산업융합연구 제22권 제8호 pp. 19-28, 2024
ISSN 2635-8875 / e-ISSN 2672-0124
DOI : https://doi.org/10.22678/JIC.2024.22.8.019

빅데이터 및 인공지능을 활용한 축구선수 연봉등급 예측

정현성[1], 김진화[2*], 현대원[2]
[1]서강대 메타버스전문대학원 학생, [2]서강대 메타버스전문대학원 교수

Predicting Soccer Players' Wage Grades Using Big Data and Artificial Intelligence

Hyeon-Seong Jeong[1], Jin-hwa Kim[2*], Dae-Won Hyun[2]
[1]Student, Graduate School of Metaverse, Sogang University
[2]Professor, Graduate School of Metaverse, Sogang University

요 약 본 연구는 빅데이터와 인공지능을 활용하여 축구선수의 연봉등급을 예측하는 새로운 방법을 제안한다. 축구선수의 연봉 예측은 선수의 성과와 잠재력을 정확하게 평가하고, 이를 연봉에 반영함으로써 축구 산업의 경제적 효율성을 높이는 중요한 과제이다. 본 연구는 FIFA 22에서 제공하는 선수 능력치 데이터를 분석하여, 다양한 빅데이터 및 인공지능 기법을 통해 선수의 연봉등급을 예측한다. 주요 연구 방법으로는 의사결정나무, 인공신경망, 랜덤 포레스트, 부스팅 등을 활용하였으며, 이를 통해 연봉등급을 예측하는 모델의 정확도를 비교 분석하였다. 연구 결과, 랜덤 포레스트와 부스팅 기법이 가장 높은 예측 정확도를 보였다. 이 연구는 빅데이터와 인공지능을 이용해 축구선수의 연봉등급을 예측하고, 축구 산업에 새로운 관점을 제공한다.

키워드 : 빅데이터, 인공지능, 축구선수, 연봉등급, FIFA 22, 예측 모델

Abstract This study proposes a new method for predicting the wage grades of soccer players using big data and artificial intelligence. Predicting the salaries of soccer players is a crucial task that involves accurately assessing players' performance and potential, and reflecting this in their salaries to enhance the economic efficiency of the soccer industry. This research analyzes player ability data provided by FIFA 22 and employs various big data and artificial intelligence techniques to predict players' salary grades. Key methodologies used include decision trees, artificial neural networks, random forests, and boosting, which were utilized to compare the accuracy of the salary prediction models. The results show that the random forest and boosting methods exhibited the highest prediction accuracy. This study demonstrates the process and utility of using big data and artificial intelligence technologies to predict soccer players' salary grades, offering a new perspective on the soccer industry.

Key Words : Big Data, Artificial Intelligence, Soccer Players, wage Grades, FIFA 22, Prediction Models

This research was supported by the MSIT(Ministry of Science and ICT), Korea, under the Graduate School of Metaverse Convergence support program(IITP-2023-RS-2022-00156318) supervised by the IITP(Institute for Information & Communications Technology Planning & Evaluation)
*Corresponding Author : Jin-Hwa, Kim(jinhwakim@sogang.ac.kr)
Received July 19, 2024 Revised August 8, 2024
Accepted August 20, 2024 Published August 28, 2024

1. 서론

현대 사회에서 빅데이터와 인공지능의 발전은 의사결정 과정을 향상시키고 예측의 정확성을 높이는 새로운 방법을 제시하고 있다. 다양한 산업 분야에서 이 기술들의 효용성이 입증되고 있는 가운데, 스포츠 산업에서도 빅데이터와 인공지능의 적용이 두드러지게 나타나고 있다. 복잡하고 동적인 축구의 성격은 빅데이터와 인공지능의 활용 가능성을 광범위하게 만드는데, 선수의 실력 평가부터 팀 전략 설정, 심지어 경기 결과 예측에 이르기까지, 이 기술들이 다양하게 활용될 수 있다. 특히, 축구 산업에서 선수의 연봉 예측은 중요한 과제로 인식되며, 이는 선수의 성과와 잠재력을 정확하게 파악하고, 이를 선수의 가치에 반영하는 데 큰 의미를 지닌다. 축구선수들의 연봉이 그들의 가치를 정확하게 반영하도록 돕는 것은, 축구 산업의 경제적 효율성을 높이는 데에 기여할 것으로 생각한다.

본 연구에서는 피파 22 게임에서 제공하는 선수 능력치 데이터를 바탕으로, 빅데이터와 인공지능 기법을 이용하여 축구선수의 연봉등급을 예측하는 방법론을 제안한다. 선수들의 실제 능력을 세밀하게 구성하고 수치화된 이 데이터는 선수의 연봉등급을 예측하는 데 있어서 매우 유용한 자료로 활용될 수 있다. 연구 목표를 달성하기 위해 다음과 같은 세부 연구 과제를 설정하였다: 첫째, 피파 22의 선수 능력치 데이터를 분석하여, 연봉등급 예측에 활용한다. 둘째, 이 데이터를 빅데이터 및 인공지능 기법에 활용하여 선수의 연봉등급을 예측하며, 그 결과를 통해 축구선수의 연봉 결정 과정의 신뢰성을 제고하고자 한다. 셋째, 이 연구를 통해 빅데이터와 인공지능의 축구 산업 내 적용 가능성과 그 가치를 탐색하고자 한다.

