A Mixed Model for Performance-Based Classification of NBA Players

Yeong Nain Chi[1], and Jennifer Chi[2]
[1]University of Maryland Eastern Shore, [2]University of Texas at Dallas
[1]ychi@umes.edu, [2]jxc126831@utdallas.edu

Corresponding author email: ychi@umes.edu

***Abstract*-Using data collected from the Basketball-Reference.com, this study examined NBA player performance values to discern patterns and to classify clusters exhibiting common patterns of player performance. Empirical results based on the K-means clustering analysis identified three NBA player clusters. Results of the K-means clustering analysis were tested for accuracy using the discriminant analysis indicated that cluster means were significantly different. The results of one-way ANOVA also showed that significant differences in all twenty-one independent variables were found within the three identified NBA player clusters. The multilayer perceptron neural network model was utilized as a predictive model in deciding the classification of NBA players based on their performance related statistics. From an architectural perspective, it showed a 21-7-3 neural network construction. Results of this study may provide insight into the understanding of the performance of NBA players for NBA management purposes.**

***Keywords*-NBA, Player Performance, Classification, K-means Clustering Analysis, Discriminant Analysis, One-Way ANOVA, Multilayer Perceptron Neural Networ**k.

1. Introduction

Basketball is an outstanding athletic sport where people all around the world can enjoy regardless of whether they are poor or rich, young or old, and even, different race or ethnicity. Basketball knows no boundaries in which any person could come up with new styles and skills that could be played in friendly games and tournaments. As the popularity of basketball continues to grow internationally, the National Basketball Association's (NBA) goals of globalization continues to market basketball towards consumers in the short term.

The NBA is a professional basketball league comprised of thirty teams across North America, where twenty-nine are located in the United States and only one in Canada, featuring the best basketball players in the world. The league originated in New York City on June 6, 1946 as the Basketball Association of America (BAA), adopted the name National Basketball Association in 1949 after merging with the rival National Basketball League (NBL) (nba.com).

Each NBA team can have a maximum of fifteen players, thirteen of which can be active each game. The NBA regular season tips off on the sixteenth of October to the tenth of April. During the regular season, each team plays a total of eighty-two games, forty-one home games and forty-one away games. A team faces opponents in its own division four times a year (sixteen games), teams from the other two divisions in its conference either three or four times (thirty-six games), and teams in the other conference twice (thirty games) (nba.com).

The NBA Playoffs begin in late April, with eight teams in each conference going for national championship. The final playoff round, a best-of-seven series between the victors of both conferences, is known as the NBA Finals, and is held every year in June. The victor in the NBA Finals wins the Larry O'Brien Championship Trophy. Each player and major contributor to the NBA season, including coaches and the general manager on the winning team, receive a championship ring. In addition, the league awards the Bill Russell NBA Finals Most Valuable Player Award to the best performing player of the season (nba.com).

NBA games are available on television in more than two hundred countries around the world, including hundreds of national broadcasts in the USA every year on ABC, ESPN, TNT and NBA channels. Fans can also watch games live and on-demand on NBA LEAGUE PASS around the world (nba.com). On the world stage, basketball is one of the most popular sports, trailing after soccer. With expanding viewership, revenue in the NBA has significantly grown. In fact, in the 2017-2018 season, the thirty NBA teams generated $7.4 billion in revenue. Basketball related income includes broadcast rights, advertising, merchandising, and concessions, among other categories [4].

NBA basketball is a highly competitive team game. In order to win the game, all effective basketball statistics and key metrics can serve as powerful tool to help players and coaches improve. Therefore, each team wants to recruit the best performance players in the team who also can put all the puzzles together for the team to win the game. For this purpose, each team gathers their general managers, scouts, and professional consultants to closely track the crucial game statistics of each player. In the analytical field, classification of the NBA players is an important task because it helps identify key vital players of the league, often a cut above the others in the league. Some are somehow hanging in the league, going on and off the court while others are in the process of becoming the next NBA star. Therefore, not every professional NBA player in the league has the same skill level, impact on the game and brand power.

With emerging fields in data science and technological advancement, teams have the ability to gather ever more data information on their players. By using deep learning, teams can classify NBA players by their performance related statistics into natural clusters that best match their skill sets. Thus, the primary purpose of this study was to be able to statistically classify the players into different clusters based on their skills and performance. This can help identify and predict a player's category on where he fits. The results of this study may provide features that are most important to an individual player, and group them in such a way that is easily interpretable and inherently understood by the players themselves, coaches, team managers, and fans alike.

2. Materials

The data for this study is available to the public in an open-access website from Basketball-Reference (http://www.basketball-reference.com/) with individual NBA player's performance related statistics during the 2018-2019 season. For the purpose of this study to classify various clusters of NBA players, the 2018-2019 Stats: Per Game was used to define a player's performance values.

