Clustering and Prediction of NBA Player Performance: A Machine Learning Approach

Chris Li, Luyang Zhang, Jonas Kempf
Department of Statistics
University of California, Davis
Davis, CA 95616
{hlcli,luyzh,jkempf}@ucdavis.edu

Abstract

The ability to accurately predict future NBA player performance is a task that has challenged teams and basketball analysts for decades and has important implications for a team's success and financial health. Our analysis utilized clustering and machine learning methods to better understand how best to predict NBA players' future success based on their early career statistics. First, a Gaussian mixture model is used to identify clusters and assign labels to both guards and forwards to differentiate top, middle, and bottom tier players. Players are clustered based on their statistics starting from their fourth season until the end of their career. We then used neural networks and random forest classifiers to predict a player's tier based on their statistics from their first three seasons. The random forest classifier outperformed the neural network model for this task, which is driven by its stronger ability to predict which players will be elite or lower tier. Furthermore, our analysis confirmed the importance of common predictors of NBA player performance such as points and minutes per game, as well as other predictors that provide interesting insight into different ways to evaluate young guards and forwards.

1 Introduction

NBA teams spend vast amounts of time and resources on player evaluation, especially evaluating the promise of young players. This paper aims to explore machine learning methods that can help predict a player's career trajectory based on their early career performance.

Every year, teams acquire young players from the draft and also sign players straight out of college. However, the reality is that there are only a select few spots on an NBA roster, and out of those on the roster, only 10 or fewer players get consistent playing time throughout the season. Some young players are never able to make the starting lineup, and therefore they either get stuck playing in the G-League (the league below the NBA), find opportunities in less competitive leagues overseas, or are forced to end their career prematurely.

On the other hand, the best young players are granted opportunities to play in the NBA, even though their playing time can often be limited in their first few years. Typically, players do not reach their peak performance until later into their career, as they gain physical maturity and valuable experience as they get older. However, early on in these players' careers, teams must determine if they wish to keep them on the roster longer term (which often means giving them a pay raise), or if they want to trade or release them. Therefore, teams are challenged with forecasting a player's future success using limited historical data and information. The long-term success of an NBA team hinges on their ability to identify and retain young players with promising futures.

2 Background

2.1 Literature Review

Given the importance of this task, several papers have explored how to evaluate the potential of NBA players by applying machine learning techniques to NBA statistics. Nguyen et al. (2022) aimed to forecast a player's performance in the next season based on the player's statistics in the current season (1). They measure player performance based on whether the player made the All Star team that season. They used a variety of methods such as logistic regression, random forests, and neural networks, and find that a balanced, under-sampling random forest performed best. Neural networks performed well but not the best, and their best performing neural network consisted of two layers. Soliman et al. (2017) also attempted to predict selections to the All Star Game using various player statistics, and find success using a random forest model (4). Chou et al. (2021) employed neural networks to tackle a slightly different task: predicting NBA players that will enter the Hall of Fame, a prestigious award that is even harder to achieve than qualifying for an All Star Game (5).

Meanwhile, other papers have attempted to predict players' NBA performance by grouping players into different clusters based on their performance. Moxley and Towne (2015) use players' college statistics and draft information to forecast the players' outcomes in their first three seasons in the NBA (3). First they used a growth mixture model to cluster young players based on their win shares in their first three years. They found that a model that categorized players into three classes fit the data best. After classification, they predicted membership in these classes and found that success in college is the strongest predictor of success in the NBA.

This analysis drew from ideas of several of these studies. Similar to Moxley and Towne (2015), this analysis classified players' ability by clustering them into three different classes, and used historical data to predict their membership in these classes. To make predictions, this analysis compared the performance of both neural networks and random forest classifiers, since previous research supported the use of both of these machine learning techniques for prediction of NBA player success (1; 4; 5).

2.2 Problem Definition

This paper differentiates itself from prior studies in several ways. Firstly, instead of using win shares to evaluate player performance and cluster players, this paper used points, assists, rebounds, and minutes (all standardized to be per game measures) to classify players. The use of the first three metrics would seem straightforward given their widespread use by teams and media to evaluate a player. Furthermore, minutes per game is also used, since it is well known that more valuable players generally play more minutes. Specifically, a Gaussian mixture model is used to cluster players based on these four statistics.

