PREDICTING NBA MOST VALUABLE PLAYER FOR THE 2021-22 SEASON USING REGRESSION MODELS

by

Stephen Nyarko

Thesis submitted to University of Plymouth
in partial fulfilment of the requirements for the degree of

MSc Data Science and Business Analytics

University of Plymouth
Faculty of Science & Engineering

September 2022

Copyright statement

Abstract

Stephen Nyarko

Predicting NBA Most Valuable Player for the 2021-22 Season
using Regression Models
(Under the direction of Dr. Davide Vadacchino)

This thesis aims to predict the NBA most valuable player for the 2021-22 season using regression models. Four regression models were used; Linear, LASSO, Random Forest, K-Nearest Neighbour and one artificial neural network model; Multi-Layer Perceptron. Three of the regression models (Linear, LASSO AND Random Forest) made good predictions of the top five MVP candidates, correctly predicting four out of five candidates. However, the ranking order was not completely correct for any of these models. Only the random forest model was able to correctly predict the actual MVP winner for the season (Nikola Jokic). The multi-layer perceptron failed entirely to predict any of the top five NBA candidates. This may have been caused by failing to tune the hyperparameters correctly. When tested across several years, Random Forest and K-Nearest Neighbour saw significant decreases in the accuracy of the model. Linear and LASSO regression saw almost negligible fall in accuracy, demonstrating that they are more robust models when providing MVP predictions over several years.

Acknowledgements

I would like to thank the University of Plymouth for the experience I have been provided and the knowledge they have bestowed on me which has allowed me to capably complete this thesis.

I’d like to thank Dr. Davide Vadacchino for his guidance and understanding whilst I completed this thesis.

I’d like to thank my parents for their financial support and love.

Finally, and most importantly, I’d like to thank God for giving me the strength to endure this process.

Table of Contents

List of Tables

List of Figures

List of Abbreviations

2P	2-Point Field Goals Per Game
2P%	2-Point Field Goal Percentage
2PA	2-Point Field Goal Attempts Per Game
3P	3-Point Field Goals Per Game
3P%	3-Point Field Goal Percentage
3PA	3-Point Field Goal Attempts Per Game
AST	Assists Per Game
BAA	Basketball Association of America
BLK	Blocks Per Game
DRB	Defensive Rebounds Per Game
eFG%	Effective Field Goal Percentage
FG	Field Goals Per Game
FG%	Field Goal Percentage
FGA	Field Goal Attempts Per Game
FT	Free Throws Per Game
FT%	Free Throw Percentage
FTA	Free Throw Attempts Per Game
K-NN	K-Nearest Neighbour
L	Losses
LASSO	Least Absolute Shrinkage and Selection Operator

MLP	Multi-Layer Perceptron
MP	Minutes Played Per Game
MSE	Mean Squared Error
MVP	Most Valuable Player
NBA	National Basketball Association
NBL	National Basketball League
ORB	Offensive Rebounds Per Game
PER	Player Efficiency Rating
PTS	Points Per Game
RF	Random Forest
SRS	Simple Rating System
STL	Steals Per Game
Tm	Team
TOV	Turnovers Per Game
TRB	Total Rebounds Per Game
VORP	Value over Replacement Player
W	Wins
WS / 48	Win Shares Per 48 Minutes
WS	Win Shares

1. Introduction

1.1 Background

The National Basketball Association was founded on 6th June 1946 at the Commodore Hotel in New York. Known at the time as the Basketball Association of America (BAA). It was not until August 3rd, 1949, when the Basketball Association of America (BAA) merged with the National Basketball League (NBL) was the name changed to National Basketball Association (NBA). The NBA consists of 30 teams with all except one team (Toronto Raptors) located in the United States (NBA, 2022a).

The NBA was a popular sport in the United States, but it was the formulation of the dream team in 1992 to compete in the Olympic games that catapulted it to the global sport it is now. Previously they would send collegiate athletes to compete in the games. That year basketball was one of the Olympics' most viewed sports and this had an enormous impact on the NBA's popularity worldwide (Jones, 2016). NBA (2014a) suggests several of the European players in the NBA watch and were inspired by the 1992 game.

Unsurprisingly this growth in popularity leads to a more diverse group of players within the NBA. NBA (2014b) found that there were 101 international players from 37 different countries playing for NBA teams and currently this number is even greater. In fact, the three frontrunners for this year's most valuable player award are all foreign players: Nikola Jokic from Serbia, Joel Embiid from Cameroon and Giannis Antetokounmpo from Greece (NBA, 2022b).

This thesis will endeavour to predict the NBA's most valuable player using different machine learning techniques. Traditional statistics will be utilised as the predictors as these are the statistics widely understood by the average NBA fan and identify the accuracy of these statistics in predicting the MVP when compared to advanced statistics.

1.2 Analytics in Sports

Sarlis and Tjortjis (2020) define sports analytics as the collection and management of data which can be used to create predictive models and computational methods which illuminate valuable information for sport-related decision-making. There are several implementations of sports analytics, Minusha (2016) indicates that sports analytics can be used to evaluate players in order to determine the best game strategy, help sports associations develop rankings of players and teams, assess present rules and study the viability of presenting new rules as well as allowing health professionals in sports use statistical methods to understand players' physical and mental conditions. In today's NBA there is a data analyst on the staff of every team with the job of assisting coaches in maximising the talent of their players and helping scouts recognise undervalued players.

In the early years of the NBA, analyst kept track of basic player statistics, such as points, rebounds, and assist. These basic statistics often called traditional statistics were the bases for evaluating a player's game. In 2010 the SportVU optical player and ball tracking system were first deployed in select NBA arenas by organisations aiming to obtain an advantage in player and team analysis (Patton et al., 2021). The NBA then adopted SportVU league-wide prior to the 2013-14 season. Since then, nearly all analysis and decision-making for NBA teams have been data-driven, utilising both raw positional data and tactical insights derived from the markings detected automatically by machine learning algorithms (Patton et al., 2021). However, there has not only been technological advancement but also advancements in statistics. Oliver (2004) popularised the idea of per possession stats proving that player and team possession stats were better predictors of team and player performance in basketball. Traditional statistics are still tracked today, but analysts also measure more advanced statistics that provide more accurate assessments of players. One such statistic is the player efficiency rating created by Hollinger (2011) which accounts for all the positives a player contributes: field goals, free throws, 3-pointers, assists, rebounds, blocks and steals, and negative ones such as missed shots, turnovers, and personal fouls.

1.3 Most Valuable Player

The National Basketball league is often considered the premier basketball league hosting a multitude of the most talented basketball players to have played the game both in the past and the present. The Most Valuable Player award (MVP) is an accolade given to the player who is judged by a voting panel to have been the best player in the league for that season, the first award was presented at the end of the 1955-56 season. Initially, the voting panel consisted of an NBA player who was part of the organisation during that season. However, following the 1980-81 season, the award has been determined by a panel of sportswriters and broadcasters throughout the United States and Canada. The voters rank five NBA players of their choice, each rank has a corresponding point. First place is worth ten points, second worth seven, third worth five, fourth worth three and fifth worth a single point (Corvo, 2021). This was a slight alteration from the system used when players decided on the MVP, voters ranked three players with respective point totals of five, three, and one. The player with the most cumulative points was declared the MVP.

The MVP award is often considered when discussing the greatest players to ever play the game. Kareem Abdul-Jabbar has received this accolade most times amassing six of them throughout his playing career followed by Michael Jordan and Bill Russell who won five. Lebron James and Wilt Chamberlain have accumulated four, and Moses Malone, Larry Bird and Magic Johnson won three. All these players are considered all-time greats, and many are within the conversation for the greatest of all time often abbreviated as GOAT. Only eight MVPs have failed to win a championship; Karl Malone, Allen Iverson, Steve Nash, Charles Barkley, Russell Westbrook, James Harden, Derrick Rose, and Nikola Jokic. However, the last four players named active players in the NBA and still have the prospect to win a championship. This means 76% of MVP winners have also won at least one championship.

1.4 Literature Review

The popularization of sports analytics is detailed in Lewis's (2003) book titled Money Ball. The book follows the use of sabermetrics by the then general manager of the Oakland Athletics Billy Beane to acquire players who were undervalued and underappreciated and fit the statistical criteria that he believed would lead to success for the franchise and winning the world series. This was the initial platform where statisticians worked with individuals and team performance data (box score) as a means of gaining a competitive advantage (Morgulev et al., 2018). The book exposed the use of analytics to identify and draft players who excelled at getting on base deviating from traditional metrics which focused on stolen bases or runs batted in. Tichy (2016) suggests that this gave them the ability to draft players who were otherwise overlooked by other teams using traditional measures. This set a new standard that other baseball teams adopted, but not only in baseball as the use of data analytics has expanded across professional sports as a whole.

There is a plethora of literature which attempts to use machine learning to predict the different outcomes within basketball. The most common project is often the use of NBA statistics in an attempt to predict the outcomes of individual matches. Torres (2013) endeavoured to predict using linear regression, a maximum likelihood classifier, and a multi-layer perceptron (MLP) the results of future NBA fixtures. He compared the results of these three methods with the naive majority vote classifier. In this method, the team with the greater win percentage from prior games within the season was chosen as the potential winner for the upcoming game. For the linear regression eight features were applied:

1. Win-Loss Percentage (Visitor Team)
2. Win-Loss Percentage (Home Team)
3. Point differential per game (Visitor Team)
4. Point differential per game (Home Team)
5. Win-loss percentage for previous eight games (Visitor Team)
6. Win-loss percentage for previous eight games (Home Team)

7. Visitor Team Win-Loss percentage (Visitor Team)
8. Home Team Win-Loss percentage (Home Team)

Torres (2013) assumed the features were highly correlated and so to eliminate multicollinearity the Principal Component Analysis was performed with Torres selecting only the three largest eigenvalues. For the maximum likelihood classifier, the best combination achieved included all the features except one and six. Finally, the MLP uses four different combinations of hidden layer and neuron and the combination with the greatest accuracy was selected. Table 1 below displays the results of each method. The MLP yielded the greatest mean accuracy at 68.41%. The MLP is the neural network

Season	**Linear Regression**	**Maximum Likelihood Classifier**	**MLP**
2007	0.6932	0.6587	0.6909
2008	0.6932	0.6888	0.6909
2009	0.6789	0.6441	0.6848
2010	0.6942	0.6789	0.6964
2011	0.6541	0.6039	0.6848
2012	0.6409	0.6776	0.6801
Mean	0.6789	0.6681	0.6841

Table 1: Results of Greatest Mean Accuracy.

