Shooting Stars: Predicting the NBA Gems of Tomorrow

Neh Desai[1][0009−0009−4441−6677], Chirag Rathi[1][0000−0003−0499−6688], and Ritu Chaturvedi[1][0000−0003−0233−674X]

School of Computer Science, University of Guelph
ndesai04@uoguelph.ca
crathi@uoguelph.ca
chaturvr@uoguelph.ca

Abstract. Predicting individual player performance, particularly scoring metrics like points per game (PPG), has become a significant area of research in sports analytics, driven by advances in data collection and machine learning. Previous studies primarily emphasized team performance or used aggregated data without in-depth feature optimization, leaving gaps in accurately forecasting individual player metrics. This paper addresses these limitations by introducing the Correlation-Optimized NBA Scoring Estimator (CONSE) model, designed to predict NBA player performance based on statistical data spanning ten NBA seasons (2013–2023). Our approach includes rigorous data preprocessing, extensive exploratory data analysis (EDA), and leverages correlation-based feature selection to enhance interpretability and predictive accuracy. Multiple regression models, including Multiple Linear Regression, K-Nearest Neighbors (KNN), Decision Tree, and Random Forest, were evaluated within the CONSE model framework using R-squared and Mean Absolute Error along with cross-validation. Random Forest Regression demonstrated superior performance due to its robustness against overfitting and outliers. Our findings provide valuable insights for coaches, analysts, and enthusiasts aiming to identify future NBA stars using a data-driven approach.

Keywords: Sports Analytics · Machine Learning · NBA Player Performance Modeling · Correlation Analysis · Feature Engineering · Regression

1 Introduction

Basketball analytics has experienced significant evolution over the past decade, driven by the growth in computational power, advanced statistical methodologies, and increased availability of granular player data. This evolution has shifted the analytical focus from traditional metrics toward more sophisticated predictive analytics capable of forecasting individual player performance. Accurately predicting player scoring capabilities, particularly points per game (PPG), is highly valuable for coaching decisions, player development, team strategies, and

even fan engagement. However, traditional methodologies often emphasize team performance, neglecting to focus explicitly on individual contributions.

This paper addresses these limitations by proposing a refined model, namely the CONSE model, centered explicitly on individual NBA player performance prediction using historical statistical data collected from Basketball-Reference and official NBA sources. We compiled data from ten recent NBA seasons (2013-14 through 2022-23), capturing a robust sample of player statistics across offensive, defensive, and advanced metrics. To enhance model interpretability and avoid biases, our preprocessing methods included addressing missing values, combining data for players traded mid-season, and converting categorical variables into numerical formats.

Following preprocessing, exploratory data analysis (EDA) played a critical role in identifying key performance predictors through correlation analysis, significantly reducing multicollinearity issues by filtering out highly correlated features. Consequently, we retained 18 meaningful predictors to improve model generalization and interpretability. The predictive capabilities of four distinct regression models, Multiple Linear Regression, K-Nearest Neighbors, Decision Tree, and Random Forest were evaluated. Model effectiveness was assessed using R-squared, Mean absolute error and cross-validation to confirm reliability and minimize overfitting.

Our findings suggest that Random Forest Regression, an ensemble approach, offered the best predictive performance within the CONSE model framework due to its robustness and ability to handle the intrinsic variability of NBA player statistics. This research not only enhances the understanding of critical predictive factors affecting player performance but also presents a robust analytical framework adaptable to other domains such as the WNBA.

2 Literature Review

The prediction of NBA player performance has been a growing focus in sports analytics, with researchers employing various statistical and machine learning models to forecast future stars. Modeling and forecasting NBA game outcomes, integrating heteroscedasticity and dynamic state-space models to capture time-variant team strengths, has been explored [8]. The study found that while betting markets are difficult to outperform, statistical models incorporating time-dependent factors can enhance predictive accuracy. These findings highlight the importance of accounting for evolving player performance over multiple seasons.

Advancements in player tracking technology have further refined predictive analytics in basketball. SportVU data has been leveraged to analyze player movement patterns and offensive-defensive strategies using deep learning techniques [14]. By employing convolutional neural networks to classify plays, the study demonstrated how machine learning can extract meaningful insights from raw movement data. This shift from traditional statistical methods to automated pattern recognition has enabled a more nuanced understanding of player potential and development.