본 연구는 축구 산업에 새로운 시사점을 제공하며, 빅데이터와 인공지능이 어떻게 선수의 연봉 예측에 활용될 수 있는지에 대한 이해를 높이는 데 이바지한다. 이로써 축구 산업의 투명성 높이는 데에도 도움이 될 것으로 기대한다. 결국, 본 연구의 목표는 빅데이터와 인공지능의 활용을 통해 축구선수의 연봉등급을 예측하고, 그 결과를 통해 축구 산업의 투명성과 경제적 효율성을 증대하는 것이다. 연구 절차는 〈Fig 1.〉과 같이 4단계로 진행된다. 먼저, 수집된 데이터를 회귀분석으로 관련 변수를 선정한 후, 실험에 필요한 데이터를 훈련용/검증용으로 구분하여 중복되지 않는 10개의 데이터 셋을 만든 다음, 빅데이터와 인공지능 기법을 이용하여 예측을 수행하여 결과에 대한 비교 및 분석한다. 이러한 연구는 빅데이터와 인공지능이 축구 산업에서 활용될 수 있음을 보여주는 사례가 될 것이다.

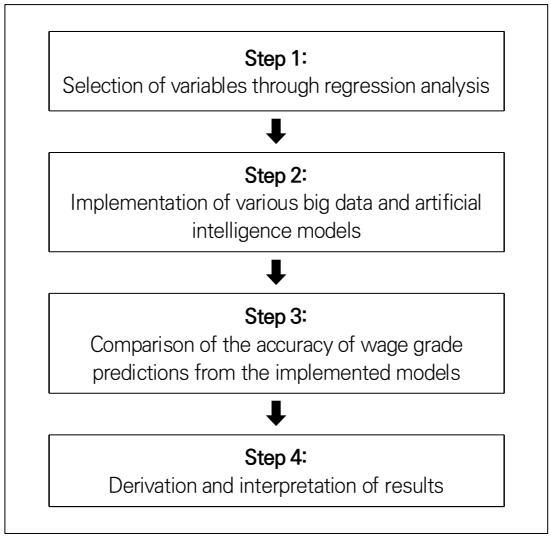

Fig. 1. Research process

2. 선행연구

2.1 빅데이터 및 인공지능 활용한 스포츠 분야 예측 관련 연구

빅데이터 및 인공지능을 활용한 스포츠 분야 예측에 관한 연구는 다양한 방식으로 진행되고 있다. 프로야구경기의 승패 예측을 위해 선수들이 기록한 날짜별 데이터를 기반으로 인공신경망을 이용하여 경기를 예측하는 모델을 제시하였으며[1], 인공신경망 기반 모델은 기존의 승률 예측 모델보다 월등히 우수한 것으로 나타났다. KBO (Korea Baseball Organization) 정규시즌에서 생성된 타자의 데이터를 바탕으로 머신러닝 알고리즘을 적용해 타자의 OPS(On-base Plus Slugging)를 예측하는 모델을 제시하였고[2], 제시된 예측기법은 타자의 OPS를 예측하는 최고의 성능을 보여주었다. 이외에도, 랜덤 포레스트 분류기를 이용하여 실질적인 분석을 시행하여 득점수 예측 기반으로 한 축구 경기 베팅 분석 모델을 만들었으며, Under/Over 예측의 경우 베팅업체의 배당률을 기반으로 한 예측에 비해 높은 성능을 보여주었다[3]. 또한, 빅데이터를 활용하여 타자 경기력과 불쾌지수의 관계 분

석하였으며[4], 이 외에도 다양한 스포츠 분야에 빅데이터 및 인공지능 기법을 활용하여 예측 및 분석 등을 진행하였고, 그 연구는 Table 1과 같다.

Table 1. Summary table of Previous Research

Index	Title	Author
1	KBO Professional Baseball Game Prediction	Noh
2	Predicting the OPS of KBO Batters	Han, Jung, Kim
3	Soccer match betting analysis based on score prediction	Chung
4	Prediction Model for Winning Rate of the Motorboat Racing	Noh
5	Determinants of Sport Participants' Injury Using Decision Tree Analysis	Myung, Park
6	Exploring the Determinants of Winner and Defeat of the Women's Water Polo World Championship	Lee, Park, Jo
7	Machine Learning Models for Predicting Swimming Competition Results	Yang
8	Predictive Model on Substitution for Starting Pitcher in Korean Professional Baseball	Cho
9	Redefining Positions for the Modern Basketball with Machine Learning	Kim
10	Prediction of Winning Horses in Horse Races	Choe, Hwang, Hwang, Song
11	Handball match Results Visualization and Prediction Comparisons Using Machine Learning	Kim
12	The Final Ranking Prediction of the Korean Professional Basketball League	Kim, Lee
13	Comparison of Prediction Performance of Machine Learning Classification Model Using 2022 FIBA Mens Basketball Asian Cup Match Results	Lee, Ni
14	Prediction and Evaluation of Keirin Competition Rankings	Kim, Lee, Jeon

2.2 축구 선수의 연봉 관련 연구

축구는 동적이고 복잡한 성격 때문에 선수의 연봉을 산정하는데 있어 타 스포츠보다 정보를 객관화하고 세분화하기에 불리한 조건이며, 현재로는 구단이 자체적으로 수립한 연봉제도(고정급, 기본급+출전승리급, 메리트 등)를 구분하여 지급하고 있지만, 문제는 그 세부기준이 비공개적이며 감독과 구단의 영향이 크게 작용하고 있다[5]. 이러한 상황에서, 빅데이터와 인공지능을 활용하여 축구선수의 능력수치를 객관적 및 정량화하여 축구선수의 연봉등급을 예측하고자 한다. 위 기술을 활용한 축구선수의 연봉등급 예측은 아직 초기 단계이며, 이러한 기술의 축구 산업에 대한 적용 가능성과 가치는 아직 연구되지 않았다. 그래서, 본 연구에서는 이런 기술들을 이용하여 축구선수의 연봉등급을 예측하고자 한다.

