From Consistency to Breakout: NBA Player Performance Analysis and Tiering

Ethan Quinones, Pardha Sai Nekkalapu, Raja Muhammad Shayan, Miles Koehnemann

Arizona State University

etquinon@asu.edu, pnekkal1@asu.edu, rshayan1@asu.edu, mkoehnem@asu.edu

***Abstract*-**This project focuses on creating a data driven system to evaluate the performance of NBA players, over seasons by categorizing them into performance tiers and recognizing players in each tier using methods, like clustering and ranking to assist teams in identifying reliable performers. The analysis supports decision-making in team strategy, player contracts, and talent development.

I. Introduction

A. Background

NBA analytics has become increasingly important in player evaluation and team strategy. Performance analysis is crucial for assessing player consistency, development, and value over time. The current focus is on identifying player performance tiers across seasons to assist teams in making informed decisions.

B. Problem

The main challenge is to group NBA players into performance tiers across seasons and rank them within these clusters to identify standout players. This involves handling multiple performance metrics, dealing with data challenges such as missing games and seasonal variability, and balancing the clustering and ranking processes.

C. Importance

This analysis is significant for several reasons:

1) It helps teams spot consistent players, emerging talent, and potential risks.

2) It supports decision-making in team strategy, player contracts, and development plans.

3) It provides a data-driven approach to player evaluation, which is crucial in the modern NBA.

D. Existing Literature

Clustering is a widely used method in data science and machine learning for grouping similar data points, with applications across diverse domains, including sports analytics. In the context of NBA player analysis, clustering methods have been employed to uncover player archetypes and analyze performance patterns. The following section reviews relevant literature and highlights the methodologies applied to NBA player clustering.

In "NBA Player Clustering: Exploring Player Archetypes in a Changing NBA" by Elam (2019) [1], clustering techniques were utilized to categorize NBA players into archetypes that reflect evolving roles in modern basketball. This study applied k-means clustering to player statistics, identifying archetypes such as "three-point specialists" and "all-round playmakers," thereby capturing the strategic shifts in player utilization.

Lutz (2012) [2], in "A Cluster Analysis of NBA Players," used unsupervised clustering methods to segment players based on performance metrics. The study demonstrated the application of hierarchical and k-means clustering to identify player categories such as "defensive anchors" and "scoring forwards," offering insights into team composition strategies.

Patel (2017) [3], in "Clustering Professional Basketball Players by Performance," extended the exploration of clustering by integrating advanced performance metrics such as efficiency ratings and usage percentages. The study employed Gaussian Mixture Models (GMM) to identify nuanced player archetypes like "high-usage scorers" and "efficient role players," emphasizing the role of probabilistic models in capturing overlapping player characteristics.

Finally, the open-source project by vjangili-26 [4] explored various clustering algorithms to determine NBA player archetypes. Using k-means, DBSCAN, and hierarchical clustering methods on datasets, this study highlighted

differences in the results produced by these algorithms and identified roles such as "rim protectors" and "perimeter defenders," underscoring the importance of algorithm selection in clustering tasks.

Overall, these studies illustrate the versatility of clustering methods like k-means, GMM, and DBSCAN for analyzing NBA player roles and archetypes. The use of these techniques highlights the evolving strategies in basketball analytics and the potential of clustering for uncovering hidden patterns in sports data.

E. System Overview

The data mining system pipeline consists of the following main steps:

1. Data Preprocessing: Addressing missing data, outliers, and feature encoding.
2. Feature Selection and Engineering: Curating key metrics.
3. Dimensionality Reduction: Reducing complexity using Principal Component Analysis (PCA)
4. Clustering: Grouping players into tiers.
5. Player Ranking within clusters: Establishing a ranking system within clusters.
6. Performance Analysis
7. Evaluation of clustering quality and identification of breakout players

F. Data Collection

The project uses NBA player statistics across multiple seasons. The dataset includes various performance metrics such as:

