Korean Journal of Sport Science 2025, 36(4), 592-604
https://doi.org/10.24985/kjss.2025.36.4.592

Original Article

Development of a Machine Learning-Based System for Classifying and Predicting Golf Players' Playing Styles

Hong-suk Kim[1], Hye-soo Cho[2], Ji-Yong Park[1], and Hyeon-su Park[1*]

[1]*Department of Sports Science, Hanyang University*
[2]*Department of Sports Science, Hanyang University ERICA*

Article Info
Received 2025. 07. 21.
Revised 2025. 11. 08.
Accepted 2025. 12. 01.

Correspondence*
Hyeon-su Park
gustn1473@naver.com

Key Words
KPGA, Clustering,
Playing style,
Performance skill factors,
Machine leaning

PURPOSE This study sought to classify the playing styles of KPGA players based on performance-related technical factors and develop a supervised learning model that automatically predicts and classifies these styles. **METHODS** Performance data were gathered from KPGA Korean Tour players between 2015 and 2024, focusing on six key technical indicators. Distinct playing styles were identified by standardizing the variables using z-scores and then clustering them using the K-means algorithm. Based on the clustering results, predictive classification models were built by applying five supervised learning algorithms—decision tree, random forest, K-nearest neighbors (KNN), support vector machine (SVM), and multinomial logistic regression. Model performance was then evaluated using accuracy, precision, recall, and F1-score, with generalizability assessed via five-fold cross-validation. **RESULTS** Four playing style clusters were obtained, each labeled according to players' technical characteristics: "overall weakness type," "distance-deficient but technically proficient type," "accuracy-oriented type," and "power and risk-management type." The multinomial logistic regression model showed the highest predictive performance, followed by SVM, KNN, random forest, and decision tree. **CONCLUSIONS** This study confirmed that KPGA players can be characterized into four distinct playing styles based on their technical performance data and that these styles can be effectively classified and predicted by supervised learning models. These findings highlight the models' practical applicability in personalizing training strategies, developing course-specific game plans, and contributing to the advancement of AI-based sports analytics systems.

서 론

골프는 선수의 신체적·심리적 상태, 경기 기술과 같은 개인적 특성과 경기장의 형태 및 상태 혹은 날씨와 같은 외적인 요인이 복합적으로 작용하는 종목이다(Kim, 2010). 골프 경기력을 구성하는 요인은 현장뿐만 아니라 학계에서도 지속적으로 논의되고 있으며, 일반적으로 기술적 능력(정확성 및 일관성), 창의적 의사결정 능력(전략적 사고와 상황 판단), 외적 능력(환경과 외부 요인), 내적 능력(심리적 안정과 집중력 등), 신체 물리적 능력(체력과 신체조건)이 있다(Smith, 2010). 이 중 경기력에 직접적인 영향을 미치는 요인은 골프 경기를 구성하는 기술 요인이다(Son & Kim, 2008).

골프의 기술 요인에는 드라이브 샷, 아이언 샷, 어프로치 샷, 벙커 샷, 퍼팅 등이 있으며, 이 외에도 대회 중 주어지는 다양한 상황에 맞춰 여러 가지 기술을 구사하여 목표 지점으로 볼을 정확히 보내기 위해 플레이한다. 이러한 선수들이 대회에 참가하여 발생하는 기술 요인의 결과를 각종 프로골프협회에서 기록하여 제공하고 있으며, 이는 선수의 내적·외적 변인을 종합적으로 포함하고 있어 그 영향력이 그대로 반영된 지표라고 할 수 있다(Min, 2011; Quinn, 2006). 기술 요인에 대한 통계적 자료는 선수들의 경기력 수준을 객관적으로 파악하고 평가할 수 있기 때문에 경기력 평가 및 향상을 위한 지표로서 사

용되고 있으며, 이를 활용한 연구가 활발히 이루어지고 있다.

기존 연구들은 기술 요인이 경기력에 미치는 영향을 분석하거나 성적 및 상금과의 관계를 규명하는 데 집중되어 왔다(Kim & Kim, 2010; Son & Kim, 2010; Min, 2011; Kim et al., 2012; Kim & Cho, 2013; Kim & Min, 2014; Pyung et al., 2015; Kwon et al., 2024). 그러나 대부분의 연구는 선수 개개인의 경기 스타일을 고려하지 않고 전체 집단을 일괄적으로 분석하였다는 한계가 있다. 평균 타수는 다양한 기술 요인의 영향을 받는 종합적인 지표이기 때문에 같은 타수를 기록한 선수라도 경기 스타일과 특성에는 차이가 존재할 수 있다(Kim & Park, 2021). 이에 본 연구자는 기존 선행 연구의 결과를 모든 선수에게 일괄적으로 적용하기에는 한계가 있을 것이라는 점에 주목하였다.

즉, 선수들의 경기 스타일과 특성을 고려하지 않고 분석할 경우 선수 개개인의 강점과 약점을 반영한 최적의 경기 전략을 도출하는 데 어려움이 있을 것이다. 반면, 선수들의 경기 스타일과 기술적 특성을 기준으로 구분하여 분석할 경우 보다 정밀하고 객관적인 해석이 가능하며, 경기력 향상을 위한 맞춤형 전략 수립에도 기여할 수 있다(Databuckets, 2015). 나아가 이러한 방식으로 도출된 분석 결과를 통해 다양한 코스 환경에서 선수의 스타일에 따른 강점과 약점을 체계적으로 파악하고, 이를 바탕으로 전략을 최적화하는 데 활용될 수 있을 것이다. 이에 본 연구는 골프 선수들의 기술 요인 데이터를 바탕으로 경기 스타일을 기준으로 군집화한 후, 머신러닝 기법을 적용하여 각 군집에 대한 자동 분류가 가능한 지도학습 모델을 구축하고자 한다.

최근 스포츠 분야에서는 머신러닝 기법을 활용한 지도학습 기반의 예측 모델을 개발하는 연구가 활발히 이루어지고 있으며(Choi, 2022; Kim & Lee, 2023; Kang, 2023; Jo et al., 2023; Kim et al., 2024a; Kim et al., 2024b), 이는 경기력 분석의 효율성과 정밀도를 높이는 데 기여하고 있다. 그러나 대부분 선행 연구는 승패 여부, 최종 순위와 같은 결과 중심의 변수 예측에 초점을 맞추고 있으며, 경기력 발현 과정이나 선수의 고유한 경기 스타일에 대한 분석은 상대적으로 미흡한 실정이다.

특히 골프와 같이 선수 개개인의 기술적 선택과 전략적 의사결정이 경기 운영 방식에 직접적인 영향을 미치는 종목에서는 결과 자체보다 어떻게 경기를 했는지에 대한 정량적·정성적 해석이 중요하다(Smith, 2010). 이에 본 연구는 2015년부터 2024년까지 한국프로골프협회(KPGA) 코리안투어에 참가한 선수들의 경기력 기술 요인 데이터를 기반으로 선수들의 경기 스타일을 군집화하고, 이를 토대로 예측·분류 지도학습 모델을 구축하여 향후 선수의 경기 스타일을 자동 예측 및 분류할 수 있는 시스템을 개발하는 것에 그 목적이 있다. 이는 기존의 결과 중심 분석의 한계를 보완하고, 정량적 경기력 분석의 새로운 방향을 제시한다는 점에서 학술적·실무적 의의가 있다.

연구방법

자료수집

본 연구에서는 KPGA 홈페이지(https://www.kpga.co.kr/tours/record/?tourId=11)에서 공개하고 있는 2015년부터 2024년까지 코리안투어 대회에 참가한 선수들의 경기력 기술 요인 자료를 수집하여 사용하였다(KPGA, 2024, Table 1). 이때 동일한 선수가 여러 연도에 중복 포함될 가능성이 있으나, 각 연도별 경기력은 해당 시즌 선수의 기술적 수준, 코스 환경, 경기 조건 등 다양한 요인에 의해 달라질 수 있으므로, 동일 선수라 하더라도 각 연도는 독립된 경기력 데이터(연도 단위의 독립 관측치)로 간주하였다. 또한, American Psychological Association(APA, 2017)의 윤리 가이드라인에 따르면 공개된 데이터를 활용한 2차 분석 연구는 별도의 기관생명윤리위원회(IRB) 승인 없이 수행할 수 있는 경우에 해당한다. 이에 따라 IRB 심의 절차는 진행하지 않았으며, 선수 개인을 식별할 수 있는 정보는 포함하지 않았다.

