|  |
| --- |
| 석사학위 논문계획서 |
| 군집분석 을 통한 K리그 포지션 세분화 |
| 명지대학교 기록정보과학전문대학원  스포츠기록분석전공 |
|  |
| 문 창 훈        지도교수 신 문 선 |
| 2022年 11月 |

|  |
| --- |
|  |
| 군집분석 을 통한 K리그 포지션 세분화 |
|  |
| 이 논문을 석사학위 논문계획서로 제출함 |
|  |
| 2022년 11월 |
|  |
| 명지대학교 기록정보과학전문대학원  스포츠기록분석전공  문 창 훈 |

목차

**제1장 서론**

제1절연구의필요성5

제2절연구목적5

제 3절 연구 문제5

제4절연구의제한점5

제 5절 용어 정의5

**제2장 이론적 배경** **4**

제1절축구에서의포메이션4

제2절K리그데이터기록실5

제 3절 국내 프로축구리그 현황5

제 4절 군집 분석5

제 5절 보루타 알고리즘6

**제3장 연구 방법** **4**

제1절 연구 대상4

제 2절 연구 방법5

제 3절 자료 처리5

제 4절 연구 절차5

**참고 문헌** **4**

**그림 목차** **4**

그림1 군집분석4

그림 2 4

그림 34

그림 44

**표 목차** **4**

표<1> 예상 수집 데이터4

표<2> 수집 공격 요인4

표<3> 수집 패스 요인4

표<4> 수집 수비 요인4

표<5> 수집된 데이터의 특성4

표<6> 연구설계4

**제1장 서론**

**1.연구의 필요성**

 축구는 전세계에 210개의 가맹국이 국제축구연맹(FIFA: Federation International de Football Association)에 이르는 대형 스포츠로 세계적으로 보편화된 스포츠이다. 2018년 러시아 월드컵의 경우 FIFA의 보고서에 따르면 지난 6∼7월 열린 러시아 월드컵 경기를 TV나 디지털 미디어, 거리 응원을 통해 일부라도 시청한 사람은 35억7천200만 명에 달했다. 이는 전 세계 4세 이상 인구의 51.%에 해당하는 것으로, 역대 최다 수준이다. 월드컵 경기를 30분 이상 본 사람은 24억9천만 명으로 4년 전 19억5천만 명보다 늘어난 수치이다. 또한 월드컵 이외에도 미디어의 발전으로 인해 잉글랜드의 프리미어리그, 스페인의 프리메라리가, 독일의 분데스리가, 이탈리아의 세리아 등, 해외축구를 실시간으로 접하고 손흥민, 김민재, 황인범 등의 해외파의 활약은 축구에 대한 국민들의 관심을 불러일으키기에 충분하다. 또한 벤투가 이끄는 대한민국 대표팀은 아시아 2차 지역 예선에서 5승1무, 최종예선에서 7승 2무 1패의 좋은 성적으로 본선행을 결정 지으며 국내 축구팬들의 기대를 받고 있다.

축구는 통계적인 지표가 부족하여, 데이터 분석을 접목시키기에 어려움이 있었다 하지만, 최근에 데이터 수집 및 데이터 분석 기술의 발전과 축구 수요의 증가로 데이터 분석을 접목하려는 시도가 증가했다. 축구시장이 수요가 많고, 급격하게 성장하고 있음에 따라서, 축구 시장의 성장으로 인한 축구 데이터의 분석 관련 연구는 소셜 네트워크 분석(Social Network Analysis)을 기반으로 한 패싱 네트워크 (Passing Network), 토너먼트 우승자 예측 등의 연구가 진행되고 있다. 또한, 독일 축구 국가대표팀은 SAP Match Insight를 활용하여 실시간으로 데이터를 분석한 뒤, 보드진의 의사결정에 도움을 주었다. 이를 활용한 독일 축구 국가대표팀은 2014브라질 월드컵에서 우승을 하는 성과를 달성 하였다. 또한 SAP Match Insight는 선수들에게 웨어러블 센서(Wearable Sensor)를 부착시킨 후, 웨어러블 센서를 통해 선수의 호흡, 심박 수 및 활동량 데이터 등을 수집했다. 데이터는 1분당 12,000개 이상의 데이터를 수집했으며, 수집된 데이터는 트레이닝 센터에 설치 된 무선 수신기로 전송된 후 대량의 데이터 분석 플랫폼인 SAP HANA를 통해 데이터 분석을 실시했다.처럼 축구 시장은 앞으로도 지속적으로 성장할 것이며, 각 팀의 목적을 위해서 다양한 데이터 분석의 필요성이 대두되고 있다.

