KSME

http://doi.org/10.21797/ksme.2022.24.3.005

ISSN 1229-4225
eISSN 2671-9134

한국체육측정평가학회지. 제24권 3호, 2022, 53-69

2022 FIBA 남자농구 아시안컵 경기결과를 활용한 머신러닝 분류 모형의 예측 성능 비교*

예원진 · 이성노* 한양대학교

국문초록

이 연구의 목적은 2022 FIBA 남자농구 아시아컵 경기대회의 공식기록(box score)을 사용하여 전통적 측면에서의 통계적 방법, 데이터마이닝 기법, 머신러닝의 기법을 활용하여 예측 성능을 비교한 것이다. 본 연구의 대상은 2022 FIBA 남자농구 아시아컵 경기대회의 공식기록을 통해 얻어지는 기록 중에서 총 72개 경기기록이었으며, 총 20개의 변수를 통해 경기 승패 결과 예측을 하였다. 남자농구 아시아컵 경기대회의 승패 결과를 예측하기 위해 KNN(K Nearest Neighbor), Decision Tree, Support Vector Machine(SVM), Logistic Regression, Random Forest 5가지 분류 모델을 사용하였다. 이 연구의 자료수집과 처리를 위하여 통계프로그램 Python 3.10.1 버전을 라이브러리와 함께 사용하였고, 얻은 결과는 다음과 같다. 첫째, 모델별 예측 결과에서는 SVM 모델이 KNN, Decision Tree, Random Forest, Logistic Regression 모델보다 최적의 예측 성능을 나타냈고 86.67%의 예측 정확도 및 0.868의 F1 점수를 보였다. 둘째, Random Forest 분류 모델을 이용하여 2022 남자농구 아시아컵 경기대회의 승패 결과를 예측했을 때, 데이터 세트의 샘플 개수가 충분하지 않기 때문에 과적화(Overfitting) 현상이 발생했다. 이 연구의 연구 결과를 토대로 향후 사례 수를 증가하여 더 많은 데이터가 모델의 정확도를 높이는 동시에 과적합 가능성을 줄일 수 있다고 사료되었다. 그리고 더 정확한 예측 결과를 얻기 위하여서 머신러닝뿐만 아니라 딥러닝과 관련한 연구도 필요할 것으로 판단되는 바이다.

주요어: 머신러닝, 남자농구 아시아컵, 예측, 분류 모형

Ⅰ. 서론

최근 사회적으로 가장 큰 변화는 세계경제포럼에서 제시한 '4차 산업혁명'의 변화가 빠르게 다가오고 있다는 점이다(최형준, 2020). 조정환(2012)은 '스포츠 빅데이터 활용과 전망'의 연구를 통하여 4차 산업혁명 시대에 스포츠 및 체육학 분야에서 활용될 빅데이터 분석에 대해서 처음 언급한 바 있었다. 특히 빅데이터 분석에 사용되는 자료는 방대하고, 다양하며, 빠르게 처리되어야 할 뿐만 아니라 분석한 내용이 정확하고, 유의미한 가치를 지녀야 한다(최형준, 2020).

FIBA 남자농구 아시안컵 경기대회는 FIBA 아시아가 주관하는 국가 대항 농구대회로서, 아시안 게임 농구 경기와 더불어 아시아 최강의 농구 국가대표팀을 가리는 대회이다. 1960년 마닐라에서 열린 1회 대회를 제외하곤 홀수 해에 2년마다 개최되었으며, 2022

* 교신저자 이성노(snl743@hanyang.ac.kr)
서울시 성동구 왕십리로 222 한양대학교

년 인도네시아 대회까지 총 30회의 대회를 치렀다. 남자농구 아시안컵 경기 대회에서는 각 팀, 선수, 경기 등에 대한 통계를 많이 만들어낸다(Gu, Foster, Shang, & Wei, 2019; Haghighat, Rastegari, Nourafza, Branch, & Esfahan, 2013). 따라서 남자농구 아시안컵 경기 대회는 빅데이터 분석의 이상적인 영역이다(Nguyen, Nguyen, Ma, & Hu, 2021; Cao, 2012; Jain & Kaur, 2017). 남자농구 아시안컵의 데이터는 서로 다른 데이터 분석 기술을 연구하여 유의미한 자료를 얻을 수 있을 뿐만 아니라(Horvat & Job, 2020; Cai, Yu, Wu, Du, & Zhou, 2019), 남자농구 아시안컵의 결과를 예측하는 데도 사용될 수 있다(Lam, 2018; Horvat, Havaš, & Srpak, 2020; Dubbs, 2018; Song, Zou, & Shi, 2020; Li & Xu, 2021).

최근 스포츠 분야에서도 4차 산업혁명시대의 키워드인 빅데이터와 인공지능을 활용하여 향후 일어날 일에 대해서 예측하고자 하는 연구가 진행되고 있다(최형준, 이윤수, 2019). 스포츠 과학의 발전은 스포츠 경기나 선수를 분석하는데 그치지 않고, 경기 결과나 선수의 동작을 과학적이고 통계적인 방법으로 예측하는 것에도 많은 학자의 관심이 집중되고 있다(최형준, 김주학, 2006). 예측은 과거에 일어난 일을 기반으로 확률적인 방법이나 비확률적인 방법에 의해서 자료를 재해석하고 분석하여 도출되는 자료의 특성에 대한 것이다(박상찬, 2017). 농구와 관련된 연구자들에게는 다양하고 객관적인 경기력을 객관적으로 분석하기 위한 토대를 마련하는 계기가 됨에 따라 다양한 통계적 분석 방법을 통하여 승패를 예측하기 위한 다양한 통계적 모형을 적용하기에 이르고 있다(허종관, 김세중, 도재현, 2016). 예를 들면 경기 결과(승·패 또는 최종 득점)를 예측하는 다중회귀분석(multiple regression), 로지스틱 회귀분석(logistic regression) 등의 통계적 모형을 규정하고 분석하는 방법이 있다(구승환, 김현수, 장성용, 2009; 김세형, 강상조, 박재현, 김혜진, 2008; 장효진, 곽현, 최승회, 2015). 이와 함께 단순히 경기력 변인들을 통한 경기 결과의 예측에 목적을 둔 머신러닝 분석이 이용되고 있다(Prasetio & Harli, 2016; Bunker & Susnjak, 2022).

머신러닝은 데이터 패턴을 학습하여 컴퓨터 프로그램의 성능을 자동으로 향상시키는 인공지능(AI)의 하위 집합으로, 다양한 분야에서 성공적으로 활용되고 있다(Nguyen et al., 2021). 머신러닝은 예제 데이터나 이전 경험을 사용하여 성능 기준을 최적화하기 위하여 컴퓨터를 프로그래밍하는 과정이다(Alpydin, 2010). 머신러닝의 목표는 미래 트렌드를 예측하고, 보이지 않는 데이터를 분류하거나, 데이터 세트의 숨겨진 패턴을 발견하는 데 활용할 수 있는 수학적 모형을 개발하는 것이다(Cao, 2012). 예를 들면 날씨, 소비자 행동 혹은 질병의 존재를 예측할 수 있는 것들을 말이다(Maier, Meister, Trösch, & Wehrlin, 2018). 스포츠 분야에서 머신러닝의 첫 번째 연구는 Purucker가 1996년 미국 내셔널풋볼리그(NFL) 경기 결과를 비지도 학습으로 예측하였으며(Purucker, 1996), 이를 통해 많은 사람이 머신러닝에 대한 관심이 높아졌다(Bunker & Susnjak, 2022). 최근 몇 년 동안 농구 분야에서 머신러닝의 활용이 많은 주목을 받고 있다(Li & Xu, 2021). 머신러닝 모형은 코칭스태프와 스포츠매니저가 경기 결과를 예측하는 데, 선수나 팀의 경기력을 분석할 때 혹은 정확한 스포츠 베팅 참조를 제공할 때 도움을 준다(Horvat et al., 2020; Tichy, 2016). 통계와 머신러닝의 진전으로 많은 첨단 예측 모형이 탄생하여 기존의 통계 도구보다 더 정확한 예측 성능을 제공할 수 있다는 것이 입증되었으며(Lopez & Matthews, 2015), 많은 기업이 머신러닝에 투자하여 농구 경기 결과를 예측하고 있다(Bunker & Thabtah, 2019).