측정 변인

본 연구에서 사용한 변인은 KPGA(한국프로골프협회) 홈페이지에서 공개하고 있는 자료 중 본 연구의 목적에 부합한 6개 변인을 선정하였다(Table 2). 또한, 본 연구에서는 기술-전략 상호작용 관점(technical-strategic interaction perspective)과 선수의 기술 수행 패턴을 통한 경기 스타일 유형화에 관한 선행연구(Ball & Best, 2007; Broadie, 2014)를 이론적 근거로 하여, 6개 기술 변인을 기반으로 한 기술적 수행 특성에 근거한 '경기 스타일'을 조작적으로 정의하였다.

통계 처리

본 연구에서는 얻어진 데이터를 통계 소프트웨어인 Python(3.13) jupyter notebook(7.3.2)의 pandas, scikit-learn 라이브러리를 활용하여 분석하였다. 먼저 KPGA 코리안투어 선수들의 기술 요인을 알아보기 위해 기술통계 분석(Descriptive Statistics Analysis)을 실시하였고, 기술 요인 간 연관성 및 다중공선성(Multicollinearity)을 검증하기 위해 피어슨의 상관분석(Pearson's Correlation Analysis)을 실시하였다. 다음으로 수집된 자료를 분석 가능한 형태로 표준화(z-점수)하였다. 이후 Kim & Park(2021)의 연구에서 군집의 수를 통계적 근거에 기반하여 설정해야 한다는 제언에 따라 선행 연구를 참고하여(Go, 2003; Kim & Choi, 2019; Choi et al., 2007) 엘보우 기법(Elbow Method)을 통해 최적의 초

Table 1. Collected data

	Year	Number of player
1	2015	105
2	2016	98
3	2017	110
4	2018	106
5	2019	104
6	2020	123
7	2021	118
8	2022	116
9	2023	104
10	2024	112
Total		1,096

Table 2. Variables

	Variables	Label	Contents
1	Driving Distance(Yds)	DD	The mean driving distan ce (yd) meas ured o n two comm ittee-designated holes among the 18 holes
2	Fairway Accuracy(%)	FA	Fairway accuracy was defined as the percentage of tee shots that landed in the fairway, excluding par-3 holes
3	Green in Regulation(%)	GIR	The percentage of holes in which the ball was on the green in regulation
4	Recovery Percentage(%)	RP	The percentage of holes in which a par or better score was recorded when the ball was not on the green in regulation
5	Sand Saves Percentage(%)	SSP	The percentage of attempts that resulted in a score in regulation when the ball was in a greenside bunker
6	Putting Average(stroke)	PA	The average number of putts on holes where the ball was on the green in regulation

기 그룹(k) 수를 결정하였다. 이때 군집 수에 따른 WCSS(Within-Cluster Sum of Squares) 값의 감소 패턴을 분석하여 WCSS의 감소 폭이 급격히 변화하는 지점인 엘보우 포인트(Elbow Point)를 찾아 최적의 그룹(k) 수를 결정하였다. 추가적으로 군집 수 결정의 객관성을 확보하기 위해 실루엣 분석(Silhouette Analysis)과 갭 통계(Gap Statistic)를 활용하였다. 이후 결정된 K값을 바탕으로 K-평균 군집분석(K-Means clustering analysis)을 실시하여 선수들의 기술 요인 특성을 반영한 경기 스타일 유형으로 분류하였다. 이때, 유사도 기반의 기술 수행 패턴으로 분류된 군집의 특성을 해석하기 위해 각 군집에 속한 선수들의 기술 요인 특성을 바탕으로 명칭을 부여하였으며, 이 과정은 골프 경력 10년 이상이면서 KPGA 프로 라이선스를 보유하고 있는 박사 3인의 자문을 통해 이루어졌다. 또한, 각 군집에 속한 선수들의 기술 요인의 차이를 알아보기 위해 일원배치분산분석(one-way ANOVA)을 실시하였고, 각 군집 별 기술 요인의 통계적으로 유의한 차이를 나타낸 경우 Scheffe를 사용하여 사후분석을 실시하였다. 마지막으로, 자동 분류 모델을 구축하기 위해 여섯 가지 기술 요인을 독립변수, 분류된 네 가지 군집을 종속변수로 설정하였다. 이때 적용된 머신러닝 기법은 총 다섯 가지로, 의사결정나무(Decision Tree), 랜덤 포레스트(Random Forest), K-최근접 이웃(K-Nearest Neighbors, KNN), 서포트 벡터 머신(Support Vector Machine, SVM), 다항 로지스틱 회귀분석(Multinomial Logistic Regression)을 활용하였다.

구축된 각 예측 모델의 성능 평가는 전체 데이터를 무작위로 배열한 후, 80%를 학습 데이터, 20%를 테스트 데이터로 설정한 뒤, 정확도(Accuracy), 정밀도(Precision), 재현율(Recall), F1-점수(F1-Score)를 산출하여 평가하였다. 또한, 모델의 일반화 성능을 확보하기 위해 k-fold 교차검증(k-fold cross validation)을 10회(5-fold) 반복 수행하여 평균 및 표준편차 값을 산출하였다. 마지막으로 다섯 가지 예측 모델 간의 성능을 비교하기 위해 골모고로프-스미르노프(Kolmogorov-Smirnov test)을 이용하여 정규성 검정을 실시한 결과, 정규성을 만족하는 것으로 확인하여(p>.05), 모수 통계 검정 방법인 일원배치분산분석(one-way ANOVA)을 실시하였다. 이때 모델 간 예측 성능이 통계적으로 유의한 차이를 나타냈을 경우 Scheffe를 사용하여 사후분석을 실시하였다. 모든 통계적 유의성 검증은 유의수준 α= .05로 설정하였다.

머신러닝 기법

본 연구에서 다루는 문제 유형은 다차원적 기술 변인을 기반으로 한 분류(classification) 문제로, 전통적인 선형 통계 기법만으로는 변수 간의 비선형적 상호작용을 충분히 반영하기 어렵다(Jeong & Choi, 2022; Choi et al., 2024). 이러한 관점에서 머신러닝의 지도학습(supervised learning) 기법은 이러한 한계를 보완하여, 고차원 데이터 내의 숨은 패턴을 인식하고 다변량 관계를 동시에 탐지할 수 있는 장점을 지닌다(Tabassum et al., 2022). 특히, 본 연구의 목적은 선수들의 경기 스타일을 예측·분류하는 자동화 모델 구축에 있으므로, 훈련 데이터로부터 목표 범주(label)를 학습하는 지도학습 접근이 적합하다(Davis et al., 2024).

따라서 본 연구에서는 이러한 이론적 근거에 따라 의사결정나무, 랜덤 포레스트, KNN, SVM, 다항 로지스틱 회귀 등 다섯 가지 지도학습 기법을 적용하였으며, 기법에 대한 설명은 다음과 같다.

1) 의사결정나무(Decision Tree)

의사결정나무는 기계학습(Machine Learning)에서 지도학습(Supervised Learning)의 알고리즘으로 데이터를 탐색하여 분류 및 예측하기 위한 목적으로 사용된다. 나무의 최상단에 위치하는 루트노드(root node)는 전체 데이터를 최초로 분할하는 기준이 되며, 그 아래 내부 노드(internal node)로 분기된다. 각 내부 노드는 특정 임계값에 따라 데이터를 둘 이상의 가지(branch)로 분할하고, 최종적으로 더 이상 분할이 필요 없거나 설정된 조건에 도달하면 리프노드(leaf node)가 되어 최종 예측 결과로 나타난다. 이를 바탕으로 의사결정나무가 내린 결론을 해석하는 단계를 거치게 된다(Berry & Linoff, 1997). 의사결정나무는 직관적인 해석이 가능하고 비교적 해석이 쉽다는 장점을 지니지만, 과적합이 발생하기 쉬우며 전체적인 선형관계 파악에 미흡하다는 한계를 지닌다(Bishop & Nasrabadi, 2006). 본 연구에서는 Scikit-Learn 라이브러리의 DecisionTreeClassifier 함수를 활용하였다.

2) 랜덤 포레스트(Random Forest)

랜덤 포레스트란, 랜덤으로 추출된 여러 개의 결정 나무가 모여 숲을 이룬 방식으로, 예측 성능을 높이는 앙상블 기법을 말한다. 랜덤 포레스트는 데이터 셋(Data set)으로부터 무작위로 데이터를 추출하여 다수의 결정 나무(Decision tree)를 구축하는데, 큰 수의 법

칙에 따라 의사결정나무의 수가 많아질수록 과적합을 방지할 수 있으며, 개별 의사결정나무를 학습시킬 때 전체 데이터 셋에서 무작위로 복원 추출된 데이터를 사용하기 때문에 잡음(noise)이나 이상치(outlier)로부터 비교적 자유롭다(Kang, 2023). 따라서 유용한 데이터 분류 및 회귀 알고리즘으로 널리 사용되고 있다(Pathak & Wadhwa, 2016). 본 연구에서는 Scikit-Learn 라이브러리의 RandomForestClassifier 함수를 활용하였으며, 결정 나무의 수를 100개로 설정하여 학습하였다.