축구 경기에서 포지션은 팀 전력과 전술을 구성하는 데에 필수적으로 고려해야할 요인이다.전통적으로 선수의 포지션은 포워드(FW),미드필더(MF),수비수(DF),골키퍼(GK)의 4개의 포지션으로 분류되어 왔다. 그러나 현대 프루 축구에서는 모든 포지션의 선수들에게 포지션을 넘나드는 다양한 역할 수행을 요구하고 있습니다. 이러한 역할 수행의 요구는 과거 전통적 포지션의 경계를 무너뜨리게 되었습니다. 따라서 단순히 선수를 FW,MF,DF,GK로 나누는 것은 더이상 그 선수의 특성을 정확히 반영한다고 할 수 없다.

축구의 경우 야구, 농구와 달리 데이터를 활용한 경기 분석,경기에 대한 이해와 활용 접근이 어려울 뿐더러 대중들에게 공개 된 시점이 얼마 되지 않아 데이터 적인 시각에서 축구를 바라보는 것이 쉽지 않다.

김윤후(2020)은 해외 축구 통계 사이트인 후스코어드의 데이터 요인들을 사용하여 군집 분석을 사용하여 포지션을 세분화하고 약팀과 강팀 상대시에 어떤 역할을 사용하여 전술을 사용하였는지 규명하고자 하였다.

김도민(2022)은 1011시즌부터 18/19시즌까지 9개의 시즌의 잉글랜드 프리미어리그, 스페인 프리라메리가, 독일 분데스리가의 공식기록 자료들의 군집 분석을 통해 유럽 빅 리그 프로 축구팀의 공식기록이 군집 분석이 가능한지와 그 특성을 파악하고자 하였다.

김도민과 최준형(2021)은 공식기록을 바탕으로 경기력을 분석하기 위해 군집분석을 통해 랭킹 별로 군집이 가능한지와 각 랭킹 별 특성을 파악하였다.

이승박(2021)은 잉글랜드 프리미어리그 경기기록과 배당률 자료를 활용하여 분류 머

신러닝 알고리즘 기반 축구 승패예측모형을 탐색 및 비교하였다.

앞서 언급한 바와 같이 선행 연구에서는 유럽의 데이터, 유럽의 대회를 중심으로 경기를 분석하거나 팀 단위의 기록을 분석하여 군집 분석하는 연구 들이 주를 이루고 있다.

때문에 유럽의 대회나 유럽의 데이터를 사용하여 연구하는 것이 아닌 K리그 공식 데이터를 사용하여 선수들에 대해 연구하는 것은 K리그 선수들의 유형을 파악하고 K리그 공식 데이터의 활용성 가능성에 대한 설명을 하기 위해 필요할 것이다.

그러므로 이 연구는 K리그 공식 데이터를 활용하여 선수들을 군집분석하고 선수들을 군집된 특성에 따라 세분화 하여, 팀별 선수 보유 유형에는 그 차이가 있는지 알아보고자 한다. 이를 통해 K리그 공식 데이터의 활용 가능성을 파악하고 포지션을 분석하는 하나의 기초 연구가 될 것이다.

**2. 연구 목적**

이 연구는 K리그 데이터 기록실에서 제공 하는 경기력 요인을 보루타 알고리즘을 통하여 각각의 포지션(FW,MF,DF)별로 특징을 선택하고, 그 특징을 기반으로하여 포지션을 군집분석하여 군집된 집단의 특징을 알아보고자 한다.

**3 연구 문제**

1) K리그 데이터 기록실의 32개의 요인을 보루타 알고리즘을 통해 각각의 포지션별 특징을 선택할 수 있을 것이다.

2) 보루타알고리즘을통해 나누어진 포지션별 특징을 통해 각각의 포지션(FW,MF,DF)별로 군집분석이 가능할 것이다.

