Machine Learning algorithms for classification of an NBA player's position

Project work for the Machine Learning course
CLAMSES - Università degli Studi di Milano-Bicocca

Nicola Pesaresi
n. 842468
n.pesaresi@campus.unimib.it

A.A. 2022/23

Abstract

In this document I apply and compare a variety of Machine Learning algorithms, ranging from the simple K-Means up to SVM and neural networks, to a set of NBA data for the season 2022-23, with the classification task of predicting the position an athlete plays in based on statistics of his scoring and gamestyle.

1 Introduction and overview

The dataset utilised is available at the following link: `https://www.kaggle.com/datasets/vivovinco/20222023-nba-player-stats-regular` and comes directly from `https://www.basketball-reference.com`, a major website for basketball statistics and related news.
It is composed by records for 679 NBA athletes who played in the 2022-23 regular season, each containing 30 different statistics about game performances, scoring and play style. Match analysis in sports is becoming increasingly data-focused every year, and NBA coaches, teams and journalists are basing more and more decisions on the conclusions of statistic analysis of the games.

The goal of my investigation is a better understanding of player positions in basketball, firstly by exploring the data with cluster analysis and then by training a classification algorithm to predict the position of a players by the exam of his season statistics.

To conduct these analyses I used R and RStudio, with a wide array of packages for the implementation of the algorithms utilised.

1.1 Positions in basketball

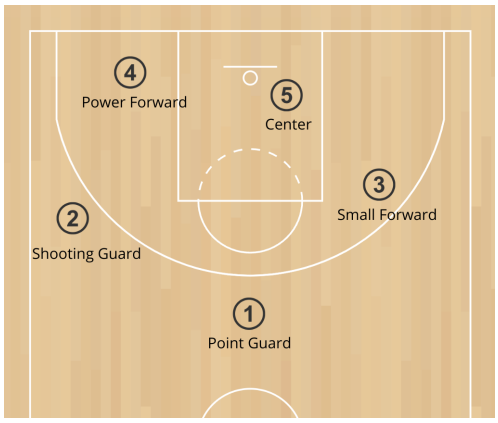

Figure 1: Positions in basketball

To better understand the reasonings of my analysis, a brief introduction on basketball and its positions is necessary. The NBA is a professional basketball league in North America, with 29 teams from the United States and 1 from Canada. It is widely considered the premier basketball league in the world and, as of 2020, its athletes are the best paid in the world by average annual salary per player.

Since the conception of the game in 1891, the same 5 positions are used to describe player roles:

1. **Point Guard (PG)**: typically the team's shortest player and best ball handler and passer. They usually are very fast and are good at driving and short-range. Therefore, they often lead their team in assists and are able to create shots for themselves and their teammates.

2. **Shooting Guard (SG)**: prolific at 3-point and long range shots, they need to be strong and athletic and their main objective is to score points.

3. **Small Forward (SF)**: quick and powerful, the most versatile of the five main basketball positions as they contribute both offensively and defensively, depending on team characteristic and game moment.

4. **Power Forward (PF)**: powerful scorers, they have good footwork in the paint and are able to force players into bad shots or blocking shots when guarding the paint.

5. **Center (C)**: usually the tallest, plays near the baseline or close to the basket. They're skilled at pulling down rebounds, contesting shots, and setting screens on players. Most centers goals are to create possessions by rebounding and trying to stop the other team from scoring in the paint.

Each of these classical positions can be interpreted and adapted in many ways, depending on the skills and weaknesses of the player and his teammates, team tactics and game strategy. Other factors to consider are that the field size is relatively small and the number players low compared to other sports, and an offensive action is limited to 25 seconds, which makes the game dynamic and exciting but also sometimes frenetic. These characteristics of the game make it so that every player needs to do a bit of everything, position wise, during a game, which makes our goal of classification more complicated.

2 Exploratory analysis and pre-processing

2.1 Variables

As mentioned, the dataset is composed by 679 records for 30 variables, which include:

- *Player*, *Age*, *Tm*: athlete's name, age and team. If someone played for more than one team, he will have a record with his statistics for the games in each roster.

- *AST*: assists; passages that lead to the scoring of a goal.

