JPEE 실천공학교육논문지
JOURNAL OF PRACTICAL ENGINEERING EDUCATION

에듀테크

J. Pract. Eng. Educ. 16(3), 291-296, 2024

직업훈련생 평가 데이터와 취업 결과의 상관관계: 머신러닝 모델을 통한 예측 방안 연구

Correlation between Vocational Training Evaluation Data and Employment Outcomes: A Study on Prediction Approaches through Machine Learning Models

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 연구는 장애인 직업훈련생의 사전 평가 데이터를 활용하여 직업 훈련 후 취업 결과를 예측하는 다양한 머신러닝 모델을 분석하였다. 연구는 훈련생의 성별, 연령, 장애 유형 등을 포함하는 다양한 개인적 특성을 포함한 데이터 세트에 기반하여, 가장 적합한 머신러닝 모델들을 선별하고 활용하였다. 이러한 분석을 통해, 사전 평가 데이터만을 사용하여 장애인 훈련생의 취업률 및 직업 만족도 향상을 목적으로 한다. 결과적으로, 장애인뿐만 아니라 다양한 배경을 가진 직업훈련생들에게도 적용할 수 있는 범용적인 접근법을 제시한다. 이는 맞춤형 직업 훈련 프로그램의 개발과 구현에 중요한 기여를 할 것으로 기대되며, 궁극적으로는 더 나은 취업 결과와 직업 만족도를 달성하는 데 도움이 될 것이다.

[Abstract]

This study analyzed various machine learning models that predict employment outcomes after vocational training using pre-assessment data of disabled vocational trainees. The study selected and utilized the most appropriate machine learning models based on a data set containing various personal characteristics, including trainees' gender, age, and type of disability. Through this analysis, the goal is to improve the employment rate and job satisfaction of disabled trainees using only pre-assessment data. As a result, it presents a universal approach that can be applied not only to people with disabilities, but also to vocational trainees from a variety of backgrounds. This is expected to make an important contribution to the development and implementation of tailored vocational training programs, ultimately helping to achieve better employment outcomes and job satisfaction.

Key Words: Big data analysis, Disability education, Employment outcomes, Linear regression, Logistic regression, Machine learning, Neural networks, Random forest, Vocational training, XGBoost

http://dx.doi.org/10.14702/JPEE.2024.291

Received 3 April 2024; **Revised** 22 April 2024
Accepted 20 May 2024

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

I. 서 론

코로나19 팬데믹 이후의 노동 시장 변화는 개인의 기술 적응력 및 평생 학습 능력의 중요성을 더욱 강조하고 있다[1]. 특히, 사회적 및 경제적으로 취약한 집단에 속하는 개인들에 대한 지원 강화의 필요성이 커지고 있는 이 시점에서, 효과적인 직업 훈련 프로그램의 설계 및 실행은 더욱 절실해졌다. 이러한 배경 하에, 장애인을 대상으로 한 맞춤형 직업 훈련 프로그램의 설계와 평가는 특히 중요한 과제로 부상하였다.

해당 연구에서는 장애인직업훈련 기관에서 근무하면서 수집된 321명의 장애인 훈련생들의 데이터를 분석하여, 다양한 기업의 규모와 직종 유형에 영향을 미치는 주요 요인들을 심층적으로 분석하고자 한다. 이 데이터는 훈련생들의 개인적 특성 및 훈련 전 평가 항목 점수를 포함하고 있으며(표 1), 이를 기반으로 다양한 머신러닝 모델을 활용하여 훈련생들의 취업처와 직종 유형을 예측한다. 머신러닝은 사전에 프로그램 되어 있지 않은 컴퓨터가 데이터로부터 패턴을 학습하고 이후 해롭게 입력되는 데이터에 대해 적절한 작업을 수행하는 일련의 처리 과정이다[2].

사용된 기계 학습 모델은 선형 회귀[3], 랜덤 포레스트[4], XGBoost[5], 신경망[6] 및 로지스틱 회귀[7]를 포함한다. 각 모델의 성능은 교차 검증을 통해 평가되며, 이를 통해 가장 적절한 모델을 선정한다. 모델링 단계는 물론, 데이터 전처리 및 변수 선택 과정 또한 중요한 역할을 하며, 최종적으로는 이 모든 과정을 통해 장애인 훈련생들의 취업 성공률을 높일 수 있는 방법을 찾을 수 가 있다.

