National College of Ireland

Basketball Performance Prediction Models and Team Efficiency Factors

MSc Research Project
Data Analytics

Karthik Nousher
Student ID: 22138668

School of Computing
National College of Ireland

Supervisor: Dr. Catherine Mulwa

National College of Ireland
Project Submission Sheet
School of Computing

Student Name:	Karthik Nousher
Student ID:	22138668
Programme:	Data Analytics
Year:	2023
Module:	MSc Research Project
Supervisor:	Dr. Catherine Mulwa
Submission Due Date:	14/12/2023
Project Title:	Basketball Performance Prediction Models and Team Efficiency Factors
Word Count:	4522
Page Count:	22

I hereby certify that the information contained in this (my submission) is information pertaining to research I conducted for this project. All information other than my own contribution will be fully referenced and listed in the relevant bibliography section at the rear of the project.

ALL internet material must be referenced in the bibliography section. Students are required to use the Referencing Standard specified in the report template. To use other author's written or electronic work is illegal (plagiarism) and may result in disciplinary action.

Signature:	Karthik Nousher
Date:	31st January 2024

PLEASE READ THE FOLLOWING INSTRUCTIONS AND CHECKLIST:

Attach a completed copy of this sheet to each project (including multiple copies).	☐
Attach a Moodle submission receipt of the online project submission, to each project (including multiple copies).	☐
You must ensure that you retain a HARD COPY of the project, both for your own reference and in case a project is lost or mislaid. It is not sufficient to keep a copy on computer.	☐

Assignments that are submitted to the Programme Coordinator office must be placed into the assignment box located outside the office.

Office Use Only	
Signature:	
Date:	
Penalty Applied (if applicable):	

Basketball Performance Prediction Models and Team Efficiency Factors

Karthik Nousher
22138668

Abstract

The project revolves around the performance-increasing factors that contribute to the efficiency of a basketball team. The study uses various algorithms primarily linear regression, random forest and decision tree, later on using the algorithm with the highest efficiency to predict the team efficiency. Here, Linear regression showed a higher efficiency value. the study critically examines different prediction models, in which it highlights the impact of linear regression, having a low MSE of 130.736, RME of 130.736 and an MAE of 9.385. The main idea of his research is to contribute to the sports world, thereby helping the team to win the game. The study also addresses some of the key elements in the basketball game such as player efficiency ratings, shooting accuracy, defensive ability, and team management. Coaches and analysts can optimize player roles, devise winning strategies, and improve overall team performance. This approach reduces the gap between data analytics and basketball performance evaluation. Knowing the factors that affect the player's efficiency and also understanding the best-fit model are some of the important factors that are needed to advance the basketball game. This research provides some valuable insights that can be used to optimize player roles, develop effective strategies, and elevate total team performance.

1 Introduction

Basketball is a popular and intensely competitive sport that necessitates a complete understanding of team dynamics. To maximise basketball team success and individual player development, organisations are using data-focused insights to predict and analyse the upcoming game, which in turn gives the player insight on how to play the game and which player should be in which position. Compared with other sports games, basketball is said to be a complicated one, because the sport contains elements that are related to each other and can affect the team's performance said by Naismith (1996). The main goal of this research project is to explore the use of data analytics to leverage players' efficiency in this field, focusing mainly on player efficiency ratings and team performance The study mainly focuses on two issues firstly, it identifies the key factors that can significantly affect the player's efficiency rating and secondly, it is also used to determine the accuracy and functionality of various performance metrics in the basketball game., prediction models are used here to increase the performance. "The quality of the team has two important components". The first component encompasses the skills of the players and is expressed as their technical knowledge. The second component is expressed as the overall team tactics. For a team to be successful, requires good coordination between the individual

players and can only be achieved with a lot of training, which is provided in the study conducted by Sarlis and Tjortjis (2020)

1.1 Motivation and Project Background

The main motivation behind this research proposal is to elevate the importance of data analytics in the field of basketball. It depends mainly on two factors, such as player accuracy and the validity of the models that are used for this evaluation. The research problem addressed in the field of data analytics focuses on the main elements, which are basketball player efficiency ratings and the accuracy and validity of various models for evaluating team performance. Basketball analytics, or any sports analytics, has become a crucial topic in the area of data-focused decision-making and sports analytics. Basketball performance is mainly dependent on two or more metrics rather than just physical ability; the game is dependent on factors such as shooting accuracy, defensive play, rebound rate, ball handle, strategy play, and team improvisation. On the other side, team winning is improved by creating more winning strategies and improving overall performance.

Furthermore, the creation of precise and trustworthy performance prediction models is crucial in basketball analytics. All three models specified improve the player's effectiveness and team outcomes using historical game data, player statistics, and several other performance metrics such as 3-point shoot, plus-minus values, and block rates. Stakeholders such as coaches and players can use this decision in this sport's field, thereby leveraging their team's performance. This insight can help them understand much better about their team and their players. Finally, it is critical to comprehend the primary factors that significantly affect player efficiency ratings in basketball and to assess the validity and precision of various performance prediction models to advance sports analytics in the basketball sector.Naismith (1996)

1.2 Research Question

The research focuses on the basketball performance indicators and the different models used for the predictive analysis. The sub-research question mainly focuses on player performance by integrating these identified metrics along with the machine learning models. This integration helps in developing the team's potential.

Research Question: *"What specific basketball performance indicators significantly impact team efficiency ratings, and how do different basketball performance prediction models compare in terms of accuracy and reliability in evaluating overall team performance?"*

The research mainly focuses on certain key indicators that help in improving basketball performance, and it also compares three prediction models to find the best fit to perform this task effectively.

Sub-Research Question: *"How can the identified performance metrics in basketball be integrated into the framework of basketball analytics to enhance the efficiency of player assessments?"*

This is done by combining statistical insights and machine learning models to evaluate player efficiency, thereby helping the basketball team generate data-driven decisions based on the previous insights, which increases player roles and overall team performance.

The following research objective is implemented by identifying the pivotal metrics of

the basketball game, evaluating the models, and integrating these metrics into basketball analytics.

1.3 Research Objectives

- Obj1: A critical review of literature on team efficiency prediction in the NBA.
- Obj2(a): Exploratory data analysis with features.
- Obj2(b): Data transformation.
- Obj2(c): Implement, evaluate, and present the results of DecisionTreeRegressor.
- Obj2(d): Implement, evaluate, and present the results of LinearRegressor.
- Obj2(e): Implement, evaluate, and present the results of RandomForestRegressor.
- Obj2(f): Comparison of developed models with appropriate metrics.

The contribution resulting from this project will help stakeholders, such as coaches and basketball team members, evaluate their team's performance and also correct the team's decisions based on the results. The research sets off with a set of explorations, including a structured framework outlined by a series of objectives:

Section 2 showcases the literature review, which explores the evolving landscape of basketball analytics. It also covers diverse topics, including predictive player performance forecasting, comparative player evaluations, tools for sports writers, and deep learning for offensive play optimisation. Section 3 mainly introduces the data acquisition process and also describes machine learning methods such as random forest, multiple linear regression, and decision trees. Section 4 of this research explains the design specification and showcases how each model is depicted for this research, and finally, Section 5 showcases the implementation perspective and details about the evaluation and results.

2 Related Work: Basketball Performance Prediction using Predictive Models over the Past Decade

The involvement of data analytics in the game of basketball improves the scope of the game and player efficiency by providing insights for decision-making. Through the use of historical data and machine learning models, this can improve the accuracy of game results. In the realm of player efficiency, data analytics allows for the identification of specific performance metrics that align with overall team effectiveness. Tailoring analytical approaches to individual players provides nuanced insights into their strengths and areas for improvement. Additionally, predictive models can be crafted to forecast player efficiency, empowering teams to make informed decisions about player roles and strategic gameplay. This comprehensive approach to data analytics in basketball not only enhances performance evaluation but also shapes dynamic strategies for success.

In the realm of ever-growing technology Machine learning and data are being used to create significant changes in the realm of basketball. The expansion of datasets, along with the mix of machine learning, has opened new horizons. In this analysis, we

explore the confluence of basketball and machine learning through an examination of four pivotal articles. These articles shed light on how advanced analytics and predictive models are revolutionising the way we perceive and evaluate basketball, providing a deeper understanding of this sport and its players.

