JKSCI

한국컴퓨터정보학회논문지
Journal of The Korea Society of Computer and Information
Vol. 29 No. 10, pp. 113-120, October 2024
https://doi.org/10.9708/jksci.2024.29.10.113

Exploring Predictive Models for Student Success in National Physical Therapy Examination: Machine Learning Approach

Bokyung Kim*, Yeonseop Lee**, Jang-hoon Shin***, Yusung Jang****, Wansuk Choi*****

*Professor, Dept. of Physical Therapy, Changshin University, Changwon, Korea
**Professor, Dept. of Physical Therapy, Daewon University, Jecheon, Korea
***Research Professor, Industry-Academy Cooperation Foundation, Sahmyook University, Seoul, Korea
****Adjunct Professor, Dept. of Physical Therapy, Gangdong University, Eumseong, Korea
*****Professor, Dept. of Physical Therapy, Kyungwoon University, Gumi, Korea

[Abstract]

This study aims to assess the effectiveness of machine learning models in predicting the pass rates of physical therapy students in national exams. Traditional grade prediction methods primarily rely on past academic performance or demographic data. However, this study employed machine learning and deep learning techniques to analyze mock test scores with the goal of improving prediction accuracy. Data from 1,242 students across five Korean universities were collected and preprocessed, followed by analysis using various models. Models, including those generated and fine-tuned with the assistance of ChatGPT-4, were applied to the dataset. The results showed that H2OAutoML (GBM2) performed the best with an accuracy of 98.4%, while TabNet, LightGBM, and RandomForest also demonstrated high performance. This study demonstrates the exceptional effectiveness of H2OAutoML (GBM2) in predicting national exam pass rates and suggests that these AI-assisted models can significantly contribute to medical education and policy.

▸ **Key words**: Machine Learning, Predictive Analysis, H2OAutoML(GBM2), Deep Learning, Educational Policy, ChatGPT

[요 약]

본 연구는 물리치료학과 학생들의 국가시험 합격률을 예측하는 데 있어 머신러닝 모델의 효과성을 검증하고자 한다. 기존의 성적 예측 방법은 주로 과거 학업 성적이나 인구 통계 데이터를 기반으로 하지만, 본 연구는 모의시험 점수를 머신러닝 및 딥러닝 기법으로 분석하여 보다 정확한 예측을 시도하였다. 한국의 5개 대학에서 총 1,242명의 학생 데이터를 수집하고 전처리한 후, 다양한 모델을 활용하여 분석을 진행하였다. ChatGPT4의 도움을 받아 생성 및 개선된 모델을 데이터셋에 적용한 결과, H2OAutoML (GBM2) 모델이 98.4%의 정확도로 가장 우수한 성능을 보였으며, TabNet, LightGBM, RandomForest 모델 역시 높은 성능을 나타냈다. 본 연구는 H2OAutoML (GBM2)이 국가시험 합격 여부를 예측하는 데 있어 뛰어난 효과를 발휘함을 보여주며, 이러한 AI 지원 모델들이 의학 교육 및 정책에 크게 기여할 수 있음을 시사한다.

▸ **주제어**: 머신러닝, 예측 분석, H2OAutoML(GBM2), 딥러닝, 교육 정책, 챗지피티

• First Author: Bokyung Kim, Corresponding Author: Wansuk Choi
*Bokyung Kim (heyuu98@gmail.com), Dept. of Physical Therapy, Changshin University
**Yeonseop Lee (bulchun325@naver.com), Dept. of Physical Therapy, Daewon University
***Jang-hoon Shin (hoon2612@naver.com), Industry-Academy Cooperation Foundation, Sahmyook University
****Yusung Jang (ysng3730@gmail.com), Dept. of Physical Therapy, Gangdong University
*****Wansuk Choi (y3korea@gmail.com), Dept. of Physical Therapy, Kyungwoon University
• Received: 2024. 08. 23, Revised: 2024. 09. 27, Accepted: 2024. 09. 27.

http://www.ksci.re.kr pISSN:1598-849X | eISSN:2383-9945

I. Introduction

의료 분야는 국가 인프라의 중요한 요소로, 신규 의료 전문가 수를 정확하게 예측하는 일은 매우 중요하다. 특히, 한국과 같은 국가에서는 한국보건의료인국가시험원(Korea Health Personnel Licensing Examination Institute; KHPLEI)이 배출하는 합격자 수가 의료 인력 규모를 주로 결정한다. 그러나 합격자 수를 예측하는 일은 시험 과정의 복잡성과 다양한 변수들로 인해 다차원적인 도전과제를 제기한다[1].

모의시험 점수를 예측 변수로 활용하는 이유는 모의시험 성적이 실제 시험 합격 결과와 높은 상관관계를 보이기 때문이다[2]. 모의시험은 실제 국가시험의 구조를 반영하며, 학생들의 준비 상태를 나타내는 중요한 지표로 작용한다. 현재의 예측 방법론은 주로 과거의 경향과 인구 통계 데이터를 기반으로 하고 있어, 시험 과정의 동적인 특성과 의료 환경의 변화하는 특성을 충분히 반영하지 못하는 경우가 많다[3]. 전통적인 통계 방법은 이러한 복잡성을 제대로 포착하지 못하여 예측의 부정확성을 초래할 수 있다[4-5]. 그러나 최근 연구들은 딥러닝(Deep Learning) 모델이 기존 통계 모델로는 발견하기 어려운 미세한 패턴과 상관관계를 데이터에서 효과적으로 식별할 수 있는 잠재력을 가지고 있음을 보여주고 있다[6-8].

