JPEE 실천공학교육논문지 JOURNAL OF PRACTICAL ENGINEERING EDUCATION

실천공학교육

J. Pract. Eng. Educ. 16(4), 409-414, 2024

인공지능(AI) 기반 직업 훈련 평가 데이터 분석 및 취업 예측 프로그램 구현

Implementation of a Job Prediction Program and Analysis of Vocational Training Evaluation Data Based on Artificial Intelligence

천재성, 문일영*

한국기술교육대학교 컴퓨터공학과

Jae-Sung Chun, Il-Young Moon*

Department of Computer Engineering Korea University of Technology and Education, Cheonan 31253, Korea

[요 약]

본 논문은 인공지능(AI)을 활용하여 장애인 직업 훈련 평가 데이터를 분석하고, 다양한 머신러닝 알고리즘을 통해 최적의 예측 모델을 선정하는 연구를 수행한다. 훈련생의 성별, 나이, 학력, 장애 유형, 기초 학습 능력 등의 데이터를 분석하여 취업 가능성이 높은 직종을 예측하고, 이를 바탕으로 맞춤형 훈련 프로그램을 설계하여 훈련 효율성과 취업 성공률을 높이는 것을 목표로 한다.

[Abstract]

This paper utilizes artificial intelligence to analyze vocational training evaluation data for people with disabilities and selects the optimal prediction model using various machine learning algorithms. It predicts the job categories most likely to employ trainees based on data such as gender, age, education level, type of disability, and basic learning abilities. The goal is to design customized training programs based on these predictions to enhance training efficiency and employment success rates.

Key Words: AI-based system, Evaluation data analysis, Job prediction, Machine learning, Vocational training

http://dx.doi.org/10.14702/JPEE.2024.409

Received 26 June 2024; **Revised** 23 July 2024
Accepted 29 July 2024

***Corresponding Author**
E-mail: iymoon@koreatech.ac.kr

I. 서 론

직업훈련은 개인의 경제적 자립을 지원하고 노동 시장의 요구에 맞는 인력을 양성하는 데 필수적인 역할을 한다. 특히 장애인과 같은 취약 계층에게 직업훈련은 사회적 자립과 경제적 독립을 이루는 중요한 수단이며, 사회적 통합을 촉진하고 편견과 차별을 줄이는 데 기여한다. 현재의 직업훈련 프로그램은 다양한 직종과 산업 군에 걸쳐 제공되고 있으나, 일률적인 운영 방식과 개인별 맞춤형 훈련의 부족으로 인해 학습 효과와 교육 효율성이 떨어지는 문제점이 있다. 이를 해결하기 위해서는 학습자의 기본 정보와 사전 평가 결과를 바탕으로 한 맞춤형 훈련 시스템이 필요하다.

코로나19 팬데믹은 노동 시장의 급격한 변화를 초래하여 원격 근무와 자동화가 가속화되고 새로운 직업 군이 등장하였다. 이에 따라 기업은 전 세계에서 인재를 채용할 수 있게 되었고, 기존 직업훈련 프로그램은 이러한 변화를 반영하여 새로운 직업 군에 필요한 역량을 배양하는 방향으로 변화할 필요가 있다. 맞춤형 훈련 프로그램을 통해 학습자의 개인적인 역량과 흥미를 반영한 훈련을 제공하는 것이 중요하다. 인공지능과 빅데이터 분석을 활용하여 학습자의 기초 역량과 학습 목표를 분석하고 이에 맞춘 맞춤형 교육 콘텐츠를 제공함으로써, 변화하는 노동 시장에 적응할 수 있는 유연하고 경쟁력 있는 인재를 양성할 수 있다.