축구 산업은 그 크기와 중요성 때문에 축구선수들의 연봉이 그들의 성과와 능력을 정확하게 반영하는 것이 중요하다. 이 연구의 목표는 빅데이터와 인공지능을 활용하여 축구선수의 연봉등급을 예측하고, 그를 통해 축구 산업에 더 큰 투명성과 공정성을 도입하는 것이다. 이를 통해, 본 연구는 축구 산업에 대한 새로운 통찰을 제공하고, 빅데이터와 인공지능이 축구선수의 연봉등급 예측에 어떻게 활용될 수 있는지 보여줄 것이다.

3. 연봉등급 예측기법

3.1 의사결정나무(Decision Tree)

의사결정나무 분석방법은 광범위한 분류 및 회귀(예측) 과제에 적용될 수 있는 강력한 도구로 인식되어 왔다[6]. 의사결정나무의 특징은 가설 없이도 예측 가능한 모든 변수로부터 결정요인을 발견할 수 있다는 장점이 있으며, '어떤 조건이 충족될 때 결과는 이런 것이다'라는 결정 규칙을 통해 의사결정 경로를 구성하기에 복잡한 데이터 전처리 작업을 최소화하면서도 이해하기 쉬운 트리 구조를 제공한다. 또한, 의사결정나무의 결과는 나무 구조로 제시되기 때문에, 결과의 가시성과 해석력 측면에서 큰 장점이 있다[7]. 이러한 특징은 데이터를 직관적으로 이해하고, 의사결정을 뒷받침하는 데 유용하기에 다양한 분야, 특히 의료, 금융, 마케팅 등에서 적극적으로 활용되고 있다.

기존 의사결정나무에서 보완된 방법으로 C-tree (Conditional Inference Tree)도 있다. C-tree는 기존 의사결정나무를 통계적 검증(p-value)에 기반하여 나무를 분할한다. 이러한 방식은 기존의 의사결정나무에 비해 분할 변수 선택에 편향이 적고, 연속 변수나 순서형 변수와 같은 다양한 자료 유형을 처리할 수 있다는 점이다. 또한, 중복 또는 누락된 데이터 처리에도 강점을 보인다. 하지만, 연구의 목적이나 데이터의 특성에 따라 기본 의사결정나무 기법이 더 적합할 수도 있기에, 기본 의사결정나무 기법과 C-tree 기법을 비교해보아야 한다.

3.2 인공신경망(Artificial Neural Networks)

인공신경망은 머신러닝의 한 방법으로, 인간의 신경망 구조를 모방한 분석법이다[8]. 이는 인간의 뇌와 유사하게 '특정 입력이 주어질 때, 출력은 이런 것이다'라는 패턴을 학습하여 최적의 출력값을 산출하는 복잡한 시스템을 통해 이루어진다. 인공신경망은 입력층, 은닉층, 출력층과 각 층(layer)에 속하는 노드들로 구성되며, 은닉층은 1개 이상으로 구성될 수 있고 입력층 및 출력층의 노드의 수 또한 속성의 표현방식에 따라 결정될 수 있다[9]. 이러한 학습 능력은 다양한 비선형 문제에 대한 해결을 가능하게 하며, 복잡한 패턴 인식에 탁월한 능력을 제공하기에, 이미지 인식, 음성 인식, 자연어 처리 등 복잡한 패턴 인식이 필요한 다양한 분야에서 널리 활용되고 있다.

3.3 랜덤포레스트(Random Forest)

랜덤포레스트는 의사결정나무 알고리즘의 발전형으로 볼 수 있다. 데이터의 변화에 따른 변동성과 높은 분산 오류로 인해 일반화가 어려운 의사결정나무의 한계를, 앙상블 기법을 도입함으로써 극복하려는 방식이다[10]. 이는 무작위로 선택한 표본과 변인을 사용하여 여러 개의 의사결정나무를 만들고, 이들의 결과를 종합하여 최종적인 예측 모델을 구축하는 원리로 작동한다. 즉, 랜덤 포레스트는 다양한 의사결정나무들의 "평균"을 이용하여 예측 오류를 줄이는 전략을 취한다[11]. 또한, 로지스틱 회귀분석과 같은 기존의 방법들과 비교했을 때, 랜덤포레스트는 독립변인의 수가 많아질수록 더욱 강력해진다. 이는 복잡한 분포와 다양한 독립변인들을 처리할 수 있는 능력을 가지고 있기 때문이다[12]. 따라서, 랜덤 포레스트는 특히 다양하고 복잡한 데이터를 다루는 문제에 매우 유용하다. 이런 특성 덕분에 랜덤 포레스트는 높은 예측 정확도와 함께 결과 해석의 용이성을 제공하므로, 데이터 분석가들 사이에서 널리 활용되는 기법이 되었다.