2.1)선별된 요인을 통해 군집분석이 가능할 것이다.

2.2) 군집별 특성을 통해 각각의 포지션을 세분화 하여 정의할 수 있을 것이다.

3) 머신러닝을 통해

**4 연구의 제한점**

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

모든 포지션 중 GK를 제외한 3개의 포지션 FW, MF, DF의 선수들을 대상으로 한다.

**2장 이론적배경**

**1. 축구에서의 포메이션**

축구에서 포지션은 11명의 선수가 팀의 전술에 따라 경기장 내에 배치된 위 치를 의미한다. 그림 2.1은 축구 포지션 및 경기장에 대한 설명이다. 포지션은 크게 공격수, 미드필더, 수비수, 골키퍼로 분류된다.

공격수는 어태킹 써드(Attacking Third) 지역에서 득점을 목표로 하는 포 지션이다.

미드필더는 미드 써드(Middle Third) 지역에서 공격수와 수비수 사이의 연결고리를 하는 역할을 맡은 선수를 의미한다. 미드필더는 팀이 공격 시에 공격수와 함께 공격에 가담해하 하고, 수비 시에 수비수들과 함께 수비에도 기여해야 한다. 즉, 경기장 내에서 가장 많은 활동량을 가진 포지션이다.

수비수는 위치에 따 라 중앙에 위치한 센터백(CB), 측면에 위치한 사이드백(SB)가 있다. 센터백은 주로 페널티 에어리어 안에서 상대 공격을 막아내는 역할을 한다. 공중볼 경합 능력, 가로채기 능력 등의 수비적 능력이 우선시된다. 사이드백의 주된 역할은 상대 공격수들이 사이드에서 크로스 공격 시도를 저지하는 역할이다. 센터백보다 수비 범위가 넓으며, 때때로 공격 시 팀에 공격가담을 함으로 팀에 기여하는 경우가 있다. 그에 따라 사이드백은 수비력은 물론 스피드와 크로스 능력이 요구된다

**2. K리그 데이터 기록실**

K리그 데이터 포털은 2018년 2월부터 8월까지 약 6개월의 개발 기간을 거쳐 개발하여 축구 팬들에게 K리그에서 발생하는 기록들에 대한 정보를 제공 하고 있다. 해외의 경우 Whoscored(이하 후스코어드)라는 옵타스포츠의 통계를 인용한 사이트를 통해 영국,스페인,프랑스,독일,이탈리아 등 500개 이상의 대회의 15,000개 이상의 팀, 250,000명 이상의 선수가 기재되어 있는 세계 최대의 축구 전문 통계 사이트를 통계 사이트를 통해서 그 기록을 살펴 볼 수 있지만 K리그의 경우 후스코어드에 등록은 되어 있으나 기록이 누락되거나 기재되어 있지 않는 경우가 많아 K리그에 대한 데이터를 수집하고 활용하기에는 어려운 상황이었다 하지만 K리그 데이터 포털이 개방되면서 한국 K리그 팬들은 보다 쉽게 K리그의 기록에 접근 할 수 있게 되었다.

K리그 데이터 포털에서 제공하는 데이터는 다음과 같다. 메인 홈페이지를 통해 클럽별 관중수 추이, 오늘의 경기 기록 결과, 방송 중계,각 팀의 오늘의 포메이션, 경기의 관전포인트를 제공하며, 매치 센터를 통해 팀 스텟, 개인 스텟, 지난 경기 결과, 라이브 이벤트 차트 보드,패스 메트릭스를 제공하고 있으며, 데이터 센터를 통해 선수의 경기 별 기록, 시즌 별 기록 을 제공하고 있다.

