석사학위 논문

머신 러닝을 통한
현대 농구 포지션 재정의

Redefining Positions for the Modern Basketball
with Machine Learning

명지대학교 기록정보과학전문대학원

스포츠기록분석전공

김 나 엘

지도교수 김 주 학

2022년 2월

머신 러닝을 통한
현대 농구 포지션 재정의

Redefining Positions for the Modern Basketball
with Machine Learning

이 논문을 석사학위 논문으로 제출함.

2022년 2월

명지대학교 기록정보과학전문대학원
스포츠기록분석전공
김 나 엘

머신 러닝을 통한
현대 농구 포지션 재정의

Redefining Positions for the Modern Basketball
with Machine Learning

명지대학교 기록정보과학전문대학원
스포츠기록분석전공
김 나 엘

상기자의 스포츠기록분석학석사 학위논문을 인준함.

심사위원장 _____ (인)

심 사 위 원 _____ (인)

심 사 위 원 _____ (인)

2022년 2월

목 차

표목차

그림목차

머신 러닝을 통한 현대 농구 포지션 재정의

김 나 엘

명지대학교 기록정보과학전문대학원 스포츠기록분석전공

지도교수 김 주 학

농구 경기의 빠른 공수전환처럼 현대 프로농구 선수들의 플레이 스타일도 빠르게 변화하고 있다. 플레이 스타일은 선수의 역할과 포지션을 의미한다. 단체종목에서의 포지션은 팀 내에서 선수들의 역할을 의미하고, 팀 전술과 전력을 구성하는데에 필수적으로 고려해야할 요인이다. 농구의 전통적인 5가지 포지션(PG, SG, SF, PF, C)은 공식적으로 정해진 객관적인 기준도 없을뿐더러, 현재 진행형으로 변화 중인 현대 프로농구 선수들의 포지션을 정의하는 것은 더욱 한계가 있다. 따라서 이 연구에서는 미국 프로농구와 한국 프로농구를 각 리그별로 비지도학습 방법 중 k-평균 군집분석을 사용해 새로운 군집으로 분류하였다. 분류된 각 군집을 군집의 특성에 맞는 새로운 이름으로 포지션을 재정의하였다. 그리고 배깅을 활용한 앙상블 기법 중 랜덤포레스트를 사용하여, 군집분석을 통해 정의된 새로운 포지션으로 분류하는 모델을 개발하였다. 그 결과의 요약은 다음과 같다.

첫째, 각 리그별로 수집된 데이터는 KBL은 364명과 10개의 변수, NBA는 899명과 18개 변수로 구성되어있다. 수집된 데이터를 정제하기 위해 결측값과 이상값을 각 변수의 평균으로 대체하였다. 또한 정제된 데이터를 로그 변환을 통하여 범위를 축소하였다. 그리고 Min-Max 정규화 과정을 통해 각 변수들의 단위를 0과 1사이로 변환하였다.

둘째, 각 리그별로 수집된 변수들을 k-means 방법을 통하여 군집화하였다. 군집화 전에 엘보우 기법과 실루엣 기법을 통하여 KBL의 k값이 4, NBA의 k값은 6으로 설정하였다. k값에 따라 분류된 군집들을 기술통계량을 통하여 군집별 특성을 파악하였고, 그에 맞는 포지션으로 군집들을 새롭게 정의하였다. KBL의 4가지 군집은 Main Ball

Handler, Linker, Finisher, 3&D로 정의되었다. NBA의 6가지 군집은 KBL의 4가지 포지션과 더불어 Long Ranger와 Traditional Bigman으로 정의하였다. 재정의된 포지션을 Bonferroni 검정을 통해서 각 포지션 간에 통계적으로 유의미한 차이가 있음($p<0.05$)을 검증하였다.

셋째, 라벨링된 포지션이 추가된 모든 데이터셋을 6:4로 학습 데이터와 테스트 데이터를 분할하였다. 각 리그별로 새롭게 정의된 포지션은 랜덤포레스트 모델의 예측변수로써 사용하였다. 나머지 변수들은 모델 생성을 위한 입력 데이터로 사용하였다. 초기 모델은 KBL은 91.0%, NBA는 88.0%의 성능을 보였다. 하지만 학습데이터에 대한 과적합 현상으로 Random Search방법을 통해 하이퍼 파라미터를 튜닝하였다. 튜닝된 최종 모델의 성능은 KBL 94.3%, NBA 92.6%로 향상되었다. 또한, 최종 모델에서 KBL은 Shoot%, NBA는 Average Dirbble per Touch 변수가 가장 높은 변수 중요도를 보였다.

이 연구는 연속형 변수들의 변수 중요도에 따라 선수들을 분류하였다. 사용 변수들은 플레이 결과에서 독립적인 변수들을 사용하여 플레이 특성을 분류했다는 점에서 기존 연구와 차별성이 있다. 또한, 기존의 주관적인 선수 포지션 정의가 아닌 과학적이고 객관적인 접근 방법으로 동일한 기준에서 농구 포지션 정의할 수 있다는 것을 제안한다는 점에서 의미가 있다고 할 수 있다.

주제어
농구, 포지션, 특성, 머신러닝, 군집분석, 랜덤포레스트

제 1장 서 론

제 1절. 연구의 필요성

농구 경기에서 포지션은 팀 전력과 전술을 구성하는데에 필수적으로 고려해야할 요인이다. 현대 미국 프로농구에서는 포지션 구분이 모호해지며 다양한 포지션의 역할을 수행하는 선수들이 많이 생겨나고 있다. 한국 프로농구도 과거에 비해 공격과 수비에서 포지션의 역할이 많이 평준화 되어 올라운드 플레이어로써의 역할을 수행하는 추세로 변화하고 있다(조성원, 2015). 김주학·최형준(2015)은 종목 특성상 농구는 불과 몇 초 안에 결과가 뒤바뀔 수 있으므로 경기 분석을 통한 과학적인 지원을 통해 전술이나 전략의 변화가 가능하다고 설명하였다. 빠르게 진행되는 농구 종목 특성처럼 선수들의 플레이 특성도 계속해서 빠르게 변화하고 있다. 따라서 이러한 시대 흐름과 변화에 맞춰 선수단 구성에 가장 기본이 되는 포지션이라는 요인을 과학적이고 객관적인 방법을 통하여 재정의할 필요가 있다.

저우차우·최형준(2019)은 세계 남자 농구 월드컵 경기대회 공식기록을 기반으로 군집 분석하여 팀별 특성을 분류하여 군집별 차이를 분석하였다. 한수철(2008)은 한국 프로농구 선수들의 특성을 군집화하였다. 또한, Zhang Shao Liang(2018)은 NBA 전체 30개 팀을 선수들의 신체조건과 경기장 각 구역에서의 성공률을 변인으로 하여 3개의 서로 다른 군집으로 분류하여 팀을 평가하였다.

하지만 앞선 연구들에서는 농구 선수와 팀의 특성을 분류하는데 사용된 변인들이 플레이 스타일보다는 개인 기량을 나타내는 지표들을 사용하였다. 득점, 어시스트, 리바운드, 턴오버, 신체조건 등과 같은 기록지에 속해있는 변인들을 입력변수로 사용했을 때의 군집분석 결과는 기량이 좋은 선수와 상대적으로 기량이 좋지 않은 선수로 군집이 형성된다는 치명적인 한계점이 있다.

이러한 앞선 연구들의 한계점을 보완하기 위해서 Advanced stat 안에 있는 선수의 플레이 특성을 나타내는 지표를 활용할 필요성이 있다. 최근 NBA와 유럽농구리그에서는 야구의 세이버메트릭스의 영향으로 Advanced stat의 발전과 이를 활용하는 사례가 늘고 있다. Julius Demenius(2017)은 농구 종목에서 Advaced stat의 활용이 팀 퍼포먼스와 리그 활성화에 미치는 영향을 분석하였고, Radiboj Mandic(2019)는 NBA와 유럽농구리그 간 Advanced stat과 박스스코어 차이를 통해 리그 수준을 비교 분석하였다.

이렇게 해외 농구 리그에서는 Advanced stat을 활용하여 기존에 존재했던 방법론에서 발전시켜 새로운 인사이트를 도출하려는 다양한 연구를 시도하고 있다. 하지만 국내 프로농구에서는 아직까지 Advanced stat의 완벽한 도입이 되지 못했고, 이를 활용한 연구가 부족하다.

Advanced stat에는 슈팅 유형별 가중치를 부여한 효율, 종합 생산성 효율 지표 등 다양한 선수 평가 지표도 있지만, 선수의 플레이 특성을 나타내는 지표들이 존재한다. 이를 군집분석의 입력변수로 사용하였을 때, 앞선 연구들에서 발견한 기량이 좋은 선수와 아닌 선수로 군집이 형성되는 한계점을 해결할 수 있다.

따라서 본 연구에서는 국내 프로농구(KBL)와 미국 프로농구(NBA)에서 사용되는 Advanced stat들 중에서 플레이 특성을 나타내는 다양한 변인들을 선택 및 가공하여 이를 군집분석의 입력변수로 설정한다. 그리고 형성된 각 군집 별로 기술 통계량을 관찰하여 각 군집의 특성을 파악하고 그에 맞는 새로운 포지션 이름을 부여한다. 이를 바탕으로 생성된 새로운 데이터셋을 기반으로 농구 선수 포지션을 분류하는 모델을 개발한다. 또한, 재정의된 포지션을 기반으로 어떤 유형의 선수들이 존재하는지 미국 프로농구(NBA)와의 비교 분석을 하여 국내프로농구(KBL)가 세계 경쟁력을 가지기 위해서 갖추어야 할 플레이 스타일을 파악하고 국내 프로농구 경기력 향상과 세계 경쟁력을 위해 나아가야 할 방향을 제시하는 하나의 지표를 개발한다.

제 2절. 연구 목적

본 연구는 군집분석 방법 중 k-means 군집 방법을 통해, KBL과 NBA 선수들을 플레이 특성을 나타내는 지표들을 기반으로 군집화 하여, 각 군집의 특성에 맞게 새로운 포지션을 정의하는 것에 목적이 있다.

또한, 배깅을 활용한 앙상블 기법 중 하나인 랜덤포레스트를 활용하여 재정의된 포지션으로 분류하는 모델을 개발한다.

제 3절. 연구 문제

연구문제 1. 수집된 데이터를 탐색한다.

1) 각 변수별 기술통계량을 통하여 데이터 타입과 단위 및 범위를 파악한다.
2) 결측값, 이상값 여부를 파악한 뒤 평균값으로 대체한다.
3) 로그 변환 후 Min-Max 정규화 과정을 통해 단위를 변환한다.

연구문제 2. 군집모델을 개발한 뒤 새로운 포지션을 정의한다.

1) 리그별 최적의 k값을 선정한다.
2) 군집 별로 특성에 맞는 새로운 포지션 이름을 정의한다.
3) 두 리그의 우승팀의 재정의된 포지션 빈도와 경향을 비교한다.

연구문제 3. 포지션을 분류하는 모델을 개발한다.

1) 연구문제 2에서 생성된 데이터를 훈련 데이터와 테스트 데이터로 나눈다.
2) 훈련 데이터로 분류 모델을 생성한다.
3) 테스트 데이터로 분류 모델의 성능을 판단한다.
4) 분류 모델의 성능을 개선하여 최종 모델을 생성한다.
5) 최종 모델의 변수 중요도를 파악한다.

제 4절. 연구의 제한점

각 팀의 팀 컬러와 감독의 전술과 같이 데이터로 객관화 할 수 없는 요소들은 고려하지 않는다.

비지도 학습을 통해 재정의된 데이터를 분류 모델의 훈련 데이터로 사용되기 때문에 모델의 정확도와 성능을 비교할 수 있는 데이터가 존재하지 않는다.

제 5절. 용어의 정의

이 연구에서 사용된 주요 용어의 정의는 다음과 같다.

(1) 1차 스탯

경기수, 출전시간, 득점, 어시스트, 리바운드, 턴오버, 야투성공률, 자유투성공률 등을 포함하는 데이터 집합

(2) Advanced stat (2차 스탯)

1차 stat을 2차 가공하여 생성된 데이터 집합

(3) Tracking data

경기장 내에 설치된 레이더를 통해 수집되는 데이터 집합

제 2장 이론적 배경

제 1절. 농구 경기 데이터 분석

1.1 농구 경기분석의 종류

농구에서 경기 분석은 크게 동영상을 활용한 전술분석과 숫자로 이루어진 지표를 활용한 데이터 분석으로 나뉜다. 전술분석은 문자 그대로 상대의 공격, 수비 전술을 파악하는 분석방법이다. 몇 초 사이에 전술이 바뀌고 움직이기 때문에 경기 앞두고 동영상을 활용한 전술분석은 단기간에 전체 팀 선수들에게 효과적인 분석방법이다(박환석, 2014).

또한, MLB(미국프로야구) 세이버 메트릭스의 영향을 받아 NBA에서는 Advanced stat을 개발하여 선수평가 지표로써 활용되고 있다. Advanced stat의 활용으로 전통적인 1차 스탯(득점, 어시스트, 리바운드, 턴오버, 야투성공률 등)으로만 선수를 평가하는 것이 아니라, 승리기여도(Win Share), 선수유형(Usage%, Individual Floor Percentage) 등으로 선수 퍼포먼스의 종합적인 평가 방법이 다양해지고 보다 정확해졌다. Julius Demenius(2017)는 Advanced stat을 활용한 농구 경기분석은 경기력 향상과 리그 흥행에도 긍정적인 영향을 미친다는 것을 연구하며 이를 적극적으로 활용할 것을 제안하였다.

최근 NBA에서는 선수들의 움직임과 위치를 감지하는 레이더를 통해 Tracking 데이터를 수집하고 있다. 이처럼 NBA에서는 다양하고 과학적인 방법으로 데이터 수집과 분석을 시도하고 있다.

1.2 농구장 구역별 명칭

이 연구에서 언급되는 주요 농구코트 명칭의 정의는 다음과 같다.

1) Back Court

3점 라인 바깥쪽 구역을 의미한다.

2) Front Court

3점 라인 안쪽 구역을 의미한다.

3) Paint Zone

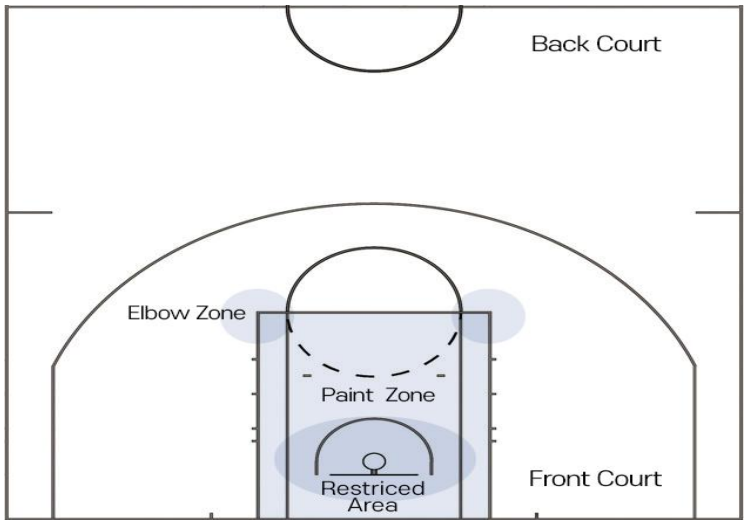
3점 라인 안쪽 페인트가 칠해져 있는 직사각형 형태의 구역을 의미한다.

4) Elbow Zone

Paint Zone 상단의 모서리 구역을 의미한다.

5) Restricted Area

Paint Zone에서 골밑으로 로우-포스트(Low-Post)라고 불리우는 구역을 의미한다.



<그림 1> 농구장 구역별 명칭

1.3 선수 평가 지표

선수 평가에 주로 사용되는 Advanced stat들은 Dean Oliver, Bob Bellotti, Hollinger 등 다양한 접근을 통해 개발 및 발전되었다. 현재 농구 경기 분석과 선수 평가에 사용되는 주요 Advanced stat은 다음과 같다.

1) Usage Rate

선수가 경기를 뛰는 동안에 팀에서 해당 선수를 활용하는 정도를 추정한 값으로 팀 내에서의 각 선수들의 플레이 비중을 파악할 수 있는 지표이다.

2) Individual Floor Percentage

해당 선수의 전체 possession중에 scoring possesion이 차지하는 비율로 득점 시도를 했을 때의 공격 효율을 판단할 수 있는 지표이다. Usage Rate와의 관계를 통해, 선수들의 기량 수준을 판단할 수 있다.

3) Touches

공격시에 공을 만진 횟수를 슈팅, 패스, 피파울, 턴오버 등으로 추정한 값이다. Touches는 Shoot%, Pass%, 등을 통해 공을 소유했을 때의 플레이 유형을 파악할 수 있는 지표이다.

4) Rebound%

해당 선수가 경기를 뛰는 동안 발생하는 리바운드가 발생할 때 해당 선수가 리바운드를 잡을 확률을 의미한다.

5) Shooting Frequency

전체 슈팅 시도에서 2점슛 시도 비율, 3점슛 시도 비율, 골밑슛 시도 비율을 통해 해당 선수의 슈팅 유형을 파악할 수 있는 지표이다.

6) Effective Field Goal Percentage%

3점슛에 일정한 가중치를 부여한 야투 성공률 지표이다. 3점슛은 가장 많은 점수를 획득하지만 확률이 가장 낮은 공격방법이다. 3점슛 빈도가 높은 선수들은 야투 성공률에서 2점슛 빈도가 높은 선수보다 낮은 평가를 받기 때문에 3점슛 빈도가 많은 선수들을 위해 개발된 지표이다.

7) True Shooting Percentage%

자유투에 일정한 가중치를 부여한 슈팅 성공률 지표이다. 기존 야투 성공률은 필드골 시도로만 구했지만, 자유투도 포함하였기 때문에 경기장에서 발생하는 모든 슈팅의 능력을 평가할 수 있다.

8) Assist%

해당 선수가 경기를 뛰는 동안 발생하는 득점 중에서, 해당 선수의 어시스트로 만들어진 득점 비율로써 어시스트 능력을 판단하는 지표이다.

9) Offensive/Defensive Rating

Offensive Rating은 해당선수 수준 100possesion에서의 득점 추정값으로 종합적인 공격능력을 평가하고, Defensive Ratin은 해당선수 수준 100possession에서의 실점 추정값으로 종합적인 수비능력을 평가한다. 두 지표의 차이는 Net Rating이라는 지표로써 해당 선수 수준 100possesion에서 발생하는 득점과 실점의 차이이다.

10) Win Share

해당 선수가 뛴 경기에서 만들어낸 승리수를 추정한 값으로 공격과 수비와 팀 승리를 모두 반영한 선수 종합 평가 지표이다.

제 2절. 농구 포지션

농구에서 포지션은 전통적으로 5가지 종류가 있다. 그리고 각 5가지 포지션마다 주어진 역할과 플레이 특성이 존재한다. 그리고 SportsRec(2021)에서는 각 포지션의 특징을 다음과 같이 설명한다. 그리고 각 포지션은 주로 1~5번 으로 칭하기도 한다.