본 연구의 목적은 장애인 직업 훈련 평가 데이터를 활용하여 인공지능(AI) 기반 예측 모델을 개발하고, 이를 통해 훈련생의 취업 가능성을 예측하는 것이다. 다양한 머신러닝 알고리즘을 비교 분석하여 최적의 모델을 선정하고, 훈련생의 성별, 나이, 학력, 장애 유형, 기초 학습 능력 등을 종합적으로 분석하여 취업 가능성이 높은 직종을 예측한다. 이를 통해 훈련생들에게 보다 실질적이고 구체적인 취업 정보를 제공하고, 훈련의 질을 높이고자 한다. 맞춤형 직업훈련 프로그램은 훈련생의 만족도와 취업 성공률을 동시에 높이는 효과를 기대할 수 있으며, 이는 직업훈련의 전반적인 품질과 효율성을 향상시키고, 더 많은 사람들이 성공적으로 취업할 수 있도록 지원하는 데 기여할 것이다. 또한, Flask 프레임워크를 사용하여 직업 훈련 평가 데이터를 기반으로 예측 결과를 제공하는 사용자 친화적인 웹 애플리케이션을 개발하여 훈련생들이 자신의 데이터를 입력하고, 이를 바탕으로 취업처 규모 및 직종을 예측할 수 있도록 설계할 것이다.

결론적으로, 맞춤형 훈련 시스템을 도입하고 인공지능과 빅데이터 분석 기술을 활용하여 직업훈련의 효율성과 효과를 높이는 것이 필요하다. 이는 변화하는 노동 시장에 적응할 수 있는 인재를 양성하는 데 중요한 역할을 할 것이다.

II. 평가 데이터 분석, 모델 개발 및 최적화

A. 데이터 수집 및 전처리

본 연구에서 사용된 데이터는 H직업훈련기관에서 제공받은 개인을 특정 지을 수 없는 1263명의 훈련생의 개인 정보, 훈련 평가 점수, 훈련 후 취업 현황 등을 포함한다. 표 1 데이터는 교육 기관을 통해 수집되었으며, 전처리 과정에서 결측치는 평균값으로 대체되고, 데이터는 표준화되어 머신러닝 모델에 적합한 형태로 조정되었다. 입력데이터는 성별, 연령대, 장애유형 등 16개의 평가 데이터이며 타겟 데이터는 표 2처럼 평가 데이터 중 대분류의 데이터이다.

B. 머신러닝 주요 알고리즘

연구에서는 다양한 머신러닝 모델의 성능을 비교하기 위해 예측 성능이 우수하고 다양한 데이터 유형에 유연하게 대응할

표 1. 장애인 훈련생들 평가 데이터

Table 1. Evaluation data of disabled trainees

연령	장애 유형	최종 학력	분야	국어	영어	수학	우세 (소)	비우세 (소)	양손 (소)	우세 (중)	비우세 (중)	우세 (대)	비우세 (대)	심리	면접	취업처 직종
20	14	5	4	52	50	42	10.9	24.3	26.9	8.3	25	9	20.5	72	60	14
20	14	5	13	46	70	42	40.4	54.4	89.8	81.3	38.5	53.2	60.2	70	47	1
20	14	8	4	63	85	77	91.1	70.4	71.2	81.3	81.5	53.2	75.5	83	78	4
30	14	5	4	52	65	77	40.4	81.3	82.7	56.3	55.8	40.4	83.8	92	66	6
20	13	5	11	100	100	96	95.6	96.6	60.3	92.8	97.6	94.8	99.3	85	73	1
20	4	1	11	100	100	96	91.1	81.3	89.8	81.3	91.8	85.9	83.8	87	72	1
20	13	5	4	100	100	99	81.4	87.7	60.3	56.3	81.5	85.9	95.4	95	72	1
40	2	8	14	54	60	59	83.7	28.1	33.9	84.8	28.1	84.8	26.9	77	60	8

수 있는 랜덤 포레스트, XGBoost, GBM, SVM, 신경망의 다섯 가지 모델을 선정하였다. 이러한 모델들은 각기 다른 접근 방식과 강점을 지니고 있어, 훈련 데이터의 특성을 효과적으로 분석하고 최적의 예측 결과를 도출하는 데 적합하다.