최근 여러 머신러닝 분류 모형들은 ANN, Random Forest, Decision tree, KNN, Bayesian method,

Logistic regression, SVM, Fuzzy Methods 등이 농구 경기 결과 예측에 사용되고 있다(Haghighatet al., 2013). 머신러닝을 활용한 선행연구는 김세형, 강상조, 박재현, 김혜진(2008)은 2006~2007시즌 KBL 리그 팀의 승/패 결과를 Logistic regression과 Decision tree를 통해 예측하였다. Zimmermann, Moorthy & Shi(2013)는 NCAA 농구 경기에 대한 몇 가지 머신러닝 방법을 비교하였는데, 그중에서 neural network, Decision trees, Rule learners (Ripper), Naïve Bayes, Random Forest 모형이 포함되었다. Thabtah, Zhang & Abdelhamid(2019)는 SVM, Logistic regression, Naïve Bayes, ANN을 사용하여 NBA 경기 결과를 예측한 결과, Logistic regression 모형이 우수하다는 것으로 확인되었다. Lin, Short & Sundaresan(2014)는 Logistic regression, SVM, aDaboost, Random Forest, Gaussian Naïve Bayes를 사용하여 NBA 경기의 승패를 예측한 결과, Random Forest 모형이 우수하다고 제언했다.

최근 농구영역에서 다양한 통계 및 데이터마이닝, 머신러닝 기법들을 대상으로 분류의 정확도, 예측력 등을 비교한 연구들이 보고되고 있다. Torres & Hu(2013)는 Multilayer perceptron—back propagation neural network, Linear Regression과 Maximum Likelihood를 사용하여 NBA의 경기결과를 예측하였고 가장 좋은 예측 정확도는 70%로 나타났다. Cao(2012)는 Logistic Regression, Naïve Bayes, SVM과 Multilayer perceptron neural network 모형을 사용하여 NBA 경기 결과를 예측하였고 가장 좋은 예측 정확도는 69.67%로 나타났다. Kravanja(2013)은 SVM 및 Logistic Regression을 이용하여 NBA 경기를 예측하였고 각각 70.01%, 69.73%의 예측 정확도를 보였다. Miljković, Gajić, Kovačević & Konjović(2010)은 Decision trees, KNN, Naïve Bayes와 SVM을 사용하여 NCAA 경기 결과를 예측하였고 가장 좋은 예측 정확도는 67%로 나타났다. Li, Wang & Li(2021)는 연구자들이 농구영역에서 경기 결과에 대한 예측 방법을 이미 구축해 놓았지만, 일반적인 단점은 예측 정확도가 낮다는 것으로 제언하였다. 따라서 농구 경기 결과에 대한 낮은 예측 정확도는 신뢰할 수 있는 예측을 얻기 위한 추가 연구가 필요성이 있다고 생각되었다(Haghighat et al., 2013; Horvat & Job, 2020).

선행연구를 종합하면 다른 기법에 비하여 KNN, Decision Tree, Logistic Regression, SVM이 농구 경기 결과를 예측하는 신뢰할 수 있는 모형으로 활용될 수 있다. 하지만 KNN, Decision Tree, Logistic Regression, Random Forest, SVM 5 가지 분류 모형을 대상으로 분류의 정확도, 예측력 등을 비교한 연구는 미비한 실정이다. 또한, 분류 모형으로 농구 경기 결과를 예측하는 연구 대상이 NBA, NCAA 등만 하였다(Zuccolotto, Manisera, & Sandri, 2018; Lam, 2018; Horvat, Havaš, & Srpak, 2020; Pai, ChangLiao, & Lin, 2017; Thabtah et al., 2019). 남자농구 아시안컵 경기대회에 대해 경기 결과를 예측하는 연구가 제한적이다(WentingWang, 2014).

따라서 이 연구의 목적은 FIBA 남자농구 아시안컵 경기대회의 공식 기록을 기반으로 승패를 예측하는데 있어서 전통적 측면에서의 통계적 방법, 데이터마이닝 기법, 머신러닝의 기법을 적용함으로써 KNN, Decision Tree, Logistic Regression, Random Forest, SVM 5 가지 분류 모형의 예측 성능을 비교한 것이다. 또한, 남자농구 아시안컵 경기대회의 경기 결과를 얼마나 예측할 수 있는지를 알아보고자 한다.

Ⅱ. 연구방법

1. 연구대상

이 연구에서는 2022년 제30회 남자 농구 아시안컵의 경기 결과를 예측하기 위하여 2022년 7월 12

일부터 7월 24일까지 인도네시아의 수도 자카르타에서 열린 남자 농구 아시안컵 경기대회에 참가한 16개 팀이 겨룬 총 36 경기의 경기내용을 대상으로 두 팀의 공식 기록을 연구 대상으로 선정하였다(n=72).

2. 자료수집

이 연구에서 사용된 연구 자료는 국제농구연맹 공식 사이트에서 제공되고 있는 공식 기록의 Box Score(https://www.fiba.basketball/asia-cup/2022)를 중심으로 수집되었다.

3. 측정변수

저우차우와 최형준(2020), Ball ı & Özdemir(2021) 등 학자들은 농구 경기 결과를 예측하는 데 box score에 있는 데이터를 활용하고 있으며 Lin et al.(2014)는 논문에서 box score 데이터의 과학성과 유효성을 검증하였다. 따라서 본 연구에서 2022년 제30회 남자 농구 아시안컵 홈페이지에서 제공된 데이터 중에서 결측값이 없는 21개의 변수를 선택하여 측정변수로 선정하였다. 선정된 측정변수를 20개 독립변수와 1개 종속변수로 구분하였고 변수의 값을 다음과 같이 구분하였다. 독립변수는 득점, 슛 성공수, 슛 시도수, 슛 성공률, 2점슛 성공수, 2점슛 시도수, 2점슛 성공률, 3점슛 성공수, 3점슛 시도수, 3점슛 성공률, 자유투 성공수, 자유투 시도수, 자유투 성공률, 공격 리바운드, 수비 리바운드, 어시스트, 개인 파울, 턴 오버, 가로채기, 블록으로 구분하였고, 종속변수는 승패(“승리”와 “패배”)로 구분하였다. 〈표 1〉은 이 연구에서 사용된 측정변수를 독립변수와 종속변수로 구분하여 정리한 표이다.

표 1. 연구 변수

내용	변수	변수의 약호
독립변수	득점	PTS
	슛 성공수	FGM
	슛 시도수	FGA
	슛 성공률	FG%
	2점슛 성공수	2PTSM
	2점슛 시도수	2PTSA
	2점슛 성공률	2PTS%
	3점슛 성공수	3PTSM
	3점슛 시도수	3PTSA
	3점슛 성공률	3PTS%
	자유투 성공수	FTM
	자유투 시도수	FTA
	자유투 성공률	FT%
	공격 리바운드	OREB
	수비 리바운드	DREB
	어시스트	AST
	개인 파울	PF
	턴 오버	TO
	가로채기	STL
	블록	BLK
종속변수	승패	Won/Lost

4. 데이터세트 분할

Kannan, Kolovich, Lawrence & Rafiqi(2018), Nguyen et al.(2021)과 Cao(2012)의 논문에서 사용했던 데이터 세트 분할 방법을 참고하였다. 이 연구의 데이터 세트 중에서 80%의 데이터는 훈련용으로(train), 20%의 데이터는 테스트용으로(test) 나누었다. 모형 예측 시 먼저 각 분류모형으로 훈련 세트에 데이터를 적합하게 하고 지도학습 실시하였으며. 예측 세트에서 예측한 후 모형 예측 성능의 비교를 통하여 최적의 예측모형을 도출하였다.

5. 예측 모형

이 연구에서는 머신러닝의 KNN, Decision Tree, Logistic Regression, SVM, Random Forest 분류모형을 선택하였으며 2022년 제30회 남자 농구 아시안컵의 경기 결과를 예측할 것이다. 각 분류 모형에 대한 간략한 설명은 다음과 같다.

1) KNN

K-NEAREST NEIGHBOUR(KNN)은 비모수(non

parametric) 지도 머신러닝 모형이며(Korhonen & Kangas, 1997; Elish, 2014), 게으른 학습 알고리즘으로 분류되어서(Chen, Jhou, Lee, & Lu, 2021) 분류 및 회귀 문제를 해결하는 데 사용할 수 있다(Horvat et al., 2020; Chen et al., 2021). KNN 모형의 설명도는 〈그림 1〉과 같다. 분류의 경우, 알 수 없는 인스턴스(instance: q)가 주어졌을 때, KNN 모형은 패턴 공간에서 q에 가장 가까운 k 개 훈련 인스턴스를 검색하였으며, 이 k개의 훈련 인스턴스는 q의 k개의 '최근접 이웃'이라고 한다(Elish, 2014). 그리고 q의 카테고리는 이웃의 영향을 기반으로 가중치를 설계할 수 있으며, 그중에서도 가장 가까운 이웃이 다른 이웃보다 더 큰 영향을 미쳤다(Horvat et al., 2020).

