Tilburg University

PREDICTING THE NBA ALL-STAR ROSTER THROUGH MACHINE LEARNING

REVEALING HIDDEN BIAS IN NBA ALL-STAR PICKS

ALEJANDRO LUIS BARCALA PAOLILLO

THESIS SUBMITTED IN PARTIAL FULFILLMENT
OF THE REQUIREMENTS FOR THE DEGREE OF
MASTER OF SCIENCE IN DATA SCIENCE & SOCIETY
AT THE SCHOOL OF HUMANITIES AND DIGITAL SCIENCES
OF TILBURG UNIVERSITY

STUDENT NUMBER

2088400

COMMITTEE

dr. Paris Mavromoustakos Blom
dr. Federico Zamberlan

LOCATION

Tilburg University
School of Humanities and Digital Sciences
Department of Cognitive Science &
Artificial Intelligence
Tilburg, The Netherlands

DATE

May 20th, 2024

WORD COUNT

8339

ACKNOWLEDGMENTS

I want to start by thanking everybody who was involved in this master thesis. After two years of studying, I hope I don"t make this list too long. First of all Paris Mavromoustakos Blom, you were exactly the thesis supervisor I needed to do get this job done, always with an ease solution to all my doubts and never answering an email in longer than two hours. To my friends who were by my side along the way, either in The Netherlands or in Uruguay, please feel this achievement in part as if it was yours. An special acknowledgment to my family back home for the unconditional support when things were not going the way I expected, and to my ace under my sleeve my sister Victoria for standing alongside me besides my annoying requests for feedback. To my second reader Federico Zamberlan, I am happy that there is another person from the other side of the puddle supervising this thesis. And to all of those whose path was not paved in the first try, as two guys I admire Rafael Nadal and Jose Mujica many times times have said, "Life does not lie in either success or failure but on standing up with bigger desire every time that one falls.".

PREDICTING THE NBA ALL-STAR ROSTER THROUGH MACHINE LEARNING

REVEALING HIDDEN BIAS IN NBA ALL-STAR PICKS

ALEJANDRO LUIS BARCALA PAOLILLO

Abstract

In the last few decades, the professionalism in the sports environment had sky rocketed in both aspects competence and financially wise, which may explain why more technical teams using methods like machine learning and artificial intelligence commence to play a role in these fields as well.

Given an NBA data set of performance of players during the last three decades and extracted information about NBA All-Star rosters for the exact period, it was predicted with a 0.94 accuracy, 0.71 true positives ratio and 0.98 true negatives ratio and a macro average of 0.83 the prediction of NBA All-Star players. Certain features like "name" and "country" were eliminated to explore other categories which may affect the players eligibility, as their names or country can be considered already as a biasing factor due to big success or xenophobia and considerable for country imbalance in the dataset. Moreover, chi-square statistical tests were performed in different categories of the true positive instances of the best performing model (Random Forest Classifier), finding a significance p-value of 0.001 and 0.004 after controlling for Bonferroni in the category "college", as well as a minor significant result for the "draft number" category of 0.015 and 0.044 after measuring with Bonferroni correction.

The results showed that besides extracting some features considered as bias like "name" and "country", categories like "college" and "draft number" had a significant p-value in the prediction of the algorithm when predicting the selected players, possibly translating into the people who chose those players who are the NBA fans, the specialized media and the NBA coaches been consciously or unconsciously driven mostly by the college feature and formulates other questions if its an institutional and/or socioeconomic condition. Aiming to unveil the bias in the sport and evidencing that there is a skewed preference for some universities for example, leads us to reflect in the casualty and causes of the "college" results, if there needs to be some institutional policies to be considered considering the outcomes of the study.

1 DATA SOURCE, ETHICS, CODE, AND TECHNOLOGY STATEMENT

1.1 *Source/Code/Ethics/Technology Statement Example*

The programming language used for the writing of the master thesis was Python and the environment used was VS Code. The data sets used for the research were publicly available online, one part was extracted from a database from the sports league being the source named stats.nba.com, whereas the rest of it was mined by the researcher for this experiment from a historical NBA statistics website, mined in an Excel worksheet and then imported to Python environment. The elaboration of the writing was enhanced by the use of Thesaurus.com, Grammarly, and Microsoft word spelling corrector. The main libraries used in VS Code environment were "sklearn", "mat.plotlib", "pandas" and "scipy.stats" for the Machine Learning, the plotting, the data framing, and the statistical models, respectively. The figures or tables that were not represented by the "matplotlib" from Python were elaborated for the research using LaTex coding.

2 INTRODUCTION

Since childhood, I have been passionate about sports, whether it was tennis, football, or basketball. Sports, especially basketball, not only benefit physical and mental health but also provide significant entertainment worldwide. Different countries practice various sports at multiple levels, which have developed greatly due to technological advancements and professional competition. This has led to multi-billion euro budgets for sports organizations. Club management and professional scouts in the NBA are particularly interested in identifying top players.

Society also generates great excitement about discovering the next great athletes. This research aims to analyze if there are any biases in the selection of NBA All-Star players, using data from NBA seasons between 1996 and 2021. These players are selected by fans, specialized media, and team coaches.

Club management often faces challenges in deciding where to allocate their budget for future investments (Guevara et al., 2021). A high-performance prediction model can mitigate these challenges, reducing financial risks and promoting long-term sustainability. This research is also beneficial for professional scouts, helping them identify talent and strategic players more effectively.

The goal of this research is to use ML models to predict the selection of NBA All-Star players, compare the performance of these models, and identify any significant biases in categorical features. For the aims of this research, big weight features such as names and country of origin were taken out to be possible to identify other bias present in this decision if we excluded big bias topics like by name selection or xenophobia or racism. To address that, statistical chi-square tests were performed on the category features in the best-performing ML model. As previously mentioned, some features with a heavy bias influence, such as "name" and "country," were eliminated for the objectives of the research.

This study addresses the decision-making bias in selecting the best players of the season. While there is existing research on decision game bias related to race or ethnicity, there are no studies on NBA All-Star selection bias involving fan ballots, NBA coaches, and specialized media votes.

Through this research, NBA directories and scouts will be able to identify which players are likely to be selected for the All-Star game using different binary classification ML models. They will also be able to detect and avoid certain biases in the modeling process, ensuring that potential talents are not overlooked. This research will compare various binary classification models, such as logistic regression, decision trees, and random forests, and examine the extent of bias interference in All-Star player recruitment (Langel & Tillé, 2011).

As a result of this, is that we formulate the following three research questions:

RQ1 *To what extent can the selection of NBA All-Star players be predicted using a machine learning model?*

RQ2 *To what extent can a categorical feature bias the prediction outcome of NBA All-Star players?*

RQ3 *How do this different machine learning models perform compared to each other given the NBA dataset?*

3 LITERATURE REVIEW

Many studies have demonstrated the potential of using Machine Learning models to predict NBA All-Star rosters. These studies, alongside investigations into biases such as racial and socioeconomic factors, provide a robust foundation for exploring categorical biases within this field. This background research underscores the feasibility of investigating societal biases in sports through the application of Machine Learning and statistical testing, which is the focus of this thesis. Understanding bias in sports is not only relevant but crucial for the broader application of data science in societal contexts, including but not limited to the NBA and other sports domains.

