Analysis of Playing Styles of NBA Players Using the K-Medoids Method

Kurnia Ramadhan Putra [a,1], *Christian Giery* [b]

[a] Department of Information System, Faculty of Industrial Technology, Institut Teknologi Nasional Bandung, 40124, Indonesia

Abstract. The evolution of basketball in the NBA has shifted from traditional five-position roles to more flexible, skill-based playstyles. This study explores the classification of NBA players based on their individual performance metrics using the K-Medoids clustering method. Data from the 2015–2016 to 2024–2025 NBA seasons were collected and processed using the CRISP-DM framework. After data standardization and dimensionality reduction with PCA, the K-Medoids algorithm was applied to group players into distinct clusters. Evaluation using Davies-Bouldin Index (DBI) and Silhouette Score confirmed that a three-cluster configuration yielded the best cohesion and separation. The identified clusters reflect distinct roles such as elite scorers, defensive big men, and versatile contributors, providing valuable insights for team composition and strategy optimization.

1 Introduction

The Gameplay in National Basketball Association (NBA) league has evolved significantly from getting used to a traditional 5 position into a new era of flexibility gameplay, which player can have a gameplay characteristic based on another position. This tactical shift also known as “Neo-Position” that prioritize player individual skill over the traditional position characteristic (1). This concept is best shown by Golden State Warriors team in the 2015 – 2016 season, their record of success in the regular season is highlighted by the individual ability by the player that not limited by the traditional position (2,3). Therefore, classifying player based on their position is no longer sufficient for evaluating their total contribution. Previous studies have indicated that specific statistic like three point shooting ability, have a positive correlation with team performance (4). So, this research proposes a clustering method using K-Medoids to identify a group of players based on their individual statistics to understand their specific playstyles.

Unlike K-Means, the K-Medoids algorithm selects actual data points as cluster center (medoids), that can handle outliers betters with players that has an unusually high or low statistics (5). This is particularly relevant in the NBA context, where some players might have extreme values in scoring, assists, or defensive metrics that could distort the clustering process. By applying K-Medoids, this study aims to uncover distinct clusters such as Playmakers, Elite Scorers, Bigman Defenders, or Sharpshooters, based on performance indicators rather than nominal roles or the traditional position. The results are expected to

[1] Corresponding author: kurniaramadhan@itenas.ac.id

provide valuable insights for coaches and analysts in constructing balanced rosters and optimizing team strategy based on actual player tendencies.

The NBA is a premier basketball league featuring 30 teams with top-tier talent (NBA Staff, 2019). While players were once confined to five standard positions—Point Guard, Shooting Guard, Small Forward, Power Forward, and Center—recent years have shown a shift toward more dynamic roles (NBA, 2025). Modern strategies often rely on "neo-positions," where players are valued more for their individual skills than their assigned role. Examples include the “Point Forward,” who blends size and ball control, or the “Score-First Point Guard,” who focuses more on scoring than distributing the ball (Phadke & Pai, 2021). This shift was exemplified by the Golden State Warriors during the 2015–2016 season. Their success using a small-ball lineup, emphasizing speed and 3-point shooting over traditional size, helped them set records for wins and shooting performance (Levy, 2015b; Gaurav, 2019). As roles evolve, relying solely on position to assess a player’s impact is no longer effective. Research has shown that strong 3-point shooting correlates with team success, while an excess of pass-first players may reduce efficiency (Pilsbury, 2023). To better understand playing styles, this study uses the K-Medoids clustering method to group players based on statistical performance. Unlike K-Means, K-Medoids selects actual data points as cluster centers, offering stronger resistance to outliers—such as players with unusually high or low stats (Berkhin, 2006). By identifying player clusters like Playmakers, Elite Scorers, or Defenders, this study aims to help teams recognize role gaps, optimize lineups, and refine strategy based on real game performance

2 Methodology

This study utilizes Cross Industry Standard Process of Data Mining (CRISP-DM) methodology as a framework that known from its flexibleness and iterative process (6,7). This methodology contained 6 processes: Business Understanding, Data Understanding, Data Preparation, Modeling, Evaluation, and Deployment. The whole process is illustrated in **Fig.1**. However, this study does not include the deployment phase.