1) 포인트가드

포인트가드는 주로 공을 다루며, 공격권의 책임을 가지고 있는 포지션이다. 공을 만질 때 팀의 공격 패턴을 설정하고, 패스와 드리블을 하면서 팀원들을 패턴 속으로 참여시키도록 하는 역할을 한다. 대표적인 포인트 가드로는 과거 Magic Johnson, 현재 Stephen Curry, Chris Paul, Russell Westbrook 등이 있다.

2) 슈팅가드

슈팅가드는 주로 짧은 시간에 많은 득점을 할 수 있는 포지션이다. 패스보다는 슈팅을 많이 하는 포지션으로 득점에 가장 큰 기여를 해야 한다. 대표적인 슈팅 가드로는 과거 Michael Jordan, Kobe Bryant, 현재 Klay Thompson 등이 있다.

3) 스몰포워드

스몰포워드는 골밑으로 드라이브인 하는 드리블과 돌파 능력과 중장거리 슛 능력이 필요한 포지션이다. 대표적인 스몰포워드로는 과거 Larry Bird, 현재 Kevin Durant, Lebron James 등이 있다.

4) 파워포워드

파워포워드는 리바운드와 인사이드 득점을 주로 하는 키가 힘이 좋은 포지션이다. 파워포워드의 슛은 주로 12피트 안쪽에서 이루어지며, 리바운드 시에는 박스아웃을 담당한다. 대표적인 파워포워드로는 과거 Karl Malone, 현재 Draymond Green, Giannis Antetokounmpo 등이 있다.

5) 센터

센터는 코트에서 가장 큰 사람이 주로 담당한다. 큰 키와 윙스팬으로 공수에서 상대보다 신체조건에서 우위를 점할 수 있는 포지션이다. 공격시에는 주로 훅 슛, 피벗 플레이로 공격을 한다. 수비시에는 특히 블락으로 상대의 득점을 막는 역할을 한다. 대표적인 센터로는 Wilt Chamberlain, Shaquille O'Neal, 현재 Anthony Davis 등이 있다.

제 3절. 머신 러닝(Machine Learning)

3.1 머신 러닝의 개념

머신 러닝(Machine learning) 또는 기계 학습은 인공지능의 한 분야로, 사람이 명시적인 논리를 직접 지시하지 않아도 컴퓨터가 데이터를 통해 스스로 학습할 수 있도록 하는 알고리즘과 기술을 개발하는 분야를 의미한다. 또한, 주어진 데이터를 통해서 입력 변수(X)와 출력 변수(Y)간의 관계를 나타내는 함수 를 만들거나 데이터 속에서 데이터의 특징을 찾아내는 함수를 만드는 것을 의미한다.

3.2 머신 러닝의 종류

머신 러닝은 크게 지도학습과 비지도학습으로 구분된다. 지도 학습은 입력 변수(X)와 출력 변수(Y)의 관계에 대하여 모델링하는 것을 의미한다. 반면 비지도학습은 출력 변수(Y)가 존재하지 않고, 입력변수(X)간의 관계에 대해 모델링하는 것을 의미한다.

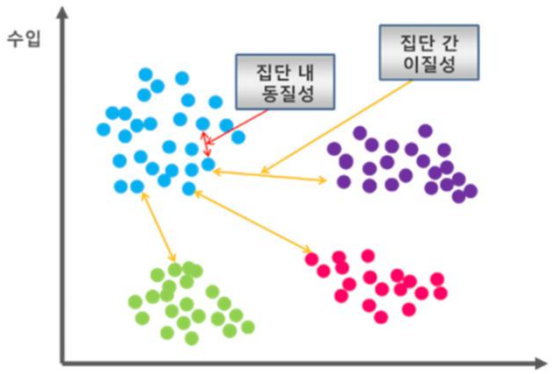
표 1. 머신러닝의 종류

구분		의미	알고리즘
머신 러닝	지도학습	학습데이터를 기반으로 새로 입력되는 데이터의 출력변수(Y)를 예측하거나 분류하는 기법	회귀분석
			의사결정나무
			랜덤 포레스트
			서포트 벡터 머신
			인공 신경망
	비지도학습	출력변수(Y)가 없는 데이터에서의 패턴과 특성을 파악하는 기법	**군집분석**
			연관 규칙 분석
			SOM (Self Organizing Map)

제 4절. 군집 분석(Clustering Analysis)

4.1 군집 분석의 개념

군집 분석은 기계 학습의 비지도 학습 기법으로 관측치에 대한 사전 지식이 없는 상황에서 전체 관측치를 여러 개의 집단으로 분류하는 것을 말한다. 군집 분석에 의해 두 개 이상의 집단으로 분류되며 분류된 각 집단을 군집(cluster)이라고 부른다. 군집 분석을 시행하면 같은 군집 내의 관측치들은 서로 동질적인 특성을 갖는 반면, 서로 다른 군집들 사이에는 이질적인 특성을 가지게 된다. 군집분석은 예측 또는 검증이 목적이 아닌 분류된 각 집단의 특성을 파악함으로써 데이터 전체의 구조에 대한 이해를 돕는 역할을 한다.



<그림 2> 군집 분석의 개념

4.2 군집 분석의 종류

4.2.1 계층적 군집분석

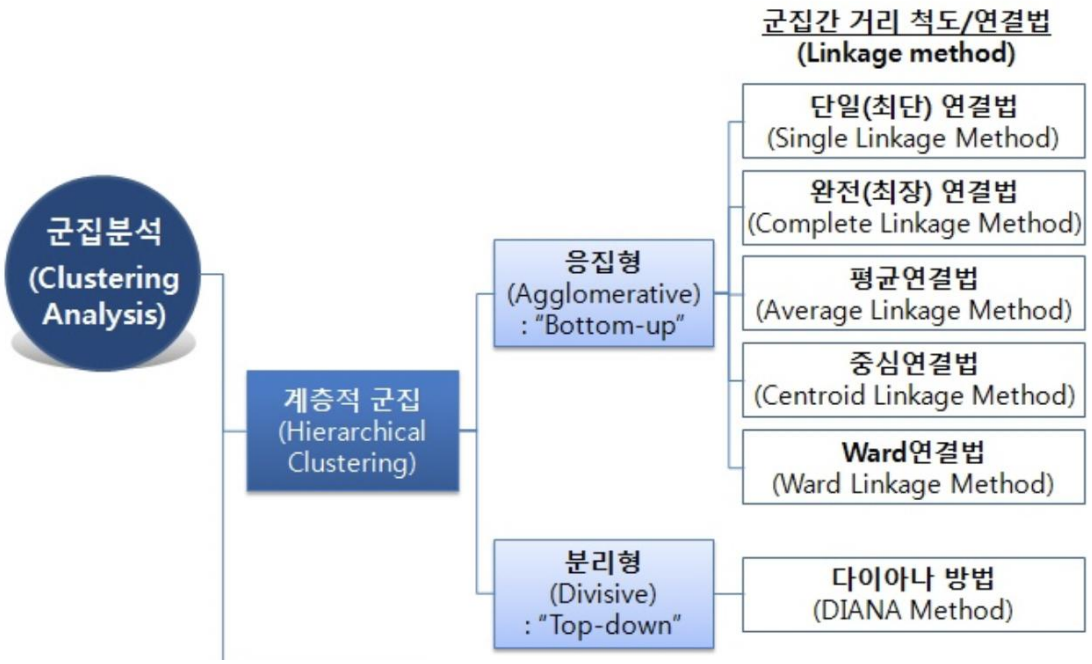
계층적 군집 분석은 모든 케이스가 각자 하나의 군집을 형성하면서 시작된다. 시작 단계에서는 케이스의 개수와 군집의 개수가 동일하다. 이후 군집 단계가 진행되면서 단계별로 유사한 군집 끼리 서로 합쳐지며 최종 단계에서는 하나의 군집만 남게 되는 구조이다.

계층적 군집 분석의 기본적인 원리는 최종 한 개의 군집이 남을 때 까지 매 단계마다 거리가 가까운 군집 끼리 합해가는 병할적 방법과 거리가 먼 군집들을 나누어 가는 분할적 방법이 있다.

병할적 방법은 매 단계마다 가장 짧은 거리를 갖는 두 군집이 결합하여 하나의 새로운 군집을 형성한다. 즉 두개의 군집이 사라지고 새로운 하나의 군집이 생성된다. 이처럼 각 단계가 진행될 때마다 군집 간의 최소 거리를 기준으로 개별 케이스들이 기존의 군집에 흡수되거나, 두 개의 케이스가 결합하여 새로운 군집을 만들거나, 기존의 두 군집이 결합하여 새로운 군집을 형성하거나 하는 등의 방식으로 군집의 개수가 줄어들게 된다. 분할적 방법은 하나의 군집으로 시작해서 거리가 먼 군집으로 분리해 나가는 방식이다.

계층적 군집 방법은 군집이 형성되는 과정을 정확하게 파악할 수 있다는 장점이 있지만, 데이터의 크기가 커지면 커질수록 소요 시간과 비용이 많이 든다는 단점이 있다.

군집 간의 거리는 일반적으로 다음과 같은 방법으로 측정한다.



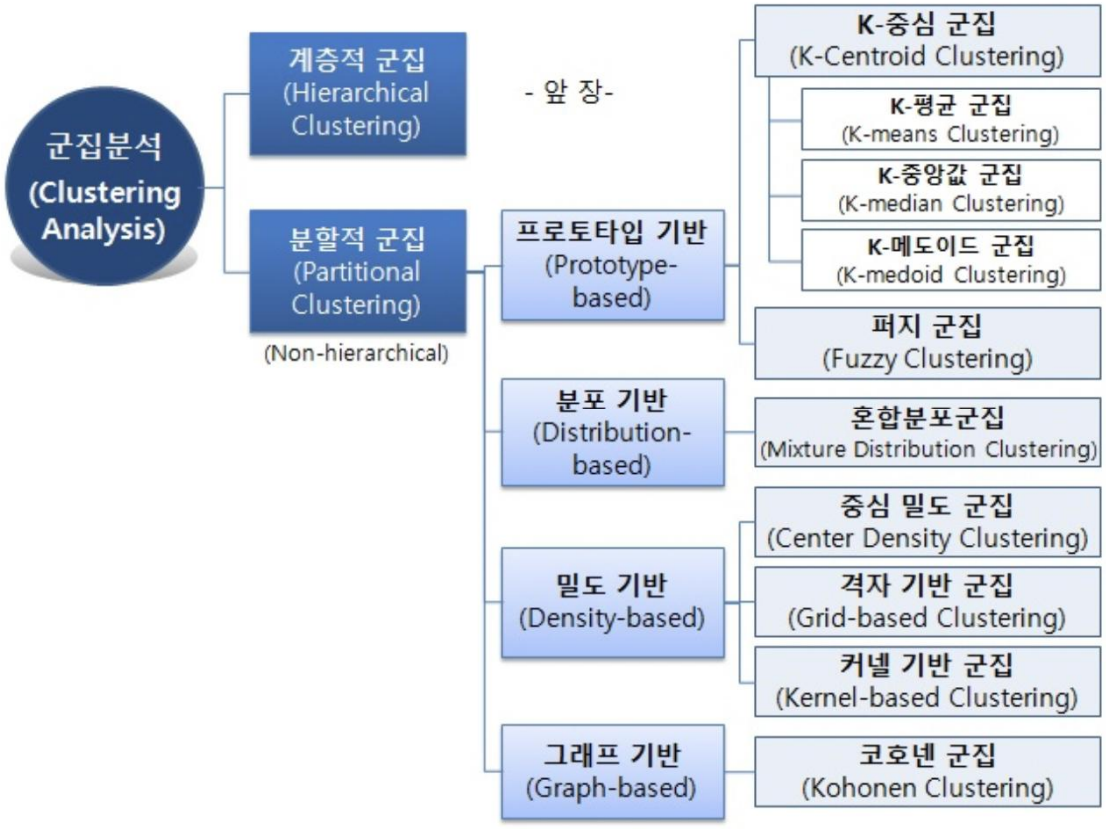
<그림 3> 계층적 군집분석의 종류

표 2. 계층적 군집분석 알고리즘

알고리즘	방법
단일 연결법	두 군집의 모든 개체 쌍의 거리 중 가장 가까운 거리를 사용
완전 연결법	두 군집의 모든 개체 쌍의 거리 중 가장 먼 거리를 사용
평균 연결법	두 군집의 모든 개체 쌍의 거리의 평균 거리를 사용
중심 연결법	두 군집의 중심(변수의 평균)간 거리를 사용
최소 분산 연결법	두 군집 내 모든 케이스 간의 총 분산을 거리로 사용하여 총 분산이 최소화되는 군집들을 결합

4.2.2 비계층적 군집분석

비계층적 군집분석은 계층적 군집분석과 달리 도출하고자 하는 군집의 개수를 사전에 결정해야 한다. 주어진 기준에 따라 모든 케이스를 사전에 정의된 개수만큼의 군집으로 할당하고 이후 가장 적합한 군집에 케이스를 재할당하는 과정을 반복 수행하여 최적화된 군집 분할을 찾아낸다.



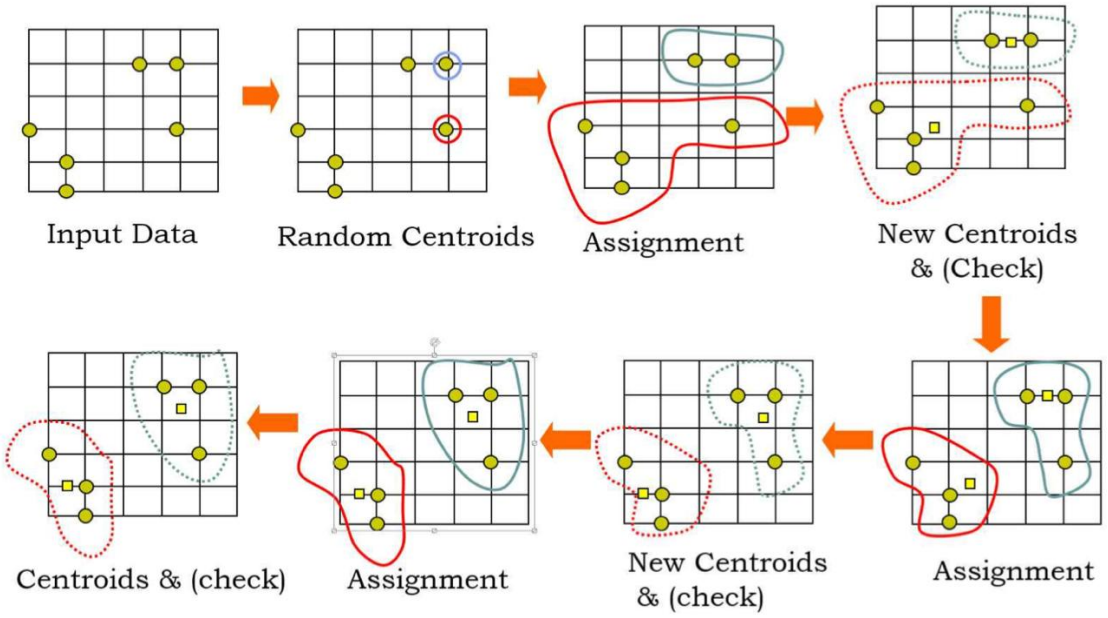
<그림 4> 비계층적 군집분석의 종류

제 5절. k-평균 군집분석(k-means clustering)

5.1 개념

비계층적 군집 분석의 대표적은 방법으로 초기에 정해지는 k개의 군집 수 만큼 군집을 생성하여 각 객체들을 분류하는 방식으로 이루어진다. 계층적 군집 분석과 마찬가지로 같은 군집 내 객체 간의 유사도는 증가하고, 다른 군집에 있는 객체와의 유사도는 감소하게 된다(서명교, 윤원영, 2017). k-평균 군집분석 알고리즘의 비용함수는 각 그룹의 중심값과 그룹 내의 객체들과의 거리의 제곱합으로 정의한다. 그리고 이 함수 값을 최소화하는 방향으로 각 객체의 군집을 재 할당하는 방법으로 군집분석을 수행하게 된다(안모하, 2019).

5.2 알고리즘



<그림 5> k-means 군집분석 알고리즘

표 3. k-means 알고리즘(전치혁, 2012)

1단계: (초기 군집 중심 선정) 어떤 규칙에 의하여 k개의 객체를 초기 군집의 중심 값으로 선정한다. **2단계:** (객체의 군집 배정) 각 객체에 대하여 k개의 군집 중심 값과의 거리를 산출 후 가장 가까운 군집에 각 객체를 배정한다. **3단계:** (군집 중심 좌표의 산출) 새로운 군집에 대한 중심 값을 산출한다. **4단계:** (수렴조건점검) 새로 산출된 중심 값과 이전 중심 값을 비교하여 수렴 조건 내에 들면 마치고, 그렇지 않으면 2~4단계를 반복한다.

5.3 선행 연구

k-평균 군집분석은 데이터에 대한 사전 정보가 필요하지 않으며, 사전에 특정 변수에 대한 역할 정의가 필요하지 않기 때문에 분석방법의 적용이 쉽다는 장점이 있다. 최근 빅데이터 마이닝(Big data mining) 기법 중 가장 대표적인 클러스터링 방법으로 인공지능, 패턴인식, 경제학, 생태학, 마케팅 등 여러 분야에서 활용 되고 있으며, 대용량 데이터에 가장 적합한 방법이다(신지은, 2016).

k-평균 군집분석은 다음과 같은 단점을 가지고 있고, 선행연구에서는 각 한계점에 대한 해결책을 제시한다.

1. 초기 중심값의 영향을 많이 받는다.

부적절한 초기값의 결정은 잘못된 군집의 생성과 군집 생성과정에서 많은 반복을 발생하며 군집 생성에 많은 시간이 소요되고, 또한 군집분석의 성능에 상당한 영향을 미치게 된다(Whasoo Bae, Se Won Rho, 2005). 이를 해결하기 위해서, Eblow Method 기법을 통해 이너셔(Inertia)라는 지표가 낮을수록 정확한 군집이라고 판단하였다. 이너셔는 각 데이터와 각 데이터에 할당된 클러스터의 중심 사이의 평균 제곱거리이다. 이외에도, MA 방법, KA 방법, Random 방법, FA 방법 등의 비교를 통해 KA 방법이 가장 적합한 결론을 내렸다(Pena, Lonzano, 1999). 여러가지 해결책이 존재하지만, 절대적인 기준이 없고 일반화된 이론은 존재하지 않는다(신성원, 2010).