기존 연구들은 주로 의사결정나무(Decision Tree), 서포트 벡터 머신(Support Vector Machine), 다중 회귀 분석(Multiple Regression Analysis) 등 전통적인 기계 학습 모델을 사용하여 물리치료사 국가시험의 합격 여부를 예측해왔다. 예를 들어, Kim et al. (2018)은 의사결정나무와 서포트 벡터 머신을 활용하여 모의시험 성적과 학습 시간을 중요한 예측 변수로 식별하였다[9]. Lee et al. (2020) 또한, 다중 회귀 분석을 통해 학업 성적과 모의시험 점수가 국가시험 성적과 유의미한 상관관계가 있음을 밝혔다[10]. 그러나 이러한 연구들은 대체로 선형적 관계를 가정하는 모델을 사용함으로써 복잡한 데이터 패턴을 충분히 포착하지 못했다[3,5]. 반면, 딥러닝 모델은 데이터를 보다 정교하게 분석하여 더 높은 정확도의 예측을 도출할 가능성이 있다[8,11].

국내에서 물리치료사 국가시험과 관련된 머신러닝(Machine Learning) 및 딥러닝 모델 연구는 부족하지만, 해외에서는 유사한 시험을 대상으로 다양한 모델들을 활용한 예측 연구들이 활발히 이루어지고 있다[3,12]. 예를 들어, 미국 의사 국가시험을 대상으로 한 연구에서는 딥러닝 모델이 학생의 시험 합격 가능성을 예측하는 데 효과적으로 활용되었으며, 기존의 통계적 예측 모델보다 더 나은 성능을 보였다[13]. ChatGPT와 같은 대규모 언어 모델을 활용한 연구에서는 미국 의사 국가시험 질문에 대한 딥러닝 모델의 성능이 60% 이상의 정확도를 기록했으며, 이는 시험 준비 과정에서 유용한 보조 도구로 사용될 가능성을 시사한다[14-15]. 또한, 간호사 시험이나 기타 의학 분야 시험에서도 딥러닝 기술을 활용한 연구들이 진행되었으며, 기존의 선형 회귀 모델을 넘어서는 성과를 보였다[5,11].

본 연구는 최신 머신러닝 및 딥러닝 모델을 활용하여 기존 통계 모델의 한계를 극복하고, 물리치료사 국가시험 합격 예측의 정확성을 개선하고자 한다. 특히 H2OAutoML, LightGBM, TabNet 등의 모델은 더 복잡한 데이터 구조를 처리하고 높은 예측 성능을 제공할 수 있는 잠재력을 가지고 있다. 본 연구에서는 모의시험 1, 2, 3교시 점수를 활용하여 예측 모델의 정확성을 높이고자 하였다.

본 연구는 챗지피티4(ChatGPT4)를 사용하여 초기 코드 생성을 수행하였으며, 구글 코랩(Google Colab)에서 실험을 진행함으로써 새로운 접근 방식을 제안한다. 이러한 접근법은 데이터 과학 및 머신러닝 교육에서 ChatGPT의 실용성을 탐구하며, Google Colab의 유연성과 접근성을 활용하여 교육 데이터 마이닝에 기여할 수 있는 방법을 제시한다. 따라서 본 연구는 KHPLEI에서 합격 여부를 예측하기 위한 새로운 접근법을 제안하며, 이는 의료 분야와 정부 정책에 중요한 영향을 미칠 수 있다. 특히, 인공지능(Artificial Intelligence; AI) 지원 모델의 도입은 기존 예측 모델의 한계를 넘어서는 정확성과 신뢰성을 제공함으로써 의료 인력 계획의 효율성을 향상시키는 데 기여할 수 있을 것이다[13].

II. Methods

학생들의 국가시험 합격을 예측하기 위해 사용된 모델은 데이터 수집, 데이터 정제, 데이터셋 분할, 모델 학습 등의 여러 단계를 거쳤다(Fig. 1). 각 단계에서 다양한 기술과 도구를 사용하여 모델의 성능을 최적화하였다.

1. Data Collection

본 연구의 데이터는 한국에 위치한 5개 대학 물리치료학과에서 수집되었다. 각 대학은 모의시험을 여러 차례 시행하였으며, 학생들이 응시한 3~5회의 모의시험 결과를 기반으로 데이터셋(Dataset)을 구축하였다. 총 1,242명의 학생 데이터가 수집되었으며, 각 학생은 3개의 교시 점수

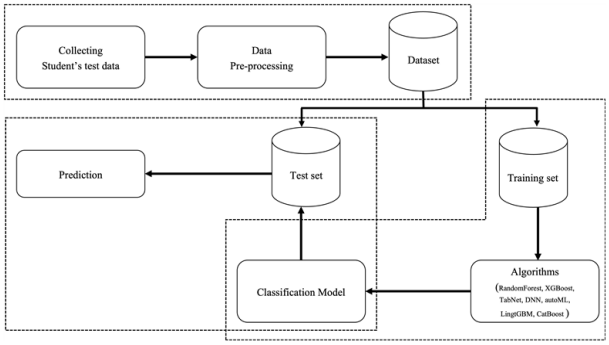


Fig. 1.The Workflow of the Prediction Model

(s1, s2, s3)를 기록하였다. 이는 각각 1교시, 2교시, 3교시의 점수를 의미하며, 모의시험의 각 과목에 대한 학생의 성취도를 반영한다.