2.1 Prediction of Basketball Performance

Terner and Franks (2021) explores the evolving landscape of basketball analytics, emphasizing the shift from traditional metrics to advanced statistical and machine-learning models.The conclusion underscores the need for structured hierarchical models, deep learning approaches, and public data availability to advance basketball analytics while advocating for greater attention to causal and game-theoretic analyses in the field. In the evolving landscape of sports analytics, Vinué and Epifanio (2019)addresses the underexplored realm of sparse functional data in basketball player performance forecasting. Leveraging ROPES and PACE methods alongside functional archetype analysis, it pioneers a methodology for predicting athletes' future trajectories. Despite the inherent uncertainties in sports, this approach, validated through comparison studies, offers a valuable tool for coaches and analysts in assessing players' potential, contributing to the broader field of sports forecasting. In the realm of basketball analytics,Martínez (2020) innovatively addresses the challenge of comparing players' performance across different playing times. By introducing PTCpred, a prediction of Player Total Contribution per game (PTC/G) relative to minutes played, the paper contributes a practical tool for analysts, media, and fans. This approach adds valuable nuance to the evaluation of players' overall performance, filling a gap in the existing literature on basketball metrics and analytics. Fu and Stasko (2022) addresses the evolving landscape of basketball journalism in the era of extensive sports data and analytics. The result is the creation of two interactive visualization tools, NBA GameViz and NBA LineupViz, designed to empower writers to conduct swift, insightful analyses. Prototypes deployed during the NBA playoffs garnered valuable feedback, informing future directions for this innovative approach. Javadpour et al. (2022) Uses deep learning techniques to simplify only the selections of offensive plays in women's division basketball. By considering factors like defender positions, court location, passes, shot history, and shot clock time, the model predicts play outcomes. Sports result prediction, popular among global fans and linked to the rise in sports betting, is addressed by Miljković et al. (2010). Focusing on NBA games, a predictive system utilizing data mining, Naive Bayes classification, and multivariate linear regression achieved 67% accuracy in predicting match winners. The MVC Model 2 pattern-based software system demonstrates satisfactory results, motivating future experiments with diverse sports domains and alternative classification methods, including neural networks. Utilizing machine learning on player and ball tracking data,Tian et al. (2019) focuses on classifying defensive strategies in basketball, specifically against pick-and-roll plays. The analytical model considers spatio-temporal patterns and features to automatically identify and label defensive plays. Support Vector Machines (SVM) prove effective, achieving a 68.9% classification accuracy for switch and trap strategies. The approach demonstrates the potential of player tracking data and machine learning in understanding defensive strategies, informing coaching tactics and player development in team sports like basketball.

2.2 Review of Data science-based prediction of basketball performance

Sports analytics prove vital for coaching teams, particularly in small sports communities with limited access to advanced technology. A data analytics project at Universiti Putra Malaysia focused on basketball players' performance and utilized an experiential learning approach Sharef et al. (2020)covered shooting performance, attack profiling, team performance, player rating prediction, and game outcome prediction. Results indicated a focus on rebounding by the PXI Basketball team, emphasizing the significance of data-driven decision-making in sports coaching and fostering students' high-order thinking and collaborative skills. Sarlis et al. (2021)investigate the impact of injuries on basketball players and team performance in the NBA from 2010 to 2020, utilizing Data Science and Sports Analytics. From this research, we can infer that injuries particularly in the musculoskeletal region are more evident.. The study suggests a holistic approach to injury analysis and emphasizes the importance of balanced rest/load management. Machine Learning techniques, such as LASSO and Ridge regression, offer insights for injury prediction and performance optimization, contributing to strategic decision-making in basketball management

2.3 Predicting per game performance through Machine Learning

Exploring the impact of sleep and training on Division-1 women's basketball during the pandemic, Senbel et al. (2022)employs machine learning to predict game performance and injuries with over 90% accuracy. Analyzing data from sleep monitors, training, and surveys, the study reveals associations between game performance and factors like sleep, training load, and heart rate variability. Notably, the prediction models show reliability, offering valuable insights for coaches to enhance athletes' performance and reduce injury risks amid challenging circumstances. Exploring basketball shooting gesture recognition,Ji (2020) employs image feature extraction and machine learning for precise classification. Extracting multi-dimensional motion features from time and frequency domains, the method demonstrates superiority in accuracy and effectiveness. Acknowledging challenges in human motion tracking due to complexity and variability, the research emphasizes the need for further exploration of diverse methods to comprehensively analyze various basketball moves in future studies. Addressing challenges in real-time basketball analysis,Yoon et al. (2019) utilizes a real-time object detection system, to track player movements and ball distribution under dynamic camera angles. The system incorporates a novel player-tracking algorithm to handle interruptions and overlaps in player visibility, enhancing accuracy. Employing network science, the research explores past relationships and player importance, acknowledging limitations in current deep learning algorithms while providing insights for future improvements in fully automated sports analytics. Examining basketball games across competitive levels, this study utilizes video-based time-motion analysis to assess the physical demands on adult male players.

2.4 Application of Artificial Intelligence in Basketball Performance

To understand the strategies and training methodologies, the use of AI was introduced by Bray and Whaley (2001), Their study on team cohesion and individual performance in high school basketball shows insight into how team dynamics affect individual and team performance. This research also points out that team members have a collective effort and also significantly impact the performance of individual players. The attributes used in a female basketball game are well explained by Ackland et al. (1997). Their study offers different sides, emphasising the importance of physical characters needed for player performance.During comparison, it is understood that Sampaio et al. (2015) approach is in a more data-driven way. The research shows how technology can be used to understand game dynamics and player performance. Application of basketball analysis using AI in Li and Xu (2021) shows the gap between basketball training methodologies and modern AI applications. Lastly.Cole (2021) uses Kirton's adaptation-innovation theory to explore team effectiveness and focuses on cognitive styles of problem-solving

In conclusion, this approach helps to understand the basketball game, predict player performance, and optimise player skills. Machine learning techniques employed for this game, help stakeholders such as coaches and basketball players to optimise or up their playing standards. Still, there are certain challenges such as data availability and human motion tracking.

3 Scientific Research Methodology

3.1 Basketball Scientific Methodology Approach

In this section, we outline the systematic methodology employed in this research on predicting future basketball team efficiency using various performance prediction models, The research here adopts the use of CRISP-DM methodology. To find this, various factors, such as team effectiveness on the court, are being evaluated. The methodology comprises certain key steps, such as data collection, model configuration, and evaluation, as illustrated in Figure 1. This structured framework aims to harness the predictive capabilities of Random Forest, Linear Regression, and Decision Tree algorithms in the dynamic context of basketball performance. By applying these models to the "Basketball Performance Prediction Models and Team Efficiency Factors" research topic, the main objective or goal is to enhance the understanding and prediction of team efficiency factors in the basketball domain.

3.2 Design flow of Predicting Basketball Performance and Team Efficiency

The research conducted is done in a set of steps as shown in Figure 2 2. All the Python libraries, such as matplotlib, seaborn, sklearn, numpy are used for this research. Here, the research trends and patterns in the basketball performance data are sorted. In this research study, linear regression, random forest, and decision tree algorithms are used. Here, the models are then evaluated, and the best model from the three is then later used in the real world to get the predictions. The main design objective of this project

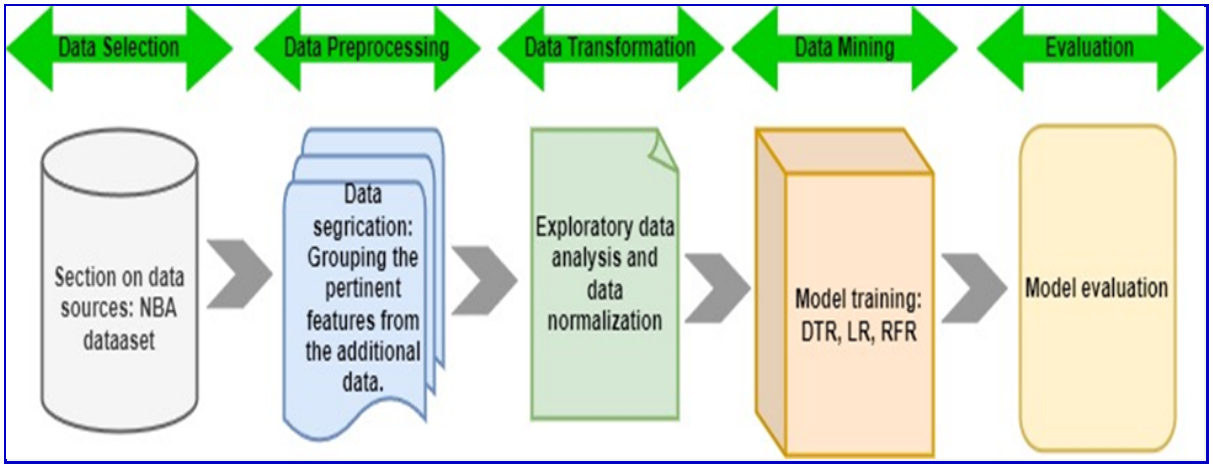

Figure 1: Research Methodology for Basketball Performance Prediction Models and Team Efficiency Factors

is to gather basketball information, clean the dataset, involve some feature engineering, and finally split the data into two sets to perform training, testing, model training, evaluation, and deployment. To get good performance results, hyperparameterization is employed, which leads to optimal model performance and finally retains the importance of regular model retraining with new data for continuous improvement. This layout provides a systematic and comprehensive methodology for developing accurate basketball performance prediction models