2. 이상값에 민감하다.

k-평균 군집분석은 이상값에 민감하게 반응한다. 이상값은 알고리즘 내에서 중심 값을 갱신하는 과정에서 군집 내의 평균값을 크게 왜곡시킬 수 있다. 즉, 군집의 중심값이 군집의 실제 중심에 있지 않고 이상값 방향으로 치우치게 위치할 수 있다(안모하, 2019). 이상값 때문에 군집이 적절하게 형성되지 않을 경우에는 군집의 평균값이 아닌, 이상값에 덜 민감한 중앙값을 군집의 대푯값으로 설정한 k-중앙개체 군집방법의 PAM 알고리즘을 해결법으로 제시하기도 하였다(한수철 외2, 2008). 또한 분석 전 탐색적 자료분석을 통해 이상값을 사전에 미리 제거하는 방법도 해결책으로 제시되었다.

3. 서로 다른 밀도의 군집을 구분하는 성능이 좋지 않다.

k-평균 군집분석은 밀도가 다른 군집 간의 구분을 찾아내는 성능이 떨어진다. 이에 대한 해결책으로는 밀도기반 군집화인 DBSCAN(Density-Based Spatial Clustering of Applications with Noise)의 사용을 제시한다. DBSCAN은 적절한 두개의 파라미터 MinPts(minimum number of points)와 EPS(e)를 지정하며, 지정한 파라미터에 기반한 밀도 조건이 충족되는 모든 개체를 하나의 군집으로 판단하는 군집방식이다(황종필, 2015).

제 6절. 랜덤포레스트(Random Forest)

6.1 붓스트랩(Bootstrap)

붓스트랩은 가설 검증이나 메트릭을 계산하기 전에 랜덤 샘플링을 적용하는 방법이다. 붓스트랩은 전체 데이터에서 일부를 랜덤 샘플링해서 만든 여러 모델 간의 성능이 얼마나 일치하는가 판단하기 위해 사용된다.

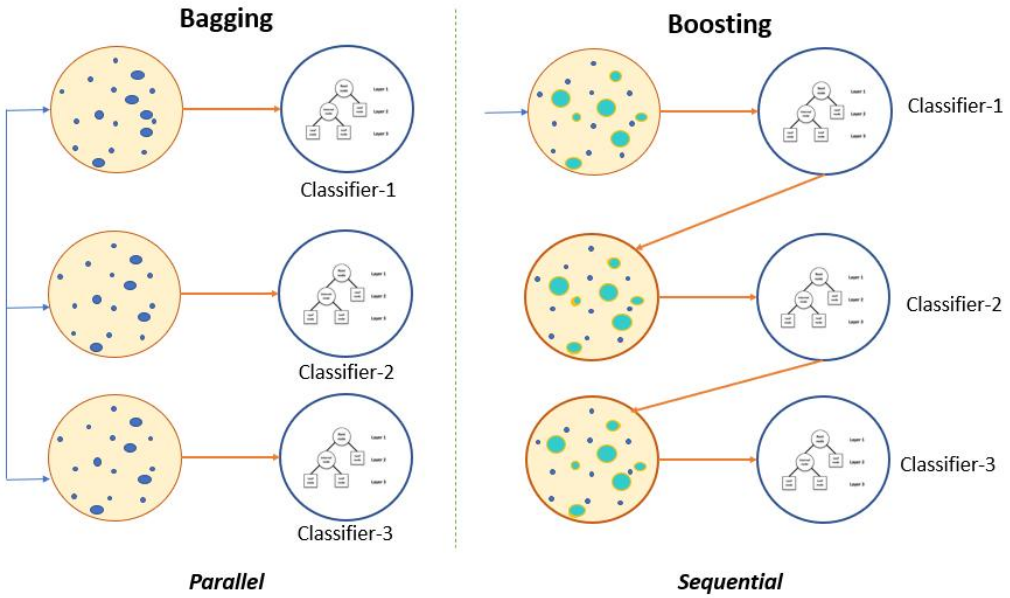
붓스트랩은 배깅(bagging)과 부스팅(boosting)으로 나뉜다. 각 용어의 설명은 다음과 같다.

6.2 배깅(Bagging)

배깅은 전체 데이터에서 랜덤 샘플링을 통해 크기가 동일한 n개의 표본 자료를 생성하고 각 표본 자료를 기준으로 병렬적으로 모델을 생성한다. 예측된 변수들을 결합하는 방법은 출력 변수가 연속형일 때는 평균(average), 범주형일 때는 보팅(voting)을 사용하여 가장 성능이 좋은 모델을 생성한다. 배깅은 예측 모형의 분산을 줄여 예측성능을 향상시키는 장점이 있다. 배깅을 이용한 대표적인 앙상블 모델은 랜덤포레스트가 있다.

6.3 부스팅(Boosting)

부스팅은 성능이 약한 모델들의 학습 에러에 가중치를 부여하여 순차적으로 다음 학습 모델에 가중치를 반영하여 강한 예측 모형을 만들어가는 알고리즘이다. 즉, 첫 번째 모델이 예측 성능이 약하더라도 이를 다음 예측 모델에 반영하여 다음 모델에 영향을 준다. 직렬, 순차적 과정으로 잘못 분류된 데이터에 집중하여 새로운 분류 규칙을 만드는 단계를 반복한다. 부스팅은 편향성을 감소시키는 장점이 있다. 부스팅을 이용한 대표적인 앙상블 모델은 XGBoost가 있다.



<그림 6> 배깅과 부스팅 과정

6.4 랜덤포레스트(Random Forest)의 장점

랜덤포레스트 알고리즘의 장점은 다음과 같다.

1. 분류(Classification)를 위한 랜덤포레스트 알고리즘은 과적합 문제를 개선한다.
2. 결측치와 이상치 존재에도 높은 정확도를 가진다.
3. 입력 변수를 분류하는 과정에서 분류 기준이 되는 변수의 중요도를 제공하며, 데이터 세트에서 가장 중요한 특징을 파악할 수 있다.

제 3장 연구 방법

제 1절. 연구 대상

KBL은 2016/17시즌부터 2020/21시즌까지 총 5년간 정규 기록 기준을 충족하는 총 364명을 대상으로 한다. KBL의 Advanced stat 기록은 시즌 600분 이상 출전해야 정규 기록으로 인정된다.

NBA는 2019/20시즌을 제외하고 2015/16시즌부터 2020/21시즌까지 총 5년간 정규 기록 기준을 충족하는 총 899명을 대상으로 한다. NBA의 Advanced stat 기록은 단일 시즌 1500분 이상 출전해야 정규 기록으로 인정된다.

NBA 2019/20시즌은 코로나 사태로 인해 시즌 중반 이후 토너먼트 형태로 리그 진행이 되어 팀별로 경기수가 다르기 때문에 연구 대상에서 제외하였다.

표 4. 각 리그와 시즌별 인원수

리그	시즌	인원수(명)
KBL	2020/21	76
	2019/20	65
	2018/19	72
	2017/18	75
	2016/17	76
	합계	364
NBA	2020/21	137
	2018/29	183
	2017/18	195
	2016/17	193
	2015/16	191
	합계	899

제 2절. 데이터 수집

본 연구에서 사용되는 데이터는 KBL·NBA 기록실, NBA Reference, Stats에서 제공하는 데이터를 수집하였다. 수집된 데이터에서 플레이 특성을 구분하는데 사용할 변수들을 2차 가공 및 추출하였다. 또한, 사용변수 선정 과정에서 프로농구 데이터분석 전문가의 의견을 참고하여 선정하였다.

표 5. 수집 데이터의 종류

사용데이터		비고
KBL	박스 스코어	시즌 누적 기록
	Advanced Stat	박스스코어 2차 가공
NBA	박스 스코어	시즌 누적 기록
	Advanced Stat	박스스코어 2차 가공
	Tracking Data	레이더기반 수집 데이터

표 6. KBL 사용변수 및 설명

리그	구분	사용변수	비고
KBL	Touches	Touches	공을 만진 추청 횟수
		PaintZone Touches	페인트존 내 Touches
		Shoot%	전체 Touches 중 슈팅 비율
		Pass%	전체 Touches 중 패스 비율
	Usage	Usage Rate	사용률
	Shooting	PaintZone Frequency	전체 야투시도에서 페인트존 야투시도 비율
		Middle Range Frequency	전체 야투 시도에서 미드레인지 야투시도 비율
		3Point Frequency	전체 야투 시도에서 3점슛 야투 시도 비율
	Rebound	Total Rebound%	경기를 뛰는 동안 리바운드를 잡을 확률
	Pace	Pace	경기 속도를 추정한 값

표 7. NBA 사용변수 및 설명

리그	구분	사용변수	비고
NBA	Touches	Touches	공을 만진 추청 횟수
		Front Court Touches	Front court에서의 Touches
		Elbow Touches	Elbow 지역에서의 Touches
		PaintZone Touches	PaintZone에서의 Touches
		Average Second per Touches	터치가 발생하는 평균 시간
		Average Dribble per Touches	터치가 발생하는 평균 드리블횟수
		Shoot%	전체 Touches 중 슈팅 비율
		Pass%	전체 Touches 중 패스 비율
	Usage	Usage Rate	사용률
	Shooting	Restricted Area Frequency	전체 야투 시도에서 Restricted 지역에서의 야투 시도 비율
		PaintZone Frequency	전체 야투 시도에서 PaintZone 지역에서의 야투 시도 비율
		Middle Range Frequency	전체 야투 시도에서 Middle range 지역에서의 야투 시도 비율
		3Point Frequency	전체 야투 시도에서 3점슛 야투 시도 비율
	Rebound	Rebound Chance	경기를 뛰는 동안 발생하는 리바운드 찬스
	Pace	Pace	경기 속도를 추정한 값
		Time of Possession	Possesion이 발생하는 시간

제 3절. 연구 방법

3.1 포지션 재정의 모델 개발

3.1.1 데이터 탐색

결측값(Missing Value), 이상값(Outlier)의 존재는 데이터를 학습하는 모델링 과정에서 특이성을 반영함으로써 모델의 성능을 저하시킬 가능성이 있다. 따라서 데이터 탐색 과정을 통해서 이를 처리하여 데이터 정제한다.

1) 결측값(Missing Value) 처리

결측값(Missing Value)은 값이 입력되지 않았거나 관측값이 없는 상태를 의미한다. 결측값의 존재는 분석 결과에 편향성을 가져오며, 결측값으로 인한 문제 발생을 식별하기 어렵다.

결측값 대체 방법으로는 삭제, 대푯값(평균, 중앙값, 최빈값)으로 대체, 예측값으로 대체 등이 있다.

2) 이상값(Outlier) 처리

이상값(Outlier)은 다른 관측치의 분포에서 벗어나는 특이성을 가진 관측값으로 모델의 성능을 왜곡할 가능성이 있다.

대표적인 이상값 판단 기준은 사분위 범위(IQR)를 넘어서는 값들을 이상치로 판단한다. 결측값과 마찬가지로 이상값 대체 방법으로는 삭제, 대푯값으로 대체, 예측값으로 대체하는 방법 등이 있다.

3.1.2 데이터 변환

수집된 데이터는 각 변수들마다 단위와 범위가 다양하다. 군집분석은 관측치 간의 거리를 이용한 알고리즘이기 때문에 이상값에 민감하다. 특히 본 연구에 사용된 k-means 군집분석 알고리즘은 군집중심이 이상값에 민감한 평균값으로 설정이 되기 때문에 데이터의 단위를 일정하게 맞추는 스케일링(scailing) 과정을 통한 정규화 작업을 선행해야 한다.

본 연구에서 사용한 정규화 과정은 다음과 같다.

- 로그(Log) 변환 : 각 변수들에 로그를 취함으로써 큰 숫자를 같은 비율의 작은 숫자로 변환하는 방법으로, 첨도와 왜도를 줄여준다.
- 최대-최소 정규화(min-max normalization) : 각 변수 값을 0과 1사이의 값으로 변환하는 방법으로 모든 변수의 단위를 일치시킨다.

3.1.3 포지션 분류 모델 설계

1) 군집분석 알고리즘

이 연구에서는 군집화는 대표적인 비계층적 군집분석 방법인 k-평균 군집분석 방법으로 실행한다.

2) 최적 k값 설정

k-means 알고리즘의 가장 중요한 단계는 초기 k값을 과정이다. k값은 임의의 자연수이며 분석하는 분야의 도메인 전문성과 반복적 실험 등을 통해서 적절한 k값을 설정해야한다. 이 연구에서는 엘보우(Elbow)기법과 실루엣(shilluete)기법 두가지를 관찰하여 적절한 k값을 설정하였다.

3) 포지션 재정의

군집분석을 통해 형성된 각 군집들의 기술통계량을 통해 각 군집의 특성을 파악한다. 프로농구 선수, 코치와 농구 데이터 분석가의 의견을 바탕으로 군집의 특성에 맞는 새로운 이름을 포지션으로 재정의한다. 또한, Bonferroni 검정을 통해서 재정의된 군집 간에 어떤 관계가 통계적으로 유의미한 차이가 있는지 검정한다.

4) 랜덤포레스트(Random Forest)

본 연구는 배깅을 이용한 앙상블 기법중 하나인 랜덤포레스트 알고리즘을 분류 모델로 사용한다. 랜덤포레스트 분류기는 예측 성능이 우수하다는 장점이 있지만, 훈련 데이터에 대해 과적합(overfitting)이 되어있다는 단점이 있다. 따라서 적절한 하이퍼 파라미터 튜닝(hyperparameter tunning)을 해야한다.

이 연구에서는 하이퍼 파라미터 튜닝(hyperparameter tunning) 방법으로 하이퍼 파라미터 범위 내에서 무작위 샘플링 방법인 Random Search 방법을 선택하였다. Random Search 방법은 반복적으로 무작위 추출된 하이퍼 파라미터에 대해 가장 좋은 Cross Validation 성능을 가진 하이퍼 파라미터를 선택하는 방식이다.

5) 모델 성능 평가 지표

본 연구에서 랜덤포레스트 모델 분류 성능을 평가하기 위해 F1-score를 활용한다. F1-score는 재현율(Recall)과 정밀도(Precision)의 조화평균을 의미한다.

제 4절. 연구 절차

본 연구의 절차는 다음과 같다.

표 8. 연구단계 및 연구내용

연구단계	연구내용
연구 계획 수립	- 연구 주제 선정 - 선행 연구, 문헌 연구 - 분석 방법 및 도구 선정
데이터 수집 및 탐색	- 데이터 수집 - 변수 선정 - 데이터 정제 및 탐색
데이터 분석	- 군집분석 구현 - 포지션 재정의 - 랜덤포레스트 분류 모델 학습 - 모델 성능 개선 및 평가 - 분석 결과 검토
논문 작성	- 최종 연구 논문 작성

제 4장 연구 결과

제 1절. 데이터 탐색

1.1 기술통계량

수집된 데이터의 각 변수들의 타입과 기술통계량은 다음과 같다.

표 9. KBL 변수 타입

리그	변수명	타입	평균	표준편차
KBL (N=364)	Touches	연속형	1155.5	623.32
	PaintZone Touches	연속형	1.25	0.48
	Shoot%	연속형	0.35	0.11
	Pass%	연속형	0.47	0.14
	Usage Rate	연속형	20.19	7.2
	PaintZone Frequency	연속형	0.46	0.19
	Middle Range Frequency	연속형	0.17	0.09
	3Point Frequency	연속형	0.36	0.22
	Total Rebound%	연속형	10.07	5.53
	Pace	연속형	40.68	2.49

표 10. NBA 변수 타입

리그	사용변수	타입	평균	표준편차
NBA (N=899)	Touches	연속형	51.79	17.12
	Front Court Touches	연속형	28.46	8.08
	Elbow Touches	연속형	1.59	1.63
	PaintZone Touches	연속형	2.71	2.65
	Average Second per Touches	연속형	2.88	1.28
	Average Dribble per Touches	연속형	2.09	1.56
	Shoot%	연속형	0.46	0.11
	Pass%	연속형	0.31	0.11
	Usage Rate	연속형	20.44	5.42
	Restricted Area Frequency	연속형	0.32	0.16
	PaintZone Frequency	연속형	0.15	0.07
	Middle Range Frequency	연속형	0.19	0.10
	3Point Frequency	연속형	0.34	0.19
	Rebound Chance	연속형	51.96	6.79
	Pace	연속형	99.2	2.72
	Time of Possession	연속형	2.71	1.96

1.2 데이터 정제

수집된 데이터의 결측값은 존재하지 않았고, 발견된 이상값은 각 변수들의 평균값으로 대체하였다.

표 11. KBL 변수별 결측값 및 이상값

리그	변수명	결측값 개수	이상값 개수
KBL (N=364)	Touches	0	3
	PaintZone Touches	0	6
	Shoot%	0	0
	Pass%	0	0
	Usage Rate	0	6
	PaintZone Frequency	0	0
	Middle Range Frequency	0	0
	3Point Frequency	0	0
	Total Rebound%	0	13
	Pace	0	5

표 12. NBA 변수별 결측값 및 이상값

리그	사용변수	결측값 개수	이상값 개수
NBA (N=899)	Touches	0	1
	Front Court Touches	0	23
	Elbow Touches	0	63
	PaintZone Touches	0	52
	Average Second per Touches	0	0
	Average Dribble per Touches	0	0
	Shoot%	0	15
	Pass%	0	9
	Usage Rate	0	6
	Restricted Area Frequency	0	46
	PaintZone Frequency	0	5
	Middle Range Frequency	0	4
	3Point Frequency	0	1
	Rebound Chance	0	2
	Pace	0	5
	Time of Possession	0	25

1.3 데이터 변환

1.3.1 로그변환

로그 변환을 통해 각 변수들의 왜도와 첨도를 줄인다.

1.3.2 최대-최소 정규화(Min-Max Normalization)

단위가 다른 각 변수들을 0과 1사이의 값으로 단위를 변환시킨다.

