한다. 그림 4에서 설정한 비율은 70%이며,

표 4. 수치 데이터 머신러닝 모델 학습 및 예측에 사용된 데이터 속성

Table 4. Data features used to train and predict numerical data machine learning models

데이터 종류	데이터명	속성	데이터 수	수치값 범위
학습데이터(표본)	A1~A6	설명변수	120	0~100
학습데이터(표본)	B1	목적변수	120	0~100
모집단테스트 랜덤데이터	A1~A6	설명변수	400	0~100

120개 표본 학습데이터 중 실제 모델 학습에 사용된 데이터 수는 84개, 테스트용 데이터 수는 36개이다.

연속적인 수치 데이터들을 분석하는 머신러닝 모델로 그림 4에서는 Linear Regression, SVM, Random Forest와 KNN 모델들을 사용하였다. 그림 5는 Predictions 위젯에서 4개 모델을 테스트한 RMSE(Root Mean Square Error) 결과를 나타낸다. 그리고 이중 가장 낮은 RMSE 8.762, 즉, 100을 기준으로 91.24%의 정확도를 나타낸 KNN 모델에 대해, B1 데이터를 예측한 값의 도수 분포를 나타낸다.

그림 4의 Test Data 위젯을 통해, A1~A6으로 구성된 모집단테스트 랜덤데이터 400개가 입력된다. 이 위젯과 연결된 Select Columns 위젯은 입력된 데이터 중 설명변수 A1~A6을 지정한다. 모집단테스트 랜덤데이터 400개가 Predictions 위젯으로 연결되고, 그림 5에서 낮은 RMSE 성능을 나타낸 SVM, Random Forest와 KNN 모델도 Predictions 위젯과 연결하였다. 이를 통해, 이 세 모델에 모집단테스트 랜덤데이터 400개를 입력하여 예측한 결과를 표 5에 나타내었다. 본 데이터는 그림 4의 Feature Statistics 위젯을 통해 추출하였다.

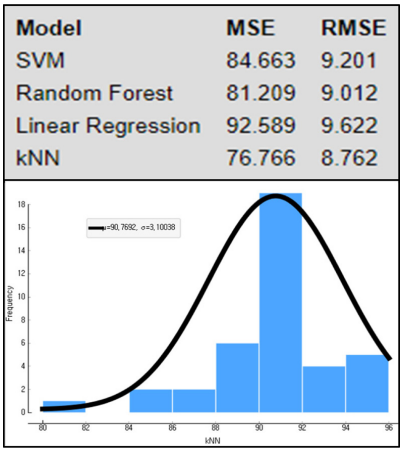

그림 5. 수치 데이터 머신러닝 모델 테스트 결과

Fig. 5. Results of numerical data machine learning model test.

표 5. 수치 데이터 모집단 예측 결과

Table 5. Prediction results of population group numerical data.

머신러닝 모델	수치데이터 B1 예측값(400개)			
	중간값	최소값	최대	평균
SVM	94.23	93.08	95.37	94.25
Random Forest	90.99	88.42	95.40	91.94
KNN	90	88.57	95.71	91.00