Another way this analysis differs from prior literature is the period over which a player's performance is evaluated, and the period of data used as predictors. One earlier study (1) predicted player success in each season given information from the previous season, while another (3) predicted early career success based on pre-NBA information. In contrast, this analysis aimed to predict players' later career success as measured by their average statistics (weighted by games played) starting from their fourth NBA season until their last NBA season. These aggregate statistics are fed into the Gaussian mixture model. Following this clustering, we used annual-level statistics from a player's first three seasons in the NBA to predict their later career success.

In addition, this analysis separated guards (point guard and shooting guard) from forwards (small forward, power forward, and center) and ran separate clustering and prediction models for these two groups. This is because the expected number of assists and rebounds guards collect compared to the number that forwards collect are quite different due to the different roles of these positions. Forwards will be closer to the basket and will collect more rebounds, while guards will dribble the ball more and have more opportunities to provide assists. Therefore, the assists per game and rebounds per game expected out of an elite guard will be different from those expected out of an elite forward. This is shown by our clustering results later in the paper. Consequently, estimating clusters separately for these two groups will make the clusters more plausible and easier to interpret. In this way, this paper is able to contribute new insights into predicting NBA player success.

2.3 Data Description

The data used is from basketballreference.com, and consists of season-level NBA statistics for players from the 1946-1947 season until the 2004-2005 season contained in file 'player_regular_season.txt'. These statistics include games played, minutes, points, rebounds, assists, steals, blocks, turnovers, fouls, field goals, free throws, and three pointers. Furthermore, this analysis also joined this dataset with player-level information (contained in 'players.txt') in order to determine the position of each player. This is needed to conduct separate analyses of guards and forwards.

This analysis only used data starting in 1979 since there are issues with data quality and incompleteness prior to 1979. These pre-1979 data issues are also documented in other papers (1). Additional filters were applied to determine which players would be included in the analysis. Players were included if they played at least 20 games in their first three seasons and if they played at least one game after their first three seasons. This first filter was put in place to ensure each player had sufficient game statistics to use as predictors of future success. If a player did not play very many games in his first three seasons (i.e. less than 20), then trying to use this limited sample to predict future success may be less effective and informative. So our model will only be externally valid for players who played a reasonable number of games in their first three seasons.

We did not impose a stringent filter for the latter, so our analysis included players that played very few games after their first three seasons. While it may be difficult to cluster players into categories based on such a small sample size of games, chances are players that do not play many games are in that situation because they don't perform well statistically, and those players will most often be clustered into the lowest tier. We also wanted our model to be able to predict which players will perform poorly and have an early end to their career. Removing these players when fitting the model would create bias in our model that would tend to overestimate the success of players. Therefore, we kept these players in.

Furthermore, other processing of the data was conducted before use in analysis. First, variables regarding shooting (field goals, free throws, and three pointers) were converted to percentages for easier interpretation. Other variables were transformed to be per game measures rather than measures over the total season. This helped accurately represent how well players played regardless of how many games they played in a season, which could be affected by injuries.

Next, we separated the dataset into two. One dataset contained players' annual per game statistics after their first three seasons. These statistics were averaged (weighted by games played) over all of these later seasons and then used for clustering. To make clustering results more interpretable, we did not standardize these data. The other dataset contained each players' first three seasons of data, and was standardized to get all variables at the same scale to help with prediction. Also, it was kept at the annual level so that the prediction models could leverage any season-to-season trends.

In both of these datasets, players were removed if they had missing values for any variables except for three point data. Several players either had missing three point data or no three point attempts in a given season, so three point percentage could not be calculated. Dropping such players from analysis would substantially reduce our sample size, so instead, we excluded three point data from our prediction model. Lastly both datasets were further split into two, one for guards and another for forwards.