The data extracted from Basketball-Reference comprised of 468 NBA players' performance related statistics in the 2018-2019 season, including Games (G), Games Started (GS), Minutes Played Per Game (MP), Field Goals Per Game (FG), Field Goal Attempts Per Game (FGA), 3-Point Field Goals Per Game (3P), 3-Point Field Goal Attempts Per Game (3PA), 2-Point Field Goals Per Game (2P), 2-Point Field Goal Attempts Per Game (2PA), Effective Field Goal Percentage (eFG%), Free Throws Per Game (FT), Free Throw Attempts Per Game (FTA), Offensive Rebounds Per Game (ORB), Defensive Rebounds Per Game (DRB), Total Rebounds Per Game (TRB), Assists Per Game (AST), Steals Per Game (STL), Blocks Per Game (BLK), Turnovers Per Game (TOV), Personal Fouls Per Game (PF), Points Per Game (PTS).

The descriptive statistics of 2018-2019 NBA Player Stats: Per Game can be shown in the Table 1. During the 2018-2019 season, average minutes played per game was 20.79 with the standard deviation of 8.51; average points per game was 9.34 with the standard deviation of 6.05; average field goals per game was 3.46 with the standard deviation of 2.18; average 2-point field goals per game was 2.51 with the standard deviation of 1.83; average 3-point field goals per game was 0.95 with the standard deviation of 0.80; and average free throws per game was 1.47 with the standard deviation of 1.34.

Table 1: Descriptive Statistics of NBA Player Stats: Per Game in the 2018-19 Season

Stats	Term	Mean	Standard. Deviation
G	Games	54.42	22.91
GS	Games Started	26.23	28.44
MP	Minutes Played Per Game	20.79	8.51
FG	Field Goals Per Game	3.46	2.18
FGA	Field Goal Attempts Per Game	7.58	4.53
3P	3-Point Field Goals Per Game	0.95	0.80
3PA	3-Point Field Goal Attempts Per Game	2.73	2.10
2P	2-Point Field Goals Per Game	2.51	1.83
2PA	2-Point Field Goal Attempts Per Game	4.86	3.36
eFG%	Effective Field Goal Percentage	0.51	0.80
FT	Free Throws Per Game	1.47	1.34
FTA	Free Throw Attempts Per Game	1.94	1.67
ORB	Offensive Rebounds Per Game	0.90	0.81
DRB	Defensive Rebounds Per Game	3.01	1.88
TRB	Total Rebounds Per Game	3.91	2.52
AST	Assists Per Game	2.08	1.80
STL	Steals Per Game	0.66	0.40
BLK	Blocks Per Game	0.43	0.41
TOV	Turnovers Per Game	1.16	0.79
PF	Personal Fouls Per Game	1.85	0.74
PTS	Points Per Game	9.34	6.05

3. Methods

In this study, a mixed model was introduced, which k-means clustering analysis for data examination, discriminant analysis for classification, and neural networks for prediction. Methodologically, K-means is one of the simplest unsupervised learning algorithms that solve the well-known clustering problem. K-means [9] is an iterative algorithm that form groups of observations around geometric centers called centroids into clusters. The algorithm calculates the centroids, which is determined by the individual conducting the analysis, and assigns a data point to that cluster that have the least distance between its centroid and the data point. K-means clustering analysis tries to detect homogeneous clusters within the data, so that the data points in each cluster consist of similarity within clusters and difference between clusters as possible, according to a similarity measure such as a Euclidean-based distance [5].

Firstly, clustering is often used as a market segmentation approach to uncover similarity among customers or uncover an entirely new segment altogether. The *K*-means clustering algorithm is used to find clusters which have not been explicitly labeled in the data. This can be used to confirm business assumptions about what types of groups exist or to identify unknown groups in complex data sets. Once the algorithm has been run and the groups are defined, any new data can be easily assigned to the correct group [2]. Thus, first, a K-means clustering analysis was conducted to find homogeneous clusters within the 530 NBA players using their performance-related statistics in the 2018-2019 season.

Secondly, discriminant analysis is often used in combination with cluster analysis. Discriminant analysis is a statistical technique used to classify the target population into specific categories or clusters based on certain attributes (independent variables) [3]. For any kind of discriminant analysis, some cluster assignments should be known beforehand. Discriminant analysis is also a method of predicting some level of a one-way classification based on known values of the responses. This method is based on how close the measurement variables are to the multivariate means of the levels being predicted. In other words, it is useful in determining whether a set of variables are effective in predicting category membership [11].

The objective of discriminant analysis is to develop discriminant functions that are nothing but the linear combination of independent variables that will discriminate between the categories of the dependent variable in a perfect manner. It examines whether significant differences exist among the groups, in terms of the independent variables. It also evaluates the accuracy of the classification [11]. Therefore, a discriminant analysis was also employed to classify the 530 NBA players into specific clusters based on their performance-related statistics in the 2018-2019 season.

Thirdly, one-way ANOVA is the most commonly used technique for comparing the groups' means of measured data. In statistics, one-way ANOVA is a technique that compares the average of two or more independent groups (using the F distribution) in order to determine whether there is statistical evidence that the associated population means are significantly different. Thus, after the formation of the identified NBA player clusters, a one-way analysis of variance (ANOVA) was used to determine whether there are any statistically significant differences between the means of the identified NBA player clusters.