1) 랜덤 포레스트 (Random Forest)

랜덤 포레스트는 배깅(Bagging) 방법을 사용하여 다수의 결정 트리(decision trees)를 학습하고, 각 트리의 예측을 종합하여 최종 결과를 도출하는 앙상블 학습 방법이다[1]. 여기서 배깅이란 Bootstrap Aggregation의 약자로 회귀분석에서 사용되는 기계 학습 알고리즘의 불안정성을 제거하고 분산을 줄이며 안정성과 정확도 예측력을 향상시키기 위해 개발된 개념이다[2].

2) XGBoost (Extreme Gradient Boosting)

XGBoost는 그래디언트 부스팅(Gradient Boosting) 알고리즘을 확장한 것으로, 경사 하강 법을 사용하여 손실 함수를 최소화하는 방향으로 모델을 학습시킨다. XGBoost는 높은 예측 성능과 빠른 학습 속도를 제공하며, 정규화(regularization)를 통해 과적 합을 방지한다[3].

3) Gradient Boosting Machine (GBM)

GBM은 그래디언트 부스팅 알고리즘을 기반으로 하는 앙상블 학습 방법이다. 여러 개의 약한 학습자(weak learners), 혹은 결정 트리(decision trees)를 결합해서 강한 학습자(strong learner)를 결성하는 머신러닝 알고리즘이다[4].

4) Support Vector Machine (SVM)

SVM은 데이터 포인트를 고차원 공간으로 매핑하여, 두 클래스 간의 최대 마진을 찾는 분류 알고리즘이다. SVM은 선형 분류와 비선형 분류 모두에 사용할 수 있으며, 커널 함수(kernel function)를 사용하여 비선형 데이터를 선형적으로 변환할 수 있다[5].

5) 신경망 (Neural Network)

신경망은 인간 뇌의 뉴런 구조를 모방한 알고리즘으로, 다층 퍼셉트론(MLP) 구조를 사용하여 입력 데이터에서 비선형 관계를 학습한다. 신경망은 대규모 데이터와 복잡한 문제 해결에 유용하며, 딥러닝의 기본이 된다. 신경세포(Neuron)를 추상화한 망(Network)이며, 일반적으로 어떠한 형태의 함수라도 근사할 수 있는 통계학적 학습 알고리즘이다[6].

C. 머신러닝 모델 적용

다양한 머신러닝 모델을 적용하여 직업 분류 예측을 수행하였다. 사용된 모델은 랜덤 포레스트, XGBoost, GBM, SVM, 그리고 신경망이며, 모든 모델은 정확도, 정밀도, 재현율, F1 점수, 평균 제곱 오차(MSE), 및 R^2 Score를 포함한 여러 평가 지표를 사용하여 다음 표 3과 같이 성능이 비교되었다.

D. 해석 및 결론

분석 결과, 모든 모델이 32%에서 37% 사이의 정확도를 보였다. 특히, 랜덤 포레스트 모델이 가장 높은 정확도(37.5%)를 기록하며 안정적인 성능을 보였다. 랜덤 포레스트 모델은 주요 특성 중요도를 명확하게 제공하여 데이터 전처리 및 특성 선택에 유리하다. 따라서 랜덤 포레스트 모델을 선택하는