1. Basic Metrics: Points, Assists, Rebounds, Steals, Blocks
2. Efficiency Metrics: Field Goal Percentage, Free Throw Percentage, True Shooting Percentage
3. Advanced Metrics: Effective Field Goal Percentage, Player Efficiency Rating

G. Components of the ML System

1. Data Preprocessing: Handling missing values, creating new features, encoding categorical variables.

2. Feature Selection and Engineering: Picking features that encompass offensive and defensive skills well as the efficiency of player performance.
3. Utilizing Principal Component Analysis (PCA) we can condense the dataset down to three components, through dimensionality reduction techniques.
4. Utilizing techniques such as K means clustering, DBSCAN and Hierarchical clustering to categorize players into performance levels.
5. Player Ranking: Ranking players within each cluster based on performance metrics.
6. Performance Analysis: Assessed cluster stability and identified breakout players.

H. Experimental Results

The team experimented with different clustering algorithms and parameters:

1. K-means Clustering:
    a. Optimal K = 5 (determined using the elbow method)
    b. Training Silhouette Score: 0.3116
    c. Testing Silhouette Score: 0.3133
    d. Training Inertia: 35687.67219092188
2. DBSCAN:
    a. Best parameters: eps=1.0, min_samples=3
    b. Clusters formed: 5
    c. Silhouette Score: 0.4363
3. Hierarchical Clustering:
    a. Best performance with 2 clusters
    b. Training Silhouette Score: 0.3582
    c. Testing Silhouette Score: 0.3772

The team also explored various feature subsets to improve clustering effectiveness:

- Role-Based features (PTS, AST, TRB, STL, BLK, POS) provided relatively low inertia with improved silhouette scores.
- Fine-tuning led to a final feature set: [PTS, AST, ORB, DRB, STL, BLK, POS, FT%]

Player ranking within clusters was initially based on points scored (PTS) but was later improved to use a composite score based on 14 weighted metrics.

II. Important Definitions and Problem Statement

The dataset that we have chosen includes comprehensive NBA player performance metrics across multiple seasons, spanning from the 1997-98 season to the 2021-22 season. Key variables represent offensive, defensive, and efficiency-based attributes that players have. These attributes are critical for player evaluation and team strategy. Key variables include Points Scored (PTS) the average number of points scored per game, Assists (AST) the average number of assists had per game, Rebounds (TRB) the average number of total rebounds per game, Steals (STL) the average number of steals had per game, Blocks (BLK) the average number of blocks per game, Field Goal Percentage (FG%) the average field goal percent per game, Minutes Played (MP) the average number of minutes played each game, Turnovers (TOV) the average number of turnovers per game, Free Throw Attempts (FTA) the average number of free throws attempted per game, Points Per Minute (PPM) the average number of points per minute per game.

The primary goal of our project is to cluster players into distinct character archetypes and then rank them within their respective clusters to identify the best potential roster constructions for a team or a coach. Each cluster represents a different style of play. For example, one cluster has a lot of very high scorers of all positions, while another cluster has a lot of high block and rebound centers.

Key concepts within the data include membership is the grouping of players based on similar performance attributes and characteristics. Ranking within each cluster gives a rank based on a player's performance within the cluster. Position refers to the position that the players is, guards, forwards, and centers.

Given multi-seasonal data containing comprehensive player performance metrics for NBA players, we want to develop an unsupervised clustering framework that groups NBA players into different player archetypes, then ranks

those players within their given archetypes. The goal is to develop a way for coaches to see potential roster constructions as well as for players to evaluate their performance compared to similar players.

III. Overview of proposed approach/system

This project uses a data-driven approach to classify and rank NBA players across seasons based on their performance metrics. By utilizing unsupervised machine learning techniques, the system creates meaningful player clusters, ranks individuals within these clusters, and visualizes their placements in performance tiers. The data pipeline begins with preprocessing and cleaning, where duplicates are removed, categorical features like player positions are encoded and new metrics, like points per minute (PPM), are introduced. Missing data is addressed through imputation, making sure there is consistent across all features. Dimensionality reduction is then performed using Principal Component Analysis (PCA), which reduces high-dimensionality data to the principal components, where we are able to retain variance while doing clustering and data visualization.

