Fifa Player Segmentation: Identifying Distinct Player Profiles

This presentation delves into the segmentation of FIFA player data using K-means clustering, aiming to identify distinct player profiles based on their attributes. We will explore the project objectives, data description, analysis, and managerial insights derived from this segmentation.

Group 5

Anish Rawat - 045007

Himanshi Saxena - 045019

Lakshit Bhattacharya - 045028

Nakul Arora - 045033

PROJECT OBJECTIVES

PO1

Segmentation of FIFA Player Data using K-means Clustering to identify distinct groups based on player attributes.

2 PO2

Identification of the optimal number of clusters (k) in the FIFA Player data to ensure meaningful and interpretable groupings.

3 PO3

Determination of the characteristics of each cluster to understand the unique features and trends within each identified player cluster by analyzing how various player positions, attributes and skills differ between clusters.

DESCRIPTION OF DATA

Index Variables

The dataset lacked a natural identifier like an "ID" column. To improve organization, an index variable was created using player_id without modifying the actual data.

Categorical Variables (CV)

Categorical variables were divided into nominal and ordinal types. Nominal variables include fifa_version, league_id, club_team_id, club_position, nationality_id, nation_team_id, preferred_foot, and body_type. Ordinal variables include league_level, weak_foot, skill_moves, international_reputation, and work_rate.

Non-Categorical Variables (NCV)

Non-categorical variables include fifa_update, overall, potential, value_eur, wage_eur, age, height_cm, weight_kg, club_jersey_number, and various attributes related to pace, shooting, passing, dribbling, defending, physic, skills, movement, power, and mentality.

DATA PRE-PROCESSING

Missing Data Statistics

The dataset was divided into categorical and non-categorical subsets. The code calculated and printed the number of missing values for each variable and record.

Missing Data Treatment

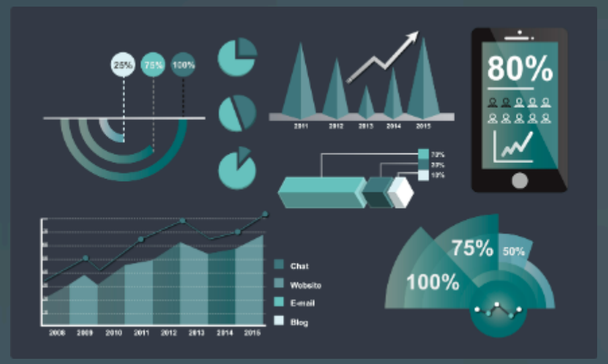
Missing categorical data was imputed using the most frequent value, while missing non-categorical data was imputed using the most frequent value. Records with more than 50% missing data were removed.

Numerical Encoding

Categorical data was encoded numerically using the Ordinal Encoder from scikit-learn.

Outlier Treatment

Non-categorical variables were normalized using Min-Max Scaler to handle outliers.



DATA ANALYSIS

Assessment Criteria

The quality of clustering was assessed using the Silhouette Score (SS) and Davies-Bouldin Index (DBI). SS measures how similar an object is to its own cluster compared to other clusters, while DBI evaluates cluster compactness and separation.

K-Means Clustering

K-means clustering was chosen due to its simplicity, scalability, interpretability, and versatility. It is well-suited for convex clusters and can handle both numerical and categorical variables.

Determining Value of 'k'

The elbow curve method was used to determine the optimal number of clusters (k). The curve showed a distinct bend around k=3, suggesting an optimal range of 2 to 4 clusters.

FORE School of Management
New Delhi

K-MEANS 4 CLUSTERING ANALYSIS

Model Performance Evaluation

The K-means clustering model with 4 clusters exhibited exceptional performance, with a very low Davies-Bouldin Index (DBI) of 0.058 and a high Silhouette Score of 0.976, indicating well-separated and compact clusters.

