NBA Archetypes Using Clustering Algorithms

Anand Addepalli
Department of Applied of Computing
Michigan Technological University
Houghton, USA
vaddepal@mtu.edu

Sahithi Bathini
Department of Applied of Computing
Michigan Technological University
Houghton, USA
sbathini@mtu.edu

Vaishnavi Jangili
Department of Applied of Computing
Michigan Technological University
Houghton, USA
vjangili@mtu.edu

***Abstract*—The NBA coaches and scouts need to identify player archetypes to build successful teams. With a vast amount of player performance data at their disposal, advanced analytical techniques can be used to find patterns that aid this process. In this paper we will be exploring clustering methods to group NBA players based on their statistical performance and identify player archetypes while investigating the traits that define them. The resulting archetypes can inform player evaluation, roster construction, and game strategy. Teams might prioritize signing players who fit a specific archetype to fill gaps in their roster. We are using a dataset consisting of player statistics from the 2018-2019 NBA season, K-means clustering, Gaussian clustering, DBSCAN will group players based on similarity in various statistical categories. This study aims to provide a more comprehensive understanding of NBA player performance using data analytics, enabling teams to make informed decisions about player evaluation, team building, and game strategies. The results of this research have implications not only for NBA teams but also for other organizations seeking to analyze large datasets and find patterns to guide decision-making.**

Keywords—archetypes, K means, Gaussian clustering, DBSCAN, data analytics.

I. INTRODUCTION

The basketball leagues require the teams to find the right players and assembling a strong team roster are essential for success in professional basketball. Accomplishing this, coaches and scouts must find the best players and develop player archetypes using the vast amounts of player performance data that are currently available and cutting-edge analytical techniques. This paper aims to identify player archetypes while examining the characteristics that define them by using clustering methods to group NBA players based on their statistical performance. It is possible to use the resulting archetypes to guide player evaluation, roster planning, and game strategy. Teams can prioritize signing players who fit a specific archetype to fill gaps in their roster by spotting common patterns in player performance data. K-means clustering, Gaussian clustering, and DBSCAN clustering will be used to group players based on similarity in different statistical categories using a dataset of player statistics from the 2018–2019 NBA season.

The aim of this paper is to use the concepts of data analytics to provide a more thorough understanding of NBA player performance, empowering teams to decide wisely about player evaluation, team building, and game strategies. In order to find patterns and similarities in large, complex datasets, clustering techniques are frequently used in data analysis [10]. This study offers important insights into the use of clustering techniques in identifying player archetypes by applying these methods to NBA player data. The findings of this study have implications for NBA teams as well as other businesses looking to analyse large datasets and identify patterns. The use of clustering techniques to analyse player performance data greatly enhance teamwork and player evaluation. The resulting archetypes can be used by teams to locate the best players to fill open roster spots and particular roles on the team. This can aid in creating game plans that play to each player's strengths to the fullest. Additionally, employing clustering techniques can considerably cut down on the time and work needed for player evaluation. It takes time to watch and analyse game footage when using the conventional methods of player evaluation. However, coaches and scouts can quickly identify player archetypes based on statistical performance by using clustering techniques, which cuts down on the time and work needed for player evaluation.

II. LITERATURE REVIEW

Clustering is a popular technique in data mining and machine learning that involves grouping data points into clusters based on their similarity. There are many different clustering algorithms and methods available, each with its own strengths and weaknesses. In this literature review, we will provide an overview of some of the most popular clustering methods and discuss their advantages and disadvantages that can be applied for our dataset and its analysis in the NBA players clustering.

In the study "Identifying and Characterizing Team Roles in the NBA using Player Tracking Data" by Lucey et al. (2014) [1], player roles are identified based on player tracking data using clustering techniques. In order to identify player roles like "rim protector," "mid-range scorer," and "spot-up shooter," this study used a k-means clustering algorithm.

In the study "Clustering and Prediction of Basketball Players Based on their Scoring Behaviours" by Jin et al. (2019) [2], player types are identified by using clustering techniques. In this study, player types like "shoot-first guards" and "rebounders and inside scorers" were identified using a Gaussian mixture model.

In their study "Clustering NBA Players Based on their Per-Game Statistics" published in 2018 [3], Kiritchenko and Mohammad used clustering techniques to categorize different player types based on per-game statistics. In this study, player types like "scoring point guards" and "versatile forwards" were identified using a k-means clustering algorithm.