The clustering framework uses DBSCAN (Density-Based Spatial Clustering) to group players based on their similarities in the PCA-reduced space. DBSCAN is particularly effective for handling nose and defining non-convex clusters. This makes it well suited for the diverse performance profiles in the NBA. Hyperparameter tuning is then used to optimize **eps** (neighborhood radius) and **min_samples** (minimum points to form a cluster), with the silhouette score being used as the metric to evaluate each of the cluster's performance. Once the players have been clustered, they are then ranked within each group using a composite score taken from standardized and weighted metrics, specifically PTS, AST, and FG%. This ranking highlights standout players of each player archetype.

Evaluation metrics for the system include the silhouette score, which measures the, and inertia, which quantifies cluster compactness by calculating the sum of square’s distances to the nearest cluster center. To enhance the visuals, clusters are shown in 2D and 3D PCA spaces, allowing further insight into performance tiers. The system outputs the cluster memberships as well as a ranking within the cluster. These are then used to display the top team based on the average player rank, the best 5-man lineup for each team each year, and an interactive section where you input a year and a team, and it outputs the best 10 5-man lineups for the given roster.

Our approach provides a robust framework for evaluating NBA players by identifying their current standing and comparing it to similar players in league history. It supports data-informed decision making in player management,

team strategy, team development and roster construction, allowing for a more strategic approach to player evaluation and optimization.

IV. Technical Details of proposed approaches/system

The proposed system leverages a data mining pipeline to analyze various NBA player statistics and cluster them into distinct performance tiers. The pipeline consists of various key steps including data preprocessing, dimensionality reduction, k-means clustering, and ranking. After utilizing the data mining pipeline to cluster and rank NBA players, the model was fine-tuned and to increase the various performance metrics.

A. Data Preprocessing

In preparation for exploratory data analysis and feature engineering, the dataset was thoroughly preprocessed to ensure consistency, reduce noise, and prepare the data for optimal clustering. Key steps in the preliminary data mining pipeline included duplicate removal, feature engineering, and missing value handling. New metrics that were computed in this section include new parameters: “Season_Start_Year” and “Points_Per_Minute”. These metrics were created to aid in temporal analyses and normalization of existing NBA player statistics.

B. Feature Selection and Normalization

After preprocessing and cleaning the data, features were normalized to aid the clustering process. For the preliminary data mining pipeline, the key performance indicators included "PTS", "AST", "TRB", "STL", "BLK", "Points_Per_Minute", "FG%", "eFG%", "MP", "TOV", "FTA", "Pos". Categorical data such as “Pos” were numerically encoded so that it could be used in the clustering model. These metrics cover the diverse skill set that is required for NBA players. Offensive, defensive, and efficiency in playmaking were all skill sets that were to be investigated by the clustering algorithm. All numeric features were then encoded using the StandardScaler Python module. Normalizing all numeric attributes included in the model ensured all attributes had comparable scales and prevented certain statistics from dominating the clustering model.

C. Principle Component Analysis

PCA was implemented in order to reduce dimensionality while retaining variance in the dataset. PCA was applied to reduce the dataset’s high dimensionality to three principal components. These components served as the input for

clustering as they captured a majority of the dataset's variance. The explained variance confirmed that meaningful information was preserved through the reduced dimensions.

D. Clustering with K-Means

K-Means clustering was performed to group the NBA players from the dataset into distinct performance tiers. We utilized the k-means clustering algorithm due to its efficient nature in partitioning data points into well-separated clusters. The algorithm assigned players to clusters while minimizing the distance between the data points and their respective cluster centroids.

1. Implementation Details

The k-means clustering was performed using the KMeans class from the sklearn.cluster module in Python. The key parameters in the algorithm included:

- Number of Clusters (K)**:** This parameter defines the number of clusters the algorithm uses to partition the data.
- Random State (random_state)**:** This parameter is a fixed random state that ensures the results from each run of the model can be reproduced.