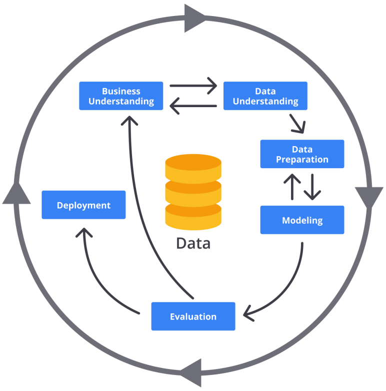

Fig. 1. Crisp-DM Methodology.

2.1 Business Understanding

The first step is to define the business and goals. In this study it aims to understand that NBA is an organization that maintain the basketball league in USA, and the team is usually containing 5 players on each position, but using clustering as an analysis it aims to provide more objective insights into player characteristics based on individual statistics, which can be valuable for coaches, team management, analysts or an NBA enthusiast. The expected outcome is to generate player groupings that reveal similarities in performance profiles, offering a more objective perspective for decision-making in team strategy, player development, and talent evaluation.

2.2 Data Understanding

NBA player statistics from the 2015 – 2016 to the 2024 – 2025 seasons from basketball-reference.com website were collected using web scrapping technique with Python and Selenium library. These datasets include statistics from the per 100 possessions dataset and advanced stats dataset. In this particular process, Exploratory Data Analysis (EDA) was performed to understand the structure of its data and identifying pattern. This process includes understanding the dataset that used in this study, dataset was obtained by merging two related tables, resulting in 52 features representing player identities and statistics. It contains 930 entries per season, providing a comprehensive overview of player performance.

2.3 Data Preparation

This stage focuses on cleaning and transforming the data for model readiness. The process includes four process such as data merging, data cleaning, data standardization and dimensionality reduction:

- Data Merging: The dataset was constructed by combining two main statistical sources, namely the Per 100 Possessions table and the Advanced Stats table. The merging process was carried out using player ID as the primary key to ensure accuracy in matching records. By integrating these two sources, each player's performance data could be represented more comprehensively, resulting in a dataset with a total of 52 features that capture both traditional and advanced metrics.
- Data Cleaning: This process is conducted to handle any inconsistencies within the dataset. For example, missing values in the "3P%" feature may occur because some players never attempted a three-point shot during the entire season. Another step is resolving duplicate entries for players who played for more than one team in a season. Lastly, redundant features that appear in both tables are removed to avoid duplication and maintain data consistency.
- Data Standardization: Z-Score Standardization was applied to transform the data into a common scale, ensuring that features with different ranges, such as "PPG" and "RB," are comparable. This step is essential to prevent features with larger values from dominating the analysis and is also a prerequisite before applying dimensionality reduction techniques such as PCA.

- Dimensionality Reduction: PCA was employed to reduce the dimensionality of the dataset while preserving as much variance as possible. To determine the optimal number of components to retain, this study utilized both the Scree Plot, which visualizes the explained variance of each principal component, and the Kaiser Criterion, which recommends retaining components with eigenvalues greater than one.

2.4 Modeling

This stage was focused on building the cluster based on the datasets. The process includes:

- Determining The Number of Cluster: The optimal number of the cluster to be built is determined using elbow method, this technique will recommend the exact number of the optimal cluster using line diagram visualization. In this particular situation, the cluster were set to 3, by using the elbow point.
- Cluster Formation: The K-Medoid clustering method was used to build the cluster, the method applied to the data that the dimension has been reduced. Players were grouped based on proximity using Euclidean Distance. The process of clustering using K-Medoids is illustrated in Fig.2..