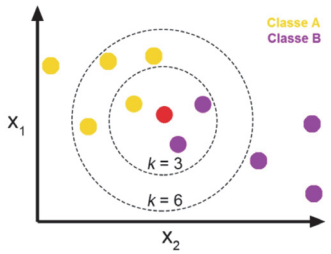

그림 1. KNN 모형의 설명도

KNN 알고리즘은 적용하기 위해서는 질의와 학습 데이터 간의 거리를 계산하는 방법이 있어야 한다. 데이터 속성이 수치인 경우 질서와 학습 데이터와 거리를 측정하기 위해 유클리드 거리(Euclidian distance)를 사용한다(Harrington, 2012). 유클리드 거리는 식(1)과 같이 정의 된다.

$$d_{Euclidean}(x,y) = \sqrt{\sum_{i}^{n}(x_i - y_i)^2} \quad (1)$$

2) Decision Tree

Decision Tree도 지도 머신러닝 모형이며(Johnson et al., 2018; Jin & Deng, 2018; Mendonça, Vieira, & Sousa, 2007), 분류 및 회귀 문제를 해결하는 데 사용할 수 있다(Tanha, Someren, & Afsarmanesh, 2017). Decision Tree도 Random Forest의 기본 구성요소이다(Liu, 2021). Decision Tree의 주요 목표는 학습된 결정 규칙에 따라 목표 변수를 예측할 수 있는 훈련 모형이다(Horvat et al., 2020). Decision Tree는 기본적으로 흐름도와 같은 트리 구조이다(Han, Kamber, & Pei, 2012). 데이터를 분류하기 위하여 전체 데이터 세트는 트리의 상단에서 트리의 하단으로 확장되는 더 작은 하위 집합으로 분할되었다. 즉, 루트 노드부터 리프 노드까지다. 그중에서 트리의 최상위에 있는 첫 번째 노드를 루트 노드라고 하며, 데이터의 분류 예측 결과를 포함하는 노드를 리프 노드라고 하며, 하위 노드가 다른 하위 노드로 분할되는지 여부를 결정하는 노드를 결정 노드라고 하다(Noor, Anwar, & Dey, 2019; Passi & Pandey, 2018). Decision Tree 모형의 설명도는 〈그림 2〉와 같다(Zhao, 2021). 선행 연구에 따르면 모든 복잡하고 방대한 양의 Decision Tree가 더 정확한 예측 성능을 얻을 수 있는 것은 아니다(Zhao, 2021). Decision Tree의 과적합을 방지하기 위하여서 가지치기해야 한다. 즉, 더 정확한 결과를 생성하기 위하여서 덜 신뢰할 수 있는 분기를 제거해야 한다(Niblett & Bratko, 1987; Osei-Bryson, 2007).

Decision Tree 분석을 수행하는 알고리즘들은 여러 가지가 있다. 대표적인 Decision Tree 알고리즘은 ID3, C4.5, CART, CHAID 등이 있다. 이 연구에서는 python의 scikit-learn 라이브러리를 사용하였다. 따라서 Decision Tree 적용 시 CART 알고리즘을 적용하였다.

CART(Classification And Regression Tree)는 지니 계수(Gini Index)를 이용하여 불순도(impurity)를 측정하며, 부모마디로부터 자식마디가 2개만 형성되는 이진분류(binary split)에 기반한 알고리즘이다. CART는 생성되는 규칙을 해석하기 쉽고, 연속형 변

수와 범주형 변수를 모두 이용할 수 있다는 장점이 있다. 지니 계수는 n개의 원소 중에서 임의로 2개를 추출하였을 때, 추출된 2개가 서로 다른 그룹에 속할 수 있는 확률을 의미한다. 지니계수의 감소량이 계산되면, 알고리즘의 마지막 과정으로 지니 계수를 가장 감소시켜 주는 분류 변수와 최적 분리를 자식 마디로 선택한다. 지니 계수의 공식은 식(2)와 같다. S는 데이터 세트이고 c는 분류의 개수이다.

$$G(S) = 1 - \sum_{i=1}^{c} p_i^2 \quad (2)$$

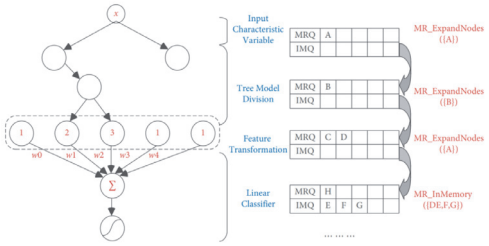

그림 2. Decision Tree 모형 설명도

출처:Sports Enterprise Marketing and Financial Risk Management Based on Decision Tree and Data Mining

3) Logistic Regression

Logistic Regression도 지도 머신러닝 모형이며(Haghighat et al., 2013), 간단한 계산과 해석, 그리고 신뢰할 수 있는 예측 결과 때문에 많은 관심을 받았다(Hosmer & Lemeshow, 2000). 이는 선형 회귀와 마찬가지로 특징적인 선형 조합에 의존하고 로지스틱 함수(Ye, 2003)를 통해 이원 분류된다. Binary Logistic Regression은 Multiple Logistic Regression으로 일반화하여 다중 분류 문제를 훈련하고 예측하는 데 사용할 수 있다. 일반적인 Logistic Regression은 2 분류(Binary Classification) 문제만 해결할 수 있다. 다중 카테고리의 분류를 구현하려면 다중 분류 문제에 적응하도록 Logistic Regression을 개선해야 한다(Chovanec, 2021).

Logistic Regression 분석은 이분화 된 데이터에 대하여 결과를 예측하는 함수식을 생성한다. 식(3)은 회귀 함수식이며 P(x)값은 0과 1사이의 값을 가지며 임의의 경계를 기준으로 이분화 된 데이터의 결과를 예측한다.

$$P(x) = \frac{1}{1 + e^{-(a + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \cdots \beta_d x_d)}} \quad (3)$$

식 (3)에서 $\beta_1 \sim \beta_d$는 입력 변수$(x_1 \sim x_d)$의 회귀계수이며, $e^{\beta_1} \sim e^{\beta_d}$ 값은 변수의 위험도를 나타낸다.

4) SVM

Support Vector Machines(SVM)는 러시아의 수학자 Vapnik(1995)이 구조적 위험 최소화(SRM: Structural Risk Minimization) 원리를 이용하여 도입된 커널(kernel) 함수 기반 학습 알고리즘이다(Boser, Guyon, & Vapnik, 1992; Bromley et al., 1993; Drucker, Burges, Kaufman, Smola, & Vapnik, 1997). 또한, 과적합이 덜 발생하는 영향력 있고 효율적인 지도 머신러닝 모형이다(Jain & Kaur, 2017; Pai et al., 2017; Bishop & Nasrabadi, 2006). 많은 전문가는 SVM을 머신러닝에서 가장 강력한 분류 기술 중 하나로 간주하며(Cortes & Vapnik, 1995; Vapnik, 1995), 선형적이거나 비선형적인 분류 문제를 해결하는 데 사용할 수 있다(Suthaharan, 2016). 비선형 데이터에 대해서는 SVM은 커널 함수 파라미터를 이용하여 기존의 비선형 특징을 더 높은 차원 공간으로 사상한다(Passi & Pandey, 2018; Nguyen et al., 2021). 다차원 공간상에서 학습 자료가 $\{x_i, y_i\}$, $i = 1, \cdots, r$, $y_i \in +1, -1$과 같을 때, 두개의 클래스를 구분하는 초평면(hyperplane)은 여러 개 있지만 최적의 초평면은 하나만 존재한다. 이러한

최적의 초평면은 각 분류 데이터 중에서 분리하는 초평면에 가장 가까운 자료 사이의 거리를 최대화할 수 있어야 하며 식(4)와 같이 정의 된다.

$$(\vec{w} \cdot \vec{x}) + b = 0 \quad (4)$$

$\vec{x}$는 초평 상의 한점, $\vec{w}$는 초평면에서의 법선이며 b는 편향이다.