The NBA, founded in 1949, was relaunched as an international entertainment endeavor in 1984. The first NBA All-Star game, a showcase of talent between the East and West Conference teams, occurred in 1951 (Grossman, 2014). The selection process for these All-Star players involves fan ballots and votes from head coaches of each conference, resulting in 24 players: 5 starters and 7 substitutes on each side.

The NBA"s longstanding engagement with data collection has supported numerous research endeavors, aiding our investigation. For instance, a study on predicting All-Star players using a random forest model identified the number of points per minute as the most significant feature, achieving a 0.925 accuracy (Soliman et al., 2017). This highlights the effectiveness of Machine Learning in player selection predictions.

Addressing biases, one study suggests racial discrimination by coaches towards players of different races, although this bias diminishes with longer team tenure, indicating a group bias rather than an individual one (Zhang, 2017). This finding aligns with another research indicating that discrimination can intensify under pressure during in-game decisions but remains present in deliberate choices like starting lineup selections (Schroffel & Magee, 2012).

Further research on NBA data sets reveals that advanced statistics, derived from elementary statistics using ML models, significantly outperform basic statistics. For example, Decision Trees applied to advanced statistics yielded a 0.91 accuracy compared to 0.75 for elementary statistics. Similarly, Random Forests and Gradient Boosting models showed improved performance with advanced statistics, achieving accuracy's of 0.88 and 0.92, respectively (Wang & Fan, 2021).

Bias extends beyond basketball. In soccer, players from higher socioeconomic backgrounds have better chances of success, influencing their education, performance, and promotions (Gandelman, 2009). Such findings can inform our understanding of bias in NBA team selections.

A study using open-source NBA statistics found that a hybrid model, combining an Artificial Neural Network with outputs from other models, was most effective in identifying true positives and negatives for All-Star selections. This model emphasized specificity and sensitivity in player selection (Wang & Fan, 2021).

Moreover, research has identified additional characteristics valued by fans for All-Star consideration. Three-point shooting prowess and "masculine attitude"—evident in rebounds and free throws per game—make players more appealing candidates (Levine, 2019).

In the realm of university sports, teams with longer winning streaks tend to experience lower acceptance rates, increased donations, and higher academic applications, ultimately enhancing the institution"s academic reputation (Stinson & Howard, 2007). This illustrates the broader societal impact of sports performance.

By weaving these studies together, we can better understand the intersection of bias, performance, and societal implications within the NBA and beyond. This thesis aims to build on this foundation, employing Machine Learning and statistical methods to uncover deeper insights into these biases.

For the following academic research, efforts were made to analyze through data science and statistical techniques if there was a possibility to track a biased category, putting aside names and nationalities due to potential confounding statistical reasons for the machine learning models.

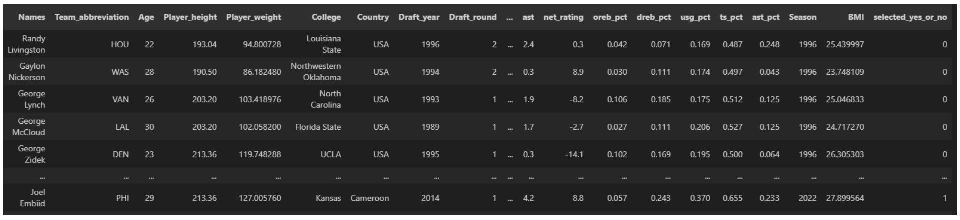
Inside an NBA dataset enables a comparison with the NBA All-Star rooster, for both in the same period from 1996 to 2021 seasons, with players performances and background information, as well as the decisions of players who were selected to be All-Star as chosen by fans, specialized media and coaches. The research gap is constructed through a critical view through this machine learning and statistical methods and tools to identify potential biases in the decision of players using these different methods and different context of study which may also be applied to other domains.

4 METHODS

4.1 *Dataset Description*

The dataset used for this research is an NBA dataset. The data frame is composed by 25 columns and 12804 rows. Each row is one of the players of the NBA players from a specific season mentioned in one of the columns. The dataset comprehends the NBA seasons from 1996 to 2021. It has many continuous features of these different players like measurements, weight and age. As well as many other categorical features as which University team the player started their youth career before being drafted by an NBA team. The round of draft of the NBA. A draft is a round where the worst teams to the best team of the past season are enabled to draft new young players out of college to balance the level of the game and replacing it by the new generations (Hussey, 2021).

Also, other features as nationality, which is one of our most relevant features as the research pretends to explore data analysis regarding nationality bias of the NBA players.



Names	Team_abbreviation	Age	Player_height	Player_weight	College	Country	Draft_year	Draft_round	...	ast	net_rating	oreb_pct	dreb_pct	usg_pct	ts_pct	ast_pct	Season	BMI	selected_yes_or_no
Randy Livingston	HOU	22	193.04	94.800728	Louisiana State	USA	1996	2	...	2.4	0.3	0.042	0.071	0.169	0.487	0.248	1996	25.439997	0
Gaylon Nickerson	WAS	28	190.50	86.182480	Northwestern Oklahoma	USA	1994	2	...	0.3	8.9	0.030	0.111	0.174	0.497	0.043	1996	23.748109	0
George Lynch	VAN	26	203.20	103.418976	North Carolina	USA	1993	1	...	1.9	-8.2	0.106	0.185	0.175	0.512	0.125	1996	25.046833	0
George McCloud	LAL	30	203.20	102.058200	Florida State	USA	1989	1	...	1.7	-2.7	0.027	0.111	0.206	0.527	0.125	1996	24.717270	0
George Zidek	DEN	23	213.36	119.748288	UCLA	USA	1995	1	...	0.3	-14.1	0.102	0.169	0.195	0.500	0.064	1996	26.305303	0
...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...
Joel Embiid	PHI	29	213.36	127.005760	Kansas	Cameroon	2014	1	...	4.2	8.8	0.057	0.243	0.370	0.655	0.233	2022	27.899564	1

Figure 1: Dataset features illustration

Hereby figure 1 illustrates some rows of the dataset, with its different columns and the predicting feature, selected "yes or no".

The NBA dataset can be found in nba.stats.com, but the predicted feature of the model was mined for this experiment and extracted as a second dataset which was consequently merged to run the different models.

The meaning of the columns names from the dataset where abbreviated to what figure 1 shows. Here the explanation of the abbreviated column names, "pts" refers to points per game in a season, "usg.pct" refers to percentage of game play during a match, "reb" refers to rebounds per game in a season, "ast" refers to assist provided per game in a season and "gp" refers to the amount of games played in a season.

On the other hand, there is a self-mined dataset from a source of statistics about the NBA. The website used for this is Basketball-Reference.com. This dataset has been merged to the first one from nba.stats.com, mentioned previously, in order to identify via dummy coding columns, which player was or was not selected for the NBA All-Star match.

The second dataset was implemented mined through different stats in the same web page and exported to Microsoft Excel worksheet and for the specific purpose of the research was then exported as a pandas data frame to the working environment.