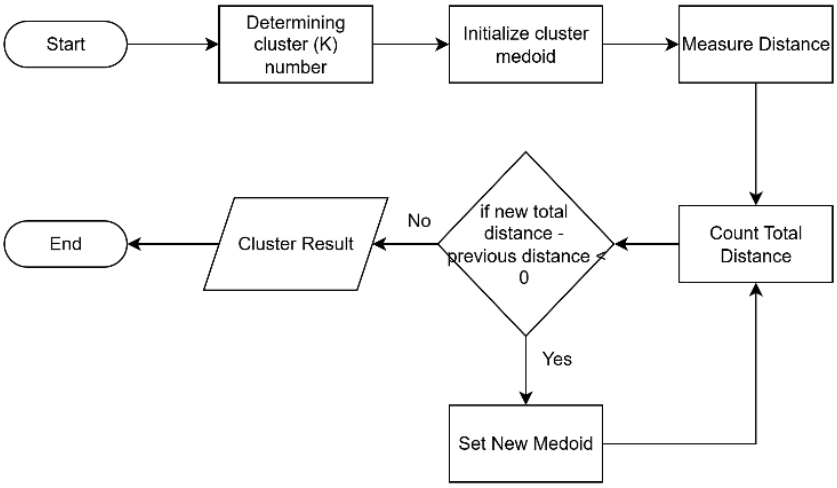
Start
Determining cluster (K) number
Initialize cluster medoid
Measure Distance
Count Total Distance
if new total distance - previous distance < 0
No
Yes
Set New Medoid
Cluster Result
End

Fig. 2. K-Medoids Process.

2.5 Evaluation

This stage was focused on evaluating the result of the model, the cluster. The evaluation was used to measure the model’s quality. The evaluation metrics used include:

- DBI: Davies-Bouldin Index (DBI) measures the ratio of intra-cluster distance (density within cluster) and inter-cluster (separation within cluster). The result of the metrics is used to evaluate the cluster result, a lower DBI score means a better cluster quality.

- Silhouette Score: This evaluation measure how each data fits within cluster, A higher score indicates more dense and well separated cluster.

3 Result and Discussion

3.1 Result of Data Preparation

Following to the previous step on preparing the data, the raw data from season 2015 – 2016 to 2024 – 2025 seasons was combined and cleaned, by handling the missing value, removing duplicate, and standardizing the data using Z-Score Standardization. And to reduce the dimension of the data, PCA was used to reduce the dimension, the result from using Scree Plot, the first 3 component was chosen to be used as illustrated in Fig.3.

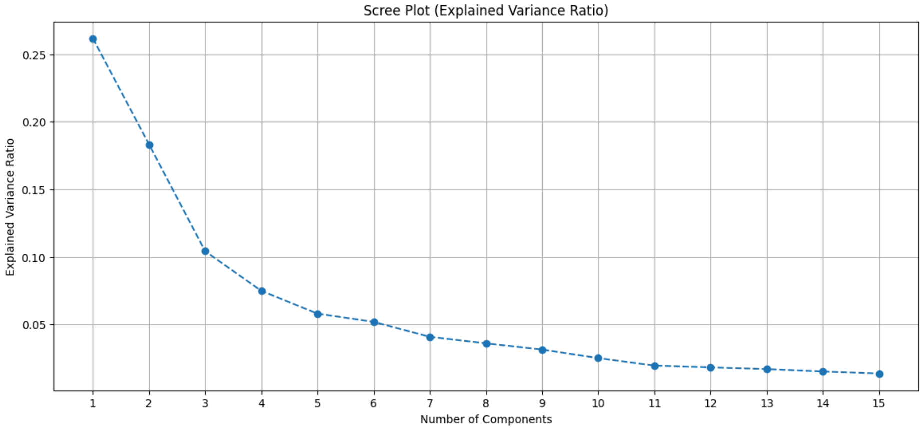

Fig. 3. Scree Plot Diagram for Number of Components.

3.2 Result of Data Preparation

To interpret the meaning of the principal components (PCs) and understand what characteristics they represent, we analyzed the loadings of the original variables onto each component. The loading values, ranging from -1 to +1, indicate the correlation between a principal component and the original features. A high absolute value signifies a strong relationship, with a positive sign indicating a direct correlation and a negative sign indicating an inverse one. This analysis is crucial for transforming abstract PC scores into meaningful insights about player attributes.

Table 1. Principal Component 1 Feature Loading.