3 Proposed Methods - Intuition and Algorithm Descriptions

3.1 Clustering Model

For clustering both guards and forwards, we used a Gaussian mixture model with three clusters, in which each cluster has its own covariance matrix. We opted to fit the model using the EM algorithm. One of the distinguishing characteristics of this method for clustering over alternatives is the soft assignment of each observation to each model during the expectation step. This is in contrast to other clustering methods which typically assign observations a single cluster label, rather than the probability it came from a given cluster. The probabilities that each observation came from each cluster are used to estimate the maximum likelihood updates of parameter estimates (6). Sports data provides a perfect application for such soft assignment, since the performance "tier" of a player, while useful, is ultimately arbitrary and fluid over time. This is in contrast to a clustering problem

concerning, for instance, the species of a flower, which is not mutable. We clustered on four measures of a player's performance: points, assists, rebounds, and minutes, which are all scaled to be per game metrics. The choice of these variables was discussed earlier in this paper.

Furthermore, some popular defensive statistics such as blocks and steals are omitted from the clustering criteria. This is for two reasons. One is that steals and blocks are rarer in games and only a select few players have high numbers of steals or blocks, so they are less telling in differentiating talent across all NBA players. In addition, those that do have higher numbers in these categories are often defensive specialists who may be less strong in offensive categories, which would add extra complexity to defining the player tiers from our Gaussian mixture model. So, for simplicity, these statistics are left out.

We also allow for non-zero covariances between variables, since it is possible that the relationships between these four variables may differ for players in different clusters. Computing BIC shows that four is the optimal number of clusters for guards (three is a close second) and three is the optimal number of clusters for forwards. These results are supported by performing cross-validation using likelihood as a metric. We stick with three clusters for both guards and forwards since that is consistent with the number of clusters used in previous mixture models (3). Furthermore, our results with three clusters make intuitive sense, as discussed later.

These fitted Gaussian mixture models produce predicted probabilities that a player falls in each cluster. To simplify our downstream analysis, we assigned each player to a cluster corresponding to the highest probability. This cluster assignment served as the measure of a player's future success, which will be our label for the classification models.

3.2 Prediction Model

Using information on each player's statistics in their first three seasons, we predicted the future performance of that player (as defined by which of the three clusters they fell into). We used all statistics available in the data as predictors with the exception of three-point data due to missing values, as discussed earlier. This includes minutes, points, rebounds, assists, steals, blocks, turnovers, and fouls (all as per game metrics), games played, field goals, and free throws.

For both guards and forwards separately, we ran two machine learning methods to predict future career success using early career statistics: (1) a two-layer neural network and (2) a random forest classifier. We chose both of these methods in order to capture some of the potential non-linear relationships between a player's early career statistics and their "tier" as a player in their mid to late career.

3.2.1 Neural Network

Both of these methods have been widely used in sports analytics to answer questions ranging from season outcomes to starting lineups (1; 7; 8; 9; 10). For neural networks, most papers have found greatest success using a two-layer network, so we also adopted this structure. Determining the optimal number of nodes in our hidden layer was more complicated, with some previous studies reporting variously 2, 3, or 10 nodes as optimal (10; 7; 8). With no clear consensus from the literature, we opted to select the number of nodes through cross validation, using predictive accuracy as our basis for selection. We fit networks with 3 (we did not want to have fewer intermediary nodes than in our output layer) through 8 nodes (to keep complexity and subsequent risk of overfitting low). Our results indicated the model with 5 nodes offered the best performance in light of these factors.

We also chose to use a ReLU activation function for our hidden layer, following standard practice in industry and literature over the last few years (12). A dropout layer with dropout proportion of 0.3 was used to further reduce overfitting. Lastly, for the output layer, we used a softmax activation function in order to output predicted probabilities for each class. Then each player's predicted cluster was determined by the cluster with the highest probability predicted by the neural network.

3.2.2 Random Forest

Previous papers have also found success running random forest classifiers on NBA player data (1; 4; 11). We performed grid search and cross-validation to tune the hyperparameters in our random forest model. Random forest is a more advanced tree-based method. Instead of training a single

classification tree, a random forest contains multiple trees. Specifically, it is an improvement from bagging in trees. In bagging, we fit trees with bootstrapped data and predicted labels are taken by averaging the results for all the trees. This method reduces variance and increases prediction accuracy from single-tree methods. Random forest further improves from bagging by only considering a random subset of predictors for each split. For our model, we choose the number of predictors considered to be the square root of the number of total features. We then used grid search to find the best combination of hyperparameters for both guards and forwards. We searched for the following hyperparameters: the number of trees in the forest and the minimum number of samples required to split an internal node. Mark R. Segal (2004) discussed the issue of overfitting in real-world data when the number of trees is too large (14). Furthermore, the larger and deeper a tree is, the more potential there is for overfitting. Therefore, tuning the number of trees in the forest and the minimum number of samples required to split an internal node helped to reduce overfitting.