The elbow method was utilized to determine the optimal value for the number of clusters (K). The inertia was plotted against different values of K. The optimal value of K was determined as the point where the curve begins to flatten. For the preliminary clustering model, we identified the optimal value of K to be 5, striking a balance between simplicity and accurate, meaningful results. Random state was set to 42 in the preliminary model.

Performance Evaluation

Evaluation metrics for the proposed model include the following:

- Inertia: This statistic measures compactness of clusters. Lower values are preferred as they indicate tighter and more well-defined clusters resulting from the dataset.
- Silhouette Score: This statistic measures the separation between clusters. Higher values are preferred and indicate better-defined and well-separated clusters.

The results of the preliminary model produced a training inertia of 35,687, a training silhouette score of 0.3116, and a testing silhouette score of 0.3113. These values indicate that the model has compact clusters with moderate within cluster variance, moderately separated clusters during training, and consistent clustering performance on unseen data (test data). However, the results indicate that there is still potential for improvement in cluster separation. Despite this, the model effectively groups players together that have similar attributes, providing valuable insights into player roles and tiers. Within each of the clusters, the top 10 players were ranked according to highest point value. This displayed the top ten performers within each cluster for both the testing and training data.

The k-means clustering algorithm, with value K=5, successfully grouped NBA players into different performance tiers based on selected statistical attributes. While the clustering performance was sufficient and valuable insights were gained, further enhancements were needed to feature selection, preprocessing, and model refinement. Only then could cluster cohesion and separation be improved. This implementation demonstrates the utility of k-means clustering in the domain of sports analytics, resulting in deeper player analysis.

V. Experiments

A. Data Description

The dataset comprises an extensive array of NBA player statistics meticulously structured to enhance performance analytics across various seasons. Each entry is uniquely identified by the player's name, with additional general information including the player's position, age, team affiliation, and the specific NBA season. Key metrics such as games played, and games started providing insights into the player's utilization and role within the team.

Performance metrics cover a broad spectrum, from minutes per game, which indicates how long players are on the court, to detailed shooting statistics like field goals made and attempted, including separate metrics for three-point and two-point attempts. These shooting metrics are complemented by percentages that provide a quick glance at shooting efficiency. Advanced scoring metrics include points per game and points per minute, offering both raw and normalized measures of scoring output. Free throw statistics are also detailed, highlighting scoring efficiency from the foul line.

Rebounding stats delineate players' abilities to recover the ball via on the offensive or defensive end and are essential for assessing contributions to team ball control. Defensive capabilities are further illustrated through steals

and blocks. These metrics show a player's defensive acumen and ability to facilitate team dynamics. Playmaking skills are quantified through turnover and personal foul counts, providing a measure of a player's discipline and aggressiveness on the court.

Notably, the dataset also includes an MVP indicator, signifying whether a player has won the Most Valuable Player award during the season, thus highlighting the peak performers. The dataset's temporal context is marked by the 'Season Start Year,' aiding in longitudinal studies and analysis of player evolution and performance trends over time.

This data has undergone rigorous preprocessing to ensure its reliability for in-depth analysis, including the removal of duplicates, comprehensive handling of missing values, and normalization of performance metrics to ensure fair comparisons across varied playing conditions. This prepared dataset serves as the foundation for deploying the DBSCAN clustering algorithm, aimed at uncovering intrinsic performance tiers and analyzing player trajectories within the NBA, offering nuanced insights into their capabilities and career progression.

B. Baseline Methods and Evaluation

1. DBSCAN

For the clustering of NBA players into distinct performance tiers, we employed the Density-Based Spatial Clustering of Applications with Noise (DBSCAN), a renowned algorithm favored for its proficiency in identifying outliers and generating clusters based on data density. This method is particularly suited for our analysis due to its ability to form clusters without pre-specifying the number of clusters, a significant advantage when dealing with diverse player performance data.

A. Implementation Details:

DBSCAN was implemented using the DBSCAN class from the sklearn.cluster module in Python. The primary parameters adjusted during the clustering process were:

- Epsilon (eps): This parameter defines the maximum distance between two samples for one to be considered as in the neighborhood of the other. It is crucial as it directly influences the growth of the clusters.