- *ORB*, *DRB*, *TRB*: offensive, defensive and total rebounds; the recovery of the ball after a missed shot.
- *X3P*, *X3PA*, *X3P.*: 3 point shots scored, attempted and score percentage; 3 points are awarded for a goal scored from outside the arc.
- *X2P*, *X2PA*, *X2P.*: 2 point shots scored, attempted and score percentage; 2 points are awarded for a goal scored from inside the arc.
- *FG*, *FG.*: field goals and field goals %; any scored shot other than a free throw.
- *STL*: steals; the defensive action of recovering the ball.
- *BLK*: blocks; the defensive action of deflecting a shot to prevent a score.
- *Pos*: position; one of the five explained in the previous section. This is the response variable in our analysis and will be used for training and testing the algorithms.

From the same source we also have the same data available for the 2022-23 playoffs and for the 2021-22 regular season. This will be useful in the testing phase, to better examine our models' performances and address eventual overfitting problems.

2.2 Exploratory Analysis

The dataset does not contain NA values, as we expect from the level of the source.

There are in the data 5 instances in which the *Pos* variable contains a mixed label, such as "PF-SF", "SF-SG" or "SG-PG". While it's true that many players play in more than one position, and often rotate during the same game, there aren't enough instances of double-labeling to factor this in the analysis. I therefore decided to remove this 5 records from the data.

All the data has been normalised to avoid scale problems when fitting the various models.

The analysis of the boxplots (fig. 2) shows there is a considerable share of outliers for most of the variables, but this is expected as a result of the small numerosity of the sample and the nature of sports in general: the best - and worst - players are always outliers compared to the rest of their field, and their style often influences the rest of the league. I decided therefore to not remove any outlier, as they reflect correctly the population of our sample.

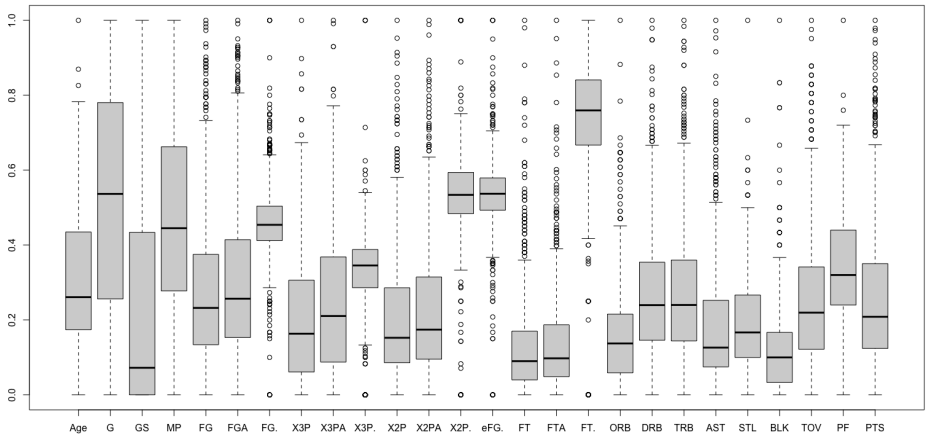

Figure 2: Normalized boxplots of the variables

2.3 Dimensionality reduction

Given the elevate number of variables, I decided to perform a dimensionality reduction to avoid fitting hypercomplicated models. To do so, I used the random forests method. This approach consists in generating a large and carefully constructed set of trees against a target attribute and then use each attribute's usage statistics to find the most informative subset of features. If a variable is often selected as best split, it is most likely an informative feature to retain.

As seen in fig. 3, *AST* is the most relevant feature for *Position* classification. The first 10 variables were kept in the data, as it is an appropriate number of dimensions for our goals and the importance of the features begins to reduce significantly after the tenth.