After an exhaustive test of combinations of these hyperparameters through grid search cross validation on the training data, we found that 30 trees and minimum of 12 samples to split a node was optimal for our guards model and 50 trees and minimum of 4 samples to split was the best for forwards.

4 Data Analysis

4.1 Questions of Interest

Through our analysis, we aim to answer three main questions:

1. Do the results of our Gaussian mixture model make sense given common knowledge about tiers of NBA guards and forwards?
2. Do neural networks or random forest classifiers work better for predicting future success of NBA players?
3. What are the most important features to consider when predicting a player's future performance?

4.2 Results

Question 1: GMM Clustering

First, we fit the Gaussian mixture model on the full guard and forward datasets. The estimated cluster means from the guard and forward Gaussian mixture models are shown in the tables below.

Table 1: Cluster Means: Guards

Cluster Number	Player Tier	Points	Assists	Rebounds	Minutes
0	Middle	9.3	3.1	2.4	23.3
1	Top	16.8	5.2	4.0	34.2
2	Bottom	4.7	1.9	1.4	14.0

Table 2: Cluster Means: Forwards

Cluster Number	Player Tier	Points	Assists	Rebounds	Minutes
0	Bottom	3.5	0.4	2.6	11.4
1	Middle	7.8	1.2	4.6	20.9
2	Top	15.2	2.5	6.9	31.2

For guards, cluster 1 represents elite players, since this cluster has the highest means across all categories (points, assists, rebounds, and minutes per game). The numerical values of these means are consistent with conventional knowledge about star players. We would expect elite players to score around 17 points per game and play about 34 minutes per game, which is well over half a game. On the other hand, cluster 2 represents bench players or players that have trouble staying on NBA rosters, as evidenced by the very low cluster means across all categories. Lastly, cluster 0 represents guards

that have moderately successful NBA careers and who may not be star players, but still contribute solid statistics and play about 23 minutes per game on average (half the game).

We saw similar results from the Gaussian mixture model fit on forwards, in terms of the clear interpretation and delineation of the three clusters. Cluster 0 represents the worst forwards, cluster 1 represents the "middle-of-the-pack", and cluster 2 consists of the most talented players. Therefore, our clusters for both guards and forwards are consistent with common practice by media and basketball analysts to label players as either elite, role players, or bench players.

Note that for each tier of player (top, middle, or bottom), we observed different cluster means for guards versus forwards. Across all tiers, guards are expected to have more assists, while forwards are expected to have more rebounds. This makes sense based on the positional responsibilities of guards versus forwards, and demonstrates the importance of running separate clustering and prediction models for guards and forwards.

To fit and test our prediction models, we randomly split the guards so that 70% were in our training set and 30% were in our test set. The same was done for forwards. Next we fit our neural network and random forest models on the training sets and tested their prediction performance on our test sets. The results for models for both guards and forwards are provided below.

Question 2: Best Predictive Model

Below are the confusion matrices for the neural networks prediction performance on the test set for both guards and forwards. The vertical axis represents the "true" cluster and the horizontal axis represents the predicted cluster. Also, let the true positive rate for cluster X be the number of players labeled cluster X and predicted as cluster X divided by all players labeled cluster X. For guards, the neural network is strong at correctly predicting both bottom and middle tier players (clusters 2 and 0), which both have true positive rates of 67%. It struggled a bit with predicting elite guards (true positive rate of 46%), as sometimes the model predicts truly elite guards to just be middle tier. For forwards, the random forest model is successful at predicting both middle tier and elite players (clusters 1 and 2), with true positive rates of 74% and 67% respectively. However, it struggled with predicting bottom tier players (true positive rate of 38%), as the model sometimes mistakes bottom tier players as middle tier players.