데이터는 각 대학의 학과장을 통해 수집되었으며, 수집 과정에서 학생들의 민감한 개인정보가 보호될 수 있도록 개인 식별 정보는 모두 제거되었다. 데이터는 보안 USB 드라이브를 통해 전달되었다. 연구에 앞서 모든 학생에게 연구의 목적과 절차, 그리고 개인정보 처리 방식에 대해 충분히 설명한 후 서면 동의를 받았으며, 학생들은 자발적으로 연구에 참여하였다. 이러한 절차는 연구 윤리 규정을 철저히 준수한 상태에서 진행되었으며, 학생들의 성적을 포함한 모든 정보는 익명화된 상태로 처리되었다.

Table 1. Passing standards for physical therapist national exams(2023)

Class	Subject Name	Number of questions	Minimum score for passing the written test(40%)	Minimum score for final pass (60%)
1	PT basics	60	24	114(written test)
	PT diagnostic evaluation	45	18	
2	PT intervention	65	26	
	Medical related laws	20	8	
3	Practical test	70		42(practical test)
Sum		260		156

Table 1은 국가 물리치료사 시험의 내용을 보여준다. 국가 물리치료사 시험은 5개의 과목으로 구성되어 있으며, 합격을 위해서는 총점 60% 이상을 획득해야 한다. 학생들은 1교시와 2교시에서 각각 40% 이상의 점수를 획득하고, 1교시와 2교시의 합산 점수가 60% 이상이어야 한다. 또한, 실기에서 60% 이상을 받아야만 합격할 수 있다. 따라서, 합격하기 위해서는 최종 점수 156/260 이상을 받아야 한다.

2. Data Pre-processing

데이터 마이닝(Data Mining)에서 데이터 전처리는 모델 학습 성능을 좌우하는 중요한 단계이다. 본 연구에서는 수집된 원시 데이터를 분석에 적합한 형태로 변환하기 위해 다양한 전처리 작업을 수행하였다.

먼저, 결측치 처리가 이루어졌다. 전체 1,242개의 데이터 중 90개(약 7.2%)는 결측치를 포함하고 있었으며, 분석에 적합하지 않다고 판단되어 삭제되었다. 최종적으로 1,152개의 데이터가 분석에 사용되었으며, 결측치 비율이 낮아 데이터 왜곡 가능성은 매우 낮다고 판단되었다.

다음으로, 특징 스케일링(Feature Scaling) 작업이 수행되었다. 머신러닝 모델에 적합한 형태로 데이터를 변환하기 위해 표준 스케일링(Standard Scaling)을 적용하여, 각 변수의 평균을 0, 표준 편차를 1로 변환하였다. 이는 특히 DNN(Deep Neural Network), TabNet 모델의 성능을 향상시키기 위해 중요한 작업이었다. 이러한 신경망 기반 모델들은 입력 변수 간의 스케일 차이에 민감할 수 있기 때문에, 모든 변수의 분포를 동일한 범위 내로 표준화함으로써 학습 효율을 높이고 모델의 예측 성능을 최적화하였다.

트리 기반 모델(RandomForest, XGBoost, LightGBM, CatBoost)의 경우 스케일링이 필요하지 않기 때문에, 스케일링을 적용하지 않고 원본 데이터를 그대로 사용하였다. 이들 모델은 데이터의 스케일에 민감하지 않기 때문이다.

모의시험 성적은 실제 국가시험과 같은 방식으로 학생들의 시험 합격 여부를 신뢰성 있게 평가할 수 있어 주요 입력 변수로 선택되었다. 세 교시의 점수(s1, s2, s3)를 활용하였으며, 각 교시의 점수는 실제 시험의 과목별 평가를 반영하며, 모델이 학생의 합격 가능성을 예측하는 데 활용되었다. SHAP(Shapley Additive Explanations) 분석을 통해 세 개의 변수(s1, s2, s3)가 모델의 예측 성능에 미치는 영향을 분석한 결과, 각 변수가 합격 여부를 예측하는 데 중요한 역할을 하는 것으로 확인되었다.

3. Dataset Partitioning Strategy

모델 학습 및 평가를 위해 전체 데이터셋을 학습 세트와 테스트 세트로 분할하였다. 본 연구에서는 80:20 비율의 계층적 분할(Stratified Split) 방식을 채택하여, 학습 세트에 80%, 테스트 세트에 20%의 데이터를 할당하였다. 계층적 분할은 학습 세트와 테스트 세트에서 합격자와 불합

격자의 비율이 전체 데이터셋의 분포를 그대로 반영하도록 한다. 이를 통해 모델이 실제 데이터 분포를 잘 학습하고, 과적합을 방지하면서도 일반화 성능을 유지할 수 있도록 하였다.