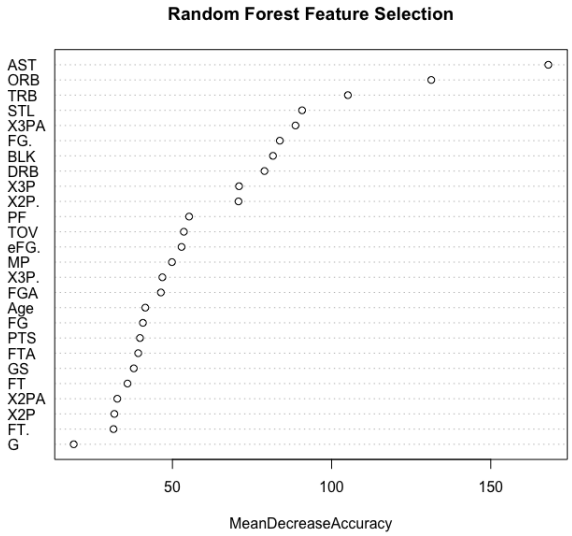

Figure 3: The 10 most influential features were kept in the data

3 Clustering

3.1 K-Means

K-Means is a well-known iterative clustering algorithm that aims to minimize within-cluster variances, assigning each observation to the cluster with the closest centroid and then recalculating the centroid as mean of the observations in the cluster. This steps are repeated until convergence is reached. Euclidean distances are used for measuring the distance from the centroid. I decided to pick the initial values for the centroids at random and execute the algorithm 10 times. K-Means needs the choice of the number of clusters K. A good candidate could be 5, as we know there are 5 positions in basket, but a better approach is to pick the value of K which maximises the silhouette score, a criterion for model choice based on intra-cluster and extra-cluster distances.

The silhouette criterion clearly shows (fig. 4) K=3 is the best choice for this data. This makes sense when examining the pairs plot (fig. 5) and the position distribution across the three clusters (fig. 7), as it can be understood how the algorithm grouped players: the third cluster is the clearest, composed by the tallest players, mostly centers, strong at blocks and rebounding and

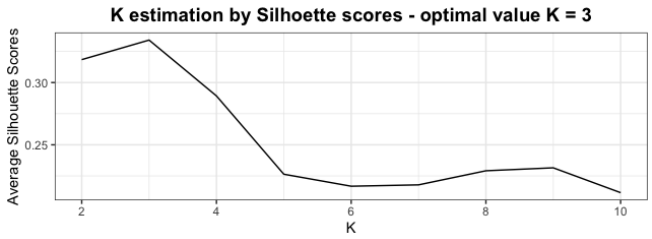

Figure 4: Silhouette scores for K-Means clustering

playing close to the hoop. The second cluster ("green") has the opposite characteristics, smaller athletes great at 3 points distance shots, assists and steals. These players are mostly guards and less physical forwards. The first cluster ("red") consists of the modern all-around kind of players, taller than the typical point guard and shorter than the typical center, good scorers, especially for 2 points. In this group there are also the players with less game time and reserves, which evens out the position distribution a little.

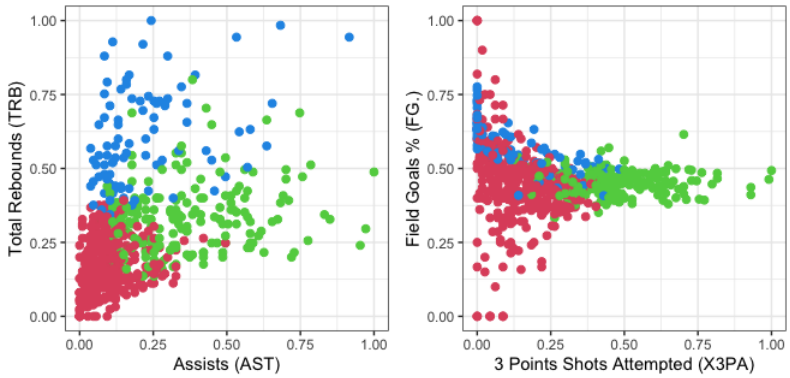

Figure 5: Plots for some of the variables as clustered by K-Means

3.2 Mixture Model

A mixture is a probabilistic model for representing the presence of subpopulations within an overall population. The idea is to assume that each cluster is generated by a different distribution, often using gaussians as in this case, and to use the iterative expectation-maximization algorithm to estimate the parameters of these distributions. This method also calculates for each obser-

vation the posterior probabilities of belonging to the clusters, which allows to switch from soft to hard clustering by picking the group with higher posterior probability.