Clustering techniques are used in the study "Unsupervised Learning Techniques for Characterizing Playing Styles in the NBA" by Akhtar et al. (2016) [4] to pinpoint player playing styles in the NBA. In order to distinguish between playing styles like "penetrators," "post players," and "perimeter shooters," this study used the DBSCAN clustering algorithm.

Overall, these studies show the value of using clustering techniques to pinpoint player archetypes and playing preferences in the NBA. The versatility of these methods for conducting various types of analyses on player data is highlighted by the use of various clustering algorithms, including k-means, Gaussian mixture models, and DBSCAN.

III. DATA EXPLORATION AND ANALYSIS

Our analysis aims to explore the advanced statistics of NBA players in the 2019 season, with a focus on identifying player archetypes based on their position. We obtained the data set from the official NBA website [5], which contains 30 different attributes for 488 instances (i.e., individual players).

The NBA statistics database is a valuable resource for evaluating players based on their performance on the court. The advanced statistics provided by the NBA go beyond basic metrics such as points, rebounds, and assists and instead include more detailed measurements of player performance, such as usage rate, true shooting percentage, and player efficiency rating. These advanced statistics can provide insights into a player's strengths and weaknesses and can be used to identify players who excel in specific areas of the game. To perform our analysis, we focused on the 2019 NBA season, which was a particularly competitive and exciting year for the league. We began by collecting data from the NBA website, which included player statistics broken down by position. We then performed exploratory data analysis to identify trends and patterns in the data, such as which positions tended to score more points or grab more rebounds.

The Exploratory Data (EDA) is one of the important steps to be performed in data mining. This step will help us visualize the data and how see how does it vary with the other instances, visualise the correlation and the distributions of the features with the given instances.

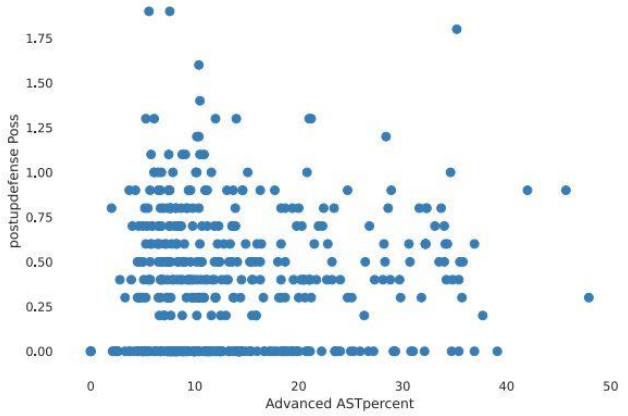

Figure:1 Defence up position v/s Advanced AST percentage.

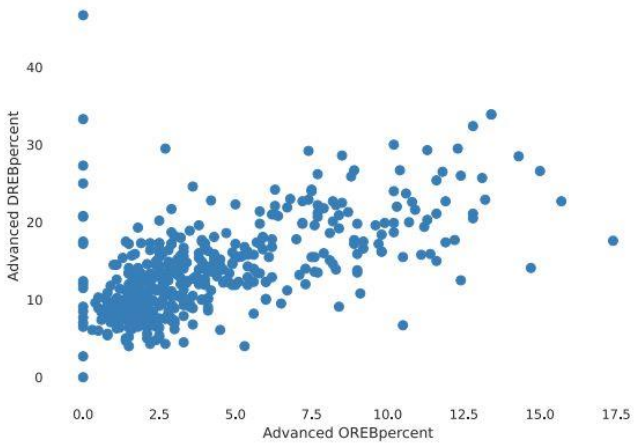

Figure:2 DREB v/s OREB percentages

The Figure-1 &2, shows the scatter plots for the features Post up defence Position v/s the Advanced AST percentage and the second shows the ball switching happening between the defence and offense teams. The first scatter plots give us information that some of the features are highly correlated and unchanged even with the changes of AST percentage.

The correlation among the features can be better visualized using the heatmap which gives us a better understanding. In particular we get range of numbers between -1 and 1. The values that were highly correlated are to be eliminated in the clustering methods. These can be shown in the Figure-3.