4.2 *Algorithms and Software*

For the research the selected computer language was Python and the environment used to run all the algorithmic procedures was VS Code (Sanner et al., 1999).

For this classification problem, the objective is to identify by the binary prediction of YES/NO of the selected players for the All-Star team, to what extent there is a nationality bias in the recruitment of these All-Star team players. For this, different models are going to be implemented to selected players, and then control for nationality in the different cases.

The models that were implemented for this parts of the research are binary classification models, one of them probabilistic and the two others tree-based algorithmic models. As it is a binary classification problem of "Yes" or "No", or also "1" or "0" respectively, is that these were the output categories.

One of the models was a probabilistic function, which was a Logistic Regression Classifier (hereafter; LR). On the other hand for the tree-based models, the work was done with a Decision Tree Classifier (hereafter; DT), and with a Random Forest Classifier (hereafter; RF).

The logistic regression model is a linear regression model that implements the sigmoid function, and regresses each candidate with a certain number in between 0 and 1 and assigns the output value of the regression to its respective class, hence a "NO" class if it is closer to a 0 or a "YES" class if it is closer to a 1, and its meant by "closer" as it may assign a number in between 0 and 100 in terms for percentage for each instance, and the one were it lies with a bigger probability to is where it will be located, either "1" or "0" or "Yes" or "No" respectively. (Solutions, 2016).

A DT is an algorithm capable of been used for either a continuous or a categorical output. For the case of this research a decision tree with a categorical outcome suits best, and more specifically as a binary classification outcome. It works by splitting the dataset with the features that have the highest entropy at the beginning of the iteration and aimed to conclude that iteration measure with the questions with the lowest one, and therefore it produces the outcome class. Entropy is a measurement for uncertainty, which quantifies the disparity of the features in the dataset. Low entropy means that our dataset is very easy to interpret within the different classes, whereas a high entropy makes classification work more hardly separable (Suthaharan & Suthaharan, 2016).

Considering the performance of each of the models it will be possible to identify if the nationality bias of the players is more notorious when more accurate, or if when higher accuracy more bias in the selection of the players. In the case there is a non-existing influence of this feature in the modeling, there can be other features with sociological relevance when having the results of the three different models, Logistic Regression, Decision Tree and Random Forest.

4.3 *Evaluation Method*

The results will be measured with sensitivity, accuracy and specificity for the selection of the NBA All-Star teams. As it is a way of considering how accurate the Yes detection are, by the implementation of sensitivity and how well our algorithms classify the No class with the use of specificity measurement for the three different models, logistic regression classifier, decision tree classifier and random forest classifier.

For the bias implementation analysis in the experiment, the best performing model was choose, and based on it the statistical analysis was conducted through a chi-square test since the categorical variables were the ones of interest. To avoid bias and weight of certain features, the columns like "name" and "country" were extracted from the data set.

The objective was identifying through statistical testing, if by the learning of the algorithm predictions, some group or feature in the dataset may be affecting the prediction of NBA All-Star players. Therefore check if there is statistical significance (p-value < 0.05) showing that the errors in a certain category of the ML model is significant after controlling for a certain variable, if so we would have found some feature or features that can be affecting the "Yes" or "No" decision (Tallarida et al., 1987).

4.4 *Preprocessing*

For the master thesis different data mining techniques were necessary to engineer the inputs for the machine learning models. Two datasets for the experiment were merged in order to perform results altogether. A dataset with season performance features and characteristics of the players towards a second data set that was mined for the purpose of this research to extract in a main dataset if each player with each specifics features and characteristics was or was not recruited to be an All-Star season player.

In figure 2, a flowchart pipeline was made to represent visually what processes were implemented for the experiment.

Due to the mentioned process it was possible to calculate algorithmicaly if the players were selected or not, performing three different binary classification models, due to the dummy columns technique used or the "One hot encoder" code in Python programming language.

Before the start of the model runs, an relevant step in the preprocessing of the names features occurred. Before the feature "name" was deleted from the dataset, it needed to be sure that each row had a correct selection with an exact match if selected or not, after running a double iteration in between "names" on each dataset and iterating for the season of each player from 1996 to 2021, the match of the whole dataset was possible. Characters from both datasets need to match identically in order that the model may output the decision class "1" or "0" correctly.

After that, every typo or byte size miss match had been properly cleaned and preprocessed to work with one whole harmonized dataset. For this part, from the package named "fuzzywuzzy" the "fuzz" function was imported. Thanks to this Python library that was installed, it was

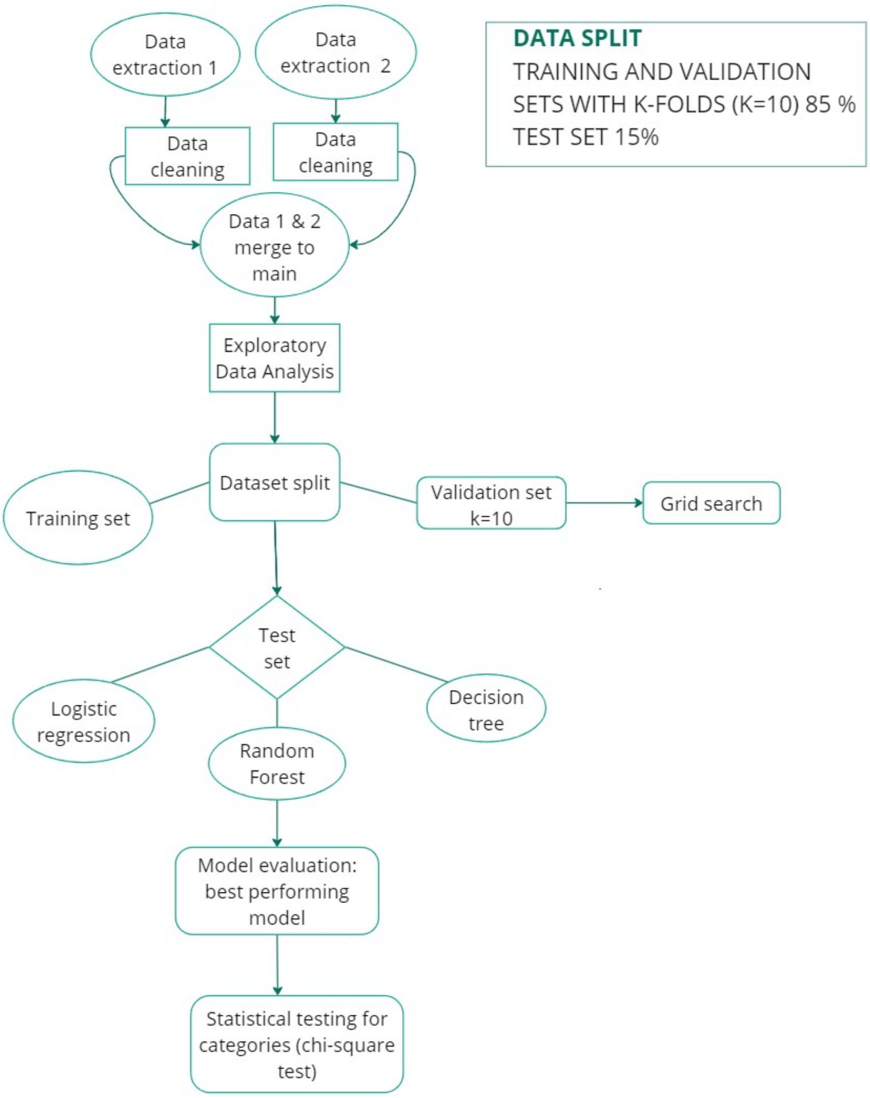


Figure 2: Data science flowchart pipeline

possible to concatenate a bigger dataset that contains 2551 original NBA names with 187 exact matches from the 188 NBA All-Star names. As for this, the following step was to create dummy columns, or binary classification for the selection or not selected for each season from 1996 to 2021.