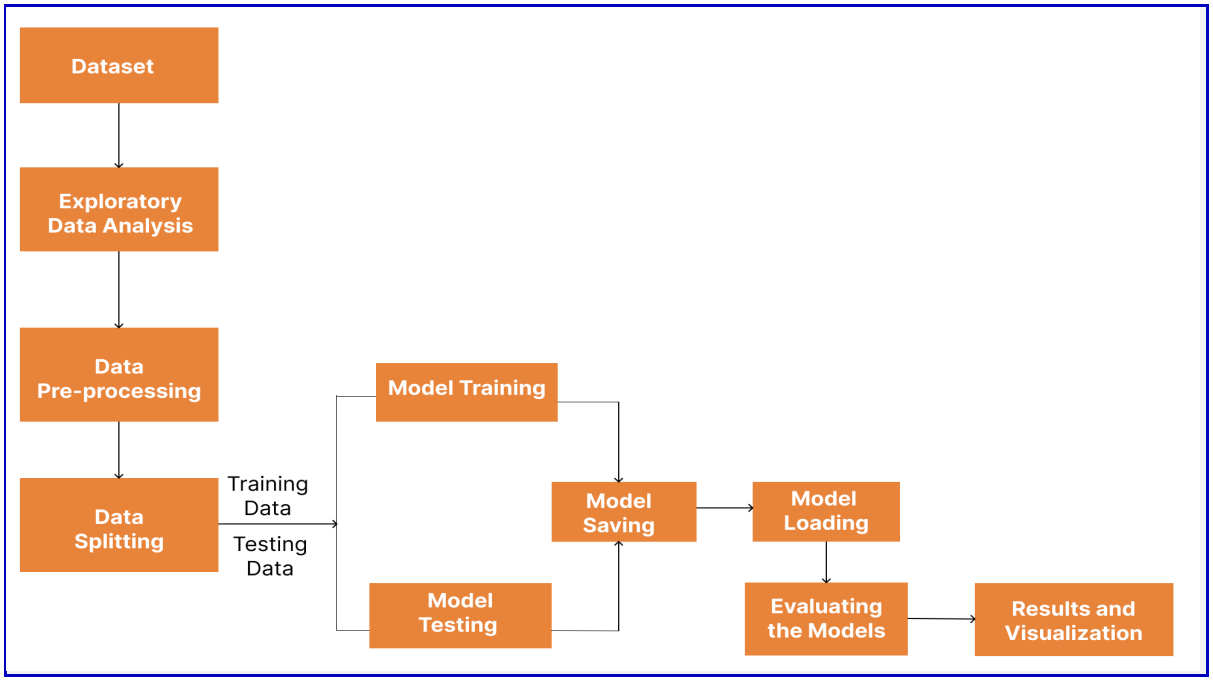

Figure 2: Design Flow for Basketball Performance Prediction Models and Team Efficiency Factors

3.3 Data Segregation and Preprocessing

For predicting player efficiency factors of an NBA player, the availability of a proper dataset is critical. Since the game has a certain unpredictability, each game statistic is required for the study. The dataset chosen for this study is NBA open stats, which contains match results from the past 10 years to the current year. Thus, there is a total of 820 games per season. Each downloaded CSV file contains 540 rows containing match details like rank, position, age, minutes played, player efficiency rating, true shooting percentage, 3-point attempt rate, and 3-point range. Data segregation is an important part of this process; it involves the separation of player performance on certain predefined criteria. The main goal of this segregation is to safeguard sensitive player information and also preserve data integrity. Organising player performance data into logical categories also makes it much easier to search, analyse, and retrieve certain player information. This efficiency is particularly crucial in large basketball datasets[1] where quick and accurate access to relevant player data[2] can significantly impact strategic decision-making processes.

The next stage is data acquisition, which involves missing values, normalising numerical features, encoding categorical variables, and handling any outliers that might affect the accuracy of models. The objective is to produce a clean and standardised dataset that provides the multifaceted aspects of basketball performance, setting the stage for robust predictive modelling.

1 2

3.4 Data Preprocessing and Visualization Used

Data preprocessing is a pivotal step of any data analytic process, it is done to make sure that the historical data that is being used is of great quality and is also able to be used effortlessly with other analytical tools as well. Data Visualisation on the other hand is used to reproduce patterns, understand insights and also to transform complex data into active information. Firstly the raw data is sent for pre-processing, where all the missing values and outliers are cleaned and standardised. The data after this is then sent for training purposes with the help of libraries such as Pandas. Here we have three datasets games.csv, games_details.csv,teams.csv, and ranking.csv, after combining the three datasets and removing the irrelevant cells we get these columns. TEAM_ID, G, W, my_season, FGM, FGA, FG_PCT, FG3M, FG3A, FG3_PCT, FTM, FTA, FT_PCT, OREB, DREB, REB, AST, STL, BLK, TO, PF, PTS, PLUS_MINUS, NEXT_SEASON are selected and saved to preprocessed_data.csv. This data contains only the required details.

The bar chart shown in Figure 3 shows/compares the number of basketball matches that are held in different seasons of the year and also shows the trends related to it. The bar shows that the average number of matches occur in spring which is around 50 to 75. whereas fall shows a further increase from 100 to 125 matches per year and winter on the other side has an average of 120 to 150 matches per year. Fall exhibits a further increase in the number of matches, with an average of 100 to 125 matches per year.

[1] https://www.kaggle.com/datasets/nathanlauga/nba-games/data

[2] https://www.kaggle.com/datasets/justinas/nba-players-data/data

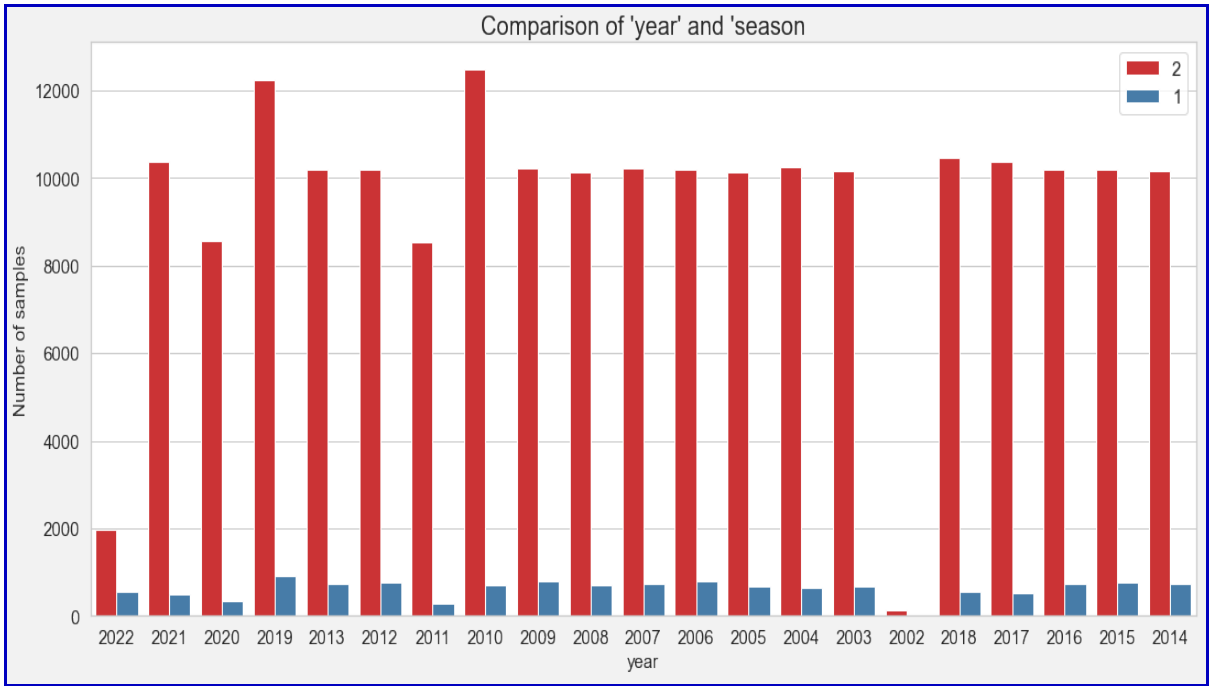

Figure 3: Comparison of year and season

3.5 Model Selection and Configuration

In the pivotal phase of model selection and configuration for the "Basketball Team Efficiency Prediction" project, delving into the characteristics of the basketball performance dataset, the models are chosen in a way that aligns with the nature of the data. Selection includes the utilisation of random forest, linear regression, and decision tree algorithms, each chosen for its specific strengths in handling diverse aspects of basketball dynamics. Beyond mere selection, meticulous model configuration and fine-tuning parameters are done to optimise predictive accuracy and bolster the models' reliability. Embracing ensemble modelling, explore the synergy of combining the predictive power of these algorithms. This phase, integral to the success of the project, establishes a robust foundation for subsequent stages.

3.5.1 Random Forest Regressor

The Random Forest Regressor as shown in Figure 4 is an ensemble learning algorithm used for regression tasks. It operates by constructing a multitude of decision trees during training and outputs the average prediction of the individual trees for regression problems. Each tree in the forest is built on a random subset of the training data and features, introducing diversity and mitigating overfitting. This approach enhances predictive accuracy and robustness. The algorithm is widely applied in various domains, including sports analytics, where it can be utilised to predict basketball performance based on historical data and relevant features.