학습 세트는 모델의 학습에 사용되었으며, 모델이 주어진 데이터에서 패턴을 학습할 수 있도록 충분한 데이터를 제공하였다. 테스트 세트는 학습되지 않은 데이터로 모델의 예측 성능을 평가하는 데 사용되었다. 이 분할 방법은 모델이 새로운 데이터에 대해 얼마나 잘 예측할 수 있는지를 정확하게 평가할 수 있는 기반을 마련하였다.

4. Model Development

본 연구에서는 다양한 머신러닝 알고리즘을 적용하여 국가시험 합격을 예측하는 모델을 개발하였다(Table 2). 각 모델의 설명은 다음과 같다.

H2OAutoML은 자동화된 머신러닝 플랫폼으로, 다양한 모델을 자동으로 학습하고 최적의 모델을 선택하는 기능을 제공한다. H2OAutoML에서 사용된 GBM2 모델은 그라디언트 부스팅 머신(Gradient Boosting Machine)의 변형으로, 약한 학습자(결정 트리)를 결합하여 성능을 극대화하는 모델이다[15]. 각 학습 단계에서 이전 모델의 오차를 줄여가며 학습하고, 이를 통해 예측 성능을 점진적으로 향상시킨다. GBM2 모델은 손실 함수를 기반으로 오차를 최소화하며, 불균형 데이터에서도 우수한 성능을 보인다. 이 모델은 결정 트리 기반의 앙상블 학습 방식 덕분에 고도의 예측 정확도를 달성할 수 있었다[15].

RandomForest는 여러 개의 결정 트리를 학습하여 예측 결과를 도출하는 앙상블 학습 기법이다. 트리 기반 모델의 장점은 직관적이며, 과적합을 방지하기 위해 다수의 트리를 학습함으로써 모델의 안정성과 예측 정확도를 높인다[16].

XGBoost는 그라디언트 부스팅 알고리즘의 일종으로, 각 단계에서 오차를 줄이는 방식으로 성능을 최적화하는 모델이다. 빠른 학습 속도와 높은 예측 정확도로 인해 널리 사용되며, 다양한 실험에서 탁월한 성능을 보인다[17].

TabNet은 표형 데이터(Tabular Data)에 특화된 딥러닝 모델로, 주목(Attention) 메커니즘을 통해 중요한 특징을 학습할 수 있다. 이는 데이터를 직접적으로 처리하면서도 특징 간의 복잡한 관계를 효과적으로 파악한다[18].

DNN은 여러 층의 신경망을 사용하여 비선형적이고 복잡한 데이터 패턴을 학습하는 모델이다. 다양한 계층에서 데이터의 복잡한 구조를 학습하며, 강력한 예측 능력을 보인다[19].

LightGBM은 대용량 데이터를 빠르게 처리할 수 있는 경량 그라디언트 부스팅 모델이다. 메모리 사용을 최소화하면서도 높은 성능을 보이며, 다양한 실험에서 효율적인 모델로 평가된다[20].

CatBoost는 범주형 데이터를 처리하는 데 특화된 부스팅 알고리즘이다. 데이터 불균형을 처리하는 데 강점이 있으며, 학습 속도와 성능 면에서 우수한 결과를 나타낸다[21].

Table 2. Code producution and prompts for creating code

Code production	Command
Get code suggestions	Using a csv file where x_label is a number, which pretraining model is best for predicting passing or failing a test?
Code production prompt (ex; CatBoost)	[Goal] Create test pass prediction model code using 'CatBoost' model. [File characteristics] x_label is a number, csv file, Target column is class (pass=1, fail=0), and three subjects are s1, s2, and s3. [Contents of the code] Visualization function (accuracy, loss, roc, auc, confusion matrix, save model, recall, precision, f1 score, logloss), Google Drive mount, file upload from Google Drive, GridSearchCV, Stratified Splitting, prediction code (3 subjects). [operating environment] Google Colaboratory

5. Hyperparameter Tuning

각 모델의 성능을 최적화하기 위해 그리드서치 교차 검증(Grid Search Cross-Validation)을 사용하여 하이퍼파라미터 조정(Hyperparameter Tuning)을 수행하였다. 이 기법은 교차 검증을 통해 최적의 하이퍼파라미터 조합을 찾고, 모델 성능을 최대화하기 위한 중요한 단계이다.