자료가 선형분리 가능한 경우에는 두개의 분류를 정의하는 초평면은 식(5)와 같이 저의 내릴 수 있으며 이러한 두개의 초평면 상의 학습 자료는 지지 벡터라 한다.

$$y_i(\vec{w} \cdot \vec{x_i} + b) - 1 \geq 0 \quad (5)$$

또한, 두 개의 초평면은 두 평면 사이의 여백(margin) $\frac{2}{\| \vec{w} \|}$을 최대화하여야 하므로, 두 분류의초평면을 구하기 위해서는 식(5)를 제약식으로 가지는 식(6)의 목적식에 대한 최적화 문제가 된다.

$$L(W) = \min\left(\frac{1}{2} \| \vec{w} \|^2\right) \quad (6)$$

최종적으로 임의의 입력 패턴 x가 주어질 때, 판별함수는 다음과 같다.

$$f(x) = \sum a_i y_i K(x_i, x_y) + b$$

여기서 a_i는 라그랑제 승수이고, K()는 커널함수이다. SVM 모형의 설명도는 〈그림 3〉과 같다.

데이터를 더 잘 분리하기 위해 몇 가지 커널 함수를 개발하였다. 가장 흔한 기존 커널 함수에는 선형 커널(linear kernel), 다항식 커널(polynomial kernel,), 가우스 커널(Gaussian (RBF) kernel) 및 시그모이드 커널(sigmoid (MLP) kernel)이 포함되었다. Schölkopf,

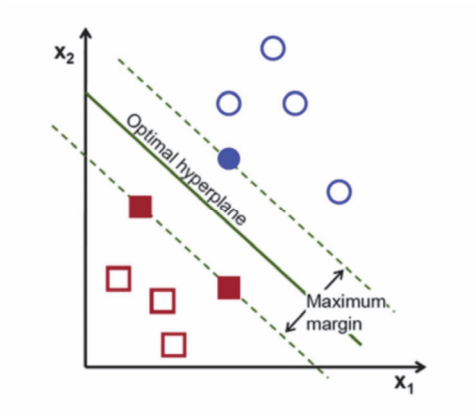

그림 3. SVM 모형 설명도
출처:The use of machine learning in sport outcome prediction: A review

Smola, & Bach(2002)는 다른 커널 함수의 일반화 능력을 연구하기 위하여서 커널 함수와 정규화 연산자 사이의 관계를 수립하였으며, RBF 커널 함수를 사용하는 SVM이 더욱 정확한 예측 성능을 얻을 수 있다고 결론지었다. 또한, 다른 학자들도 논문에서 이러한 결론을 지지하였다(Levandoski & Lobo, 2017 ; Cai et al., 2019). 그중에서 Levandoski & Lobo(2017)의 연구결과는 〈그림 4〉와 같다.

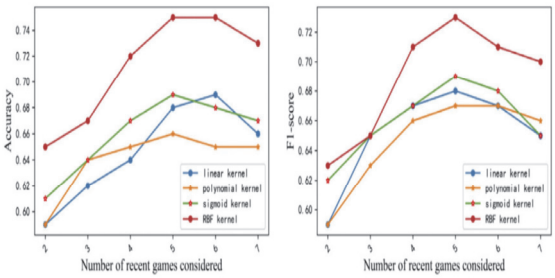

그림 4. 네 가지 커널 함수의 예측 정확도 비교
출처:A hybrid ensemble learning framework for basketball outcomes prediction

5) Random Forest

통계학과 머신러닝에서 앙상블 방법은 몇 개의 모

형을 종합하여 각각의 기본 모형의 장점을 이용하여 더 나은 모형을 구축하였다(Mendes-Moreira, Soares, Jorge, & Sousa, 2012; Sagi & Rokach, 2018). Liu(2021)는 논문에서 앙상블 방법은 단일 모형에 비하면 더 나은 예측 성능을 제공한다고 제시하였다. 2001년 Breiman이 제안한 Random Forest(Groll, Schauberger, & Van Eetvelde, 2019)는 인기가 있고 매우 효과적인 앙상블 학습 방법이다(Soliman, Misbah, & Eldawlatly, 2017; Noor, Anwar, & Dey, 2019; Pathak & Wadhwa, 2016; Mueller, 2020). Random Forest 모형의 강점은 특히 복잡한 결정 경계를 고려할 수 있다는 점이며, 회귀 및 분류 문제에서 정확한 예측을 생성하는 것으로 나타났다(Lin et al., 2014; Lessmann, Sung, & Johnson, 2010; Nedellec, Cugliari, & Goude, 2014). Random Forest는 많은 수의 Decision Tree를 기본 모형으로 구축하였으며, 서로 다른 Decision Tree는 훈련할 때 서로 다른 랜덤 하위 집합을 훈련하며, 그의 예측 결과는 여러 트리의 예측 평균값이다(Levandoski & Lobo, 2017; Mueller, 2020). 이러한 랜덤성(randomness)은 Random Forest 모형이 단일 Decision Tree보다 더 우수한 예측 성능을 갖도록 도와준다. 또한 매우 깊이 성장한 Decision Tree는 훈련 데이터에 대해 과적합 (overfitting)하게 되므로 Random Forest는 bagging(bootstrap aggregating)을 통하여 Decision Tree의 과적합 문제를 해결할 수 있고 Decision Tree 예측의 분산도 감소시킬 수 있다(Nguyen et al., 2021; Pathak, Wadhwa, 2016; Groll et al., 2019). Random Forest 모형의 설명도는 〈그림 5〉와 같다.

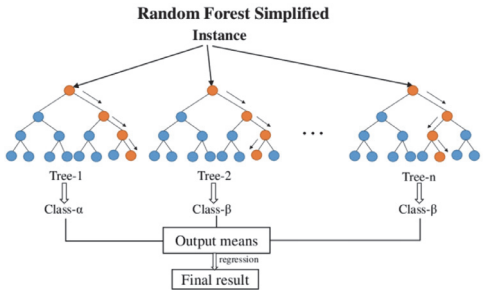

그림 5. Random forest 모형 설명도
출처:The use of machine learning in sport outcome prediction: A review

6. 자료처리

1) 통계 프로그램

이 연구는 2022년 제30회 남자 농구 아시안컵 경기 대회에 참가한 팀이 겨룬 총 36경기의 경기내용을 바탕으로 머신러닝 분류 모형을 사용하여 승패 결과를 예측하기 위하여 통계 프로그램인 python 3.10.1 버전을 사용하였다. python 3.10.1 버전은 데이터를 효율적으로 처리할 수 있는 모듈을 제공하는 것뿐만 아니라, 자료처리에 관한 모듈도 사용이 가능하다(최형준, 이윤수, 2020). 또한, 유용한 라이브러리를 포함하고 있어서 데이터 추출과 훈련/테스트 머신러닝 분류모형이 가능하다(Yezus, 2014).

따라서 Yezus(2014)의 논문에 사용되는 python 라이브러리를 참고하였다. 이 연구에서 사용한 python의 라이브러리는 scikit-learn, Pandas, Matplotlib이었다.

2) 모형 예측 정확도 평가 지표

이 연구의 정확도 평가지표는 남자 농구 아시안컵 팀의 잠재적인 경기 결과를 예측하였고 실제 발생한 경기 결과에 따라 모형의 예측 정확도를 계산하고 평가하였다(Levandoski & Lobo, 2017). 이 연구에서는 머신러닝 분야에서 사용되는 모형을 평가지표로 선정하였고 서로 다른 머신러닝 모형의 예측 정확도를 평가하고 비교하였다. 평가지표에는 훈련 세트에 대한 모형의 예측 정확도(Jadhav & Channe, 2016; Hassonah, Rodan, Tamimi, & Alsakran, 2019), 테스트 세트에 대한 모형의 예측 정확도(Jadhav & Channe, 2016; Hassonah et al., 2019), F1 점수(Nguyen et al., 2021)가 포함되었다.