3.4 부스팅(Boosting)

부스팅 알고리즘은 앙상블 학습의 한 형태로, 여러 개의 약한 분류기(weak classifier)를 결합하여 더 나은 최종 예측을 출력하는 방법이다[13]. 앙상블 학습은 다수의 모델을 통합하는 방법으로, 일반적으로 단일 모델을 사용하는 것보다 평균적인 예측 성능을 향상시키는 효과를 가지고 있다[14]. 이는 모델별로 다르게 발생하는 예측 오류를 상호 보완함으로써 성능 향상을 이루는 방식이다. 특히, 부스팅 알고리즘은 의사결정트리와 같은 기계학습 알고리즘과 결합될 때 예측력을 더욱 향상시킬 수 있다. 여러 의사결정트리를 함께 사용하되, 각각의 트리가 학습하는 과정에서 오류를 최소화하는 방향으로 학습하게 된다[15]. 이렇게 학습된 모델들은 각자의 장점을 결합하여 더욱 강력한 예측 모델을 형성하게 된다. 부스팅의 중요한 특징 중 하나는 약한 학습기가 순차적으로 학습되며, 각 학습 단계에서 이전 학습기의 오류를 보완하려는 경향이 있다는 점이다. 부스팅은 동시에 병렬 학습하는 배깅기법과 다르게 순차적으로 앞의 모델들을 보완해 나가며 직렬로 학습한다. 이러한 방식으로, 부스팅 알고리즘은 각각의 학습기가 독립적으로 학습하는 다른 앙상블 방법들과는 달리, 강한 예측력을 가진 하나의 강력한 모델을 구축한다. 이런 특성 덕분에 부스팅 알고리즘은 복잡한 문제를 해결하는 데 널리 활용되며, 특히 높은 정확도가 요구되는 분야에서 많은 성공을 거두었다.

3.5 KNN(K-Nearest Neighbors)

KNN(K-Nearest Neighbors)은 지도학습의 한 방법으로, 거리기반 분류분석 모델이다[16]. 새로운 데이터의 분류나 예측을 위해 가장 가까운 훈련 데이터 포인트 'k' 개를 참고하는 알고리즘으로, 이 알고리즘은 학습 데이터 셋 내에서 새로운 데이터 포인트와 가장 가까운 'k'개의 이웃을 찾아 그들의 클래스 또는 값에 따라 예측값을 결정한다. KNN은 선택된 K값에 따라 성능이 크게 좌우되며, 너무 작은 K값은 과적합, 너무 큰 K값은 과소적합의 위험을 내포하고 있다[17]. 적절한 K값의 선정은 이 알고리즘의 성능을 크게 좌우하는 중요한 요소이기에, 다양한 실험을 통해 최적의 K값을 찾는 것이 중요하다. 이처럼 KNN 알고리즘은 거리 기반의 분류 방법이기에 간결하고 직관적이다. KNN은 원리가 단순하나 다양한 문제에 대한 높은 성능을 자랑하기에 많은 분야에서 사용된다.

3.6 딥러닝(Deep Learning)

딥러닝은 인공신경망을 활용하여 데이터로부터 복잡한 특징과 패턴을 추출하는 고급 기계학습 방법론이다. 이 기술은 여러 비선형 변환 과정을 거쳐 데이터의 고차원적인 추상화를 달성하며, 인간의 뇌 구조에 영감을 받아 다층적인 네트워크 아키텍처를 사용한다. 딥러닝은 사

람의 개입 없이 스스로 학습하고 문제를 해결할 수 있도록 설계되어, 자동화된 패턴 인식과 의사결정을 가능하게 하기에, 이러한 점에서 딥러닝은 데이터에서 복잡한 통찰을 추출하고, 이를 기반으로 인공지능이 자율적으로 작동할 수 있는 능력을 제공한다[18]. 딥러닝은 입력 데이터를 이용하여 모델을 훈련시키는 과정에서 피드포워드와 역전파 단계를 번갈아 수행함으로써, 모델의 예측값과 실제값 사이의 차이를 점진적으로 줄여가는 방식으로 진행하며, 이 과정을 통해 모델은 점차 최적화된다[19].

본 연구에서는 다양한 예측 기법 중에서 의사결정나무, 딥러닝, 랜덤포레스트, 부스팅, K-최근접 이웃(KNN), C-tree 총 6가지 기법을 선택하여 사용하였다. 이러한 선택은 각 기법이 축구선수의 연봉 예측에 필요한 다양한 데이터 특성과 복잡한 상호작용을 효과적으로 처리할 수 있는 능력을 갖추고 있기 때문이다. 의사결정나무와 랜덤포레스트는 변수의 중요도를 평가하고 해석하기 용이하며, 딥러닝은 비선형 관계를 학습하는 데 강점을 지닌다. 부스팅 기법은 예측 정확도를 향상시키고, K-최근접 이웃은 데이터의 지역적 패턴을 반영하는 데 유리하다. 이러한 다양한 특성을 가진 기법들을 조합함으로써, 본 연구는 모델의 예측 성능을 비교하고, 데이터의 다양한 특성을 효과적으로 반영하고자 하였다. 아래의 Table 2는 적용한 기법과 각 기법의 주요 특징을 요약한 것이다.