No	Attribute Name	Loading
1	PER	0.26
2	2PT	0.24
3	FG	0.23
4	WS	0.22
5	PPG	0.22
6	BPM	0.22
7	VORP	0.21
8	OWS	0.21
9	FG%	0.20
10	DWS	0.20
11	TS%	0.19
12	MPG	0.17
13	Games_start	0.17
14	2PT%	0.16
15	Games_played	0.14

Table 1. explains that PC1 represents the overall productivity and contribution of a player. High positive loadings on stats like PER, PPG, and WS show that a player's PC1 score is strongly correlated with their scoring ability and positive impact on team success.

Table 2. Principal Component 2 Feature Loading.

No	Attribute Name	Loading
1	TRB%	0.27
2	RB	0.27
3	ORB%	0.27
4	ORB	0.27
5	3PT	0.27
6	3PTA	0.26
7	3PTA%	0.22
8	DRB%	0.22
9	DRB	0.22
10	BLK	0.20
11	BLK%	0.20
12	3Par	0.20
13	DRtg	0.15
14	AST%	0.15
15	PF	0.15
16	AST	0.13
17	FT%	0.12

Table 2 explains that PC2 acts as a clear discriminator between two primary player archetypes: interior players and perimeter shooters. The component has high positive loadings on rebounding and defensive metrics, such as Total Rebound Percentage (TRB%), Rebounds (RB), and Blocks (BLK). Conversely, it has strong negative loadings on three-point shooting variables, including Three-Pointers Made (3PT), Three-Point Attempts (3PTA), and Three-Point Attempt Rate (3PAr). This bipolar distribution indicates that a high positive PC2 score represents a player whose style is defined by rebounding and defense, while a high negative score signifies a player focused on three-point shooting.

Table 3. Principal Component 3 Feature Loading.

No	Attribute Name	Loading
1	USG%	0.29
2	TOV	0.28
3	FTA	0.28
4	ORtg	0.27
5	2PTA	0.26
6	FT	0.26
7	WS/48	0.25
8	fga	0.23
9	DBPM	0.22
10	FTr	0.18
12	TOV%	0.14
13	Age	0.9
14	STL%	0.1
15	STL	0.1

Table 3 explains that PC3 appears to represent a player's offensive volume and ball-handling role. This is evident from the high positive loadings on metrics such as Usage Rate (USG%), Turnovers (TOV), Free Throw Attempts (FTA), and Field Goal Attempts (fga). These loadings collectively indicate a player who is the primary offensive engine for their team, handling the ball frequently, and taking a high volume of shots. The strong correlation with Turnovers is a natural consequence of this high-volume role. Furthermore, the component also has notable loadings on Offensive Rating (ORtg) and Win Shares per 48 minutes (WS/48), suggesting that while this style is high-volume, it is also linked to overall offensive efficiency and positive team contribution.

3.3 K-Medoids Clustering Method

To determine the optimal number of clusters, the Elbow Method was used as shown in **Fig.4.** The plot of the Sum of Squared Error (SSE) showed a clear "elbow" point at k=3 represent the most effective number of clusters.

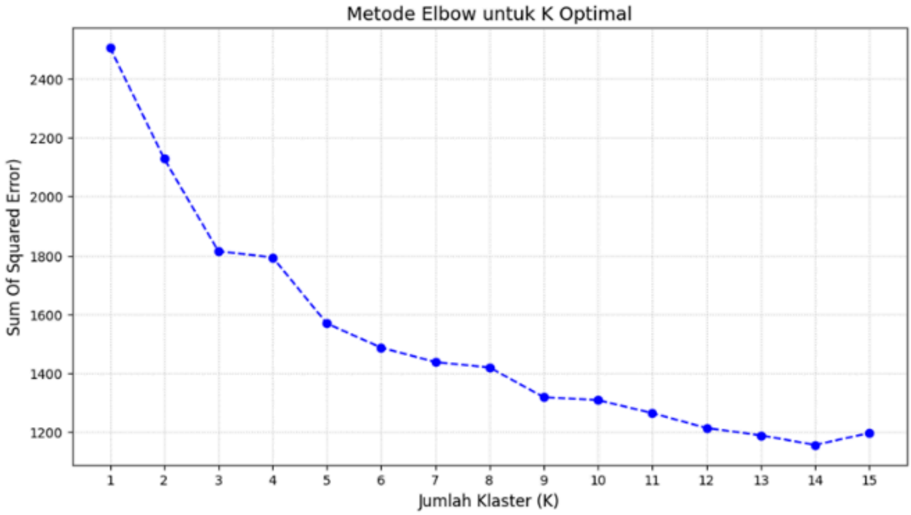

Fig. 4. Elbow Method.