이번 연구는 장애인 직업 훈련생을 대상으로 진행되었지만, 그 결과와 방법론은 모든 직업 훈련 프로그램에 적용될 수 있는 범용성을 지닌다. 이는 다양한 배경을 가진 개인들에게 적합한 직업 훈련 프로그램을 설계하고, 이를 통해 더 나은 취업 결과와 개인의 직업 만족도를 달성하려는 연구의 목적을 반영한다. 따라서 이 연구의 접근 방식과 결과는 향후 다양한 직업 훈련 프로그램의 개선과 평가에 중요한 기초 자료를 제공할 것이며, 모든 훈련생이 시장에서 요구하는 기술과 역량을 개발하고, 평생 학습의 경로를 걷도록 지원하는 데 기여할 것으로 기대된다.

II. 이론적 배경

A. 데이터 처리와 기계 학습

데이터 처리는 모든 기계 학습 프로젝트의 핵심이다. 본 연구에서 사용된 pandas, numpy, 및 openpyxl 라이브러리는 대량의 데이터를 효과적으로 처리하고 분석하는 필수 도구이다. 이러한 도구들은 장애인 직업 훈련생 데이터의 관리와 전처리에 필수적이며, SimpleImputer를 사용한 결측치 처리는 데이터 분석의 정확성을 보장하고 데이터의 품질을 향상시킨다.

B. 모델 선택과 평가

선형 회귀, 랜덤 포레스트, 로지스틱회귀, XGBoost 및 신경망 모델과 같은 다양한 기계 학습 알고리즘을 활용한다. 이 모델들은 장애인 훈련생의 취업 가능성 및 직종 유형을 예측하며, 직업 훈련 프로그램의 개선과 개인화에 중요한 정보를 제공한다.

표 1. 장애인 훈련생들의 데이터

Table 1. Data from disabled trainess

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 규모	취업처 직종
20	14	5	4	52	50	42	10.9	24.3	26.9	8.3	25	9	20.5	72	60	3	14
20	14	5	13	46	70	42	40.4	54.4	89.8	81.3	38.5	53.2	60.2	70	47	3	1
20	14	8	4	63	85	77	91.1	70.4	71.2	81.3	81.5	53.2	75.5	83	78	1	4
30	14	5	4	52	65	77	40.4	81.3	82.7	56.3	55.8	40.4	83.8	92	66	3	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	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	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	1
40	2	8	14	54	60	59	83.7	28.1	33.9	84.8	28.1	84.8	26.9	77	60	0	8

C. 데이터 전처리와 인코딩

데이터 전처리 과정에는 StandardScaler 및 OneHotEncoder를 사용하여 수치 데이터의 스케일링 및 범주형 데이터의 변환이 포함된다. 이는 모델이 데이터를 더 잘 이해하고 학습 과정에서의 성능을 향상시킨다. 데이터의 결측치는 평균값으로 대체되어 데이터 세트의 일관성을 유지하고 모델 학습의 효율성을 보장한다.

D. 딥러닝과 신경망

TensorFlow를 사용한 신경망 구축은 복잡한 데이터 패턴을 학습할 수 있는 능력을 제공한다. 다양한 차원의 데이터에서 중요한 특성을 추출하고, 이를 통해 훈련 프로그램의 효과를 분석하며, 향후 훈련 방향성을 제시할 수 있다.

E. 모델 평가 지표

모델의 성능 평가는 mean_squared_error, r2_score, accuracy_score 및 classification_report와 같은 지표를 사용하여 수행된다. 이 지표들은 모델의 예측 정확도, 적합성 및 오류 수준을 평가하며 모델 개선 방향을 제시한다.

표 2에서와 같이 데이터 처리, 결측치 처리, 데이터 스케일링, 데이터 인코딩, 기계학습모델, 로지스틱 회귀, XGBoost, 신경망, 모델평가 지표를 정리하였다.

III. 연구방법

A. 데이터 전처리

본 연구에서는 데이터 전처리 과정을 통해 데이터의 질을 향상시킨다. 장애유형과 (표 3), 최종학력 (표 4)은 레이블인코딩으로 처리하고, 취업처 대분류를 범주별로 분류하여 숫자데이터로 매핑한다(표 5). SimpleImputer를 활용하여 결측치가 있는 데이터를 평균값으로 대체하며, 이는 데이터의 일관성 및 신뢰성을 보장한다. 이어서, train_test_split 함수를 이용하여 데이터를 훈련 세트와 테스트 세트로 분할한다. 이 분할은 모델의 학습과 평가를 분리하여 신뢰도 있는 성능 평가를 가능하게 한다.