Table 2. Methods and Their Features

Methods	Features
KNN	거리 기반 분류, Input 데이터를 학습 데이터 기반으로 분류 및 예측 수행
Random Forest	여러 개의 의사결정나무를 결합하여 오류 최소화
Boosting	약한 분류기를 결합하여 강 분류기로 구축
Decision Tree	직관적인 트리구조, 각 변수에서 결정요인을 발견하여 결정
Deep Learning	뇌의 신경망 구조를 모방하며, 입력에 따른 최적의 출력값 산출, 스스로 패턴을 찾아 분류

4. 연구결과

4.1 자료수집

본 연구에서 사용된 데이터는 피파 22 게임에서 제공하는 선수 능력치 데이터이다. 피파 시리즈는 실제 축구 경기에 기반하여 각 선수의 기술, 체력, 스피드 등을 포괄적으로 정량화하며, 이러한 데이터는 선수들의 실제 경기력을 기반으로 평가된다. 또한, EA Sports에서 전문가의 평가를 바탕으로 일관된 기준으로 선수들의 능력을 평가하며 제공하므로 데이터의 일관성을 유지한다. 또한, 선수들에 대한 광범위한 데이터를 제공함으로써 다양한 통계적 분석 및 빅데이터와 인공지능기법을 적용할 때 필요한 대규모 샘플 크기를 확보할 수 있게 해 주기 때문에 본 연구에서 피파 2022 데이터를 사용하였다.

4.2 변수선정

FIFA 2022 데이터셋은 총 19,240명의 선수에 대한 110개의 다양한 변수를 포함하고 있다. 이 변수들에는 선수의 이름, 팀 이름, 유니폼 번호, 포지션, 플레이 스타일, 나이, 키, 몸무게, 생년월일, 연봉, 시장가치, 잠재력, 슈팅 능력, 패스 능력, 몸싸움 능력 등이 있다. 본 연구의 초기 단계에서는 연구 목적에 부합하는 변수를 선정하는 작업이 수행되었다. 수치화할 수 없는 변수들인 이름, 생년월일, 팀 이름 등은 제외하였으며, 나머지 변수들 중에서 연봉(wage_eur)과 높은 상관관계를 보이는 변수들을 체계적으로 선별하기 위해 회귀분석을 실시하였다. 이 분석을 통해 연봉에 가장 밀접한 영향을 미치는 9개의 변수를 선정하였다. 10개 이상의 변수를 포함시킬 경우 모델의 과적합 문제가 발생하고 예측 정확도가 저하되는 경향이 관찰되었다. 따라서, 최적의 모델 성능을 달성하기 위해 상위 9개의 변수만을 최종적으로 선택하였다. 이렇게 선정된 9개의 독립변수는 아래의 Fig. 2와 같다.

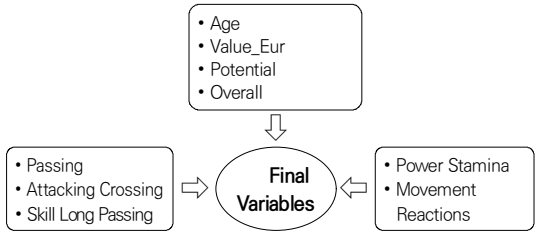

Fig. 2. Final Variable Selection

이를 통해 선수의 연봉에 영향을 미치는 주요 변수들을 파악하고, 이러한 변수들을 활용하여 연봉을 예측하는 모델을 구현하였다. 이러한 변수 선정 과정은 연봉 예측 모델의 특별한 접근 방식과 독창성을 강조하며, 이는 본 연구의 연봉 예측 정확도를 더욱 높이는 요소로 작용한다. 선정된 9개의 변수에 대한 회귀분석 결과는 아래의 Table 3에 나타나 있다. 회귀분석에서 R^2는 0.713으로,

모델이 설명하는 데이터의 변동성의 71.3%를 나타내며, 조정된 R^2는 0.713으로 계산되어, 변수의 수를 고려한 모델의 효율성을 나타내었다.

Table 3. Regression analysis of variables on wage_eur

회귀분석		
1	value_eur	p-value < 0.000
2	overall	p-value < 0.001
3	age	p-value < 0.002
4	potential	p-value < 0.003
5	passing	p-value = 0.001238
6	attacking crossing	p-value = 0.001239
7	skill long passing	p-value = 0.001240
8	power stamina	p-value = 0.001241
9	movement reactions	p-value = 0.001242

선수의 시장가치(value_eur), 종합능력치(overall), 잠재력(potential)과 같은 변수들은 선수의 전반적인 능력과 장래가치를 나타내는 중요한 지표로, 선수의 연봉 결정에 큰 영향을 미친다. 여기서 '종합능력치'란 선수의 공격력, 수비력, 체력, 속도, 패스 능력과 같은 개별 요소들을 종합적으로 고려하여 계산된 수치이다. 또한, 패스능력, 공격 크로스 능력, 롱패스 능력, 체력, 반응속도과 같은 변수들은 선수의 경기 내 성과와 직결되는 요인으로, 연봉 산정에 있어 핵심적인 역할을 하기에 연봉과 가장 밀접하게 연관이 되어 있었다. Fig 3은 이러한 변수들 각각이 연봉과의 상관계수가 시각적으로 표현된 그래프로 연봉에서 각 변수의 영향력을 직관적으로 이해할 수 있다.