- Minimum Samples (min_samples): This parameter represents the number of samples in a neighborhood for a point to be considered as a core point. This includes the point itself, thus playing a key role in the sensitivity of the model to noise.

The initial settings of eps=0.5 and min_samples=5 were selected based on preliminary experiments which suggested these values allowed for a reasonable initial segregation of performance tiers.

B. Hyperparameter Tuning:

To refine our clustering approach and improve the meaningfulness of the performance tiers, we conducted a systematic tuning of the DBSCAN parameters. The fine-tuning process involved a grid search over a range of values for eps (from 0.1 to 1.0 in increments of 0.1) and min_samples (from 3 to 10). Each configuration was evaluated based on its silhouette score, which measures the cohesion and separation of the generated clusters:

- Silhouette Score: A metric used to evaluate the quality of clusters created by DBSCAN. A higher silhouette score indicates better-defined clusters.

The tuning revealed that an eps of 1.0 and min_samples of 3 provided a balanced outcome, generating five distinct clusters with a silhouette score of 0.4352. This configuration offered an optimal balance by producing a higher number of meaningful clusters with substantial separation, thereby enabling a more nuanced segmentation of player performances.

C. Performance Evaluation:

The optimized parameters led to the formation of clusters that were both meaningful and interpretable, with the silhouette score significantly improved from the initial trials, indicating more appropriate clustering with less overlap and clearer boundaries between clusters.

The DBSCAN clustering with these tuned parameters successfully categorized players into performance tiers that reflect their playing statistics and roles on the field, validating the effectiveness of our approach in

the sports analytics context. The results underscore the utility of DBSCAN in handling complex, real-world datasets where the relationships between observations are not straightforward or well-defined.

2. Hierarchal Clustering

Hierarchical clustering was employed to segment NBA players into distinct performance tiers based on their statistics. NBA players were grouped into performance categories using clustering based on their stats instead of using centroid based methods like before. This technique arranges the stats in a tree shaped structure called a dendrogram that shows how players are related in terms of performance levels clearly. It's useful because it lets you analyze stats at levels without needing to decide on the number of groups.

Implementation Details:

Implemented clustering by utilizing the AgglomerativeClustering class from the Python sklearn.cluster module. The following key choices were made during the implementation:

Linkage Criterion: The ward linkage method was used, as it minimizes the variance of clusters being merged, making it well-suited for the analysis of numerical player statistics.

Distance Metric: The Euclidean distance was chosen to measure the similarity between players based on their performance metrics.

In order to find the number of groups needed for analysis we tested numbers (ranging from 2, to 15) using silhouette scores to assess how well the clusters are connected and separated. The silhouette score provides insights into the effectiveness of the groupings created.

Results and Observations:

The silhouette analysis for different values of n_clusters yielded the following insights:

1) For n_clusters = 2, the highest average silhouette score of 0.3582 was observed, indicating the best-defined clusters.

2) Increasing the number of clusters to 3 and beyond resulted in a noticeable decline in the silhouette score, suggesting reduced cluster separability and overlap among clusters.

Based on this analysis, two clusters were selected as the optimal configuration for hierarchical clustering.

Performance Evaluation:

The silhouette score on the test data was 0.3772, indicating that the hierarchical clustering approach effectively captured the patterns in player performance across the dataset.

The hierarchical clustering method provided meaningful segmentation of NBA players into performance tiers, with two clusters offering the most interpretable grouping. The method's ability to adaptively adjust granularity and visualize relationships via dendrograms proved advantageous for the analysis of complex player performance data.

C. Evaluation Metrics and Performance

This part discusses the measurements used to assess how well the clustering models. K Means, DBSCAN and Hierarchical Clustering. We were able to differentiate between performance levels of NBA players.