There are several works which attempt to predict the NBA most valuable player using different techniques. Chen et al. (2019) used neural networks to create an MVP forecasting system. They chose the neural network model as it performs best when using large amounts of data. They collected three sets of data one containing player totals data, advanced statistics and a hybrid dataset compiled from the totals and advanced datasets. There were 16 variables within each dataset, with three shared statics between each dataset: players, games played, and minutes played. Each dataset was separated into three kinds of sets by individual season, test set, validation set and training set. They trained their model using a mini-batch gradient descent algorithm. The validation set was used to adjust the hyperparameter and the test set was used to measure the accuracy of the model prediction. After training three different models, they were compared, and it was found that all effectively forecasted the most valued player in the NBA 2009-10 season and 2016-17 season. They concluded that the mixed dataset has the greatest possibility and performed best.

Chen (2017) created four different predictive models to predict NBA MVP winners using JMP statistics software Z transformation, Descriptive statistics, Power Transformation, Best Subset, and Data Mining Discriminant analysis to create. The ‘power=3’ had the greatest prediction capability. This indicated that team performance was extremely important in the MVP selection process. When the team factor was removed from consideration, the Data Mining Discriminant Model outperformed the Uniform Model and Weighted Model. Chen (2017) concluded that the ‘power=3’ model has the potential to make the MVP selection process more objective and that the current method of voting is too subjective. Furthermore, The MVP selection should weigh more on team performance over individual achievement. Table 2 displays the accuracies of each model.

Algorithm	**Accuracy**
Uniform Model	47%
Weighted Model	52%
Power=3 Model	69%
Discriminant Model	55%

Table 2: Accuracy Percentage of each Model.

The literature when trying to predict NBA's most valuable player winner tends to use a neural network. For this thesis, regression models will be used as the primary means of attempting to predict the MVP. Neural network techniques will be used, and their results compared to the regression techniques. Furthermore, traditional statistics will be employed as literature tends to use more advanced statistics which are generally better at contextualising the game of basketball. The MVP voting is largely in the hands of media and sports journalists who are much more knowledgeable than the average sports fan. However, most sports fans would use traditional statistics to determine their choice of most valuable. The thesis will investigate if the average fans can arrive at similar results to those of the current voters.

Freire's (2021) work resembles this thesis. He attempts to predict the NBA MVP for the 2020-21 season using data collected from 'Dribble Analytics' consisting of the top ten players in MVP voting of each season, from 1979–80 to 2019–20. He then collected MVP tracking data for the 2020-21 season from 'www.basketball-reference.com'. The variables in the data set consisted of a mixture of traditional statistics, advanced statistics, team statistics and MVP voter share data.

Freire (2021) used three machine learning models random forest, K-NN and MLP. Table 3 displays the mean squared error (MSE) and r-squared values achieved with each technique. K-NN was the best performing model having both the smallest mean squared error as well as the greatest r-squared value. All three models predicted Jokic as the eventual winner of the award and this was correct. For this thesis, all three models are applied and discussed along with two extra models. However, it was decided to use average precision as the error metric it seemed better at displaying the predictive capabilities of each model especially as the placement of MVP candidates were most important in this work.

Model	**MSE**	**R-squared**
MLP	0.027	0.681
K-NN	0.023	0.726
Random Forest	0.034	0.605

Table 3: Mean Squared Error and R-Squared.

McCorey (2021) set out to create an algorithm that could be used during every season to accurately predict the MVP. Seven underlying models were formulated, and three separate techniques (linear regression, K-NN and ANN) were used to determine the optimal model to forecast the Most Valuable Player. The primary analysis discovered the linear regression model was the optimal model and four of the seven models formulated were effective dependent on the technique used. Three linear regression techniques Simple Average, Ordinary Least Squares, and Constrained Least Squares along with the four effective underlying models were used to find the best combination forecast to track the players who are most deserving of the MVP accolade. Simple Average had the greatest accuracy, however, the results showed that all three methods ranked the players correctly.

The literature focused on this topic is vast, the few that is mentioned here have endowed the research with greater knowledge around the topic of discussion and assisted in the formulation of this thesis

2. Theoretical Background

2.1 Linear Regression

Schroeder et al. (2017) explain that a linear regression analysis, analyses cases in which the relationship between the dependent variable and the independent variable can be summarized by a straight line. Multiple Linear Regression is a method for estimating the effects of several factors $x_1, x_2, \dots x_n$ (for this thesis' case this would be the predictor variables) on the dependent variable y (share). The model will have the form of Eq. 1.

$$y_i = \beta_1 x_1, \beta_2 x_2, \beta_3 x_3, \dots \beta_i x_i + \varepsilon_i$$

Eq. 1

where ε_i is the error term and β denotes the unknown parameters. In our model, there will be a lot of explanatory variables. A system of model simplification will create a model where only the variables that are statistically significant in determining the win/loss percentage remain. This will make it less likely to overfit the model. The three main techniques often used are backward elimination, forward elimination, and stepwise elimination.

Linear regressions can be used to determine the statistical significance of the independent variables on the dependent variable. They can also be utilised for the prediction of trends and future values for interpretation and estimations. Linear regressions work best with datasets that contain little noise. It is also necessary to remove highly correlated features to reduce collinearity as much as possible and reduce the likelihood of overfitting. For the model to be reliable the variables must be normally distributed so that the data points will tend to cluster around the mean. Finally, in linear regressions, the independent variables are often standardised as a means to create a common scale for the variables without distorting the differences in the range of the values.

These types of regression are easy to implement, and overfitting can be mitigated using regularisation techniques. However, they are susceptible to underfitting and sensitive to outliers. This is the first regression technique is used to perform the analysis.

2.2 LASSO Regression

LASSO (Least Absolute Selection Shrinkage Operator) regression. This is a penalised regression model. The LASSO imposes a constraint on the sum of the absolute values of the model parameters where the sum has a specified constant as an upper bound. This constraint causes regression coefficients for some variables to shrink toward zero. This is the shrinkage process. The shrinkage process allows for better interpretations of the model and identifies the variables most strongly associated with the target or response variable. That is the variable selection process. The goal is to obtain the subset of predictors that minimises prediction error.

Regression is useful because shrinking the regression coefficient can reduce variance without a substantial increase in bias. Second, LASSO regression can increase model interpretability oftentimes at least some of the explanatory variables in an OLS multiple regression analysis are not associated with the response variable. As a result, it often ends up with a model that is overfitted and more difficult to interpret. The regression coefficients for unimportant variables are reduced to zero, which effectively removes them from the model and produces a simpler model that selects only the most important predictors.

In LASSO regression a tuning parameter called λ is applied to the regression model to control the strength of the penalty. As lambda increases more, coefficients are reduced to zero. That is fewer predictors are selected and there is more shrinkage of the non-zero coefficients with LASSO regression. When lambda is equal to zero, there is an OLS regression analysis. Bias increases and variance decreases as lambda increases. This technique was used to assess whether better a predictive model could be created through regularisation.

2.3 K-Nearest Neighbour (K-NN)

K-nearest neighbour (K-NN) is a simple, established, fundamental classification method. It is a supervised machine learning algorithm that predicts the classification of unspecified data by using the characteristics of the training data. K-NN classifies datasets through training models using the k nearest training data points (e.g., neighbours) and then performs a majority voting rule to settle the classification (Uddin et al., 2022).

K-NN was developed to evaluate the uncertainties in datasets and define a reliable classification. This method is customizable and eliminates challenges such as different data sizes, quantities of data, irrelevance, range of data and context (Uddin et al., 2022).

K-NN is a commonly used method due to its implementation simplicity and debugging. Unique noise reduction techniques are available for this method to improve the accuracy of the classifier. However, the performance of the algorithm in run-time can be reduced due to large training sets. The method is also sensitive to redundant features due to the similarities in datasets and hence classification (Cunningham and Delany, 2022).

K-NN method is achieved in two steps. The initial step locates the nearest neighbours, and the second step determines the class using the neighbours (Cunningham and Delany, 2022). This classification is used by assigning arbitrary variables into classes with the most similar data models. Characteristics are initially accumulated for training and test datasets. The quantity of predictors is optional and can incorporate any quantity of characteristics (Zhang, 2016).

One of the main methods used is via Euclidean distance. This method calculates the distance between characteristics by Eq. 2 below:

$$D_{(p,q)} = \sqrt{(p_1 - q_1)^2 + (p_2 - q_2)^2 + \cdots + (p_n - q_n)^2} \qquad \text{Eq. 2}$$

where p and q are the NBA players to be compared with n characteristics.

Another method commonly used is the k parameter. This parameter decides the number of neighbours used for the K-NN algorithm. The k-number is key as it will influence the performance of the algorithm. A large k-number will reduce the effect of variance due to error but will hold a possibility of disregarding trivial yet crucial patterns (Zhang, 2016). K-NN was chosen as another technique to conduct analysis for this thesis.

2.4 Multi-Layer Perceptron (MLP)

An example of an artificial neural network is the Multi-Layer Perceptron (MLP) architecture. This architecture is comprised of neurons and connections to make up an input layer, one or more hidden layers and an output layer. The MLP classification method is contingent on the initial values of the input layer that include weights and biases. Neurons would then calculate the sum of the input weights and apply an activation function to communicate with the next neuron (Deyasi et al., 2021; Castro et al., 2017; Wang et al., 2006).

MLP is used due to its better mapping with functions, great accuracy, and controlled output. However, a smaller quantity of nodes is safer used to avoid complex computations (Deyasi et al., 2021).

Castro et al. (2017) mentioned in their journal article two major struggles with MLP. Time inefficiency for training large datasets and overfitting the training data by having an abundance of connections or underfitting the training data by having a deficiency of connections, thus preventing the network from attaining the desired result.

This neural network structure demonstrates intricate non-linear interactions between the input layer neurons and other layered connections (Bikku, 2020). MLP is done by entering the weight and bias of a set number of attributes into the input layer, choosing the number of hidden layers and neurons in each hidden layer, the class of the output layer and the learning rate (Deyasi et al., 2021).

A general diagnostic classification method is a feed-forward back-propagation MLP. This method is trained by utilizing the input and target neurons by updating the weights and biases of the neurons until a mapping function is formed with its output neurons. The generation of further mapping functions enables an idealistic classifier for diagnostic classification (Wang et al., 2006).

