Prediction the best shooters in the NBA

Ahmet HACIOĞLU - 041801123

MEF University, Engineering Faculty, Computer Engineering

COMP 466, BUSINESS INTELLIGENCE

Prof. Adem KARAHOCA

1 June, 2023

Abstract

This research aims to predict the best shooters in the NBA. The dataset contains 29 different attributes and 217 different instances(these attributes and instances consist of statistical information from NBA 2021-2022 playoffs). The dataset has special metrics for player performance and this research purpose of various machine learning algorithms using the weka tool. The classification part including Naive Bayes, J48, Random Forest, SMO, K-nearest neighbor and Random Tree classify algorithms and the clustering algorithm are also applied for the dataset and include Simple k-means, Make Density Based Clustering, Farthest First and Filtered Clusterer clustering algorithms. The Cluster algorithm aims to group players based on data for their shooting abilities.By conducting this research, good shooters can be predicted using different machine learning algorithms with the help of analysis of NBA playoff data and also this project could be useful for future research.

Keywords: NBA, machine learning algorithms, Weka, Classification, Clustering, Shooting, Naive Bayes, J48, Random Forest, SMO, K-nearest neighbor, Random Tree, Simple k-means, Make Density Based Clusterer, Farthest First, Filtered Clusterer

1. Introduction

Data mining is a computer-assisted technique used for analytics to process large data sets[1]. It is known that data mining and data science play an active role in many industries such as healthcare, sports, BFSI, automotive, cyber security, digital marketing and many more[2]. Data mining plays a crucial role in extracting meaningful insights and from large datasets in the field of sports[3].

Basketball is a complex game and math, data mining, data analysis, data visualization, data science and similar subjects play a very important role and these topics are heavily supported in the NBA league, so the game evolves over time[4]. Data mining has gained an important place in the basketball industry. Thus, NBA teams make significant investments. Data mining provides significant assistance to issues such as individual athlete performance and team analysis. Important results are obtained by many trainers and managers using data mining and algorithms.

In this paper, machine learning algorithms were used to predict the best shooter in the NBA from NBA Play-off statistics using the weka tool. With the help of this prediction modeling, we can arrive at the algorithm that will predict the best shooter. This study has lots of classification and clustering algorithms such as Naive Bayes, J48, Random Forest, SMO, K-nearest neighbor, Random Tree, Simple k-means, Make Density Based Clustering, Farthest First and Filtered Clusterer. This paper comprises four parts. The first part is the introduction. In the second part, deep into the research method and how algorithms work and compare algorithms to each other. In the third part, this paper analyzes the results and the fourth part gives a conclusion.

2. Research Method

Different machine learning algorithms were tested using the Weka tool. The data set contains 29 attributes and 217 instances from the NBA 2021-2022 play-off. Attributes contain statistical information and these statistics have important metrics. The subsections contain information about the data set and the algorithms used.

Table 1. Initial data set attributes

Attribute	Description
FULL NAME	Player’s name and surname
TEAM	Player’s team
POS	Player’s position
AGE	Player’s gender
GP	Player’s games played
MPG	Player’s minutes per game
MIN	Player’s minutes
USG	Player’s usage rate
TO	Player’s turnovers
FTA	Player’s free throws attempted
FT%	Player’s free throw percentage
2PA	Player’s 2-point attempted
2P%	Player’s 2-point percentage
3PA	Player’s 3-point attempted
3P%	Player’s 3-point percentage
eFG	Player’s effective field goal percentage
TS	Player’s true shooting
PPG	Player’s points per game
RPG	Player’s rebounds per game
TRB	Player’s total rebound percentage
APGA	Player’s average efficiency rate
AST	Player’s assist per game
SPG	Player’s steals per game
BPG	Player’s blocks per game
TOPG	Player’s turnovers per game
VI	Player’s versatility index
ORTG	Player’s offensive rating
DRTG	Player’s defensive rating
Results	Player’s are good or bad shooter

2.1 Data gathering and processing

The data set was taken from the kaggle site by applying web scraping operations with python. Firstly, data set comes in .csv file and then .csv file converted to .arff file according to some operations on weka tool. After operations, the data set has 217 instances from nba in different teams and 29 different attributes. These attributes are almost great statistics variables. Table 1 lists these initial data sets attributes with their description.

2.2 Classification algorithms and clustering algorithms

This part gives short explanations about algorithms which are used in this paper. First six algorithms will be classification algorithms and then clustering algorithms come from after the first six algorithms.