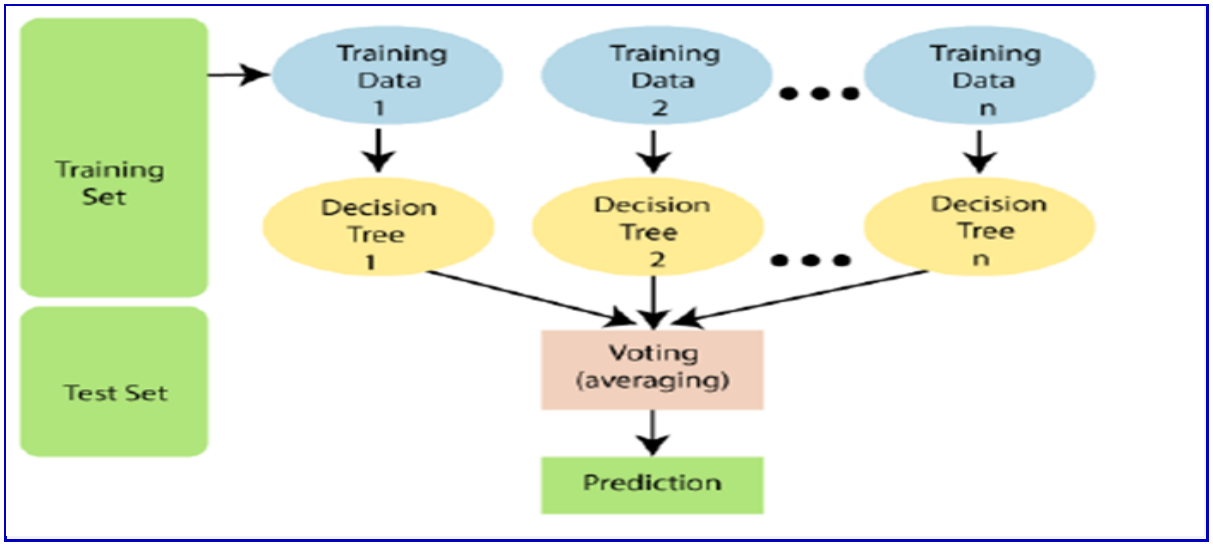

Figure 4: Random Forest Regressor

3.5.2 Linear Regression

Linear regression as shown in figure 5 is a statistical method used to model the relationship between a dependent variable and one or more independent variables by fitting a linear equation to the observed data. The goal is to find the best-fit line that minimises the difference between actual and predicted values, as shown in Figure 18. In the context of basketball performance prediction, linear regression analyses how various factors (independent variables) contribute to the overall performance (dependent variable). The algorithm estimates the coefficients of the equation, allowing for predictions based on input variables and facilitating a straightforward understanding of the relationship between the predictors and the predicted outcome.

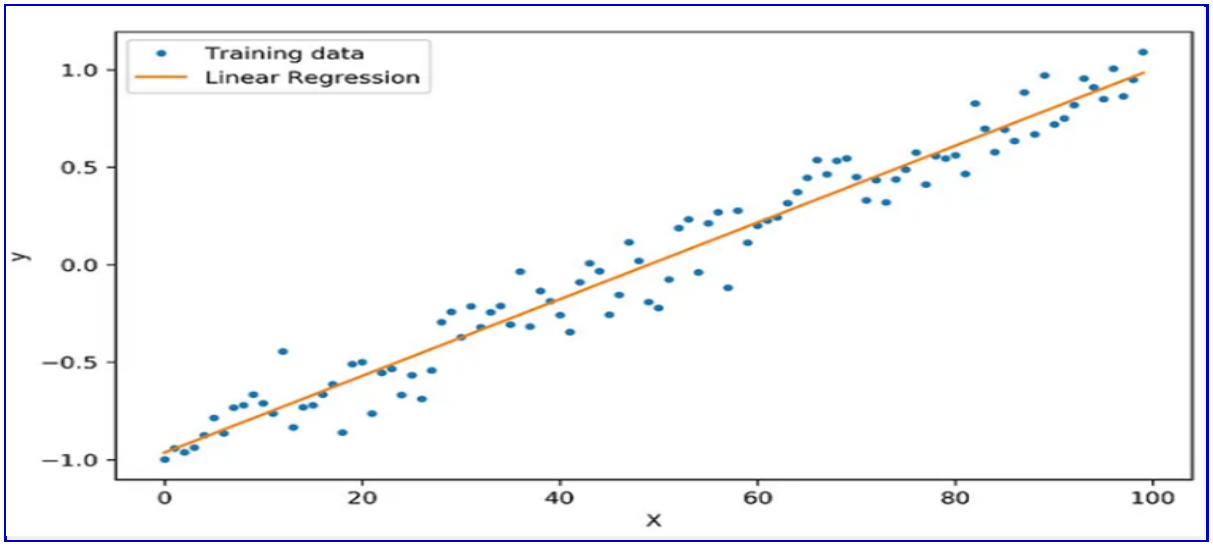

Figure 5: Linear Regression

3.5.3 Decision Tree Regressor

The Decision Tree Regressor as shown in figure 6 is a machine-learning algorithm designed for regression tasks. It creates a tree-like structure by recursively partitioning the data into subsets based on 8 feature conditions. Each leaf node represents a predicted output. Decision trees are adept at capturing complex relationships in the data but are prone to overfitting. Regularisation techniques and ensemble methods, like random forests, are often employed to enhance their generalisation capabilities. In basketball performance prediction, a Decision Tree regression can analyse relevant features to forecast outcomes, making it valuable for understanding and predicting player or team performance.

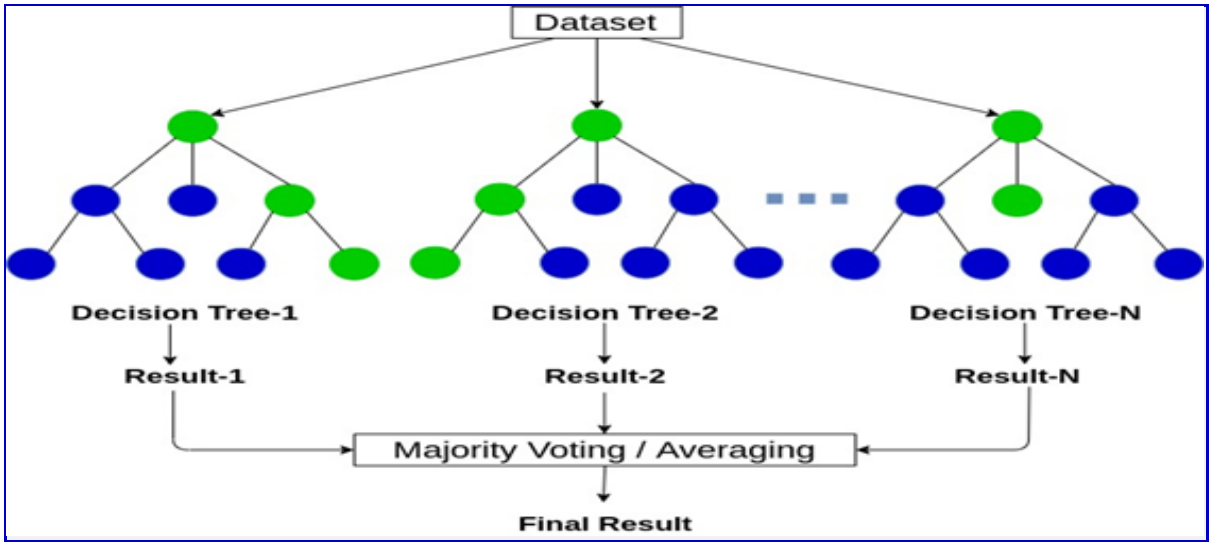

Figure 6: Decision Tree

4 Implementation

4.1 Player Position Analysis

The distribution of players in their respective positions based on the player's calibre is pivotal for the success of the game. Here using the data, visualization techniques are used to analyze and represent the prevalence of different start positions among players. The output graph provides a clear overview, which gives an idea about the composition of player roles within the dataset. This analysis provides a foundational step towards understanding team dynamics and strategic decision-making in the realm of basketball. The 'chart_data' dictionary holds the counts of each start position category.

4.1.1 Player Position on centre

In Figure 7 depicting player performance at the Center position reveals a distribution of samples for each player on the x-axis, ranging from Marc Gasol to Yao Ming. The y-axis, ranging from 0 to 800, illustrates the number of samples. Notably, players such as Karl-Anthony Towns, Dwight Howard, and Jonas Valanciunas have more than 800 samples, indicating a substantial dataset for these players. Conversely, several players,

including Roy Hibbert and ZydrunasIlgauskas, have fewer than 600 samples. The distribution further breaks down with varying sample counts, delineating the performance data for each player in the context of the centre position, offering valuable insights for your basketball team efficiency prediction project.