3) K-최근접 이웃(K-Nearest Neighbors, KNN)

K-최근접 이웃은 분류 및 회귀 알고리즘의 한 종류로 N차원의 공간에서 가까운 거리에 위치한 데이터를 동일한 범주로 분류하는 지도학습 모델이다(Alonso & Babac, 2022). 다른 지도학습 모델에 비해 원리가 단순하여 구현이 쉽고 훈련 단계에서 빠른 수행이 가능해 분류 및 회귀 모델로 많이 사용되고 있다(Horvat, Havaš et al., 2020). 그러나, 선택하는 k의 수에 따라 분류되는 그룹의 속성 및 특성이 달라지기 때문에 데이터의 특성에 따른 적절한 k의 수를 선택하는 것이 중요하다. 본 연구에서는 Scikit-Learn 라이브러리의 KNeighbors Classifier 함수를 활용하였으며, 최근접 이웃의 수(k)는 선행 연구에서 제시한 과적합(overfitting)과 과소적합(underfitting)의 균형을 고려한 기준값인 5로 설정하였다(Pedregosa et al., 2011; Tan et al., 2019).

4) 서포트 벡터 머신(Support Vector Machine, SVM)

서포트 벡터 머신은 초평면을 이용하여 다차원(N차원)의 공간을 N-1 차원으로 나누는 분류 알고리즘이다(Kim, Lee et al., 2024). 각 변인의 데이터 중 클래스 간 경계에 가까운 데이터 간의 거리가 최대가 되는 구분선(초평면)을 추산하여 새로운 점이 나타날 때 경계의 어느 쪽에 속하는지 분류하는 방법이다(Horvat & Jod, 2020). 서포트 벡터 머신은 선형이나 비선형 문제에 모두 적용 가능하며, 과적합의 위험이 낮아 분류 학습 모델로 효과적인 기법이다(Kim & Lee, 2023). 본 연구에서는 Scikit-Learn 라이브러리의 SVC 함수를 활용하였으며, 최대 반복 횟수는 100으로 설정하였다(Kang, 2023).

5) 다항 로지스틱 회귀분석(Multinomial Logistic Regression)

다항 로지스틱 회귀분석은 세 개 이상의 범주형 종속변수를 다루는 모형으로, 기준 범주(reference group)와 비교하여 각 범주에 속할 로즈 오그(log-odds)를 추정하는 통계적 분석 기법이다(Lee, 2020). 본 연구에서는 엘보우 기법을 통해 최적의 군집 수를 4로 설정하고, 선수들의 기술 요인 수준에 따른 경기 스타일을 분류하기 위해 다항 로지스틱 회귀분석을 실시하였다. 또한, Scikit-Learn 라이브러리의 LogisticRegression 함수를 활용하였으며, multi_class=multinomial과 solver=lbfgs 옵션을 적용하여 다항 로지스틱 회귀 분석을 수행하였다.

연구결과

KPGA 코리안투어 선수들의 기술 요인 기술통계 분석

KPGA 코리안투어 선수들의 기술 요인을 알아보기 위해 기술통계 분석을 실시한 결과는 다음 〈Table 3〉과 같다.

기술 요인 간 상관관계 분석

본 연구에서 선수들의 경기 스타일을 분류하기 위해 사용된 기술 요인 간 연관성 및 다중공선성을 검증하기 위해 피어슨의 상관분석을 실시한 결과, 선행 연구에서 제시한 기준(Dormann et al., 2013)에 따라 다중공선성 문제는 없는 것으로 판단하였다. 이에 따라 각 기술 요인은 독립성을 유지한 상태에서 군집화 분석에 활용하기에 적절한 것으로 판단하였다(Table 4).

군집 수 결정을 위한 통계 기법

최적의 군집 수를 결정하기 위해 엘보우 기법을 실시하여 WCSS 값의 변화량을 분석하였다. 그 결과, 군집의 수가 4개까지는 WCSS 값이 급격히 감소하였으며, 군집의 수가 5개 이상부터는 감소 폭이 완만해지는 경향을 보여 해당 지점을 엘보우 포인트로 설정하였다(Table 5, Fig. 1). 또한, 군집 수 결정의 객관성을 확보하기 위해 실루엣 지수와 갭 통계를 병행 산출하였다. 실루엣 분석 결과, k=4까지 WCSS 값이 급격히 감소하다가 이후부터는 감소 폭이 점차 완만해지는 경향을 나타냈으며(Table 6, Fig. 2), 갭 통계에서는 k=3에서 가장 높은 값을 보였다(Table 7, Fig. 3). 최종적으로 군집 간 분리도 해석 가능성을 종합적으로 고려한 결과 k=4가 가장 안정적이고 해석 가능한 군집 구조로 판단되어, 최종적으로 4의 군집을 최적 군집 수로 설정하였다.

Table 3. Descriptive statistics of skill factors

	Variables	M±SD
1	Driving Distance(Yds)	283.643±9.959
2	Fairway Accuracy(%)	63.712±6.965
3	Green in Regulation(%)	69.282±4.038
4	Recovery Percentage(%)	52.769±5.883
5	Sand Saves Percentage(%)	56.514±11.274
6	Putting Average(stroke)	1.811±.041

Table 4. Correlation analysis among skill factors

	DD	FA	GIR	RP	SP	PA
DD	1	-.600**	.166**	-.020	.284**	-.112**
FA		1	.355**	.117**	-.137**	.024
GIR			1	.154**	.294**	-.142**
RP				1	.365**	-.334**
SSP					1	-.251**
PA						1

**p<.01

DD: Driving Distance, FA: Fairway Accuracy, GIR: Green In Regulation, RP: Recovery Percentage, SSP: Sand Saves Percentage, PA: Putting Average

Table 5. WCSS by number of clusters (Elbow Method)

K(Number of clusters)	WCSS	Reduction
2	5498.604	-
3	4801.022	697.582
4	3337.467	1463.555
5	3032.817	304.65
6	2783.397	249.42
7	2600.733	182.664
8	2482.073	118.66

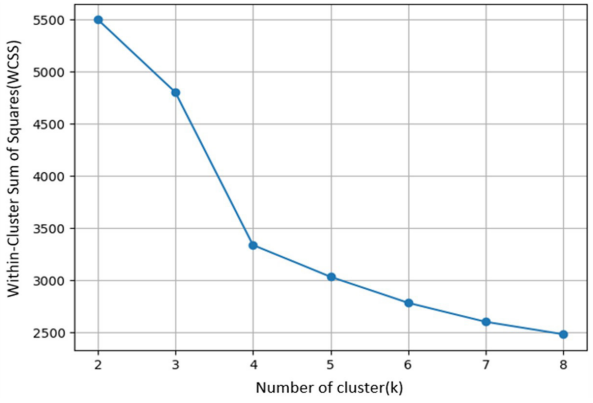

Fig. 1. WCSS by number of clusters (Elbow Method)

Table 6. Silhouette Score by number of clusters (Silhoette Analysis)

K(Number of clusters)	Silhiette Score	Reduction
2	5272.051	-
3	4428.162	843.892
4	4062.601	365.561
5	3741.332	321.275
6	3536.492	204.848
7	3353.481	183.013
8	3180.607	172.889

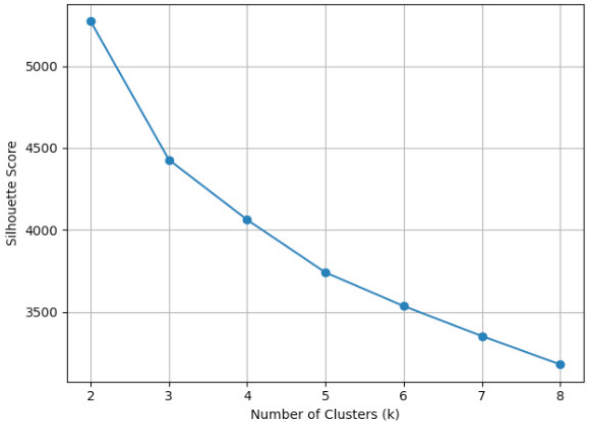

Fig. 2. Silhouette Score by number of clusters (Silhoette Analysis)

Table 7. Gap Statistic by number of clusters (Gap Statistic)

K(Number of clusters)	Gap Statistic	s(k)
2	1.6762	.0038
3	1.7170	.0103
4	1.6955	.0182
5	1.7008	.0090
6	1.6994	.0108
7	1.6870	.0139
8	1.6936	.0139

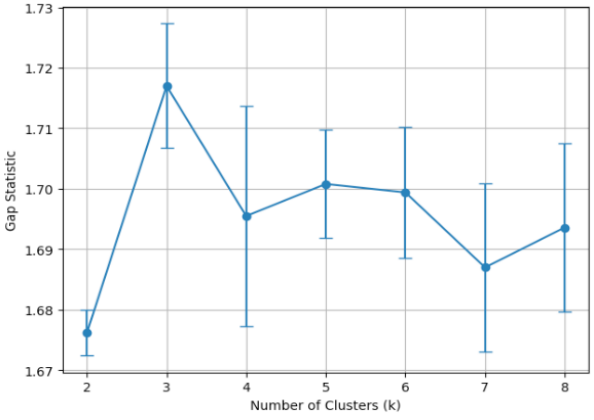

Fig. 3. Gap Statistic by number of clusters (Gap Statistic)

군집 별 특성

세 가지 군집 수 설정 통계기법을 토대로 산출된 4개의 군집에 따라 K-평균 군집분석을 실시하여, 각 군집에 속한 선수 수와 기술 요인을 알아보기 위해 기술통계 분석을 실시한 결과와 군집에 속한 선수들의 기술 요인 특성에 따라 명명된 군집은 다음 〈Table 8〉, 〈Fig. 4〉와 같다.