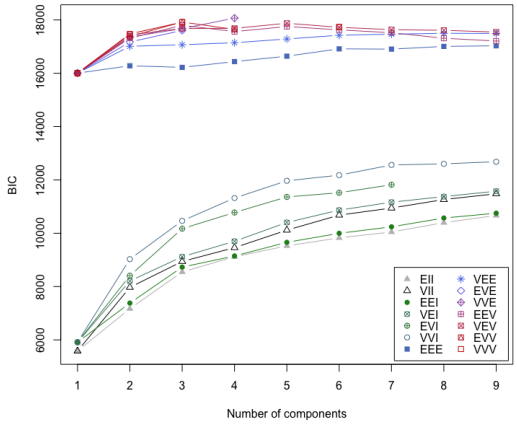

Figure 6: BIC scores for mixture models with different number of components

In this case, we used the BIC criterion (fig. 6) to pick the number of components of the gaussian mixture: the best model fitted had 4 components, closely followed by many 3 components models. The 4 components model resulted to be dividing players more based on skills than position, grouping together the best scorers on one side and the reserves on the other. While this isn't a wrong criteria of separation, for the scope of our analysis the 3 components mixture model is better suited. This model gave a very similar result to the K-Means output, recognising the same pattern of tall and powerful centers ("blue" cluster), agile and good at long distance shots guards ("green"), and all around versatile modern players ("red") (fig. 7). A confirmation that this interpretation of the clustering is solid is the fact that almost no centers have been grouped in the "guards" green cluster, and no point guards in the "centers" blue cluster.

3.3 Hierarchical clustering

The last clustering method we utilised was agglomerative hierarchical clustering. In this approach each observation starts in its own group, and pairs of clusters are merged as one moves up the hierarchy. In this case the best

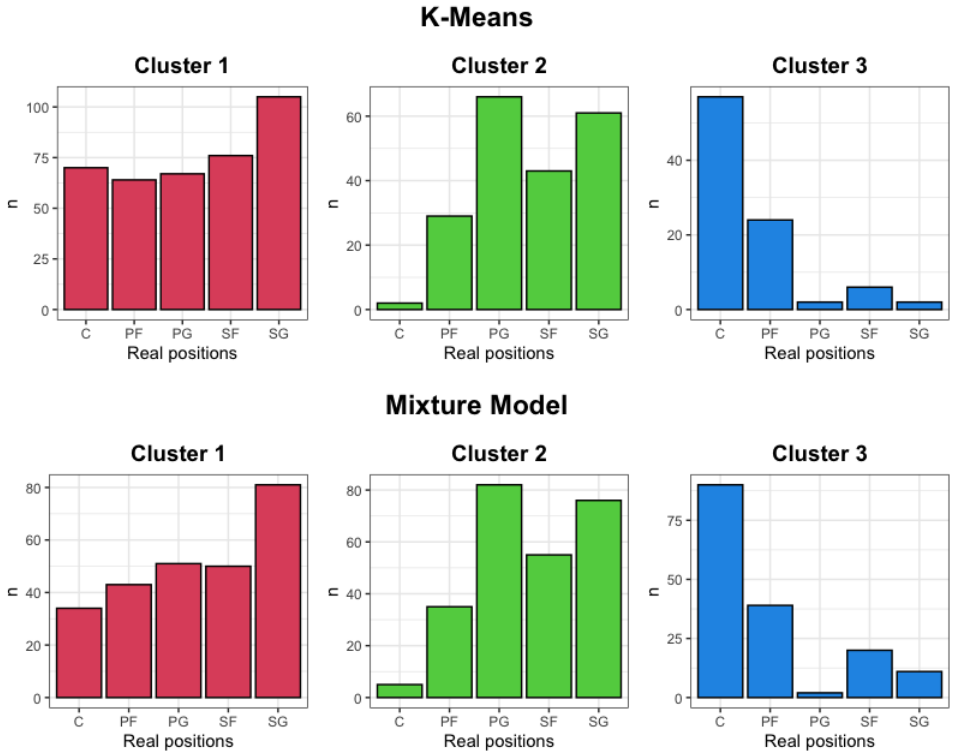

Figure 7: Position distribution across clusters, comparison

silhouette score (0.45) was obtained when cutting for two clusters and utilising the complete link distance for the dendrogram splits (fig. 8). Results for three clusters where close (0.40), while the cut for five clusters fell short (0.26).