정확도(Accuracy)는 실제 데이터에서 예측 데이터가 얼마나 같은지를 판단하는 지표이다. 또한, 직관적으로 모델 예측 성능을 나타낼 수 있다(Thabtah et al., 2019). 즉 정확도가 높을수록 예측 성능이 정확하다(Jadhav & Channe, 2016; Hassonah et al., 2019). 정확도의 계산식은 식(7)과 같다. 계산식에서는 TP는 True Positivie, TN은 True Negative, FP는 False Positive, FN은 True Negative이다.

$$accuracy = \frac{TP + TN}{TP + TN + FP + FN} \quad (7)$$

F1 점수는 정밀도와 재현율의 가중조화 평균값이다(Luu et al., 2020). F1의 최댓값은 1, 최솟값은 0으로 F1 점수가 높을수록 모형 예측 성능이 효과적이다(Nguyen et al., 2021; Thabtah et al., 2019). F1 점수의 계산식은 식(8)과 같다. 계산식에서는 P는 정밀도 Precision이고 R은 재현율 Recall이다.

$$F_1 = \frac{2PR}{P + R} \quad (8)$$

Ⅲ. 연구결과

1. 모형의 파라미터 결정

1) KNN

KNN의 초 파라미터(Hyperparameters) 조정은 먼저 k를 1-20으로 설정하였고 5겹 교차 검증(5-fold cross validation)을 실시하였다. k의 각 값에 대한 평균 예측 정확도(mean accuracy)를 도출하였다. 그 결과 〈그림 6〉은 k=13일 때 KNN 모형의 예측 성능이 가장 좋은 것으로 확인되었다. 따라서 K의 값을 13으로 조정하였고 이웃마다 예측한 영향을 유클리드 거리에 따라 가중되었다(Levandoski & Lobo, 2017).

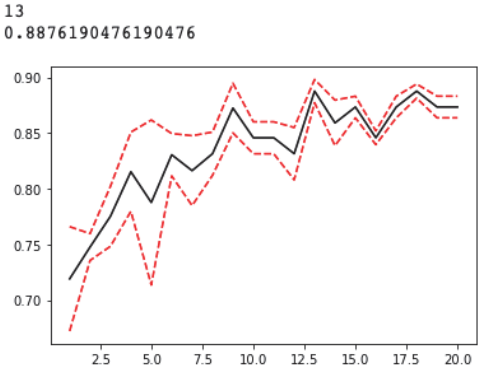

그림 6. 각 k의 평균 예측 정확도

2) Decision Tree

Decision Tree의 초 파라미터 조정은 최대 깊이(max-depth)를 1-7로 설정하였고 각 깊이에서 Decision Tree의 예측 성능을 도출하였다. 그 결과 〈표 2〉는 최대 깊이가 2, 3, 4, 5, 6, 7인 때 Decision Tree의 예측 성능이 가장 좋은 것으로 확인되었다. 하지만, Decision Tree 모형은 매우 깊이 성장하면 과적합(overfitting) 문제가 생길 수 있다. 따라서 Decision Tree의 최대 깊이를 2로 조정하였다. 또한, Horvat et al.(2020)은 논문에서 Decision Tree에 대한 가중치 설정을 참고하였고 criterion을 "entropy"로 결정하였다.

표 2. Decision Tree의 파라미터 조정

최대 깊이	예측 정확도
1	0.73
2	0.8
3	0.8
4	0.8
5	0.73
6	0.8
7	0.8

3) Logistic Regression

Logistic Regression의 초 파라미터 조정은 Logistic

Regression의 최적 예측 성능을 얻기 위하여서 Maier, Meister, Trösch, & Wehrlin(2018)의 논문에서 Logistic Regression에 대한 초 파라미터 설정을 참고하였고 본 연구의 초 파라미터 C를 0.01로(C=0.01) 확정하였다. 또한 Logistic Regression의 초 파라미터 중에서 solver 파라미터는 Logistic Regression의 손실 함수에 대한 최적화 방법을 결정한다. solver 파라미터에는 newton-cg, lbfgs, liblinear, sag, saga 5개의 알고리즘이 포함된다. 본 연구에서는 기본 liblinea 알고리즘을 사용하기로 선택하였다. 그는 좌표 하강(coordinate descent) 방법을 사용하였고 손실 함수를 반복적으로 최적화하였다.

4) SVM

SVM의 초 파라미터 조정은 SVM의 최적 예측 성능을 얻기 위하여서 SVM의 선형 커널, 다항식(POLY) 커널, RBF커널, 시그모이드(MLP) 커널을 별도로 훈련하였다. 각 커널은 훈련세트와 예측세트에 대한 예측 정확도를 도출하였다. 훈련 결과 〈그림 7〉은 다항식 커널은 다른 커널 함수보다 예측 성능이 더 우수하다는 것으로 확인되었다. 따라서 이 연구에서 선택한 SVM 커널 함수는 다항식 커널 함수이다.

```
the accuracy under kernel linear Train set Accuracy is 1.000000
the accuracy under kernel linear Test set Accuracy is 0.800000
the accuracy under kernel poly Train set Accuracy is 0.947368
the accuracy under kernel poly Test set Accuracy is 0.800000
the accuracy under kernel rbf Train set Accuracy is 1.000000
the accuracy under kernel rbf Test set Accuracy is 0.666667
the accuracy under kernel sigmoid Train set Accuracy is 0.912281
the accuracy under kernel sigmoid Test set Accuracy is 0.800000
```

그림 7. 각 커널의 예측 정확도

5) Random Forest

Random Forest의 초 파라미터 조정은 최적의 예측 성능을 얻기 위하여 Decision Tree 수를 1-200으로 설정하였고 10겹 교차 검증(10-fold cross validation)을 실시하였다. 그 결과 〈그림 8〉은 Decision Tree 수=90일 때 Random Forest 모형의 예측 성능이 가장 좋은 것으로 확인되었다. 따라서 90개의 Decision Tree를 설정하였고 Random Forest 모델을 구축하였다.

2. 모형 훈련세트와 예측세트에서 예측 정확도

KNN, Decision Tree, Logistic Regression,

```
0.9017857142857144 90

Out[26]: [<matplotlib.lines.Line2D at 0x7f8c6833c130>]
```

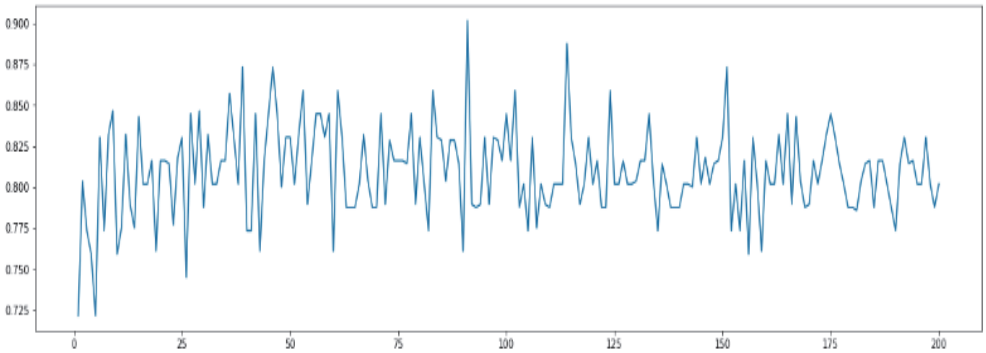

그림 8. Decision Tree 수의 예측 정확도

표 3. 각 모형 훈련세트와 테스트세트에서 예측 정확도

분류 모형	Train Accuracy	Test Accuracy
KNN	0.947	0.733
Decision Tree	0.912	0.800
Random Forest	1.00	0.800
SVM	0.930	0.867
Logistic Regression	0.912	0.867

SVM, Random Forest 분류 모형을 사용하였으며 2022년 제30회 남자 농구 아시안컵에 대한 경기 결과를 예측한 정확도는 〈표 3〉과 같다.

KNN 모형은 훈련 세트에서 예측 정확도는 94.7%, 테스트 세트에서 예측 정확도는 73.3%로 나타났다. Decision Tree 모형은 훈련 세트에서 예측 정확도는 91.2%, 테스트 세트에서 예측 정확도는 80%로 나타났다. Random Forest 모형은 훈련 세트에서 예측 정확도는 100%, 테스트 세트에서 예측 정확도는 80%로 나타났다. SVM 모형은 훈련 세트에서 예측 정확도는 93.0%, 테스트 세트에서 예측 정확도는 86.7%로 나타났다. Logistic Regression 모형은 훈련 세트에서 예측 정확도는 91.2%, 테스트 세트에서 예측 정확도는 86.7%로 나타났다.

Random Forest 분류 모델을 이용하여 예측했을 때, 데이터 세트의 샘플 개수가 충분하지 않기 때문에 과적화(Overfitting) 현상이 발생했다.

3. 모형 예측의 정밀도, 재현율, F1 점수

2022년 제30회 남자 농구 아시안컵에 대한 경기 결과를 예측한 정밀도와 재현율 및 F1 점수는 〈표 4〉와 같다.