군집 별 기술 요인 차이 분석

1) 드라이브 비거리

각 군집에 속한 선수들의 드라이브 비거리의 차이를 검증한 결과 군집 별 차이가 있는 것으로 나타났다(F=357.405, p<.001). 결과는 〈Table 9〉와 같다.

2) 페어웨이 안착률

각 군집에 속한 선수들의 페어웨이 안착률의 차이를 검증한 결과 군집 별 차이가 있는 것으로 나타났다(F=403.933, p<.001). 〈Table 10〉과 같다.

3) 그린 적중률

각 군집에 속한 선수들의 그린 적중률의 차이를 검증한 결과 군집 별 차이가 있는 것으로 나타났다(F=216.294, p<.001). 〈Table 11〉과 같다.

4) 리커버리율

각 군집에 속한 선수들의 리커버리율의 차이를 검증한 결과 군집

Table 8. Skill factor characteristics and cluster labeling

Cluster	Label	n	DD(Yard)	FA(%)	GIR(%)	RP(%)	SSP(%)	PA(Stroke)
C1	Overall Low-Performance Type	156	280.890±7.896	59.234±5.320	63.563±3.792	47.583±5.837	47.098±10.548	1.833±.044
C2	Distance-Limited but Skill-Dominant Type	296	281.269±6.751	67.229±4.608	71.474±3.034	57.577±4.849	63.927±8.339	1.787±.033
C3	Accuracy-Specialist Type	319	276.600±7.292	68.940±4.393	69.753±2.971	51.032±4.538	49.341±9.317	1.829±.036
C4	Long-Hitter with Risk-Management Type	325	294.039±6.753	57.528±5.088	69.568±3.236	52.584±4.688	61.321±7.818	1.806±.036

DD: Driving Distance, FA: Fairway Accuracy, GIR: Green In Regulation, RP: Recovery Percentage, SSP: Sand Saves Percentage, PA: Putting Average

별 차이가 있는 것으로 나타났다(F=168.914, p<.001). 〈Table 12〉와 같다.

5) 벙커 세이브율

각 군집에 속한 선수들의 벙커 세이브율의 차이를 검증한 결과 군집 별 차이가 있는 것으로 나타났다(F=230.860, p<.001). 〈Table 13〉과 같다.

6) 퍼트 수

각 군집에 속한 선수들의 퍼트 수의 차이를 검증한 결과 군집 별 차이가 있는 것으로 나타났다(F=87.368, p<.001). 〈Table 14〉와 같다.

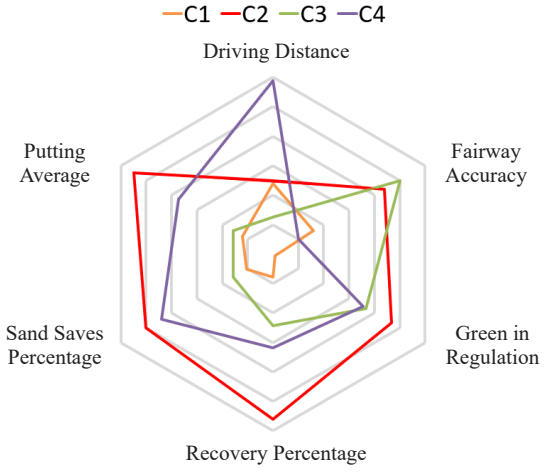

Fig. 4. Characteristics by cluster

Table 9. Driving distance (Yds)

Cluster	M	SE	$F(p)$	post-hoc (Scheffe)
C1	280.890	7.896	357.405***	C4 > C2*** = C1 > C3***
C2	281.269	6.751		
C3	276.600	7.292		
C4	294.039	6.753		

***p<.001

Table 10. Fairway accuracy (%)

Cluster	M	SE	$F(p)$	post-hoc (Scheffe)
C1	59.234	5.320	403.933***	C3 > C2*** > C1** > C4*
C2	67.229	4.608		
C3	68.940	4.393		
C4	57.528	5.088		

*p<.05, **p<.01, ***p<.001

Table 11. Green in regulation (%)

Cluster	M	SE	$F(p)$	post-hoc (Scheffe)
C1	63.563	3.792	216.294***	C2 > C3*** = C4 > C1***
C2	71.474	3.034		
C3	69.753	2.971		
C4	69.568	3.236		

***p<.001

Table 12. Recovery percentage (%)

Cluster	M	SE	$F(p)$	post-hoc (Scheffe)
C1	47.583	5.837	168.914***	C2 > C4*** > C3*** > C1***
C2	57.577	4.849		
C3	51.032	4.538		
C4	52.584	4.688		

***p<.001

Table 13. Sand saves percentage (%)

Cluster	M	SE	$F(p)$	post-hoc (Scheffe)
C1	47.098	10.548	230.860***	C2 > C1*** > C4* > C3***
C2	63.927	8.339		
C3	49.341	9.317		
C4	61.321	7.818		

*p<.05, ***p<.001

Table 14. Putting average (stroke)

Cluster	M	SE	$F(p)$	post-hoc (Scheffe)
C1	1.833	.044	87.368***	C2 < C4*** < C3*** = C1
C2	1.787	.033		
C3	1.829	.036		
C4	1.806	.036		

***p<.001

Decision Tree 지도학습 모델의 예측 성능 평가

K-평균 군집 방법을 활용하여 선수들의 경기 스타일 그룹으로 분류된 결과를 학습 데이터로 사용하여 구축 된 지도학습 모델(Decision Tree)의 예측 성능을 평가한 결과, 정확도 .818, 정밀도 .813±.033, 재현율 .802±.065, F1-점수 .807±.045로 나타났다. 또한 K-fold 교차검증을 진행하여 각 모델 성능 평가 지표의 값을 산출한 결과, 정확도 .792±.022, 정밀도 .784±.028, 재현율 .778±.023, F1-점수 .779±.026로 나타난 것을 확인하였다(Table 15).

Random Forest 지도학습 모델의 예측 성능 평가

K-평균 군집 방법을 활용하여 선수들의 경기 스타일 그룹으로 분류된 결과를 학습 데이터로 사용하여 구축 된 지도학습 모델(Random Forest)의 예측 성능을 평가한 결과, 정확도 .900, 정밀도 .908±.038, 재현율 .892±.053, F1-점수 .899±.022로 나타났다. 또한 K-fold 교차검증을 진행하여 각 모델 성능 평가 지표의 값을 산출한 결과, 정확도 .904±.022, 정밀도 .909±.023, 재현율 .890±.024, F1-점수 .897±.024로 나타난 것을 확인하였다(Table 16).

KNN 지도학습 모델의 예측 성능 평가

KNN 지도학습 모델의 예측 성능을 평가한 결과, 정확도 .945, 정밀도 .954±.038, 재현율 .936±.048, F1-점수 .944±.015로 나타났다. 또한 K-fold 교차검증을 진행하여 각 모델 성능 평가 지표의 값을 산출한 결과, 정확도 .894±.020, 정밀도 .898±.020, 재현율 .879±.023, F1-점수 .886±.021로 나타난 것을 확인하였다(Table 17).

Table 15. Performance evaluation metrics of the decision tree model

Cluster	Accuracy	Precision	Recall	F1-Score
C1	-	.767	.697	.730
C2	-	.825	.810	.817
C3	-	.805	.873	.838
C4	-	.857	.828	.842
Average	.818	.813±.033	.802±.065	.807±.045
K-fold	.792±.022	.784±.028	.778±.023	.779±.026

Table 16. Performance evaluation metrics of the random forest model

Cluster	Accuracy	Precision	Recall	F1-Score
C1	-	.966	.848	.903
C2	-	.891	.845	.867
C3	-	.893	.945	.918
C4	-	.885	.931	.908
Average	.900	.908±.038	.892±.053	.899±.022
K-fold	.904±.022	.909±.023	.890±.024	.897±.024

SVM 지도학습 모델의 예측 성능 평가

SVM 지도학습 모델의 예측 성능을 평가한 결과, 정확도 .977, 정밀도 .98±.019, 재현율 .972±.028, F1-점수 .976±.007로 나타났다. 또한 K-fold 교차검증을 진행하여 각 모델 성능 평가 지표의 값을 산출한 결과, 정확도 .959±.015, 정밀도 .959±.015, 재현율 .954±.018, F1-점수 .956±.016로 나타난 것을 확인하였다(Table 18).