Finally, a multilayer perceptron (MLP) neural network model was utilized as a predictive model in deciding the classification of NBA player performance values in the 2018-2019 season. *Neural networks* are algorithms used to recognize patterns in a data set, both labeled and unlabeled data. They take input data, process the data through hidden layers, and return an output. Neural networks seek to classify an observation as belonging to some discrete class as a function of the inputs. The input data (independent variables) can be categorical or numeric types, however, we require a categorical feature as the dependent variable [6] [8].

The MLP neural networks comprise of distributed neurons and weighted links. Arranged in a multi-layered structure, each neuron contains a simple processing function (i.e., activation function) that individually handles pieces of complex problems; the weighted links between neurons determine the direction of data flow and the contribution of the "from" neuron to the "to" neuron. These weights can be determined through an iterative back-propagation training process that learns from known samples and adjusts the weights between neurons until the minimum error of the performance function is achieved [1].

The classification and clustering of these data sets are significant. The data set is divided into training set and testing set. With the help of these datasets, the network first goes through the training process in order to produce results that are later used for testing. The training set is taken from two-thirds of the dataset, while the remaining is used for the test set. This is made through the assessment of accuracy achieved through testing against these data sets. The network then is simulated with the same data [1].

4. Results

4.1. K-means Clustering Analysis

The K-means clustering analysis was conducted to identify a solution with the specified number of clusters of 468 NBA players using their performance related statistics in the 2018-2019 season.

Consequently, a three-cluster solution was agreed upon the distance, computed using simple Euclidean distance, from the cluster centers to every object with the shortest distance to the cluster center. The clusters were labeled as *Key Player*, *Bench Player*, and *Supporting Player* clusters (Table 2).

Table 2: Cluster Analysis of NBA Players in the 2018-19 Season

Stats	*Key Player*	*Bench Player*	*Supporting Player*
G	72	22	62
GS	67	3	14
MP	30.2	12.0	19.9
FG	5.7	1.7	3.0
FGA	12.2	4.0	6.7
3P	1.5	0.4	0.9
3PA	4.2	1.5	2.5
2P	4.2	1.3	2.1
2PA	8.0	2.5	4.2
eFG%	0.534	0.469	0.521
FT	2.6	0.7	1.2
FTA	3.4	1.0	1.6
ORB	1.3	0.6	0.8
DRB	4.6	1.8	2.7
TRB	5.9	2.4	3.5
AST	3.4	1.0	1.8
STL	1.0	0.4	0.6
BLK	0.6	0.3	0.4
TOV	1.9	0.6	1.0
PF	2.4	1.3	1.8
PTS	15.6	4.5	8.1
n = 468	136	124	208
Percentage	29.1	26.5	44.4

The *Key Player* cluster, with about 29 percent of all NBA players in the 2018-2019 season, was named because of the highest value of all performance related statistics, G = 72, GS = 67, MP = 30.2, PTS = 15.6. Thus, the NBA players in this *Key Player* cluster demonstrated more active performances when they were playing basketball.

The *Bench Player* cluster was the smallest group, comprising of approximately 26.5 percent of all NBA players in the 2018-2019 season, named because of the lowest value of all performance-related statistics, particularly G = 22, GS = 3, MP = 12.0, PTS = 4.5. Furthermore, the NBA players in this *Bench Player* cluster demonstrated more inactive performances when they were involved in the game.

The *Supporting Player* cluster was the largest group comprising of approximately 44.4 percent of all NBA players in the 2018-2019 season. These NBA players had the value of all performance-related statistics between the *Key Player* cluster and the *Bench Player* cluster, for example, G = 62, GS = 14, MP = 19.9, PTS = 8.1. Furthermore, the NBA players in this *Supporting Player* cluster demonstrated more preferences for supporting the members of the *Key Player* cluster when they were playing the game.

4.2 Discriminant Analysis

Results of the K-means cluster analysis were tested for accuracy using the discriminant analysis, which is used primarily to predict membership in two or more mutually exclusive groups. In this case, the Wilk's Lambda scores were 0.051 ($\chi^2 = 1362.737$, $df = 34$, $p < 0.001$) and 0.417 ($\chi^2 = 399.477$, $df = 16$, $p < 0.001$) for both discriminant functions, respectively, indicating that group means were significantly different. The canonical correlation results were both above 0.7, supporting that there were strong relationships between the discriminant score and the cluster membership (Table 3).

Table 3: Canonical Correlation of Discriminant Functions

Function	Eigenvalue	% of Variance	Canonical Correlation
1	7.228*	83.8	0.937
2	1.397*	16.2	0.763

* First two canonical discriminant functions were used in the analysis.

Two discriminant functions were formulated shown in Table 4. The first function is for discriminating between the combined *Key Player*, *Bench Player* and *Supporting Player* clusters, and the second one for discriminating between *Bench Player* and *Supporting Player* clusters, respectively. The first function is the most powerful differentiating dimension, but the second function may also represent additional significant dimensions of differentiation. Though mathematically different, each discriminant function is a dimension which differentiates a case into categories of the dependent variables (three identified NBA player clusters). Those clusters are based on its values on the independent variables of the twenty-one performance-related statistics. Furthermore, the territorial map is a tool for assessing discriminant analysis results by plotting the group membership of each case on a graph (Figure 1).