3.4 Clustering summary

We utilised the K-Means, mixture model and hierarchical clustering approaches to investigate the nature of this dataset, in particular concerning the position-related characteristics of players. Across the three method, the 3-cluster split performed consistently well, and made sense interpretation wise as it tended to separate very well centers from point guards, the two most distinct positions. In the group with centers we found some power forwards, which are the most similar physically, while point guards have been

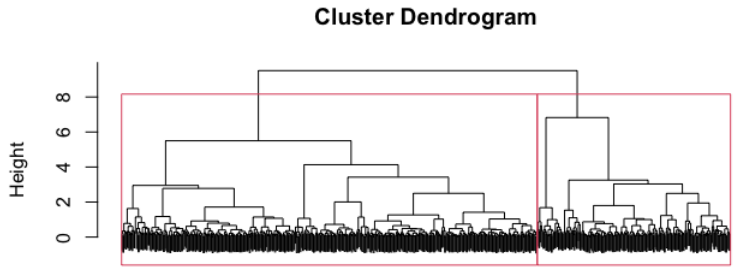

Figure 8: Dendrogram for agglomerative hierarchical clustering

grouped all the smaller players and long range shooters. The other cluster represented more versatile and all around players, together with reserves and athletes with less game time. As it's clear, a 2-group split also makes sense, ad it would split this last cluster in the other two, but it would probably be too simple of a view of the characteristics of NBA players. The mixture model even individuated a 4-cluster split more focused on players' scoring, gametime and overall skill level, but it was not the objective of this analysis to further elaborate on this. Finally, it is evident that a precise split in 5 groups, as the number of traditional positions, it's not easily achievable. NBA players are simply very good at many things, and are required to play in more than one position during the season and even during single matches, so the distinction between roles isn't as clear as one could expect.

4 Classification

The goal of this section is the training of a series of model whose goal is to predict the position of a player based on his season statistics. The data was split in a training set, consisting in 70% of the original records, and test set. Since the classes to predict have roughly the same numerosity, the training set was picked at random.

After the training phase is concluded, the models estimated were also tested on two different datasets, one with data for the 2022-23 NBA playoffs and one with data for the 2021-22 regular season.

4.1 K - Nearest Neighbours

The KNN algorithm searches for the K closest records in the data to every observation, and generates the prediction based on target values of those neighbours. The only hyperparameter to tune , the number of neighbours K, was estimated as 27 by cross validation (fig. 9).

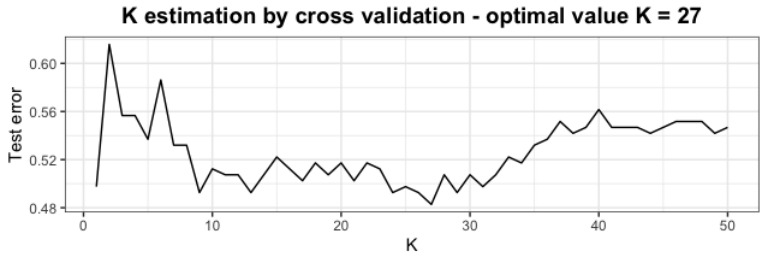

Figure 9: Test error values calculated for different values of K during KNN cross validation

This model performs a 0.50 accuracy on the test set. This is not terrible for a simple algorithm like KNN, but confirms the difficulties of detecting the more versatile players, small and power forwards.

KNN -Accuracy: 0.5025

	C	PF	PG	SF	SG
Sensitivity	0.65	0.40	0.64	0.24	0.62
Specificity	0.92	0.82	0.87	0.99	0.79

4.2 Support Vector Machine

A Support Vector Machine works by constructing a set of hyperplanes in a high dimensional space so that the distance from the hyperplanes to the nearest data point on each side is maximized. New examples are then mapped into that same space and predicted to belong to a category based on which side of the gap they fall.

A strength of this family of algorithms is the ability to perform non linear classification by using a kernel function. In this case, we compare the results obtained using a radial kernel and a polynomial kernel.