먼저, RandomForest 모델에서는 'gini' 기준을 사용하고, 최대 깊이를 5로 설정하였으며, 최대 특징 수는 자동으로 선택되도록 하고, 500개의 결정 트리를 사용하도록 하였다. XGBoost 모델에서는 'colsample_bytree'를 0.5로 설정하여 트리 구성 시 사용할 피처의 비율을 제한하였고, 학습률을 0.01로 낮춰 점진적으로 학습을 진행하였다. 또한, 최대 깊이는 5로, 결정 트리의 개수는 500개로 설정하였으며, 목적 함수로는 'binary'을 사용하여 이진 분류 문제를 해결하도록 하였다. 'subsample'은 1.0으로 설정하여 모든 데이터 포인트를 활용하였다. TabNet 모델의 경우, 'n_a'를 16으로, 'n_d'를 8로 설정하여 TabNet의 아텐션 레이어와 결정 레이어의 크기를 최적화하였다. LightGBM 모델에서는 'lambda_l1'을 0으로, 'lambda_l2'를 1로 설정하여 정규화 파라미터를 조정하였고, 'min_data_in_leaf'를 50으로 설정하여 최소 잎사귀

크기를 조정하였다. 또한, 'num_leaves'를 31로, 'reg_alpha'를 0.1로 설정하여 모델의 복잡성을 관리하였다. 마지막으로, CatBoost 모델에서는 최대 깊이를 5로, 학습률을 0.05로, 반복 횟수를 500회로 설정하여 최적의 성능을 도출하였다.

H2OAutoML의 경우, 수동으로 하이퍼파라미터를 설정할 필요 없이 자동으로 다양한 모델을 학습하고, 주어진 시간 내에서 최적의 모델을 선택하였다.

III. Results

1. Performance Evaluation of Models

본 연구에서는 5개의 머신러닝 모델(H2OAutoML, RandomForest, XGBoost, LightGBM, Catboost)과 2개의 딥러닝 모델(TabNet, DNN)을 비교 분석하였다. 각 모델의 성능은 네 가지 주요 지표(Precision, Recall, F1 score, Log Loss)를 사용하여 평가되었다. 정밀도(Precision)는 예측된 양성 사례 중 올바르게 식별된 양성 사례의 수를 측정하는 지표로, False Positive 비용이 높은 상황에서 중요하다. 재현율(Recall)은 실제 양성 사례 중 올바르게 식별된 양성 사례의 수를 정량화하는 지표로, False Negative 비용이 높은 경우에 유리하다. F1 score는 정밀도와 재현율의 조화 평균으로, 두 지표 간의 균형을 제공하며, 클래스 분포가 불균형한 경우 특히 유용하다. 마지막으로, Log Loss는 예측의 불확실성을 고려하여 모델 성능을 정밀하게 평가하는 지표로, 실제 레이블(Label)과 예측값 사이의 차이를 반영한다.

2. Comparison of Model's Performance

모델 성능 평가 결과, H2OAutoML 모델은 Precision 98.4%, Recall 99.1%, F1 score 98.8%, Log Loss 0.071로, 전체적으로 가장 뛰어난 성능을 기록하였다. 특히, 낮은 Log Loss는 예측의 불확실성을 잘 처리하고 있음을 보여준다. TabNet 모델은 Precision 98.1%, Recall 92.2%, F1 score 95.1%, Log Loss 3.172를 기록하였다. Precision은 높았으나, 낮은 Recall과 높은 Log Loss는 음성 사례 예측의 한계를 보여준다. LightGBM과 RandomForest 모델은 안정적인 성능을 보였다. LightGBM은 Precision 98.2%, Recall 95.7%, F1 score 96.9%, Log Loss 0.211을 기록하였으며, RandomForest는 Precision 97.4%, Recall 96.5%, F1 score 96.9%, Log Loss 2.018을 기록하였다. XGBoost는 Precision 97.3%, Recall 95.7%, F1 score 96.5%, Log Loss 0.185를 기록하며, 안정적인 성능을 보였으나 H2OAutoML보다는 다소 낮은 성능을 보였다. CatBoost는 Recall에서 100%를 기록하며 양성 사례를 정확히 예측하는 강점을 보였으나, Precision이 94.3%로 다소 낮아 일부 False Positive가 발생하였다. DNN 모델은 Precision 94.1%, Recall 96.5%, F1 score 95.3%, Log Loss 1.403을 기록하며, 다른 모델에 비해 낮은 성능을 보였다. DNN은 복잡한 데이터 패턴 학습에 강점을 가지지만, 본 연구의 데이터셋에서는 다른 모델에 비해 성능이 부족하였다.

Table 3. Performance comparison of deep learning models

Model	Precision (%)	Recall (%)	F1 score(%)	Log Loss
H2OAuto ML(GBM2)	0.984	0.991	0.988	0.071
Random Forest	0.974	0.965	0.969	2.018
XGBoost	0.973	0.957	0.965	0.185
TabNet	0.981	0.922	0.951	3.172
DNN	0.941	0.965	0.953	1.403
LightGBM	0.982	0.957	0.969	0.211
CatBoost	0.943	1.0	0.970	0.526

2.1 H2OAutoML Model

본 연구에서는 GBM2 모델의 성능 분석에 중점을 두었으며, 이를 평가하는 중요한 도구는 학습 곡선 플롯이다[13,15]. Fig. 2는 모델의 학습 과정을 이해하기 위한 필수적인 도구로, x축에 표시된 반복 횟수에 따라 두 개의 곡선을 매핑(Mapping)한다. 첫 번째 곡선(Blue)은 반복 횟수가 증가함에 따라 감소하는 훈련 손실을 나타내며, 이는 모델이 각 반복마다 훈련 데이터에 대한 적합성을 지속적으로 향상시키고 있음을 의미한다. 두 번째 곡선(Orange)은 검증 손실을 나타내며, 초기에는 감소하지만 최적의 반복 횟수에 도달한 후 증가하는 양상을 보인다. 이 지점 이후의 추가적인 훈련은 과적합을 초래할 수 있으므로, 학습 곡선은 모델 훈련 중 적절한 중단 시점을 결정하는 데 중요한 역할을 한다.