In this work, 20 input neurons are entered into the input layer which consists of general NBA statistics. One hidden layer is applied and holds a quantity of 100 neurons. The result is outputted as a class in the output layer. This method is used due to its capability of choosing the input node attributes and selecting a suitable quantity of hidden nodes which would increase the number of connections produced and improve the training data. Figure 1 below illustrates how the diagram would look with the selected input and output attributes for the MLP architecture.

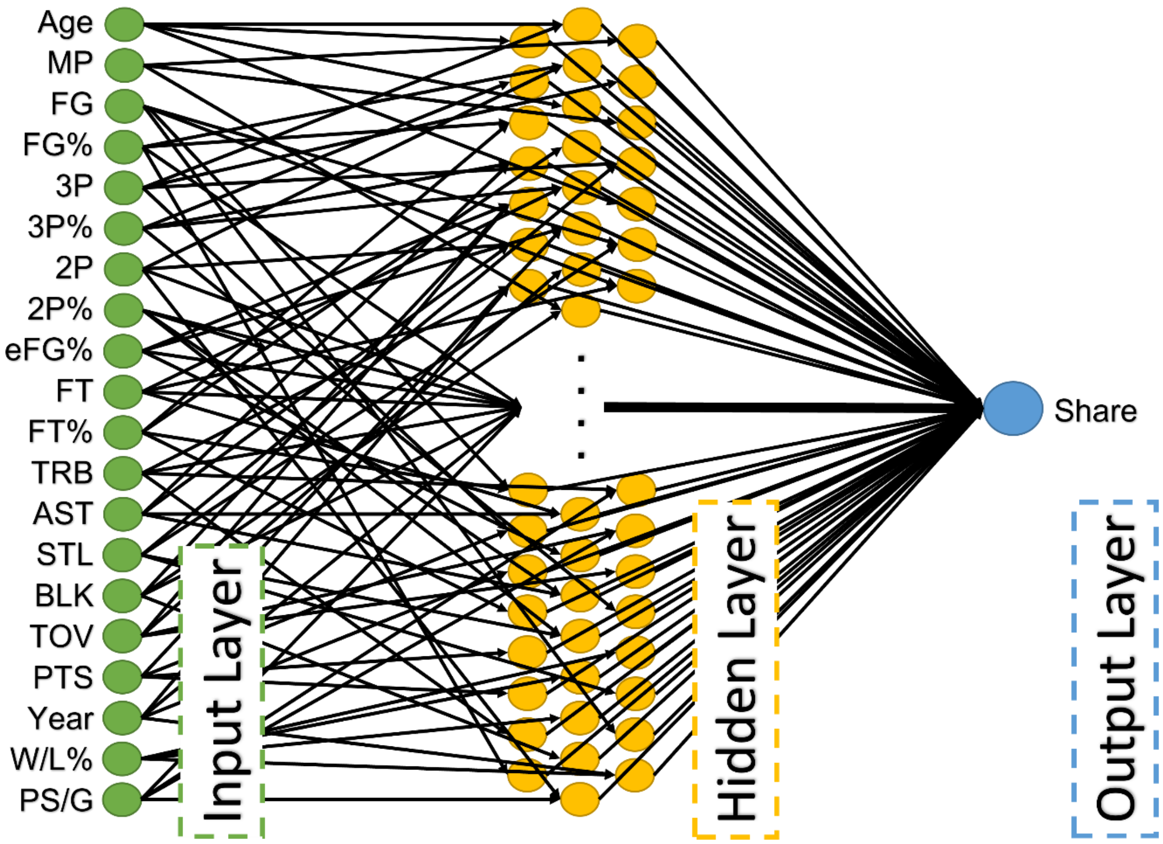


Figure 1: MLP Diagram

2.5 Random Forest (RF) Regression

Random forest (RF) consists of many accumulated decision trees. It is a meta estimator that uses averaging to enhance predictive accuracy. It controls overfitting and can decrease inconsistency compared to single decision trees (Couronné, Probst and Boulesteix, 2018; Shreyas et al., 2016).

A decision tree is built by predicting all possible answers through the options provided. This is done by dividing the dataset into subsets and therefore a data tree is developed. This tree would look like a flowchart comprising decision nodes and leaf nodes. The root node would fall into decision nodes that should normally consist of two or more branches whereas a leaf node is a numerical or categorical decision that comes after the decision node.

The decision tree approach does not allow backtracking but uses the standard deviation reduction technique from information gain. The standard deviation tells how spread a numerical dataset is from each other. The closer it is to zero, the homogenous it is, and therefore the ideal result. The variation coefficient is a formula that determines when to stop sub-setting into more data.

To calculate the standard deviation for a singular attribute and the coefficient of deviation for all attributes, an example is shown below:

Step 1: Find the mean value Average (Eq. 3) of X.

$$X = [\,4, 5, 9, 6, 12, 1\,]$$
$$n = 6$$
$$\mathrm{Avg} = \bar{\mathrm{x}} = \frac{\Sigma x}{n} = \frac{37}{6} = 6.16$$

Eq. 3

Step 2: Find the standard deviation (Eq. 4) of the variance of the Avg (Eq. 5).

$$S = \sqrt{V} \quad and \quad V = \frac{\Sigma(\mathrm{x}-\bar{\mathrm{x}})^2}{n}$$

Eq. 4 and Eq. 5

$$\Sigma(\mathrm{x_n} - \bar{\mathrm{x}})^2$$
$$= (4-6.16)^2 + (5-6.16)^2 + (9-6.16)^2 + (6-6.16)^2 + (12-6.16)^2 + (1-6.16)^2$$
$$= 4.67 + 1.35 + 8.07 + 0.03 + 34.11 + 26.63$$
$$= 74.86$$

$$V = \frac{\Sigma(\mathrm{x}-\bar{\mathrm{x}})^2}{n} = \frac{74.86}{6} = 12.48$$
$$S = \sqrt{V} = \sqrt{12.48} = 3.53$$

Step 3: Find the coefficient of deviation (Eq. 6).

$$CV = \frac{S}{\bar{x}} * 100\% = \frac{3.53}{6.16} * 100\% = 57.31\%$$

Eq. 6

Where:

S = Standard Deviation (to branch out decision trees)

CV = Coefficient of Deviation (used to stop branching out decision trees)

Avg = Average (represents the value in the leaf nodes)

n = The number of nodes or data in a dataset.

V = Variance

The standard deviation (Eq. 4) for the single attribute, 3.53 (S) is the value determining how far the numbers are from each other, which is 57.31% (CV) away from being homogenous.

To calculate the standard deviation for two attributes, known as a target and a predictor, is as below (Eq. 7):

$$S_i = 6.3 \qquad S_j = 2.4 \qquad S_k = 9.8$$

$$n_i = 8 \qquad n_j = 7 \qquad n_k = 3$$

$$n_{Total} = n_i + n_j + n_k = 8 + 7 + 3 = 18$$

$$S_{Result} = \left(\frac{n_i}{n_{Total}} * S_i\right) + \left(\frac{n_j}{n_{Total}} * S_j\right) + \left(\frac{n_k}{n_{Total}} * S_k\right)$$

$$= \left(\frac{8}{18} * 6.3\right) + \left(\frac{7}{18} * 2.4\right) + \left(\frac{3}{18} * 9.8\right)$$

$$= 2.8 + 0.93 + 1.63$$

$$= 5.36$$

Eq. 7

This method is used in the calculation of decision trees in linear regression.

RF is comparatively robust against parameter-based environments. Although it was mentioned in Couronné et al. (2018) that LR is much better than RF due to its prediction process, explanation, and requirement of only the fitted values of the regression coefficient to run, whereas RF is said to be biased in variable selection disregarding the significance for prediction.

Schonlau and Zou (2020) stated how RF are like black boxes as they do not allow any insight into how the predictions were achieved. However, he opposes Couronné's hypothesis and believes RF to be much better than LR based on linearity. He mentions how the linearity of the model allows interpretability but opposes flexibility for prediction. However, RF adjusts to non-linearities in data leading to accuracy in predictions. RF also works well with medium to large datasets, using only the required predictor variables. LR will struggle in this scenario when the parameters are more than the observations.

In both decision trees and RF algorithms, it is important to adjust the values of iterations and the number of variables (numvars) concerning the dataset. Decision trees tend to overfit, directing to lower predictive accuracy and error rate compared to RF. The decision tree model observes the habits of old test datasets thus performing inadequately on new test datasets (Schonlau and Zou, 2020).

An RF training model process is initialised by setting the data points in a random order to eliminate potential dependencies on sorting the test data observations. The dataset is then split equally into two groups for training and testing. A 50-50 split is ideal for large datasets but not small datasets due to the reduction of training data size (Schonlau and Zou, 2020).

Hyperparameters are then adjusted to find the greatest testing accuracy of the models. The main hyperparameters include the number of subtrees(iterations) and the number of variables (numvar) to randomly explore each node split (Schonlau and Zou, 2020). Other parameters that can be entered include the number of trees in the forest (ntree) and the minimum size of terminal nodes (nodesize). A large node size value is believed to generate smaller trees (Couronné, Probst and Boulesteix, 2018).

Out-Of-Bag (OOB) error and validation error can finally be calculated to verify the ideal model. OOB error is tested against the training data and validation error is tested against the testing data (Schonlau and Zou, 2020).

RF is used in this work as a regression model to output continuous outcomes of the top five player share predictions. There will be 20 independent variables obtained from the NBA statistics, which will be known as the predictors, and one dependent variable will be predicted (player shares). The iterations will be set to 100 with a node size of five to output the top five NBA players with the highest probable share value for the 2021-22 season.

3. Methodology

3.1 Data Collection

In order to retrieve data for the predictive model, it was necessary to retrieve data from ‘www.basketball-reference.com’. It was decided to scrape the data from the website although there was the option to download the data as .csv files directly from the website. However, considering the amount of that of tables that needed to be downloaded web scraping seemed like a more viable method of downloading and extracting the data. The decision was made to extract data from 1986 to 2022. This was because the 3-point line was introduced to the NBA in 1980 however it was not a vital strategy, and 3-point volume only began to see significant growth in the years following the 1986 season. Adding the years prior to this would have seen a substantial number of empty rows in the 3-point columns and this irregularity in data may have made it more difficult for the model to predict the MVP winner.