다항 로지스틱 회귀분석 지도학습 모델의 예측 성능 평가

다항 로지스틱 회귀분석 지도학습 모델의 예측 성능을 평가한 결과, 정확도 .982, 정밀도 .984±.008, 재현율 .981±.01, F1-점수 .982±.003로 나타났다. 또한 K-fold 교차검증을 진행하여 각 모델 성능 평가 지표의 값을 산출한 결과, 정확도 .982±.011, 정밀도 .984±.010, 재현율 .981±.012, F1-점수 .982±.011로 나타난 것을 확인하였다(Table 19).

Table 17. Performance evaluation metrics of the KNN model

Cluster	Accuracy	Precision	Recall	F1-Score
C1	-	1.000	.871	.931
C2	-	.966	.933	.949
C3	-	.940	.984	.962
C4	-	.912	.954	.932
Average	.945	.954±.038	.936±.048	.944±.015
K-fold	.894±.020	.898±.020	.879±.023	.886±.021

Table 18. Performance evaluation metrics of the SVM model

Cluster	Accuracy	Precision	Recall	F1-Score
C1	-	1.000	.935	.967
C2	-	.983	.983	.933
C3	-	.955	1	.977
C4	-	.984	.969	.977
Average	.977	.98±.019	.972±.028	.976±.007
K-fold	.959±.015	.959±.015	.954±.018	.956±.016

Table 19. Performance evaluation metrics of the logistic regression model

Cluster	Accuracy	Precision	Recall	F1-Score
C1	-	.975	.983	.979
C2	-	.981	.991	.986
C3	-	.984	.982	.983
C4	-	.994	.967	.980
Average	982±.011	984±.008	.981±.01	.982±.003
K-fold	982±.011	.984±.010	.981±.012	.982±.011

Table 20. Analysis of differences in predictive performance among supervised learning models

	SS Between	*df*	MS	*F*(*p*)	post-hoc (Scheffe)
Accuracy	.159	4	.040	929.094***	a<b***=c<d***<e***
Precision	.157	4	.039	971.922***	a<b***=c<d***<e***
Recall	.159	4	.040	929.094***	a<b***=c<d***<e***
F1-score	.161	4	.0404	913.288***	a<b***=c<d***<e***

***p<.001
a= Decision Tree, b=Random Forest, c= K-Nearest Neighbors, d= Support Vector Machine, e= Multinomial Logistic Regression

지도학습 모델 간 예측 성능 차이 분석

본 연구에서 활용한 다섯 가지 지도학습 모델 간 예측 성능(정확도, 정밀도, 재현율, F1-점수)의 차이를 알아보기 위해 일원배치분산분석을 실시한 결과, 지도학습 모델 간 모든 예측 성능 지표에서 통계적으로 유의한 차이가 있는 것으로 나타났다. 사후분석 결과, 모든 예측 성능 지표에서 다항 로지스틱 회귀분석, 서포트 벡터 머신, K-최근접 이웃, 랜덤 포레스트, 의사결정나무 순의 예측 성능을 나타냈다(Table 20).

논 의

본 연구에서 KPGA 선수들의 경기력 기술 요인 데이터를 기반으로 선수들의 경기 스타일에 따라 군집화하고, 이를 토대로 예측·분류 지도학습 모델을 구축하여 각 모델의 성능을 평가한 결과에 대한 논의는 다음과 같다.

첫째, 본 연구에서는 KPGA 선수들의 기술 요인 데이터를 기반으로 경기 스타일에 따라 K-평균 군집화를 실시하기 전, 최적의 군집 수를 설정하기 위해 엘보우 기법을 사용하였으며, 보완 지표로 실루엣 지수와 갭 통계를 병행 사용하였다. 엘보우 기법을 실시한 결과, k=4까지는 WCSS 값이 급격히 감소하였고, k=5 이상부터는 감소 폭이 완만해지는 양상을 나타냈다. 이는 K=4로 설정하였을 때 군집화의 효율성과 해석 가능성 간 균형을 잘 반영하는 지점이라 판단할 수 있으며, K>4부터는 오히려 의미 없는 군집의 수가 늘어나는 것으로 판단할 수 있다. 이후 실루엣 분석에서도 K=4까지는 군집의 품질이 비교적 높게 유지되다가 이후 감소 폭이 점차 완만해지는 경향을 보였으며, 갭 통계에서는 K=3에서 가장 높은 값을 나타냈다. 이러한 결과를 기반으로 통계적 안정성과 군집 간 분리도 해석 가능성을 종합적으로 고려하였을 때, 최종적으로 K=4가 최적의 군집 수로 판단하였다. 이는 선행 연구(Kim & Park, 2021; Go, 2003; Kim & Choi, 2019; Choi et al., 2007)에서 제시한 바와 같이, 군집 수 결정 시 통계적 기준과 실질적 해석 가능성을 모두 고려해야 한다는 주장과 일치한다. 따라서 본 연구에서 도출된 k=4는 수치적 타당성과 해석적 타당성을 확보한 결과로서, 향후 KPGA 선수들의 경기 스타일을 기술 요인 수준에서 유형화하는 기초 근거로 활용할 수 있을 것으로 판단된다.

둘째, 엘보우 기법을 토대로 산출된 4개의 군집을 바탕으로 K-평균 군집 분석을 통해 선수들을 4개의 군집으로 분류한 뒤, 6가지 기술 요인에 대한 군집 간 차이를 분석하였다. 그 결과, 모든 기술 요인에서 군집 간 통계적으로 유의한 차이가 나타났으며(p<.001), 이는 분류된 각 군집의 구조적 타당성을 뒷받침해주는 결과로 해석할 수 있다. 이는 골프 경기력이 단순히 평균 타수나 상금 순위와 같은 결과 지표만으로 설명되기 보다는 선수의 기술적 역량과 경기 운영 방식의 복합적 작용에 의해 형성된다는 점을 시사한다.

각 군집에 속한 선수들의 경기 스타일을 고려하여 각 군집을 명명하였으며, 그 결과는 다음과 같다. 먼저, C1 군집은 '전반적 약세형'으로 명명되었다. 이 군집은 드라이브 거리에서는 중간 정도의 수준을 나타냈으나, 페어웨이 안착률, 그린 적중률, 리커버리율, 벙커 세이브율, 퍼트 수까지 모든 기술 요인에서 가장 낮은 수준을 보였다. 이러한 결과는 C1 군집의 선수들이 특정 기술 요소의 향상만으로는 경기력 개선이 어려운, 즉 전반적 기술 역량의 부족 상태에 있음을 시사한다. 골프는 다양한 기술 요소가 상호작용하여 성과를 결정하는 종목으로, 일정 수준 이상의 기술 균형이 확보되지 않으면 경기력 향상에 한계가 존재한다(Shin, 2015). 또한, 기술 요인 간 불균형은 경기 내 일관성을 저하시켜 전략적 의사결정의 폭을 제한할 수 있다. 따라서 C1 군집에 속한 선수들은 단일 기술의 강화보다는 전반적인 기술 요인 간 상호보완적 균형을 중점으로 한 트레이닝 접근이 필요하다. 이는 기술·전략 상호작용 관점에서 볼 때, 전략적 선택의 폭을 넓히기 위해서는 기술적 안정성이 선행되어야 함을 시사한다(McPherson & Kernodle, 2007).