This left the data with 25 year rows, so there was one column missing from 2021 until 1996, as there was a lockout for season 1998, for such a reason is that there were no measures available for the NBA All-Star game for that particular year of 1998 (Ringer, n.d.).

The output of this process is going through a data frame which concatenates the different player features from 1996 to 2021, with the different columns such as the point scoring in their games, assists, games played, percentage of minutes played, their country of origin, names, age, their previous university of competition, their weight, their height, their draft pick and their draft round, and to all this information is that we add the All-Star selection or not as a last column dataset.

Furthermore, each row of this 12844 that represents a specific player are related to an amount of 2551 unique NBA player names, as players participate in the NBA for several seasons. From all this 2551 different names, 187 are a selected group that participated in the NBA All-Star teams at the end of the season, and its what the model was based on.

The aim of this ML models is to have accurate outputs of accuracy with our true labels, in order to check statically to what extent there is a bias, and if some what kind, in the selection of the All-Star players rosters.

For the modelling, the dataset was broke down to a train and validation set with a data amount of 85 percent, leaving the latter 15 percent for the test set. Inside the 85 percent split for training and validation set, a k-folds of k=10 was made within a grid search to show the best hyper parameters for each model. The data split made was of 85 for validation and train with a k-fold = 10, and the remaining 15 percent for the test, given the size of the dataset of 12800 times 25. Since it is not that huge, a model with a smaller test set is better tailored for a better performing predicting model (Vabalas et al., 2019).

The validation set was carried out through a k-fold equaled 10 for the three different models (Decision Tree Classifier, Random Forest and Logistic Regression) with a grid search, as well as 15 percent of the data set for the test sets of the three models (Shekar & Dagnew, 2019)..

Firstly, the training process was done which finds a way in the model to transform the input looking for a specific instance in the dataset which is known by the algorithm, hence accuracy is 100. It works as a mathematics function in which there is an "X" input and the function finds an equation to achieve the output "Y", named as the true values.

This exercise is done several times, as many "Y" instances in each split of the data. For this research it is 85 percent for the training set and validation set with k = 10 folds, so 10917 rows times 25 columns, and the remaining 15 percent for the test set, so 1927 rows and 25 columns. The dataset is composed by 12844 rows and 25 columns in total. Given the dimension of it, was that the split was managed for a bigger size in the training to conclude a more robust test set (Vabalas et al., 2019).

The results of the three different models were analyzed using "sklearn metrics" package, utilizing learning performance graphs for the three models, confusion matrices and a feature importance analysis for the best performing model.

To neutralize the problem of class imbalance, a normalization of the features was done across all the numerical columns in the dataset. As mentioned in previous sections, and even after deleting some of the variables like "name" and "country", the results thrown were better than previous literature review papers.

To work with the error analysis, and with it apply statistical testing to see how different groups may affect the errors in the best performing model of the research (random forest), this may infer that after controlling for specific categories from the dataset, the prediction model as well as the people involved themselves may be giving too much relevance to a certain or certain categories.

This made the experiment find insights about the random forest model performance in this data set. This positively impacted the expectations in the research as the features with the most weight in our best performing model were all related to game performance and scoring attributes of the game like points, usage percentage of players, rebounds per game and assists per game.

Although it can also be considered that the usage percentage is already been studied by one of the literature review materials as some racial bias in the selection of players and in minor impact the decision making through out the substitute game play decision which already shows scientific conscious in certain types of bias in the named sports league (Schroffel & Magee, 2012).

Due to the interests of this academic research, was that two column categories were deleted. The aim of the research was to explain the prediction and finding insights that can unveil the decision process of stakeholders and help from society perspective the decision making of management teams of this institutions. Due to that, was that the categories of "name" and "country" were decided to be left out, as they were considered to have a heavy influence.

In the first place, the variable "Name" can affect heavily the decision making as players like "Michael Jordan" or "Stephen Curry" who had more than 10 selections for this team consecutively, they would be automatically calculated as selected because of the name feature, which in reality is not the case.

For the case of country, most of the players are from the United States but at the same time when foreign players succeed they might have a chance in being part of the prestigious All-Star group. As well and as studied in the literature review, different nations in the NBA league could be associated with a certain race, so would be helping less in the unveiling of a bias for reciting the players. To look into a dataset decreasing strong biases to acquire insights through implementation of data science techniques.

Many other literature reviews point out the "players popularity", as well as "social media impact" or "press conference characters", "personality types", etc. This named circumstances can be considered as important weight variables but unfortunately were not measured precisely. Because of the mentioned situation, this can be considered a limitation in the research process, and something to be approached by other researchers in similar fields for future experiments.

This can explain some of the accuracy gap from 0.94 accuracy in our test set to even higher values, as well as the 0.71 for specificity in our best ML model, and even increase the 0.98 performance for the sensitivity measure, giving a total macro average of 0.83.

Moreover, the analysis of this master thesis consisted in leaving some of the features aside and then elaborate in the results of the best model (random forest) and analyze if any categories can be inclining the balance to a side in the decision of the people involved.

4.5 *Models and algorithms used*

The prediction outcome was a binary classification, meaning that the outputs of the model were either number 1 answering to a positive instance “selected”, and number 0 answering to a negative instance of “not selected”.

As in the purposes of answering the third research question, one of the objectives was to work with model comparison performance. Due to that, and our classification problem of player selection, is that the three ML models that were chosen were the following, Decision Tree Classifier (hereafter; DT), a Random Forest Classifier (hereafter; RF), and a Logistic Regression Classifier (hereafter; LR).

The DT is a ML model that does its computational findings through yes or no statements, dividing in each step of the tree or each node with statement with the lowest entropy to the highest entropy. Entropy is a coefficient between 0 and 1 which shows how different our outputs respect to a certain feature are, so decision tree classifiers start the process of a prediction output by splitting the data from more general instances, with less division of the data, and aiming to the next node into the variable with the following highest entropy, and so on. The process ceases when the feature with the highest level of entropy stands. This instance will be the one with the more random outputs.

Also a LR classifier was implemented. A logistic regression classifier model has other particularities worth pointing out. This model works as a linear regression, but outputs categories and a brief explanation of how that happens will introduce it. When working in linear regression models, different from logistic regression, variables have basically a certain ratio and given the input of each of the variables to that ratio, is that a function can regress the continuous output for a certain instance in a linear regression model. Reminding once again that a linear regression works for continuous values, hence numerical output that may not have theoretically speaking any intermediate values between two outputs.