Three main web pages needed to be downloaded. The first page downloaded contained the most valuable players for each year and the share of votes received by every player who received a vote using the request library in python. The next pages downloaded were those containing all NBA per game statistics for each player in the NBA for each year. These stats are essential in determining a player’s individual achievement. In order to download the player statistics data, it was necessary to use selenium, a python library as using requests alone meant not all the data could be reached as the website uses javascript to render the rest of the rows in the table after the page has loaded. Selenium allows python to automate the browser so it can download the webpage in the browser and render it using the browser's javascript execution environment. The last pages downloaded contained team statistics primarily the wins and losses for each team for every season from 1986 to 2022. These stats are important to learn from Chen 2017 that team performance is vital in the MVP selection process.

Thereafter, it was required to extract the data in the table from each .html file. To do that, 'BeautifulSoup' was a library used in python to parse webpages. This allowed the table to be extracted from each page, concatenated, and then converted into a data frame using pandas. Each data frame was then turned into three separate .csv files using pandas 'to_csv' function. The first .csv file contained the MVP votes received between the target years, the second contained the traditional or per-game statistics of each player and the final .csv file contained the team's record for the season.

3.2 Data Cleaning

The data needed to be cleaned before it was even possible to run any machine learning algorithms or regressions on it. Firstly, the MVP .csv file was cleaned by keeping only the Player, Year, Pts Won, Pts Max, and Share columns. The advanced statistics (WS and WS/48) and player information (Age, G, MP, PTS, TRB, AST, STL, BLK, FG%, 3PT%, FT%) were removed as these statistics are already present in the player statistics file.

To clean the player statistics file, two unnecessary columns (Unnamed 0 and RK) were initially removed. In the player's column, some of the players had an asterisk next to their names indicating that they were a member of the NBA Hall of

Fame. The end goal was to merge the player's data frame with the MVPs data frame using the two mutual columns (Player and Year). This meant the asterisk had to be removed so the name would match and be merged. Whilst inspecting the data frame, some players had several rows for the same year as they had played for multiple teams in that season. For these players, their total stats for that season were kept, the other rows were removed and the last team they played in that season was assigned as the team named read TOT which was not an NBA team. Once this was done, the two data frames were merged using the player names and years. After merging the two data frames all the players who were not in the MVP data frame had missing values in the Pts Won, pts Max and Share columns. Realistically, these are not actually missing values these players just failed to receive any votes, so the missing values (interpreted as 'NaN') were replaced with zero.

The next step was to merge the team's data. After inspecting the data, it was necessary to remove the division rows. Some team names had asterisks signifying that they made the playoffs for that season, the asterisks were removed. The greatest issue before being able to merge the team data frame was that in the combined data frame, they use the team abbreviations whereas, in the team data frame they use the team's full name. This meant that it was required to create a .csv file named 'team_names' to map between the abbreviation and the full team names. The data frame was then merged with the previously combined data frame using the team names and year columns. Once this was finished, the data types of the columns needed to be checked and changed to the correct data types, then converted this new combined data frame to a .csv file called 'all_stats'.

Finally, before running any machine learning algorithm, it was necessary to check if there were any more missing values in the data as some machine learning models such as K-NN struggle when there are missing values in the data. After inspecting the ‘all_stats’ data frame, missing values were present within the 3-point percentage, the free throw percentage and the field goal percentage. Upon further inspection, it was understood that this was large because these players did not attempt any 3-pointers, free throws or field goals. Usually, when faced with missing values, it is necessary to either remove the affected rows, replace the missing values with the average of the column or median if there is a heavy skew, or replace the missing values with zero. It was required to replace the missing values with zero because removing players could lead to the removal of an MVP candidate whose game revolves around scoring closer to the basket and changing the values to the average or median could make a player appear far better than he is.

3.3 Exploratory Analysis

The MVP is a prestigious award sought after by all great players along with winning the championship as these two accolades can help cement a player’s name in the history books. To win the greatest majority of voting shares it is imperative that a player performs well both individually and as a team considering winning is the primary objective of the game. Therefore, statistics separate an MVP calibre player from the rest and help them attain the greatest number of votes. This segment will consider the statics that assists in the determination of MVP.

Scoring is recorded by points (PTS) whereas a player can score points by a 2-pointer, 3-pointer, or a free throw for 1-point. One statistic that is constantly compared to the greatest player is their points per game. This statistic is used as a metric to measure a player's ability to score throughout their career. Precisely 86% of all MVPs so far have averaged over 20 points per game often greater than the league average for their respective year. Offensive stats such as points per game are often prioritised over defensive stats when deciding MVP. Of all the MVPs so far only two can be considered defensive players (Bill Russel and Wes Unsled) there are other great defensive MVPs, however, they played a more hybrid role being both good scorers as well as great defenders, players such as Hakeem Olajuwon, Tim Duncan and David Robinson.

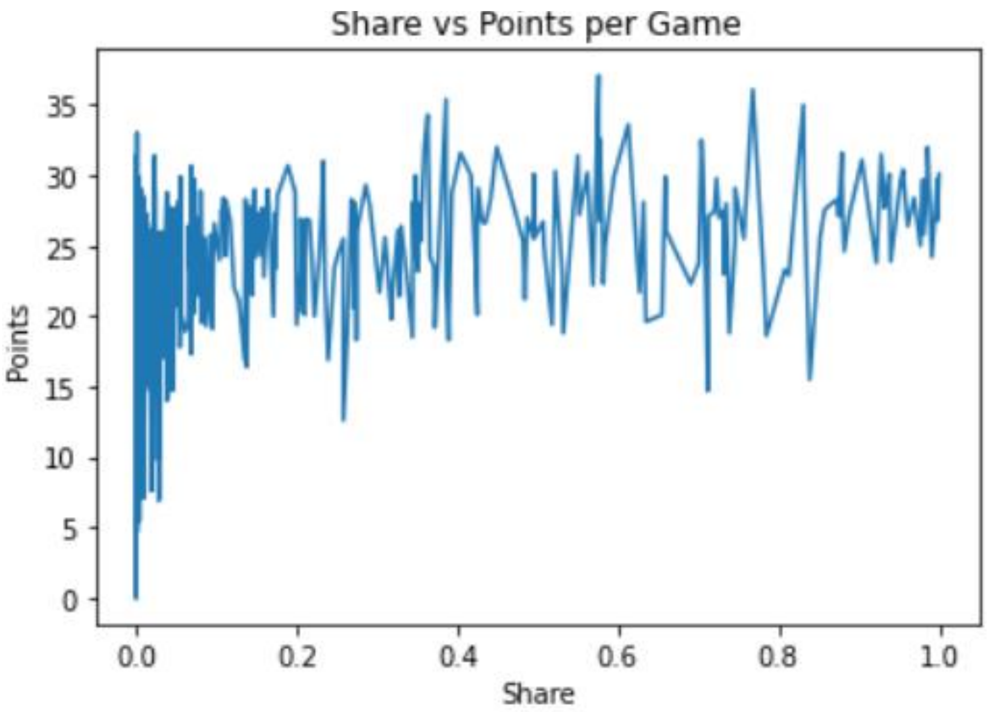


Figure 2: Share vs. Points per Game

Figure 2 shows a slight correlation between points per game share. The graph shows a slight increase in voter shares as points per game rise. This is likely because this statistic does not fully contextualise a player's ability to score as. More advanced statistics such as VORP and PER are better at determining MVP winners because they contextualise the player's offensive ability better (Freire, 2021; Chen et al., 2019).

The team win/loss percentage is presented as a ratio of the number of games won by a team over the number of games played in that season. It is rare that a player with a win/loss percentage less than 0.5 to win MVP. A player would need to have a historic season to be in the conversation of being an MVP nominee if his team has a poor win/loss percentage. Only Kareem Abdul-Jabbar and Bob Pettit have won the award whilst having a losing record. Bob Pettit won in 1955-56, his team had a win/loss percentage of 0.458. Bob Pettit was the league leader in PER, FG, FGA, FT, FTA, REB, and PPG. In the 1975-76 season, Kareem won his fourth MVP award, the Los Angeles Lakers had a win/loss percentage of 0.488. Kareem had an exceptional season ranking first in the league in MP, REB, BLK, PER, WS, and PM whilst, finishing second in FG, FGA, and PTS. In fact, only five players have won MVP with a win/loss percentage under 0.6; Bob Pettit (1955-56), Kareem Abdul-Jabbar (1975-76), Russell Westbrook (2016-17), Bob McAdoo (1974-75) and Moses Malone twice (1978-79, 1981-82).

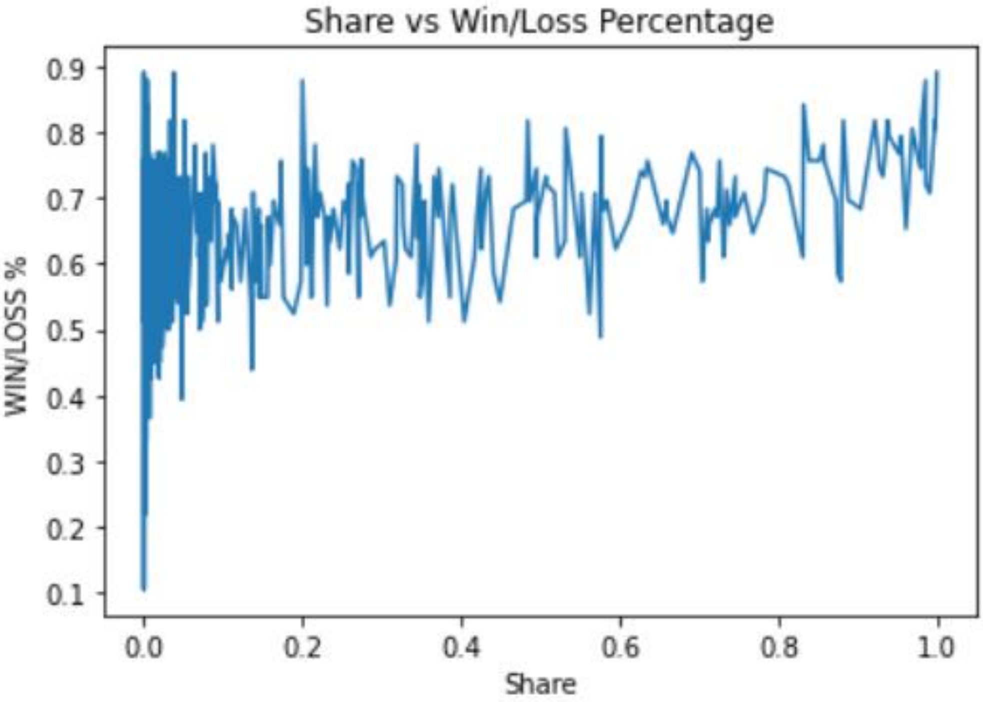


Figure 3: Share vs. Win/Loss Percentage

Figure 3 shows a positive correlation between win/loss percentage. This is expected as team performance is often taken into account when determining MVP because the aim of basketball is to win. Putting up outstanding numbers but losing is often referred to as putting up empty stats as the effort does not make the team better, maybe reducing the scoring and getting other teammates involved may make the team better.