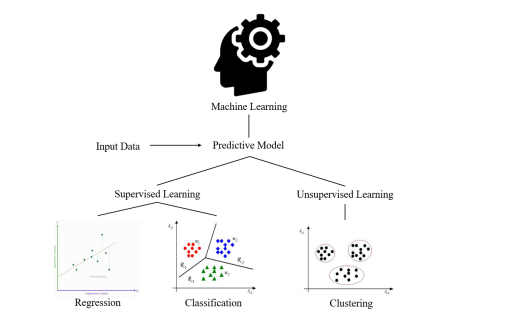
**3. 국내 프로축구리그 현황**

1983년 ‘수퍼리그’라는 이름으로 출범한 K리그는 국내 축구팬에게 즐거움을 선 사하였고 한국 축구의 수준을 한 단계 끌어올리는데 기여하였다. 수퍼리그는 프로와 아마추어를 통합하는 성인 축구 리그로 프로와 아마추어의 공동 발전을 이루고자 수 준 높은 경기를 통해 붐을 일으키면서 본격적인 프로시대를 마련하는 기획이었다. 처 음 출범 당시 프로팀은 할렐루야, 유공 외, 실업리그의 포항제철, 국민은행, 대우 등 상위 랭커 3팀을 묶어 연고제에 의한 초청과 원정 경기를 통해 한국 축구의 정상을 가린다는 취지였다. 한국 프로축구는 1983년 세미프로 형태로 탄생하여, 2개의 프로팀과 3개의 실업팀 으로 초창기 프로리그를 이끌어 갔다. 1984년에는 대우, 포항제철, 현대 등 3개의 실업팀이 프로로 전환하고, 럭키금성이 창단되면서 프로축구의 열기는 더해졌다. 이러한 배경과 뜨거운 관심 속에 1954년 월드컵 본선 진출 이후 32년 만에 1986년 멕시코 월 드컵에 진출하게 되었다. 1987년 아마추어 팀들을 제외한 프로 5개 팀(대우, 포항제철, 유공, 현대, LG)만으로 진행되는 순수 프로리그로 재출범하였고 이어, 89년에는 일화 천마가 추가로 창단되면서 1993년까지 6개 팀으로 운영되었다. 월드컵 본선 연속진출, 2002 월드컵 유치운동 등 축 구 활성화에 힘을 쏟은 한국 프로축구는 1994년에 전북 다이노스와 전남드래곤즈, 1995 년 수원 삼성의 잇따른 창단으로 크게 발전하게 되었다. 그 이후에 1997년 10번째 구단 인 대전 시티즌이 창단되었고, 2003년 대구FC 창단과 상무의 참여, 2004년에는 인천 유 나이티드가 창단되었다. 이어서 2005년 경남FC 창단, 2009년 강원FC 창단, 2011년 광주 FC가 창단하였으며 2012년 이후에는 마지막까지 결과를 예측하기 힘든 경기를 보여주며 리그의 흥행을 도모하며 팬들을 유입시키고자 승강제 및 스플릿 제도를 도입하였다. 2015년 서울 이랜드가 창단하며 2016년에 K리그 1부 리그인 K League Classic은 12팀, 2 부 리그인 K League Challenge에는 11팀으로 진행되며 리그 운영이 순탄하게 흘러가는 듯 보였지만 구단 운영의 재정적 어려움으로 인해 충주 험멜과 고양 자이크로 2팀이 시 즌 종료 후 해체가 되었다. 2017년 안산 그리너스FC가 창단하며 이후로는 K리그 1부 리 그는 12팀, 2부 리그는 10팀 체제로 나뉘어 지금까지 유지되고 있다. 결과적으로 최초 5 개 팀으로 시작한 한국 프로축구는 2022년 현재 총 22팀으로 구성되어 운영되고 있다.

**4. 군집분석**

 데이터를 사용하여 컴퓨터가 학습하는 방법인 기계 학습은 지도 학습, 비지 도 학습으로 나뉜다.

지도 학습은 데이터에 라벨이 정해져 있는 상태에서 학습하는 방법이고, 비지도 학습은 학습데이터에 라벨이 정해져 있지 않은 상 태에서 컴퓨터가 학습하는 방법이다 아래의 그림은 지도학습과 비지도학 습을 분류하여 기계학습을 설명하는 그림이다.