1. Evaluation Metrics:
    a. The silhouette score is used in all types of clustering models to evaluate how closely the clusters are formed and how distinct they are, from one another by measuring their cohesion and separation levels based on the proximity of points within and between clusters. A higher silhouette score signifies defined clusters that are both tightly packed and easily distinguishable, from one another.

b. When we talk about inertia, in relation to the K Means model in particular we are essentially looking at how it measures the squared distances between every data point and the centroid closest to it. Lower values of inertia indicate that the clusters, in the model are together and distinct, which is what we aim for when working with algorithms.

2. Performance Overview:
    a. The K-Means clustering achieved moderately good silhouette scores (0.3116 for training and 0.3113 for testing), indicating reasonable separation and cohesion within the clusters. However, the inertia pointed to potential overlaps and less distinct cluster boundaries, suggesting room for improvement in cluster tightness and separation.
    b. DBSCAN showed significant improvement in the clustering quality with the highest silhouette score of 0.4352 after fine-tuning parameters. This model excelled in delineating player tiers with greater accuracy, particularly beneficial for datasets with complex and overlapping data distributions like ours.
    c. Hierarchical Clustering provided valuable insights into the data structure at various levels of granularity, with a silhouette score of 0.3772 on testing data, offering a robust method for understanding nested relationships and performing well in scenarios where cluster number is not predefined.

3. Comparative Analysis:
    a. DBSCAN outperformed other models in terms of the silhouette score, establishing it as the most effective method for this dataset due to its flexibility and robustness in handling outliers and varying densities within the data.
    b. The performance metrics collectively indicate that while each model has its strengths, DBSCAN's ability to adapt to the data's intrinsic structure makes it particularly suitable for the diverse and dynamic nature of NBA player statistics.

D. Ranking

This section outlines our approach and methods for ranking NBA players after they've been grouped into performance tiers through DBSCAN clustering. Ranking players within these tiers is crucial because it shifts our focus from simply categorizing players to meaningfully assessing and ordering them based on key performance indicators. This phase is essential for deriving practical insights from the clustered data and for assessing the clustering approach's effectiveness.

1. Implementation of Ranking:

   After the optimal clustering of players using DBSCAN, the ranking system was developed to evaluate players within each cluster. The primary goal was to assign a composite score to each player that reflects their overall performance, considering a variety of statistics such as points per game, assists, rebounds, and defensive metrics.

2. Aggregate Scoring System:

   The aggregate scoring system was introduced to calculate a weighted score for each player. The system considers multiple aspects of a basketball game, assigning different weights to various statistics:

   - Scoring Metrics (e.g., Points, Field Goal Percentage) received higher weights as they are direct indicators of a player's offensive capabilities.
   - Playmaking and Defensive Metrics (e.g., Assists, Steals, Blocks) were also weighted to reflect contributions beyond scoring.
   - Efficiency Metrics (e.g., Points Per Minute, Turnover Rate) were adjusted to penalize or reward players based on their efficiency and decision-making on the court.

   This scoring methodology ensures that the ranking captures a holistic view of player performance, emphasizing the multifaceted nature of basketball where different roles and skills contribute to a player's value.

3. Normalization:

   All metrics were normalized to ensure that each statistic contributes proportionately to the final score. This step prevents any single statistic from disproportionately affecting the player's score due to scale differences.

4. Ranking Within Clusters:

Players were then ranked within their respective clusters based on their aggregate scores. This intra-cluster ranking allows for a nuanced comparison among similarly performing players, providing insights into who stands out even among peers with comparable overall performance.

5. Application of Rankings:

In the refined analysis of our data mining project, we integrated a comprehensive ranking system post-DBSCAN clustering to deepen our understanding of player and team performances across various NBA seasons. Here's a detailed look at the functionalities incorporated into our ranking procedures:

Top Clusters: The system meticulously identifies and showcases top performers within each DBSCAN cluster based on their aggregate scores. This segment of our analysis serves to highlight the standout players in their respective performance tiers, providing a clear demarcation of elite performance within the clusters.

Team Ranking: Grouping per Team and Season: This function groups players by their respective teams and the season they played in, offering an analytical view of each team's roster for specific seasons. It assesses the collective effectiveness of team rosters, providing insights into the strategic assembly of players and their potential impact on the team's success during league competitions.