Cluster 4 Profile Analysis

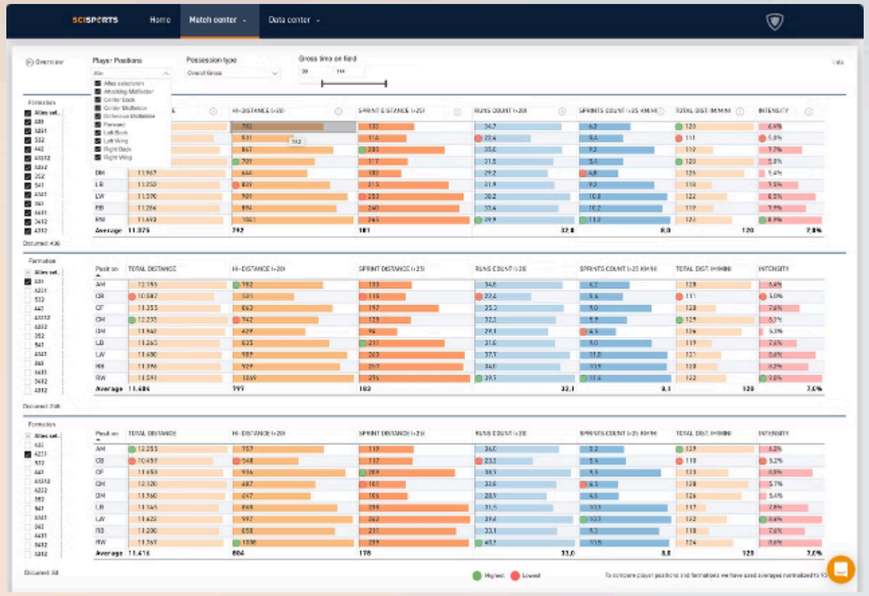
ANOVA and Chi-square tests revealed strong associations between categorical variables and cluster assignments, highlighting their importance in defining player clusters. Most non-categorical variables showed significant differences between clusters, suggesting their crucial role in distinguishing player groups.

Centricity Analysis

Cluster 0 represents specialized goalkeepers, Cluster 1 comprises top-tier, well-rounded players, Cluster 2 encompasses a large group of balanced, mid-tier players, and Cluster 3 includes less developed or specialized players.

Cluster Sizes

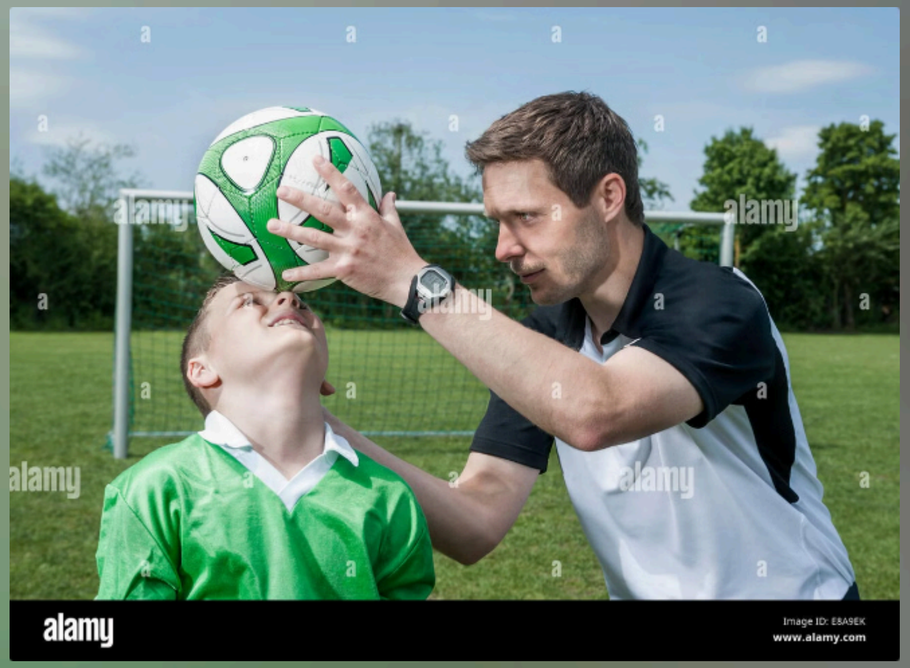
Cluster 0 (Goalkeepers) had 132 players, Cluster 1 (High-Rated Players) had 10,122 players, Cluster 2 (Mid-Tier Players) had 14,612 players, and Cluster 3 (Lower-Rated Players) had 987 players.



RESULTS | OBSERVATIONS

Clusters	Silhouette Score	Davies-Bouldin Index	Memory Usage (MiB)
k=2	0.908	0.165	1048.59
k=3	0.968	0.201	1078.74
k=4	0.976	0.058	1094.34

MANAGERIAL INSIGHTS



Player Archetype Identification

The segmentation provides valuable insights into distinct player archetypes, allowing managers to identify players with specific skill sets and potential.

Team Building Strategies

Managers can utilize these insights to build balanced teams by strategically selecting players from different clusters based on their roles and attributes.

Player Valuation and Transfer Decisions

The clustering analysis can inform player valuation and transfer decisions by providing a framework for understanding the relative value of players within different clusters.

Player Development and Training

The clustering analysis can help in player development by analysing their weakness, pain points and strengths and help in providing them the required training.

CONCLUSION

The FIFA player segmentation using K-means clustering has successfully identified distinct player profiles, providing valuable insights for managers. The analysis reveals meaningful differences in player attributes, skills, and potential across clusters, offering a framework for informed decision-making in team building, player valuation, and development.