<그림 1> 군집 분석

 본 연구에서 사용된 축구선수 데이터는 선수들의 실제 기록들을 바탕으로 분석하기 때문에, 각 데이터에 주어진 라벨이 없다. 그래서 본 연구에서는 비지도 학습 중 하나인 군집을 사용하여 선수들의 군집을 분류하였다. 집 분석은 크게 n개의 군집으로 시작하여 점차 군집의 개수를 줄여나가는 방식인 계층적 군집분석 (Hierachical Clustering)과 n개의 개체를 K개의 군집으로 나눌 수 있는 모든 방법을 고려하여 군집을 최적화하는 분할적 군집분석(Partitional Clustering) 이 있다. 축구 분야에서는 경기 결과를 예측하고 클러스터링을 통해 예측 결 과를 검증하는 연구와 축구의 시공간 데이터를 군집화 하여 패스 패턴에 대한 연구 등 다양한 연구들이 진행되었다 하지만, 선수의 세부 역할을 정 의한 선행 연구는 거의 이루어지지 않았다. 본 연구에서는 K-평균 군집 분석 기법을 적용하여, 특정 선수들의 세부 역할을 검증했다. 특정 선수의 감독들과 축구 기자들은 같은 포지션이더라도 선수들의 세부 역할에 대해 비공식적으로 인지하고 있다. K-평균 군집 분석의 알고리즘은 먼저, K개의 군집의 수를 정의해야 한다. 군집마다 최초 임의의 점을 군집의 중심점으로 설정한다. 그 후, 각 관찰 데이터를 K개의 중심점과 비교하여 가장 가까운 점의 군집으로 분류한다. 모든 관찰 데이터를 K개의 군집으로 분류 후, 각 군집의 중심점을 재계산한다. 이와 같은 단계를 반복하며, 관찰 데이터들의 소속된 군집이 더 이상 변경되지 않으면 알고리즘을 종료한다

**5.** **보루타 알고리즘**

보루타 알고리즘은 랜덤 포레스트(Random Forest)의 Z값(Z-Score)을 활용한 특징 엔지니어링 기법으로서 다양성이 높고, 예측력이 높은 특징의 알고리즘이다. 임의의 다양한 예측 결과를 바탕으로 특징을 선택하기 때문에, 안정성과 예측력을 보장한다. 보루타 알고리즘은 NBA (National Basketball Association)의 결과를 예측하기 위한 연구에서 선수들의 기록 데이터의 중요한 변수를 추출을 위해 사용되었고, 아이스 하키에서 관객들의 수요 예측을 위해 변수 추출 시 사용되었다. 또한 축구 분야에서는 MLS(Major League Soccer) 관객들의 수요를 예측하기 위하여 경기장 상태에 관한 변수에 보루타 알고리즘을 적용하여 연구를 진행했다. 본 연구에서는 기존 연구들과 달리, 축구 선수 데이터에 보루타 알고리즘을 적용하여 중요한 변수를 선택했다. 보루타 알고리즘의 작동방식은 먼저, 모든 특징을 복사하여 새로운 칼럼인 쉐도우 특징(Shadow Feature)을 생성한다. 그 후 쉐도우 특징들을 랜덤하게 섞는다. 쉐도우 특징들에 대해서 랜덤 포레스트를 실시하고, 얻어진 Z값 중 가장 큰 값인 MSZA(Max Z-Score among shadow feature)를 계산한다. 또한, 기존의 특징들에 대해서도 랜덤 포레스트를 실행하여 Z값을 추출한다. 추출된 Z값이 MSZA보다 큰 경우 해당 특징 변수를 중요한 변수로 선택한다

**제3장 연구방법**

**1. 연구대상**

2021K리그1 12개팀의 1라운드부터 38라운드에 걸쳐 치러진 정규 시즌 경기에서 출전한 골키퍼를 제외한 모든 포지션의 선수 중 정규 시즌 경기 출장 시간이 10%이상인 모든 선수를 대상으로 한다.

1.1)문헌 연구를 통한 요인 추출: 축구경기의 경기력과 관련되어 구조화했다고 평가된 연구를 채택하고 이를 토대로 축구선수의 경기력에 관련된 요인들을 추출하였다. 국내, 외 전문서적은 축구전문 단체에 의해 검증절차를 거친 후 발행된 서적으로 제한하였다. 인터넷관련 자료는 국내 5개 스포츠신문과 대한축구 협회의 홈페이지를 기술된 내용을 파악하였으며, 선행연구논문은 경기 내용분석 관련논문, 경기기록 분석논문, 축구 경기력 관련논문, 기타경기종목 지수산출 관련논문 등을 중점적으로 다룬 학위논문 및 학회지를 중심으로 요인을 탐색하였다