C2 군집은 '거리 약세·기술 우수형'으로 명명되었다. 해당 군집 선수들의 드라이브 거리는 비교적 낮은 수준에 속했으나, 그린 적중률, 리커버리율, 벙커 세이브율, 퍼트 수 등 모든 기술 요인에서 가장 뛰어난 수준을 나타내 정확성 중심의 경기 운영 특성을 나타냈다. 이러한 결과는 공격적인 장타보다는 안정적인 샷 운영을 통해 스코어 손실을 최소화하려는 전략적 의사결정의 결과로 해석될 수 있다. 따라서 C2 군집에 속한 선수들이 비거리가 다소 부족하더라도 뛰어난 정확성과 위기 대응 능력을 바탕으로 안정적인 경기 운영이 가능함을 시사한다. Broadie and Rendleman(2013)은 프로 골퍼의 경기 전략을 분석한 연구에서, 정확성과 위험을 관리하는 전략이 경기력 향상에 핵심적 요인임을 제시하였으며, Broadie(2014) 역시 거리보다 정확성과 퍼팅 효율의 조합이 스코어 개선에 더 큰 영향을 미친다고 주장하였다. 이러한 맥락에서 C2 군집의 선수들은 장타보다는 정확성과 실수 최소화에 초점을 둔 보수적 경기 운영을 통해 경기 흐름을 관리하는 경향을 보인다. 이와 더불어 선행 연구(Lee & So, 2019; Kim, 2016; Lee, 2018)에서는 일반적으로 긴 드라이브 거리가 경기력 향상에 유리한 것으로 보고되었으나, Kwon et al.(2024)의 연구에서는 드라이브 거리가 평균 타수(par 4, par 5)에 유의한 영향을 미치지 않는 것으로 밝힌 바 있다. 이러한 점을 종합해 봤을 때, C2 군

집의 선수들은 무리한 장타를 시도하기보다는 현재의 정확성 기반 전략을 유지하되, 짧은 비거리를 보완할 수 있는 정교한 코스 매니지먼트 전략을 강화함으로써 경기력의 지속적 향상을 도모할 필요가 있을 것으로 판단된다

C3 군집은 '정확성 특화형'으로 명명되었다. 이는 Ball and Best(2007)가 제시한 기술적 효율성과 재현성을 강조하는 경기 스타일 유형과 유사한 특성을 나타내는 것을 확인하였다. 이 군집에 속한 선수들은 드라이브 거리가 가장 짧았으나, 페어웨이 안착률과 그린 적중률에서 높은 수준을 나타냈다. 이는 정확성 중심의 보수적 경기 운영 전략을 활용하는 스타일로 해석된다. 다수의 선행 연구에서는 그린 적중률과 페어웨이 안착률이 평균 타수에 유의한 영향을 미치는 핵심 기술 요인으로 보고하였으며(Hur, 2005; Heo et al., 2006; Kim & Seo, 2015; Kim, 2016; Kwon et al., 2024), 이러한 관점에서 C3 군집의 경기 운영 방식은 기술적 효율성이 높은 유형으로 평가될 수 있다. 따라서 C3 군집에 속한 선수들은 현재의 정확성을 유지하는 동시에 숏게임과 퍼팅 능력 향상을 통한 마무리 완성도를 높이는 방향으로 훈련 전략을 설정해야 할 것으로 판단된다.

마지막으로 C4 군집은 '장타 및 위험 관리형'으로 명명되었다. 이 군집에 속한 선수들은 가장 긴 드라이브 거리를 기록했으며, 리커버리율과 벙커 세이브율 또한 높은 수준을 나타냈다. 반면, 페어웨이 안착률은 가장 낮아 티 샷 정확도에서 다소 취약한 특성을 보였다. 이러한 결과는 단순한 기술적 구분을 넘어 골프 경기력의 결정 요인을 설명하는 기존 이론 틀과도 연계해 해석될 수 있다. Broadie(2014)의 경기력 결정요인론에서 제시한 거리(distance)와 리스크 관리(risk management) 요소의 상호작용을 반영한다. 또한, 이는 경기 전략적 의사결정이 기술 수행에 직접적인 영향을 미친다는 기술 수행과 전략·전술적 의사결정의 상호작용 관점(McPherson & Kernodle, 2007)과도 부합한다. 장타는 경기 중 클럽 선택과 전략적으로 중요한 장점으로 작용하지만(Lee, 2018) 동시에 정확도가 낮을 경우 스코어의 변동성이 커질 수 있는 위험 요인으로 작용할 수 있다(Wiseman et al., 2011). 본 연구 결과에서도 C4 군집의 선수들은 장타 기반의 공격적인 전략과 더불어 위기 상황에서의 회복 능력을 통해 이러한 변동성을 보완하고 있는 것으로 해석된다. 따라서 본 군집의 선수들은 장타 능력을 유지하면서도 정확도 향상과 위기 상황에서의 안정적 대응을 통해 경기 운영의 일관성을 제고할 필요가 있을 것으로 판단된다.

셋째, K-평균 군집 분석을 통해 분류된 선수들의 경기 스타일을 기반으로 지도학습 모델을 구축하고, 다양한 지도학습 모델의 예측 성능의 차이를 검증하기 위해 일원배치분산분석을 실시하였다. 그 결과, 지도학습 모델 간 예측 성능은 통계적으로 유의한 차이를 나타냈으며(p<.05), 사후검정을 통해 이러한 차이를 구체적으로 확인하였다. 그 중 다항 로지스틱 회귀분석이 정확도 .982, 정밀도 .984±.008, 재현율 .981±.01, F1-점수 .982±.003로 모든 모델 중 가장 뛰어난 성능을 기록하였으며, K-fold 교차검증에서도 유사한 양상이 반복되어 본 모델이 우수한 일반화 성능을 갖추고 있음을 확인할 수 있었다. 이는 로지스틱 회귀가 다중 클래스 분류 문제에서 안정성과 해석 용이성을 갖춘 기법임을 강조한 선행 연구(Kim & Lee, 2023; Kim, Park et al., 2024)의 결과와도 일치한다. 다음으로 서포트 벡터 머신 또한 정확도 .977, 정밀도 .98±.019, 재현율 .972±.028, F1-점수 .976±.007 다항 로지스틱 회귀분석과 근접한 수준의 매우 높은 성능을 나타냈다. SVM은 과적합 가능성이 낮고, 고차원 정형 데이터 분류에 적합한 기법으로, 이러한 특성을 강조한 선행 연구(Cortes & Vapnik, 1995; Noble, 2006; Pai et al., 2017; Kim & Lee, 2023)의 결과와 본 연구의 결과가 일치하는 것을 확인하였다. 반면, 의사결정나무 모델은 정확도 .818, 정밀도 .813±.033, 재현율 .802±.065, F1-점수 .807±.045로 가장 낮은 성능을 보였다. 의사결정나무는 해석이 용이하고 직관적 분류가 가능한 장점(Horvat & Job, 2020)이 있음에도 불구하고 본 연구와 같이 다차원적인 기술 요인이 복합적으로 작용하는 문제 상황에서는 과적합 및 분류 불안정성이 발생할 수 있는 문제로 인한 것으로 해석된다. 이러한 특성은 의사결정나무의 구조적 한계를 지적한 선행 연구와 일치한다(Bishop & Nasrabadi, 2006). 이와 비교해 랜덤 포레스트 모델은 의사결정나무의 불안정성을 보완하여 정확도 .900, 정밀도 .908±.038, 재현율 .892±.053, F1-점수 .899±.022로 의사결정나무보다 향상된 성능을 보였다. KNN 모델 역시 정확도 .945, 정밀도 .954±.038, 재현율 .936±.048, F1-점수 .944±.015로 우수한 성능을 나타냈다. 다만, KNN은 학습 속도가 빠른 장점을 가지고 있으나, 예측 시 연산량이 많고 데이터 밀도나 분포에 따라 성능 변동이 클 수 있어, 실시간 예측 시스템에는 상대적으로 불리하게 작용할 수 있다(Horvat, Stanojević et al., 2020).

본 연구에서 활용한 5개 지도학습 모델의 예측 성능을 비교·분석한 결과를 정리해 보자면, 다항 로지스틱 회귀분석, 서포트 벡터 머신, K-최근접 이웃, 랜덤 포레스트, 의사결정나무 순의 예측 성능을 나타냈으며, 모든 모델이 군집화된 경기 스타일을 일정 수준 이상으로 정확하게 분류할 수 있는 성능을 확인하였다. 특히 다항 로지스틱 회귀분석과 서포트 벡터 머신이 가장 높은 예측력과 안정성을 나타낸다는 점에서 KPGA 선수들의 경기 스타일을 분류 및 예측하는 데 있어 우선적으로 고려될 수 있는 모델임을 시사한다. 이러한 결과는 골프 종목에서 선수들의 기술 요인 데이터를 기반으로 경기 스타일을 자동으로 분류하고 예측하는 지도학습 기반 모델의 적용 가능성을 뒷받침한다.

결론 및 제언

본 연구는 KPGA 투어 선수들의 6가지 경기력 기술 요인 데이터를 바탕으로 선수들의 경기 스타일을 군집화하고, 이를 기반으로 지도학습 모델을 구축하여 향후 선수의 경기 스타일을 자동으로 예측 및 분류할 수 있는 시스템을 개발하여 적용 가능성을 검증하고자 진행되었다. 그 결과, 다음과 같은 결론을 도출하였다.

첫째, KPGA 코리안투어 선수들의 경기 스타일은 기술 요인에 따라 4개의 군집으로 분류되었다.