표 13. KBL 정규화 결과

리그	변수명	최소	평균	최대	표준편차
KBL (N=364)	Touches	0	0.29	1	0.2
	PaintZone Touches	0	0.29	1	0.19
	Shoot%	0	0.48	1	0.2
	Pass%	0	0.5	1	0.22
	Usage Rate	0	0.32	1	0.21
	PaintZone Frequency	0	0.44	1	0.21
	Middle Range Frequency	0	0.29	1	0.17
	3Point Frequency	0	0.43	1	0.26
	Total Rebound%	0	0.31	1	0.24
	Pace	0	0.32	1	0.18

표 14. NBA 정규화 결과

리그	사용변수	최소	평균	최대	표준편차
NBA (N=899)	Touches	0	0.42	1	0.2
	Front Court Touches	0	0.35	1	0.17
	Elbow Touches	0	0.13	1	0.14
	PaintZone Touches	0	0.2	1	0.2
	Average Second per Touches	0	0.33	1	0.25
	Average Dribble per Touches	0	0.29	1	0.24
	Shoot%	0	0.46	1	0.17
	Pass%	0	0.49	1	0.12
	Usage Rate	0	0.36	1	0.17
	Restricted Area Frequency	0	0.33	1	0.18
	PaintZone Frequency	0	0.35	1	0.19
	Middle Range Frequency	0	0.35	1	0.2
	3Point Frequency	0	0.39	1	0.22
	Rebound Chance	0	0.53	1	0.16
	Pace	0	0.49	1	0.17
	Time of Possession	0	0.25	1	0.22

제 2절. 군집 모델 개발

2.1 k값 설정

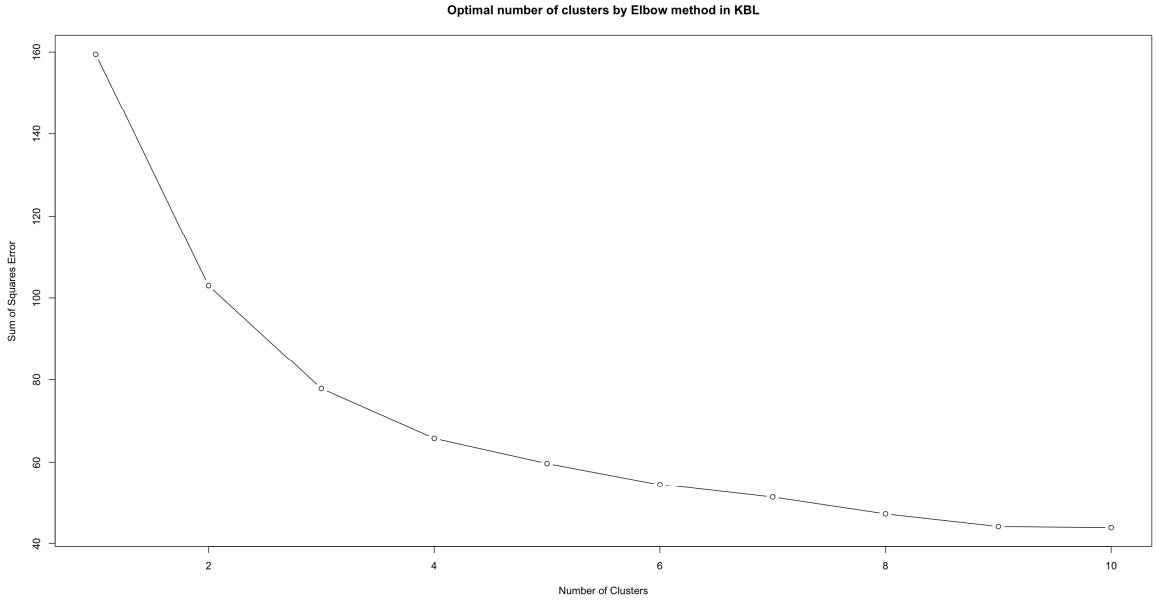
k-means 군집분석을 실행하기 전 최적의 k값을 설정한다. k값을 결정하는 방법은 엘보우 기법과 실루엣 기법을 참고하였다. 엘보우 기법과 실루엣 기법을 구현한 연구 환경은 다음과 같다.

표 15. 군집분석 연구환경

언어	기법	라이브러리	함수
R	엘보우기법	-	kmans()
R	실루엣기법	clustMixType, factoextra	fviz_nbclust()

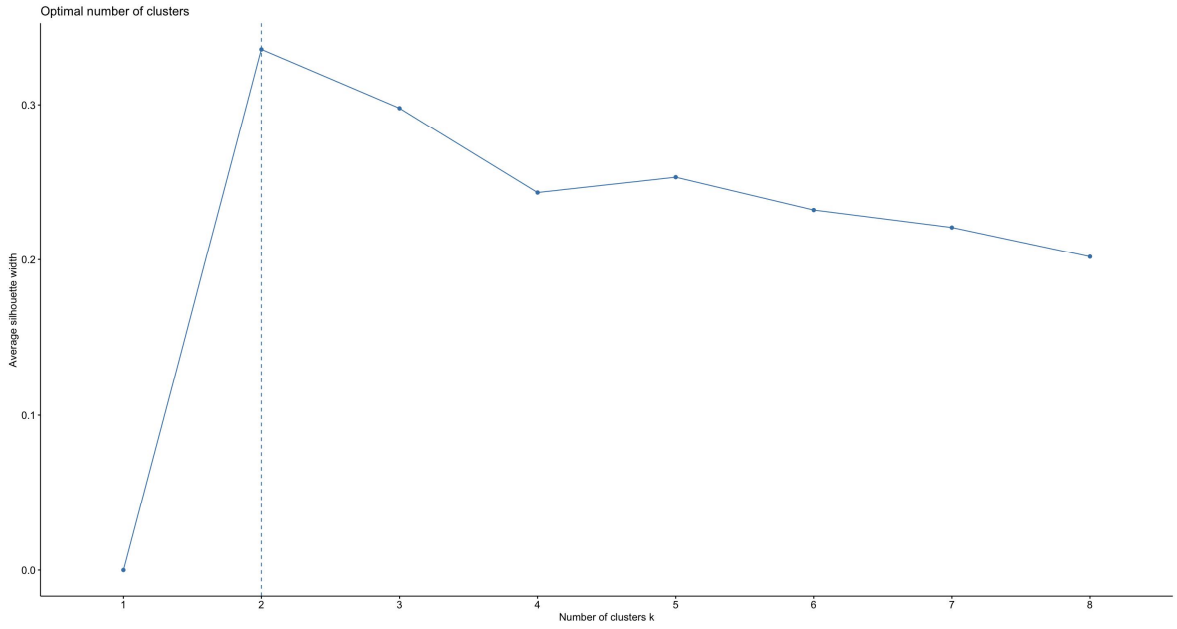
엘보우 기법은 각 군집의 중심과의 거리의 합을 계산하였을 때, 더 이상 큰 변화가 없는 곳을 최적의 k값으로 추천하는 방법이다. 실루엣 기법은 엘보우 기법의 한계점을 보완하여 군집내의 거리의 합 뿐만아니라 군집간의 거리의 합도 고려한 방법이다. 최적의 k값을 구하는 방법은 다양하고, 각 기법마다 k값의 결과도 다른 것을 확인 할 수 있다. 따라서 각 기법에서 나온 결과를 기반으로 적절한 해석이 필요하다. 본 연구에서 군집분석에 사용할 KBL과 NBA 데이터의 각 기법의 결과는 다음과 같다.

2.1.1 KBL의 군집개수



<그림 7> KBL 엘보우 기법

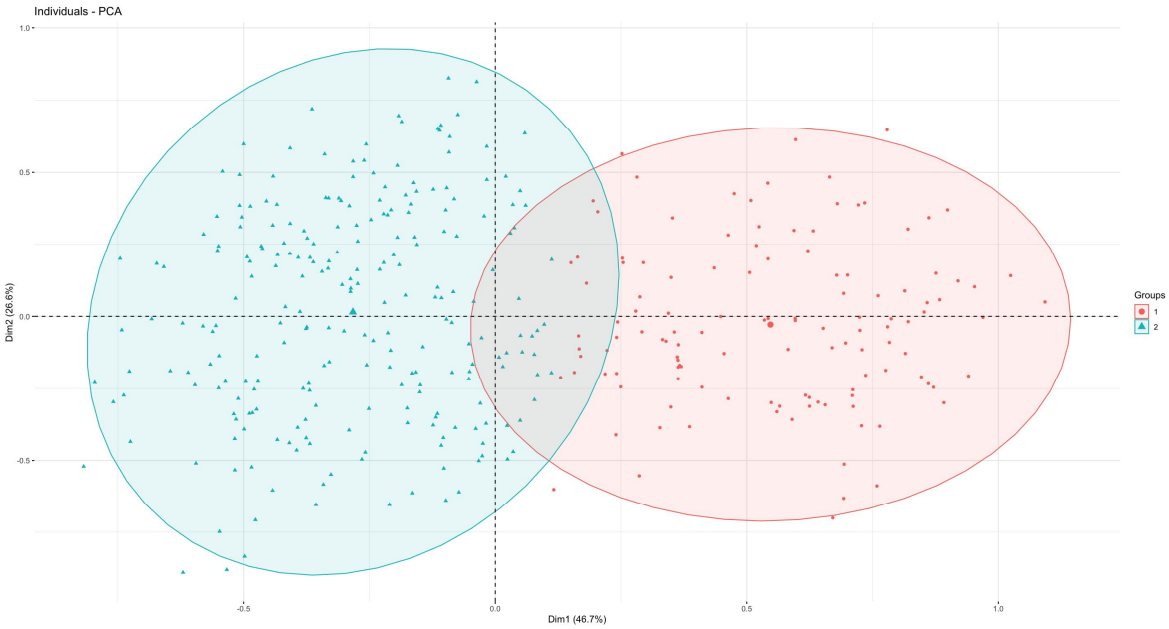
엘보우 기법을 통해서 KBL 데이터의 최적의 k값은 3과 5사이에서 결정하는 것이 적절하다고 판단된다.



<그림 8> KBL 실루엣 기법

실루엣 기법을 통해서는 k값이 2일 때가 가장 적절한 군집이라고 판단된다.

1) k=2 일때



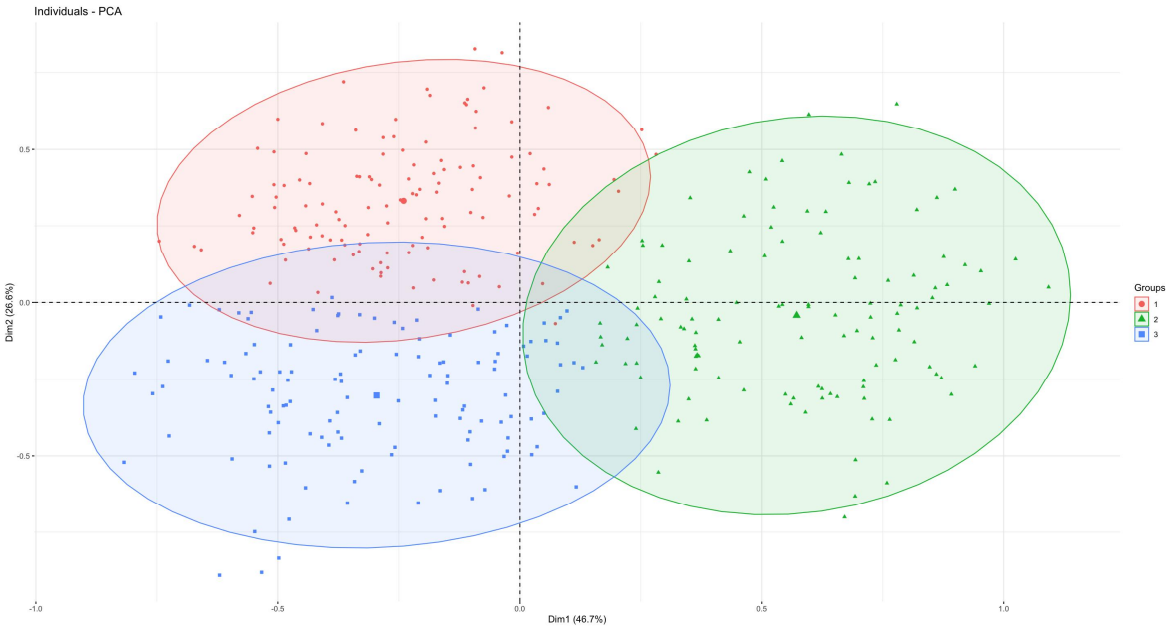
<그림 9> KBL 군집 시각화 (k=2)

표 16. KBL 군집 결과 (k=2)

군집	개체수	국내선수(명)	외국선수(명)
1	124	27	97
2	240	240	0

k값이 2일 때 외국 선수들은 모두 같은 군집으로 형성된 것을 확인할 수 있다. 위의 결과를 통해 최근 5년간 KBL 리그는 국내선수와 외국선수로 구분되는 모습을 확인할 수 있다. k값이 2일때는 외국 선수와 국내선수 내의 집단을 구분하지 못했기 때문에 과소군집화로 판단되었다.

2) k=3 일때



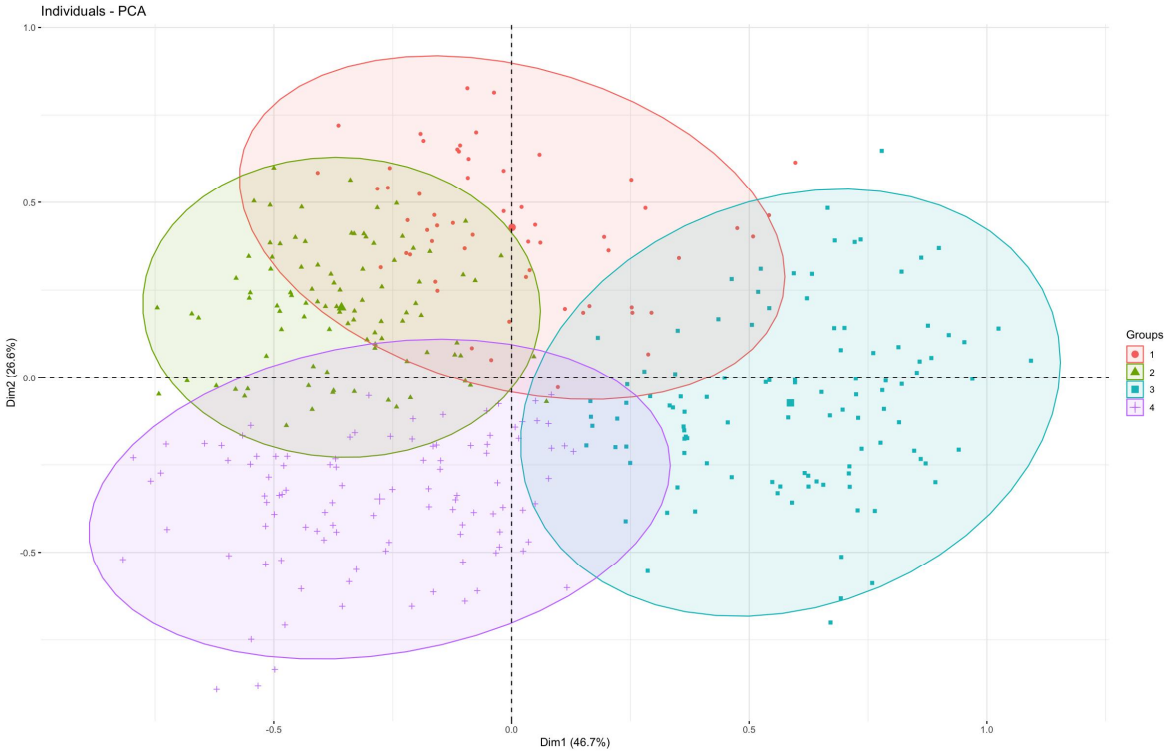
<그림 10> KBL 군집 시각화 (k=3)

표 17. KBL 군집 결과 (k=3)

군집	개체수	국내선수(명)	외국선수(명)
1	126	126	0
2	116	19	97
3	122	122	0

k값이 3일 때, 국내 선수 집단이 두 집단으로 분류되었다. 하지만, 외국 선수들은 기존에 포함된 군집에서 다른 군집으로 이동하지 않았기 때문에 과소군집화로 판단하였다. 위 그림에서 2번군집 내 좌, 우측으로 군집이 분류될 것을 생각하여 다음 k값을 4로 수정하였다.

3) k=4 일 때



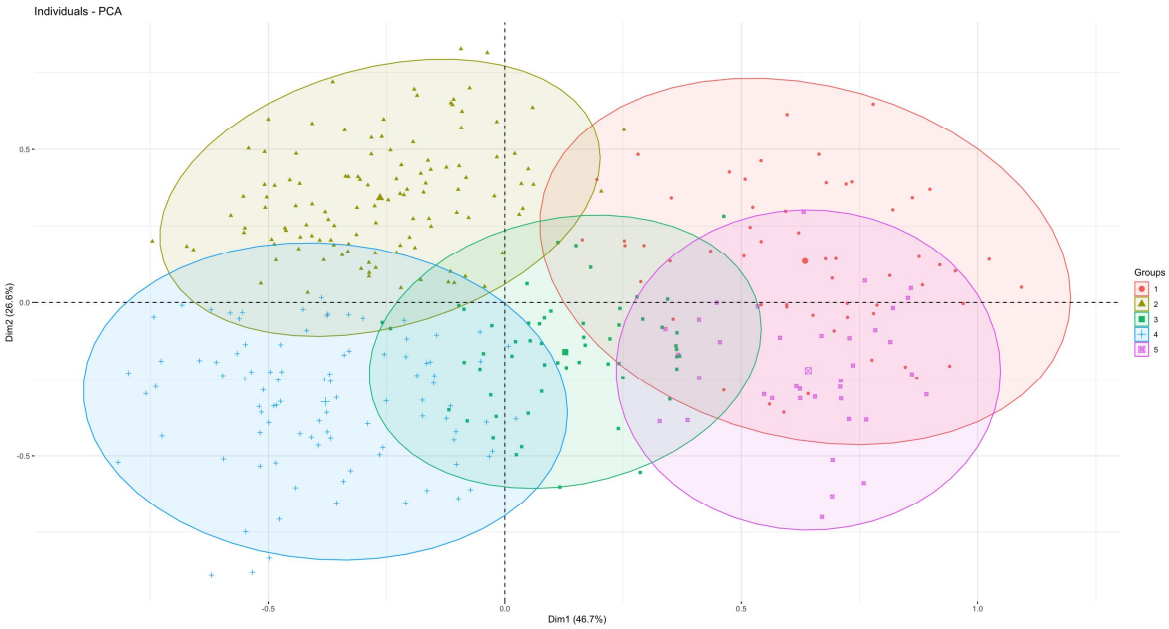
<그림 11> KBL 군집 시각화 (k=4)

k값이 4 일 때, 외국 선수들이 분리되었고 일부 국내 선수들과 같은 그룹으로 형성된 것을 확인할 수 있다. 하지만 여전히 국내 선수들의 100명이 넘는 두 집단과 외국 선수들을 분류하기에는 소군집화로 판단하였다.