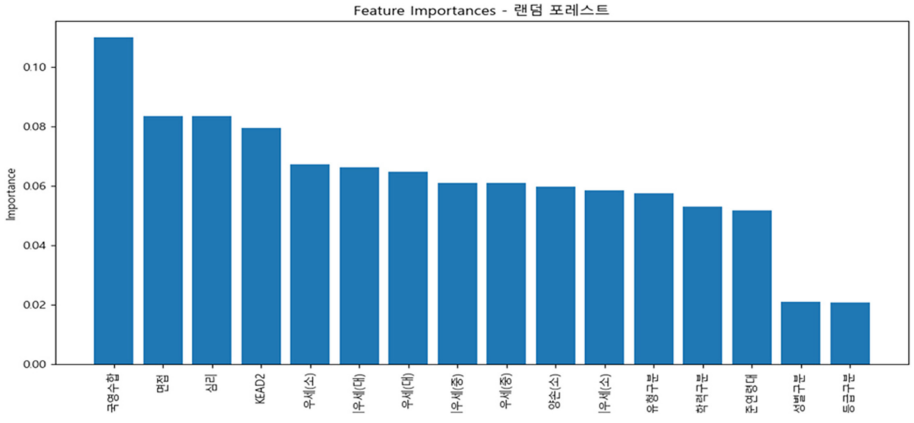

그림 1. 초기 Feature Importances – 랜덤포레스트

Fig. 1. Initial feature importances – random forest.

것이 가장 적합하다고 판단된다.

랜덤 포레스트 모델의 중요도(Feature Importances) 그래프를 해석해보면 (그림 1) 국영수 피처가 가장 높은 중요도를 갖고 있다. 이는 국영수합 점수가 취업 예측에서 가장 큰 영향을 미친다는 것을 의미한다. 이어 면접과 심리 점수 순으로 높은 영향을 미치는 것으로 알 수 있다.

III. 성능 평가

A. 데이터 정제

처음에는 1263명의 데이터를 사용하여 머신러닝 모델을 평가하였다. 그러나 데이터의 세부 분석을 통해 몇 가지 문제점을 발견하였다.

1) 점수가 0점이거나, 의미 없는 점수들이 포함된 데이터 존재함.
2) 특정 피처(장애등급구분, 성별 구분, 입학기준연령대)가 모델의 예측 성능에 큰 영향을 미치지 않음.

따라서 이러한 불필요한 데이터를 제거하고 640명의 유효한 데이터를 남겼다. 또한, 타겟 값에서 '교육&공공행정'과 '정보기술&통신'은 고수준의 교육 및 스킬을 요구하는 점이 공통되기에 통합하여 표 2처럼 기존 5개의 타겟에서 총 4개의 타겟으로 재설정 하였다. 변경된 타겟 값은 다음 표 4와 같다.

B. 데이터 전처리

범주형 변수와 수치형 변수를 분리하여 각각 적합한 전처리 과정을 수행하였다. 파이프라인을 사용하여 데이터 전처리를 일관되게 처리하였다.

C. 모델 성능 비교

다섯 가지 머신러닝 모델(랜덤 포레스트, XGBoost, GBM, SVM, 신경망)을 사용하여 초기 1263개의 데이터로 실험을 진행한 결과, 랜덤 포레스트 모델이 가장 높은 정확도를 기록하였다. 그러나 데이터 정제 후 640개의 데이터를 사용한 실험에서는 모델의 성능이 다음 표 5와 같이 향상되었다.

1) 성능지표
   a) 정확도(Accuracy) : 모델이 올바르게 예측한 비율.
   b) 정밀도(Precision) : 모델이 긍정으로 예측한 것 중 실제 긍정의 비율.
   c) 재현율(Recall) : 실제 긍정 중 모델이 긍정으로 예측한 비율.
   d) F1 점수(F1 Score) : 정밀도와 재현율의 조화 평균.
   e) 평균 제곱 오차(MSE) : 예측 값과 실제 값 간의 차이를 제곱하여 평균낸 값.
   f) R^2 점수(R^2 Score) : 모델의 설명력(예측의 분산 비율).