Ranked Team Average Per Year: Following the grouping of players by team and season, the system calculates an average score per player to rank teams. This ranking is instrumental in performing comparative analysis across teams and seasons, shedding light on the effectiveness of team strategies and roster decisions over time.

Best Roster for Each Team Each Year: In an effort to provide actionable insights for team management and coaching staff, our system also pinpoints the top five players for each team for every season, proposing an "ideal lineup". This optimal player combination is based on performance scores and is pivotal for strategic planning, helping to maximize team performance based on data-driven decisions.

Roster Creation Given a Year and a Team: Tailored for strategic team planning and game simulations, this functionality allows for the generation and ranking of all possible 5-player combinations for a specified team and year. It empowers teams to explore various lineup possibilities and their potential impacts on game outcomes, enhancing tactical decisions.

Through the implementation of these ranking methodologies, the project not only evaluates individual player performance but also offers a granular view of team dynamics and performance trends that can prove critical for:

- Team Management: Assisting team executives and coaches in understanding the strengths and weaknesses of their roster.
- Player Comparison: Offering a data-driven method to compare players, which is useful for trades, drafting, and player development.
- Trend Analysis: Examining how player rankings change over time and how different player archetypes evolve in the NBA.

The ranking section thus serves as a bridge between the theoretical clustering output and practical applications, enhancing the value of the data mining project by providing actionable insights that are accessible to both technical and non-technical stakeholders in the realm of basketball analytics.

D. Studies of the Need of Proposed Technical Components

In the context of sports analytics, particularly within the NBA, the ability to discern performance patterns across seasons is crucial for strategic decision-making related to player contracts, trades, and team building. Our project capitalizes on the integration of various technical components within our data mining pipeline to address these needs effectively.

The preliminary data pipeline incorporates Principal Component Analysis (PCA) to reduce the high dimensionality of NBA player statistics while retaining the variance necessary for accurate clustering. This reduction is pivotal, as it simplifies the data without losing critical information, enabling more effective clustering and visualization.

Clustering methods such as DBSCAN and Hierarchical Clustering are then employed to segment NBA players into performance tiers. The choice of DBSCAN was particularly motivated by its ability to handle outliers and noise within data sets, which is common in sports statistics due to varied player performances and roles. This method's flexibility in forming clusters based on data density allows for more natural grouping of players based on performance metrics without the constraint of pre-defined cluster sizes, unlike K-Means.

The hierarchical clustering approach provides insights into the data's nested structures, offering a granular view of player tiers that can be crucial for detailed analyses such as identifying potential breakout players or those at risk of declining performances.

Fine-tuning these models, as demonstrated through iterative adjustments in our experiments, ensures that the clusters formed are both meaningful and statistically significant, as indicated by improved silhouette scores. These scores are crucial as they provide a measure of how similar an object is to its own cluster compared to other clusters, thereby validating the cohesion and separation of the clustering approach.

VI. Related Work

The exploration of player clustering within the NBA has seen significant contributions through various scholarly efforts. Elam (2019) employed Principal Component Analysis (PCA) and K-Means clustering to delineate NBA players into eight distinct archetypes using detailed statistics from the 2018-2019 season. This study illuminates the power of advanced statistical methodologies in capturing player roles and efficacy, moving beyond traditional basketball positions to offer a deeper understanding of player contributions. [1]

Lutz (2012) explores the application of cluster analysis to evaluate player performance in the NBA, utilizing both hierarchical and k-means clustering to segment player data into meaningful performance categories. The study meticulously analyzes player statistics to group similar performance traits, thereby offering an objective method to gauge player effectiveness and team fit. The findings reveal clusters that categorize players not just by traditional positions but by their playing style and contributions, suggesting a model that can significantly enhance team strategies and player evaluations. This methodological approach helps in identifying undervalued players and optimizing team compositions, ultimately contributing to more strategic decisions in player recruitment and game planning. [2]

