# 데이터 마이닝\_동양 vs 서양

작성자: 2091104 이찬희

이 수행 과제를 통해 단순한 서양인과 동양인의 신체적 차이 뿐만 아니라 기술적 기록의 차이를 비교하여 MLB 선수와 KBO 선수들 사이의 수치적 비교를 통해 끝내 유사도를 비교하는 것이 목적이다.  
이를 통해, 서양인과 동양인의 단순한 육안적 차이 뿐만이 아닌 실질적 수치적 차이를 정량화하여 KBO만의 맞춤형 전략의 수립 필요성에 대해 강조할 수 있다.

## 1. 실험 환경 및 데이터

본 프로젝트는 Google Colab에서 1차적으로 수행되었으며, 이후 Jupyter Notebook으로 후속 분석을 진행하였다.  
이는 개인 노트북에서의 성능 제약으로 인한 처리 효율성 때문이다.  
  
사용한 데이터:  
- kbo\_dataset\_2018\_2024.csv (KBO 국내 선수 기록)  
- batting\_2023\_data.csv (MLB 선수 기록)  
- baseball\_players.csv (MLB 신체 정보)  
- kbo 타자 신체정보.xlsx (KBO 신체 정보)  
- kbo\_batter\_physical\_data\_cleaned.csv (엑셀 정제 버전)  
  
2023년 데이터를 중심으로 비교를 수행하였다. 이는 데이터 양과 신뢰성 측면에서 가장 적절하다고 판단되었기 때문이다.

## 2. 데이터 병합 및 전처리

- MLB 기록 데이터와 신체 정보 병합 (자료2 + 자료3)  
- KBO 기록 데이터와 신체 정보 병합 (자료1 + 자료5)  
- 공통 컬럼 정리 및 리그 구분 컬럼 추가  
- 2023년도 기록 필터링

## 3. 기술통계 및 시각화

KBO와 MLB의 주요 기록 지표 평균 및 표준편차를 비교한 결과, MLB는 평균값이 높고 표준편차는 작아 성과가 안정적임을 확인하였다.  
KBO는 전반적으로 분산이 크고 평균값이 낮아 실력의 일관성보다 다양성이 특징임을 시사하였다.

## 4. 주요 분석 실험

기술력 비교:

- PCA 분석: OPS, HR, BB 등 직접지표 기반 주성분 분석  
- t-검정: OPS, OBP, SLG, HR 등에서 통계적으로 유의한 평균 차이 확인 (p < 0.0001)  
- RandomForest 분류: 정확도 84%, 주요 변수는 HR, SO, BB

성향 분류:

- KMeans 군집화 (n=4): MLB는 장타 중심 클러스터에 집중, KBO는 다양성이 높은 군집에 분포

리그 스타일 해석:

- 전략적 지표와 간접 지표 평균 비교  
- KBO는 희생타, 도루 중심 / MLB는 장타, 출루 중심

## 5. 결론

이 분석을 통해 KBO와 MLB는 단순히 실력 차이만 존재하는 것이 아닌, 리그 구조와 선수군 유형에서도 뚜렷한 차이를 보이고 있음을 확인하였다.  
MLB는 성향이 군집화되어 있는 반면, KBO는 다양한 유형이 혼재된 구조를 갖는다.   
이는 KBO가 MLB 전략을 모방하기보다는 KBO 자체 성향에 맞춘 전략 수립이 필요하다는 점을 강하게 시사한다.