SVM - radial - Accuracy: 0.5419

	C	PF	PG	SF	SG
Sensitivity	0.73	0.67	0.53	0.24	0.64
Specificity	0.97	0.86	0.91	0.96	0.73

SVM - polynomial - Accuracy: 0.4926

	C	PF	PG	SF	SG
Sensitivity	0.46	0.50	0.61	0.44	0.48
Specificity	0.92	0.82	0.88	0.88	0.87

In both cases the hyperparameters to tune are two: *Gamma* and *Cost*. The first represents how much curvature we want in the decision boundary, the second is the penalty we want to set for not classifying sample points correctly. These were estimated via grid search over supplied parameter ranges.

The two accuracies on the test set are pretty close, with the radial SVM being a little better. While its sensitivity values are pretty close to the KNN model, the polynomial svm is notably worse in detecting centers, while compensating a little with small forwards, the main flaw of the other two.

4.3 Random Forest

The random forest method, used in the pre-processing phase of this analysis for dimensionality reduction, can also be used for classification, by predicting the class selected by most trees. The model we fitted uses 1000 decision tree, because during testing the accuracy did not improve when increasing the number after this threshold. The results are close to those of KNN and radial SVM, with the detection of small forwards being its biggest problem.

Random Forest - Accuracy: 0.532

	C	PF	PG	SF	SG
Sensitivity	0.59	0.57	0.69	0.32	0.56
Specificity	0.95	0.83	0.90	0.92	0.82

4.4 Neural Networks

Artificial neural networks are a branch of machine learning algorithms which loosely model the biological brain. They are based on a collection of nodes, called neurons, connected to each other and passing on signals. Neurons are organized in layers: an input layer, an output layer, and a number of hidden layers inbetween.

This model has a single hidden layer of 8 neurons and a decay parameter of 0.1, where this last parameter serves to penalize complexity when estimating the weights of the net to avoid problems of overfitting. Size and decay were tuned with a 10-fold cross validation on the training set, each performing a grid search over supplied parameter ranges.

Neural Network - Accuracy: 0.5911

	C	PF	PG	SF	SG
Sensitivity	0.88	0.48	0.46	0.36	0.66
Specificity	0.91	0.94	0.94	0.95	0.74

The neural network model was the best performer by test set accuracy so far, fruit of the ability of detecting centers with 0.88 accuracy, but does not solve the problems of detecting the less defined roles like small forwards.

Figure 10: Confusion matrices on of the prediction on the test set of 4 of the models trained

4.5 Model comparisons on different test sets

To validate our testing results, we tested the models on data from the 22-23 playoffs and 21-22 regular season. When examining this comparison, it must be kept in mind that the game style in the playoffs is slightly different from the regular season, as matches are more competitive, many teams (and therefore their tactics) are eliminated, and the sample pool is simply smaller. Different seasons also present some differences in game style, as coaches are replaced, players change teams and new tactics can emerge. All of this makes a slight drop in the accuracy of the model expected, but still the core features of the game remain the same and thus this testing has its significance.

Model performances of different test sets

	KNN	SVM - rad	SVM - poly	RF	NNet
Original test set	0.52	0.53	0.51	0.53	0.59
Playoffs 22-23	0.47	0.45	0.38	0.52	0.52
Regular season 21-22	0.46	0.49	0.46	0.52	0.50

As expected, all the models have a drop in performance on the new test sets, some more significant than others. Everything considered, the neural net model and the random forest one seem to be the best among the fitted models, as they can consistently classify correctly half of the observations, but this results confirms the interpretation emerged during the clustering analysis: most players aren't strictly specialised in only one position, instead they can and often will play more than one position, adapting its interpretation to his characteristics. Centers remain the easiest to detect, being the position which differentiates itself the most from the others.

5 Classification in three classes

To further investigate the results emerged in the previous analyses, I decided train one more model, still with the task of prediction the position of a player, but this time grouped in only 3 categories, as suggested by the cluster analysis.

- category **C**, composed only by centers.
- category **G**, composed by Point Guards and Shooting Guards.

- category **F**, composed by Small Forwards and Power Forwards.

This way I hope to reduce the "confusion" created by players who can play more than one position, and let the algorithm focus more on the different typology of athlete that emerged from the cluster analysis.

This last model is a neural network with a 6 neurons single hidden layers and a decay rate of 0.2. These hyperparameters where tuned via 10 fold cross validation grid search over supplied parameter ranges. The training set is again composed by 70% of the original dataset normalised. The testing is done on the remaining 30% of the records and on the data of playoffs 22-23 and regular season 21-22.