둘째, 각 군집은 기술 특성에 따라 '전반적 약세형'(C1), '거리 약세·기술 우수형'(C2), '정확성 특화형'(C3), '장타·위험 관리형'(C4)으로 명명되었으며, 각 유형은 뚜렷한 기술 요인의 조합으로 구분되었다.

셋째, 선수들의 경기 스타일을 자동으로 예측·분류할 수 있는 지도학습 모델을 검증하기 위해 5가지 지도학습 기법을 적용한 결과, 다항 로지스틱 회귀분석이 가장 높은 성능을 나타냈으며, SVM, KNN, 랜덤 포레스트, 의사결정나무 순의 예측 성능을 나타냈다.

결론적으로, 본 연구에서 설정한 군집화 기반 분석은 기존의 평균 타수 중심의 일률적 접근 방식과 달리, 선수 개인의 기술적 조합과 경기 스타일을 정밀하게 구분할 수 있는 분석 방법임을 시사한다. 아울

러, KPGA 코리안투어 선수들의 경기 스타일은 기술 요인 데이터를 기반으로 4개의 유형으로 명확히 분류될 수 있으며, 이를 자동으로 분류·예측할 수 있는 지도학습 기반 모델의 적용 가능성도 확인되었다. 이러한 결과는 향후 선수 맞춤형 훈련 전략 수립, 코스 환경에 따른 경기 운영 전략 도출, 나아가 AI 기반 스포츠 분석 시스템 개발에 있어 실질적인 기여 가능성을 기대한다.

다만, 본 연구에서는 군집화 분석을 통해 선수들의 경기 스타일을 분류하였으나, 각 스타일 유형이 평균 타수와 같은 실질적 경기 성과에 어떤 영향을 미치는지에 대한 직접적인 분석은 수행하지 못했다. 따라서 후속 연구에서는 각 경기 스타일 별로 평균 타수에 유의한 영향을 미치는 핵심 기술 요인을 규명함으로써, 스타일 특성에 기반한 구체적 기술 역량 도출이 이루어질 필요가 있을 것으로 사료된다.

CONFLICT OF INTEREST

논문 작성에 있어서 어떠한 조직으로부터 재정을 포함한 일체의 지원을 받지 않았으며 논문에 영향을 미칠 수 있는 어떠한 관계도 없음을 밝힌다.

AUTHOR CONTRIBUTION

Conceptualization: Hong-suk Kim, Data curation: Hong-suk Kim & Ji-yong Park, Formal analysis: Hong-suk Kim & Hye-su Cho, Methodology: Hong-suk Kim & Hyun-su Park, Project administration: Hyun-su Park, Visualization: Ji-yong Park, Writing – original draft: Hong-suk Kim, Writing – review & editing: Hong-suk Kim, Hye-su Cho, Ji-yong Park & Hyun-su Park

참고문헌

Alonso, R. P., & Babac, M. B. (2022). Machine learning approach to predicting a basketball game outcome. *International Journal of Data Science, 7*(1), 60-77. https://doi.org/10.1504/IJDS.2022.124356

American Psychological Association. (2017). *Ethical principles of psychologists and code of conduct*. American Psychological Association. https://www.apa.org/ethics/code

Ball, K. A., & Best, R. J. (2007). Different centre of pressure patterns within the golf stroke I: Cluster analysis. *Journal of Sports Sciences, 25*(7), 757-770. https://doi.org/10.1080/02640410600874971

Berry, M. J. A., & Linoff, G. S. (1997). *Data mining techniques: For marketing, sales, and customer support*. John Wiley & Sons. https://doi.org/10.5555/2543983

Bishop, C. M., & Nasrabadi, N. M. (2006). *Pattern recognition and machine learning*. Springer. https://doi.org/10.1117/1.2819119

Broadie, M. (2014). *Every shot counts: Using the revolutionary strokes gained approach to improve your golf performance and strategy on the course*. Gotham Books.

Broadie, M., & Rendleman, R. (2013). Risk and reward: Course management in professional golf. *Operations Research, 61*(6), 1248-1260. https://doi.org/10.1287/opre.2013.1229

Choi, C.-I., Moon, J.-Y., & Lee, E.-C. (2024). Application and analysis of nonlinear regression models using bat tracking and hitting metrics for baseball batting performance prediction. *Journal of Next-generation Convergence Technology Association, 8*(12), 2881-2890. https://doi.org/10.33097/JNCTA.2024.08.12.2881

Choi, H. J. (2022). Comparison of machine learning methods for predicting match outcomes in soccer. *Journal of the Korean Society of Measurement and Evaluation in Physical Education and Sport, 24*(4), 81-91. https://doi.org/10.21797/ksme.2022.24.4.007

Choi, H. J., Kim, J. H., & Go, B.-G. (2007). Application of self-organizing maps for cluster analysis. *Journal of the Korean Society of Physical Education, 46*(5), 553-564.

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

DataBuckets. (2015). *Classifying players on the PGA tour with clustering methods*. DataBuckets blog. https://databuckets.org/databucket/2015/05/classifying-types-of-players-on-pga.html

Davis, J., Wiemeyer, J., & Baca, A. (2024). Methodology and evaluation in sports analytics: Challenges, approaches, and lessons learned. *Machine Learning, 113*(4), 1741-1762. https://doi.org/10.1007/s10994-024-06585-0

Dormann, C. F., Elith, J., Bacher, S., Buchmann, C., Carl, G., Carré, G., … & Lautenbach, S. (2013). Collinearity: A review of methods to deal with it and a simulation study evaluating their performance. *Ecography, 36*(1), 27-46. https://doi.org/10.1111/j.1600-0587.2012.07348.x

Go, B. G. (2003). Cluster analysis of sport events based on morphological similarity among elite athletes. *Journal of the Korean Society of Physical Education, 42*(6), 1007-1018.

Heo, C., Cho, G. K., Jeong, W. J., & Choi, S. B. (2006). Analysis of technical factors related to performance of professional golfers - Focused on PGA tour and nationwide tour players. *Korean Sport Research, 17*(3), 647-656. UCI: I410-ECN-0102-2009-690-000421354

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), e1380. https://doi.org/10.1002/widm.1380

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. https://doi.org/sym12030431

Horvat, T., Stanojević, D., & Petrović, S. (2020). *Predicting NBA outcomes using machine learning methods: KNN, Decision Trees, SVM, and Random Forest*. ArXiv Preprint. Retrieved from https://arxiv.org/abs/2010.02059

Hur. N. Y. (2005). Analysis of technical factors affecting golf scores. *Korean Journal of Physical Education, 44*(4), 617-623.

Jeong, Y. T., & Choi, Y. J. (2022). Statistical research in physical education: Focusing on regression analysis. *Korean Journal of Physical Education, 61*(5), 125-138. https://doi.org/10.23949/kjpe.2022.9.61.5.125

Jo, E., Kim, M., & Lee, S. (2023). Development of a supervised learning–based performance prediction model in elite sports. *International Journal of Sports Science & Coaching, 18*(3), 742-754.

Kang, M. S. (2023). Building a machine learning model to predict the outcome of international soccer matches. *The Korean Journal of Physical Education, 62*(5), 45-56. https://doi.org/10.23949/kjpe.2023.9.62.5.4

Kim, D. M., & Choi, H. J. (2019). Cluster analysis of performance indicators based on official records of the English Premier League. *Korean Journal of Sport Science, 28*(2), 1237-1245.

Kim, J. E., Park, J. M., & Park, J. C. (2024). Application of machine learning for position and match outcome prediction using physical data in soccer games. *Sport Science, 42*(2), 13-21. https://doi.org/10.46394/ISS.42.2.2

Kim, M. S., & Seo, E. C. (2015). A model exploration for explaining golf performance based on KPGA records. *Korean Journal of Sports Science, 13*(2), 337-346. UCI: G704-SER000001967.2015.13.2.012

Kim, N. J., & Min, D. G. (2014). A study on changes in performance of KLPGA players using growth curve modeling. *Journal of the Korean Data and Information Science Society, 25*(4), 847-855. https://doi.org/10.7465/jkdi.2014.25.4.847

Kim, P. S., & Lee, S. H. (2023). Predicting final rankings in the Korean professional basketball league regular season using machine learning: A sports analytics approach. *Korean Journal of Measurement and Evaluation in Physical Education and Sport Science, 25*(2), 103-115. https://doi.org/10.21797/ksme.2023.25.2.008

Kim, P., Lee, S. H., & Jeon, S. S. (2024). Predicting the outcomes of K-League matches using machine learning algorithms. *Journal*

of Digital Contents Society, 25(4), 1027-1037. https://doi.org/10.9728/dcs.2024.25.4.1027

Kim, P., Lee, S. H., & Jeon, S. S. (2024). Machine learning–based prediction and evaluation of competition rankings in professional sports. *Korean Journal of Sport Management, 29*(1), 55-72.