Machine learning models have also been applied to predict game outcomes based on individual and team performance metrics. Logistic regression, deep neural networks, and random forests have been compared in predicting NBA game results, identifying field goal percentage, rebounds, and assists as key determinants [16]. The study reinforced the effectiveness of data-driven feature selection, suggesting that specific performance indicators can be strong predictors of long-term success. Multiple classification models, including Gaussian Naïve Bayes and XGBoost, have also been tested to evaluate their performance, emphasizing the importance of feature engineering and hyper-parameter tuning in improving model accuracy [7]. These studies collectively demonstrate the growing role of machine learning in identifying high-potential players and optimizing prediction models for sports analytics.

However, none of the previously mentioned methods focus solely on individual player performance metrics but rather emphasize team dynamics. Additionally, they do not optimize features based on correlation analysis to determine which features should be inputted into models. In this paper, we propose a methodology through our CONSE model that focuses on predicting individual player performance metrics, mainly points per game, by preprocessing the data in a way that eliminates biases among the features, making the results more interpretable and reliable.

3 Methodology

3.1 Data

Data Collection The required data was compiled using Basketball-Reference [13] and the official NBA statistics webpage [9]. These sources were cross-referenced to ensure consistency and eliminate any potential biases that could impact our model's performance. The dataset covered ten NBA seasons, from 2013-14 to 2022-23, including only regular season games in which active players participated. We selected the most recent ten seasons to ensure accuracy and relevance, as maintaining consistency within a single decade minimizes the impact of changes in rules and regulations that could otherwise skew the results. Each season featured approximately 400-500 active players, which resulted in a total of 5,218 rows of data over the analyzed period.

Data Features Initially, 31 variables were collected, encompassing a range of player statistics, some of which are illustrated along the x- and y-axes of Figure 1. These included offensive metrics like shooting percentage, shots made, and shots attempted, as well as defensive statistics such as blocks, steals, and rebounds. General statistics were also gathered, including the number of minutes played, games played, and the primary position, which was determined by the position played in the majority of games. Additionally, advanced metrics like effective field goal percentage were included to provide deeper insights.

Data Preprocessing The raw data, which had 31 features, was refined by filtering out irrelevant features, such as the team a player played for, as the focus of this project is on individual player performance rather than team dynamics. Additionally, some features were renamed to enhance the dataset's readability so that it could be more clear as to what each feature described or represented. To address duplicates, particularly for players who were traded or bought out, which resulted in them playing for multiple teams within a season, their separate entries were combined into a single row that summed their totals. This ensured that the data remained reliable and unbiased. Missing values in features such as three-point, two-point, or field goal percentage columns were handled by applying a simple mathematical calculation which involved dividing the number of specific shots made by the number of specific shots attempted to maintain consistency across the dataset. For example, the three-point percentage was computed by dividing the number of three-pointers made by the number of three-pointers attempted in total. Finally, categorical data, such as player positions (e.g., Point Guard, Shooting Guard, Small Forward etc), were transformed into an integer encoding to enable effective use by the future models. After preprocessing the data, 18 features remained (excluding the target variable, points per game, or PPG), as shown in Figure 2, along with a total of 5,218 rows available for analysis.

3.2 Feature Selection With the Aid of Exploratory Data Analysis

After preparing the data, an exploratory data analysis (EDA) was conducted to help uncover patterns, trends, and outliers, as well as to examine the relationships between variables to identify any meaningful insights. This analysis included evaluating statistical measures such as mean, median, mode, and standard deviation, as well as examining the distributions of various features. However, one of the most valuable aspects of this stage was the correlation analysis, which was visualized using heat maps with the features as shown in Figure 1. A correlation analysis is a statistical method used to determine whether a relationship exists between two sets of variables, helping to identify significant connections between them [6]. It is typically conducted using a heat map, which is a data visualization tool that presents data in a grid format, where colors represent different levels of magnitude. This allows for quick identification of patterns or anomalies within a dataset [5]. The heat map was particularly insightful, revealing that some features were highly correlated with the target variable, points per game (PPG). For example, field goals made (FG) and field goals attempted (FGA) had correlations of 0.99 and 0.98, respectively, with PPG as seen in Figure 1. This high multicollinearity posed a problem, as it could obscure the importance of other individual features and lead to a model that overly focused on those with higher weights. In a heat map, a weight is a value assigned to each data point within a square, determining its influence. The higher the weight, the darker the color [4]. To address this issue, we removed all features with a correlation of 0.8 or higher with the target variable, as a correlation coefficient of

this magnitude between two regressors is considered a serious problem [12]. Removing these features would help achieve a clearer interpretation of the model's decision-making process and ensure the model would be more robust, with a better ability to generalize to new data without over or under-fitting. After this refinement, we were left with 18 features, excluding the target variable, out of the original 31, as shown in Figure 2. Most of the removed features were filtered out due to high correlation, but all non-numeric features irrelevant to the target variable, such as season identifiers or player indices, were also removed as they were solely for organizing the datasets and had no predictive value in determining any performance metrics. The sequence of data-related steps as outlined in Section 3.1 and mentioned earlier in this section are illustrated in the first row of Figure 3, highlighted by the gray-shaded blocks.