There has been a big shift in NBA game of basketball since the introduction of the 3-point line in the 1980 season. Initially, the 3-point line was seen as a gimmick by most teams with 227 being shot in the 1980 season. It was until after the 1986 season that 3-point shot attempts started steadily rising, comprising 20% of all scoring attempts in the mid-2000s and skyrocketing over the last decade to make up over 33% of all true shot attempts

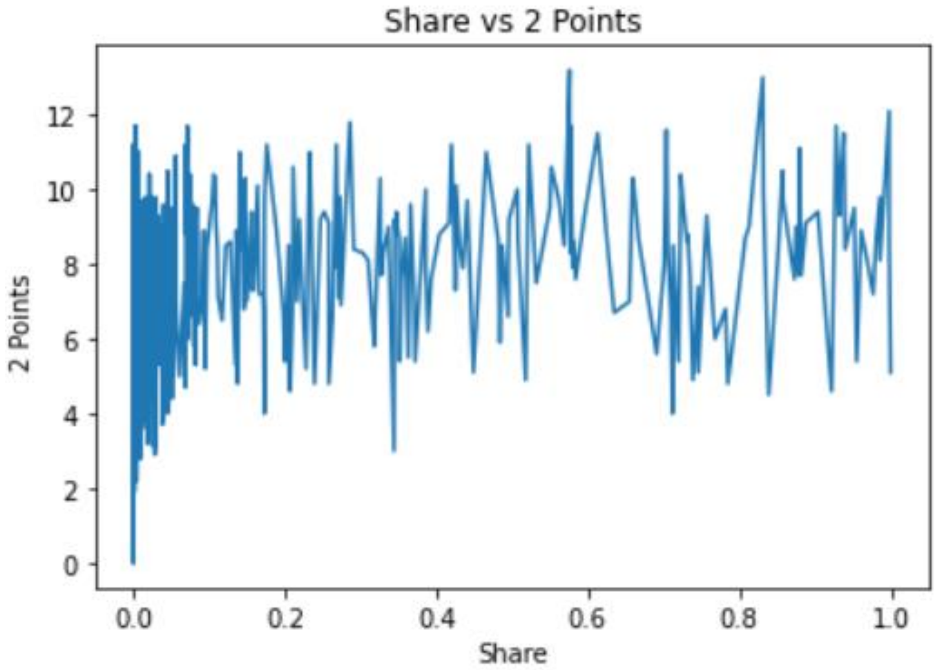


Figure 4: Share vs. 2 Points

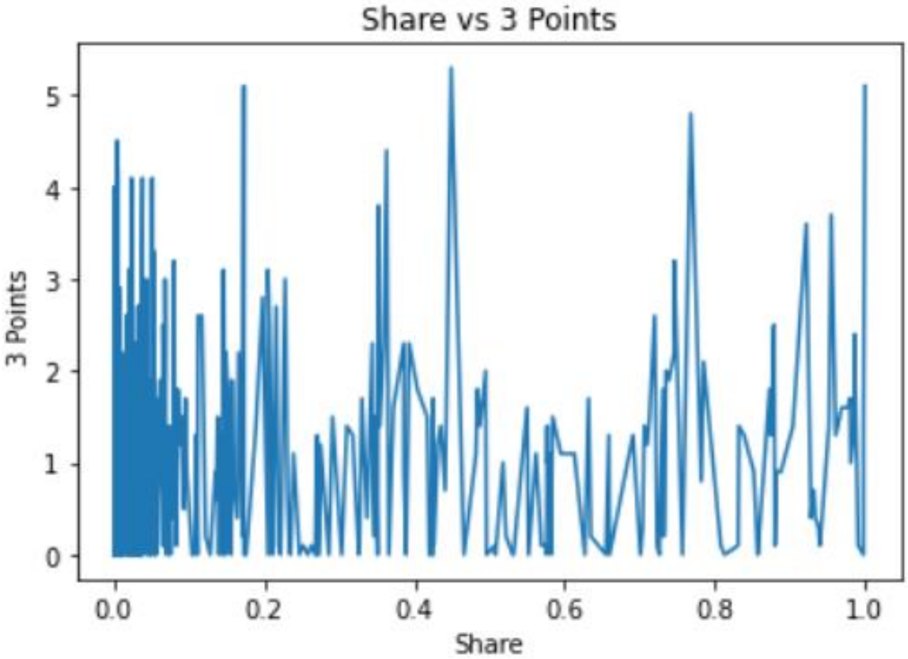


Figure 5: Shares vs. 3 Points

Figure 4 shows a slight positive correlation between the 2-point made per game and voter shares. However, Figure 5 is far noisier and difficult to tell if a correlation exists. There seems to be little correlation between a player's 3-point made per game and their voter share. This may be because the way in which defensive guard players who are prolific 3-point scorers opens up space for the offence to exploit and score 2-pointers more easily. Figure 6 shows a rise in both 2 and 3-point percentages which may be explained by this.

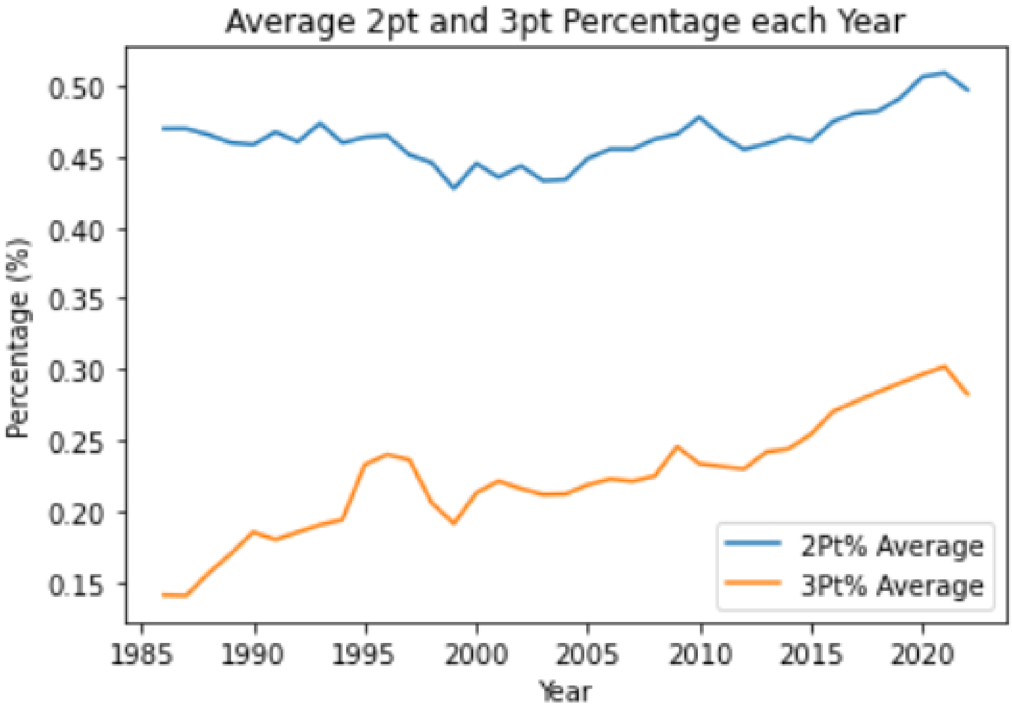


Figure 6: Average 2pt and 3pt Percentage each Year

Field Goal percentage is the ratio of shots made by both 2 and 3-pointers over the number of shots taken. Figure 7 illustrates the field goal percentages against the voter share the illustration shows that overall accuracy seems to have little impact on MVP votes. This is likely because field goal percentage does not paint a very complete picture of scoring efficiency. After all, the statistic fails to account for whether those shots were 2 or 3-pointers. Since 3-pointers are worth 50% more than 2-pointers, a player only needs to shoot 33% from 3-point range to match the number of points, he would need to score from 2-point range.

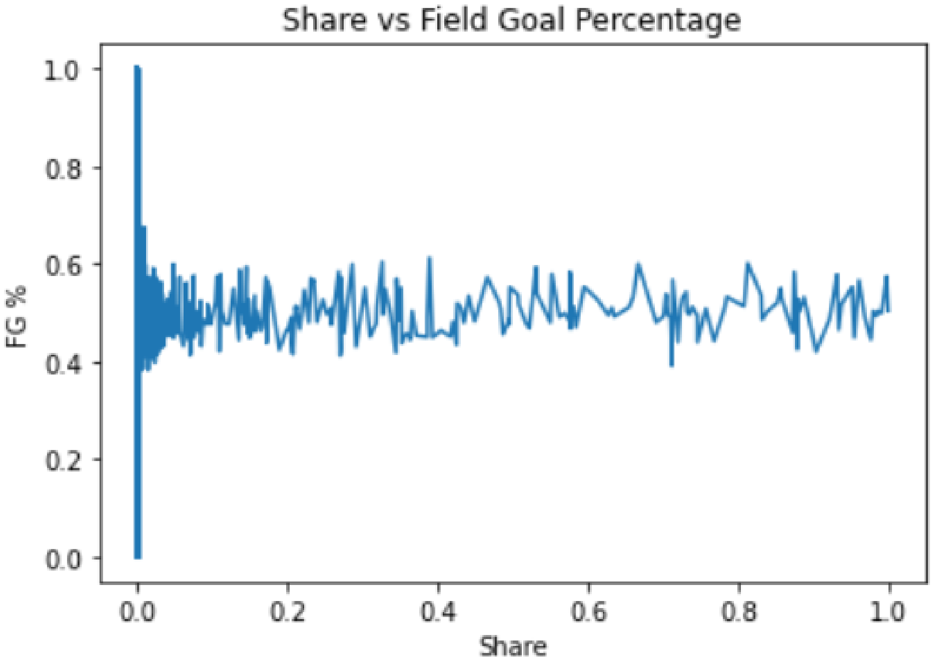


Figure 7: Share vs. Field Goal Percentage

Figure 8 displays the minutes played against the voter shares. From the graph, a small positive correlation can be seen. The graph also shows that only players who played between 30 and 46 minutes per game received a significant share.