Kim, P. S., Lee, S. H., & Woo, S. B. (2024). Sociological implications of applying machine learning algorithms to predict KLPGA players' performance in the sports industry. *Korean Journal of Sociology of Sport, 37*(Special Issue), 122-145. https://doi.org/10.22173/ksss.2024.37.SpecialIssue.8

Kim, S. H., & Cho, J. H. (2013). Multi-group analysis using professional golf performance data: Application of path analysis. *Journal of the Korean Data and Information Science Society, 24*(3), 543-555. https://doi.org/10.7465/jkdi.2013.24.3.543

Kim, S. H., Lee, J. W., & Lee, M. S. (2012). The effect of PGA players' performance factors on scoring average. *Journal of the Korean Data and Information Science Society, 23*(3), 505-514. https://doi.org/10.7465/jkdi.2012.23.3.505

Kim, S. I. (2016). Analysis of technical factors in golf performance on LPGA Tour events (1993-2015). *Journal of Golf Studies, 10*(4), 49-59. UCI: G704-SER000002249.2016.10.4.008

Kim, S. I. (2010). Analysis of skill factors for the improvement of golf performance in golf tour. *Journal of Coaching Development, 12*(3), 103-112. UCI: G704-001507.2010.12.3.012

Kim, S. M., & Kim, S. Y. (2010). Effects of performance factors on scoring average and tour performance in Korean female professional golfers. *Korean Journal of Women in Physical Education, 24*(2), 129-140. UCI: G704-001368.2010.24.2.008

Kim, Y. J., & Park, H. W. (2021). Cluster analysis of Korean female professional golfers based on performance records. *Korean Journal of Sport Science, 30*(2), 1025-1032. https://doi.org/10.35159/kjss.2021.4.30.2.1025

Korea Professional Golfers' Association. (2024). KPGA Records. https://www.kpga.co.kr/tours/record/?tourId=11

Kwon, T. W., Cho, H. S., Kim, H. S. (2024). Analysis of golf skill factor of KPGA Korean tour according to the type of hole. *The Korean Journal of Measurement and Evaluation in Physical Education and Sports Science, 26*(1), 67-77. http://dx.doi.org/10.21797/ksme.2024.26.1.067

Kwon, T., Cho, H., & Kim, H. (2024). Performance analysis of KPGA Korean Tour players according to hole type. *Korean Journal of Sport Measurement and Evaluation, 26*(1). https://doi.org/10.21797/ksme.2024.26.1.067

Lee, J. W. (2018). A meta-analysis on the effects of training programs on driver distance among golf participants. *Journal of Golf Studies, 12*(3), 15-32. https://doi.org/10.34283/ksgs.2018.12.3.15

Lee, S. K., & So, J. M. (2019). Kinematic characteristics based on club head acceleration in golf driver swings. *Journal of Golf Studies, 13*(2), 29-40. https://doi.org/10.34283/ksgs.2019.13.2.03

McPherson, S. L., & Kernodle, M. W. (2007). Tactical and technical decision-making in sport: The role of knowledge. *Journal of Sports Sciences, 25*(11), 1315-1328. https://doi.org/10.1080/02640410601129780

Min, D. G. (2011). A path analysis of factors affecting scoring average using 2010 PGA Tour statistics. *Journal of the Korean Data and Information Science Society, 22*(1), 65-71. UCI: G704-000605.2011.22.1.005

Min, D. K. (2011). The study for effectiveness of golf skills to adjust average score using path analysis in 2010 PGA. *Journal of the Korean Data & Information Science Society, 22*(1), 65-71. UCI : G704-000605.2011.22.1.005

Noble, W. S. (2006). What is a support vector machine?. *Nature Biotechnology, 24*(12), 1565-1567.

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*, 4159-4167.

Pathak, N., & Wadhwa, H. (2016). Applications of modern classification techniques to predict the outcome of ODI cricket. *Proceedings of the International Conference on Computational Science (ICCS), Procedia Computer Science, 87*, 55-60. https://doi.org/10.1016/j.procs.2016.05.329

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., Vanderplas, J., Passos, A., Cournapeau, D., Brucher, M., Perrot, M., & Duchesnay, É. (2011). Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research, 12*, 2825-2830.

Pyung, C. S., Yeo, I. S., Chung, S. H. (2015). An analysis on the tour performance determinants of KPGA players. *The Korea Journal of Sports Science, 24*(6), 1-10. UCI: G704-001369.2015.24.6.062

Quinn, R. J. (2006). Exploring correlation coefficients with golf statistics. *Teaching Statistics, 28*(1), 10-13. https://doi.org/10.1111/j.1467-9639.2006.00229

Shin, S. N. (2015). Analysis of the relationship between technical factors and points scored among Korean middle and high school golf players. *Journal of Golf Studies, 9*(4), 77-86. UCI: G704-SER000002249.2015.9.4.003

Smith, M. F. (2010). The role of physiology in the development of golf performance. *Sports Medicine, 40*, 635-655.

Son, S. B., & Kim, Y. G. (2008). Analysis of performance determinants in the 2007 PGA and LPGA Tour. *Journal of Sport and Leisure Studies, 32*(2), 1185-1194. https://doi.org/10.51979/kssls.2008.05.32.1185

Son, S. B., & Kim, Y. G. (2010). Analysis of performance determinants using 2008 PGA and LPGA Tour statistics. *Journal of Coaching Development, 12*(1), 151-160. UCI: G704-001507.2010.12.1.010

Tabassum, H., Fiaz, M., Fiaz, M. A., & Akram, T. (2022). Supervised machine learning algorithms in the high-dimensional setting. *Computational Intelligence and Neuroscience, 2022*, Article 5816145. https://doi.org/10.1155/2022/5816145

Tan, P.-N., Steinbach, M., Karpatne, A., & Kumar, V. (2019). *Introduction to data mining* (2nd ed.). Pearson.

Wiseman, F., Habibullah, M., & Friar, J. (2011). The importance of driving distance and driving accuracy on the PGA and Champions Tours. *The Sport Journal, 14*(1), 1-4.

머신러닝 기반 골프선수의 경기 스타일 분류 및 예측 시스템 개발

김홍석[1], 조혜수[2], 박지용[3], 박현수[4*]

[1]한양대학교 ERICA, 스포츠과학과, 박사
[2]한양대학교 ERICA, 스포츠과학부, 조교수
[3]한양대학교, 스포츠과학과, 박사과정
[4]한양대학교, 스포츠과학과, 박사

*교신저자: 박현수(gustn1473@naver.com)

[목적] 본 연구는 KPGA 선수들의 경기력 기술 요인 데이터를 기반으로 선수들의 경기 스타일을 유형화하고, 이를 자동으로 분류할 수 있는 지도학습 기반 모델을 구축하여 적용 가능성을 검토하는 데 목적이 있다.
[방법] 2015년부터 2024년까지 KPGA 코리안투어에 참가한 선수들의 6가지 경기력 기술 요인 데이터를 수집하여 분석에 활용하였다. 기술 요인을 z-점수로 표준화한 뒤, K-평균 군집분석을 통해 선수들의 경기 스타일을 군집화하고, 이를 바탕으로 다섯 가지 지도학습 기법(의사결정나무, 랜덤 포레스트, K-최근접 이웃, 서포트 벡터 머신, 다항 로지스틱 회귀)을 적용하여 자동 분류 모델을 구축하였다. 각 모델의 예측 성능은 정확도, 정밀도, 재현율, F1-점수를 기준으로 평가하였으며, 5-fold 교차검증을 통해 일반화 성능을 검증하였다.
[결과] 첫째, KPGA 선수들의 경기 스타일은 4개의 유형으로 군집화되었으며, 각 군집은 기술적 특성에 따라 '전반적 약세형', '거리 약세·기술 우수형', '정확성 특화형', '장타·위험 관리형'으로 명명되었다. 둘째, 지도학습 기법 비교 결과, 다항 로지스틱 회귀분석이 가장 높은 예측 성능을 보였으며, SVM, KNN, 랜덤 포레스트, 의사결정나무 순의 예측 성능을 나타냈다.
[결론] KPGA 코리안투어 선수들의 경기 스타일은 기술 요인 데이터를 기반으로 4개의 그룹으로 명확히 구분될 수 있으며, 이를 자동으로 분류·예측할 수 있는 지도학습 기반 모델의 적용 가능성 또한 확인되었다. 이러한 결과는 향후 선수 맞춤형 훈련 전략 수립, 코스 환경별 경기 운영 전략 도출, 나아가 AI 기반 스포츠 분석 시스템 개발에 있어 실질적인 기여를 할 수 있을 것으로 기대된다.

주요어
한국프로골프협회, 군집화, 경기 스타일, 경기력 기술 요인, 머신러닝