Table 4: Standardized Canonical Discriminant Function Coefficient

Stats	Function 1	Function 2
G	0.457	0.799
GS	0.837	-0.704
MP	0.023	0.537
FG	0.231	-0.614
FGA	-0.130	0.552
3P	-0.048	0.147
3PA	-0.006	-0.269
eFG%	-0.035	0.086
FT	0.022	-0.154
FTA	0.003	-0.061
ORB	-0.152	0.138
DRB	0.036	-0.015
AST	-0.143	0.083
STL	0.048	-0.053
BLK	-0.041	0.071
TOV	0.064	0.053
PF	-0.022	-0.028

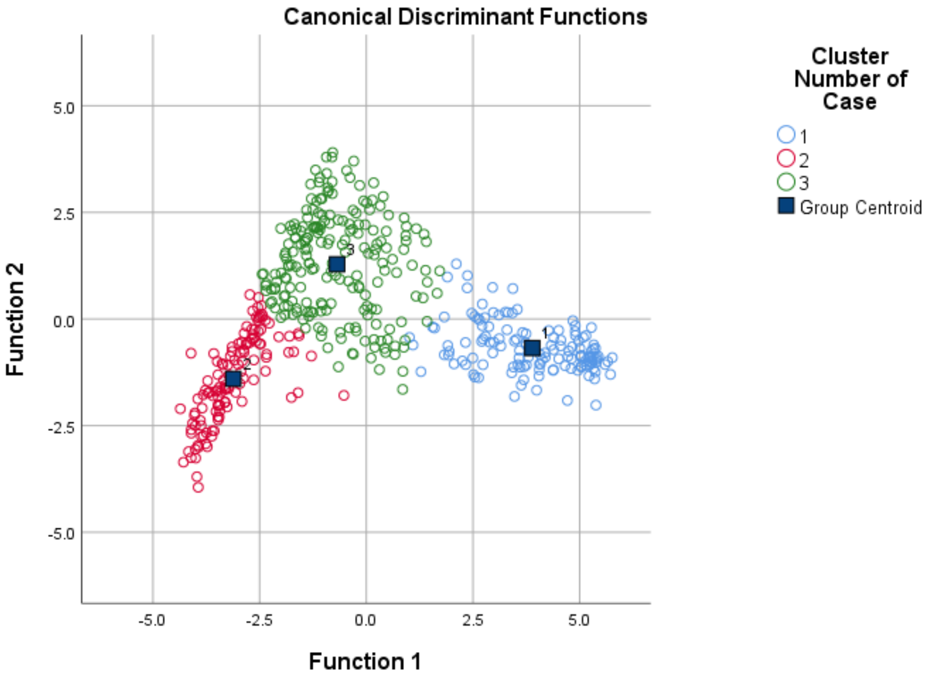

Figure 1: Territorial Map (1 = *Key Player* cluster; 2 = *Bench Player* cluster; 3 = *Supporting Player* cluster)

The classification results based on discriminant analysis (Table 5), 136 cases fell into the *Key Player* cluster, 124 fell into the *Bench Player* cluster, and 208 fell into the *Supporting Player* cluster in the original row total, which is the groups' frequencies found in the data. Across each row, how many of the cases in the group can be classified by this analysis into each of the different groups. For example, of the 136 cases that were in the *Key Player* cluster, 134 were predicted correctly and two were predicted incorrectly (two were predicted to be in the *Supporting Player* cluster). Predicted group membership indicates the predicted frequencies of groups from the analysis. The numbers going down each column indicate how many were correctly and incorrectly classified. For example, of the 135 cases that were predicted to be in the *Key Player* cluster, 134 were correctly predicted, and one were incorrectly predicted (one case was in the *Supporting Player* cluster). It explained that 99.1% of original grouped cases correctly classified (Table 5).

Table 5: Classification Results[a] Based on Discriminant Analysis in the 2018-19 Season

		Cluster Number of Case	Predicted Group Membership			Total
			Key Player	*Bench Player*	*Supporting Player*	
Original	Count	*Key Player*	134	0	2	136
		Bench Player	0	123	1	124
		Supporting Player	1	0	207	208
	%	*Key Player*	98.5	0.0	1.5	100
		Bench Player	0.0	99.2	0.8	100
		Supporting Player	0.5	0.0	99.5	100

a. 99.1% of original grouped cases correctly classified

4.3 One-Way ANOVA

The results of one-way ANOVA showed that significant differences in all twenty-one performance related statistics and player's salary were found within the three identified NBA player clusters statistically (Table 6).