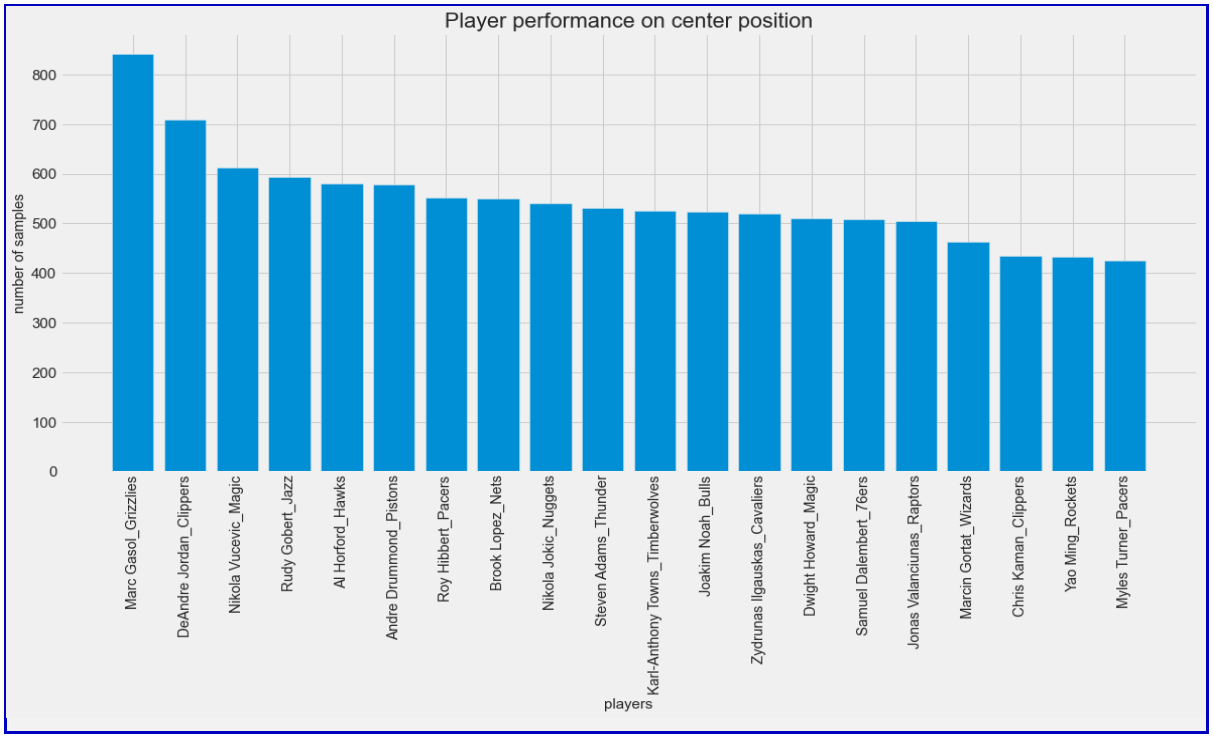

Figure 7: Player Performance on Centre Position

4.1.2 Player Position on Forward

Figure 8 illustrates player performance in the forward position in the NBA, with the x-axis representing players and the y-axis depicting the number of samples. Dirk Nowitzki leads in goals for the Mavericks, trailed by Tim Duncan, LeBron James, Tayshaun Prince, and Giannis Antetokounmpo. The graph reveals a diverse range in forward performance; consistent high performers include Nowitzki and Duncan, while others like Josh Smith and Kevin Durant exhibit more variable results. The overall depiction underscores the intense competition among forwards, showcasing various players with the capability to score significant goals. This comprehensive overview aids in understanding the dynamic landscape of forward player performance in the NBA.

4.1.3 Player Position on Guard

Figure 9 delineates the career performance of 18 guard players, assessing factors such as scoring, assists, and rebounds. Topping the chart are Tony Parker, Dwyane Wade, and Stephen Curry, renowned as some of the greatest guards in history. Notably, Parker secured five NBA championships, Wade earned one with the Miami Heat, and Curry boasted four championships with the Golden State Warriors. Russell Westbrook, Kobe Bryant, and Mike Conley follow, each recognized among the premier guards of their era. The subsequent performers, including Damian Lillard, Klay Thompson, James Harden,

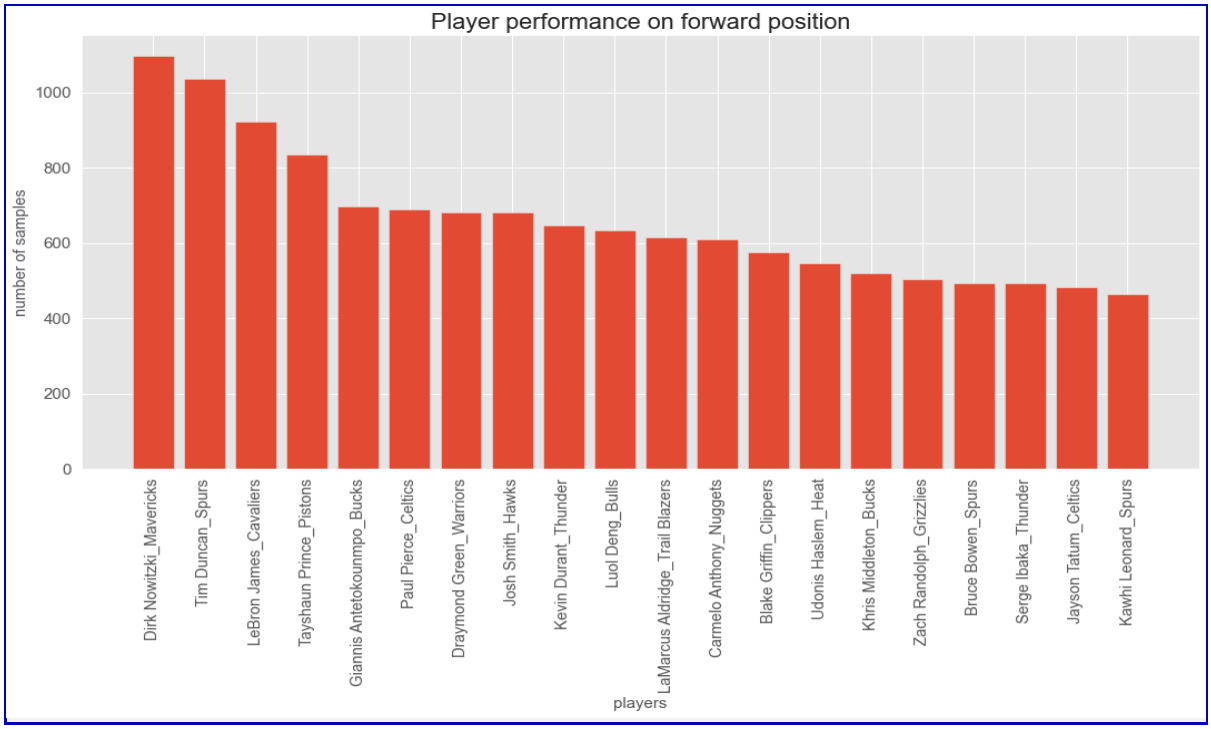

Figure 8: Player Performance on forward Position

DeMarDeRozan, and Kyle Lowry, represent the contemporary elite. The graph collectively portrays the significant impact and competitiveness of guard players, shaping the dynamic landscape of the NBA over decades.

4.1.4 Player performance on Average

Figure 10 illustrates the average performance of the top 20 NBA players, sorted by descending player performance scores. Giannis Antetokounmpo boasts the highest average, while Dirk Nowitzki holds the lowest. The x-axis denotes players and teams, while the y-axis represents average player performance scores. Notably, Western Conference players exhibit higher averages than their Eastern Conference counterparts, reflecting the conference's perceived competitiveness. Additionally, early draft picks and younger players generally correlate with elevated average performance scores, attributed to enhanced talent and athleticism. This graph provides insights into performance trends among top NBA players, shedding light on factors influencing their average performance.

4.2 Influential factor in determining the next Winnings

In the Figure 11 provided graph delves into the pivotal factors shaping the future winnings of NBA basketball players, showcasing a comprehensive analysis of influential metrics. The x-axis meticulously details each factor, while the y-axis quantifies their influential ranks. Notably, PLUS MINUS claims the highest influence rank at 0.27, underscoring its significance in predicting player success. Conversely, FTM holds the lowest influence rank at 0.03. The cascade of factors unfolds, revealing the intricate impact of metrics such as FG3A, BLK, REB, and more, each assigned a specific influence rank. This detailed breakdown elucidates the nuanced landscape of performance indicators affecting players'.