The LR model required some other steps in the feature engineering given the algorithm processing. For the data preparation, besides handling missing values which have already been fixed in previous steps, it is important to handle an encoding of the categorical variables as well as normally distributing the continuous variable values for this specific Logistic Regression model. To deploy the model calculations, the "Logistic Regression" function from "Scikit-Learn" package in Python environment was imported.

A relevant outlook in the LR model and that it differentiates out by the previous RF and DT classifiers is the error function of the algorithm. The error function in LR is the Log loss function or logarithmic loss, it quantifies the difference from predicted to observed values, and it discrepancy for the other tree-based models in that its classification lies in a probabilistic measurement, rather than in an entropy split classification.

For this calculation, also the correction or penalization of the model itself varies, as the wrong classifications will be heavily penalized aiming to neutralized the mistakes to produce a bigger number of true outcomes.

Given the different mathematical natures of this models, it is that they behave and translate the errors of the models with different expressions. Another specific difference lies when the algorithm produces an error, the penalization is bigger as it needs to adjusts its parameters to iterate more precisely for future instances. Whereas the decision tree and the random forest, both tree bases methods, process data to divide it in homogeneous splits working with the concept of entropy and division in its features, and aiming to classify the category outcomes logically.

5 RESULTS

To answer the three research questions the machine learning models comparison in the prediction of the selected players for the NBA All-Star matches were studied together with chi-square statistical tests results.

For it answering of the research question is that two ML approaches were selected, a probabilistic with a logistic regression classifier and two different tree-based methods a random forest classifier and a decision tree classifier.

To initialize the running of models with the LR, the following libraries were needed. The main libraries were "sklear.linear.model" to import the Logistic Regression function as well as the "sklearn.preprocessing" library and "label encoder" to preprocess the data. are the "sklearn.preprocessing" where "StandardScaler" and "LabelEncoder" . This functions worked with the data transformation to normalize and fix the problem of feature imbalance and skewness of values, boosting the model performance.

The columns that were normalized in the data set for the LR model were six, "player weight", "player height", "number of games played per season", "assists per season", "net rating" in a season, "points per game" in that season, as well as their "draft number" (position in the draft), and "number of rebounds" per game in that season.

The "draft number" category is the hiring process for NBA basketball players, in where teams pick the newcomers or more socially known as "rookies" as well as the international players of the NBA season and in other sports leagues alike in the US. New players and internationals players also are hired by the NBA franchise and are paid to do a preseason and then a process of selection takes place. It is a system of rounds, in where the worst ranked team of the previous season gets the first draft pick and so on using the last seasons positions ranking in order to balance the future in between the best and the worst teams of the league, and in the NBA there are just two rounds for the draft pick (Hussey, 2021)o.

On the other hand, the dataset had other two categorical values such as "college", and "draft round". As previously mentioned features such as name and country were took out to extract insights from a dataset without a nationality or name influence, which enabled the experiment satisfactory results.

College indicates which educational institution they had played before becoming professional NBA players, and draft rounds involve only two rounds, meaning there are two turns to select the newcomers after or previous to the start of a new season. Draft number references to the order they were elected, meaning a lower number in this list a better prospect or promise for the future of the league. For these categorical variables is that the features were encoded, whereas the continuous variables that were normalized to neutralize the feature imbalance.

After the feature engineering for the data processing in the LR model, the performance of the algorithm was of 0.90 for the best cross-validation score and for the test set. The sensitivity performance of the model situated in 0.97 for the true negatives or "0" category instances in (not) selected players, and the sensitivity for positive instances or "1" was situated in 0.53. Respectively, the weighted average of this ML model was of of 0.90 and the macro average a 0.75, in the measurements of our interest (sensitivity and specificity). The best hyper parameters found for the LR model after performing the grid search were "C" = 10, "solver" = "newton-cg".

The performance of the Logistic Regression model in the train and validation sets and a confusion matrix of its results were calculated. Every confusion matrix in the research were represented using a color blind tolerant plot display.

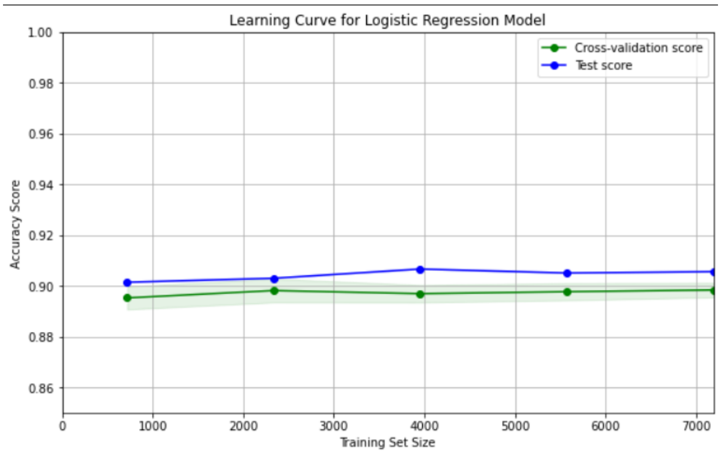


Figure 3: Logistic regression validation and test sets performance

Secondly, another two models were conducted to discuss the model performance in between different models. On the one hand, a Decision Tree Classifier, imported from the ‘sklearn.tree’ package in Python which provided an accuracy of 0.89 for the test set, with a sensitivity of 0.93 for the negative or "0" instances and a 0.64 for the positives or "1" instances. The macro average for the model was 0.77, and the weighted average for the DT was 0.89. The best hyper parameters found after performing the grid search for the DT model were "criterion" = "gini", "max depth" = 10, "min samples leaf" = 1, "min samples split" = 2.

Hereby figure showed the results of the different model sets as well as another image with a confusion matrix of its results.