표 2. 타겟 데이터(대분류 데이터)

Table 2. Taget data(Major Categories)

타겟	제조&건설업	교육&공공행정	의료&사회 복지	정보기술&통신	서비스&기타
값	1	2	3	4	5

표 3. 모델 성능 비교 및 특성 중요도

Table 3. Model performance comparison and feature importance

모델	정확도	MSE	R^2	정밀도	재현율	F1-점수	주요 특성 중요도
랜덤포레스트	37.5%	3.28	−0.50	0.358	0.375	0.336	국영수합 (0.10), 심리 (0.084), 면접 (0.083), KEAD2 (0.077)
XGBoost	34.4%	–	–	0.343	0.343	0.329	장애유형구분 (0.097), 비우세(대) (0.065), 성별구분 (0.063), 양손(소) (0.063)
GBM	37.2%	–	–	0.365	0.371	0.353	국영수합 (0.1387), 면접 (0.1004), 장애유형구분 (0.0957), 심리 (0.0862)
SVM	32.0%	–	–	0.193	0.320	0.239	장애유형구분 (0.501), 비우세(소) (0.492), 우세(대) (0.391)

표 4. 변경된 타겟 값

Table 4. Modified target values

타겟	제조&건설업	교육&공공행정&정보기술&통신	의료&사회 복지	서비스&기타
값	1	2	3	4

표 5. 데이터 정제 후 결과 값

Table 5. Results after data cleaning

지표	초기 실험	데이터 정제 후
정확도	39.1%	86.72%
MSE	3.51	0.703125
R^2 점수	−0.61	0.0473
정밀도	36.2%	88.55%
재현율	34.3%	86.72%
F1−점수	32.9%	83.78%

D. 온라인 평가 예측 시스템 구축

1) 시스템 기능

온라인 평가 예측 시스템의 설계는 사용자 친화적인 웹 인터페이스를 제공하고, 훈련생의 입력 데이터를 기반으로 예측 결과를 제공하는 것을 목표로 한다. 시스템은 데이터 입력 폼, 예측 결과 시각화, 사용자 맞춤형 피드백과 같은 주요 기능을 포함한다.

2) Flask 기반 시스템 구현

Flask는 파이썬의 micro framework이며 웹 서버 게이트웨이 인터페이스(WSGI)인 Werkzeug와 템플릿 언어인 Jinja2를 기반으로 만들어졌다. 확장성이 좋고, 다양한 엔진이나 다른 컴포넌트를 같이 사용할 수 있기에 많이 사용되어 진다[7]. 이런 특성으로 평가 데이터로부터 머신러닝 모델을 생성하고, 이를 사용자에게 제공하는 백엔드로 사용된다.

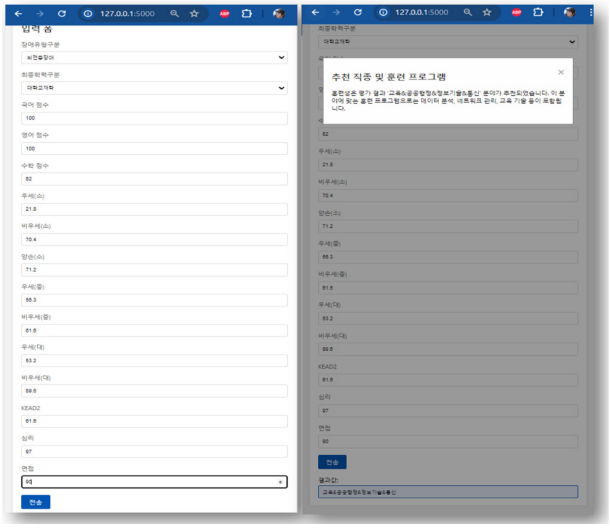

그림 2. 평가시스템 결과 화면

Fig. 2. Evaluation system result screen.

3) 시스템 결과

그림 2 화면처럼 Flask 기반 웹사이트에서 데이터를 입력하고 전송 버튼을 눌렀을 때, 랜덤포레스트 모델로 학습된 예측 결과가 출력되었음을 확인하였다.