표 18. KBL 군집 결과 (k=4)

군집	개체수	국내선수(명)	외국선수(명)
1	58	44	14
2	95	95	0
3	107	24	83
4	104	104	0

3) k=5 일 때



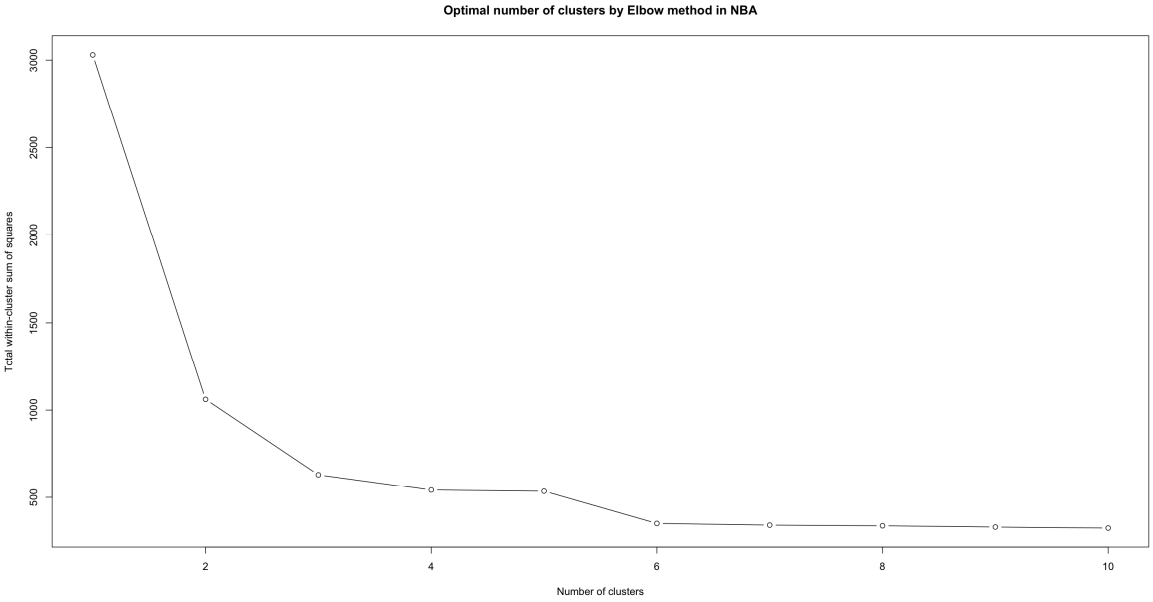
<그림 12> KBL 군집 시각화 (k=5)

k값이 5 일 때부터 더 이상 외국 선수들의 집단은 새로운 군집을 형성하지 않는 것을 확인 할 수 있다. 따라서 k값이 5이상 일때는 과군집화라고 판단하였다. KBL 데이터에서는 실루엣 기법을 통해 대분류를 하였고, 엘보우 기법을 통해 소분류 개수를 설정하였다. 이에 따라 KBL의 k값은 4로 정하였다.

표 19. KBL 군집 결과 (k=5)

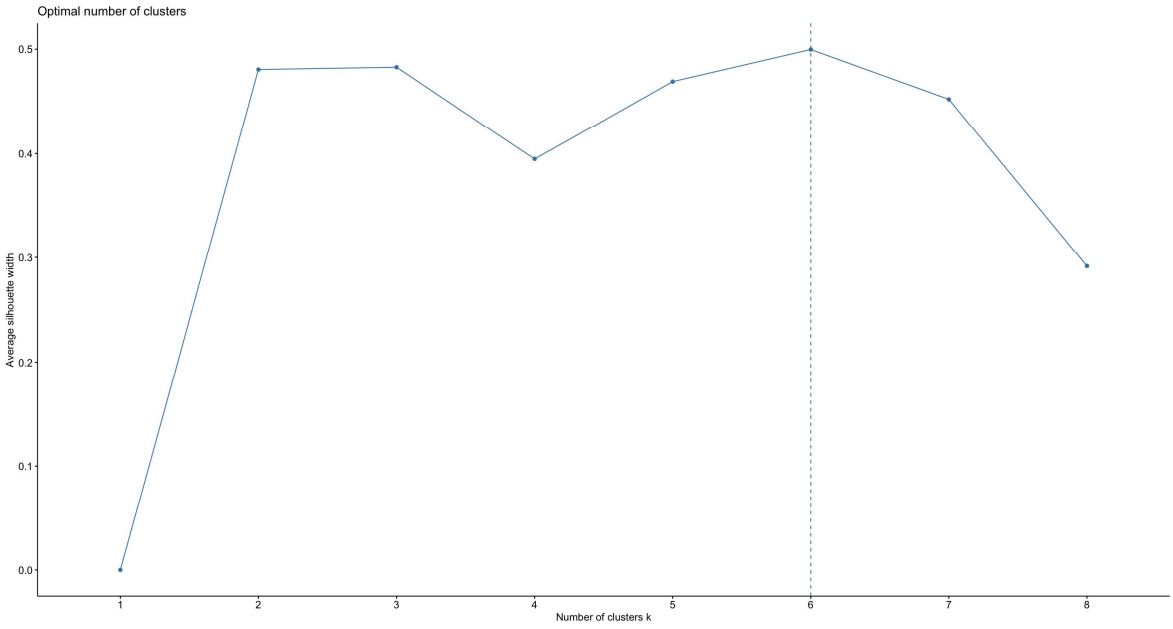
군집	개체수	국내선수(명)	외국선수(명)
1	82	1	81
2	117	114	3
3	59	58	1
4	62	62	0
5	44	32	12

2.1.2 NBA의 군집개수



<그림 13> NBA 엘보우 기법

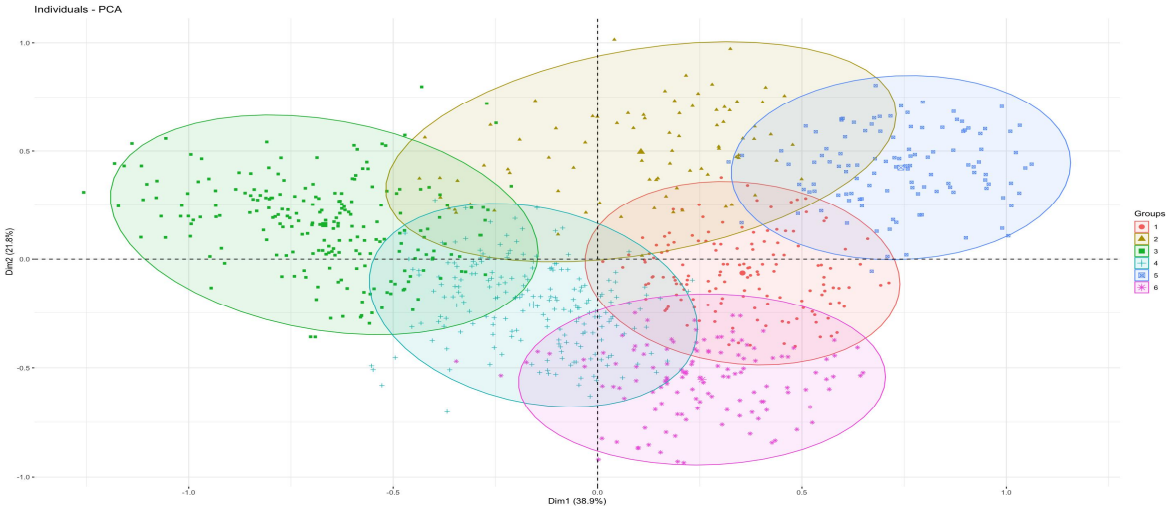
엘보우 기법 결과 k값은 3이상 6이하로 설정하는 것이 적절하다고 판단된다.



<그림 14> NBA 실루엣 기법

실루엣 기법 결과 k값이 6일때가 최적의 값이라고 판단된다.

엘보우 기법을 통해서 오차제곱합(SSE)의 감소하는 추세가 더 이상 보이지 않는 지점은 k값이 6일때이고, 실루엣 기법을 통해 실루엣 점수가 가장 높은 k값이 6인 것을 확인 할 수 있다. 외국 선수에게 공격 의존도가 높은 KBL리그와 달리 NBA는 다양한 플레이 특성을 가지고 기량이 좋은 선수들이 존재하는 리그이다. 이러한 리그 특성을 고려했을 때에도 KBL 리그보다 많은 6개의 군집으로 분류하는 것은 타당하다고 판단된다. k값이 6일 때, 형성된 군집분포와 관측된 개체수는 다음과 같다.



<그림 15> NBA 군집 시각화(k=6)

표 20. NBA 군집분석 결과(k=6)

군집	관측치 개수
1	145
2	82
3	220
4	206
5	116
6	130

제 3절. 포지션 정의

포지션 단어 선정은 프로농구선수, 코치, 농구 데이터 분석가 등 전문가들의 의견을 수렴하여 정의하였다. 또한, 각 리그별 Bonferroni 검정[1]을 통하여, 모든 변수에서 군집 간에 유의수준 0.05에서 통계적으로 유의미한 차이가 있다고 검증하였다.

3.1 kbl 군집특성

<표 21>부터 <표 24>는 k=4일 때 각 군집의 변수들의 기술 통계량이다.

1) 군집1 (58명)

표 21. KBL 군집1 특성

군집	변수명	최소	평균	최대	표준편차	구분
1 (N=58)	Touches	0.22	0.57	0.97	0.15	High
	PaintZone Touches	0	0.1	0.32	0.06	Low
	Shoot%	0.04	0.32	0.65	0.14	Low
	Pass%	0.32	0.67	0.97	0.15	High
	Usage Rate	0.09	0.42	0.93	0.17	Low
	PaintZone Frequency	0.2	0.41	0.65	0.11	Low
	Middle Range Frequency	0.08	0.32	0.87	0.14	Low
	3Point Frequency	0.15	0.45	0.68	0.14	Low
	Total Rebound%	0.05	0.45	0.68	0.14	Low
	Pace	0.22	0.46	0.94	0.13	Low

1) 부록 참조

군집1은 4가지 군집중에서 Touches가 가장 많고, 그 중 패스로 Touches를 마무리하는 빈도가 가장 많다는 대표적인 특징이 있다. 3번 군집 다음으로 Usage Rate가 가장 높은 것으로 보아 패스 뿐만아니라 득점으로 포제션을 마무리하는 역할도 담당하고 있다는 것을 알 수 있다. PaintZone Touches 보다는 주로 Back Court에서 플레이가 이루어지며 슈팅 타입은 3점슛, 페인트존, 미들슛 순서로 빈도가 많다. 군집1의 대표적인 선수로는 허훈(2021), 허웅(2021), 송교창(2019), 함지훈(2018) 등이 있다. 주로 팀 내에서 공을 가장 오래 소유하며, 패싱 능력이 좋고 3점슛과 페인트존 득점력도 갖추고 있다는 특징이 있다.

이 연구에서는 이러한 특징을 고려하여 군집1을 Main Ball Handler라고 정의하였다.

2) 군집2 (95명)

표 22. KBL 군집2 특성

군집	변수명	최소	평균	최대	표준편차	구분
2 (N=95)	Touches	0.06	0.25	0.47	0.1	Low
	PaintZone Touches	0.05	0.23	0.51	0.1	Low
	Shoot%	0	0.27	0.5	0.11	Low
	Pass%	0.5	0.73	1	0.11	High
	Usage Rate	0	0.18	0.47	0.11	Low
	PaintZone Frequency	0.09	0.32	0.62	0.12	Low
	Middle Range Frequency	0.04	0.27	0.52	0.13	Low
	3Point Frequency	0.27	0.57	0.78	0.15	High
	Total Rebound%	0.02	0.15	0.5	0.1	Low
	Pace	0	0.15	0.5	0.1	Low

군집2는 Touches와 Usage Rate가 모두 낮은 것으로 보아 공을 오래 소유하지도 않고, 포제션을 마무리하는 역할을 담당하는 유형은 아니다. 공을 소유할 때에는 슈팅보다는 주로 패스위주의 플레이를 하는 선수이고, 슈팅 타입은 3점슛을 주로 던지는 특징이 있다. 군집2에 속한 대표적인 선수로는 김영환(2021), 오재현(2021), 양우섭(2021), 정창영(2021)등이 있다. 공격 비중이 높지는 않고, 자신의 공격보다는 BackCourt에서 같은 팀의 공격을 위해 패스를 하며 연결고리 역할을 하는 특징을 가지고 있다.

이 연구에서는 이러한 특성을 고려하여 군집2를 Linker라고 정의하였다.

3) 군집3 (107명)

표 23. KBL 군집3 특성

군집	변수명	최소	평균	최대	표준편차	구분
3 (N=107)	Touches	0.03	0.35	0.5	0.19	Low
	PaintZone Touches	0.03	0.22	0.61	0.12	Low
	Shoot%	0.31	0.58	0.91	0.13	High
	Pass%	0	0.33	0.63	0.15	Low
	Usage Rate	0.13	0.52	1	0.19	High
	PaintZone Frequency	0.43	0.69	1	0.13	High
	Middle Range Frequency	0	0.34	0.81	0.19	Low
	3Point Frequency	0	0.13	0.5	0.13	Low
	Total Rebound%	0.25	0.62	1	0.2	High
	Pace	0.01	0.44	0.97	0.2	Low

군집3은 Usage Rate가 가장 높으며 Main Ball Handler 다음으로 Touches가 많다. Touches를 주로 슈팅으로 마무리하는 특징과 함께 팀내에서 공격을 마무리하는 역할을 담당하는 특징이 있다. 슈팅 타입은 페인트존 빈도가 가장 많으며 3점슛 빈도가 가장 적다. 또한 4가지 군집중에서 리바운드 창출 능력이 가장 좋다는 특징이 있다. 군집3에 속한 대표적인 외국 선수로는 자밀 워니(2021), 숀롱(2021), 국내 선수로는 양홍석(2021), 송교창(2021), 오세근(2021)등이 있다. 군집1에서 단신용병 선수들을 제외한 나머지 외국인 선수들은 모두 군집3에 속하였다. 이를 통해, 국내에서 뛰는 외국 선수들의 유형은 다양하지 않다는 것을 확인할 수 있다. 이 연구에서는 이러한 특성을 고려하여 군집3을 Finisher라고 정의하였다.

4) 군집4 (104명)

표 24. KBL 군집4 특성

군집	변수명	최소	평균	최대	표준편차	구분
4 (N=104)	Touches	0	0.13	0.37	0.08	Low
	PaintZone Touches	0.25	0.51	1	0.15	High
	Shoot%	0.41	0.64	1	0.13	High
	Pass%	0.09	0.39	0.66	0.12	Low
	Usage Rate	0	0.19	0.42	0.09	Low
	PaintZone Frequency	0	0.31	0.61	0.15	Low
	Middle Range Frequency	0.04	0.24	0.63	0.14	Low
	3Point Frequency	0.04	0.61	1	0.2	High
	Total Rebound%	0	0.2	0.47	0.11	Low
	Pace	0.04	0.22	0.55	0.1	Low

군집4는 Touches와 Usage Rate가 낮은 것으로 보아 공을 많이 소유하지 않고, 포제션을 마무리하는 역할을 담당하는 유형은 아니다. 또한, 3점슛 빈도가 높다는 것으로 보아 Linker와 유사한 특성이 있지만, 군집4는 공을 잡았을 때 패스보다는 슈팅을 하는 빈도가 더 많다는 점에서 차이가 있다. PaintZone Touches가 가장 높은 군집임에도 불구하고 슈팅 타입은 3점슛 빈도가 가장 많다. 슈팅은 주로 3점을 던지지만, 슛을 제외한 플레이는 주로 Front Court 플레이를 하는 선수임을 알 수 있다. 군집4의 대표적인 선수로는 허일영(2021), 박준영(2021), 안영준(2021), 문성곤(2021) 등이 있다. 3점슛을 제외한 플레이가 주로 PaintZone에서 이루어지는 선수들이다.

이 연구에서는 이러한 특성을 고려하여 군집4를 3&D(3Point & Defense)로 정의하였다.

3.2 NBA 군집특성

<표 25>부터 <표 30>는 k=6일 때 각 군집의 변수들의 기술 통계량이다.

1) 군집1 (145명)

표 25. NBA 군집1 특성

군집	변수명	최소	평균	최대	표준편차	구분
1 (N=145)	Touches	0.08	0.33	0.65	0.16	Low
	Front Court Touches	0.08	0.31	0.59	0.12	Low
	Elbow Touches	0.02	0.13	0.35	0.07	Low
	PaintZone Touches	0.05	0.21	0.49	0.09	Low
	Average Second per Touches	0	0.14	0.35	0.08	Low
	Average Dribble per Touches	0.01	0.12	0.31	0.07	Low
	Shoot%	0.38	0.55	0.78	0.09	High
	Pass%	0.11	0.36	0.61	0.11	Low
	Usage Rate	0.06	0.31	0.55	0.1	Low
	Restricted Area Frequency	0.1	0.34	0.64	0.1	Low
	PaintZone Frequency	0.11	0.25	0.51	0.13	Low
	Middle Range Frequency	0.04	0.32	0.81	0.18	Low
	3Point Frequency	0.2	0.54	0.8	0.13	High
	Rebound Chance	0.34	0.63	0.93	0.12	High
	Pace	0.15	0.51	0.89	0.16	High
	Time of Possession	0.02	0.12	0.25	0.05	Low

군집1에 속한 선수들은 Average Sceond per Touches, Average Dribble per Touches 변수를 통해서 공을 소유하는 시간이 짧다는 것을 확인할 수 있다. 공을 소유할 때에는 패스보다는 주로 슈팅으로 Touches를 마무리하는 선수들이 속해있다. 슈팅 타입은 3점슛 빈도가 가장 많다. 또한, Rebound Chance변수를 통해서는 군집1에 속한 선수들은 리바운드 기회를 많이 창출하는 적극성을 가진 선수들이라고 판단하였다. 대표적인 선수들로 Brook Lopez(2021), Jeff Green(2021) 등이 있다.

이 연구에서는 이러한 특성을 고려하여, 군집1을 3&D라고 정의하였다.

2) 군집2 (82명)

표 26. NBA 군집2 특성

군집	변수명	최소	평균	최대	표준편차	구분
2 (N=82)	Touches	0.34	0.6	0.9	0.12	High
	Front Court Touches	0.4	0.69	1	0.15	High
	Elbow Touches	0.09	0.35	0.73	0.14	Low
	PaintZone Touches	0.09	0.39	0.76	0.16	Low
	Average Second per Touches	0.07	0.24	0.56	0.13	Low
	Average Dribble per Touches	0.03	0.17	0.44	0.12	Low
	Shoot%	0.39	0.55	0.74	0.09	High
	Pass%	0.12	0.35	0.7	0.12	Low
	Usage Rate	0.3	0.58	0.86	0.11	High
	Restricted Area Frequency	0.06	0.33	0.66	0.12	Low
	PaintZone Frequency	0.13	0.49	0.89	0.16	Low
	Middle Range Frequency	0.04	0.5	1	0.2	Mid
	3Point Frequency	0	0.23	0.46	0.13	Low
	Rebound Chance	0.41	0.67	0.93	0.12	High
	Pace	0.07	0.44	0.88	0.2	Low
	Time of Possession	0.12	0.26	0.54	0.11	Low

군집2는 6개의 군집중에서 Usage Rate가 가장 높다. Usage뿐만 아니라 Touches도 많고 그중에서 주로 Front Court Touches 빈도가 가장 높은 특징을 보인다. 그러나 터치를 끝내는 동안 공을 가지고 있는 시간과 드리블 숫자는 상대적으로 적다. 슈팅은 주로 페인트존과 미들슛을 시도하며 리바운드 기회를 많이 창출하는 특징을 가지고 있다. 낮은 Pass%를 통해서 패스를 주기보다는 주로 받아서 슈팅으로 해결하는 선수들이라는 것을 알 수 있다. 대표적인 선수들로 Giannis Antetokounmpo(2021), Joel Embiid(2021), Nikola Jokic(2021)등이 있다. 성공률이 높은 Front Court에서의 슈팅을 주로 하면서, 팀내에서 주 득점원 역할을 하는 선수들이 속한 그룹이다.