KNN 모형은 정밀도는 0.733, 재현율은 0.733, F1 점수는 0.733으로 나타났다. Decision Tree 모형은 정밀도는 0.810, 재현율은 0.800, F1 점수는 0.796으로 나타났다. Random Forest 모형은 정밀도는 0.810, 재현율은 0.800, F1 점수는 0.796으로 나타났다. SVM 모형은 정밀도는 0.900, 재현율은 0.867, F1 점수는 0.868로 나타났다. Logistic Regression 모형은 정밀도는 0.867, 재현율은 0.867, F1 점수는 0.867로 나타났다.

따라서 모형의 평가지표를 종합적으로 분석한 결과, SVM 모형은 예측 성능이 다른 모형보다 우수하게 나타났다.

표 4. 모형 예측의 정밀도, 재현율, F1 점수

분류 모형	정밀도	재현율	F1 점수
KNN	0.733	0.733	0.733
Decision Tree	0.810	0.800	0.796
Random Forest	0.810	0.800	0.796
SVM	0.900	0.867	0.868
Logistic Regression	0.867	0.867	0.867

Ⅳ. 논의

이 연구에서는 2022년 제30회 남자 농구 아시안컵 홈페이지에서 제공된 box score 데이터 중에서 20개의 변수를 선정하여 KNN, Decision Tree, SVM, Logistic Regression, Random Forest 5가지 머신러닝 분류 방법을 사용하여 남자 농구 아시안컵의 승패 결과를 예측하였다. 각각 모형 평가지표를 분석한 결과, KNN 모형은 훈련 세트에서 예측 정확도는 94.7%, 테스트 세트에서 예측 정확도는 73.3%, 정밀도는 0.733, 재현율은 0.733, F1 점수는 0.733으로 나타났다. Decision Tree 모형은 훈련 세트에서 예측 정확도는 91.2%, 테스트 세트에서 예측 정확도는 80%, 정밀도는 0.810, 재현율은 0.800, F1 점수는 0.796으로 나타났다. Random Forest 모형은 훈련 세트에서 예측 정확도는 100%, 테스트 세트에서 예측 정확도는 80%, 정밀도는 0.810, 재현율은 0.800, F1 점수는 0.796으로 나타났다. SVM 모형은 훈련 세트에서 예측 정확도는 93.0%, 테스트 세트에서 예측 정확도는 86.7%, 정밀도는 0.900, 재현율은 0.867, F1 점수는 0.868로 나타났다. Logistic Regression 모형은 훈련 세트에서 예측 정확도는 91.2%, 테스트 세트에서 예측

정확도는 86.7%, 정밀도는 0.867, 재현율은 0.867, F1 점수는 0.867로 나타났다. 각 모형의 예측 정확도에 따라 SVM 모형은 KNN, Decision Tree, Random Forest, Logistic Regression 모형보다 가장 정확한 예측 성능을 보였으며, SVM 모형의 예측 정확도는 86.7%로 나타났다.

NBA 농구 경기 결과 예측에 대한 선행 연구를 살펴보면, Torres & Hu(2013)는 Multilayer perceptron—back propagation neural network, Linear Regression과 Maximum Likelihood를 사용하여 NBA의 경기결과를 예측하였고 가장 좋은 예측 정확도는 70%로 나타났다. Horvat et al.(2020)은 다양한 분류와 회귀의 머신러닝 방법을 사용하여 NBA 경기의 결과를 예측하였고 가장 좋은 예측 정확도는 65%를 넘어서였다. Lieder(2018)는 Logistics Regression, Linear Regression, ANN 모형을 사용하여 NBA의 경기 결과를 예측하였고 가장 좋은 예측 정확도는 70%로 나타났다. Cao(2012)는 Logistic Regression, Naïve Bayes, SVM과 Multilayer perceptron neural network 모형을 사용하여 NBA 경기 결과를 예측하였고 가장 좋은 예측 정확도는 69.67%로 나타났다. Kravanja (2013)은 SVM 및 Logistic Regression을 이용하여 각각 70.01%, 69.73%의 예측 정확도를 보였다. Lin et al.(2014)은 Logistic Regression, Adaboost, Random Forest, SVM, GNB 모형을 이용하여 NBA 경기 결과를 예측하였고 가장 좋은 예측 정확도는 65.20%로 나타났다. 따라서 선행연구에서 NBA 경기결과에 대하여 67~70%의 예측 정확도를 보였다.

NCAA 농구 경기 결과 예측에 대한 선행 연구를 살펴보면, Miljković et al.(2010)은 Decision trees, KNN, Naïve Bayes와 SVM을 사용하여 NCAA 경기 결과를 예측하였고 Naïve Bayes가 가장 좋은 결과를 얻었고, 67%의 예측 정확도를 보였다. Zimmermann et al.(2013)은 Decision Trees, Random Forest, Naïve Bayes, Multilayer perceptron neural network를 사용하여 NCAA 농구 경기 결과를 예측하였고 Decision Trees와 Random Forest가 가장 좋은 예측 결과를 얻었고 각각 70.42%, 71.37%의 예측 정확도를 보였다. Levandoski & Lobo(2017)는 Random Forest, KNN, Logistics Regression, Neural Network, Naïve Bayes, SVM과 Adaboost를 사용하여 NCAA 농구 경기 결과를 예측하였고, Neural Network 모형이 가장 좋은 예측 결과를 얻었고 79.4%의 예측 정확도를 보였다. 따라서 선행연구에서 NCAA 경기 결과에 대하여 62~80%의 예측 정확도를 보였다. Li, Wang & Li(2021)는 연구자들이 NBA와 NCAA에서 경기 결과에 대한 예측 방법을 이미 구축해 놓았지만, 일반적인 단점은 예측 정확도가 낮다는 것으로 확인되었다.

이 연구에서는 SVM 모형의 예측 정확도는 KNN, Decision Tree, Random Forest, Logistic Regression 모형보다 정확한 예측 성능을 가지고 있다고 예상할 수 있다. 또한, 이 연구에서 SVM 모형의 예측 정확도(86.7%)는 농구 영역에서 예측모형의 비교를 수행한 다양한 선행연구의 결과(NBA: 67~70%, NCAA: 62~80%)에 비하면 더 우수하다는 것으로 확인되었다. 따라서 SVM 모형이 남자 농구 아시안컵의 승패 결과를 예측하는데 효과적이고 정확한 머신러닝 예측 모형이라고 사료된다.

V. 결론

이 연구는 2022년 제30회 남자 농구 아시안컵 경기대회의 공식기록을 기반으로 전통적 측면에서의 통계적 방법, 데이터마이닝 기법, 머신러닝의 기법을 활용하여 남자 농구 아시안컵의 승패 결과 예측 및 비교의 결론은 다음과 같다.

첫째, 모델별 예측 결과에서는 SVM 모델이 KNN, Decision Tree, Random Forest, Logistic Regression 모델보다 최적의 예측 성능을 나타냈고 86.67%의 예측 정확도 및 0.868의 F1 점수를 보였다.

둘째, Random Forest 분류 모델을 이용하여 2022 남자농구 아시아컵 경기대회의 승패 결과를 예측했을 때, 데이터 세트의 샘플 개수가 충분하지 않기 때문에 과적화(Overfitting) 현상이 발생했다.

이 연구의 연구 결과를 토대로 향후 사례 수를 증가하여 더 많은 데이터가 모델의 정확도를 높이는 동시에 과적합 가능성을 줄일 수 있다고 사료되었다. 그리고 더 정확한 예측 결과를 얻기 위하여서 머신러닝뿐만 아니라 딥러닝과 관련한 연구도 필요할 것으로 사료된다.

참고문헌

구승환, 김현수, 장성용(2009). 국내 남자 프로농구 승패 예측 모형 비교 연구. **체육과학연구, 20**(4), 704-711.

김세형, 강상조, 박재현, 김혜진(2008). 한국프로농구 경기기록 분석에 의한 승패결정요인. **한국체육측정평가학회지, 10**(1). 1-12.

김세형, 박재현, 김혜진, 강상조(2008). 한국프로농구 경기기록 분석에 의한 승패결정요인. **한국체육측정평가학회지, 10**(1), 1-12.

박상찬(2017). 제4차 산업혁명과 데이터 과학. **한국콘텐츠학회지, 제15권 제1호**, 21-28.