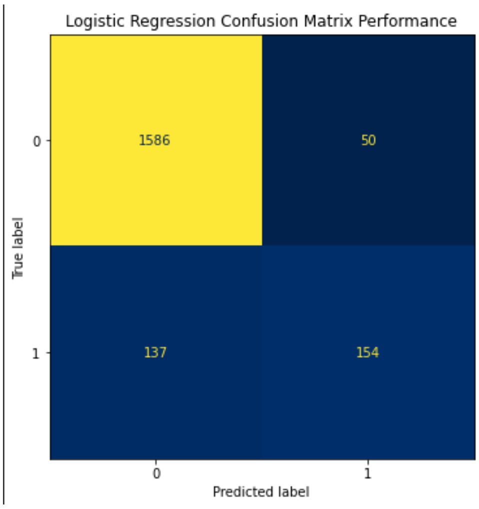


Figure 4: Confusion matrix performance for the logistic regression model

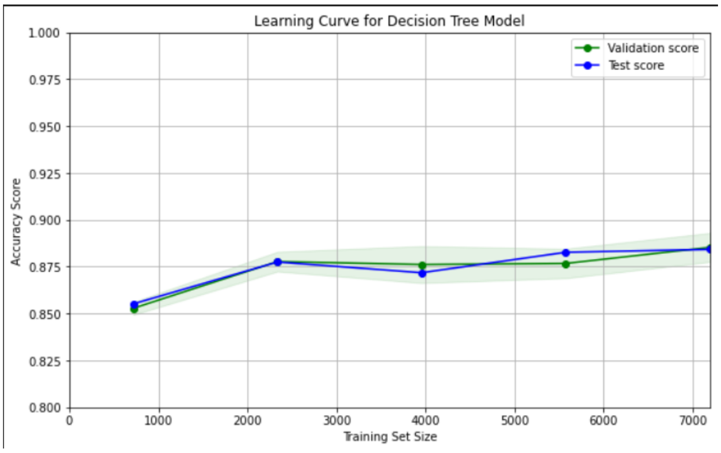


Figure 5: Decision tree validation and test sets

Last but not least, the random forest classifier was ran. It was implemented by the "sklearn.ensemble" package which enables the function Random Forest Classifier, and "sklearn.model selection" to import a grid search and a k-fold (k = 10) for the validation set. The two tree-based models are such because of the nature of its calculation, a RF performs what a DT calculates multiple times, reducing the erroneous instances for a more generalized and balanced model.

After the feature engineering, the grid search and a k-fold (k = 10) for the best model performance was found in the random forest classifier. The accuracy of the RF presents an average of 0.94 in its predictions, and a performance of 0.98 for the negative or "0" class instances, and of 0.71 for the positive or "1" class instances. Therefore, the weighted average

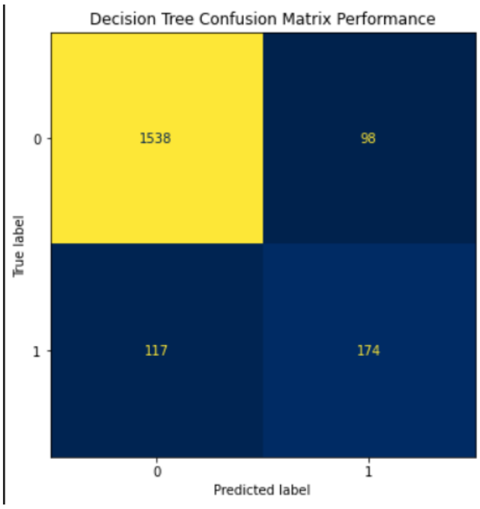


Figure 6: Decision tree confusion matrix

and macro average presented its best performances with a 0.94 and a 0.83 respectively. The best hyper parameters found by the grid search for the RF model were "max depth" = 20, "max features" = "auto", "min samples leaf" = 1, "min samples split" = 2, "n estimators" = 300.

Hereby, the learning graph for the validation and test sets and a confusion matrix for the results of the random forest classifier.

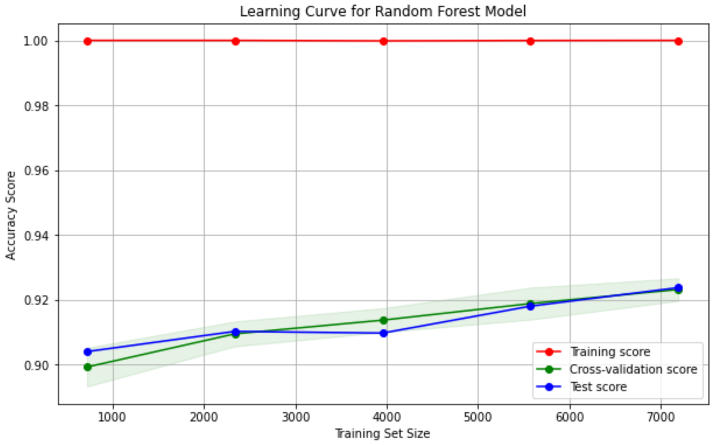


Figure 7: Random forest validation and test sets performance

After the mining and concatenating the two data sets, cleaning all its instances, taking features like country and names to reduce bias, implementing the output "YES/NO" column, and implementing the three different ML models using a grid searches for the validation sets to get best tests accuracy, the best performing model for the study using NBA

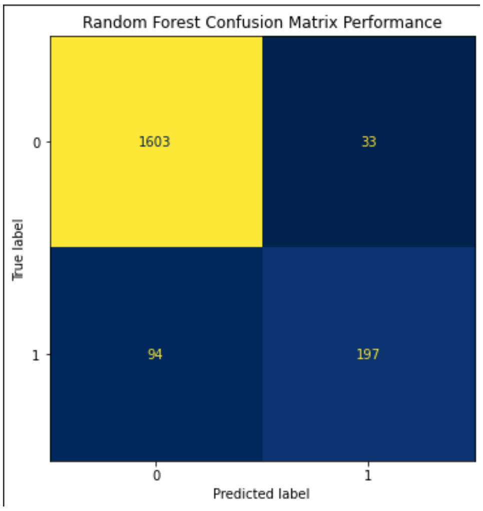


Figure 8: Random forest confusion matrix

players data from 1996 to 2021 All-star players was shown by the random forest classifier.

It is relevant to point out that the data set used was intended to avoid bias while trying to execute the predictions. Applying less categories related to features like race and direct preference like name, to manifest a more conscious decision making from the stakeholders.

The statistical test used was the chi-square, since the categories were considered a variable that may been having more weight in the selection process than it ma actually matter. For this, three categories were remaining in the data set, "draft round", "draft number", and "college".

The statistical test were conducted over the dataset with the true positive values in the random forest classifier model. This model was the best performing model, so selected to continue with the experiment (accuracy = 0.93), and the highest flaw in the calculations were found in the true positive instances with 0.68.

The chi-square tests were performed in the experiment, performed the following values. First, a chi-square statistical test and in accordance to its p-value result a Bonferroni correction was applied to reduce the chances for type I error. The category draft round got a non-significant p-value of 0.404 and a 1.213 after Bonferroni correction. Draft number got a significant p-value of 0.015 and a 0.044 after using Bonferroni correction. The category college got a significant p-value of 0.001 and a 0.004 after applying the Bonferroni correction.

Given these results, there is a significance in both "draft number" and "college" after testing statistically and correcting by Bonferroni with the selection of players and these two categories. The variable "college" has a more clear result respecting the model decision of yes or no, therefore it has more impact as well as sociological relevance.

From another angle, and related to what was previously discussed in the literature review, different players may have a different career success, influenced by different stakeholders and are expected by the model to be selected, conditioned by the college they attended in the early stages of their basketball careers. By this, it is interesting to formulate another question weather this inequality of college success for players may be a policy to be considered in the future for less inequality in the opportunity of outcome, or it is a gap that may be increased if not questioned by policy makers in this sports field.

Considering other research papers related to similar aspects suggested that tuition fees, application number and rejection ratio may all increase together when a college performs better for college football, suggesting that success may increase this gap rather than decrease it (Stinson & Howard, 2007).

In figure 1 there is a feature importance graphic of the best performing model (random forest classifier) with the selected features for the experiment.