Table 6: Cluster Means of the NBA Player Clusters in the 2018-19 Season

Stats	*Key Player*		*Bench Player*		*Supporting Player*	
	Mean	Std. Dev.	Mean	Std. Dev.	Mean	Std. Dev.
G	72.22	8.65	22.42	12.80	61.85	12.68
GS	66.60	12.20	2.52	5.23	13.96	12.37
MP	30.17	3.70	12.03	5.81	19.87	5.45
FG	5.73	2.03	1.70	1.14	3.02	1.34
FGA	12.20	4.21	4.00	2.38	6.70	2.96
3P	1.51	0.94	0.44	0.44	0.89	0.63
3PA	4.17	2.42	1.46	1.23	2.54	1.69
2P	4.22	1.93	1.26	1.01	2.13	1.24
2PA	8.04	3.44	2.54	1.76	4.16	2.36
eFG%	0.53	0.04	0.47	0.12	0.52	0.05
FT	2.64	1.65	0.70	0.76	1.17	0.78
FTA	3.37	2.02	0.99	0.99	1.57	1.00
ORB	1.29	1.00	0.58	0.57	0.84	0.67
DRB	4.58	2.04	1.79	1.44	2.71	1.18
TRB	5.87	2.84	2.38	1.88	3.54	1.72
AST	3.42	2.12	0.99	0.86	1.84	1.41
STL	1.00	0.40	0.37	0.25	0.60	0.31
BLK	0.65	0.53	0.26	0.28	0.39	0.32
TOV	1.87	0.85	0.63	0.41	1.00	0.54
PF	2.40	0.52	1.30	0.74	1.81	0.59
PTS	15.62	5.85	4.54	3.18	8.09	3.60
Salary	12783779.76	10036423.69	2858263.53	4419920.55	5529678.12	5498780.45

According to the post-hoc comparisons with the LSD test, significant clustering pairwise differences were obtained in all twenty-one performance related statistics and player's salary between the *Key Player* cluster and both *Bench Player* and *Supporting Player* clusters, except the eFG% between *Key Player* and *Supporting Player* clusters (mean difference = 0.012, p = 0.142).

4.4 Multilayer Perceptron Neural Network

After the formation of three identified NBA player clusters, a multilayer perceptron (MLP) neural network model was employed as a predictive model in deciding the classification of NBA players associated with their performance related statistics. The Multilayer Perceptron Module of IBM SPSS Statistics 26 was used to build the neural network model and test its accuracy. The MLP neural network model, trained with a back-propagation learning algorithm which uses the gradient descent to update the weights towards minimizing the error function.

The aim of this analysis was to examine whether a MLP neural network model can help NBA managers to correctly classify NBA players from their performance values, by analyzing data obtained from NBA player performance. The data were randomly assigned to training (n_1 = 320, 68.4%) and testing (n_2 = 148, 31.6%) subsets. The training dataset was used to find the weights and build the model, while the testing data was used to find errors and prevent overtraining during the training mode.

In order to find the best neural network, disparate possible networks were tested and it was concluded that neural network with a single input layer, a single hidden layer, and a single output layer was the best option for this study. Previous studies have found that using neural network with a single input layer, a single hidden layer, and a single output layer is advantageous. Sheela and Deepa [10] pointed out that as the number of neurons or the number of layers of a neural network increase, the training error also increases due to overfitting. It is clear that using a single input layer, a single hidden layer, and a single output layer in the neural network will help decrease the probability of overfitting and will require relatively lower computational time.

One of the most salient considerations in the construction of neural network is choosing activation function for hidden and output layers that are differentiable. The results showed that in this study, a hyperbolic tangent activation function would be used for the single hidden layer of the model and linear activation function would be used for the output layer. The Multilayer Perceptron Module of IBM SPSS Statistics 26 was used as the tool to choose the best architecture model automatically, and it built the network with one hidden layer.

From the twenty-one independent variables, the automatic architecture selection chose seven nodes for the hidden layer, while the output layer had three nodes to code the depended variable, *Cluster*. For the input layer, standardized option, subtract the mean and divide by the standard deviation, was used for rescaling input covariates. For the hidden layer the activation function was the hyperbolic tangent, while the output layer used a softmax function. Cross entropy was used as an error function because of the use of softmax function. Intuitively, the cross-entropy loss function is used to measure the error at a softmax layer, typically the final output layer in a neural network. In the architectural point-of-view, it was a 21-7-3 neural network, means that there were total twenty-one independent (input) variables, seven neurons in the hidden layer and three dependent (output) variables.

The model summary provided information related to the results of training and testing sample (Table 7). Cross entropy error is displayed because the analysis is based on the softmax activation function, and is given for both training and testing sample since the error function is given that neural network minimizes during training phase. The value of cross entropy error (= 1.361) indicated the power of the model to predict the three identified NBA player clusters. The cross entropy error was less for the testing sample compared with the training data set, meaning that the neural network model had not been over-fitted to the training data, and learned to generalize from the trend. The result justified the role of testing sample which was to prevent overtraining.

In this study, the percentage of incorrect prediction was equal to 0.0% in the training sample. As a result, the percentage of correct prediction was 100% which is an excellent prediction in a qualitative study for determining management results of NBA players' performance. The learning procedure was performed until one consecutive step, with no decrease in error function, was attained from the training sample.