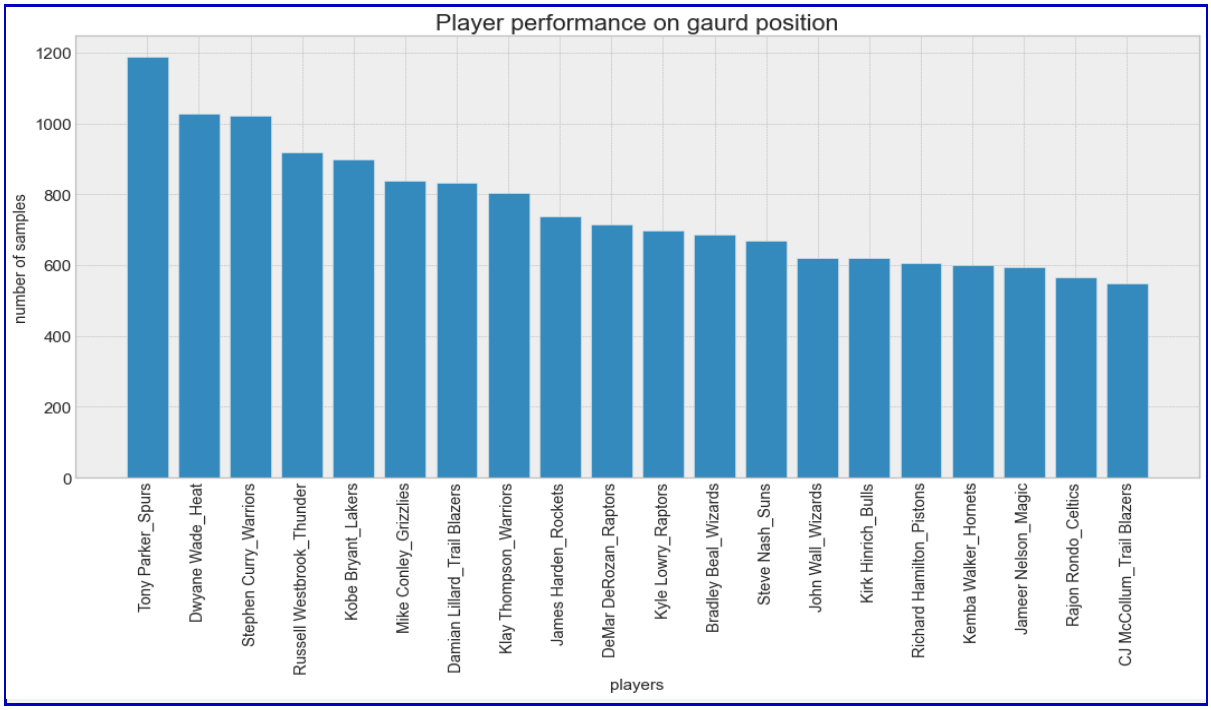

Figure 9: Player Performance on Guard Position

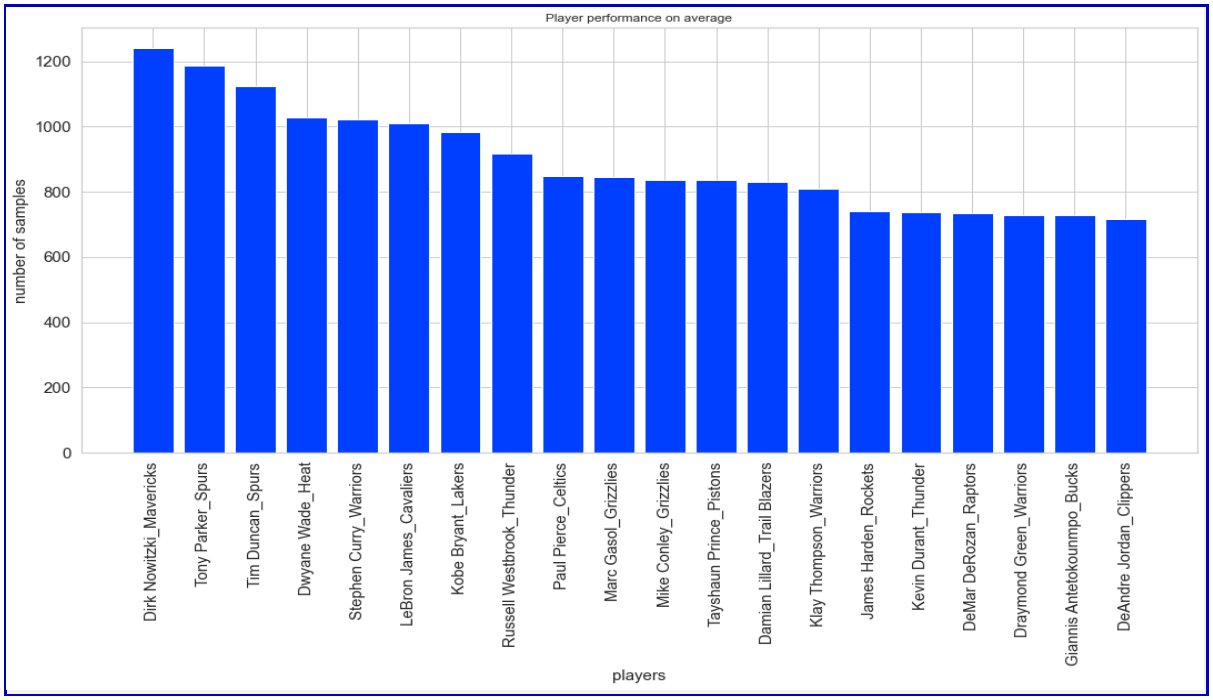

Figure 10: Player performance on Average

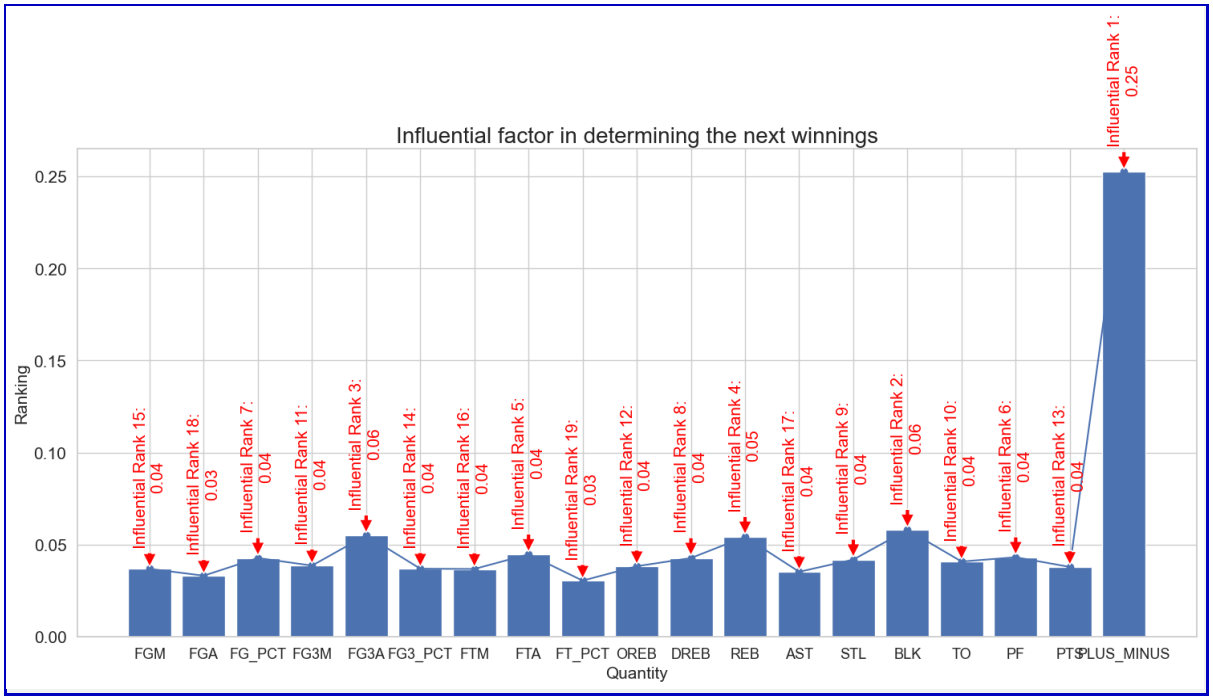

Figure 11: Influential factor in determining the next Winnings

4.3 Model Loading and Implementation

To prepare the dataset for use in machine learning, several pre-processing processes are carried out during the first stage of the project. To confirm that the data acquired should be useful, that is it can be used for analysis purposes, it must undergo a severe data preparation step. The project first uses a sci-kit-learn library for building a predictive model. The algorithms used here are random forest, decision tree, and linear regression, which provide interpretability and predictive power. Here the models are then introduced with certain parameters, and hyperparameter tuning is conducted using GridSearchCV to optimise performance. Training is done, which facilitates the learning of datasets. To marginally increase efficiency, pre-trained models are saved to Python libraries like Pickle, which helps in loading and integration into future applications. The 'GridSearchCV' function provides different combinations of hyperparameter values. This dictionary encompasses parameters such as the number of trees in the forest ('n_estimators'), the maximum depth of the trees ('max_depth'), and the minimum number of samples required to split an internal node ('min_samples_split'), among others.

4.3.1 Hyperparameter Tuning for RandomForestRegressor: A Grid Search Approach

In Figure 12 developed using Python code utilizes the scikit-learn library to perform a grid search for hyperparameter tuning on a RandomForestRegressor. The RandomForestRegressor is a machine-learning model used for regression tasks. The hyperparameters under consideration include the number of trees in the forest ('n_estimators'), the maximum depth of the trees ('max_depth'), the minimum number of samples required to split an internal node ('min_samples_split'), and the minimum number of samples required to be a leaf node ('min_samples_leaf'). The GridSearchCV function is employed for an

exhaustive search over a specified parameter grid to find the optimal combination of hyperparameters. The search is conducted using 5-fold cross-validation ('cv=5'). The scoring metric used is the negative mean squared error 'scoring='neg_mean_squared_error'. Finally, the model is fitted to the training data (X_train and y_train) to determine the best hyperparameter configuration.

```
GridSearchCV(cv=5, estimator=RandomForestRegressor(random_state=42), n_jobs=-1,
             param_grid={'max_depth': [None, 10, 20, 30],
                         'min_samples_leaf': [1, 2, 4],
                         'min_samples_split': [2, 5, 10],
                         'n_estimators': [10, 50, 100, 200]},
             scoring='neg_mean_squared_error')
```

Figure 12: Hyperparameter tuning on Random Forest Regressor

4.3.2 Hyperparameter Tuning for Linear Regressor: A Grid Search Approach

In Figure 13, developed using Python code, the scikitlearn library is employed to perform a grid search for hyperparameter tuning on a Linear Regression model. Linear Regression is a commonly used algorithm for regression tasks. The hyperparameter being tuned in this case is 'fit_intercept,' which determines whether to calculate the intercept for the model. The grid search is conducted using two options for 'fit_intercept': 'False' and 'True'. The 'GridSearchCV' function is employed for an exhaustive search over the specified parameter grid, utilizing 5-fold cross-validation ('cv=5') to evaluate the model's performance. The scoring metric used is the negative mean squared error ('scoring='neg_mean_squared_error''). The model is then fitted to the training data ('X_train' and 'y_train') to identify the optimal 'fit_intercept' setting that minimizes the mean squared error, thereby enhancing the performance of the Linear Regression model

```
GridSearchCV(cv=5, estimator=LinearRegression(),
             param_grid={'fit_intercept': [False, True]},
             scoring='neg_mean_squared_error')
```