장효진, 곽현, 최승회(2015). 회귀모형을 이용한 한국프로농구 승부결과 분석. **한국지능시스템학회 논문지, 25**(5), 489-494.

저우차우, 최형준(2020). 2019 세계 남자 농구 월드컵 경기대회의 공식기록에 기반한 경기내용에 관한 군집 분석. **한국체육학회지, 59**(3), 397-411.

조정환(2012). 스포츠 빅데이터 활용과 전망. **한국체육측정평가학회지, 14**(3), 1-12.

최형준(2020).국내 스포츠 빅데이터 분석 연구의 현황. **한국체육측정평가학회지, 22**(2), 63-69.

최형준, 김주학(2006). 인공신경망(Artificial Neural Network)을 이용한 2005년도 영국 윔블던 테니스 대회의 경기결과 예측에 관한 연구. **한국체육학회지, 45**(3), 459-468.

최형준, 이윤수(2019). 축구 월드컵대회의 경기기록 기반 경기결과 예측. **한국체육과학회지, 28**(1), 1317-1325.

허종관, 김세중, 도재현(2016). 로지스틱/마르코프 체인 모델을 이용한 한국프로농구 순위 예측. **한국체육과학회지, 25**(5), 1269-1276.

Alpydin, E. (2010). *Introduction to machine learning* (2nd ed.). Cambridge, MA, London, England: MIT Press, 201-205.

Ballı, S., &zdemir, E. (2021). A novel method for prediction of EuroLeague game results using hybrid feature extraction and machine learning techniques. *Chaos, Solitons & Fractals, 150*, 111119.

Bishop, C. M., & Nasrabadi, N. M. (2006). *Pattern recognition and machine learning*. New York: springer.

Boser, B. E., Guyon, I. M., & Vapnik, V. N. (1992). A training algorithm for optimal margin classifiers. *In Proceedings of the fifth annual workshop on Computational learning theory*, 144-152.

Bromley, J., Bentz, J. W., Bottou, L., Guyon, I., Lecun, Y., Moore, C., Säckinger, E., & Shah, R. (1993). Signature verification using a "siamese" time delay neural network. *International Journal of Pattern Recognition and Artificial Intelligence, 7*(04), 669–688.

Bunker, R. P., Thabtah, F. (2019). A machine learning framework for sport result prediction. *Applied Computing and Informatics, 15*(1), 27–33.

Bunker, R., Susnjak, T. (2022). The Application of Machine Learning Techniques for Predicting Match Results in Team Sport: A Review. *Journal of Artificial Intelligence Research, 73*, 1285-1322.

Cai, W., Yu, D., Wu, Z., Du, X., & Zhou, T. (2019). A hybrid ensemble learning framework for basketball outcomes prediction. *Physica A: Statistical Mechanics and its Applications, 528*, 121461.

Cao, C.(2012). *Sports data mining technology used in basketball outcome prediction*. Master's Thesis, Dublin Institute of Technology, Ireland.

Chen, W. J., Jhou, M. J., Lee, T. S., & Lu, C. J. (2021). Hybrid basketball game outcome prediction model by integrating data mining methods for the national basketball association. *Entropy, 23*(4), 477.

Chovanec, P. (2021). *Predicting winners in the NCAA Basketball Tournament*. Kansas State Universiteit; Manhattan, Kansas.

Cortes, C., Vapnik, V. (1995). Support-vector networks. *Machine learning, 20*(3), 273-297.

Drucker, H., Burges, C. J., Kaufman, L., Smola, A., & Vapnik, V. (1997). Support vector regression machines. *Advances in Neural Information Processing Systems, 9*(1), 155-161.

Dubbs, A. (2018). Statistics-free sports prediction. *Model Assisted Statistics and Applications, 13*(2), 173-181.

Elish, M. O. (2014). A comparative study of fault density prediction in aspect-oriented systems using MLP, RBF, KNN, RT, DENFIS and SVR models. *Artificial Intelligence Review, 42*(4), 695-703.

Groll, A., Ley, C., Schauberger, G., & Van Eetvelde, H. (2019). A hybrid random forest to predict soccer matches in international tournaments. *Journal of Quantitative Analysis in Sports, 15*(4), 271-287.

Gu, W., Foster, K., Shang, J., & Wei, L. (2019). A game-predicting expert system using big data and machine learning. *Expert Systems with Applications, 130*, 293-305.

Haghighat, M., Rastegari, H., Nourafza, N., Branch, N., & Esfahan, I. (2013). A review of data mining techniques for result prediction in sports. *Advances in Computer Science: an International Journal, 2*(5), 7-12.

Han, J., Kamber, M., & Pei, J. (2012). *Data mining: concepts and techniques*, Waltham, MA. Morgan Kaufman Publishers.

Harrington, P. (2012). *Machine Learning in Action*; Manning Publications Co.: Shelter Island, New York, USA.

Hassonah, M. A., Rodan, A., Al-Tamimi, A. K., & Alsakran, J. (2019). *Churn Prediction: A Comparative Study Using KNN and Decision Trees*. In 2019 Sixth HCT Information Technology Trends (ITT), 182-186. IEEE: the Institute of Electrical and Electronics.

Horvat, T., & Job, J. (2020). The use of machine learning in sport outcome prediction: A review. *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery, 10*(5), 138-151.

Horvat, T., Havaš, L., & Srpak, D. (2020). The impact of selecting a validation method in machine learning on predicting basketball game outcomes. *Symmetry, 12*(3), 431.

Hosmer, D. W., & Lemeshow, S. (2000). *Applied Logistic Regression*, John Wiley&Sons Inc. New York, NY.

Jadhav, S. D., & Channe, H. P. (2016). Comparative study of K-NN, naive Bayes and decision tree classification techniques. *International Journal of Science and Research (IJSR), 5*(1), 1842-1845.

Jain, S., & Kaur, H. (2017). Machine learning approaches to predict basketball game outcome. *In 2017 3rd International Conference on Advances in Computing, Communication & Automation (ICACCA)(Fall)*, 1-7. IEEE: The Institute of Electrical and Electronics.

Jin, M., Deng, W. (2018). Predication of different stages of Alzheimer's disease using neighborhood component analysis and ensemble decision tree. *Journal of neuroscience methods, 302*(1), 35-41.

Johnson, K. C., Whelton, P. K., Cushman, W. C., Cutler, J. A., Evans, G. W., Snyder, J. K., ... & Wright Jr, J. T. (2018). Blood pressure measurement in SPRINT (systolic blood pressure intervention trial). *Hypertension, 71*(5), 848-857

Kannan, A., Kolovich, B., Lawrence, B., & Rafiqi, S. (2018). Predicting National Basketball Association success: A machine learning approach. *SMU Data Science Review, 1*(3), 7.

Korhonen, K. T., Kangas, A. (1997). Application of nearest-neighbour regression for generalizing sample tree information. *Scandinavian Journal of Forest Research, 12*(1), 97-101.

Kravanja, A. (2013). *Napovedanje zmagovalcev košarkaških tekem*. Undergraduate Thesis, Faculty of Computer and Information Science, University of Ljubljana, Ljubljana, Slovenia.

Lam, M. W. (2018). One-match-ahead forecasting in two-team sports with stacked Bayesian regressions. *Journal of Artificial Intelligence and Soft Computing Research, 8*(1), 87-99.

Lessmann, S., Sung, M. C., & Johnson, J. E. (2010). Alternative methods of predicting competitive events: An application in horserace betting markets. *International Journal of Forecasting, 26*(3), 518-536.

Levandoski, A., Lobo, J. (2017). *Predicting the NCAA Men's*

Basketball Tournament with Machine Learning, 1-15.

Li, B., Xu, X. (2021). Application of artificial intelligence in basketball sport. *Journal of Education, Health and Sport, 11*(7), 54-67.

Li, Y., Wang, L., & Li, F. (2021). A data-driven prediction approach for sports team performance and its application to National Basketball Association. *Omega, 98*, 102123.

Lieder, N. (2018). *Can machine-learning methods predict the outcome of an NBA game?*. Available at SSRN. doi: 10.2139/ssrn.320810

Lin, J., Short, L., & Sundaresan, V. (2014). Predicting National Basketball Association winners. *CS 229 FINAL PROJECT*, 1-5.

Liu, Y. (2021). Prediction for NCAA championship. *In 2021 the 5th International Conference on Big Data Research (ICBDR)*, 8-14.