3 class Neural Network classifier

	Accuracy	C - Sens.	G - Sens.	F - Sens.
Original test set	0.78	0.68	0.79	0.96
Playoffs 22-23	0.67	0.64	0.93	0.99
Regular season 21-22	0.72	0.67	0.64	0.93

This model performs at 0.78 accuracy on his relative test set, a significant improvement over 5-levels classification. Results on the additional test sets for 22-23 playoffs and 21-22 regular season score lower, but still maintain a good degree of accuracy, which is satisfying considering it is data relative to a different set of games. Surprisingly, centers are now the class with lower sensitivity values. This could be caused by the fact that, having grouped together two of the original positions in each of the other two classes, the training sample ends up being the one with a smaller sample, but experimenting with a bootstrap resampling to even the classes' sizes does has not shown an improvement on this front.

6 Summary and conclusions

The goal of this document was the exploration of the relationship between the statistics of the 22-23 NBA regular season and the position of the players, and the fitting of a series of models which could predict an athlete's position based on such data. This analysis has found that the nature of the game, which asks many players to be as versatile and all around skilled as possible, to be effective in different roles on the pitch instead of hyper-specialising in

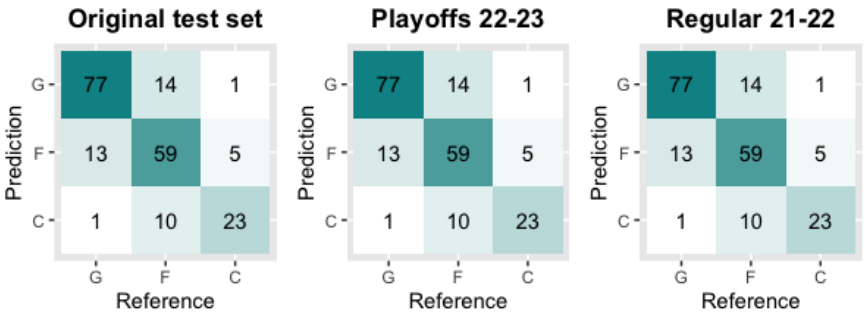

Figure 11: Confusion matrices of 3-levels classification on the different test sets

a few fundamentals, makes it hard to box the athletes in the five traditional positions. Instead, it is easier and probably better suited to differentiate in three main categories: the first group is that of taller and more static players, typically centers, who play close the hoop and are very good at rebounding. The second is composed by smaller and more agile players, with great long range shooting abilities, closer to the guard label. The third is the group of forwards, the most versatile and all around players, good at stealing and scoring.

The final model that has been trained is a neural network with a 6-units single hidden layer, capable of classifying player in these three categories with a 78% accuracy on the main test set provided. This result is indicative of the power of machine learning techniques and the usefulness of integrating this processes with traditional sports analytics to boost our understanding of the games, as well as the vastity of the domain of applicability of this methods.

References

[1] *Basketball positions*. https://en.wikipedia.org/wiki/Basketball_positions.

[2] *Basketball Reference*. https://www.basketball-reference.com/leagues/NBA_2023_totals.html.

[3] Andrew Baumann. "A Multi-Stage Clustering Algorithm to Re-Evaluate Basketball Positions and Performance Analysis". PhD thesis. Dublin, National College of Ireland, 2022.

[4] F. Bianchi and T. Facchinetti. *Towars a new meaning of modern basketball players positions*. 2016.

[5] A. Candelieri. *Machine Learning course*. https://elearning.unimib.it/course/view.php?id=44929.

[6] Alexander L Hedquist. "Redefining NBA Basketball Positions Through Visualization and Mega-Cluster Analysis". PhD thesis. Utah State University, 2022.

[7] M. Mohri, A. Rostamizadeh, and A. Talwalkar. *Foundations of machine learning*. MIT press, 2018.

[8] *NBA player stats 21-22*. https://www.kaggle.com/datasets/vivovinco/nba-player-stats/.

[9] *NBA player stats 22-23*. https://www.kaggle.com/datasets/vivovinco/20222023-nba-player-stats-regular.

[10] *NBA Stats Glossary*. https://www.nba.com/stats/help/glossary#efgpct.