이 연구에서는 이러한 특성을 고려하여, 군집2를 Finisher라고 정의하였다.

3) 군집3 (220명)

표 27. NBA 군집3 특성

군집	변수명	최소	평균	최대	표준편차	구분
3 (N=220)	Touches	0.31	0.64	0.97	0.15	High
	Front Court Touches	0.09	0.37	0.73	0.14	Low
	Elbow Touches	0	0.05	0.14	0.03	Low
	PaintZone Touches	0.01	0.08	0.2	0.04	Low
	Average Second per Touches	0.43	0.68	0.97	0.1	High
	Average Dribble per Touches	0.39	0.63	0.88	0.11	High
	Shoot%	0.2	0.43	0.65	0.08	Low
	Pass%	0.32	0.57	0.88	0.11	High
	Usage Rate	0.11	0.49	1	0.17	Low
	Restricted Area Frequency	0.04	0.28	0.51	0.09	Low
	PaintZone Frequency	0.08	0.39	0.84	0.17	Low
	Middle Range Frequency	0.06	0.41	0.81	0.17	Low
	3Point Frequency	0.01	0.38	0.74	0.15	Low
	Rebound Chance	0	0.36	0.76	0.19	Low
	Pace	0.09	0.5	0.9	0.17	Mid
	Time of Possession	0.33	0.57	0.98	0.15	High

군집3은 6개의 군집중에서 가장 Touches가 많은 군집이라는 대표적인 특징이 있다. 또한 Finisher 다음으로 가장 Usage Rate가 높은 군집이다. 터치 중에서 패스 빈도가 가장 높게 나타나지만, Usage가 높은 것으로 보아 득점으로 포제션을 마무리하는 역할도 담당한다. Finisher 군집과 다르게 군집3은 Average Second per Touches, Average Dribble per Touches값이 상위권에 분포되어 있다. 또한, Front Court보다는 Back Court에서 터치가 이루어지는 특징이 있다. 이러한 특성을 보아 군집3은 주로 Back Court에서 공을 오래 소유하며 경기 운영과 득점을 주로 하는 그룹이라고 할 수 있다. 대표적인 선수로는 Stephen Curry(2021), Lebron James(2021), Luka Doncic(2021) 등이 있다.

이 연구에서는 이러한 특성을 고려하여, 군집3을 Main Ball Handler로 정의하였다.

4) 군집4 (206명)

표 28. NBA 군집4 특성

군집	변수명	최소	평균	최대	표준편차	구분
4 (N=206)	Touches	0.02	0.32	0.58	0.11	Low
	Front Court Touches	0.05	0.35	0.74	0.14	Low
	Elbow Touches	0	0.05	0.13	0.03	Low
	PaintZone Touches	0	0.07	0.18	0.04	Low
	Average Second per Touches	0.09	0.35	0.66	0.13	Low
	Average Dribble per Touches	0.08	0.33	0.65	0.12	Low
	Shoot%	0.3	0.49	0.68	0.08	Low
	Pass%	0.24	0.48	0.74	0.09	Low
	Usage Rate	0.08	0.37	0.66	0.11	Low
	Restricted Area Frequency	0.04	0.22	0.51	0.09	Low
	PaintZone Frequency	0	0.3	0.66	0.14	Low
	Middle Range Frequency	0.04	0.4	0.81	0.18	Low
	3Point Frequency	0.18	0.5	0.79	0.12	Mid
	Rebound Chance	0.09	0.48	0.84	0.15	Low
	Pace	0.09	0.5	0.9	0.15	Mid
	Time of Possession	0.03	0.2	0.43	0.09	Low

군집4는 Touches, Usage Rate, Average Second, Dribble per Touches가 낮은 것으로 보아 공을 오래 소유하거나 공격 의존도가 높지 않은 특징을 가지고 있다. 플레이 스타일은 주로 Back Court에서 공을 소유하며, 슈팅과 패스는 비슷한 빈도를 나타낸다. 또한, 미들슛과 3점슛과 같이 중장거리 슈팅을 주로 던지는 특징이 있다. 대표적인 선수로는 Andrew Wiggins(2021), Seth Curry(2021), Buddy Heild(2018)등이 있다.

이 연구에서는 이러한 특성을 고려하여 군집4를 Long Ranger라고 정의하였다.

5) 군집5 (116명)

표 29. NBA 군집5 특성

군집	변수명	최소	평균	최대	표준편차	구분
5 (N=116)	Touches	0.11	0.32	0.58	0.11	Low
	Front Court Touches	0	0.28	0.59	0.12	Low
	Elbow Touches	0.08	0.33	0.6	0.11	Low
	PaintZone Touches	0.31	0.59	1	0.18	High
	Average Second per Touches	0.01	0.09	0.23	0.04	Low
	Average Dribble per Touches	0	0.05	0.13	0.03	Low
	Shoot%	0.26	0.59	0.88	0.12	High
	Pass%	0	0.28	0.6	0.13	Low
	Usage Rate	0	0.28	0.58	0.12	Low
	Restricted Area Frequency	0.35	0.67	1	0.16	High
	PaintZone Frequency	0.13	0.52	0.92	0.19	High
	Middle Range Frequency	0	0.2	0.63	0.14	Low
	3Point Frequency	0	0.02	0.14	0.03	Low
	Rebound Chance	0.3	0.58	0.85	0.12	High
	Pace	0.13	0.42	0.7	0.13	Low
	Time of Possession	0.02	0.09	0.2	0.04	Low

군집5 페인트존 터치가 가장 많고, Restriced 구역과 페인트존에서의 슈팅 빈도가 가장 많은 군집이다. 공을 잡으면 슈팅을 가장 많이 선택하는 특징이 있다. 공을 소유하는 시간이 많지는 않고, 슈팅 타입도 다양하지 않고, 골밑슈팅을 주로 선택한다. 또한 리바운드 기회를 많이 창출하는 특징이 있다. 대표적인 선수로는 Clint Capela(2021), Ivica Zubac(2021), Rudy Govert(2021)등이 있다.

이 연구에서는 이러한 특징을 고려하여 군집5를 Traditional Bigman이라고 정의하였다.

6) 군집6 (130명)

표 30. NBA 군집6 특성

군집	변수명	최소	평균	최대	표준편차	구분
6 (N=130)	Touches	0	0.23	0.51	0.11	Low
	Front Court Touches	0.01	0.22	0.54	0.12	Low
	Elbow Touches	0	0.04	0.1	0.2	Low
	PaintZone Touches	0.01	0.09	0.24	0.05	Low
	Average Second per Touches	0.01	0.12	0.31	0.07	Low
	Average Dribble per Touches	0.01	0.11	0.28	0.06	Low
	Shoot%	0.12	0.32	0.56	0.09	Low
	Pass%	0.22	0.57	0.94	0.15	High
	Usage Rate	0	0.17	0.38	0.09	Low
	Restricted Area Frequency	0	0.22	0.48	0.11	Low
	PaintZone Frequency	0	0.13	0.32	0.07	Low
	Middle Range Frequency	0.02	0.2	0.5	0.11	Low
	3Point Frequency	0.35	0.69	1	0.13	High
	Rebound Chance	0.18	0.61	0.87	0.14	High
	Pace	0.15	0.53	0.89	0.16	High
	Time of Possession	0	0.07	0.18	0.04	Low

군집6은 3&D와 유사한 특성을 가지고 있다. 하지만 군집6은 Touches를 주로 패스로 마무리한다는 특징이 있다. 슈팅 타입은 3점슛 비중이 가장 높으며, 다른 슈팅을 거의 시도하지 않는다. 자신의 공격보다는 패스를 통해 공의 흐름을 이어주는 역할을 한다. 대표적인 선수로는 Danny Green(2021), Jae Crowder(2021), Joe Ingles(2021)등이 있다.

이 연구에서는 이러한 특성을 고려하여 군집6을 Linker로 정의하였다.

3.3 비교

각 리그의 시즌별 정규리그 우승팀 주전 라인업의 포지션 분포는 다음과 같다.

표 31. KBL 정규리그 우승팀 포지션 분포

KBL					
시즌	팀명	Main Ball Handler	Linker	Finisher	3&D
2020-21	이지스	이정현 유현준	김지완	송교창 라건아	-
2019-20	나이츠	김선형	최준용	자밀 워니	김민수 안영준
2018-19	모비스	이대성 함지훈	양동근 박경상	라건아	-
2017-18	프로미	두경민	-	디욘테 버튼	김태홍 서민수 윤호영
2016-17	인삼공사	이정현 박찬희	양희종	사이먼 오세근	-

표 32. NBA 파이널 우승팀 포지션 분포

NBA							
시즌	팀명	Main Ball Handler	Linker	Finisher	3&D	Long Ranger	Traditional Bigman
2020-21	MIL	-	미들턴 할로데이	아테토쿰보	로페즈 디빈첸조	-	-
2018-19	TOR	레너드	라우리	시아캄	그린	이바카	
2017-18	GSW	커리 그린	-	듀란트	이궈달라	탐슨	-
2016-17	GSW	커리 그린	-	듀란트	이궈달라	탐슨	-
2015-16	CLE	어빙 제임스	델라베도바	-	-	스미스	탐슨

제 4절. 랜덤포레스트 분류 모델

4.1 데이터셋

3절 과정을 통하여 모든 선수들에게 새로운 포지션을 부여하고, 이를 기존 데이터셋에 새로운 칼럼으로 추가한다. 포지션이라는 새로운 칼럼이 생성되었고, 이를 랜덤포레스트 분류 모델의 데이터셋으로 활용한다.

4.2 데이터 분할

4.1에서 생성된 데이터셋을 모델 생성과 학습을 위한 Train data와 모델의 성능개선, 테스트를 위한 Test data로 6:4 비율로 무작위 분할한다.

표 33. KBL 데이터 분할 결과

구분	변수명	Train data(N=218)		Test data(N=146)	
		평균	표준편자	평균	표준편차
KBL	Touches	1140.48	612.65	1177.94	640.39
	PaintZone Touches	1.11	0.41	1.14	0.45
	Shoot%	0.36	0.11	0.35	0.11
	Pass%	0.49	0.15	0.5	0.14
	Usage Rate	20.49	7.47	19.73	6.8
	PaintZone Frequency	0.47	0.19	0.44	0.19
	Middle Range Frequency	0.17	0.09	0.18	0.09
	3Point Frequency	0.35	0.22	0.38	0.22
	Total Rebound%	10.53	5.71	9.39	5.22
	Pace	40.74	2.61	40.58	2.31

표 34. NBA 데이터 분할 결과

리그	변수명	Train data(N=539)		Test data(N=360)	
		평균	표준편차	평균	표준편차
NBA	Touches	51.75	16.86	51.87	17.52
	Front Court Touches	28.65	8.23	28.19	7.87
	Elbow Touches	1.6	1.63	1.59	0.32
	PaintZone Touches	2.67	2.63	2.77	2.69
	Average Second per Touches	2.86	1.26	2.91	1.32
	Average Dribble per Touches	2.07	1.53	2.12	1.6
	Shoot%	0.46	0.11	0.47	0.11
	Pass%	0.31	0.11	0.32	0.11
	Usage Rate	20.29	5.36	20.65	5.52
	Restricted Area Frequency	0.32	0.16	0.32	0.16
	PaintZone Frequency	0.15	0.07	0.16	0.07
	Middle Range Frequency	0.19	0.1	0.19	0.11
	3Point Frequency	0.33	0.19	0.33	0.19
	Rebound Chance	51.85	6.89	52.13	6.65
	Pace	99.22	2.74	99.16	2.68
	Time of Possession	2.68	1.9	2.78	2.01

4.3 랜덤포레스트 분류 모델 생성

랜덤포레스트 기법을 구현한 연구 환경은 다음과 같다

표 35. 랜덤포레스트 연구 환경

언어	기법	라이브러리	함수
Python	RandomForest	sklearn.ensemble	RandomForestClassifier()
Phthon	RandomSearch	sklearn.model.selection	RandomizedSearchCV()

Python에서 제공하는 랜덤포레스트 함수의 주요 하이퍼 파라미터는 다음과 같다. 적절한 하이퍼 파라미터 설정은 Train data에 대한 과대적합(overfitting)을 방지한다.

표 36. 랜덤포레스트 주요 하이퍼 파라미터

하이퍼 파라미터	의미	입력값 예시
N_estimators	생성할 트리의 개수	120, 300, 500, 800, 1200
Max_depth	트리의 최대깊이	5, 8, 15, 25, 30, None
Min_samples_split	노드를 분할하기 위한 최소 데이터 개수	1, 2, 5, 10, 15, 100
Min_samples_leaf	리프노드가 되기 위해 필요한 최소 샘플 데이터 개수	1, 2, 5, 10
Max features	데이터 feature 최대 개수	1, 3, 5, 10

4.3.1 KBL 랜덤포레스트 모델

1) 초기 모델

랜덤포레스트 초기 모델 생성에 사용될 input data는 4.2절에서 분할된 Train data(N=218)을 사용한다.

표 37. 랜덤포레스트 파이썬 모듈 및 함수

언어	기법	모듈	함수
Python	RandomForest	sklearn.ensemble	RandomForestClassifier()
Phthon	RandomSearch	sklearn.model.selection	RandomizedSearchCV()

초기 모델 생성에 사용된 일부 코딩은 다음과 같다.

```
rf_model = RandomForestClassifier(random_state = 47)
rf_model.fit(x_train, y_train)
```

<그림 16> KBL 초기 모델 생성 코딩

랜덤포레스트는 하이퍼 파라미터를 설정하지 않은 초기 모델에서도 성능이 좋다는 장점이 있다. 그러나 Train data에 대한 정확도가 100%이므로 초기 모델은 Train data에 대한 과대적합으로 판단한다.

표 38. KBL 초기 모델 성능

구분	정확도
Train data	1.000
Test data	0.918
f1-score	0.91

2) 하이퍼 파라미터 튜닝

이 연구에서는 하이퍼 파라미터 튜닝 방법으로 Random Search 기법을 사용했다. Random Search 기법을 구현하기 위한 일부 코딩은 다음과 같다.

```
hyperparameter_space = {'n_estimators' : randint(low=1, high=200),
                        'max_depth':randint(low=1, high=10),
                        'min_samples_leaf':randint(low=1, high=30),
                        'min_samples_split':randint(low=1, high=10),
                        'max_features' : randint(low=1, high=10)}

from sklearn.model_selection import RandomizedSearchCV
rs = RandomizedSearchCV(rf_model, param_distributions=hyperparameter_space,
                        n_iter=10, scoring="accuracy", random_state=47,
                        n_jobs=-1, cv=2, return_train_score=True)
```

<그림 17> KBL 하이퍼 파라미터 수정 코딩

Random Search 결과 최적의 하이퍼 파라미터는 다음과 같다.

표 39. KBL 최적 하이퍼 파라미터

하이퍼 파라미터 튜닝	
하이퍼 파라미터	**최적값**
N_estimators	80
Max_depth	4
Min_samples_split	7
Min_samples_leaf	6
Max features	4

3) 모델 튜닝

Random Search 기법을 통해 얻은 하이퍼 파라미터 값들을 입력하여 모델을 다시 생성한다. 하이퍼 파라미터 튜닝을 통해 Train data에 대한 과적합을 줄이고 Test data에 대한 성능이 향상되었다. 따라서 최종모델은 다음과 같이 설정하였다.

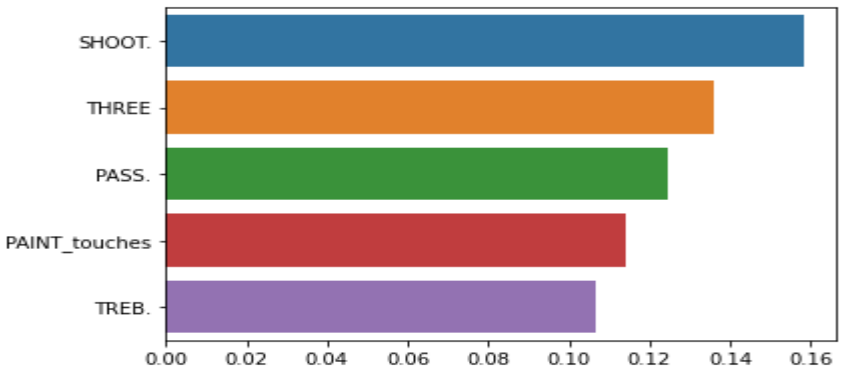
```
RandomForestClassifier(max_depth=4,
                       max_features=5,
                       min_samples_leaf=6,
                       min_samples_split=7,
                       n_estimators=80,
                       random_state=47)
```

<그림 18> KBL 최종 분류 모델

표 40. KBL 최종 모델 성능

구분	정확도
Train data	0.972
Test data	0.943
f1-score	0.943

4) 변수 중요도



<그림 21> KBL 분류 모델 변수 중요도

KBL 랜덤포레스트 모델에서는 Touches의 요소중 Shoot% 변수의 변수 중요도가 가장 높다. 변수 중요도를 통해서 KBL 리그에서는 공을 잡았을 때, 슈팅 빈도가 높은 선수와 그렇지 않은 선수로 가장 먼저 분류된다는 것을 확인할 수 있다.

4.3.2 NBA 랜덤포레스트 모델

1) 초기 모델

랜덤포레스트 초기 모델 생성에 사용될 input data는 4.2절에서 분할된 Train data(N=539)을 사용한다.

랜덤포레스트는 하이퍼 파라미터를 설정하지 않은 초기 모델에서도 성능이 좋다는 장점이 있다. 그러나 Train data에 대한 정확도가 100%이므로 초기 모델은 Train data에 대한 과대적합으로 판단한다.

표 41. NBA 초기 모델 성능

구분	정확도
Train data	1.000
Test data	0.881
f1-score	0.88

2) 하이퍼 파라미터 튜닝

이 연구에서는 하이퍼 파라미터 튜닝 방법으로 Random Search 기법을 사용했다. Random Search 기법을 구현하기 위한 일부 코딩은 다음과 같다.

```
hyperparameter_space = {'n_estimators' : randint(low=1, high=200),
                        'max_depth':randint(low=1, high=10),
                        'min_samples_leaf':randint(low=1, high=30),
                        'min_samples_split':randint(low=1, high=10),
                        'max_features' : randint(low=1, high=10)}

from sklearn.model_selection import RandomizedSearchCV
rs = RandomizedSearchCV(rf_model, param_distributions=hyperparameter_space,
                        n_iter=10, scoring="accuracy", random_state=47,
                        n_jobs=-1, cv=2, return_train_score=True)
```

<그림 20> NBA 하이퍼 파라미터 수정 코딩

Random Search 결과 최적의 하이퍼 파라미터는 다음과 같다.

표 42. NBA 최적 하이퍼 파라미터

하이퍼 파라미터 튜닝	
하이퍼 파라미터	**최적값**
N_estimators	73
Max_depth	8
Min_samples_split	9
Min_samples_leaf	8
Max features	7

3) 모델 튜닝

Random Search 기법을 통해 얻은 하이퍼 파라미터 값들을 입력하여 모델을 다시 생성한다. 하이퍼 파라미터 튜닝을 통해 Train data에 대한 과적합을 줄이고 Test data에 대한 성능이 향상되었다. 따라서 최종모델은 다음과 같이 설정하였다.