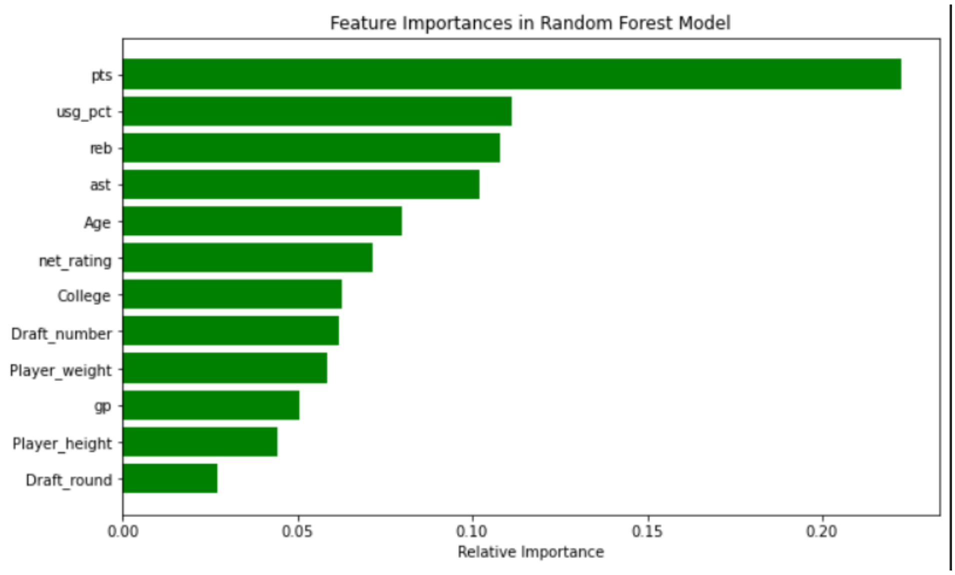


Figure 9: Feature Importance for Random Forest Model

In the following table it is possible to visualize for the three different models proposed for the third research questions as well as answering the first question if a machine learning model could predict such findings.

The table shows the most relevant measurements used in this research which were accuracy, true positives and true negatives percentages and the macro average.

The random forest classifier over performs the rest of the models, considering the most relevant measure to answer research question three, based on its accuracy. The model that showed the best results after the RF was the logistic regression and then followed by the decision tree. In Table 1 a comparison of the results obtained in the experiment.

Table 1: Model performance comparison

Model	Accuracy	True Positives	True Negatives	Macro Average
Logistic Regression	0.90	0.53	0.90	0.75
Decision Tree	0.89	0.60	0.94	0.77
Random Forest	0.94	0.71	0.98	0.83

As for the second research question, the following table showed the results of the p-values in the chi-square tests as well as the Bonferroni corrections for error type I. The variable college was identified as the only one with a significant p-value score for the three different categories analyzed in the true positive instance data set for the best performing model, the random forest classifier (from research question three). Hereby in Table 2 the results.

Table 2: Chi-square tests results for categories in true positives of the random forest model

Category	P-value	Bonferroni correction
Draft round	0.404	1.213
Draft number	0.015	0.044
College	0.001	0.004

6 DISCUSSION

The second research question of this Master thesis was to identify if there was any bias for the selection of players, or any relevant insight in order to help sport management teams, sport data driven departments, mainstream media channels, the players,or either fans themselves. For this regard, the feature importance was relevant in the fact that the biggest weight variables were those ones strictly related with the player performance or players scoring for a season, like points, player usage percentage, rebounds and assists, representing around the 0.50 of the model calculations. Producing a model with less categories that may affect heavily on the bias like "name" and "country", the variables "college" came somehow lower in the RF split as it was at the ending of the feature importance graph.

On the other hand, the RF performed more than well for negative or "0" class instances, with a result of 0.98 for this outcome. However, the accuracy score drops when calculating the positive instances or "1" predictions to a much lower 0.71. So is that the next step of the experiment was focused on testing statistically in the positive instances to check if there were variables that have been indeed affecting this decision in players careers.

By this, the research tries to reveal this selection of players from a conscious perspective, unveiling decision making by these stakeholders like the specialized media, NBA coaches and fans together with club management who may elaborate better decisions when considering scouting.

As mentioned in the methods section and as it was practiced, the focus remained in the sensitivity (true positives) and the specificity (true negatives). Consecutively, the model performed excellent in the specificity aspect, but showed some flaws in the sensitivity prediction process.

Given the best performance, and answered the first research question of which algorithm can identify this task more accurately, is that the random forest results were analyzed into further detail, and continuously as the research has to be conducted, the results were tested statistically to explain insights from a sociological perspective.

Besides, other studies had satisfactory results (92.5 accuracy) measuring a similar problem regarding research question one, about predicting through machine learning, it is interesting to consider how this results may continue to be improved due to some research limitations in the master thesis.

These limitations could be to work on a bigger data set for future researches, as well as consider other relevant topics like "personality", "popularity", "interactions with the media", using different measures for it like social media platform followers and amount of press interviews per season.

In the literature review section of this academic research there were different bias regarding sports in different disciplines of study, socioeconomic, racial, and how college sports may affect rating improvements in different universities from the United States.

The after math of these thesis results were that a player who went to a certain college has a better shot on becoming very successful than the others, but for future researchers the question to analyze could be if it is a problem that is continuously feed by the same issue or not. Is because some players that come from certain institutions receive more money from other related stakeholders, or should a mechanism to neutralize this bias in success of youngster players be conducted to equalize the opportunity outcome in the future of athletes. It is indeed an institutional approach.

As well, another interesting remark found by another researcher was that in the league of study, the NBA, managers with a certain ethnicity are prone to consider better certain player of an ethnicity alike their current coach himself, shown with a bigger correlation at the moment of the team assembling previous to the game, and in lower measurement while taking decisions in the process of the game like performing player changes during the game from the technical directors.

Agreeing with the last statement and with other ones such as the improvement of college requests and funding in different universities due to winning streaks in American football, is that also question if there is a societal flaw in the conscious sense that sports in universities and university budget imbalance may be study from a more fundamental approach and if that is making society take action in the outcomes of this professionals, either in sports or other domains.

The experiment threw quantitative evidence in the p-value after correcting for Bonferroni of 0.004 of the chi-square test of the true positive results (0.71) in the random forest model (0.94 accuracy) after controlling for the college category in the dataset. The most important remark of the academic research, besides proving a point for this specific discipline or even sports as a whole, is that the same methodology may be extended for other processes in society decision making such as buying a house, receiving a loan, getting hired for certain jobs, or getting accepted by a certain university.

The variable college was shown to be the category to present the most significant p-value after correcting for Bonferroni 0.004 compared to the 0.044 for draft number. This is logic, since the order of the draft is related to the college origin as it is keen to influence the success of players too, therefore they are correlated (Hildebrand et al., 1977).

The chi square test takes different categories and compares in between groups to see if there is a significant effect, p-value of under 0.05, in the effect of a certain group on between categories compared to the rest in the results of the studies. If the p-value is bigger than 0.05 means that there is no significant effect also understood like the outcomes for those variables are just randomly given in a specific way, but if the p-value is inferior to 0.05 the interpretation changes to a significant effect meaning that it is not randomizing the result of the experiment respect to a certain categorical variable.