|  |  |
| --- | --- |
| 연구자 | 주요내용 |
| 박종준(1999) | 득점관련연구, 공격방향,득점지역 |
| 엄한주(1999) | 드리블, 슛,패스,킥,스피드,축구기능측정관련 |
| 신문선 | 프로축구선수연봉산정관련, 경력,연령,출장수,득점,방어,교체수,도움,팀기여도,팀 성적 |
| Hughes와Frank(1997) | 부호화분석관련, 패스,슛,크로스,코너킥,스로인,패널티킥,프리킥,선수,위치,행동,시간 |
| 홍성진(2010) | 포지션별경기력요인관련,  득점,도움,슈팅,패스,드리블,태클,인터셉트,클리어링,블락,경합,세컨볼,크로스,유효슈팅률,어시스트율,패스성공률,슛블록률,패스차단률,스틸률 |
| 홍성진(2017) | 축구경기력평가요인관련, 득점,도움,슈팅,패스,드리블,태클,인터셉트,클리어링,블락,경합,세컨볼 |
| 한필수,백은호,장상영(2010) | 부호화분석관련, 기입방법관련  4개의 범주를 통한 81개의요인에 대한 가산점 부여 |
| 유강원(2013) | 축구 경기력 평가요인관련  슈팅, 패스,드리블,차단,태클,파울 |

**2. 연구방법**

1) 데이터 수집

이 논문을 위해 2021 K리그 시즌의 대상 선수의 데이터를 수집한다. 데이터 수집을 위하여 K리그 공식 기록관리 사이트인 data.kleague.com을 통해 데이터를 수집한다. 수집된 데이터는 2021 K리그 전체 선수 870명 중 정규 시즌 경기 출장 시간이 10%이상인 선수와 요인 30개로 이루어진 데이터 9960개를 수집하였다.

|  |  |
| --- | --- |
| **선수의 수** | 332 명 |
| **요인의 수** | 30개 |
| **데이터의 수** | 9960 개 |

<표 1> 예상 수집 데이터

2) 수집 요인

이 논문을 수행하기 위해 수집 요인은 다음과 같다.

|  |  |  |  |
| --- | --- | --- | --- |
| 분류 | 요인 | 정의 | 유형 |
| 공격 |  |  |  |
|  | 득점 | 상대 골대에 공을 넣어 점수를 얻는것 | 연속형 |
|  | 도움 | 득점을 도운 기록 | 연속형 |
|  | 유효슈팅 | 슈팅중 공대를 향한 슈팅 | 연속형 |
|  | PA내 슈팅 | 패널티 에어리어 내에서 기록한 슈팅 | 연속형 |
|  | PA외 슈팅 | 패널티 에어리어 외에서 기록한 슈팅 | 연속형 |
|  | 드리블 | 혼자 공을 가지고 상대 선수를 돌파하는 행위 | 연속형 |
|  | 공격지역패스 | 상대 진영에 도달한 패스 | 연속형 |
|  | 크로스 | 상대팀 진영에서 패널티 박스로 길게 띄워주는 패스 | 연속형 |
|  | 탈압박 | 공을 소유한 선수가 상대 선수의 압박에서 벗어나 소유권을 계속 유지하는 행위 | 연속형 |
|  | 피파울 | 상대 선수로부터 파울을 당하여 프리킥 또는 패널티킥을 얻은 경우 | 연속형 |

<표 2> 수집 공격 변수

|  |  |  |  |
| --- | --- | --- | --- |
| 분류 | 요인 | 정의 | 유형 |
| 패스 |  |  |  |
|  | 프리킥 | 축구 규칙에 따른 프리킥 | 연속형 |
|  | 코너킥 | 축구 규칙에 따른 코너킥 | 연속형 |
|  | 후방패스 | 자기측 진영쪽으로 향한 패스 | 연속형 |
|  | 전방패스 | 상대 진영쪽으로 향한 패스 | 연속형 |
|  | 횡패스 | 패스 방향이 좌우 측면 120도 범위 내로 향한 패스 | 연속형 |
|  | 중앙지역패스 | 전체 경기장을 3등분 햇을때, 중원징영에서 시작된 패스 | 연속형 |
|  | 중거리패스 | 패스거리가 15미터 이상 30미터 미만의 패스 | 연속형 |
|  | 롱패스 | 패스 거리가 30미터 이상인 패스 | 연속형 |
|  | 숏패스 | 패스 거리가 15미터 미만의 패스 | 연속형 |
|  | 키패스 | 슈팅으로 연결된 마지막 패스 | 연속형 |