2.2.1 Naive Bayes

Naive Bayes is a simple and effective statistical machine learning algorithm. This algorithm is based on Bayes theorem and commonly used for classification tasks for predicting the probability for real-world applications.

2.2.2 J48

J48 is another popular machine learning algorithm for classification. Also, J48 known as a decision tree algorithm and decision tree is most commonly used for recursive data sets whose values are different. Decision tree algorithms could easily give some features about missing values.

2.2.3 Random Forest

Random forest is useful if we need to combine multiple decision trees. All subsets are combined for the final decision.It is useful in high-dimensional and complex datasets.

2.2.4 SMO

SMO means vector machine. Generally, used to train SVM (Support Vector Machine) with Sequential Minimal Optimization (SMO) and SMO usually used for optimization. If we are going to do binary classification, it is an important algorithm that has been proven to be correct.

2.2.4 Random Tree

Random forest and random tree are pretty close to each other. Random tree uses random sets of tree algorithms to make predictions and makes inferences over the sets. Finally, a random tree has computational efficiency for a large data set. If the data set is large, the Random tree algorithm could be effective.

2.2.6 K-nearest neighbor

Another popular machine learning algorithm is K-nearest neighbor. It works by trying to find the nearest k points and this algorithm is an intuitive algorithm. If the data set is simple and if research wants to quickly review, using the K-nearest neighbor algorithm could be effective. Importantly, accuracy rate could be lower than other algorithms.

2.2.6 Simple k-means

This algorithm is an unsupervised algorithm for clustering. With k the algorithm randomly pulls a specific initial point and algorithm then iteratively assigns each data point to the nearest center.

2.2.6 Make density based clustering

Also this algorithm selects starting points randomly and connects specific remote points.

2.2.6 Farthest First

This algorithm selects the first point from the dataset using k-means sets and this algorithm does not guarantee an optimal solution set.

2.2.6 Filtered Clusterer

If we need a clustering algorithm for data analysis and pattern, this algorithm will be fine. When we need filters in general, filtered clusterer can come to our rescue.

2.3 Comparing algorithms

In general, the dataset used is not complex. This data set has a certain order because of this situation, we could have continued with certain algorithms, but this research paper uses different algorithms and focuses on different metrics.

While working on algorithms, focused on important metrics such as RMSE, ROC, accuracy rate and sensitivity. These metrics will be sufficient to classify the performance of algorithms and also whenever we look at the cluster part, the focus is on how clustering is done there.

RMSE is Root Mean Square Error; it measures the average difference between prediction values. Roc is Receiver Operating Characteristic; graphical representation of the performance when we use it for binary classification. Accuracy rate is the most strong metrics for algorithms because these metrics give general overview. Sensitivity means recall or true positive rate, sensitivity metrics could show us how to project positive instances correctly.

2.4 Model training

The weka[6] tool was used to run the classification and clustering algorithms. The results obtained are explained comparatively in the next section.

3. Results and analyze

As a result, this part will dive into deep for this project results via Weka tool. Focused on accuracy rate, RMSE, ROC, Precision parameters. These metrics provide great information about the success of machine learning algorithms.

Table 2. Data mining/machine learning algorithms & techniques for Accuracy Rate

Algorithm name	Accuracy Rate
Naive Bayes	%80.18
J48	%100
Random Forest	%99.53
SMO	%88.01
K-nearest neighbor	%64.05
Random Tree	%81.1

Table 3. Data mining/machine learning algorithms & techniques for RMSE

Algorithm name	RMSE
Naive Bayes	0.4266
J48	0
Random Forest	0.2741
SMO	0.3461
K-nearest neighbor	0.5965
Random Tree	0.3856

Table 4. Data mining/machine learning algorithms & techniques for ROC

Algorithm name	ROC
Naive Bayes	0.847
J48	1.000
Random Forest	1.000
SMO	0.874
K-nearest neighbor	0.628
Random Tree	0.865

Table 5. Data mining/machine learning algorithms & techniques for Precision

Algorithm name	Precision
Naive Bayes	0.802
J48	1.000
Random Forest	0.995
SMO	0.884
K-nearest neighbor	0.639
Random Tree	0.825

In the current model, six different algorithms are used for classification. When we look at table two, we can see the accuracy rates and when we look at the accuracy rates, we see that the J48 algorithm works extremely well in the current model and also Random forest was awesome for this model. Looking at the remaining models, the k-neighbor accuracy rate is lower than other algorithms but mostly machine learning algorithms performed well in this model.