B. 모델 학습 및 평가

연구는 다음과 같은 다양한 기계 학습 모델을 사용하여 진행된다.

1) 선형 회귀(Linear Regression): 이 모델은

$$Y = \beta 0 + \beta 1X1 + ... + \beta nXn + \varepsilon$$

의 수식을 기반으로 작동한다. 모델의 성능은 MSE와 R^2지표를 사용해 평가된다.

2) 랜덤 포레스트(Random Forest): 여러 개의 결정 트리를 통합하여 사용한다. 평균 예측값 또는 다수의 투표로 최종

표 2. 데이터기술 및 모델요약

Table 2. Summary of data techniques and models

	데이터처리	결측치처리	데이터스케일링	데이터인코딩	기계학습모델	로지스틱회귀, XGHBoost	신경망	모델평가
기술/모델	Pandas, numpy, openpuxl	SimpleImputer	StandardScaler	OneHotEncoder	선형회귀, 랜덤포레스트	분류모델해결	TensorFlow 신경망 구축	Accuracy_score 등
목적/적용	data입력, 전처리	data정확성 및 품질향상	수치데이터 정규화	범주형데이터 변환	취업가능성 및 직종유형예측		복잡한 패턴학습및 특성추출	모델예측 정확도 및 적합성 평가

표 3. 장애유형별 레이블인코딩

Table 3. Label encoding by disability type

구분	간	국가유공	뇌병변	뇌전증	시각	신장	심장	안면	언어	자폐	장루요루	정신	지적	지체	청각	호흡기
코드	0	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15

표 4. 최종학력 레이블인코딩

Table 4. Label encoding by highest level of education

구분	무학	초졸	중졸	고교재학	고교중퇴	고졸	초대재학	초대중퇴	초대졸	대학재학	대학중퇴	대졸	석사중퇴	석사졸업	박사졸업	특수학교고교재학	특수학교고교졸업
코드	0	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16

표 5. 머신러닝 모델별 결과값(취업처 규모)

Table 5. Major categories of places of employment

코드	취업처 직종
1	제조업, 기타 반도체소자 제조업
2	교육서비스업, 예술, 스포츠 및 여가관련 서비스업, 숙박 및 음식점업, 의료, 서양식 음식점업, 의료기관, 한식 일반 음식점업
3	응용 소프트웨어 개발 및 공급업, 정보통신업, 소프트웨어사업자(패키지소프트웨어개발.공급사업), 정보통신업
4	물류, 운수 및 창고업, 교통 항공, 택배업
5	종합 건설업, 토목건축공사업, 건설업, 부동산업, 건 설 업, 건 설 업
6	시장조사 및 여론조사업, 교육 서비스업, 학술.연구용역, 일반 교과 학원
7	금융, 금융 및 보험업
8	국가기관, 공공 행정, 국방 및 사회보장 행정
9	의료기관세탁물처리업
10	도매 및 소매업, 전자상거래업, 통신 판매업
11	시설경비업, 보안시스템 서비스업
12	수도, 하수 및 폐기물 처리, 원료 재생업
14	사업시설 관리, 사업 지원 및 임대 서비스업, 사업시설 유지 · 관리 서비스업, 건물 및 산업설비 청소업, 사업시설 관리 및 조경 서비스업
15	출판, 영상, 방송통신 및 정보서비스업, 포털 · 컨텐츠 · 커뮤니티
16	증기, 냉 · 온수 및 공기조절 공급업, 전기, 가스, 증기 및 공기 조절 공급업
17	보건업 및 사회복지 서비스업, 중증장애인생산품시설, 노인요양시설
18	기타 엔지니어링 서비스업, 엔지니어링사업(도로, 공항)
19	일반 통신 공사업
20	전문, 과학 및 기술 서비스업, 기타자유업종, 서비스, 그 외 기타 분류 안된 사업지원 서비스업, 전문, 과학 및 기술서비스업, 협회 및 단체, 수리 및 기타 개인서비스, 협회 및 단체, 수리 및 기타 개인 서비스업, 협회 및 단체, 수리 및 기타 개인 서비스업, 상용 인력 공급 및 인사관리 서비스업, 승강기설치공

결정을 내리는 방식으로 평가된다.