With the optimal number of clusters established, the K-Medoids model was applied to the data, which had been reduced to 3 principal components. The clustering results visualize the players' data grouped into 3 clusters, where each player is associated with their closest *medoid* as shown in **Fig. 4**.

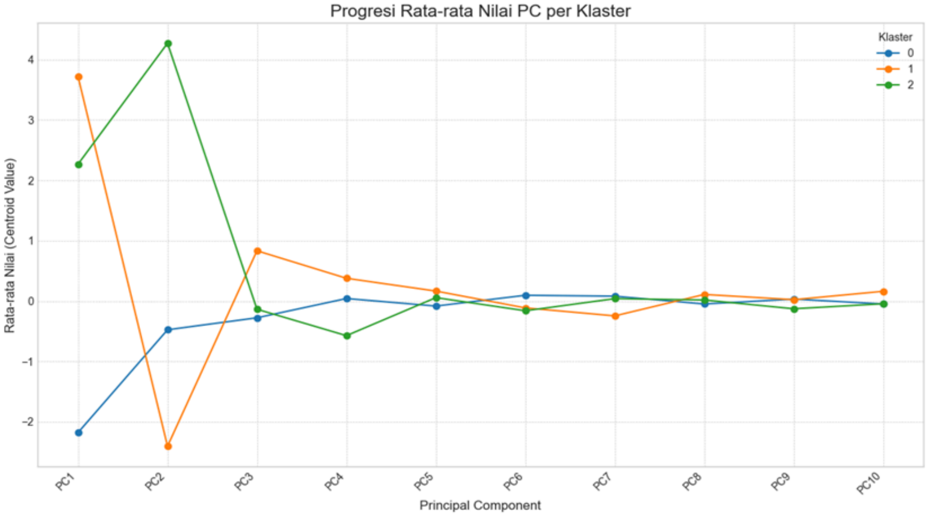

Fig. 5. PC Score Visualization Each Cluster.

As shown in **Fig.5**, this line chart shows the differences of PC score from each cluster. Cluster with ID 0 is dominant on PC 8, other cluster with ID 1 is dominant on PC 1 and lastly cluster with ID 2 is dominant on PC 2. The average PC's score each cluster can be seen in **Table 4**.

Table 4. Average PC Score By Each Cluster.

Id	PC1	PC2	PC3
0	-2.285053	-0.486420	-0.336017
1	3.287489	-2.369835	0.783290
2	2.178930	4.287421	-0.053103

To gain insight on how each cluster has pc scores in average we can look into the average statistic by each cluster that shown in, these statistics were already standardize using Z-Score, so if it’s above 0, it means the statistic is above the average of that spesific column. To gain more insight cluster 0 have a strong negative relation with PC1, Cluster 1 have a strong positive relation with PC1 and lastly cluster 2 has a strong positive relation with PC2. To gain more insight about the result of the clustering, in tables below, these table includes head of each cluster to give a representation about data on each cluster.

Table 5. Cluster ID 0

Id	Nama	PC1	PC2	PC3
agbajoc01	Ochai Agbaji	- 0.366060	- 1.046707	- 2.458745
alexani01	Nickeil Alexander-Walker	- 0.194930	- 1.950591	- 2.053835
alexatr01	Trey Alexander	- 5.135658	- 0.621947	1.803660
allengr01	Grayson Allen	- 0.925671	- 2.817510	- 1.565221
alvarjo01	Jose Alvarado	- 1.610677	- 3.319547	- 0.122540

Table 5 presents a sample of player data that serves as the basis for clustering analysis. This cluster has a uniqueness, described by the PC scores, player within this cluster doesn’t have any significant strong positive score on each PC, but having a negative strong score mostly in PC 1.