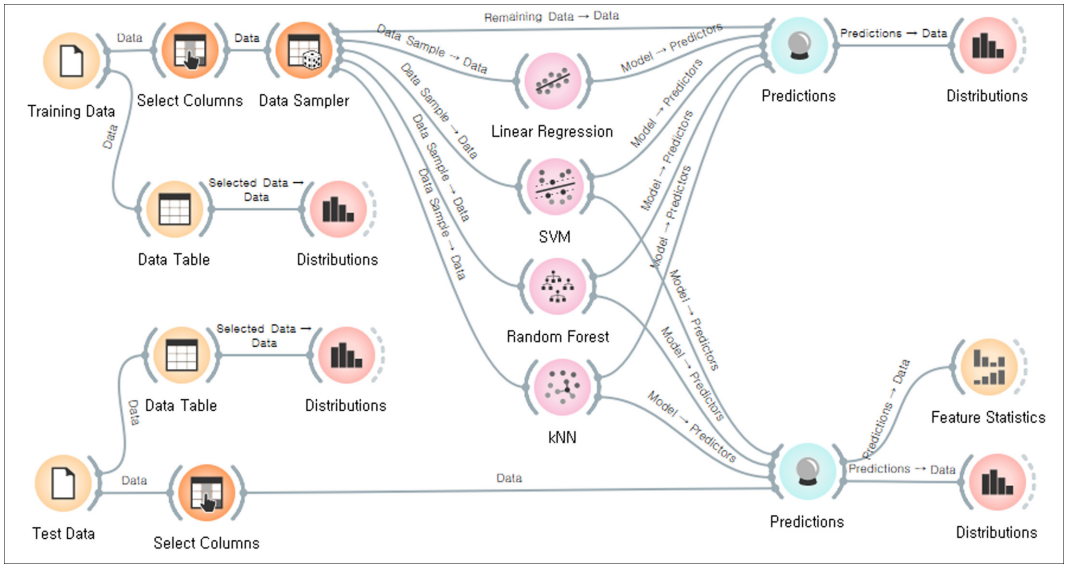

그림 4. 수치 데이터 머신러닝 모델 학습 및 예측 워크플로우

Fig. 4. Numerical data machine learning model training and prediction workflow.

표 6. 비전공자 대상 만족도 조사 결과(60명)

Table 6. Results of satisfaction survey for non-majors

커리큘럼 만족도 평가 영역	매우 불만족 (1점)	불만족 (2점)	보통 (3점)	만족 (4점)	매우 만족 (5점)	5점 척도 평균
실생활문제 적용사례 선정	0%	3%(2)	28%(17)	45%(27)	23%(14)	3.9
수치예측 모델교육	0%	2%(1)	25%(15)	45%(27)	28%(17)	4.0
수치데이터분류 모델교육	0%	3%(2)	27%(16)	38%(23)	32%(19)	4.0
이미지데이터분류 모델교육	0%	2%(1)	30%(18)	38%(23)	30%(18)	4.0
데이터 군집화 모델교육	0%	2%(1)	33%(20)	33%(20)	32%(19)	4.0
분류데이터 모델활용 과제	0%	3%(2)	23%(14)	32%(19)	42%(25)	4.1
수치데이터 모델활용 과제	0%	3%(2)	27%(16)	38%(23)	32%(19)	4.0

표 5와 같이 세 모델은 모집단 400명이 나타내는 목적변수 B1 수치 데이터들을 예측하였고, Random Forest와 KNN 모델은 유사한 최소값, 최대값, 중간값 및 평균값 특성을 나타내었다. 이 세 모델이 모집단에 대해 예측한 값들 중 어느 모델의 값을 사용할 것인가에 대해, 표 5에서 유사한 결과를 나타낸 두 모델, 즉, Random Forest와 KNN 모델이 예측한 통계값을 모집단 수치 데이터 B1 예측 결과로 선정할 수 있다.

표 6은 Orange를 활용한 기초 머신러닝 모델 학습 및 활용교육 커리큘럼의 만족도 결과를 나타낸다. 이 결과는 22년도 2학기 비전공자 학부생 및 현직교원 60명을 대상으로 교육 만족도를 조사한 결과이다. 만족도 평가 영역은 본 커리큘럼에서 제시한 실생활문제 적용사례 선정, 수치예측 모델교육, 수치데이터분류 모델교육, 이미지데이터분류 모델교육, 데이터 군집화 모델교육, 분류데이터 모델활용 과제 그리고 수치데이터 모델활용 과제로 선정하였다. 모든 평가 영역에 대해 60%이상 만족한다는 결과를 나타내었다. 특히, 3장과 4장에서 제안한 교육 사례에 따라 부과한 '분류데이터 모델활용 과제' 및 '수치데이터 모델활용 과제'에 대해서는 70%이상 만족하였다. 표 6에 제시한 만족도 조사 결과에 5점('매우 만족')부터 1점('매우 불만족')까지 점수를 부과하였다. 그리고 각 만족도 평가 영역별로 5점 척도 환산 총점을 산출하고, 전체 60명 평균값을 도출하여 표 6에 나타내었다. 모든 커리큘럼 만족도 평가 영역에서 개인별 평균값이 '만족한다'는 의미를 갖는 4.0 값에 근접한 값을 나타내고 있다. 이 결과는 제안한 커리큘럼이 비전공자 대상 교육에 적합하다는 것을 나타낸다.