3) XGBoost: Gradient Boosting 기술을 기반으로 하며, 다양한 문제에 대한 뛰어난 예측 능력과 확장성을 가진 결정 트리 기반의 앙상블 학습 알고리즘이다. 목적 함수를 최소화하려고 시도한다.

4) 신경망(Neural Networks): 이는 입력층, 은닉층, 출력층으로 구성되며, 각 층은

$$f(x) = Wx + b$$

위 수식을 따른다. 모델의 정확도를 통해 성능을 평가한다.

5) 로지스틱 회귀(Logistic Regression): 이 모델은 신경망 모델의 변수 간 인과관계를 규명하지 못하는 한계점을 보완해 주는 방법이다.

$$\log it(p) = \beta 0 + \beta 1X1 + ... + \beta nXn$$

위 수식을 사용하여 분류 문제를 해결한다. 정확도와 분류 리포트를 통해 모델의 성능을 평가한다.

IV. 연구 결과

본 연구는 321명 장애인 직업 훈련생들의 데이터를 바탕으로, '취업처 규모' 및 '취업처 직무' 예측에 관한 다양한 머신러닝 모델의 성능을 비교 분석하였다. 연구의 목적은 훈련생들의 취업 결과를 보다 정확히 예측할 수 있는 모델을 찾아내어, 직업 훈련 프로그램의 설계 및 취업 매칭 서비스의 개선에 기여하는 것이다.

'취업처 규모' 예측을 위해서, 선형회귀, 랜덤포레스트, 신경망, 로지스틱회귀 및 XGBoost와 같은 모델들이 평가되었다. 분석 결과에 따르면, **XGBoost 모델이 약 58.5%의 정확도로 가장 우수한 성능을 보였다**(표 6). 이러한 결과는 XGBoost의 강력한 데이터 처리 능력과 복잡한 상호작용을 효과적으로 모델링할 수 있는 능력 때문으로 해석된다. XGBoost 모델은 특히 대규모 데이터셋 및 다차원의 변수 간 상호작용을 다룰 때 뛰어난 성능을 발휘하는 것으로 알려져 있다.

표 6. 머신러닝 모델별 결과값(취업처 규모)

Table 6. Results for each machine learning model (size of employment)

모델	정확도	MSE	R^2	정밀도	재현율	F1-점수
선형회귀	-	1.525	-0.022	-	-	-
랜덤포레스트	-	1.468	0.016	-	-	-
신경망	0.474	-	-	0.39	0.47	0.42
로지스틱 회귀	0.508	-	-	0.45	0.51	0.46
XGBoost	0.585	-	-	0.55	0.58	0.56

표 7. 머신러닝 모델별 결과값(취업처 규모)

Table 7. Results for each machine learning model (employment location)

모델	정확도	MSE	R^2	정밀도	재현율	F1-점수
선형회귀	-	42.164	0.144	-	-	-
랜덤포레스트	-	44.428	0.098	-	-	-
신경망	0.340	-	-	0.27	0.34	0.28
로지스틱 회귀	0.369	-	-	0.29	0.37	0.32
XGBoost	0.323	-	-	0.33	0.32	0.31

반면, **선형회귀 모델은 R^2 점수가 음수인 -0.022081554830588 38를 기록하여, 본 데이터 세트에 대한 낮은 설명력을 보였다.** 이는 선형회귀 모델이 본 연구에서 사용된 데이터의 복잡성과 다양성을 충분히 반영하지 못했음을 의미한다.

'취업처 직무' 예측 분야에서는 **Logistic Regression 모델이 36.92%의 정확도로 비교적 우수한 성능을 보였으나**, 이는 여전히 낮은 수치이다(표 7). 이 결과는 '취업처 직무'의 예측이 '취업처 규모' 예측보다 더 많은 변수와 더 복잡한 데이터 구조를 요구할 수 있음을 시사한다. 따라서 이 예측 분야는 더 많은 변수의 추가나 다른 모델링 접근 방식을 요구할 수 있다.

신경망 모델은 특히 낮은 정확도 0.340을 보여, 해당 모델의 구조와 훈련 방법에 대한 재평가가 필요함을 보여준다. 신경망의 성능은 적절한 데이터 전처리, 적합한 아키텍처 선택, 충분한 훈련 데이터의 확보 등 다양한 요소에 의해 영향을 받는다.