Table 7: Model Summary

Training	Cross Entropy Error	1.361
	Percent Incorrect Predictions	0.0%
	Stopping Rule Used	1 consecutive step(s) with no decrease in error^a
	Training Time	0:00:00.06
Testing	Cross Entropy Error	5.852
	Percent Incorrect Predictions	2.7%

Dependent Variable: Cluster

a. Error computations are based on the testing sample.

Using the training sample only, MLP neural network utilized synaptic weights to display the parameter estimates that showed the relationship between units in a given layer to the units in the following layer (Table 8). Note that the number of synaptic weights can become rather large, and these weights are generally not used for interpreting neural network results [7].

Table 8: Parameter Estimates

Predictor		Predicted									
		Hidden Layer 1							Output Layer		
		H(1:1)	H(1:2)	H(1:3)	H(1:4)	H(1:5)	H(1:6)	H(1:7)	Key Player	Bench Player	Supporting Player
Input Layer	(Bias)	1.489	1.686	-1.170	0.852	-0.082	-1.633	0.695			
	G	1.915	-0.698	0.460	2.137	-0.057	-2.423	0.706			
	GS	0.790	-2.857	1.840	-0.394	-0.946	-0.975	-0.252			
	MP	0.453	-0.349	0.424	0.630	-0.319	-0.439	0.538			
	FG	-0.206	0.031	0-.407	-0.319	-0.093	-0.053	0.190			
	FGA	0.227	-0.031	0.425	0.062	-0.114	0.068	0.403			
	3P	-0.033	0.008	0.268	0.352	0.137	-0.173	0.187			
	3PA	0.192	-0.101	-0.181	-0.236	-0.275	0.155	0.086			
	2P	0.008	0.014	-0.183	0.224	-0.425	0.226	-0.249			
	2PA	0.000	-0.166	0.242	0.119	-0.221	-0.202	0.029			
	eFG%	0.020	0.255	0.186	0.075	0.439	-0.244	0.271			
	FT	-0.074	0.401	-0.133	-0.451	-0.296	0.254	-0.304			
	FTA	-0.384	0.059	0.077	0.029	-0.101	-0.229	0.207			
	ORB	0-.061	0.051	-0.123	-0.064	-0.426	-0.240	-0.102			
	DRB	-0.137	0.003	-0.143	0.331	-0.344	0.148	0.438			
	TRB	0.259	-0.101	0.311	-0.228	0.106	0.225	-0.234			
	AST	0.269	0.133	-0.159	-0.254	-0.065	-0.126	0.494			
	STL	-0.179	-0.045	0.009	0.198	0.243	-0.406	0.202			
	BLK	0.266	0.150	0.036	0.025	0.374	0.446	0.333			
	TOV	-0.356	0.012	0.051	0.150	-0.147	0.200	-0.395			
	PF	-0.304	-0.104	0.055	0.026	0.109	-0.022	0.054			
	PTS	0.307	-0.218	-0.235	-0.007	-0.174	0.455	0.221			
Hidden Layer 1	(Bias)								-0.012	-0.467	0.816
	H(1:1)								0.897	-2.530	1.076
	H(1:2)								-3.068	0.963	2.377
	H(1:3)								1.874	-0.396	-1.329
	H(1:4)								0.653	-1.778	1.214
	H(1:5)								-1.010	0.666	0.627
	H(1:6)								-1.187	3.071	-1.372
	H(1:7)								0.062	-0.908	0.403

Table 9: Predictive Ability and Classification Results

Classification					
Sample	Observed	Predicted			
		Key Player	*Bench Player*	*Supporting Player*	Percent Correct
Training	*Key Player*	100	0	0	100.0%
	Bench Player	0	82	0	100.0%
	Supporting Player	0	0	138	100.0%
	Overall Percent	31.3%	25.6%	43.1%	100.0%
Testing	*Key Player*	34	0	2	94.4%
	Bench Player	0	41	1	97.6%
	Supporting Player	0	1	69	98.6%
	Overall Percent	23.0%	28.4%	48.6%	97.3%

Dependent Variable: Cluster

Based on the MLP neural network, a predictive model developed and displayed a classification table (i.e. confusion matrix) for categorical dependent variable – the three identified NBA players' clusters – by partition and overall (Table 9). Shown in the table below, the MLP neural network correctly classified 320 NBA players out of 320 in the training sample and 144 out of 148 in the testing sample. Overall, 100% of the training cases were correctly classified. The predictive model developed had excellent classification accuracy.

Using the training sample only, it was able to classify 100 NBA players as the *Key Player* into the *Key Player* cluster, out of 100. It held 100% classification accuracy for the *Key Player* cluster. Similarly, the same model was able to classify 82 NBA players as the *Bench Player* into the *Bench Player* cluster out of 82, and 138 NBA players as the *Supporting Player* into the *Supporting Player* cluster out of 138. It was able to generate 100% classification accuracy for both the *Bench Player* and the *Supporting Player* clusters (Table 9).