Table 6. Cluster ID 1

Id	Nama	PC1	PC2	PC3
adebaba01	Bam Adebayo	6.053220	0.087275	0.672382
aldamsa01	Santi Aldama	2.398692	- 0.498708	- 1.642948
antetgi01	Giannis Antetokounmpo	13.597815	- 0.011413	4.049855
anthoco01	Cole Anthony	1.550211	- 1.195354	1.229648
anunoog01	OG Anunoby	2.729140	- 2.283649	- 1.583260

Table 6 above presents a selection of NBA players identified within one of the formed clusters ID 1, highlighting their unique grouping based on statistical similarities. It includes five players from different teams and positions, such as Giannis Antetokounmpo who have a strong PC1 scores and Bam Adebayo on the second as a player who has a highest PC1’s score.

Table 7. Cluster ID 2

Id	Nama	PC1	PC2	PC3
achiupr01	Precious Achiuwa	0.873804	3.798287	- 0.444613

adamsst01	Steven Adams	2.286608	7.360474	0.608919
allenja01	Jarrett Allen	8.430815	3.458788	- 1.714211
anderky01	Kyle Anderson	1.221516	1.257097	- 1.306115
aytonde01	Deandre Ayton	3.428702	3.896133	0.921751

Table 7 cluster highlights a diverse group of NBA players from various teams and positions, showing how statistical similarities can connect athletes despite their different roles on the court. It includes Steven Adams who has the highest PC2's score.

3.4 Cluster Labelling

After the clustering model successfully grouped the players, the next crucial step was to interpret and label each cluster to provide meaningful insights. This was achieved by analyzing the average Principal Component (PC) scores for each cluster, as well as examining the standardized statistics of the players within them. By identifying the attributes with the highest positive or negative values, a clear profile for each cluster could be established. For example, a cluster with a high average PC2 score would be labeled based on the features with the highest loadings on that component (e.g., rebounding and blocks), while a cluster with a low average score on PC1 would be characterized by attributes with negative loadings on PC1 (e.g., lower offensive efficiency). This process was informed by domain knowledge of basketball analytics, allowing us to assign descriptive names to each cluster, such as "All-around Playmaker" or "Non-Scoring Bigman." The goal of this labeling was to transform the raw clustering results into an understandable narrative of player playing styles.

- Cluster 0 : Low Volume Three Point Specialists
  This cluster is characterized by very low average scores, particularly on PC1 and PC2. Based on the loadings analysis, the strong negative values on PC1 and PC2 indicate players with low offensive productivity and shooting ability. However, they show slightly above-average values in statistics such as 3PT% and Steals (STL). This implies that they are not primary scorers, but rather role players who shoot three-pointers with high efficiency when the opportunity arises and contribute through defense.
- Cluster 1 : All-Around Playmaker
  This cluster exhibits very high average scores on PC1 and very negative values on PC2. This combination very clearly identifies players with high offensive productivity (based on PC1) and a playing style that is the opposite of interior players (PC2). With their positive value on PC1, players in this cluster excel at scoring points (PPG), team contribution (WS), and efficiency (PER). Furthermore, the strong negative value on PC2 indicates that they are not players who dominate in rebounding or blocking, but rather tend to possess playmaking abilities and perimeter shooting skills.
- Cluster 2 : Non-Scoring Bigmen
  This cluster has a very high average score on PC2. As explained by the loadings, the high PC2 value indicates that this cluster is dominated by players who excel at rebounding (RB, TRB%) and interior rim defense (BLK, DRB%). Although they

have a high score on PC2, their PC1 value is relatively low compared to Cluster 1. This confirms that they are players whose focus is not on scoring points, but on physical dominance and defense around the basket.