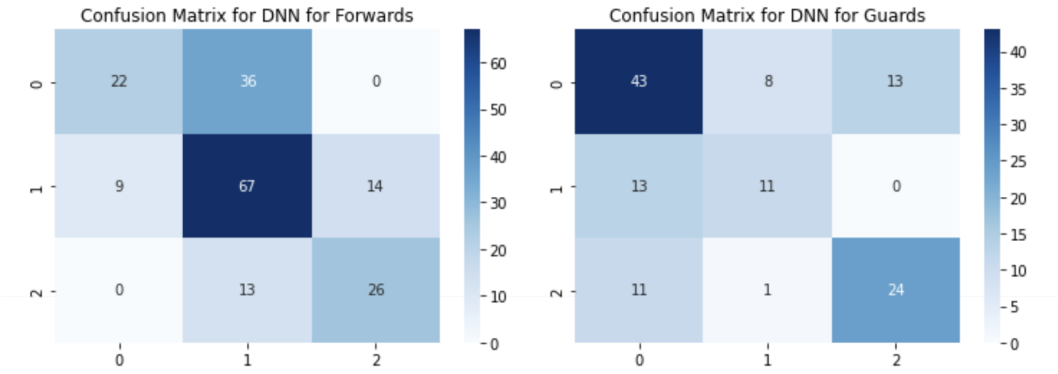

Figure 1: Confusion matrices for deep neural networks results.

Below are the confusion matrices for the random forest prediction performance on the test set for both guards and forwards. These results are quite similar to the neural network results, but provide a slight upgrade in performance. For guards, the true positive rates for bottom, middle, and top tier players are 81%, 61%, and 50% respectively. For forwards, the true positive rates for bottom, middle, and top tier players are 45%, 72%, and 69% respectively.

For guards, the random forest model has a higher overall accuracy than the neural networks (65% versus 63%). For forwards, the random forest model also has a higher overall accuracy than neural networks (63% versus 61%). Furthermore, the random forest was better at predicting bottom and top tier players, while the neural network was better at predicting middle tier players. This applies for both guards and forwards, and is based on comparing true positive rates across the different clusters.

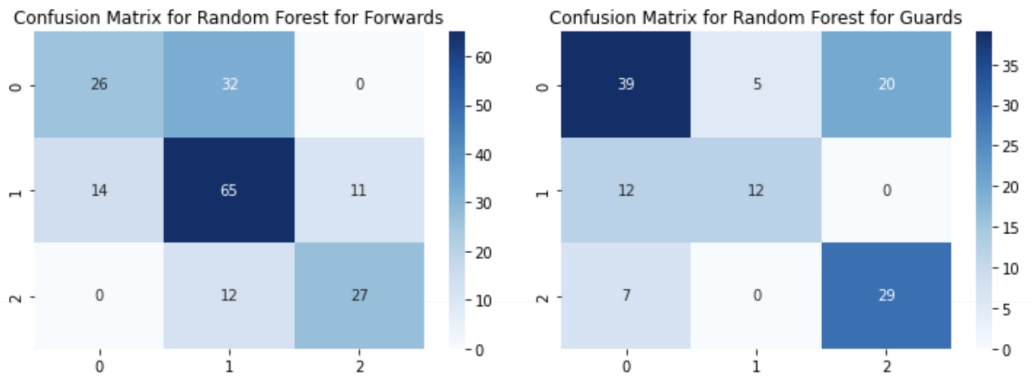

Figure 2: Confusion matrices for random forest results.

Question 3: Most Important Features

Given that the random forest model has better prediction performance than neural networks, we proceed to show visualizations of the importance of various features in the random forest models. This is illustrated by an aggregate measure of the mean decrease in impurity across all the trees in our random forests. Higher values for mean decrease impurity indicate more influential features. One can observe some common themes for both guards and forwards. First, players' statistics from their second and third seasons are better predictors of future performance than their statistics from their first season. Second, minutes per game and points per game seem to be important indicators. Some other notable findings are that turnovers and assists seem to be important predictors for forwards and rebounds seem to be important predictors for guards.