V. 결 론

본 논문에서 제안한 기초 머신러닝 모델 학습 및 활용교육 커리큘럼의 특징은 실생활문제 적용 교육, 수치예측 모델교육, 수치데이터분류 모델교육, 이미지데이터분류 모델교육, 데이터 군집화 모델교육, 분류데이터 모델활용 사례교육 및 과제수행 그리고 수치데이터 모델활용 사례교육 및 과제수행이다. 제안한 기초 머신러닝 모델 학습 및 활용교육 커리큘럼은 비전공자들을 위한 대학 교양 교과목에 적합하도록, 프로그래밍 없이 시각적인 위젯을 사용하여, 데이터를 학습시켜 머신러닝 모델을 만드는 도구로 Orange를 사용하였다. 이로 인해 프로그래밍 대신 탐색적 데이터 분석에 집중할 수 있고, 기본값을 통해 데이터 분석 워크플로우의 빠른 프로토타이핑이 가능하다. 본 커리큘럼을 통해 기초 머신러닝 모델 학습 및 활용 이론을 학습한 학생은 파이썬을 활용한 심화 머신러닝 데이터분석 내용을 쉽게 이해할 수 있다. 추후 연구로, 본 커리큘럼에서 소개하지 않은 다양한 머신러닝 모델을 교육하고 심화 데이터분석 방법을 Orange로 교육하는 사례 개발이 가능하다.

참고문헌

[1] Ministry of Science and ICT, "The 4th Industrial Revolution in History," R & D KIOSK, No. 40, September 2017.

[2] Ministry of Science and ICT, "The Various Aspects of the Fourth Industrial Revolution, The Realized Future," R & D KIOSK, No. 41, October 2017.

[3] Ministry of Science and ICT, "Beyond an IT powerhouse to an AI powerhouse," Report Material, 2019.

[4] Joint Ministries, "Artificial Intelligence National Strategy," Report Material, 2019.

[5] Orange, University of Ljubljana, 2023, [Online]. Available: https://orangedatamining.com/.

[6] Y. J. Jang, “Searching for the direction of data science education in the era of the 4th industrial revolution,” *Integrated Humanities Research*, vol. 9, no. 10, pp. 155-180, 2017.

[7] Y. S. Park and S. J. Lee, “Study on the direction of universal big data and big data education-based on the survey of big data experts,” *Journal of the Korean Association of Information Education*, vol. 24, no. 2, pp. 201-214, 2020.

[8] Y. M. Kim and J. H. Kim, “Effect of data science education program using spreadsheet on improvement of elementary school computational thinking,” *Journal of the Korean Association of Information Education*, vol. 21, no. 2, pp. 219-230, 2017.

[9] J. S. Lee, “A study on visualization methods and expressions of information design for big data,” *Basic Formulation Studies*, vol. 14, no. 3, pp. 259-269, 2013.

[10] K. Hur, “Curriculum of basic data science practices for non-majors,” *Journal of Practical Engineering Education*, vol. 12, no. 2, pp. 265-273, 2020.

[11] ADDIE Model, Wikipedia, 2020, [Online]. Available: https://ko.wikipedia.org/wiki/ADDIE _%EB%AA%A8%ED%98%95.

[12] J. S. Lim, B. C. Jang, M. R. S, and J. H. Jeong, “I analyze data with Orange : Learning AI with Orange3,” Seoul : Cmath, 2022.

허 경 (Kyeong Hur)_종신회원

1998년 : 고려대 전자공학과 학사
2000년 : 고려대 전자공학과 석사
2004년 8월 : 고려대 전자공학과 통신공학박사
2004년 8월 ~ 2005년 8월 : 삼성종합기술원(SAIT) 전문연구원
2005년 9월 ~ 현재 : 경인교대 컴퓨터교육과 교수
〈관심분야〉 네트워크 MAC QoS, IoT, SW교육, AI교육, 데이터과학교육