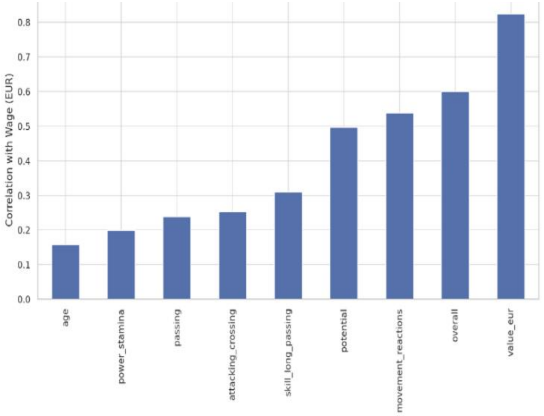

Fig. 3. Correlation Coefficients between Variables and Wage

4.3 빅데이터 및 인공지능 모델링

본 연구에서는 선수의 연봉등급 예측을 목표로 하여, 다양한 기계학습 방법들을 적용하여 성능을 비교하였다. K-Nearest Neighbors(KNN), Random Forest, Boosting, Decision Tree, Decision Tree(Ctree), Deep Learning 방법이 포함되어 있으며, 이러한 다양한 모델들은 각기 다른 원리와 특징을 지니고 있어, 선수의 연봉등급 예측에 어떠한 방법이 가장 적합한지를 파악하는 데 중요한 역할을 한다. 본 연구에서는 대부분의 모델링 작업을 R 언어 환경에서 수행하였다. R 언어는 통계 및 데이터 분석 작업에 널리 사용되는 프로그래밍 언어로, 여러 데이터 분석 패키지와 라이브러리를 지원하여 연구에 적합하다. 특히 복잡한 계산과 큰 데이터셋에 대한 학습을 위한 Deep Learning은 Google Colab 환경에서 구현하였다. Google Colab은 클라우드 기반의 무료 개발 환경으로, 복잡한 deep learning 모델의 효율적인 학습과 실험을 가능하게 한다.

4.4 평가방법

본 연구 연봉등급을 기반으로 한 분류 문제에서 정확도(Accuracy)를 활용하여 모델의 성능을 평가하였다. 정확도는 모델이 올바르게 분류된 항목의 비율을 나타내는 지표로, 분류 문제에서 모델의 성능을 판단하는 기본적인 척도이다. 값이 높을수록 모델의 성능이 좋다고 판단할 수 있으며, 본 연구에서는 선수의 연봉등급을 정확하게 분류하는 능력을 측정하는 데 사용되었다. 실제 연봉이 아닌 연봉등급으로 분류하여 정확도를 측정한 이유는 월드클래스 선수들의 연봉이 일반 선수들과 비교하여 지나치게 높은 경우가 많아, 이로 인해 전체 연봉 데이터는 '긴 꼬리' 분포를 보이거나 이상치의 영향을 크게 받았다. 이러한 데이터 분포의 특성은 일반적인 회귀 모델에서 가정하는 '정규성'을 침해하게 되므로, 모델의 예측 성능에 부정적인 영향을 미칠 수 있다. 따라서 본 연구에서는 이러한 문제를 해결하기 위해 연봉을 총 6등급(상위 1% 등급, 1~5% 등급, 5~10% 등급, 10~20% 등급, 20~50% 등급, 50~100% 등급)으로 구분하여 분석하였다. 각 등급은 연봉 분포의 특성과 선수들의 실제 연봉을 고려하여 설정하였고, 이를 통해 모델의 예측 성능을 더욱 견고하게 만들 수 있었다.

5. 연구결과

5.1 데이터 추출

본 연구에서 사용된 데이터는 총 19,240명의 선수 데이터를 기반으로 하며, 교차 검증 방법을 적용하기 위해 7:3의 비율로 학습 데이터와 테스트 데이터를 총 10개로 분할하였다. 이는 각 실험마다 서로 다른 70%의 학습 데이터(약 13,468명)와 30%의 테스트 데이터(약 5,772명)가 사용되도록 설정하여, 모델의 예측 성능을 검증하고자 하였다. 이러한 방식은, 한 번의 분할로 인한 특정 데이터 패턴에 의존하지 않고, 모델의 과적합 위험을 최소화하며 일반화 능력을 극대화하는데 기여한다. 또한, 여러 번의 실험을 통해 각 기법의 성능의 안정성과 일관성 또한 평가할 수 있게 되었다.

5.2 등급 기반 모델 학습과 정확도 측정

종속변수와 선택된 9가지 독립변수들을 활용하여 각 모델을 학습시킨 후, 학습된 모델에 10개의 테스트 데이터에 대한 연봉 등급을 예측하였다. 그 결과를 바탕으로 모델의 정확도를 측정하였고, 이를 통해 등급 기반의 연봉 예측에서 어떤 기법 모델이 더 우수한 성능을 보이는지를 균일한 기준으로 평가할 수 있었다.

5.3 기법별 정확도 지표

각 기법의 정확도는 다음의 Table 4와 같다.

본 연구에서 측정한 각 기법별 정확도는 Deep Learning을 제외하고 일관적인 결과를 보였다. Random Forest와 Boosting은 각각 약 70%의 평균 정확도로 가장 높은 정확도를 보였다. Random Forest는 여러 개의 Decision Tree를 결합하여 예측을 수행하는 방식으로, 복잡한 변수 간의 상호 작용을 잘 파악하고 이를 모델에 반영한 결과라고 볼 수 있다. Boosting 역시 약한 학습기를 순차적으로 조합하여 오차를 줄이는 방식으로 작동하며, 이를 통해 높은 정확도를 달성하였다. Decision Tree와 Decision Tree(Ctree)는 각각 66%의 평균 정확도로 괜찮은 성능을 보였지만, 다른 기법들에 비해 상대적으로 약간 낮았다. 이는 단일 트리구조의 한계와 과적합 문제 때문으로 추측된다. 반면에, Deep Learning은 48%의 낮은 평균 정확도를 보였다. 이는 본 연구에서 사용된 데이터셋의 크기와 복잡성이 Deep Learning 모델이 학습하면서 과적합이 발생하여 실제 예측 정확도가 낮은 것으로 생각된다.