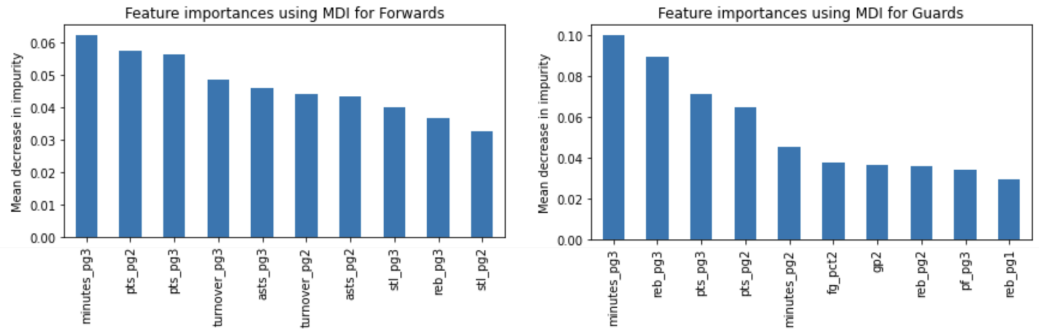

Figure 3: Feature Importance for Guards and Forwards.

5 Conclusion

This paper presented deep neural network and random forest models used to classify and predict NBA guards and forwards' future performance based on early career statistics. Random forest slightly outperformed neural networks in terms of prediction accuracy. The slightly improved performance of random forest may be due to its strong handling of small sample sizes, which is confirmed in the literature. For example, Xu et al.(2021) conducted a conceptual and empirical comparative study of random forests and deep neural networks with ReLU activation function on multi-classification tasks(13). They achieved a consistent finding that random forest performed better for small sample sizes, whereas neural network performed better for large sample sizes. This is aligned with our results as we have a small sample size. Therefore, if NBA teams wish to use machine learning techniques to predict the future performance of their players using player-level data, we would recommend the use of random forest classifiers.

In addition, for both guards and forwards, random forests are slightly better at correctly predicting bottom and top tier players, who generally have more outlier statistics. This makes sense given that literature has shown that random forest classifiers are quite robust to noise and outliers (15). Our analysis of feature importance also provides some helpful insights for NBA player evaluation. Consistent with conventional basketball knowledge, a players' points and minutes per game in their second and third seasons are strong predictors of future success. This is consistent with the theory that current performance is a good indicator for future performance, since better players are generally entrusted with more playing time and score more points.

Our finding that turnovers and assists seem to be important predictors for forwards is interesting, especially since forwards generally do not collect many assists relative to guards. Our results suggest that forwards who are good passers early on in their career may have a higher chance for overall career success, and NBA teams should try and focus more attention on forwards' passing ability when scouting. The importance of turnovers suggests that teams should identify forwards who play responsibly and do not give the ball away often. Furthermore, we find that rebounds seem to be important predictors for guards, which may also be contrary to popular belief since getting rebounds is usually not seen as a primary role for a guard. Therefore, our analysis provides some interesting ideas for young NBA player talent assessment, namely, to identify guards and forwards that possess unique qualities that are not as common for their position. In addition, our results suggest that it may be extra challenging to correctly predict which guards are elite and which forwards will have unsuccessful careers.

The accuracy of our random forest model may be further improved by combining the test and training set together to train the models and using out-of-bag error as a metric for validation. With more samples included in training the model, we could potentially reach a higher accuracy. Another area of further exploration could be boosting methods, such as AdaBoost or gradient boosting. These ensemble methods iteratively fit trees using the unexplained residuals from the previous trees. More research would need to be done to see if these methods can provide an upgrade on random forest for NBA player analysis.

In addition to a small sample size, this analysis is also limited by the player statistics at our disposal from our dataset. Our dataset contains very traditional NBA statistics, but recently many advanced statistics have emerged such as player efficiency rating and win shares. Having these more detailed statistics could serve as alternate measures of player performance and could also help provide more accurate predictions.