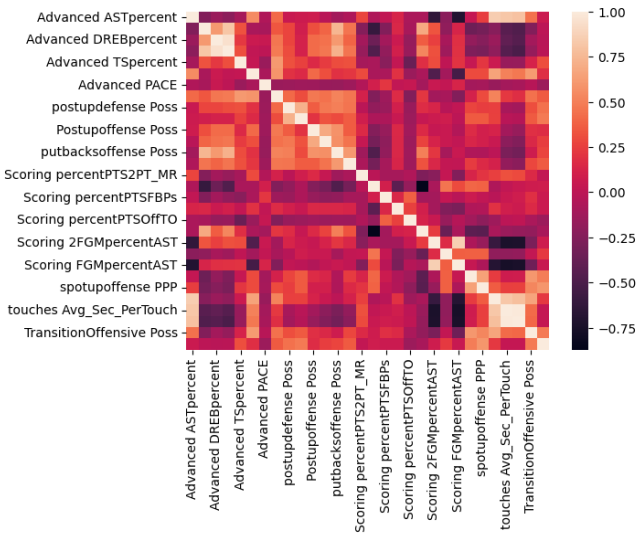

Figure:3 Heatmap representing the correlation among the features.

IV. DATA MODELLING

The first step in any Machine learning process is the data cleaning which involves pre-processing steps like exploring the dataset to visualize the distributions for various attributes and their correlation with each other and the other features. This step will give us insights about the data. From this initial analysis, we found that some of the features are highly correlated. The distributions of most of the features are found to be normally distributed with some peaks also been observed in some of them. For the elimination of the highly correlated features and some of the duplicated data we have utilized the concept of Principal Component Analysis (PCA). PCA is a statistical technique that is commonly used for dimensionality reduction. When dealing with a high-dimensional dataset, PCA can be applied to reduce the number of dimensions while retaining as much information as possible. This is achieved by finding a new set of orthogonal variables, called principal components, that capture the most significant variations in the data. The first principal component explains the largest amount of variance in the data, followed by the second principal component, and so on. By retaining the first few principal components, we can effectively reduce the dimensionality of the dataset.

Moreover, PCA is also useful in dealing with highly correlated features. When two or more features are highly correlated, it means that they are measuring similar aspects of

the data. In such cases, PCA can be used to transform these features into a smaller set of uncorrelated features, i.e., principal components. These principal components represent the underlying patterns in the data and can be used instead of the original features. By doing this, we can reduce the redundancy in the dataset and avoid overfitting while still retaining most of the important information. Explained variance ratio is an important metric in PCA that measures the amount of variance in the data that is explained by each principal component. The explained variance ratio of a principal component is equal to the proportion of the total variance in the data that is accounted for by that component.

By examining the explained variance ratios of the principal components, we can determine how much of the total variance in the data is being captured by each component. In our project principal components explain a high percentage of the total variance (nearly 0.93), and we are confident that these components are important and can be used to effectively reduce the dimensionality of the data.

The next step in our procedure would be to find the number of components in the PCA. For this we have initially plotted the explained variance ratio and the number of components we could get a graph as shown in the Figure- 4a. From this graph we could not to any meaningful conclusions, therefore we have taken the difference of the slopes and plotted the differential graph and it can be show in Figure-4b. This graph is popularly known as the **elbow curve** and we could come to a conclusion that the optimal value of number of components required for the PCA analysis is around 13.

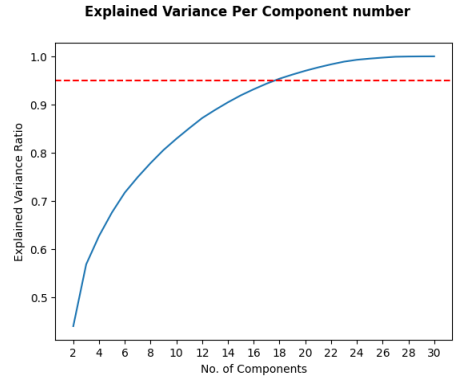

Figure 4a: Variance v/s number of components

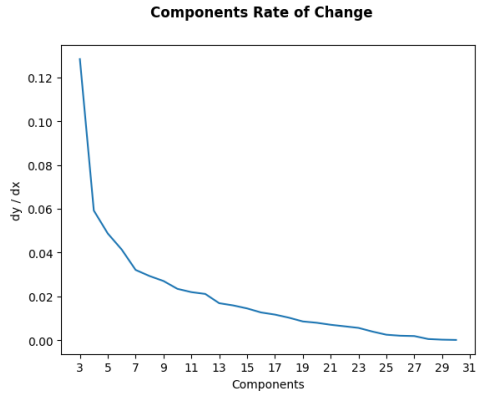

Figure 4b: Difference in Variance v/s number of components

We have used the 13 components and divided the scaled data into 13 groups and verified for the total value of explained variance ratio and found to be 0.93. This verified that the objects and the number of groups are highly related. We could therefore proceed to the next step which involved the evaluation of the Kmeans for which we have used the silhouette score.