Figure 13: Hyperparameter tuning on Linear Regressor

4.3.3 Optimizing Decision Tree Regressor through Hyperparameter Tuning

In Figure 14, crafted through Python coding, the scikit-learn library is harnessed to execute a meticulous grid search, refining hyperparameters for the DecisionTreeRegressora specialized machine-learning model designed for regression tasks. The parameters undergoing refinement encompass 'max_depth' (representing the utmost depth of the tree), 'min_samples_leaf' (the minimum required samples for splitting an internal node) (the minimum samples mandated for a leaf node), 'max_features' (indicating the maximum features deliberated for node division), and 'criterion' (denoting the function gauging split

quality). Employing the GridSearchCV function facilitates a comprehensive exploration across the specified parameter grid, incorporating a 5-fold cross-validation ('cv=5') methodology to meticulously evaluate the model's efficacy. The performance metric employed is the negative mean squared error ('scoring='neg_mean_squared_error"). Subsequently, the model is adeptly applied to the training dataset (X_train and y_train) to pinpoint the optimum hyperparameter amalgamation that systematically minimizes mean squared error, thereby elevating the DecisionTreeRegressor's prognostic proficiency. The inclusion of the 'n_jobs=-1' parameter enables parallelization, expediting computational processes.

```
GridSearchCV(cv=5, estimator=DecisionTreeRegressor(random_state=42), n_jobs=-1,
             param_grid={'criterion': ['mse', 'mae'],
                         'max_depth': [None, 10, 20, 30],
                         'max_features': ['auto', 'sqrt', 'log2'],
                         'min_samples_leaf': [1, 2, 4],
                         'min_samples_split': [2, 5, 10]},
             scoring='neg_mean_squared_error')
```

Figure 14: Hyperparameter tuning on Decession tree Regressor

4.4 Conclusion of Implementation

The implementation was done by training models like Random Forest, Decision Tree, and Linear Regression to get player performance analysis across various positions. Data visualisation was done to get an understanding of the impact of player positions on team dynamics and game strategy. The hyperparameter tuning was carried out to refine the model, hence increasing the predictive power of the models. Pickle function was used to save the models so that they could be loaded easily for seamless application in practical settings. With the models trained and their parameters finely tuned, the research has fulfilled the objective of not just predicting team efficiency but also offering actionable insights into basketball analytics. Thus, the implementation goals of this study have been fully realised.

5 Results and Evaluation

The evaluation of basketball performance prediction models is critical for optimizing team strategies and informing decision-making in the sports sector. Mean Absolute Error (MAE), Mean Squared Error (MSE), and Root Mean Squared Error (RMSE) serve as fundamental indicators of predictive accuracy. These metrics assess the disparities between predicted and actual values, offering insights into the models' overall performance. Within the scope of "Basketball Team Efficiency Prediction," a meticulous examination of model performance unveils Linear Regression's supremacy among Random Forest Regressor and Decision Tree Regressor. The metrics, encompassing Mean Squared Error, Root Mean Squared Error, and Mean Absolute Error, uniformly validate Linear Regression's exceptional accuracy in forecasting team efficiency.

5.1 Model Comparison through Mean Squared Error (MSE) Loss Analysis

In Figure 15, the examination of Mean Squared Error (MSE) loss across three models reveals distinctive insights. The Random Forest Regressor shows an MSE of 146.666, Linear Regression records 130.736, and the Decision Tree Regressor demonstrates 207.017. MSE loss serves as a critical metric for evaluating the accuracy of regression models in predicting the target variable, with diminished values indicating superior predictive performance. Significantly, the graph illustrates that Linear Regression attains the lowest MSE loss, outshining both the Random Forest and Decision Tree Regressors. This underscores that Linear Regression excels in predicting the target variable compared to the other models.

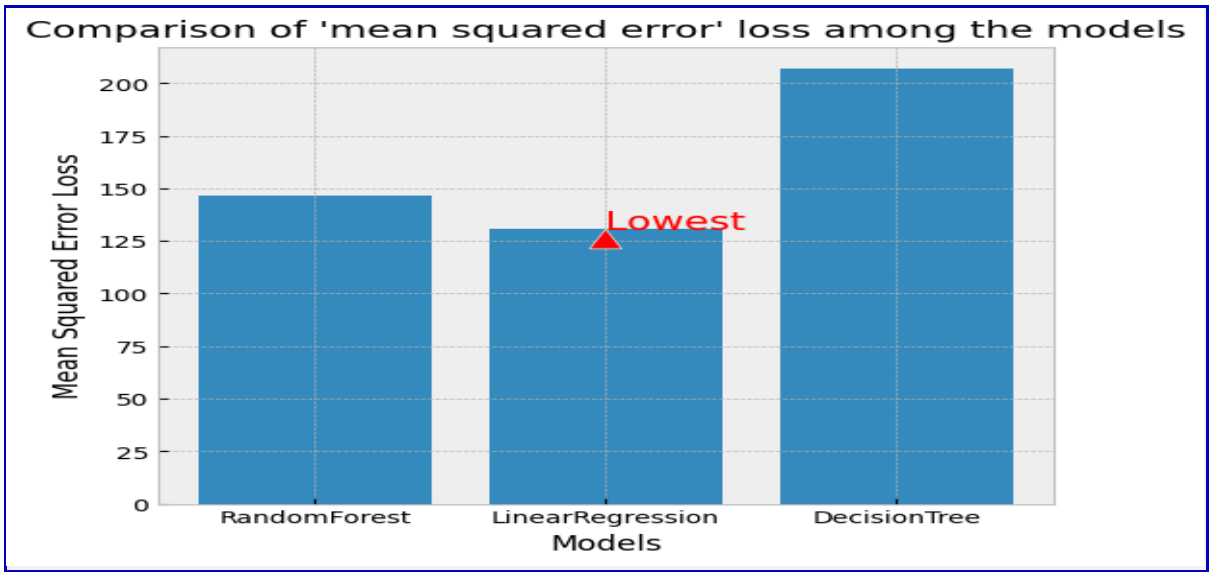

Figure 15: Mean Squared Error

5.2 Model Comparison through Mean Absolute Error (RMSE) Analysis

In Figure 16, the analysis of Root Mean Squared Error (RMSE) loss across three models reveals distinctive patterns. The Random Forest Regressor shows an RMSE of 146.666, Linear Regression records 130.736, and the Decision Tree Regressor demonstrates 207.017. RMSE is a crucial metric for assessing the precision of regression models, with lower values denoting superior predictive accuracy. Notably, Linear Regression emerges with the lowest RMSE loss, followed by the Random Forest Regressor and the Decision Tree Regressor. This implies that Linear Regression excels in minimizing prediction errors, further affirming its superior performance in achieving precision and reliability compared to the other two models in the realm of regression analysis.

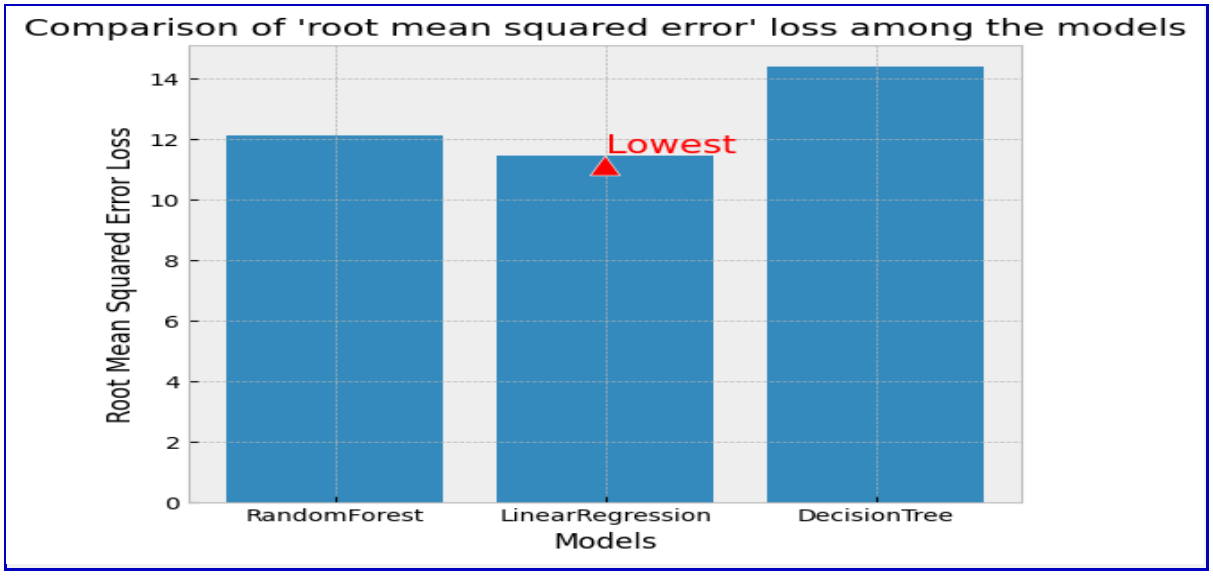

Figure 16: Root Mean Squared Error

5.3 Model Comparison through Mean Absolute Error (MAE) Analysis

In Figure 17, the examination of Mean Absolute Error (MAE) loss across three models reveals distinct performance trends. The Random Forest Regressor shows an MAE of 9.543, Linear Regression records 9.385, and the Decision Tree Regressor demonstrates 11.421. MAE is a crucial metric for evaluating the accuracy of regression models, with lower values indicating superior predictive precision. Notably, Linear Regression exhibits the smallest MAE loss, followed by the Random Forest Regressor and the Decision Tree Regressor. This suggests that Linear Regression excels in minimizing the absolute differences between predicted and actual values, establishing it as the model with the highest accuracy and precision in predicting the target variable among the three considered models.