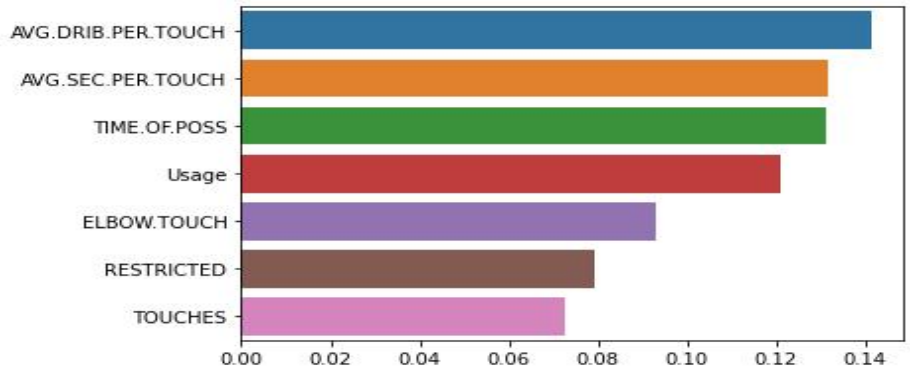
```
RandomForestClassifier(max_depth=8,
                       max_feature=7,
                       min_sample_leaf=8,
                       min_sample_split=9,
                       n_estimators=73,
                       random_state=47)
```

<그림 21> NBA 최종 분류 모델

표 43. NBA 최종 모델 성능

구분	정확도
Train data	0.954
Test data	0.931
f1-score	0.926

4) 변수 중요도



<그림 22> NBA 분류 모델 변수 중요도

NBA 랜덤포레스트 모델에서는 Touch 대비 평균 드리블 개수 변수의 변수 중요도가 가장 높다. 변수 중요도를 통해서 자신이 공을 소유하는 동안 드리블을 하는 선수인지 또는 공을 소유하는 시간에 따라서 가장 먼저 선수를 분류하는 것을 확인할 수 있다.

제 5장 논의

이 연구에서는 최근 5년 KBL과 NBA의 Advanced stat을 이용하여 각 리그의 새로운 포지션을 재정의하여 리그와 포지션별 특성을 분석하고자 하였다. KBL 데이터는 정규 기록 조건을 충족하는 총 364명의 선수와 10개의 변수로 구성되어있다. NBA 데이터는 정규 기록 조건을 충족하는 총 899명의 선수와 16개의 변수로 구성되어있다.

1. 데이터 탐색

k-means 군집분석은 군집 간 거리를 계산하여 군집을 형성하는 방법이기 때문에, 데이터의 단위와 스케일에 민감하다.

첫째, 이 연구에서는 두 리그의 데이터들의 기술 통계량을 통하여 각 변수들의 특성과 단위 및 범위를 파악과 정규화의 필요성을 판단하였다. 수집된 데이터의 타입은 모두 연속형 변수이다.

둘째, 수집된 데이터의 이상치 및 결측치 여부를 파악하여 평균값으로 대체하였다. 수집된 데이터의 결측치는 존재하지 않았다. 두 데이터 모두 Front Court에서 발생하는 이벤트들과 관련이 있는 변수들에 대해서 이상치가 상대적으로 많이 발견되었다.

셋째, 각 변수들의 단위와 범위는 매우 다양하다. 유클리디안 거리를 구하는 과정에서 단위가 다르면 적절하지 않은 값들이 나오기 때문에 로그변환을 통하여 값의 크기를 줄인 뒤, Min-Max 정규화 방법을 통해서 모든 값의 범위를 0~1사이로 통일시켰다.

2. 군집 모델 개발

k-means 군집분석에서 가장 중요한 부분은 k값 설정이다. 최적의 k값은 다양한 방법으로 구해지고 있다. 이 연구에서는 그중에서 엘보우 기법과 실루엣 기법을 사용하여 얻은 k값을 관찰하여, 군집이 적절하게 형성하는지 파악하였다.

KBL은 엘보우 기법을 통해 얻은 k값을 통하여 두 가지 군집으로 나눴을 때, 국내 선수와 외국 선수로 구분되는 것을 확인하였다. 외국 선수 97명은 단 한명도 다른 군집으로 분류되지 않았다. k값이 3일때에도 외국 선수 97명은 다른 군집으로 분류되지 않았다. 이를 통하여 최근 5년간 국내에서 뛰는 외국 선수들의 유형은 매우 유사하다는 것을 파악할 수 있다. 실루엣기법을 통해 얻은 k값을 통하여 4가지 군집으로 나누었을 때에 비로소 외국 선수 14명이 다른 군집으로 분류가 되는 것을 확인할 수 있었다. 하지만 이또한 외국 단신 가드라고 불리는 선수들이 주로 포함되었다. k값이 5일 때부터는 외국 선수들이 더 이상 다른 군집으로 형성되지 않았고, 오히려 국내 선수들의 군집이 과군집화가 되는 현상이 발생하였다. 따라서 KBL 군집의 최적 k값은 실루엣 기법을 통해 나온 k=4로 설정하였다.

NBA는 엘보우 기법과 실루엣 기법을 통하여 얻은 결과가 k=6이라는 동일한 결과가 나왔고 이를 최적의 k값으로 설정하였다. k값이 3일때에도 실루엣 기법에서 높은 값이 나오지만 이는 소군집화라고 판단하였다.

3. 포지션 정의

k-means를 통해 형성된 각 군집은 연속적인 팔레트 형식으로 군집들이 분포되어 있다. 연속적으로 얽혀있는 집단 간의 특징을 이산적으로 구분한다는 것 자체가 한계가 있다. 따라서 Bonferroni 검정을 통해서 통계적으로 유의미한 차이가 있음을 검증한 뒤에 각 군집의 특성을 기술 통계량을 통하여 파악하였다. 모든 변수들의 단위를 0과 1사이로 정규화했기 때문에, 0.5를 기준으로 1에 가까울수록 강한 특성, 0에 가까울수록 약한 특성을 가진다고 판단하였다.

이를 통하여 KBL은 Main Ball Handler 58명, Linker 95명, Finisher 107명, 3&D

104명으로 포지션을 새롭게 정의하였다. NBA는 Main Ball Handler 220명, Linker 145명, Finisher 82명, 3&D 130명으로 정의되었고, Long Ranger 206명, Traditional Bigman 116명으로 추가적인 포지션을 정의하였다.

Main Ball Handler는 Touches, Pass% 변수값이 가장 큰 집단으로써, 공을 가장 오래 소유하면서 경기를 운영하고 패턴을 지시하는 포지션이다. 뿐만 아니라, Finisher 다음으로 가장 높은 Usage Rate를 기록하면서 경기 운영뿐만 아니라 일부 Finisher 특성을 가지고 있는 포지션이다.

Linker는 Usage Rate가 높지 않아 자신의 공격 보다는 다른 선수들의 공격 기회를 위해 존재하는 선수로써 연결고리의 역할을 하는 포지션이다. 또한 Linker는 Shoot% 빈도도 낮으며 주로 패스를 하는 특징을 가지고 있다. 주로 던지는 슈팅은 3점슛이다. Linker는 Main Ball Handler와는 다르게 어시스트를 동반한 패스보다는 연결고리 역할로써의 패스를 주로 하는 특징을 가지고 있다.

Finisher는 가장 Usage Rate가 높은 포지션으로 팀내에서 주로 공격을 마무리하는 역할을 담당하고 있다. 많은 공격을 담당하기 위해서는 높은 공격 성공률을 가지고 있어야하기 때문에 Back Court보다는 Front Court슈팅의 빈도가 높은 특징이 있다. KBL에서는 외국 선수들이 Finisher에 주로 포함되어있다.

3&D는 현대농구에서 흔히 언급되는 단어로써, 팀내에서 수비, 허슬과 3점슛 능력을 가지고 있는 특징이 있다. 수비와 허슬은 숫자로 표현할 수 없는 영역이 있기 때문에, 3&D 포지션으로 구분된 선수들의 일부 영상분석을 참고하여 플레이 스타일을 파악하였다. NBA 데이터에서는 Rebound Chance라는 변수를 통해서 리바운드에 참여하려는 적극성을 파악할 수 있는 변수라고 판단하여 이를 3&D를 정의하는 데에 사용하였다.

Long Ranger는 공을 오래 소유하지 않으면서 소유 했을 시에는 미들, 3점슛을 주로 던지는 유성의 포지션이다. 공을 소유했을때에 짧은 시간내에 슈팅을 하거나, 패스를 받아서 짧은 시간 내에 슈팅을 하는 특징이 있다.

Traditional Bigman은 미들, 3점슈팅 시도보다는, 대부분의 슈팅과 플레이가 PaintZone과 Restricted Area에서 이루어지는 정통 빅맨이다.

4. 랜덤포레스트 분류 모델 개발

랜덤포레스트는 훈련데이터의 단위와 범위가 달라도 분류 성능이 좋다는 강점이 있다. 따라서 랜덤포레스트 모델의 훈련데이터는 정규화 과정을 거치지 않은 원본 데이터를 사용하였다. 테스트 데이터로 평가한 초기 분류 모델의 성능은 KBL 91.0%, NBA는 88.0%의 정확도가 나왔다. 하지만 두 가지 모델 모두 훈련데이터에 대한 정확도가 100%로 과적합 상태라고 판단하였다. 훈련데이터에 대한 정확도 손실이 있더라도 테스트 데이터에 대한 성능을 향상시키기 위해 하이퍼 파라미터를 조정하였다. 하이퍼 파라미터 설정은 Random Search 방법을 사용하였고 주요 하이퍼 파라미터는 N_estimators, Max_depth, Min_samples_split, Min_samples_leaf, Max features가 있다. 두 모델의 하이퍼 파라미터는 각각 KBL은 80, 4, 7, 6, 4, NBA는 73, 8, 9, 8, 7로 설정하였다. 수정된 모델의 테스트 데이터의 성능은 94.3%, 92.6%로써 초기 모델보다 더 좋은 성능을 보였다. 또한, 포지션 분류 모델에서 KBL은 Shoot%, NBA는 Average Dribble per Touch 변수가 가장 높은 변수 중요도를 보였다.

제 6장 결론 및 제언

1. 결론

단체 스포츠에서 포지션은 팀의 선수구성과 경기 전력에서 가장 중요한 요인이다. 최근 농구 경기에서는 기존 5가지 포지션으로는 정의하기 힘든 유형이 선수들이 생겨나고 있다. 따라서 이 연구에서는 기존 포지션에 대한 개념과는 별개로 Advanced Stat을 활용하여 오로지 선수들의 플레이 스타일을 나타내는 데이터를 수집하였다. 이를 활용하여 군집화하고 군집의 특성에 맞는 새로운 포지션을 정의하였다. 그 결과의 요약은 다음과 같다.

첫째, 이상치와 단위에 민감한 군집분석을 실시하기 전 데이터를 정제 및 변환하였다. 수집된 5년의 데이터 중에서 정규 기록 기준을 충족시키는 선수들의 데이터가 많지 않기 때문에 이상치를 제거하는 방법보다는 평균치로 대체하는 방법을 선택하였다. 그 후에 로그 변환을 통하여 각 변수들의 범위를 비율 범위로 변환한 뒤, Min-Max정규화 방법을 통하여 모든 변수들의 범위를 0과 1 사이로 변환하였다.

둘째, k-means 군집분석의 최적 k값을 찾기 위해 엘보우 기법과 실루엣 기법을 사용하였다. 이 연구에서는 군집내 거리의 합만 고려한 엘보우 기법의 한계를 군집간 거리를 고려한 실루엣 기법을 같이 사용하여 관찰하였다. 두 기법을 관찰한 결과 KBL과 NBA의 최적 k값은 각각 4, 6으로 설정 하였다. KBL에서는 97명의 전체 외국 선수들이 k값이 4가 되기전까지는 다른 군집에 속하지 않는 현상이 발생하였다. 이는 국내에 뛰는 외국선수들의 유형이 5년간 매우 유사하다는 것을 의미하고, 국내 선수들과는 매우 구분된 플레이를 한다는 것을 파악할 수 있었다. k값이 4로 증가했을 때에 12명의 외국 선수들이 분리되어 군집을 형성한 것을 확인하였다. NBA는 엘보우 기법과 실루엣 기법에서 추천 받은 k값이 6으로 일치하였다. 군집화 결과 NBA는 상대적으로 KBL보다 연속적으로 군집 분포가 형성되었다는 것을 시각화를 통해 확인하였다.

셋째, 각 군집의 특성을 각 변수들의 기술통계량을 통하여 파악하였고, 특성에 맞는 포지션을 새롭게 정의하였다. KBL은 Main Ball Handler, Linker, Finisher, 3&D, 총 4가지 포지션으로 정의하였다. NBA는 Main Ball Handler, Linker, Finisher, 3&D, Long Ranger, Traditional Bigman, 총 6가지 포지션으로 정의하였다. 포지션에 속한 선수들은 KBL은 각각 58명, 95명, 107명, 104명이고 NBA는 각각 145명, 85명, 220명, 206명, 116명, 130명이 속하였다. 또한 각 리그의 최근 5년간 우승팀(KBL:정규리그, NBA:파이널)의 주전 라인업의 포지션 분포를 관찰하였다. NBA는 2015-16시즌에 어빙과 제임스라는 강력한 Main Ball Handler를 앞세워 우승을 하였다. 어빙과 제임스가 번갈아가며 공을 운반하고 아이솔레이션으로 공격까지 마무리하며 나머지 선수들에 대한 공격 의존도는 낮았다. 2015-16시즌 이후 NBA 파이널 우승팀에서는 Traditional Bigman 포지션의 선수가 없었다. 최근 정통센터라고 불리는 선수들은 빠른 템포의 현대농구에서 플레이스타일을 변화하지 않으면 경쟁력이 떨어진다는 것을 유추해 볼 수 있다. 변화에 성공한 대표적인 예로 2020-21시즌 우승팀에 속한 로페즈 선수는 2016-17시즌까지는 Traditional Bigman 포지션으로 분류되었지만, 2017-18시즌 이후로는 3&D 포지션으로 분류되었다. 이는 리그의 흐름에 맞게 플레이 스타일의 변화가 발생했다고 할 수 있다. 그리고 최근 시즌으로 거듭할수록 우승팀에서는 Main Ball Handler의 숫자가 줄어들고 있다. 빠른 템포의 공격을 하는 현대농구에서 한 명의 선수가 공을 오래 소유하며 공격 템포를 늦추는 것은 경쟁력이 떨어진다고 할 수 있다. 2020-21시즌 우승팀인 밀워키는 Main Ball Handler가 없다. 이는 공을 오래 끌지 않으며 트랜지션을 통한 빠른 공격 또는 패스 플레이 위주로 경기를 운영한다는 것을 파악할 수 있다. 또한 3&D의 존재로 확실한 Finisher가 존재한다면, 그만큼 수비에 치중된 선수들의 존재가 필요하다. 하지만 최근 KBL은 Main Ball Handler의 역할이 아직까지 중요한 것으로 파악된다. 리그 특성상 KBL의 대부분의 팀들은 상대적으로 느린 템포의 패턴 플레이에 의존한다. 그렇기 때문에 Touches가 많은 Main Ball Handler의 선수가 많이 분포된 것을 확인할 수 있다. KBL에서 Finisher 포지션은 외국 선수들이 주로 속해있는 그룹이다. 송교창, 오세근은 우승시즌에 외국 선수들의 집단으로 분류될 정도의 퍼포먼스를 보였다고 파악된다. KBL은 외국 선수가 한 쿼터에 두명이 뛸 수 없기 때문에, Finisher 포지션의 국내선수의 존재는 매우 강력하다고 할 수 있다. 우리나라의 국제 무대 경쟁력을 위해서는 NBA의 포지션 트렌드를 참고해야 할 필요

성이 있다. 국제 경기에서는 공을 오래 소유하는 Main Ball Handler를 두명이상 동시 출전 시키는 것은 비효율적이다. 오히려 공을 오래 소유하지 않으면서 짧은 시간에 슈팅할 수 있는 Finisher 유형의 선수들을 포함 시키는 것이 더욱 경쟁력이 있을 것이라 판단한다. 또한 국내 엘리트 또는 프로 선수들을 육성할 때 기존 농구 1-5번의 포지션 보다는 매년 NBA리그의 포지션 변화를 고려하여 육성에 적용할 필요가 있다.

넷째, 초기 모델 생성과 하이퍼 파라미터를 수정한 최종 모델을 개발하였다. KBL과 NBA의 초기 모델의 성능은 각각 91.0%, 88.0%였다. 두 모델 모두 훈련데이터에 대한 정확도가 100%라는 과적합 현상을 보이기 때문에, 훈련데이터에 대한 성능의 약간의 손실을 보더라도 테스트 데이터에 대한 성능을 향상시키기 위해 하이퍼 파라미터를 수정하였다. Random Search 방법을 통하여 두 모델의 N_estimators, Max_depth, Min_samples_split, Min_samples_leaf, Max feature는 각각 KBL은 80, 4, 7, 6, 4, NBA는 73, 8, 9, 8, 7로 설정하였다. 수정된 모델의 테스트 데이터의 성능은 94.3%, 92.6%로써 초기 모델보다 더 좋은 성능을 보였다. 또한 포지션을 분류하는데 가장 중요한 요인으로 KBL은 Shoot%이며, Touches의 마무리를 어떻게 하는지에 따라 가장 먼저 분류한다. NBA는 Average Dribble per Touch로 터치를 마무리할 때 평균 드리블 개수로 가장 먼저 포지션을 분류한다.

이 연구는 Advanced Stat, Tracking data 중 플레이 특성만을 나타내는 양적인 변수들을 추출한 뒤 머신러닝을 통해 포지션을 새롭게 정의하였다. 변수 선정 과정에서 개인 능력의 수준을 포지션 정의에서 최대한 제외했다는 점에서 독창적이며 이 점에서 선행 연구들과의 차별성이 있다. 또한, 스포츠에서 포지션을 양적인 변수들을 활용하여 분류할 수 있다는 새로운 접근과 장을 열었다는 점에서 이 연구는 의의가 있다.

2. 제언

이 연구에서 얻은 결과와 논의를 통하여 얻은 몇 가지 제한점과 후속 연구를 위한 제언을 하고자 한다.

첫째, 이 연구는 현대 농구라는 개념에 집중하여 각 리그의 최근 5시즌 데이터만을 사용하였다. 따라서 시즌이 거듭될수록 현대 농구의 범위와 트렌드는 변화할 것이고, 이 연구에서 개발한 모델도 거듭 개발해야 할 필요성이 있다.