Table 3 gives us RMSE values and RMSE values can give average value for error according to the prediction model but in this table we can see that correlation between table 2 and table 3 are great.

Other metrics ROC(table 4) and Precision(table 5) give us correlation between all other algorithms. When we looked at these tables, we can see that mostly all algorithms performed well. According to this model, J48 and also random forest machine learning algorithms are great for choosing the best algorithm. Also Naive Bayes, Random Tree and SMO algorithms are good enough for prediction for this project.

Table 6. Data mining/machine learning algorithms & techniques for Simple k-means

Real-Project value - 0	Real-Project value - 1	Cluster Algorithm-0	Cluster Algorithm-1
99	118	124	93

Table 7. Data mining/machine learning algorithms & techniques for Make Density Based Clusterer

Real-Project value - 0	Real-Project value - 1	Cluster Algorithm-0	Cluster Algorithm-1
99	118	122	95

Table 8. Data mining/machine learning algorithms & techniques for Farthest First

Real-Project value - 0	Real-Project value - 1	Cluster Algorithm-0	Cluster Algorithm-1
99	118	139	78

Table 9. Data mining/machine learning algorithms & techniques for Filtered Clusterer

Real-Project value - 0	Real-Project value - 1	Cluster Algorithm-0	Cluster Algorithm-1
99	118	124	93

In this project, the data set has 118 great shooters and 99 instances for not great shooters. When we look at clustering algorithms for creating good sets. Table 6, table 7, tables 8 and table 9 show us the result of the clustering method. We could easily say simple k-means clustering and filtered clustered were the same result and make density based clustering methods almost close to these methods. The worst clustering algorithm was performed by the Farthest First algorithm.

5. Conclusion

In this research, many different clustering and classification machine learning algorithms were used. The aim of the project was to group the best shooting players using clustering algorithms such as Simple k-means, Make Density Based Clustering, Farthest First, Filtered Clusterer and provide their analysis with using classification algorithms such as Naive Bayes, J48, Random Forest, SMO, K-nearest neighbor and Random Tree. All methods include awesome statistical points from the NBA 2021-2022 season play off.

According to this study, J48 and random forest performed very well. These algorithms achieved high accuracy to predict the best shooters in the NBA from 2021-2021 play-offs. Random tree and SMO also performed well, especially accuracy rate and other metrics such as Sensitivity, ROC and RMSE. These algorithms can be used for future prediction of choosing the best shooters and other research topics. Using these algorithms can be useful features. The clustering algorithms showed good results, but they did not come very close to the result. These algorithms tried to separate players based on their shooting abilities. Finally, Simple k-means and Filtered Clusterer performed better than the other clustering algorithms.

Thoroughly, this project indicates the capabilities of machine learning algorithms in predicting NBA player shooting skills. Using the algorithms, teams can identify the right strategies for themselves. Finally, the algorithms used in the study can be developed in real-world basketball applications.

References

[1] J. D. Kelleher and B. Tierney, “Data Science,” MA: The MIT Press, 2018, pp. 1–39.

[2] R. Karan, “Top Industries Hiring Data Scientists in 2022,” naukri.com, March 22, 2022. [Online] Available: https://www.naukri.com/learning/articles/top-industries-hiring-data-scientists/. [Accessed: May 5, 2022].

[3] Demenius, J. and Kreivytė, R., 2017. “THE BENEFITS OF ADVANCED DATA ANALYTICS IN BASKETBALL: APPROACH OF MANAGERS AND COACHES OF LITHUANIAN BASKETBALL LEAGUE TEAMS,” Baltic Journal of Sport and Health Sciences, 1(104), pp.8-13.

[4] Schuhmann, J., 2021, October 14., “NBA’s 3-point revolution: How 1 shot is changing the game,” [Online]. Available: https://www.nba.com/news/3-point-era-nba-75. [Accessed: May 5, 2022].

[5] Li, B., & Xu, X., 2021, “Application of Artificial Intelligence in Basketball Sport. Journal of Education,” Health and Sport, 11(7), pp.54–67.

[6] I. H. Witten, et al., “The WEKA Workbench. Online Appendix for Data Mining: Practical Machine Learning Tools and Techniques,” Morgan Kaufmann, 2016.