Independent variable importance analysis provides the sensitivity analysis, by computing the importance of each independent variable which in turn determines the structure of the neural network. The analysis has been based on the combined training and testing samples. Normalized importance is the importance value divided by the largest importance value and is expressed as a percentage [7]. The importance of independent variables (factors influencing NBA player performance) is a measure of how much the neural network model predicted value changes for different independent variables. The input parameters – NBA player performance related statistics, which influenced the three identified NBA players' clusters – have been ranked by the neural network model were given in the following Table 10. The first three significant dominant factors that have been found were "G" (100%), contributed the most in the neural network model construction, followed by "GS" (79.9%), and "MP" (37.5%), had the greatest effect on how NBA player performance. The next two important factors were "FT" (30.4%) and "eFG%" (29.3%). The other factors were relatively not as important, such as "3PA" (10.4%), "FTA" (10.2%), "DRB" (10.0%), "2P" (7.4%), and the least important factor which has been identified was "PF" (6.7%).

Table 10: Independent Variable Importance Analysis

Stats	Importance	Normalized Importance	Rank
G	0.210	100.0%	1
GS	0.168	79.9%	2
MP	0.079	37.5%	3
FG	0.022	10.5%	16
FGA	0.033	15.9%	9
3P	0.029	13.7%	11
3PA	0.022	10.4%	17
2P	0.016	7.4%	20
2PA	0.028	13.3%	13
eFG%	0.062	29.3%	5
FT	0.064	30.4%	4
FTA	0.022	10.2%	18
ORB	0.025	11.7%	14
DRB	0.021	10.0%	19
TRB	0.033	15.9%	8
AST	0.030	14.1%	10
STL	0.024	11.4%	15
BLK	0.036	17.0%	7
TOV	0.036	17.3%	6
PF	0.014	6.7%	21
PTS	0.028	13.5%	12

Independent variable importance chart showed the impact of each independent variable in the MLP neural network model in terms of relative and normalized importance [7]. Independent variable importance chart also depicted the importance of the independent variables, i.e. how sensitive is the model is the change of each input variable (Figure 2).

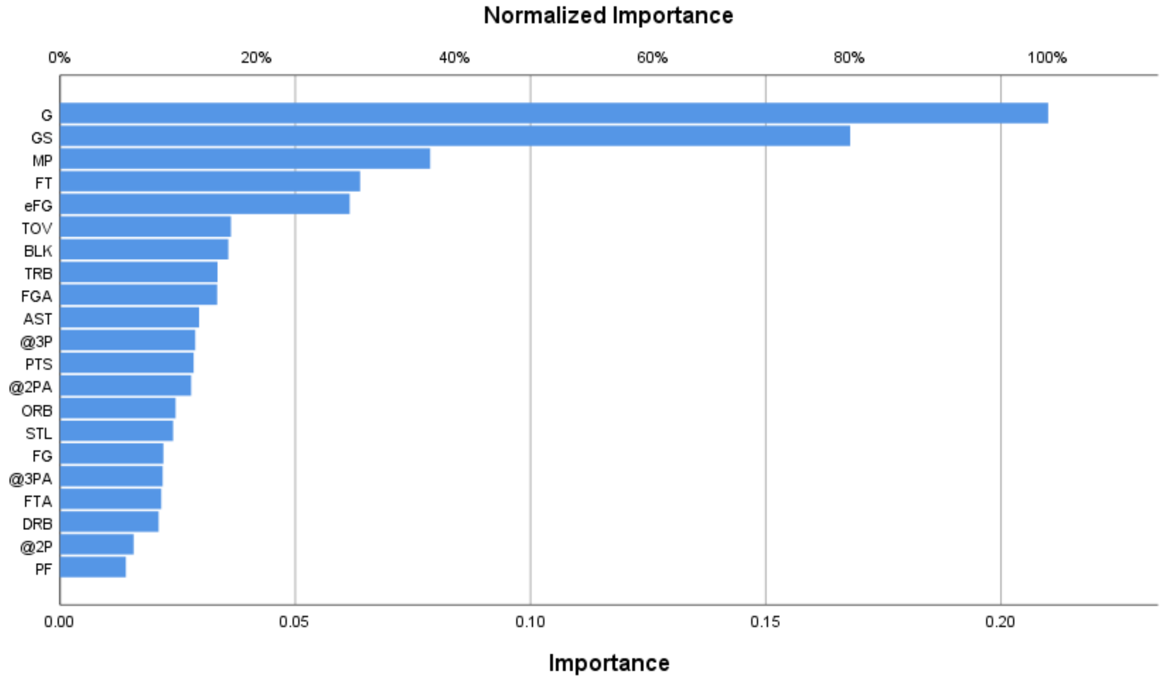

Figure 2: Independent Variable Importance Chart

5. Discussion and Conclusions

Overall, this study adopted K-means clustering analysis to identify three NBA player clusters, named *Key Player* cluster (29.1% of 468 NBA players in the 2018-19 season), *Bench Player* cluster (26.5%), and *Supporting Player* cluster (44.4%). This study also showed that almost all of the notable players like Stephen Curry, LeBron James, Chris Paul, Russell Westbrook, Blake Griffin, Kyle Lowry, Paul George, Mike Conley, James Harden, Kevin Duran, Anthony Davis, etc. were classified together as *Key Players*. These famous players have been among the top performers of the league for a long time and produced best results. The centers of all the NBA player performance-related statistics were way larger for *Key Players* than the other two clusters.