The chi square statistical test was indeed conducted for these three categories, obtaining the following p-values. After controlling for the category "draft round" the p-value equaled 0.234 hence non-significant. After controlling for category "draft number" the p-value equaled 0.109, so non-significant as well. Whereas for the category "college" the p-value equaled 0.011 meaning there is statistical significance with the experiment of the performance of positive instances for a random forest, and this variable "college".

Realising this, the feature college is receiving a weight such that can decide either in the selection or non selection of these players. As an institutional approach and more inside academic research another study could be conducted explaining how this universities receive more sports success, and controlling for different variables such as geography, students features, funding of the institutions.

This meaning that college history from athletes might be receiving a bigger relevance than it actually could have for such a group, if we considered a bias exists. If the decision making can be studied outside the biased, it may lead to an information gain to develop other plans.

This can be correlated to previous literature review of socioeconomic background, or even certain bias or statistical sensitive affinity from the ones who make the decisions. Instead of, in this case, the responsible being the manager, for this All-Star selection the cause would be fan ballot, players and the selected media specialists who all together choose the NBA All-Star players.

As an aftermath, it can be interpreted that in the world of NBA stars, the college history of players it is influencing the perspective that, coming from a certain institution as a former athlete might incline the balance to the All-Star path or the non-All-Star selection path.

For further investigations it would also be interesting to find out, to what extent are competition college affecting career performances in long term, and also to what extent career opportunities, as in a personal view the career prospect of this individuals from different colleges might have an overrated weight, as the level of competition in National Competition American Association (college level sports) is already considered as professional competitiveness, only having aside the economic aspect of world sports retribution (Sandy et al., 2004).

7 CONCLUSION

The experiment conducted showed a bias in the selection of players due to a college history of students previous to their professional career. Due to the numbers threw in the analysis it is quantitatively confident to assure that a bias in the decision making due to the variable college exists in the selection. For other studies, and relating to the socioeconomic backgrounds of students in college, it may be interesting for future researchers to investigate to what extent this difference in colleges athletes that may lead to more successfully careers was already settled or influenced by other specifics in athletes that may shorten their careers outcome.

From a different perspective, in other mainstreams sports such as football, many players become super stars coming from humble backgrounds, meaning that youth background and beginning in these disciplines may not have a big participation in what the player will develop to be transformed into, and from society prospect this is very contributing to the discussion but might be more insightful if conducted by other studies into further details and analysis. According to this reasoning, surpassing difficulties may also relate to higher achievements in a professional athletes career as it may develop a more resilient behaviour.

With this been mentioned, is also relevant to consider the identification of bias also in sports controlling for different aspects of the individuals taking part. This can also be adapted to many society domains, which can improve our decision making as thanks to machine learning and statistics we can support this findings from a quantitative perspective.

The college history of athletes may affect the outcome of more successful players in regards of the NBA All-Star team selection, in a significant quantitative way which may lead as to other questions such as, do the policies for university development in different areas such as sports receive similar stimulus across different geographies or if this outcomes might be influenced by society's influence due to cultural significance in different geographies which may also affect the influence of the decision making stakeholders such as public or private institutional funding. To answer these spin off questions, crossing the results from this experiment with information regarding socioeconomic and cultural background of the different geographies and individuals that constitute these same college institutions would be needed.

Numerous studies have demonstrated the capability of Machine Learning models, with lower performing accuracy's, to predict NBA All-Star rosters, and in this research well performing predictions were possible to appreciate. Previous researchers in similar areas of study, had also revealed biases like racial and socioeconomic influences for sport making decisions, and sport success careers. This evidence supports the feasibility of exploring decision biases in sports through the field of data science.

Ultimately, investigating such biases is crucial and applicable not only to the NBA and other sports but to various societal domains as well, and can driven us to analyze different sociological circumstances in society such as the opportunities of outcome and equality of opportunities.

REFERENCES

Gandelman, N. (2009). Selection biases in sports markets. *Journal of Sports Economics*, *10*(5), 502–521.

Grossman, N. (2014). What is the nba. *Marq. Sports L. Rev.*, *25*, 101.

Guevara, J., Martín, E., & Arcas, M. (2021). Financial sustainability and earnings management in the spanish sports federations: A multi-theoretical approach. sustainability 2021, 13, 2099.

Hildebrand, D. K., Laing, J. D., & Rosenthal, H. (1977). *Analysis of ordinal data*. Sage.

Hussey, A. (2021). *Nba draft analysis* [Doctoral dissertation, Dublin, National College of Ireland].

Langel, M., & Tillé, Y. (2011). Statistical inference for the quintile share ratio. *Journal of Statistical Planning and Inference*, *141*(8), 2976–2985.

Levine, G. R. (2019). *All-nba team voting patterns: Using classification models to identify how and why players are nominated* [Doctoral dissertation, Ohio University].

Ringer, T. (n.d.). *Lockout seasons 1999* [citation of source the ringer].

Sandy, R., Sloane, P. J., Rosentraub, M. S., Sandy, R., Sloane, P. J., & Rosentraub, M. S. (2004). College sports in the usa and the role of the ncaa. *The Economics of Sport: An International Perspective*, 257–284.

Sanner, M. F., et al. (1999). Python: A programming language for software integration and development. *J Mol Graph Model*, *17*(1), 57–61.

Schroffel, J. L., & Magee, C. S. (2012). Own-race bias among nba coaches. *Journal of Sports Economics*, *13*(2), 130–151.

Shekar, B., & Dagnew, G. (2019). Grid search-based hyperparameter tuning and classification of microarray cancer data. *2019 second international conference on advanced computational and communication paradigms (ICACCP)*, 1–8.

Soliman, G., Misbah, A., Eldawlatly, S., et al. (2017). Predicting all star player in the national basketball association using random forest. *2017 Intelligent Systems Conference (IntelliSys)*, 706–713.

Solutions, S. (2016). What is logistic regression. *Retrieved from*.

Stinson, J. L., & Howard, D. R. (2007). Athletic success and private giving to athletic and academic programs at ncaa institutions. *Journal of Sport Management*, *21*(2), 235–264.

Suthaharan, S., & Suthaharan, S. (2016). Decision tree learning. *Machine Learning Models and Algorithms for Big Data Classification: Thinking with Examples for Effective Learning*, 237–269.

Tallarida, R. J., Murray, R. B., Tallarida, R. J., & Murray, R. B. (1987). Chi-square test. *Manual of pharmacologic calculations: with computer programs*, 140–142.

Vabalas, A., Gowen, E., Poliakoff, E., & Casson, A. J. (2019). Machine learning algorithm validation with a limited sample size. *PloS one*, *14*(11), e0224365.

Wang, J., & Fan, Q. (2021). Application of machine learning on nba data sets. *Journal of Physics: Conference Series*, *1802*(3), 032036.

Zhang, L. (2017). A fair game? racial bias and repeated interaction between nba coaches and players. *Administrative Science Quarterly*, *62*(4), 603–625.