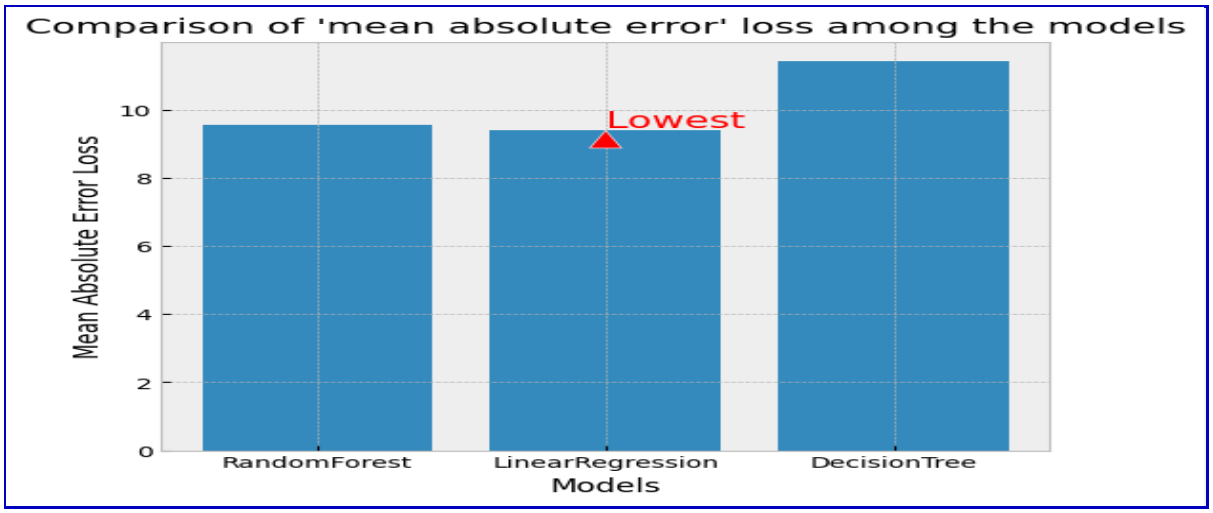

Figure 17: Mean Absolute Error

5.4 Cases Studies

5.4.1 Case 1: 2017 Clippers Basketball Team Performance Prediction

While undergoing a detailed examination to predict the performance factors in the basketball game, two case studies are taken, In the case of the study shown in Table 1, the focus was mainly on the performance done by the Clippers in the year 2017, and the model used for the prediction (regression, decision tree and random forest), produced reasonable accuracy, which nearly aligned to the actual values in terms of wins and efficiency. From this, it is understandable that the model can produce its output within a specific time. Case Study 1 predicts the 2017 performance of the "Clippers" basketball team. Actual wins were 42, while the model predicted 41. True efficiency stood at 51.22%, with the model forecasting 50.0%.

Team	True Wins	Predicted Wins	True Efficiency (%)	Predicted Efficiency (%)
Clippers	42	41	51.21%	50.0%

Table 1: Clippers Basketball Team Performance

5.4.2 Case 2: 2021 Wizards Basketball Team Performance Prediction

Whereas, in the case study shown in Table 2, the emphasis lies on forecasting key performance indicators. A comparative analysis between actual and predicted values is conducted to assess the model's accuracy in anticipating specific metrics during the designated period, with a focus on the Wizards. Case Study 2 delves into forecasting the 2021 performance of the basketball team "Wizards." Actual wins for the Wizards amounted to 35, while the model predicted 31 wins. In terms of efficiency, the true value stood at 42.68%, contrasting with the model's prediction of 37.80%.

Team	True Wins	Predicted Wins	True Efficiency (%)	Predicted Efficiency (%)
Wizards	35	31	42.68%	37.80%

Table 2: Wizards Basketball Team Performance

5.5 Discussion

Here, from the above, we can infer that linear regression stands out with the rest of the models as all the others showed comparatively low performance, all the metrics shown provide the reliability of the models used. such as Case Study 1, which focuses on the Clippers game shows how it aligns between the true values and the accuracy of each prediction. By leveraging the model's efficacy, it can be used for various sports purposes. Even though there are limitations, such as the limited availability of data.

The comparison done with different models shows strengths and weaknesses. The discussion puts more strength on coaches, analysts, and researchers. Future researchers can

direct attention towards ensemble techniques, thereby increasing the predictive capacity within the field of basketball. The metrics attained suggest the need for sophistication in modelling to achieve accuracy in predicting basketball team efficiency.
In conclusion, through the use of algorithms such as linear regression, random forest, and decision trees, it is clear that linear regression in particular showcases accurate predictions. By comparing this outcome with a real-world basketball game, it shows how the model is effective in analysing the results. The study successfully integrated performance metrics into the framework of basketball analytics. With the help of the models that were studied, the research offers practical support for basketball teams to make data-driven decisions.

6 Conclusion and Future Work

The outcomes of this research delved into the application of machine learning to predict team efficiency based on performance metrics, aiming to bridge existing gaps in basketball statistics. Through a critical examination of various prediction models, the study highlighted the efficacy of Linear Regression in optimising predictive accuracy and precision. The values from Mean Squared Error (MSE) show that Linear Regression performed better than Random Forest and Decision Tree Regressor, with the lowest MSE of 130.736. The values from the Root Mean Square Error(RMSE) and also from the Mean Absolute Error(MAE) output an updated value for the Linear Regression, as they show lower values when compared with the other two models. With an RMSE of 130.736 and an MAE of 9.385, Linear Regression exhibited a remarkable ability to minimise prediction errors and absolute differences between predicted and actual values. The research done not only contributes to the basketball game but also provides practical support to increase team efficiency.Finally, by including real-time data and adding the study to several sports, we can gain an understanding of performance dynamics.

7 Acknowledgement

I would like to express my sincere gratitude to Dr Catherine Mulwa for her supervision, guidance and support throughout the research. Her patience, motivation and immense knowledge helped me at the time of the research. I would like to acknowledge my mother, father and Colleagues for their support and motivation during the coursework.

References

Ackland, T. R., Schreiner, A. B. and Kerr, D. A. (1997). Absolute size and proportionality characteristics of world championship female basketball players, *Journal of Sports Sciences* **15**(5): 485–490.

Bray, C. D. and Whaley, D. E. (2001). Team cohesion, effort, and objective individual performance of high school basketball players, *The Sport Psychologist* **15**(3): 260–275.

Cole, T. A. (2021). Exploring team effectiveness of a collegiate women's basketball team using kirton's adaption-innovation theory.

Fu, Y. and Stasko, J. (2022). Supporting data-driven basketball journalism through interactive visualization, *Proceedings of the 2022 CHI Conference on Human Factors in Computing Systems*, pp. 1–17.

Javadpour, L., Blakeslee, J., Khazaeli, M. and Schroeder, P. (2022). Optimizing the best play in basketball using deep learning, *Journal of Sports Analytics* **8**(1): 1–7.

Ji, R. (2020). Research on basketball shooting action based on image feature extraction and machine learning, *IEEE Access* **8**: 138743–138751.

Li, B. and Xu, X. (2021). Application of artificial intelligence in basketball sport, *Journal of Education, Health and Sport* **11**(7): 54–67.

Martínez, J. A. (2020). Predicting per game performance through per minute performance in basketball, *Journal of Physical Education and Sport* **20**(2): 686–689.

Miljković, D., Gajić, L., Kovačević, A. and Konjović, Z. (2010). The use of data mining for basketball matches outcomes prediction, *IEEE 8th international symposium on intelligent systems and informatics*, IEEE, pp. 309–312.

Naismith, J. (1996). *Basketball: Its origin and development*, U of Nebraska Press.

Sampaio, J. et al. (2015). Exploring game performance in the national basketball association using player tracking data, *PloS one* **10**(7): e0132894.

Sarlis, V., Chatziilias, V., Tjortjis, C. and Mandalidis, D. (2021). A data science approach analysing the impact of injuries on basketball player and team performance, *Information Systems* **99**: 101750.

Sarlis, V. and Tjortjis, C. (2020). Sports analytics—evaluation of basketball players and team performance, *Information Systems* **93**: 101562.

Senbel, S., Sharma, S., Raval, M. S., Taber, C., Nolan, J., Artan, N. S., Ezzeddine, D. and Kaya, T. (2022). Impact of sleep and training on game performance and injury in division-1 women's basketball amidst the pandemic, *Ieee Access* **10**: 15516–15527.

Sharef, N. M., Mustapha, A., Azmi, M. B. N. and Nordin, R. (2020). Basketball players performance analytic as experiential learning approach in teaching undergraduate data science course, *2020 International Conference on Advancement in Data Science, E-learning and Information Systems (ICADEIS)*, IEEE, pp. 1–7.

Terner, Z. and Franks, A. (2021). Modeling player and team performance in basketball, *Annual Review of Statistics and Its Application* **8**: 1–23.

Tian, C., De Silva, V., Caine, M. and Swanson, S. (2019). Use of machine learning to automate the identification of basketball strategies using whole team player tracking data, *Applied Sciences* **10**(1): 24.

Vinué, G. and Epifanio, I. (2019). Forecasting basketball players' performance using sparse functional data, *Statistical Analysis and Data Mining: The ASA Data Science Journal* **12**(6): 534–547.

Yoon, Y., Hwang, H., Choi, Y., Joo, M., Oh, H., Park, I., Lee, K.-H. and Hwang, J.-H. (2019). Analyzing basketball movements and pass relationships using realtime object tracking techniques based on deep learning, *IEEE Access* **7**: 56564–56576.