5.4 성능 비교를 위한 일원배치 분산분석(ANOVA)

랜덤포레스트와 부스팅 기법을 포함하여, 다른 기법들과의 성능 차이도 평가하기 위해 일원배치 분산분석(ANOVA)을 수행하였다. 이 분석을 통해 각 기법 간의 평균 정확도에 통계적으로 유의미한 차이가 있는지를 검증하였다. 그 결과는 Table 5와 같이 나타났다.

Table 5: One-way ANOVA by Method

Source	SS	df	MS	F	*p*
Between Groups	0.347	5	0.069	70.982	0.000
Within Groups	0.053	54	0.001		
Total	0.399	59			

Table 4. Prediction Accuracy Results by Method

	KNN	Random Forest	Boosting	Decision Tree	Decision Tree(Ctree)	Deep Learning
Dataset1	0.659	0.696	0.682	0.652	0.657	0.352
Dataset2	0.677	0.698	0.698	0.658	0.661	0.493
Dataset3	0.663	0.695	0.692	0.659	0.656	0.498
Dataset4	0.665	0.693	0.695	0.655	0.656	0.582
Dataset5	0.666	0.690	0.686	0.656	0.661	0.467
Dataset6	0.668	0.697	0.698	0.662	0.664	0.442
Dataset7	0.674	0.699	0.698	0.661	0.662	0.563
Dataset8	0.677	0.708	0.708	0.665	0.663	0.470
Dataset9	0.662	0.694	0.694	0.647	0.649	0.526
Dataset10	0.672	0.696	0.699	0.661	0.661	0.364
Average	0.668	0.696	0.695	0.657	0.659	0.475

일원배치 분산분석(ANOVA)을 통해 각 기법 간의 평균 정확도 차이를 분석한 결과, 그룹 간 변동성(Sum of Squares = 0.347, df = 5, Mean Square = 0.069)과 그룹 내 변동성(Sum of Squares = 0.053, df = 54, Mean Square = 0.001)을 비교한 F 값은 70.982로 나타났으며, 이는 통계적으로 매우 유의미한 결과($p < 0.001$)였다. 이 결과는 기법 간에 통계적으로 유의미한 성능 차이가 있음을 시사한다. 구체적으로 어느 기법에 유의한 차이가 있는지 알아보기 위하여 Tukey 방법에 의해 사후검증을 실시하였다, Table 6을 보면 Deep Learning 기법과 Deep Learning을 제외한 기법 사이에 유의한 차이가 있다. Deep Learning을 제외한 기법은 서로 유사한 성능을 보이며, 통계적으로 유의미하지 않았으나, Deep_Learning은 다른 기법들에 비해 상당히 낮은 정확도를 보여, 성능이 상대적으로 떨어짐을 알 수 있다. 이는 95% 신뢰구간에서도 확인되었으며, 딥러닝의 정확도는 평균 0.4757로, 신뢰구간은 0.4217부터 0.5297까지로 나타났다. 다른 기법들(KNN, Random Forest, Boosting, Decision Tree, Decision Tree_C)의 평균 정확도는 0.657에서 0.696 사이로, 이들의 95% 신뢰구간은 대략 0.649에서 0.708 사이로 유사하게 나타났다. 딥러닝을 제외한 다른 기법들은 서로 유사한 성능을 보이며, 통계적으로 유의미한 차이는 없었다.

Table 6: Post-hoc Test Results by Method

	Group A	Group B
Deep Learning	0.475	
Decision Tree		0.657
Decision Tree Ctree		0.659
KNN		0.668
Boosting		0.695
Random Forest		0.696

6. 결론

본 연구는 FIFA 22 데이터를 활용하여 축구선수의 연봉등급을 예측하기 위해 다양한 빅데이터 및 인공지능 기법을 적용하고 그 성능을 평가하였다. 데이터 분석 결과, 선수의 시장가치, 종합능력치, 나이, 잠재력 등의 변수들이 연봉 예측에 중요한 역할을 하였으며, 여러 기법 중에서 Random Forest와 Boosting 기법이 가장 높은 성능(70%)을 보였다.

본 연구는 빅데이터 및 인공지능 기술을 이용한 축구선수 연봉 예측 모델의 가능성을 보여줌으로써, 관련 학문 분야에 새로운 연구 방향과 기법의 적용 가능성을 제시하였다. 특히, 기존 연구에서 다루지 않았던 다양한 기법의 비교 분석은 향후 연구자들에게 데이터 처리 및 모델 선택의 기준을 제공할 것이다. 또한, 축구 선수들의 연봉 책정 과정에서 이 연구의 결과를 통해 얻은 데이터 기반 예측 모델은 선수의 성능과 잠재력을 객관적인 데이터로 평가함으로써 공정하고 투명한 연봉 협상이 가능하다. 이러한 연봉 책정은 구단의 재정 건전성 유지에 기여하고, 선수와 구단 간의 신뢰를 증진시키며, 전체 축구 산업의 경쟁력을 강화하는 결과를 가져올 것으로 생각한다.