Evaluation of the **silhouette score** has been done by taking a range of number of cluster values. After this we have applied to K means algorithm [6] to plot the various values of the scores for each number. This can be shown from the Figure-6. In the given figure coming to any meaningful conclusion was another challenge since there can be seen that lot of spikes made it difficult to identify the optimal number of clusters to be fed to the algorithm. To overcome this problem, we have taken the difference of the slopes from the graph and plotted the differential graph. Although, we could see that again there is no smooth graph observed even from the differential graph we could somehow see the optimal number of clusters is around 14. The final step would be to apply to the k means algorithm and divide the data into the different clusters. After this was plotted using the matlabplotlib and seaborn functions.

The other algorithm that we explored is the Gaussian modelling Models (GMMs) [8]. For this algorithm the first and simple main approach is that the data is assumed to be distributed normally or it follows the gaussian distribution. We have seen this condition in the beginning of the exploratory data analysis that the data follows the Gaussian distribution. We have taken the number of clusters here 14 as well from the values calculated from the above score.

Density based [7] is another clustering method that we utilized in the project. The algorithm then iterates through each data point, checking whether it has already been assigned to a cluster. If the point has not been assigned, the algorithm finds its neighbors within a distance of **eps**. If the number of neighbors is less than **min_samples**, the point is labelled as noise and the iteration continues with the next point. If the number of neighbors is greater than or equal to **min_samples**, a new cluster is formed, and the point and its neighbors are assigned to the cluster. The algorithm then expands the cluster by recursively adding more neighbors to the cluster until there are no more neighbors to add

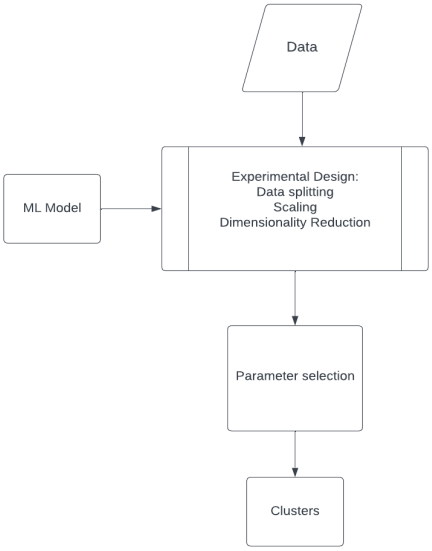

Figure:5 The flowchart represents the modelling of the algorithm development.

V. RESULTS AND DISCUSSION

After having performed the modelling and diving the objects into clusters it is important to analyze the results and see how the different algorithms have performed on the dataset. Generally, we have got numerous evaluation metrics for the classification problems like accuracy, precision, etc. On the other hand, for the clustering problems we can see a limited metrics to evaluate the performance of the model.

The first and most important metric for the evaluation is the silhouette score. Silhouette score [9] is a metric used to determine the quality of clustering in K-means algorithm. It measures how similar an object is to its own cluster compared to other clusters. The silhouette score ranges from -1 to 1, where a higher score indicates better clustering.

The formula for silhouette score is as follows:

$$\text{silhouette score} = \frac{b-a}{\max(a,b)} \quad (1)$$

where:
a is the average distance between a data point and all other points in the same cluster.
b is the average distance between a data point and all points in the nearest cluster.

A silhouette score of 1 indicates that the clustering is dense and well-separated, whereas a score of -1 indicates that the clustering is not appropriate. A score close to 0 indicates that the clustering is overlapping. In K-means algorithm, the optimal number of clusters is often determined by finding the highest silhouette score. This can be achieved by computing the silhouette score for different values of k, and selecting the value of k that gives the highest score.

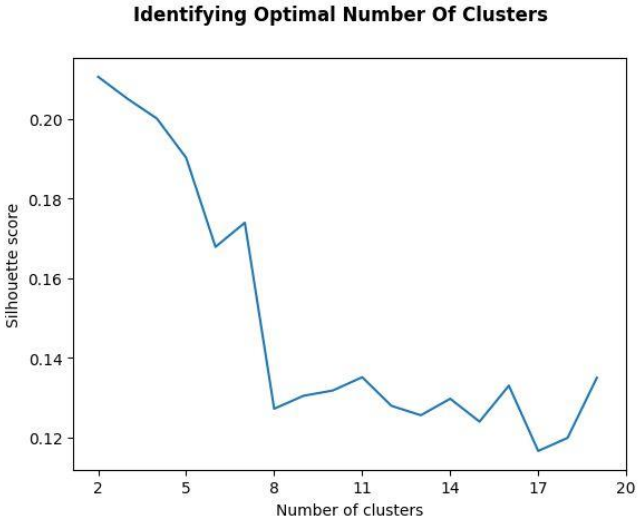

Figure: 6 The graph shows the variation of silhouette score and the number of clusters.