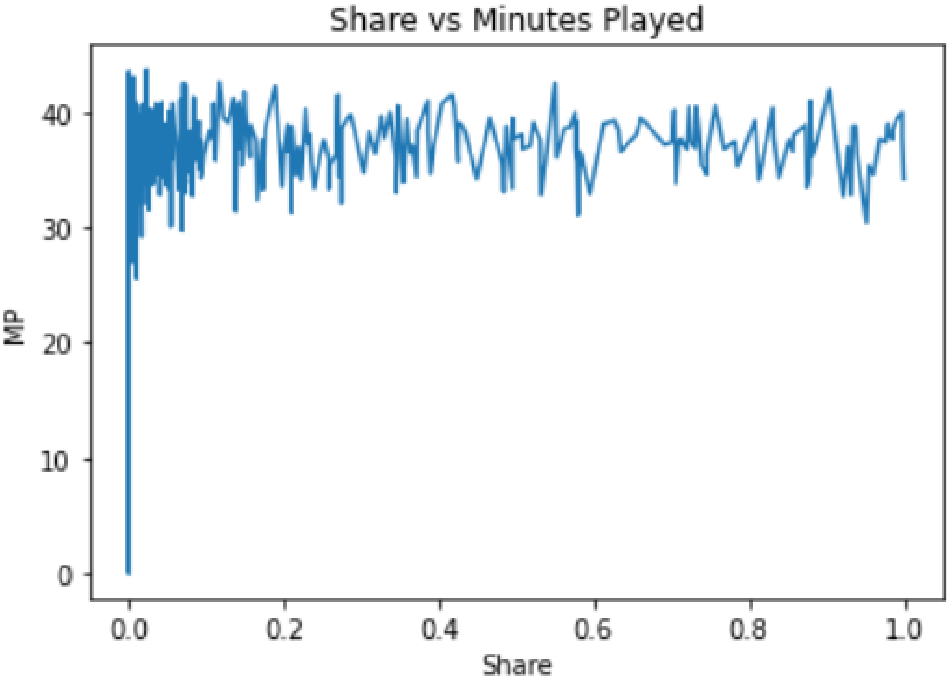


Figure 8: Share vs. Minutes Played

Appendix A shows a correlation map. From this map, all the features which are highly correlated with each other can be seen. It can then be removed from the model as a means of avoiding feeding the model duplicate information and avoiding high levels of collinearity that could lead to the overfitting of the model. After investigating the correlation heatmap, G, GS, FGA, 3PA, FTA, ORB, DRB, WL, GB, PA/G and SRS from the model was removed.

3.4 Model

There are five machine learning techniques chosen to use as predictive models in the determination of linear regression, LASSO, K-NN, random forest and MLP. The first technique employed was the linear regression model. The 'scikit-learn' library in python was used to run each regression. Instead of using the mean squared error as the error metric as is often the case with most regressions, the average precision error was used instead as this error works better when dealing with rankings.

The objective of these regressions is to predict the most valuable player for the NBA season 2021-22. The error metric will gauge the accuracy of each models' ability to correctly identify the five players with the most voting shares. Each models' chosen MVP will be compared with the true winner of the MVP award to ascertain if the model was able to correctly identify the MVP.

4. Results

Five machine learning techniques were used to predict the NBA's most valuable player for the 2020-21 season. Table 4 shows the expected MVP winner determined by each algorithm. The linear, LASSO and K-NN all predicted Joel Embiid as the winner of the award for the 2020-21 season. Random Forest predicted Nikola Jokic and MLP predicted Emanuel Terry.

Machine Learning Technique	**MVP Winner**
Linear Regression	Joel Embiid
LASSO	Joel Embiid
Random Forest	Nikola Jokic
K-NN	Joel Embiid
MLP	Emanuel Terry

Table 4: Predicted MVP Winners

Figures 9, 10, 11, 12 and 13 illustrate the top five MVP candidates and their expected predicted shares. The top three MVP candidates remain the same for the linear, LASSO, Random Forest and K-NN regressions. However, their rankings are different in order. All three regressions correctly identified four of the top five MVP candidates. However, the MLP failed catastrophically in its determination of any of the top five candidates.