<표 3> 수집 패스 변수

|  |  |  |  |
| --- | --- | --- | --- |
| 분류 | 요인 | 정의 | 유형 |
| 수비 |  |  |  |
|  | 태클 | 몸을 사용하여 상대 공격을 저지하는 수비 기술 | 연속형 |
|  | 경합(지상) | 지상에서 상대방과 볼의 소유권을 경합하는 행위 | 연속형 |
|  | 경합(공중) | 공중에서 상대방과 볼의 소유권을 경합하는 행위 | 연속형 |
|  | 클리어링 | 상대가 공격중인 상황에서 볼을 걷어내는 행위 | 연속형 |
|  | 인터셉트 | 상대방의 공을 가로채어 볼 소유권을 획득하는 행위 | 연속형 |
|  | 차단 | 상대 선수가 소유한 볼이나 패스를 막아 루즈볼로 만드는 수비 행위 | 연속형 |
|  | 획득 | 루즈볼이 된 공의 소유권을 가져오는 행위 | 연속형 |
|  | 블락 | 필드플레이어가 상대 슈팅을 막아내는 수비 행위 | 연속형 |
|  | 수비진형패스 | 우리 진영으로 도달한 패스 | 연속형 |
|  | 스로인 | 축구 규칙에 따른 스로인 | 연속형 |

<표 4> 수집 수비 변수

3) 수집 데이터의 특성



<표 5> 수집 데이터의 특성

이 논문에서 수집되는 데이터는 한 선수에 대한 30가지의 변수의 데이터로 공격 요인 패스 요인 수비 요인을 포함하고 있다. 한 선수의 슈팅 횟수, 패스 방향, 수비 빈도 수 등에 대한 다양한 데이터들이 수집된다. 수집된 데이터는 그림과 같이 한명의 선수에 대한 요인별 데이터가 수집되는 것을 확인 할 수 있다.

4)데이터 전처리

수집된 데이터는 엑셀 파일로 정리하여 저장 및 관리하고, GK를 제외한 모든 포지션에 대한 데이터를 저장한다. 데이터의 결측 치는 평균 값으로 대체하고 각 변수들의 단위와 범위를 LOG기법을 활용하여 스케일링하고 MIN-MAX기법을 활용하여 정규화 하는 과정을 거친다.

**3. 자료 처리**

연구 목적을 달성하기 위한 자료처리 방법은 파이썬과 SPSS를 사용하였다.

첫 번째 연구는 문제 해결을 위해 자료처리 방법은 파이썬과 SPSS를 사용하여 비모수 통계기법 중 하나인 군집분석 K-means를 사용하여 군집이 가능한지 알아보고 각 군집의 차이, 특성을 파악하기 위해 일원 분산 분석ANOVA를 사용하고 유의수준은 0.05로 설정하였다. 또한 초기의 객관적인 최적의 초기 군집수 K를 선정하기 위해서 엘보우 기법을 사용하여 타당한 초기 군집수를 결정하였다. 군집 별로 특성에 맞는 포지션으로 선수를 재분류 한다.

두 번째 연구 문제인 1차 분석으로 나누어진 군집별로 다시 요인을 선별하여 군집을 진행하면 세분화를 위해서 랜덤포레스트를 기반으로 하는 보루타 알고리즘을 사용하여 각 군집 별 군집에 형성을 주는 요인을 파악한 후 가장 큰 영향을 주는 요인을 제외 하고 선별된 요인으로 다시 재 군집한다. 각 군집의 차이, 특성을 파악하기 위해 분산분석ANOVA를 사용하고 유의수준은 0.05로 설정하였다. 또한 초기의 객관적인 최적의 초기 군집수 K를 선정하기 위해서 엘보우 기법을 사용하여 타당한 초기 군집수를 결정하였다. 군집 별로 특성에 맞는 세분화된 이름을 정의한다.