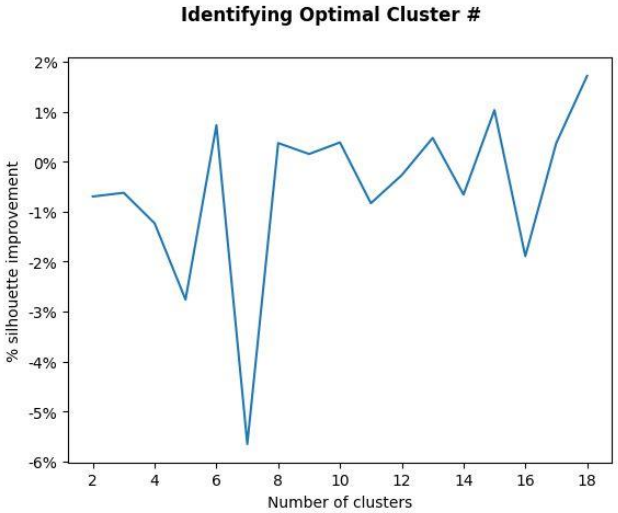

Figure: 7 Dereferences of the slopes of silhouette score and the number of clusters

The above two graphs help us in determining the number of clusters that is explained in the previous section

After having calculated the given scores for the all models an analysis was done which is described in the given table. Although, the values of the silhouette scores have been calculated individually. At the end we have calculated the mean of all the scores and reported it. This gives us the better analysis for different models.

Table:1 Performance Analysis

S No.	Analysis of the performance of models		
	Clustering model	***Metric***	***Score***
1	K-means	Silhouette	0.13
2	DBSCAN	Silhouette	0.12
3	Gaussian Models	Silhouette	0.11

The clusters of the given algorithms have been mentioned below which describes classes into which the objects have been divided into.