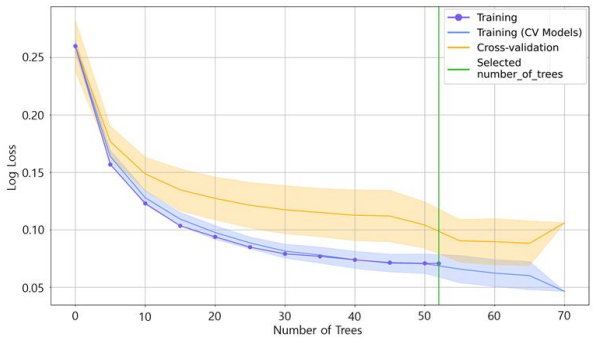


Fig. 2. Learning Curve for GBM2

2.2 SHAP Summary of the H2OAutoML Model

SHAP 요약 플롯(Fig. 3)을 통해 모델 예측에 미치는 세 교시(s1, s2, s3) 값을 평가하였다. 높은 SHAP 값을 통해 s1은 모델 출력과 복잡한 상호작용을 형성하며, 값이 높거나 낮을 때 모두 예측에 긍정적 또는 부정적 영향을 미칠 수 있다. 이는 다른 특징들과의 상호작용에 따라 다양한 결과를 초래한다. s2는 모델 출력에 긍정적인 영향을 미쳤으며, 높은 s2 값은 긍정적인 예측에 기여하며, 낮은 값은 예측 성능을 감소시키는 경향이 있다. s3는 모델 예측에 상대적으로 작은 영향을 미친다. s3의 영향은 s1과 유사한 방식으로 나타나지만, 그 영향력은 훨씬 제한적이다.

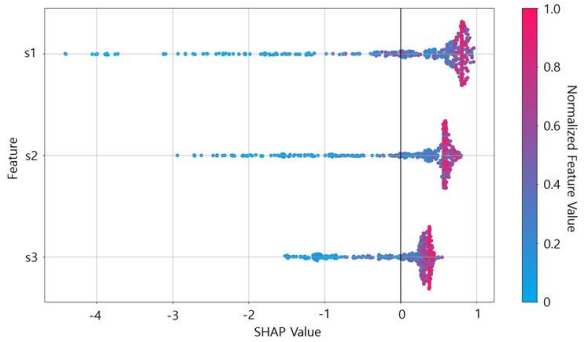


Fig. 3. SHAP Summary Plot for GBM2
(s1=1교시, s2=2교시, s3=3교시)

결론적으로, s1은 비선형적인 관계를 통해, s2는 주로 긍정적인 관계를 통해, 그리고 s3는 상대적으로 적은 영향을 통해 모델 예측에 기여하며, 중요도는 s1 > s2 > s3 순서로 나타났다.

IV. Discussion

본 연구는 물리치료사 국가시험 합격 예측을 위해 여러 머신러닝 및 딥러닝 모델을 적용하고 그 성능을 비교 분석하였다. 연구 결과, H2OAutoML이 가장 높은 성능을 보였다. 특히, H2OAutoML은 다양한 머신러닝 알고리즘을 자동으로 통합하여 최적의 모델을 선택하고, 예측 성능을 극대화하는데 성공하였다. H2OAutoML의 GBM2 모델은 그라디언트 부스팅 방법을 사용하여 복잡한 비선형 관계를 학습하고, 낮은 Log Loss와 높은 F1 score를 기록하면서 예측 정확성과 불확실성 측면에서 우수한 성능을 보였다[15].

TabNet은 높은 Precision을 보였으나, Recall과 Log Loss에서 상대적으로 낮은 성능을 기록하였다. 이는 TabNet이 양성 사례는 잘 예측하지만 음성 사례에 대해서는 예측력이 떨어짐을 시사한다. 따라서, TabNet은 데이터셋의 특성에 따라 성능이 달라질 수 있으며, 모든 예측 문제에 일관된 성능을 보이지 않을 수 있다[18].

CatBoost는 양성 사례를 100% 정확도로 예측하여 높은 Recall을 기록했지만, Precision이 다소 낮아 False Positive가 발생하였다. 이는 모델이 긍정적인 사례를 과대 평가할 가능성을 보여준다. 이는 CatBoost가 양성 사례를 과대 평가할 가능성이 있음을 나타내며, 특히 의료 응용에서 False Positive의 비용이 높을 경우 주의가 필요하다[21].

SHAP 분석을 통해 각 변수의 중요성을 평가하였다. 1교시 점수(s1)가 예측에서 가장 중요한 변수로 나타났으며, 이는 시험 초반 성적이 학생의 합격을 결정하는 중요한 지표임을 시사한다. 이러한 변수들의 상호작용은 머신러닝 모델이 시험 성적을 어떻게 예측하는지를 종합적으로 이해하는 데 중요한 단서를 제공한다.