Lopez, M. J., Matthews, G. J. (2015). Building an NCAA men's basketball predictive model and quantifying its success. *Journal of Quantitative Analysis in Sports, 11*(1), 5-12.

Luu, B. C., Wright, A. L., Haeberle, H. S., Karnuta, J. M., Schickendantz, M. S., Makhni, E. C., & Ramkumar, P. N. (2020). Machine learning outperforms logistic regression analysis to predict next-season NHL player injury: an analysis of 2322 players from 2007 to 2017. *Orthopaedic journal of sports medicine, 8*(9), 2325967120953404.

Madarame, H. (2018). Defensive rebounds discriminate winners from losers in European but not in Asian women's basketball championships. *Asian Journal of Sports Medicine, 9*(1), 13-18.

Maier, T., Meister, D., Trösch, S., & Wehrlin, J. P. (2018). Predicting biathlon shooting performance using machine learning. *Journal of sports sciences, 36*(20), 2333-2339.

Mendes-Moreira, J., Soares, C., Jorge, A. M., & Sousa, J. F. D. (2012). Ensemble approaches for regression: A survey. *Acm computing surveys (csur), 45*(1), 1-40.

Mendonça, L. F., Vieira, S. M., & Sousa, J. M. C. (2007). Decision tree search methods in fuzzy modeling and classification. *International Journal of Approximate Reasoning, 44*(2), 106-123.

Miljković, D., Gajić, L., Kovačević, A., & Konjović, Z. (2010). The use of data mining for basketball matches outcomes prediction. *In IEEE 8th International Symposium on Intelligent Systems and Informatics*, Subotica.

Mueller, S. Q. (2020). Pre-and within-season attendance forecasting in Major League Baseball: a random forest approach. *Applied Economics, 52*(41), 4512-4528.

Nedellec, R., Cugliari, J., & Goude, Y. (2014). GEFCom2012: Electric load forecasting and backcasting with semi-parametric models. *International Journal of forecasting, 30*(2), 375-381.

Nguyen, N. H., Nguyen, D. T. A., Ma, B., & Hu, J. (2021). The application of machine learning and deep learning in sport: predicting NBA players' performance and popularity. *Journal of Information and Telecommunication*, 1-19.

Niblett, T., Bratko, I. (1987). Learning decision rules in noisy domains. In Proceedings of Expert Systems' 86, *the 6th Annual Technical Conference on Research and development in expert systems III*, 25-34.

Noor, N. B., Anwar, M. S., & Dey, M. (2019). Comparative Study Between Decision Tree, SVM and KNN to Predict Anaemic Condition. *In 2019 IEEE International Conference on Biomedical Engineering, Computer and Information Technology for Health (BECITHCON)*, 24-28.

Osei-Bryson, K. M. (2007). Post-pruning in decision tree induction using multiple performance measures. *Computers & operations research, 34*(11), 3331-3345.

Pai, P. F., ChangLiao, L. H., & Lin, K. P. (2017). Analyzing basketball games by a support vector machines with decision tree model. *Neural Computing and Applications, 28*(12), 4159-4167.

Passi, K., Pandey, N. (2018). *Increased Prediction Accuracy in the Game of Cricket using Machine Learning*. ArXiv Preprint ArXiv:1804.04226.

Pathak, N., Wadhwa, H. (2016). Applications of modern classification techniques to predict the outcome of ODI cricket. *Procedia Computer Science, 87*(2), 55-60

Prasetio, D.; Harlili, D. (2016). Predicting football match results with logistic regression. *In Proceedings of the International Conference on Advanced Informatics: Concepts, Theory and Application (ICAICTA)*, Penang, Malaysia, 16–19.

Purucker, M. C. (1996). Neural network quarterbacking. *IEEE Potentials, 15*, 9–15.

Russell, S., & Norvig, P. (2002). *Artificial Intelligence: A Modern Approach. 3rd ed*. Upper Saddle River, NJ :Prentice Hall.

Sagi, O., & Rokach, L. (2018). Ensemble learning: A survey. *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery, 8*(4), e1249.

Schölkopf, B., Smola, A. J., & Bach, F. (2002). *Learning with kernels: support vector machines, regularization, optimization, and beyond*. MIT Press, 96-97.

Soliman, G., Misbah, A., & Eldawlatly, S. (2017). Predicting all star player in the national basketball association using random forest. *In 2017 Intelligent Systems Conference (IntelliSys)* . IEEE. 706-713

Song, K., Zou, Q., & Shi, J. (2020). Modelling the scores and performance statistics of NBA basketball games. *Communications in Statistics-Simulation and Computation, 49*(10), 2604-2616

Suthaharan, S. (2016). Machine learning models and algorithms for big data classification. *Integr. Ser. Inf. Syst, 36*(3), 1-12.

Tanha, J., van Someren, M., & Afsarmanesh, H. (2017). Semi-supervised self-training for decision tree classifiers. *International Journal of Machine Learning and Cybernetics, 8*(1), 355-370.

Thabtah, F., Zhang, L., & Abdelhamid, N. (2019). NBA game result prediction using feature analysis and machine learning. *Annals of Data Science, 6*(1), 103.

Tichy, W. (2016). Changing the Game: Dr. Dave Schrader on sports analytics. *Ubiquity*, 1-10.

Torres, R. A., & Hu, Y. H. (2013). *Prediction of NBA games based on Machine Learning Methods*. University of Wisconsin, Madison.

Vapnik, V. N. (1995). *The nature of statistical learning theory.Springer*, New York, 232-233.

WentingWang, R. M. (2014). Predicting winners of NCAA women's basketball tournament games. *International Journal of Sports Science, 4*(5), 173-180.

Ye, N. (2003). *The handbook of Data Mining. New Jersey: Lawrence Erlbaum Associates*, 125-126.

Yezus, A. (2014). *Predicting outcome of soccer matches using machine learning*. Saint Petersburg University; Saint Petersburg, Russia.

Zhao, Y. (2021). Sports enterprise marketing and financial risk management based on decision tree and data mining. *Journal of Healthcare Engineering, 8*(1), 2-7

Zimmermann, A., Moorthy, S., & Shi, Z. (2013). *Predicting college basketball match outcomes using machine learning techniques: some results and lessons learned*. arXiv preprint arXiv:1310.3607.

Zuccolotto, P., Manisera, M., & Sandri, M. (2018). Big data analytics for modelling scoring probability in basketball: The effect of shooting under high-pressure conditions. *International Journal of Sports Science & Coaching, 13*(4), 569–589.

저자정보

예원진(Yuan-Zhen NI)
한양대학교 박사과정
yuanzhen_ni@naver.com

이성노(Seong-No LEE)
한양대학교 예술체육학과 교수
snl743@hanyang.ac.kr

논문투고일	2022년 09월 07일
심사완료일	2022년 09월 22일
게재확정일	2022년 09월 26일

Abstract

The Korean Journal of Measurement and Evaluation in Physical Education and Sport Science. 2022, 24(3), 53-69

Comparison of Prediction Performance of Machine Learning Classification Model Using 2022 FIBA Men's Basketball Asian Cup Match Results

Yuan-Zhen NI · Seong-No LEE *Hanyang Univ.*

The purpose of this study is to compare predictive performance using traditional statistical methods, data mining techniques, and machine learning techniques using the box scores of the 2022 FIBA Men's Basketball Asian Cup Games. The subject of this study was a total of 72 match records among the records obtained through the official records of the 2022 FIBA Men's Basketball Asian Cup, and the outcome of the match was predicted through a total of 20 variables. Five classification models were used to predict the outcome of the men's basketball Asian Cup competition: K Nearest Neighbor (KNN), Decision Tree, Support Vector Machine (SVM), Logistic Regression, and Random Forest. For the data collection and processing of this study, the statistical program Python 3.10.1 version was used together with the library, and the results obtained are as follows. First, in the prediction results for each model, the SVM model showed the optimal prediction performance than the KNN, Decision Tree, Random Forest, and Logistic Regression models, and showed 86.67% prediction accuracy and 0.868 F1 score. Second, when the Random Forest classification model was used to predict the win/loss result of the 2022 Men's Basketball Asian Cup, overfitting occurred because the number of samples in the data set was not sufficient. Based on the results of this study, it was considered that more data by increasing the number of cases in the future can increase the accuracy of the model and reduce the possibility of overfitting. And in order to obtain more accurate prediction results, it is judged that research related to deep learning as well as machine learning is necessary.

Keywords: Machine Learning, Men's Basketball Asian Cup, Prediction, Classification Model