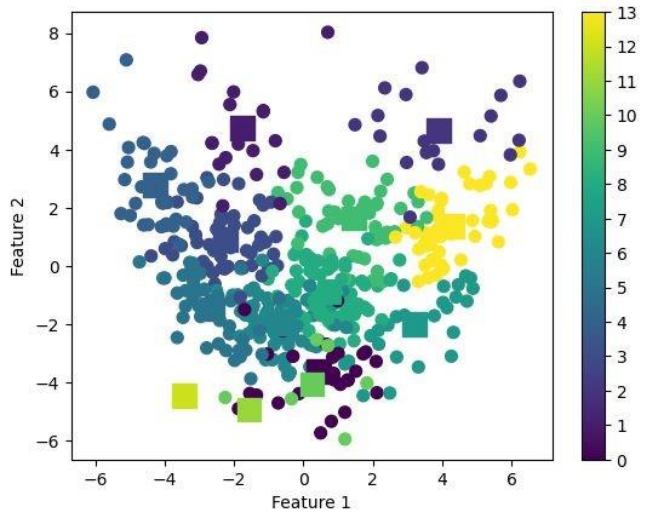

Figure 8a: Kmeans Clusters

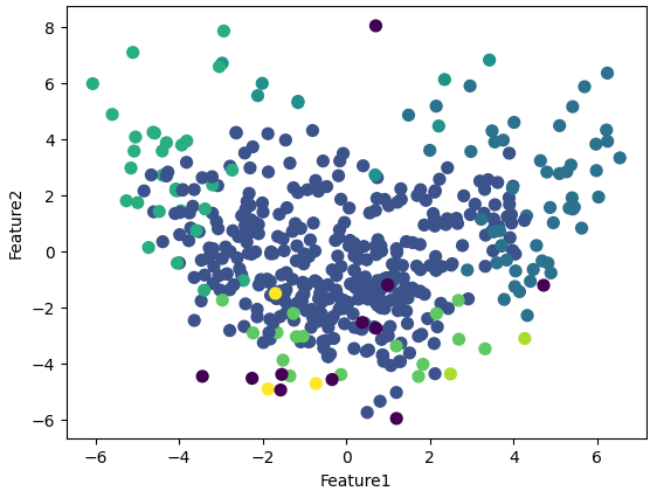

Figure 8b: DBSCAN Clusters

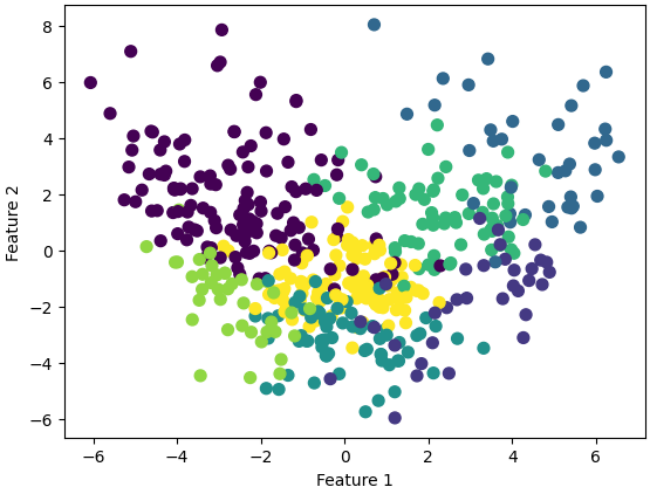

Figure 8c: Gaussian Model Clusters

Figure-8

VI. CONCLUSION AND FUTURE WORK

We have successfully implemented the clustering algorithms to the NBA dataset and the performance of the models was assessed as well. We have observed that, although the DBSCAN model is performing good on the given dataset and had also clustered well with reasonably good score. But it could have performed better if the metrics were good and the hyperparameters were better optimized. We are currently studying the underlying mathematics behind the models and the implementation will be done soon. Apart from that, we are also exploring other models which are better suitable and can perform better than the current ones.

VII. ACKNOWLEDGEMENT

While all the models have been successfully implemented and executed well, we show our gratitude towards our professor for all his hard work and the opportunity he has given us to select such an interesting topic and explore such areas. The theory was explicitly taught well which helped us to understand and manipulate the algorithms.

REFERENCES

[1] J. Lucey, A. Bialkowski, P. Carr, S. Matthews, and I. Simon, "Identifying and characterizing team roles in the NBA using player tracking data," in IEEE Transactions on Visualization and Computer Graphics, vol. 20, no. 4, pp. 549-558, April 2014.

[2] Y. Jin, S. Li, and X. Li, "Clustering and prediction of basketball players based on their scoring behaviors," in Journal of Sports Analytics, vol. 5, no. 3, pp. 139-152, Sep. 2019.

[3] S. Kiritchenko and S. M. Mohammad, "Clustering NBA players based on their per-game statistics," in Proceedings of the First Workshop on NLP and Sports, 2018, pp. 30-39.

[4] M. S. Akhtar, S. S. Ahmed, and A. El Saddik, "Unsupervised learning techniques for characterizing playing styles in the NBA," in Proceedings of the 2016 IEEE International Conference on Image Processing, 2016, pp. 3051-3055.

[5] NBA Advanced Stats: Player Analysis" by NBA.com: https://stats.nba.com/players/advanced/ Nicole.

[6] A. Stern, "HoopDown," [Online]. Available: https://alexcstern.github.io/hoopDown.html. [Accessed: Apr. 26, 2023].

[7] G. Hu, Y. Xue, and W. Shen, "Multidimensional heterogeneity learning for count value tensor data with applications to field goal attempt analysis of NBA players,"arXiv preprint arXiv:2205.09918, May 2022 arXiv preprint arXiv:2205.09918, May 2022.

[8] F. Yin, G. Hu, and W. Shen, "Analysis of professional basketball field goal attempts via a Bayesian matrix clustering approach,"arXiv preprint arXiv:2010.08495, Oct. 2020.

[9] K. R. Shahapure and C. Nicholas, "Cluster Quality Analysis Using Silhouette Score," *2020 IEEE 7th International Conference on Data Science and Advanced Analytics (DSAA)*, Sydney, NSW, Australia, 2020, pp. 747-748, doi: 10.1109/DSAA49011.2020.00096.

[10] Zhang, Shaoliang & Lorenzo Calvo, Alberto & Ruano, Miguel & Mateus, Nuno & Gonçalves, Bruno & Sampaio, Jaime. (2018). Clustering performances in the NBA according to players' anthropometric attributes and playing experience. Journal of Sports Sciences. 36. 1-10. 10.1080/02640414.2018.1466493.