둘째, NBA Tracking Data의 존재로 Advanced Stat보다 더 정밀하게 플레이 특성을 파악할 수 있었다. 두 리그의 같은 포지션끼리 유사한 특성을 나타내긴 하지만 NBA에서 사용된 변수들은 KBL에 존재하지 않기 때문에 두 리그를 비교한다는 것에 한계가 있다. 따라서 KBL도 NBA의 Tracking data와 유사한 유형의 데이터를 수집한다면 더 다양한 특성들을 발견할 수 있을 것이다.

셋째, 이 연구에서 형성된 군집들은 팔레트 형식으로 연속적으로 연결되어 있다. 연속적인 집단들을 이산적으로 분류하는데에 한계점이 있다. 따라서 추후 연구에서는 이를 해결하기 위한 새로운 군집화, 분류 방법론이 동시에 필요하다.

참 고 문 헌

한수철, 전수영, 진서훈(2008). k-중앙개체 군집방법을 이용한 한국 프로농구선수의 군집화. 한국자료분석학회, 10.6, 3423-3433.

저우차우, 최형준(2019). 세계 남자 농구 월드컵 경기대회의 공식기록에 기반한 경기내용에 관한 군집 분석, 한국체육학회지 59.3, 397-411.

조성원, 한형래, 백경완, 박정준. (2019). 국내 남자 프로농구 선수의 포지션별 체력 및 등속성 근기능 비교. 코칭능력개발지, 21(2), 55-63.

김주학, 최형준(2015). 자료범주에 따른 경기기록 승패의 재해석. 한국체육측정평가학회지, 17(1), 1-12.

김종원, 최형준(2021). 군집분석을 통한 K리그 축구팀 플레이스타일 분류. 한국체육측정평가학회지, 23(1), 1-9.

강지연(2019). 기계학습 기반 축구경기 기대득점 모델 개발. 명지대학교 기록정보과학전문대학원 석사학위논문.

황규인(2018). 의사결정나무 분석 기법을 활용한 프로야구 외국인 투수 재계약 확률 예측, 고려사이버대학교 융합정보대학원 석사학위논문.

홍준호, 오민지, 조용빈, 이경희, 조완섭(2020). 다차원 데이터의 군집분석을 위한 차원 축소 방법: 주성분분석 및 요인분석 비교. 한국빅데이터학회 학회지, 5(2), 135-143.

황종필(2015). 밀도기반 군집화의 시각화. 고려대학교 정책대학원 석사학위논문.

신성원(2011). 이상치를 고려한 결측 자료의 모형 기반 군집 분석. 고려대학교 대학원 석사학위논문.

서명교, 윤원영(2017) .마할라노비스거리를 이용한 군집기반 열간 조압연설비 상태모니터링과 진단.대한산업공학회지,43(4),298-307.

Shin, J. E., Jung, B. H., & Lim, D. H(2015). Big data distributed processing system using RHadoop. Journal of the Korean Data and Information Science Society. Korean Data and Information Science Society.

Julius Demenius, Rasa Kreivytė(2017). The benefits of advanced data analytics in

basketball: Approach of managers and coaches of Lithuanian basketball league teams. Lithuanian Sports University, Kaunas, Lithuania

Zhang S, Lorenzo A, Gómez MA, Mateus N, Gonçalves B, Sampaio J(2018). Clustering performances in the NBA according to players' anthropometric attributes and playing experience. J Sports Sci. 2018 Nov;36(22):2511-2520.

부 록

<부록 1> KBL 변수별 Bonferroni 검정 결과

KBL Touches 사후분석

변수명	군집1	군집2	군집2-군집1(평균)	p-value
Touches	1	2	-1001.96	<0.05
	1	3	-687.53	<0.05
	1	4	-1370.41	<0.05
	2	3	314.43	<0.05
	2	4	-368.44	<0.05
	3	4	-682.873	<0.05

KBL PaintZone Touches 사후분석

변수명	군집1	군집2	군집2-군집1(평균)	p-value
PaintZone Touches	1	2	0.29	<0.05
	1	3	0.26	<0.05
	1	4	0.91	<0.05
	2	3	-0.02	0.88
	2	4	0.62	<0.05
	3	4	0.65	<0.05

KBL Shoot% 사후분석

변수명	군집1	군집2	군집2-군집1(평균)	p-value
Shoot%	1	2	-0.02	0.13
	1	3	0.14	<0.05
	1	4	0.17	<0.05
	2	3	0.16	<0.05
	2	4	0.19	<0.05
	3	4	0.03	<0.05

KBL Pass% 사후분석

변수명	군집1	군집2	군집2-군집1(평균)	p-value
Pass%	1	2	0.04	<0.05
	1	3	-0.22	<0.05
	1	4	-0.18	<0.05
	2	3	-0.26	<0.05
	2	4	-0.22	<0.05
	3	4	0.04	<0.05

KBL Usage Rate 사후분석

변수명	군집1	군집2	군집2-군집1(평균)	p-value
Usage Rate	1	2	-8.05	<0.05
	1	3	3.63	<0.05
	1	4	-7.71	<0.05
	2	3	11.68	<0.05
	2	4	0.34	0.9
	3	4	-11.34	<0.05

KBL PaintZone Frequency 사후분석

변수명	군집1	군집2	군집2-군집1(평균)	p-value
PaintZone Frequency	1	2	-0.08	<0.05
	1	3	0.26	<0.05
	1	4	-0.09	<0.05
	2	3	0.33	<0.05
	2	4	-0.01	0.9
	3	4	-0.35	<0.05

KBL Middle Range Frequency 사후분석

변수명	군집1	군집2	군집2-군집1(평균)	p-value
Middle Range Frequency	1	2	-0.02	0.37
	1	3	0.01	0.85
	1	4	-0.04	<0.05
	2	3	0.04	<0.05
	2	4	-0.02	0.62
	3	4	-0.05	<0.05

KBL 3Point Frequency 사후분석

변수명	군집1	군집2	군집2-군집1(평균)	p-value
3Point Frequency	1	2	0.11	<0.05
	1	3	-0.27	<0.05
	1	4	0.13	<0.05
	2	3	-0.37	<0.05
	2	4	0.02	0.51
	3	4	0.39	<0.05

KBL Total Rebound% 사후분석

변수명	군집1	군집2	군집2-군집1(평균)	p-value
Total Rebound%	1	2	-0.59	0.66
	1	3	9.88	<0.05
	1	4	0.55	0.69
	2	3	10.48	<0.05
	2	4	1.15	0.06
	3	4	-9.33	<0.05

KBL Pace 사후분석

변수명	군집1	군집2	군집2-군집1(평균)	p-value
Pace	1	2	-0.59	0.66
	1	3	9.88	<0.05
	1	4	0.55	0.69
	2	3	10.48	<0.05
	2	4	1.15	0.57
	3	4	-9.33	<0.05

<부록 2> NBA 변수별 Bonferroni 검정 결과

NBA Touches 사후분석

변수명	군집1	군집2	군집2-군집1(평균)	p-value
Touches	1	2	-2.21	0.39
	1	3	21.25	<0.05
	1	4	-4.15	<0.05
	1	5	32.71	<0.05
	1	6	11.02	<0.05
	2	3	23.37	<0.05
	2	4	-2.03	0.46
	2	5	34.83	<0.05
	2	6	13.14	<0.05
	3	4	-25.40	<0.05
	3	5	11.46	<0.05
	3	6	-10.22	<0.05
	4	5	36.86	<0.05
	4	6	15.17	<0.05
	5	6	-21.68	<0.05

NBA Front Court Touches 사후분석

변수명	군집1	군집2	군집2-군집1(평균)	p-value
Front Court Touches	1	2	3.91	<0.05
	1	3	15.58	<0.05
	1	4	-1.39	0.51
	1	5	6.73	<0.05
	1	6	-0.17	0.9
	2	3	11.67	<0.05
	2	4	75.31	<0.05
	2	5	2.82	<0.05
	2	6	-4.08	<0.05
	3	4	-16.98	<0.05
	3	5	-8.84	<0.05
	3	6	-15.76	<0.05
	4	5	8.13	<0.05
	4	6	1.12	0.59
	5	6	-6.91	<0.05

NBA Elbow Touches 사후분석

변수명	군집1	군집2	군집2-군집1(평균)	p-value
Elbow Touches	1	2	-2.81	<0.05
	1	3	0.12	0.9
	1	4	-2.79	<0.05
	1	5	-2.78	<0.05
	1	6	-3.14	<0.05
	2	3	2.92	<0.05
	2	4	0.01	0.9
	2	5	0.02	0.9
	2	6	-0.33	<0.05
	3	4	-2.91	<0.05
	3	5	-2.90	<0.05
	3	6	-3.26	<0.05
	4	5	0.01	0.9
	4	6	-0.35	<0.05
	5	6	-0.36	<0.05

NBA PaintZone Touches 사후분석

변수명	군집1	군집2	군집2-군집1(평균)	p-value
PaintZone Touches	1	2	-5.97	<0.05
	1	3	-2.40	<0.05
	1	4	-5.48	<0.05
	1	5	-5.92	<0.05
	1	6	-6.30	<0.05
	2	3	3.57	<0.05
	2	4	0.49	<0.05
	2	5	0.05	0.9
	2	6	-0.32	0.35
	3	4	-3.07	<0.05
	3	5	-3.51	<0.05
	3	6	-3.90	<0.05
	4	5	-0.44	<0.05
	4	6	-0.82	<0.05
	5	6	-0.38	0.22

NBA Average Second per Touches 사후분석

변수명	군집1	군집2	군집2-군집1(평균)	p-value
Avg Sec per Touches	1	2	0.88	<0.05
	1	3	0.55	<0.05
	1	4	0.13	0.2
	1	5	3.11	<0.05
	1	6	2.46	<0.05
	2	3	-0.32	<0.05
	2	4	-0.75	<0.05
	2	5	2.23	<0.05
	2	6	1.58	<0.05
	3	4	-0.42	<0.05
	3	5	2.56	<0.05
	3	6	1.91	<0.05
	4	5	2.98	<0.05
	4	6	2.33	<0.05
	5	6	-0.65	<0.05

NBA Average Dribble per Touches 사후분석

변수명	군집1	군집2	군집2-군집1(평균)	p-value
Avg Drib per Touches	1	2	1.22	<0.05
	1	3	0.58	<0.05
	1	4	0.34	<0.05
	1	5	3.79	<0.05
	1	6	3.18	<0.05
	2	3	-0.64	<0.05
	2	4	-0.87	<0.05
	2	5	2.57	<0.05
	2	6	1.96	<0.05
	3	4	-0.23	<0.05
	3	5	3.21	<0.05
	3	6	2.60	<0.05
	4	5	3.44	<0.05
	4	6	2.83	<0.05
	5	6	-0.60	<0.05

NBA Shoot% 사후분석

변수명	군집1	군집2	군집2-군집1(평균)	p-value
Shoot%	1	2	-0.05	<0.05
	1	3	-0.01	0.65
	1	4	-0.11	<0.05
	1	5	-0.12	<0.05
	1	6	-0.15	<0.05
	2	3	0.04	<0.05
	2	4	-0.04	<0.05
	2	5	-0.06	<0.05
	2	6	-0.09	<0.05
	3	4	-0.08	<0.05
	3	5	-0.105	<0.05
	3	6	-0.134	<0.05
	4	5	-0.02	0.45
	4	6	-0.04	<0.05
	5	6	-0.03	0.13

NBA Pass% 사후분석

변수명	군집1	군집2	군집2-군집1(평균)	p-value
Pass%	1	2	8.97	<0.05
	1	3	3.42	<0.05
	1	4	13.81	<0.05
	1	5	16.19	<0.05
	1	6	20.93	<0.05
	2	3	-5.54	<0.05
	2	4	4.83	<0.05
	2	5	7.22	<0.05
	2	6	11.96	<0.05
	3	4	10.8	<0.05
	3	5	12.77	<0.05
	3	6	17.51	<0.05
	4	5	2.38	0.13
	4	6	7.12	<0.05
	5	6	4.73	<0.05

NBA Usage Rate 사후분석

변수명	군집1	군집2	군집2-군집1(평균)	p-value
Usage Rate	1	2	2.74	<0.05
	1	3	7.61	<0.05
	1	4	-2.38	<0.05
	1	5	8.83	<0.05
	1	6	0.78	0.48
	2	3	4.87	<0.05
	2	4	-5.12	<0.05
	2	5	6.09	<0.05
	2	6	-1.95	<0.05
	3	4	-9.99	<0.05
	3	5	1.12	0.07
	3	6	-6.82	<0.05
	4	5	11.22	<0.05
	4	6	3.16	<0.05
	5	6	-8.05	<0.05

NBA Restricted Area Frequency 사후분석

변수명	군집1	군집2	군집2-군집1(평균)	p-value
Restriced Area Frequency	1	2	-0.38	<0.05
	1	3	-0.28	<0.05
	1	4	-0.33	<0.05
	1	5	-0.31	<0.05
	1	6	-0.34	<0.05
	2	3	0.09	<0.05
	2	4	0.05	<0.05
	2	5	0.06	<0.05
	2	6	0.03	<0.05
	3	4	-0.04	<0.05
	3	5	-0.02	0.23
	3	6	-0.05	<0.05
	4	5	0.02	0.73
	4	6	-0.01	0.89
	5	6	-0.03	0.23

NBA PaintZone Frequency 사후분석

변수명	군집1	군집2	군집2-군집1(평균)	p-value
PaintZone Frequency	1	2	-0.07	<0.05
	1	3	-0.01	0.73
	1	4	-0.11	<0.05
	1	5	-0.04	<0.05
	1	6	-0.05	<0.05
	2	3	0.07	<0.05
	2	4	-0.03	<0.05
	2	5	0.03	<0.05
	2	6	0.02	<0.05
	3	4	-0.1	<0.05
	3	5	-0.03	<0.05
	3	6	-0.04	<0.05
	4	5	0.07	<0.05
	4	6	0.05	<0.05
	5	6	-0.01	0.52

NBA Middle Range Frequency 사후분석

변수명	군집1	군집2	군집2-군집1(평균)	p-value
Middle Range Frequency	1	2	0.11	<0.05
	1	3	0.13	<0.05
	1	4	-0.01	0.9
	1	5	0.10	<0.05
	1	6	0.07	<0.05
	2	3	0.02	0.08
	2	4	-0.11	<0.05
	2	5	-0.003	0.9
	2	6	-0.03	<0.05
	3	4	-0.14	<0.05
	3	5	-0.03	<0.05
	3	6	-0.06	<0.05
	4	5	0.11	<0.05
	4	6	0.08	<0.05
	5	6	-0.03	0.0504

NBA 3Point Frequnecy 사후분석

변수명	군집1	군집2	군집2-군집1(평균)	p-value
3Point Frequency	1	2	0.35	<0.05
	1	3	0.16	<0.05
	1	4	0.45	<0.05
	1	5	0.25	<0.05
	1	6	0.32	<0.05
	2	3	-0.19	<0.05
	2	4	0.09	<0.05
	2	5	-0.10	<0.05
	2	6	-0.02	0.28
	3	4	0.29	<0.05
	3	5	0.09	<0.05
	3	6	0.14	<0.05
	4	5	-0.09	<0.05
	4	6	-0.12	<0.05
	5	6	0.07	<0.05

NBA Rebound Chance 사후분석

변수명	군집1	군집2	군집2-군집1(평균)	p-value
Rebound Chance	1	2	-3.99	<0.05
	1	3	3.54	<0.05
	1	4	-0.57	0.9
	1	5	-4.43	<0.05
	1	6	-4.18	<0.05
	2	3	7.53	<0.05
	2	4	3.41	<0.05
	2	5	-0.43	0.9
	2	6	-0.19	0.9
	3	4	-4.11	<0.05
	3	5	-7.97	<0.05
	3	6	-7.72	<0.05
	4	5	-3.85	<0.05
	4	6	-3.60	<0.05
	5	6	0.247	0.9

NBA Time of Possession 사후분석

변수명	군집1	군집2	군집2-군집1(평균)	p-value
Time of Possession	1	2	0.57	<0.05
	1	3	1.25	<0.05
	1	4	-0.02	0.9
	1	5	4.93	<0.05
	1	6	2.55	<0.05
	2	3	0.67	<0.05
	2	4	-0.59	<0.05
	2	5	4.36	<0.05
	2	6	1.98	<0.05
	3	4	-1.27	<0.05
	3	5	3.68	<0.05
	3	6	1.30	<0.05
	4	5	4.95	<0.05
	4	6	2.57	<0.05
	5	6	-2.37	<0.05

Redefining Positions for the Modern Basketball with Machine Learning

Kim Nael

Major of Sport Recording and Analysis

Graduate School of Records, Archives & Information Science, Myongji University

Directed by Professor Kim Joohak

The play style of modern professional basketball players is changing rapidly like the fast pace of a basketball game. Playstyles refer to the player's role and position.. Position in the team sports means the role of players in the team and is an essential factor to consider in constructing team tactics. But the five traditional basketball positions (PG, SG, SF, PF, C) do not have an official objective standard. .In addition, Through traditional position, defining the positions of modern professional basketball players that are currently changing is even more limited. Therefore, in this study, KBL and NBA were classified as new clusters using that one of the unsupervised learning methods is k-means cluster analysis for each league. Each cluster was redefined with a new name suitable for the characteristics of the cluster. In addition, a model for classifying new positions defined through cluster analysis was developed using random forest among ensemble techniques using bagging. The results are summarized as follows.

First, the data collected by each league consists of 364 people and 10 variables in KBL, 899 people and 16 variables in NBA. In order to refine the collected data, missing values and outliers were replaced by the average of each variable. In addition, the range was reduced through log conversion. And through the Min-Max

normalization, the units of each variable were converted between 0 and 1.

Second, data for each league were clustered through the k-means method. Before clustering, the k value of KBL was set to 4 and the k value of NBA was set to 6 through the elbow technique and silhouette technique. The characteristics of each cluster were identified through descriptive statistics for the clusters classified by each k value. And clusters were newly defined as appropriate positions. The four clusters of KBL were defined as Main Ball Handler, Linker, Finisher, and 3&D. The six clusters of NBA were defined as Long Ranger and Traditional Bigman along with the four positions of KBL. Through the Bonferroni test, it was verified that there was a statistically significant difference between each position ($p<0.05$).

Third, all datasets to which the labeled position was added were divided into 6:4 learning data and test data. The redefined position for each league was used as a predict variable of the random forest model. The remaining variables were used as input data for model creation. The initial model showed 91.0% performance for KBL and 88.0% performance for NBA. However, hyper parameters were tuned through the Random Search method as an overfitting phenomenon for learning data. The performance of the tuned final model improved to 94.3% performance ofr KBL and 92.6% performace for NBA. In addition, in the final model, Shoot% in KBL and Average Dribble per Touch in NBA showed the highest importance of variables.

This study does not consider subjective things. However, this study is meaningful because it proposes defining basketball positions on the same basis in a scientific and objective approach by classifying players according to the importance of variables. In addition, it is meaningful because it attempted to classify play characteristics using independent variables in the play results.

Keyword
Basketball, Position, Characteristic, Machine Learning, Clustring, Random Forest