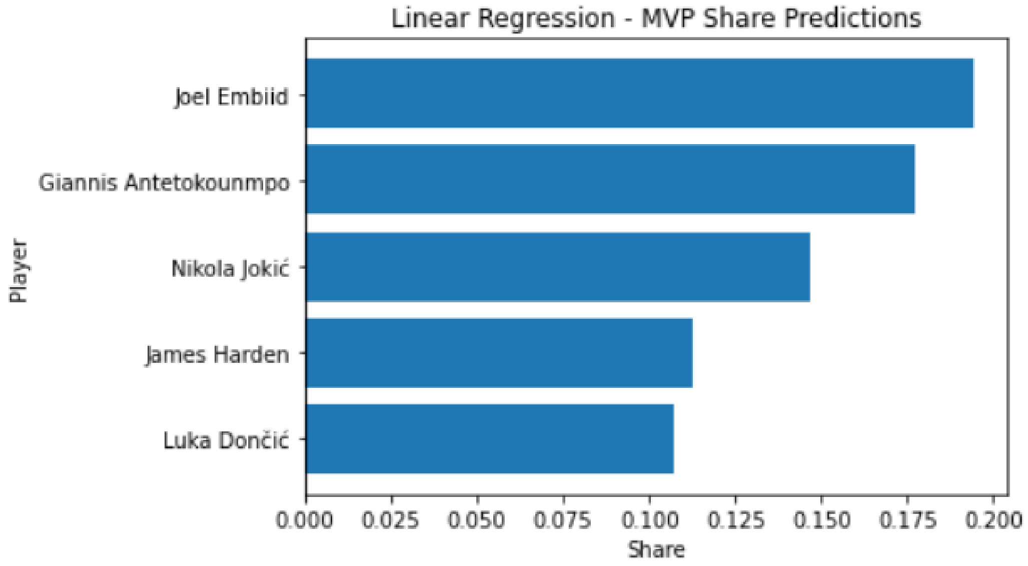


Figure 9: Linear Regression MVP Share Predictions

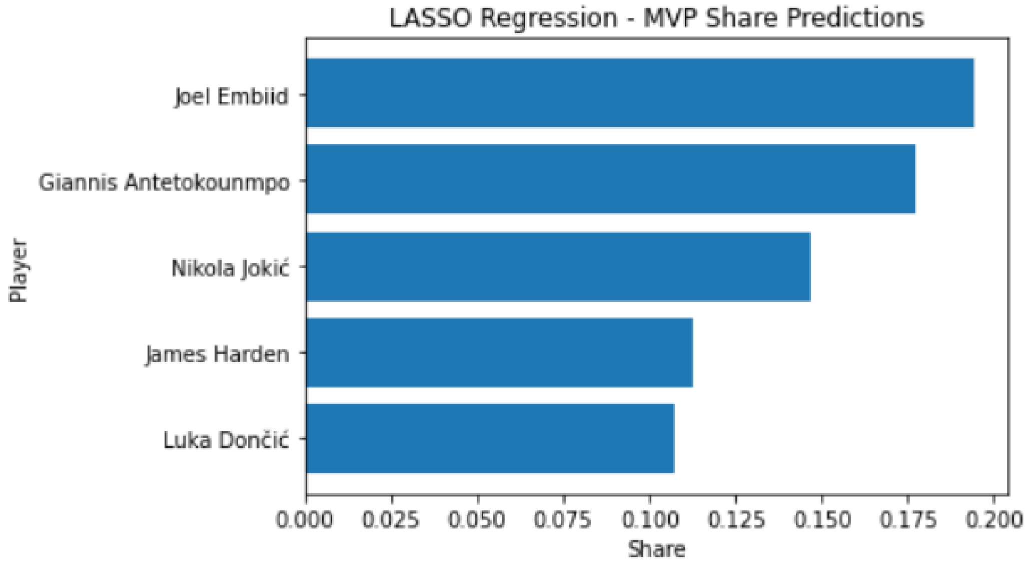


Figure 10: LASSO Regression MVP Share Predictions

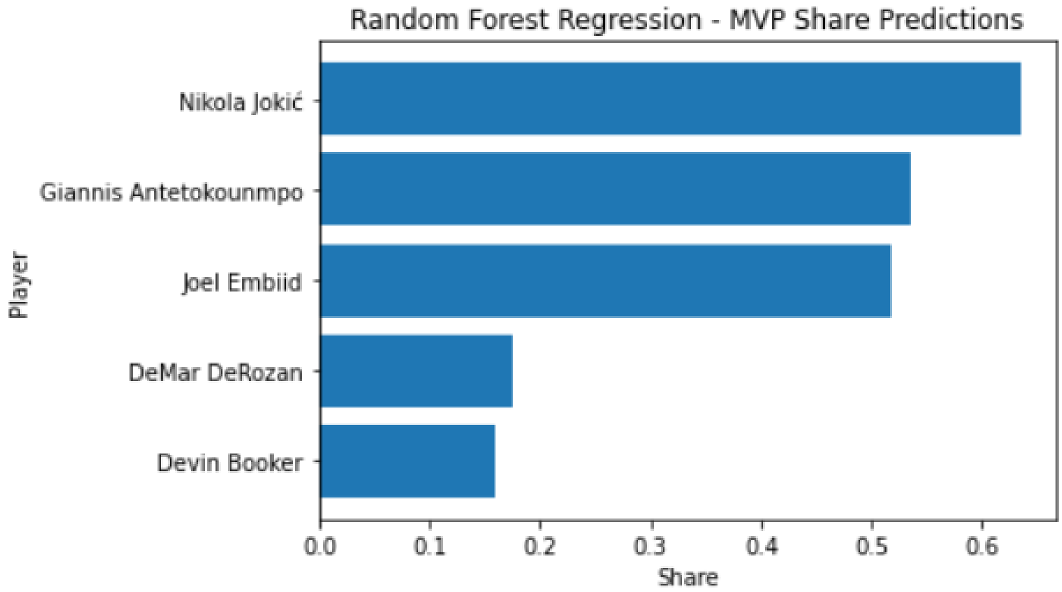


Figure 11: RF Regression MVP Share Predictions

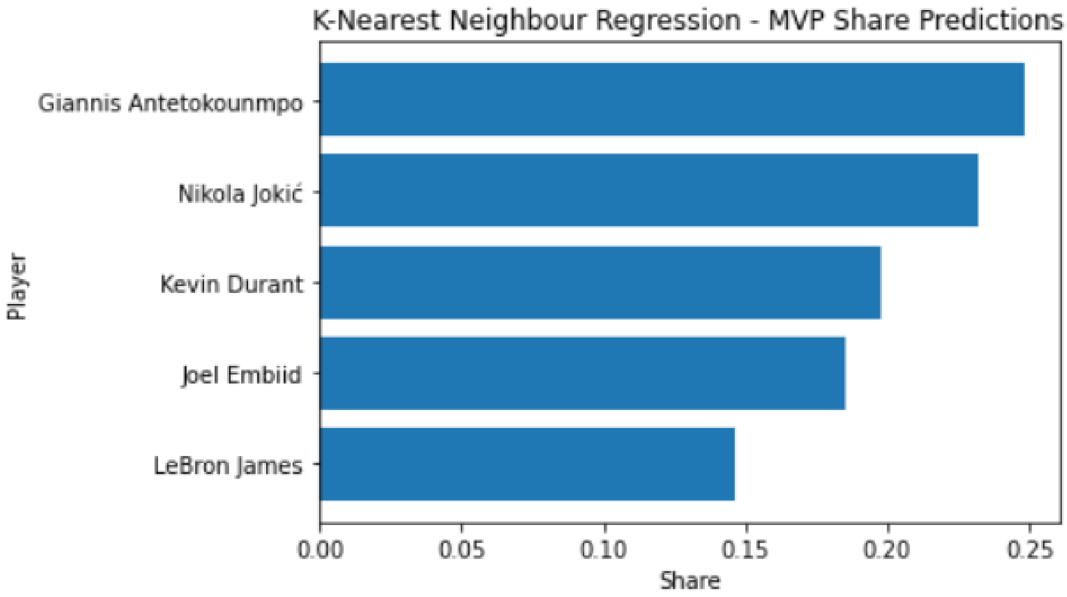


Figure 12: K-NN Regression MVP Share Predictions

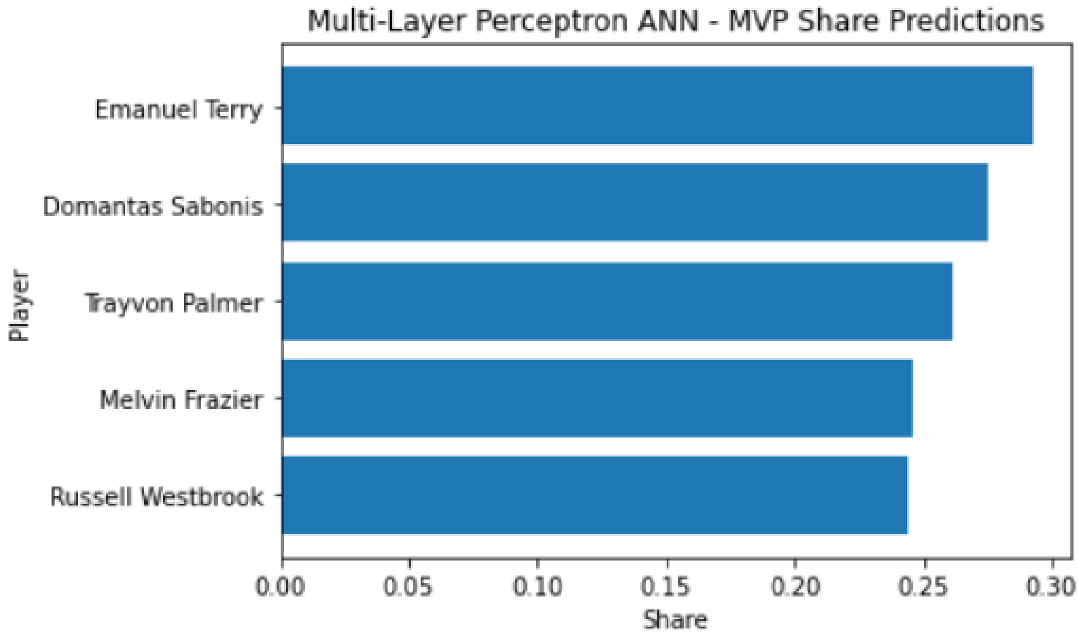


Figure 13: MLP-ANN MVP Share Predictions

From Table 5 it can be identified that the Random Forest had the greatest accuracy when trying to identify the top five MVP candidates. This is evident by its average precision error (0.903) being the greatest out of the techniques employed. The linear (0.797) and LASSO (0.797) regressions was also very successful when predicting the top five MVP candidates followed by the K-NN (0.713) regression. The MLP (0.040) was completely unsuccessful in its predictions of MVP candidates.

Machine Learning Technique	Average Precision Error
Linear Regression	0.797
LASSO	0.797
Random Forest	0.903
K-NN	0.713
MLP	0.040

Table 5: Average Precision Errors

In order to test the capabilities of these techniques, it was essential that each technique was tested against several years (Table 6). In doing so, a huge drop in the error metric for the Random Forest regression can be seen. Testing the technique against several years saw a massive drop in the accuracy produced by the Random Forest model from 0.903 to 0.742. The linear and LASSO regression accuracy remained relatively similar when tested across several years. K-NN also saw a large drop in accuracy falling from 0.883 to 0.676. The regression models outperformed the artificial neural network model considerably.

Machine Learning Technique	Mean Average Precision Error
Linear Regression	0.766
LASSO	0.765
Random Forest	0.742
K-NN	0.676
MLP	0.208

Table 6: Mean Average Precision Error

5. Discussion

The results of this thesis illustrated that several of the techniques were able to accurately predict the top five MVP candidates for the 2021-22 season. The Random Forest regression yielded the greatest prediction, and accuracy in comparison with the other four techniques. No technique was able to fully predict the exact placement of the top five MVP candidates. This is likely because the MVP award is determined by multiple variables naturally voters consider a player's statistics when identifying the choice for the award. Nonetheless, other factors such as narratives, influence, or voter fatigue can be determinants when choosing the award winner. Flom (2016) suggests that narratives portrayed by the media and basketball pundits can sway voters. Russell Westbrook famously won the MVP award in the 2016-17 season; the media portrayed him favourably constantly reminding viewers of the historic achievement of averaging a triple-double throughout the season. This had only been accomplished once before that season. However, James Harden was equally having a historic season, and yet the MVP award was easily taken by Russell Westbrook. The narrative aided in his efforts to win the accolade as for the next two seasons he averaged a triple-double again and was hardly in the MVP conversation. Anderson (2017) suggests that voter fatigue can influence the choice for MVP. This is because voters become uninterested when seeing the same player winning the award constantly. An example of this was in 1993, Charles Barkley won the MVP although Michael Jordan was statistically the best player in the league. This was likely because Jordan had won the previous two MVPs.

Furthermore, during the formulation of the K-NN regression model, the tuning parameters were randomly chosen in this case. It would have been best to test multiple K-values to produce the optimal value for the predicted model. Similarly, for the MLP, multiple hidden layers and the number of hidden neurons should have been tested to create a more optimised model. This may have been the reason for the poor predictive capacity of the MLP. However, when testing the MLP for multiple years, due to the high number of iterations and hidden layer size, obtaining the result was time-consuming. This made it difficult to test out different parameters that may have improved the overall predictive capacity of the model.

6. Conclusion

The objective of this thesis was to predict the winner of the NBA's most valuable player award for the current year. The Random Forest regression was able to correctly predict the winner of the 2022 MVP award. All the techniques with the exception of the MLP were able to predict the top five players in the MVP race. The Random Forest, Linear and LASSO regressions were the most suitable methods for predicting MVP through multiple years. The K-NN and MLP hyperparameters should have been tuned to achieve the greatest possible predictions. The regression models significantly outperform the artificial neural network

Several steps can be taken in order to improve this work. The first would be to devise an error metric which penalises the misplacement of an MVP candidate more strictly. The average precision metric used suggests that most of the models were good at predicting MVP candidates. The metric largely ignores the accuracy of the model ranking of those candidates within the top five and rewards the model as long as they are within the top five. A stricter metric may have revealed a worse predictive model; however, it would better represent the findings of this thesis.

Another method for improving the predictability of the model would have been to formulate additional statistics using the traditional statistics which may have been better predictors for winning MVP than the features in the model. Often multiple statistical columns are taken into account when determining the MVP, the aggregation of all the offensive statistics (PTS, TRB, AST) to create a player offence stat could have been made. The best players would have a greater offence stat number and this feature would likely have been a good predictor of players' offensive ability. The same could be done to create a player defence stat aggregating STL, BLK, and DRB. Furthermore, it would have been beneficial to collect more team statistics. Weiner (2021) and Chen (2017) suggest that team performance is more indicative of both winning basketball games and winning MVP than individual statistics.

Using a smaller more selective data set may have created significant improvements in prediction. Freire (2021) found success in predicting the MVP winner for the 2020-21 season where he only collected player and team statistics for the top ten players in that season.

Nevertheless, most of the techniques utilised in this study produced fairly accurate predictions of the top five MVP candidates for the current season, however only one was able to predict the eventual winner of the award. The modifications discussed can be applied to improve the accuracy and prediction of the model and the techniques discussed used to create predictive models for the forecasting of essential information in basketball and other sports.

List of References

Anderson, B., 2017. What Happens When We Get the NBA MVP Wrong?. [online] Medium. Available at: <https://medium.com/sportsraid/what-happens-when-we-get-the-nba-mvp-wrong-f78effb25114> [Accessed 9 September 2022].

Bikku, T., 2020. Multi-layered deep learning perceptron approach for health risk prediction. Journal of Big Data, [online] 7(1). Available at: <https://doi.org/10.1186/s40537-020-00316-7>.

Castro, W., Oblitas, J., Santa-Cruz, R. and Avila-George, H., 2017. Multilayer perceptron architecture optimization using parallel computing techniques. PLOS ONE, [online] 12(12), p.e0189369. Available at: <https://doi.org/10.1371/journal.pone.0189369>.

Chen, M., 2017. Predict NBA Regular Season MVP Winner. Proceedings of the International Conference on Industrial Engineering and Operations Management, [online] pp.pp.25-26. Available at: <http://ieomsociety.org/bogota2017/papers/9.pdf> [Accessed 1 September 2022].