연구의 한계로는 데이터셋의 다양성이 부족하다는 점이 있다. 연구에 사용된 데이터는 한국 내 특정 대학에 한정되어 있어, 다른 국가나 교육 시스템에 일반화하기 어렵다. 또한, 데이터 불균형 문제를 해결하기 위해 오버샘플링(Oversampling)이나 언더샘플링(Undersampling) 같은 추가적인 데이터 조정 방법이 사용되지 않았다는 점도 한계로 작용한다[11].

향후 연구에서는 더 다양한 데이터셋을 사용하여 모델의 일반화 성능을 평가할 필요가 있으며, 특히 딥러닝 모델의 성능을 개선하기 위해 하이퍼파라미터 최적화나 복잡한 네트워크 구조를 탐색하는 것도 필요하다. 또한, 모의시험 점수 외에도 학생의 학업 성취도나 학습 습관 등의 변수를 포함하여 예측 정확성을 더욱 향상시킬 수 있을 것이다[2].

이 연구는 H2OAutoML과 같은 자동화된 머신러닝 도구가 복잡한 예측 문제에서 수동으로 조정된 모델보다 효

과적일 수 있음을 시사하며, 특히 의료 및 교육 분야에서 의사결정 과정을 개선하는 데 중요한 역할을 할 수 있음을 보여준다. 또한, 특정 예측 문제에 맞는 모델을 선택하는 것이 성능을 최적화하는 데 중요하며, 이를 위한 정책적 가이드라인을 마련하는 것이 필요하다.

V. Conclusions

본 연구는 국가 물리치료사 시험에서 학생의 합격 여부를 예측하기 위해 다양한 머신러닝 및 딥러닝 모델을 비교 분석하였다. 연구 결과, H2OAutoML(GBM2) 모델이 가장 우수한 예측 성능을 보였으며, CatBoost의 인상적인 재현율이 확인되었다. 이는 교육 데이터 마이닝 및 예측 모델링 분야에서 자동화된 도구의 잠재적 유용성을 보여준다. 또한, 교육과 의료 분야에서 예측 모델을 활용하여 의사결정 과정을 개선하고, 보다 효과적인 학습 및 준비 전략을 개발하는 데 기여할 수 있다. 향후 연구에서는 더 다양한 데이터셋을 사용하여 모델의 일반화 성능을 평가하고, 데이터 불균형 문제를 보다 효과적으로 해결할 수 있는 방법을 탐색할 필요가 있다.

ACKNOWLEDGEMENT

This research was supported by a Research Grant of Kyungwoon University in 2024.

REFERENCES

[1] S. Batool, J. Rashid, M.W. Nisar, J. Kim, H.Y. Kwon, and A. Hussain, “Educational data mining to predict students' academic performance: A survey study,” Education and Information Technologies, Vol. 28, No. 1, pp. 905-971, Jan 2023. DOI: 10.1007/s10639-022-11152-y

[2] C. Ha, U. Ahmed, M. Khasminsky, M. Salib, and T. Andey, “Correlative and comparative study assessing use of a mock examination in a pharmaceutical calculations course,” American Journal of Pharmaceutical Education, Vol. 87, No. 1, pp. 8654, Jan 2023. DOI: 10.5688/ajpe8654

[3] R.R. Utzman, D.L. Riddle, and D.V. Jewell, “Use of demographic and quantitative admissions data to predict performance on the national physical therapy examination,” Physical Therapy, Vol. 87, No. 9, pp. 1181-1193, Sep 2023. DOI: 10.2522/ptj.20060222

[4] S.H. Kim, and S.H. Cho, “Exploring the predictive factors of passing the Korean physical therapist licensing examination,” Journal of The Korean Society of Integrative Medicine, Vol. 10, No. 3, pp. 107-117, Sep 2022. DOI: 10.15268/KSIM.2022.10.3..107

[5] A. Parhizkar, G. Tejeddin, and T. Khatibi, “Student performance prediction using datamining classification algorithms: Evaluating generalizability of models from geographical aspect,” Education and Information Technologies, pp. 1-19, Sep 2023. DOI: 10.1007/s10639-022-11560-0

[6] M.S. Kiran, E. Siramkaya, E. Esme, and M.N. Senkaya, “Prediction of the number of students taking make-up examinations using artificial neural networks,” International Journal of Machine Learning and Cybernetics, Vol. 13, No. 1, pp. 71-81, Jan 2022. DOI: 10.1007/s13042-021-01348-y

[7] Y. Chen, and L.A. Zhai, “A comparative study on student performance prediction using machine learning,” Education and Information Technologies, Vol. 28, No. 9, pp. 1-19, Sep 2023. DOI: 10.1007/s10639-023-11672-1

[8] M. Yağcı, “Educational data mining: prediction of students' academic performance using machine learning algorithms,” Smart Learning Environments, Vol. 9, No. 1, pp. 11, Jan 2022. DOI: 10.1186/s40561-022-00192-z