The classification results based on discriminant analysis showed that 99.1% of original grouped cases are correctly classified. After the formation of the three identified clusters, a MLP neural network model was employed as a predictive model in deciding the classification of NBA players associated with their performance related statistics. As a result, 100% of the training cases were correctly classified, revealing that the predictive model developed had excellent classification accuracy. This study was intended to provide a snapshot of today's NBA players. It reveals a general clustering effect that deep learning algorithms are able to create to specifically fit the game of basketball.

With some of the top talented players in the NBA, every team is constantly searching for an edge, and with the success of sports analytics, NBA teams are looking to advanced technologies like machine learning to gain a competitive advantage. There are many opportunities to assess player performance. At the most basic level, basketball is about scoring more points than the opponent, so naturally points-per-game is a potential starting point to look into. From optimizing a player's workload to determining what drives performance on an individual basis, methods of building highly accurate models for predicting player performance can be the feature that coaches and executives are searching for.

Theoretically, a cluster is a collection of items that are similar among themselves and are dissimilar to the items belonging to other clusters. It can be shown that there is no absolute best criterion, which would be independent of the final aim of the clustering. Hence, the structure of the clusters should be finalized by the user depending on the physical requirements. Thus, there is no unique approach to correctly classify NBA player clusters.

Recently, LeBron James has been voted the athlete of the decade by USA TODAY Sports. Zillgitt [12] also pointed out that LeBron James could change the NBA in the 2010s, such as (1) finals and titles, (2) big three/player empowerment movement, (3) mastery of media, (4) philanthropy, and (5) social and political issues. However, measuring greatness in basketball is extremely difficult, depending on defining what the term means, which is the most important part of "the greatest player ever" debate.

Trends in the NBA are constantly changing and there will always be new players that completely defy prior expectations of positional roles. Further empirical research into play-by-play data and an advanced analysis of every player in NBA may enrich and improve current results. Furthermore, a more in-depth exploration into each cluster may reveal insight into how teams can scout and develop the next talent players.

This study could be a useful application in determining and identifying the NBA players in the same level of performance. This will help the team managers to identify players when planning a transfer or player exchange deal with a franchise. Team managers used to wish a player of at least the same skill and performance level as the one they are trading off. Furthermore, the coaches can use these results to identify the weaker players in their opponent's line-up. It will help them plan their in-game strategies if they can classify the performance levels of the players in their opponent team. This can also help them to estimate their winning odds. Also, they want their key players to spend maximum time on the court and use strugglers as fillers.

Team managers can also use these results to do several types of pre-match and post-match analysis. Analysts can use the application of similar kind of analysis on the data of the previous seasons to conduct a time series analysis of each player's performance. This can take the analysis a step deeper to understand the performance levels of a player across the NBA seasons.

References

[1] M. Buscema. Back propagation neural networks, Substance Use & Misuse, Vol. 33, No. 2, pp. 233-270, 1998.

[2] G. A. Churchill, Jr., D. Iacobucci, Marketing Research: Methodological Foundations, 9th ed., Mason, OH: Thomson/South-Western, 2005.

[3] R. A. Fisher. The use of multiple measurements in taxonomic problems, Annals of Eugenics, Vol. 7, pp. 179-188, 1936.

[4] Forbes Press, Forbes releases 20th annual NBA team valuation, 2018. Retrieved from https://www.forbes.com/sites/forbespr/2018/02/07/forbes-releases-20th-annual-nba-team-valuations/#6959bb1734e6

[5] E. W. Forgy. Cluster analysis of multivariate data: efficiency versus interpretability of classifications, Biometrics, Vol. 21, pp. 768–769, 1965.

[6] M. W. Gardner, S. R. Dorling. Artificial neural networks (the multilayer perceptron) - a review of applications in the atmospheric sciences, Atmospheric Environment, Vol. 32, No. 14, pp. 2627-2636, 1998.

[7] IBM, IBM SPSS neural networks 26, Armonk, NY: IBM Corporation, 2019.

[8] S. S. Haykin, Neural Networks and Learning Machines, 3rd ed., Upper Saddle River, New Jersey: Pearson Education, Inc., 2009.

[9] J. MacQueen. Some methods for classification and analysis of multivariate observations, Proceedings of the 5th Berkeley Symposium on Mathematical Statistics and Probability, Vol. 1, pp. 281-297, 1967.

[10] K. G. Sheela, S. N. Deepa. Review on methods to fix number of hidden neurons in neural networks, Mathematical Problems in Engineering, Volume 2013, Article ID 425740, 11p, 2013.

[11] B. G. Tabatchnick, L. S. Fidell, Using Multivariate Statistics, 6th ed., Boston: Pearson Education, Inc., 2013.

[12] J. Zillgitt. How LeBron James changed NBA this decade, 2019. Retrieved from USA Today, https://www.usatoday.com/story/sports/nba/2019/12/19/lebron-james-changed-nba-this-decade/2665096001/