Chen, Y., Dai, J. and Zhang, C., 2019. A Neural Network Model of the NBA Most Valued Player Selection Prediction. Proceedings of the 2019 the International Conference on Pattern Recognition and Artificial Intelligence - PRAI '19,.

Corvo, M., 2021. How Voting is Done for the NBA MVP and Its Evolution. [online] ClutchPoints. Available at: <https://clutchpoints.com/how-voting-is-done-for-the-nba-mvp-and-its-evolution/> [Accessed 1 September 2022].

Couronné, R., Probst, P. and Boulesteix, A., 2018. Random forest versus logistic regression: a large-scale benchmark experiment. BMC Bioinformatics, [online] 19(1). Available at: <https://doi.org/10.1186/s12859-018-2264-5>.

Cunningham, P. and Delany, S., 2022. k-Nearest Neighbour Classifiers - A Tutorial. ACM Computing Surveys, [online] 54(6), pp.1-25. Available at: <https://doi.org/10.1145/3459665>.

Deyasi, A., Bhattacharjee, A., Mukherjee, S. and Sarkar, A., 2021. Multi-layer Perceptron based Comparative Analysis between CNTFET and Quantum Wire FET for Optimum Design Performance. Solid State Electronics Letters, [online] 3, pp.42-52. Available at: <https://doi.org/10.1016/j.ssel.2021.12.003>.

Flom, R., 2016. The Power of Narrative in the NBA. [online] Clips Nation. Available at: <https://www.clipsnation.com/2016/9/9/12815810/the-power-of-narrative-in-the-nba> [Accessed 9 September 2022].

Freire, D., 2021. Predicting 2020–21 NBA’s Most Valuable Player using Machine Learning. [online] Medium. Available at: <https://towardsdatascience.com/predicting-2020-21-nbas-most-valuable-player-using-machine-learning-24aaa869a740> [Accessed 2 September 2022].

Jones, E., 2016. PREDICTING OUTCOMES OF NBA BASKETBALL GAMES. North Dakota State University of Agriculture and Applied Science, [online] (Fargo, North Dakota). Available at: <https://library.ndsu.edu/ir/bitstream/handle/10365/28084/Predicting%20Outcomes%20of%20NBA%20Basketball%20Games.pdf?sequence=1&isAllowed=y> [Accessed 1 September 2022].

McCorey, J., 2021. FORECASTING MOST VALUABLE PLAYERS OF THE NATIONAL BASKETBALL ASSOCIATION. University of North Carolina at Charlotte, [online] Available at: <https://island1.uncc.edu/islandora/object/etd%3A2482/datastream/PDF/download/citation.pdf> [Accessed 1 September 2022].

Minusha Silva, R., 2016. Sports Analytics. Sam Houston State University, [online] Available at: <https://summit.sfu.ca/_flysystem/fedora/sfu_migrate/16939/etd9888_RSilva.pdf> [Accessed 1 September 2022].

NBA, 2014a. Ten ways David Stern helped grow the game of basketball. [online] Nba.com. Available at: < https://www.nba.com/pelicans/news/ten-ways-david-stern-helped-grow-game-basketball> [Accessed 14 February 2022].

NBA, 2014b. NBA sets record with 101 international players from 37 countries and territories. [online] Nba.com. Available at: < https://pr.nba.com/nba-international-players-2014-15/> [Accessed 14 February 2022].

NBA, 2022a. This Date in the NBA: June. [online] Nba.com. Available at: <https://www.nba.com/news/history-this-date-in-nba-june> [Accessed 14 February 2022].

NBA, 2022b. Kia NBA MVP Tracker: Latest performances from leading contenders. [online] Nba.com. Available at: < https://www.nba.com/news/kia-nba-mvp-tracker-2022> [Accessed 15 August 2022].

Patton, A., Scott, M., Walker, N., Ottenwess, A., Power, P., Cherukumudi, A. and Lucey, P., 2020. Predicting NBA Talent from Enormous Amounts of College Basketball Tracking Data. 15th Annual MIT Sloan Sports Analytics Conference, [online] Available at: <https://www.researchgate.net/profile/Andrew-Patton-8/publication/354131580_Predicting_NBA_Talent_from_Enormous_Amounts_of_College_Basketball_Tracking_Data/links/6126b6da035d5831d7725350/Predicting-NBA-Talent-from-Enormous-Amounts-of-College-Basketball-Tracking-Data.pdf> [Accessed 1 September 2022].

Sarlis, V. and Tjortjis, C., 2020. Sports analytics — Evaluation of basketball players and team performance. Information Systems, 93, p.101562.

Schonlau, M. and Zou, R., 2020. The random forest algorithm for statistical learning. The Stata Journal: Promoting communications on statistics and Stata, [online] 20(1), pp.3-29. Available at: <https://doi.org/10.1177/1536867X20909688>.

Shreyas, R., Akshata, D., Mahanand, B., Shagun, B. and Abhishek, C., 2016. Predicting popularity of online articles using Random Forest regression. 2016 Second International Conference on Cognitive Computing and Information

Processing (CCIP), [online] pp.pp.1-5. Available at: <https://doi.org/10.1109/CCIP.2016.7802890>.

Torres, R.A., 2013. Prediction of NBA games based on Machine Learning Methods. University of Wisconsin, Madison.

Uddin, S., Haque, I., Lu, H., Moni, M. and Gide, E., 2022. Comparative performance analysis of K-nearest neighbour (KNN) algorithm and its different variants for disease prediction. Scientific Reports, [online] 12(1). Available at: <https://doi.org/10.1038/s41598-022-10358-x>.

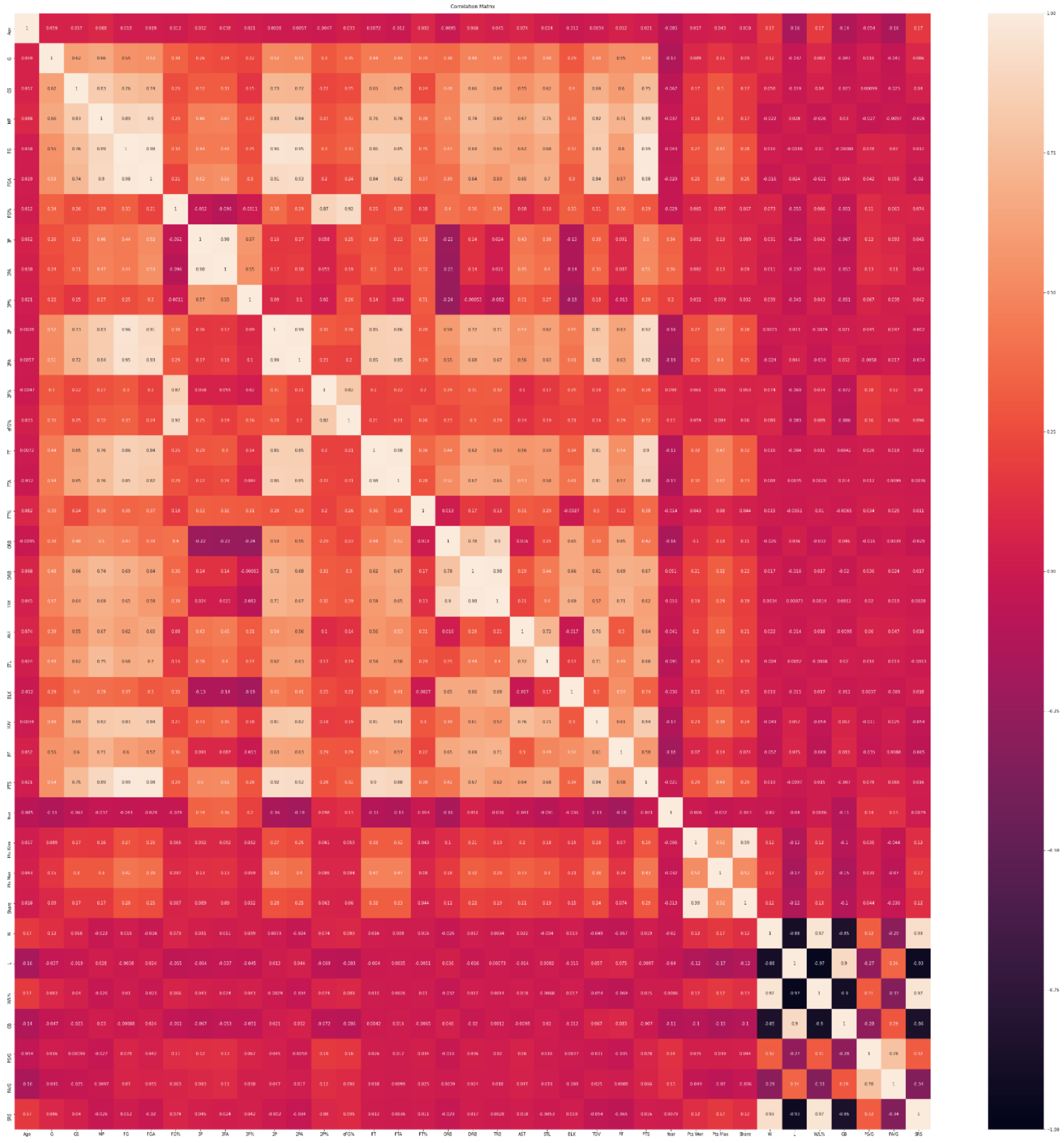
Wang, Z., Wang, Y., Xuan, J., Dong, Y., Bakay, M., Feng, Y., Clarke, R. and Hoffman, E., 2006. Optimized multilayer perceptrons for molecular classification and diagnosis using genomic data. Bioinformatics, [online] 22(6), pp.755-761. Available at: <https://doi.org/10.1093/bioinformatics/btk036>.

Weiner, J., 2021. Predicting the outcome of NBA games with Machine Learning. [online] Towards Data Science. Available at: <https://towardsdatascience.com/predicting-the-outcome-of-nba-games-with-machine-learning-a810bb768f20> [Accessed 14 February 2022].

Zhang, Z., 2016. Introduction to machine learning: k-nearest neighbors. Annals of Translational Medicine, [online] 4(11), pp.218-218. Available at: <https://doi.org/10.21037/atm.2016.03.37>.

Appendices

Appendix A: Correlation Heat Map



Appendix B: Codes and .csv Files Uploaded onto GitHub

https://github.com/stnyarko/Dissertation