[9] S.H. Kim, J.Y. Park, and S.H. Cho, “Predicting the performance of students on the Korean national licensing examination for physical therapists using decision trees and support vector machines,” Journal of the Korean Society of Integrative Medicine, Vol. 6, No. 2, pp. 117-125, Jun 2018. DOI: 10.15268/KSIM.2018.6.2.117

[10] H.J. Lee, M.S. Kim, and H.Y. Choi, “Exploring the relationship between academic performance and mock test scores using multiple regression analysis: A study on the Korean national physical therapist examination,” Korean Journal of Educational Measurement and Evaluation, Vol. 12, No. 3, pp. 203-217, Sep 2020. DOI: 10.15268/KJEME.2020.12.3.203

[11] P. Nayak, S. Vaheed, S. Gupta, and N. Mohan, “Predicting students’ academic performance by mining the educational data through machine learning-based classification model,” Education and Information Technologies, pp. 1-27, Sep 2023. DOI: 10.1007/s10639-023-11706-8

[12] P. Sharma, K. Thapa, D. Thapa, P. Dhakal, M.D. Upadhaya, S. Adhikari, and S.R. Khanal, “Performance of ChatGPT on USMLE: Unlocking the Potential of Large Language Models for AI-Assisted Medical Education.” PLOS Digital Health. Jul 2023. DOI: 10.48550/arXiv.2307.00112

[13] P. Probst, B. Bischl, and A.L. Boulesteix, “Tunability: importance of hyperparameters of machine learning algorithms,” Journal of Machine Learning Research, Vol. 20, No. 53, pp. 1-32, Feb 2019.

[14] J. Bayliss, R.M. Thomas, and M. Eifert-Mangine, “Pilot study: what measures predict first time pass rate on the national physical therapy examination?” Internet Journal of Allied Health Sciences and Practice, Vol. 15, No. 4, pp. 1, Jan 2017. DOI: 10.46743/1540-580X/2017.1693

[15] E. LeDell, and S. Poirier. "H2o automl: Scalable automatic machine learning." Proceedings of the AutoML Workshop at ICML. Jul 2020. San Diego, CA, USA. DOI: 10.1145/3429136

[16] P. Geurts, D. Ernst, and L. Wehenkel, “Extremely randomized trees,” Machine Learning, Vol. 63, No. 1, pp. 3-42, Jun 2006. DOI: 10.1007/s10994-006-6226-1

[17] 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, August 2016, San Francisco, USA. DOI: 10.1145/2939672.2939785

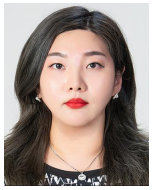
[18] S.Ö. Arik, and T. Pfister, “Tabnet: Attentive interpretable tabular learning,” Proceedings of the 35th AAAI Conference on Artificial Intelligence, pp. 6679-6687, February 2021, virtually. DOI: 10.1609/aaai.v35i8.16826

[19] K. He, X. Zhang, S. Ren, and J. Sun, “Deep residual learning for image recognition,” Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 770-778, June 2016, Las Vegas, USA. DOI: 10.48550/arXiv.1512.03385

[20] G. Ke, Q. Meng, T. Finley, T. Wang, W. Chen, W. Ma, Q. Ye, and T.Y. Liu, “LightGBM: A highly efficient gradient boosting decision tree," Proceedings of the 31st Conference on Neural Information Processing Systems,” pp. 3149-3157, Long Beach, USA, Dec, 2017.

[21] L. Prokhorenkova, G. Gusev, A. Vorobev, A.V. Dorogush, and A. Gulin, “CatBoost: unbiased boosting with categorical features,” Proceedings of the 32nd International Conference on Neural Information Processing Systems, pp. 6638-6648, December 2018, Montreal, Canada. DOI: 10.48550/arXiv.1706.09516

Authors



Bokyung Kim received her Ph.D., MS., and BS degrees in Physical Therapy from Yongin University. She is currently an Associate Professor in the Department of Physical Therapy at Changshin University, Korea. Her research interests include human-computer interaction (HCI), with a focus on applications to companion animal rehabilitation and healthcare.

Yeonseop Lee majored in physical therapy at Daejeon Health College and received his doctorate and master's degrees from Daegu University. He is currently an associate professor in the Department of Physical Therapy at Daewon University. His research areas include neurological diseases and pediatric diseases, and he focuses on the competencies of physical therapists in the administration of health care.

Jang-hoon Shin is a Research Professor at the Industry-University Cooperation Foundation of Sahmyook University, located in Seoul, South Korea. He holds a doctorate in physical therapy and is actively involved in research projects funded by Samsung Electronics and the National Research Foundation of Korea. His primary research interests include wearable robotics and rehabilitation ultrasound imaging.

Yusung Jang is currently pursuing his PhD in Physical Therapy at Yongin University. He is serving as an Adjunct Professor in the Department of Physical Therapy at Gangdong University, Korea. His research interests include Artificial Intelligence (AI) and its applications in manual and exercise therapy.

Wansuk Choi received his PhD, MS, and BS in Physical Therapy from Yongin University. He is currently an Assistant Professor in the Department of Physical Therapy at Kyungwoon University, Korea. His research interests include Human-Computer Interaction (HCI), with a focus on its applications in healthcare and rehabilitation technology.