본 연구는 몇 가지 한계점을 가지고 있다. 첫째로, 사용된 데이터셋의 규모와 복잡도 그리고 정형화된 데이터를 예측하기엔 Deep Learning과 같은 모델의 성능이 제한적일 수 있다는 점이다. Deep Learning 모델은 스스로 패턴을 찾아 예측하기에 정형화된 데이터에서는 성능을 발휘하기에 제한적일 수 있다. 또한, 지나치게 학습을 하려는 경향이 있기에, 과적합이 발생하여 정확도 예측 저하로 이어진 것으로 생각된다. Deep Learning 모델 향상을 위해 정규화, 드랍아웃 등 추가적인 기법을 통하여 성능율 올려는 것이 필요하다. 둘째로, 최근 사우디 축구리그에서 스타급 선수들을 고액의 연봉으로 유혹하는 오일머니 현상이 연봉 예측의 복잡도를 높여, 예측의 정확도에 영향을 줄 수 있다. 향후 연구에서는 다양한 변수를 선정하고 모델의 하이퍼파라미터를 세밀하게 조정하여 예측 성능을 향상시키는 방안을 탐구해야 할 것이다. 또한, 여러 알고리즘을 조합한 앙상블 기법을 적용함으로써 모델의 강점을 극대화하고 단점을 보완하는 전략을 개발하는 것이 중요하다.

REFERENCES

[1] Noh, E. S. (2016). A Study of KBO Professional Baseball Game Prediction using Artificial Neural Networks. Graduate School of Software at Soongsil University.

[2] Han, J. S., Jung, D. H., & Kim, J. J. (2022). Predicting the OPS of KBO Batters through Big Data Analysis Using Machine Learning. *International Next-generation Convergence Technology Association, 6(1),* 12-18.

[3] Jung, Y. H. (2020). Soccer match betting analysis based on score prediction. Konkuk University.

[4] Kim, S. M., & Yoo, K. S. (2020). Analysis of the Relationship between a Batter's Performance and Discomfort Index using Big Data: focusing on the Number of Pitches and On Base Percentage. *Journal of Industrial Convergence, 18(4),* 61-66.

[5] Shin, M. S. (2003). A Comparative Study on the Salary Systems of Korean Professional Soccer and Baseball. *Korean Society for Sport Management, 7(2),* 141-155.

[6] Kang, H. C., Han, S. T., Choi, J. H., Lee, S. G., Kim, E. S., & Um, I. H. (2014). Data Mining Methodology: A Case Study Approach Using SAS Enterprise Miner. Paju: Freedom Academy.

[7] Kim, G. S. (2015). Big Data Analysis and Meta-analysis. Seoul: HanaNarae.

[8] Alfaro, E., Garcia, N., Gamez, M., & Elizondo, D. (2008). Bankruptcy forecasting: An empirical comparison of AdaBoost and neural networks. *Decision Support Systems, 45,* 110-122.

[9] Yoon, H. S. (2023). An Integrated Model of k-Means and Neural Network - For House Price Prediction. *Kyungsung University, 39(2),* 34-41.

[10] Hong, G. H. (2020). Prediction and Analysis of Suicidal Thoughts among Male and Female Adolescents Based on Random Forest Machine Learning Algorithm. *Korean Academy of Social Welfare, 72(3),* 157-180.

[11] Park, S. Y., & Jung, H. W. (2020). Exploring Predictive Factors for Middle School Students' Career Decisions: Application of Machine Learning Techniques. *Asia Journal of Education, 21(3),* 727-753.

[12] Lim, H. R., & Hong, S. P. (2023). Analysis of Factors Predicting Graduate School Enrollment among University Graduates Using Random Forest. *The Korean Society for the Study of Career Education, 36(1).*

[13] Jung, J. W., Kim, J. Y., & Oh, S. G. (2023). Comparative Study of Waste Plastic Data Pattern Classifiers Using Boosting Algorithm. *Korean Institute of Intelligent System, 33(3),* 242-248.

[14] Park, S. Y. (2022). Malicious Insider Detection Techniques Using Boosting Method of Ensemble Learning. *Korea Institute of Information Security & Cryptology, 32(2),* 267-277.

[15] Alfaro, E., Gamez, M., & Garcia, N. (2007). Multiclass corporate failure prediction by AdaBoost.MI. *Advanced Economic Research, 13,* 301-312.

[16] Kim, J. G., Kim, H. G., & Choi, S. W. (2023). A kNN Algorithm-Based Driver Facial Identification Model Applicable to Car-Sharing Services. *Korean Institute of Information and Communication Engineering, 27(1),* 658-660.

[17] Lim, H. C., & Lee, S. S. (2022). Interference Mitigation Techniques Among Ultrasound Sensors Using KNN Algorithm. *Institute of Korean Electrical and Electronics Engineers, 26(2),* 169-175.

[18] H. S. Choi, Y. H. Cho. (2019). Analysis of Security Problems of Deep Learning Technology. *Journal of the Korea Convergence Society, 10(5),* 9-16.

[19] H. J. Mooi, G. H. Kim. (2019). A Survey on Deep Learning based Face Recognition for User Authentication. *Journal of Industrial Convergence, 17(3),* 23-29.

정현성(Hyeon-Seong, Jeong) [정회원]

• 2024년 3월 ~ 현재 : 서강대학교 메타버스학과(박사과정)
• 관심분야 : 빅데이터, 인공지능, 메타버스, 블록체인
• E-Mail : j018550@gamil.com

김진화(Jin-Hwa Kim) [정회원]

• 2003년 3월 ~ 현재 : 서강대학교 경영학과 교수
• 관심분야 : 빅데이터, 인공지능
• E-Mail : jinhwakim@sogang.ac.kr

현대원(Dae-Won Hyun) [일반회원]

• 2003년 2월 ~ 현재 : 서강대학교 메타버스학과 교수
• 관심분야 : 메타버스, 미디어
• E-Mail : dyhyun@sogang.ac.kr