IV. 결 론

본 연구는 장애인 직업훈련 평가 데이터를 활용하여 머신러닝 기반의 취업 예측 모델을 개발하고, 이를 통해 훈련생에게 맞춤형 훈련 프로그램을 제공하는 시스템을 구축하는 것을 목표로 하였다. H직업훈련 기관에서 제공된 1263개의 데이터를 수집하고, 결측치와 의미 없는 값을 제거하여 640개의 유효한 데이터로 정제하였다. 다양한 머신러닝 알고리즘을 사용하여 취업 예측 모델을 개발하고 성능을 비교한 결과, 랜덤 포레스트 모델이 가장 높은 성능을 보였다. 랜덤 포레스트 모델은 다양한 변수 처리 능력, 과적합 방지, 변수 중요도 제공, 높은 예측 성능 등의 장점을 지니고 있으며, 정확도 86.72%, 정밀도 88.55%, 재현율 86.72%, F1-Score 83.78%를 기록하여 다른 모델들보다 우수한 성능을 나타냈다. 이러한 이유로 랜덤 포레스트 모델이 장애인 직업훈련 평가 데이터를 기반으로 취업 예측을 수행하는 시스템에서 가장 적합한 알고리즘으로 선정되었다. 또한, Flask를 기반으로 한 웹 애플리케이션을 개발하여 훈련생이 자신의 데이터를 입력하면 예측 결과를 제공받을 수 있도록 하였으며, 예측 결과에 따라 맞춤형 훈련 프로그램을 추천하는 기능을 포함하였다. 결론적으로, 본 연구는 맞춤형 훈련 시스템을 도입하고 인공지능과 빅데이터 분석 기술을 활용하여 직업훈련의 효율성과 효과를 높이는 방안을 연구하였다. 이를 통해 장애인 훈련생들에게 보다 구체적이고 실질적인 취업 정보를 제공하고, 그들의 성공적인 취업을 지원할 수 있을 것으로 기대된다.

참고문헌

[1] S. H. Moon, “Head pose estimation by using histogram and random forest,” Master’s thesis, Jeonnam National University, Gwangju, p. 13, 2016.

[2] J. H. Yoo, “Peak load forecasting method for Jeju Island on alternative holiday using random forest,” Master’s thesis, Graduate School of Engineering, Korea University, p. 16, 2023.

[3] T. Chen and C. Guestrin, “XGBoost: A scalable tree boosting system,” *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining,* pp. 785-794, 2016.

[4] S. Y. Park, J. H. Baek, S. H. Park, and J. Hur, “Implementation of a short-term wind power output forecasting model based on gradient boosting machine(GBM) algorithms,” *Proceedings of the Korean Institute of Electrical Engineers Conference*, Gyeongbuk, pp. 305-306, May 26, 2022.

[5] J. H. Kim and J. M. Won, “A development of the road surface decision algorithm using SVM (Support Vector Machine) clustering methods,” *Journal of Korean ITS Society*, vol. 12, no. 5, pp. 1-12, 2013.

[6] H. Choi, T. K. Kim, G. R. Heo, S. D. Choi, and J. W. Hur, “Study of fuel pump failure prognostic based on machine learning using artificial neural network,” *Journal of the Korean Society of Manufacturing Process Engineers,* vol. 18, no. 9, pp. 52-57, 2019.

[7] J. Heo, W. K. Choi, J. M. Son, H. C. Park, and D. G. Yoon, “A study on performance improvement with minio file server on flask web server,” *Proceedings of the Korean Institute of Communication and Information Sciences Conference,* Jeju, pp. 914-915, June 20, 2018.

천 재 성 (Jae-Sung Chun)_정회원

2011년 2월 : 한국기술교육대학교 컴퓨터공학부 졸업
2022년 8월 ~ 2024년 8월 : 한국기술교육대학교 컴퓨터공학과 석사
2022년 8월 ~ 현재 : 한국기술교육대학교 컴퓨터공학과 박사 과정
〈관심분야〉 웹, 앱, AI, 빅데이터, 직업훈련

문 일 영 (Il-Young Moon)_종신회원

2005년 2월 : 한국항공대학교 항공통신정보공학과 공학박사
2005년 3월 ~ 현재 : 한국기술교육대학교 컴퓨터공학부 정교수
〈관심분야〉 AI, 무선인터넷 응용, 무선 인터넷, 모바일 IP