In the study conducted by Patel (2017), advanced statistical methods are used to redefine how NBA players are clustered, moving away from traditional position labels to more dynamic performance-based groupings. The study employs techniques like t-SNE to complement Principal Component Analysis (PCA), enhancing the ability to visualize and interpret the multidimensional data of player performance metrics. This approach not only underscores the diversity in player roles but also assists team strategists and coaches in identifying and leveraging unique player strengths that might be obscured by traditional statistical categories. The findings advocate for a tailored approach to team composition, emphasizing the strategic alignment of player skills with team needs, which can lead to more effective and competitive team dynamics. [3]

In their study, Addepalli et al. explore the utility of clustering algorithms in creating NBA player archetypes using statistical performance data from the 2018-2019 NBA season. They utilized K-means, Gaussian Mixture Models, and DBSCAN to analyze players' performance and identify distinct types, which could significantly enhance team strategies and player evaluations. Their research demonstrates how different clustering techniques can reveal complex patterns in sports analytics, making it a valuable tool for teams to optimize their rosters and game strategies based on robust data-driven insights. [4]

Together, these studies exemplify the expanding role of data analytics in sports, particularly basketball, where strategic and tactical decisions are increasingly driven by deep dives into statistical data, enabling a richer, more nuanced approach to player evaluation and team management.

VII. Conclusion

The data mining project aimed at clustering NBA players into performance tiers using DBSCAN and subsequent ranking of players and teams has provided deep insights into the complexities and dynamics of player performance across different seasons. By leveraging advanced clustering techniques and a detailed ranking system, the project was successful in identifying distinct performance groups, aiding in the nuanced understanding of player roles and capabilities within the NBA.

The use of DBSCAN allowed for the adaptive identification of clusters based on data density, which proved to be effective in dealing with the variability and complexity of NBA data. This method facilitated the discovery of natural groupings without the need for predefining the number of clusters, which is particularly advantageous given the diverse nature of the dataset. Fine-tuning the parameters of DBSCAN helped in refining the cluster quality, ensuring that the clusters were meaningful and representative of actual performance tiers.

The ranking segment of the project was instrumental in further dissecting the clusters to rank players within each performance tier and across the NBA landscape. By evaluating individual and team performances, the project could deliver strategic insights into player selection and team assembly that go beyond traditional metrics. The ranking systems deployed could enable stakeholders to identify top performers, optimal team rosters, and effective team strategies based on comprehensive data-driven assessments.

Furthermore, the project has laid a strong foundation for future enhancements and research. The methodologies and findings can be expanded with additional data, such as player health, in-game decisions, and more granular

performance metrics to refine player tiers and predictive models further. Additionally, incorporating machine learning techniques could automate some of the analysis, leading to real-time insights during seasons.

To conclude, this project highlighted the importance of advanced data analytics in sports and showed how data-driven strategies can be successfully applied in sports management and decision-making. The insights obtained from this project are crucial for teams aiming to improve their strategic planning and operational choices, which can ultimately enhance performance and provide a competitive edge.

IX. References

[1] M. M. Elam, "NBA Player Clustering: Exploring Player Archetypes in a Changing NBA," Asu.edu, May 2019. https://keep.lib.asu.edu/items/132157 (accessed Nov. 26, 2024).

[2] D. Lutz, "A CLUSTER ANALYSIS OF NBA PLAYERS." Accessed: Nov. 26, 2024. [Online]. Available: https://cdn.prod.website-files.com/5f1af76ed86d6771ad48324b/65552e92646d15a7a613b7a4_SSAC12%20-%20Lutz_cluster_analysis_NBA.pdf

[3] R. Patel, "Clustering Professional Basketball Players by Performance," escholarship.org, 2017. https://escholarship.org/uc/item/917739k8

[4] vjangili-26, "GitHub - vjangili-26/Clustering-methods-for-identifying-NBA-player-archetypes," GitHub, 2024. https://github.com/vjangili-26/Clustering-methods-for-identifying-NBA-player-archetypes

Section 9: **Link to code**

Code

(The main final code can be found in 'CSE_572_Final_Group_Project.ipynb' file uploaded at the link.)