이러한 분석 결과는 현재 사용된 모델의 성능 개선이 필요하다는 것을 강조한다. 특히, 데이터의 불균형한 분포와 복잡한 모델 구조에 대한 고려는 예측 성능을 향상시키기 위해 중요하다. 또한, 하이퍼파라미터 최적화를 위한 교차 검증 및 그리드 탐색 같은 방법은 모델의 예측 능력을 높이는 데 중요한 역할을 할 수 있다.

또한, 취업처 규모 및 직무 분야 예측을 위한 변수 선택과 데이터 전처리의 중요성을 강조한다. 취업 성공 예측에 영향을 미치는 다양한 요인들을 정확히 식별하고 이를 모델에 통합하는 것은 높은 예측 정확도를 달성하기 위해 필수적이다. 이 과정에서, 변수 간의 상호작용, 중요도 분석 및 다차원 데이터의 처리 등이 중점적으로 고려되어야 한다.

V. 결 론

이번 연구를 통해, 장애인 직업 훈련생들의 취업처 규모 및 직종 분야 예측에 있어서 XGBoost와 Logistic Regression 모델이 유용하게 적용될 수 있음을 확인하였다. 특히, XGBoost 모델은 취업처 규모 예측에서 높은 정확도를 보임으로써, 복잡한 데이터 구조와 다양한 변수 간 상호작용을 효과적으로 처리할 수 있음을 보여주었다. 이는 향후 장애인뿐만 아니라 다양한 배경을 가진 직업 훈련생들의 취업 지원 서비스 개선에 중요한 시사점을 제공한다.

하지만, 한계점으로는 사용된 데이터 세트의 크기가 상대적으로 작다는 점이 있으며, 이는 모델의 신뢰성과 일반화 가능성에 영향을 줄 수 있다. 따라서, 향후 연구에서는 보다 다양하고 방대한 데이터를 수집 및 활용하여 모델을 재평가하고, 추가적인 변수 선택과 하이퍼파라미터 최적화를 통해 모델의 예측 성능을 더욱 향상시킬 필요가 있다.

결론적으로, 본 연구의 결과와 접근 방식은 장애인 직업 훈련생들에게 국한되지 않고, 다양한 배경을 가진 모든 직업 훈련생들의 취업 성공률을 향상시키는 데 기여할 수 있다. 이를 통해, 개인의 능력과 적성에 맞는 맞춤형 직업 훈련 기

회를 제공하고, 궁극적으로 더 나은 취업 결과와 직업 만족도를 달성하는 데 도움이 될 것으로 기대된다.

참고문헌

[1] J. H. Lee and K. L. Cho, “A study on the prediction model for the ratio of mathematics low-performing students in middle school using machine learning,” *Journal of Educational Technology,* vol. 37, no. 1, pp. 95-129, 2021. DOI: 10.17232/KSET.37.1.095

[2] The Hankyoreh, “How COVID-19 Has Transformed the Labor Market,” The Hankyoreh, 2021. Available from: https://www.hani.co.kr/arti/economy/economy_general/1028657.html.

[3] S. E. Kim, K. B. Kim, W. Y. Kim, and J. H. Jin, “Inference of drone VLOS distance using machine learning,” *Proceedings of the Korean Society of Intelligent Transportation Systems Fall Conference*, pp. 297-299, 2023.

[4] E. J. Lee, Y. S. Song, J. H. Kim, and S. H. 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.

[5] S. M. Cho, “Modeling the machine learning-based XGBoost for prediction of Korean professional baseball pitcher casey P. Kelly's situational pitch-type,” *Korean Journal of Convergence Science*, vol. 12, no. 10, pp. 87-98, 2023. DOI: 10.24826/KSCS.12.10.6.

[6] H. J. Choi and J. H. Kim, “The prediction of game results using ANN (Artificial Neural Networks) within the Wimbledon Tennis Championship 2005,” *The Korean Journal of Physical Education*, vol. 45, no. 3, pp. 459-468, 2006.

[7] J. B. Lim and K. S. Jeong, “A study on the change of quality in a residential sector of single person households in Seoul during the COVID-19: Analyze variable importance and causality with artificial neural networks and logistic regression analysis,” *LHI Journal of Land, Housing, and Urban Affairs*, vol. 14, no. 1, pp. 67-82, 2023.

천 재 성 (Jae-Sung Chun)_정회원

2011년 2월 : 한국기술교육대학교 컴퓨터공학부 졸업
2022년 8월 ~ 현재 : 한국기술교육대학교 컴퓨터공학과 석사과정
〈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