References

[1] Nguyen, H. N., Nguyen, D. T. A., Ma, B., & Hu, J. (2022). The application of machine learning and deep learning in sport: predicting NBA players' performance and popularity. Journal of Information and Telecommunication, 6:2, 217-235, DOI: 10.1080/24751839.2021.1977066

[2] Nguyen, N., Ma, B., & Hu, J. (2020). Predicting National Basketball Association Players Performance and Popularity: A Data Mining Approach. In: Nguyen, N.T., Hoang, B.H., Huynh, C.P., Hwang, D., Trawiński, B., Vossen, G. (eds) Computational Collective Intelligence. ICCCI 2020. Lecture Notes in Computer Science(), vol 12496. Springer, Cham. https://doi.org/10.1007/978-3-030-63007-2_23

[3] Moxley, J. H. & Towne, T. J. (2015). Predicting success in the National Basketball Association: Stability & potential. Psychology of Sport and Exercise, 16(1), 128-136. ISSN 1469-0292. doi: 10.1016/j.psychsport.2014.07.003.

[4] Soliman, G., El-Nabawy, A., Misbah, A., & Eldawlatly, S. (2017). Predicting all star player in the national basketball association using random forest. Intelligent Systems Conference (IntelliSys), London, UK, pp. 706-713, doi: 10.1109/IntelliSys.2017.8324371.

[5] Chou, P-H., Chien, T-W., Yang, T-Y., Yeh, Y-T., Chou W., Yeh C.H. (2021) Predicting Active NBA Players Most Likely to Be Inducted into the Basketball Hall of Famers Using Artificial Neural Networks in Microsoft Excel: Development and Usability Study. Int J Environ Res Public Health. Apr 16;18(8):4256. doi: 10.3390/ijerph18084256. PMID: 33923846; PMCID: PMC8072800.

[6] Hastie, T., Tibshirani, R., & Friedman, J. (2017). *The Elements of Statistical Learning: Data Mining, Inference, and Prediction* (12th printing with corrections, Jan 2017). Springer.

[7] Kahn, J. (2003). Neural network prediction of NFL football games. World Wide Web electronic publication, 9, 15.

[8] McCabe, A. & Trevathan, J. (2008). Artificial Intelligence in Sports Prediction. Fifth International Conference on Information Technology: New Generations (itng 2008), Las Vegas, NV, USA, 2008, pp. 1194-1197, doi: 10.1109/ITNG.2008.203.

[9] Loeffelholz, B., Bednar, E., & Bauer, K. W. (2009). Predicting NBA Games Using Neural Networks. Journal of Quantitative Analysis in Sports: Vol. 5: Iss. 1, Article 7.

[10] Barron, D., Ball, G., Robins, M., & Sunderland, C. (2018) Artificial neural networks and player recruitment in professional soccer. PLOS ONE 13(10): e0205818. https://doi.org/10.1371/journal.pone.0205818

[11] Albert, AA., de Mingo López, L.F., Allbright, K., Gómez Blas, N. 2022. A Hybrid Machine Learning Model for Predicting USA NBA All-Stars. Electronics. 11(1):97. https://doi.org/10.3390/electronics11010097

[12] Ramachandran, P., Zoph, B., & Le, Q. V. (2017). Searching for Activation Functions. Retrieved from https://arxiv.org/abs/1710.05941

[13] Xu, H., Kaleab A. K, Will L, Sambit P, Jayanta D, Michael A, Yu-Chung P, Madi K, Florian E, Christopher M. W, Joshua T. V, & Carey E. Priebe. (2021). When are Deep Networks really better than Decision Forests at small sample sizes, and how? Retrieved from https://arxiv.org/abs/2108.13637

[14] Segal, M. R. (2004). Machine Learning Benchmarks and Random Forest Regression. UCSF: Center for Bioinformatics and Molecular Biostatistics. Retrieved from https://escholarship.org/uc/item/35x3v9t4

[15] Rodriguez-Galiano, V. F., Ghimire, B., Rogan, J., Chica-Olmo, M., & Rigol-Sanchez, J. P. (2012). An assessment of the effectiveness of a random forest classifier for land-cover classification ISPRS Journal of Photogrammetry and Remote Sensing, vol. 67, pp. 93–104. doi:10.1016/j.isprsjprs.2011.11.002.