3.5 K-Medoids Clustering Evaluation

Looking into the result of the clustering as shown in **Fig.5** each cluster has its own unique characteristics, some has higher on and some has lower on each statistic comparing into other cluster. After completing the modelling, evaluation is required to make sure indeed it was the best possible outcomes. The evaluation is conducted using DBI and Silhouette Score to assess the quality of the clustering results, particularly in cluster separation and cohesion. The DBI measures the average similarity between each cluster and its most similar one, where lower DBI values indicate better clustering. A lower DBI suggests that clusters are compact and well-separated from each other, which is desirable in clustering analysis.

Table 8. DBI's Score

No	Cluster	DBI
1	2	2.3437
2	3	1.3580
3	4	1.4889
4	5	1.5153
5	6	1.5605

Table 8 indicate that three cluster with three PCA components configurations, have the best compactness and the cluster are well separated between each other, achieving 1.3580 score which is superior compared to other clustering options.

Table 9. Silhouette Score First Attempt

No	Cluster	DBI
1	2	0.2818
2	3	0.3285
3	4	0.1783
4	5	0.1949
5	6	0.1981

Second evaluation using Silhouette Score, as shown in Tabel 9 When K=3, the clustering configuration produces higher silhouette score compared into other values of K, indicating better cohesion within cluster separation. The score achieve 0.3285 which means that the data is reasonably well assigned to their respective cluster.

Table 10. Silhouette Score Second Attempt

No	Cluster	DBI
1	2	0.1482
2	3	0.1035

3	4	0.1248
4	5	0.0857
5	6	0.0633

Second attempt on evaluating cluster result using Silhouette Score by using configuration of three cluster and ten PCA components. Although this configuration provided a decent result, the evaluation indicated that the silhouette score yielded a better configuration when applying two clusters, suggesting that the data structure is more naturally represented with fewer cluster divisions.

4 Conclusion

This study successfully classified NBA players into distinct playstyles based on their statistical performance, moving beyond traditional positional roles. By employing the CRISP-DM methodology, data standardization, PCA for dimensionality reduction, and the K-Medoids clustering algorithm, a robust analytical framework was established. Evaluation using the Davies-Bouldin Index (DBI) and Silhouette Score confirmed that a three-cluster configuration yielded the most optimal results, demonstrating strong cohesion within clusters and clear separation between them.

The three identified clusters represent unique player archetypes:

- Non-Scoring Bigmen: Players who excel in rebounding and interior rim defense but are not focal points for scoring.
- All-Around Playmakers: Highly productive offensive players who dominate in scoring, efficiency, and team contribution, often possessing playmaking and perimeter shooting skills.
- Low-Volume Three-Point Specialists: Role players who are not primary scorers but contribute through high-efficiency three-point shooting and defensive efforts.

These insights provide valuable, data-driven guidance for coaches, analysts, and team management in constructing balanced rosters, identifying player talent, and optimizing strategic gameplay based on empirically derived playstyles

References

1. Phadke V, Pai O. The Rise of Neo-Positions in Basketball [Internet]. 2021. Available from: https://www.bruinsportsanalytics.com/post/neo_positions

2. Levy I. How the Warriors evolved small ball and, in the process, the NBA [Internet]. 2015. Available from: https://www.si.com/the-cauldron/2015/10/12/golden-state-warriors-nba-title-small-ball-steph-curry

3. Gaurav. NBA records broken during the 2015-16 season [Internet]. 2019. Available from: https://www.essentiallysports.com/nba-records-broken-during-the-2015-16-season/
4. Pilsbury D. Evaluating Combinations of Play Styles in the NBA by. 2023;
5. Berkhin P. A survey of clustering data mining techniques BT - Grouping Multidimensional Data. Group Multidimens Data [Internet]. 2006;(c):25–71. Available from: http://dx.doi.org/10.1007/3-540-28349-8_2%5Cnpapers2://publication/doi/10.1007/3-540-28349-8_2
6. Wirth R, Hipp J. CRISP-DM: towards a standard process model for data mining. Proceedings of the Fourth International Conference on the Practical Application of Knowledge Discovery and Data Mining, 29-39. Proc Fourth Int Conf Pract Appl Knowl Discov Data Min [Internet]. 2000;(24959):29–39.