세번째 연구 문제인 세분화된 포지션을 이용하여 팀별로 나누었을 때 팀 간의 차이가 존재할 것이다를 해결하기 위해서 세분화된 포지션으로 선수를 분류하고, 상 하위 팀별 분산분석 ANOVA를 사용하고 유의수준은 0,05로 설정하였다.

**4. 연구절차**

이 논문을 수행하기 위한 열구절차는 아래와 같다.

|  |  |  |
| --- | --- | --- |
| 서론 | 연구설계 | 2M |

|  |  |  |
| --- | --- | --- |
| 이론적 배경 | 축구의 포메이션  K리그 데이터 기록실  국내 프로축구 현황  군집분석  보루타 알고리즘 | 1M |

|  |  |  |
| --- | --- | --- |
| 연구절차 | 연구 대상 선정 | 3M |
| 2021 K리그 선수 공식 기록 |
| 데이터 수집 및 전처리 |
| 1. K리그 공식 사이트 K 리그 데이터 기록실을 통한 요인 수집  2. 선행연구 고찰을 통한 요인 통합, 정제  3. 데이터의 결측치 처리  4. 데이터의 단위 통합, 스케일링, 정규화 |
| 자료 처리 |
| 1. 군집분석을 통한 선수의 포지션 재분류  -기술 통계 분석, ANOVA,시각화  2. 보루타 알고리즘을 통한 군집 결정 요인 선정  3. 결정 요인을 통한 군집 분석 및 포지션 세분화 정의  - 기술 통계 분석, ANOVA,시각화  4. 팀별 세분화된 포지션 보유 빈도 경향 확인  -기술 통계 분석, ANOVA |

|  |  |  |
| --- | --- | --- |
| 결과 | K리그 공식 데이터를 통한 포지션 재정의,세분화, 차이분석, 팀별 분석 | 1M |

|  |  |  |
| --- | --- | --- |
| 결론 | 결론 및 논의 | 1M |

<표 6> 연구설계

**참고문헌**

한재수 (2017), 「축구경기의 공격 효율성 지수 개발과 적용 : K리그 클래식을 중심 으로」, 명지대학교 기록정보과학전문대학원 석사학위 논문.

황정욱 (2012), 「K-리그, LFP, EPL의 공격지역 내 경기내용 비교 분석」, 명지대학 교 기록정보과학전문대학원 석사학위 논문.

홍성진 (2009), 「축구 선수의 경기력 평가를 위한 요인 및 규준 개발」, 명지대학교 기록과학대학원 박사학위 논문

홍성진(2017)- 계층적분석방법을 통한 축구 경기력 평가 요인 가중치 산출: 한국체육측정평가학회지. 제19권 1호 2017,1-12

최경호(2022)- 군집분석을 이용한 K리그1 경기력 요인 탐색: 한국체육과학회지, 2022, 제31권 제2호, PP.699~709

김종원(2021)- 군집분석을 통한 K리그 축구팀 플레이스타일 분류: 한국체육측정평가학회지. 제23권 1호, 2021, 1-9

김주학(2007)- 축구 팀 경기력 평가를 위한 내용분석: 한국체육학회지, 2007년, 제46권 제2호, 201-211

홍성진(2018)- 축구 경기력 요인의 포지션별 차이분석과 가중치 적용: 한국체육측정평가학회지. 제20권 4호, 2018, 175-184

홍성진(2010)-축구 선수의 포지션별 평가를 위한 경기력 결정요인 규준 개발: 체육과학연구. 2010, 제21권, 제2호, 1172-1182

홍성진(2017)-축구 기술 요인의 포지션별 상대적 중요도: 한국체육측정평가학회지. 제19권 4호, 89-98

김나엘(2022)-머신 러닝을 통한 현대 농구 포지션 재정의: 명지대학교 기록정보과학전문대학원 석사학위 논문

김윤후(2002)-빅데이터 분석을 기반으로 한 축구 전술 및 포메이션 패턴 분석: 중앙대학교 대학원 석사학위 논문

김도민, 최형준 (2019). 잉글랜드 프리미어리그 공식기록을 통한 경기력의 군집분석. 한국체육과학회지, 28(2), 1237-1245

김도민,최형준(2021). 군집분석을 통한 K리그 축구팀 플레이스타일 분류.

한국체육측정평가학회지. 제23권 1호, 2021, 1-9