4 Modeling Information

4.1 Model Implementation and Evaluation

The goal of our models is to predict a continuous numerical value related to a specific performance metric for an individual player, so we employed regression models throughout this project. There are various types of regression models, each with their own strengths and weaknesses, but we focused on four main types. We began with a multiple linear regression model to predict our dependent variable (PPG in this case) based on the independent variables or features we provided. We then progressed to more complex models, such as Decision Tree regression. These models are further explained in Section 4.2 and are illustrated in the second row of Figure 3, which corresponds to the modeling stage highlighted by the blue-shaded boxes.

Initially, all of the models were evaluated on the training dataset, which had a shape of (5218, 18), representing data from the past ten seasons, where each row contained values for a player across the 18 selected features. The coefficient of determination (R^2) and Mean Absolute Error (MAE), along with the 5-fold cross-validation technique, were used to assess baseline model performance. R^2 measures the proportion of variance explained by the model, with higher values indicating better fit [2]. On the other hand, MAE calculates the average absolute difference between predicted and actual values [11]. Together, these metrics offer complementary insights into how well the model captures variability and how close its predictions are to the actual values as shown in Table 1. To further ensure reliability and prevent overfitting, we also performed 5-fold cross-validation. This helped confirm that the models were consistent and capable of generalizing patterns across different parts of the dataset by training on multiple subsets and evaluating performance on unseen data in each fold, which is also shown in Table 1 [3]. Following the initial evaluation, hyperparameter tuning was performed using RandomizedSearchCV to efficiently explore the hyperparameter space and identify optimal settings for each model. After several iterations, the best-performing configurations were selected, and the tuned models were evaluated on a separate validation dataset with a shape of (572, 18),

which included player data from the most recent 2023–24 NBA regular season at the time of writing. The final post-tuning evaluation used the R^2 and MAE metrics, as discussed earlier, to better assess generalization performance. These model evaluation steps are also visually represented in the final row of Figure 3, shown in the green-shaded blocks.

4.2 Explanation of Models

Multiple Linear Regression Multiple Linear Regression (MLR) is a statistical technique that models the relationship between a dependent variable, which in our case is points per game (PPG) and multiple independent variables. The term multiple refers to the inclusion of more than one predictor which allows the model to estimate the combined effect of these variables on the outcome [15]. This model was chosen as our baseline because player statistics depend on several interacting factors, such as playing time and shooting efficiency, which we had previously collected in our dataset. Another reason for selecting MLR is its ability to quantify the influence of each factor on a player's performance. Using this approach we can further determine how much each variable contributes to the final prediction, making it a valuable tool for analyzing player performance trends and identifying key factors that drive a player's success.

KNN Regression Although the K-nearest neighbours (KNN) algorithm is most commonly associated with classification tasks, it can also be used for regression. This type of regression makes minimal assumptions about the underlying data and its distribution and is a non-parametric method for predicting continuous values [15]. This is one reason why this model was chosen, as player statistics often exhibit complex relationships. Another reason for choosing this model was due to its ability to make locality-based predictions. Since KNN relies on nearby data points, it is particularly useful in the NBA, where groups of players exhibit similar playstyles. This allows the model to identify clusters of comparable players and predict a new player's statistics based on those with similar characteristics.

Decision Tree Regression Decision Tree Regression is a machine learning technique that predicts continuous values by splitting data into smaller subsets based on features, continuing until leaf nodes provide the final predictions [10]. Since this type of model makes it easy to interpret results through its tree-like structure, it was chosen to help us visualize and understand the decision-making process while identifying the key factors that most strongly influence player performance, making it a valuable tool for data-driven decision-making. Additionally, Decision Tree Regression is more robust to outliers because it partitions features into distinct regions, meaning extreme values have a minimal impact on overall performance [10]. This was particularly important for our NBA dataset, as some elite players exhibit exceptional statistical outliers, such as extremely high scoring averages or rebound counts that do not align with league averages.

By using decision trees, we can ensure that our model effectively handles such variations without being overly influenced by extreme cases.

Random Forest Regression Random forest regression extends the concept of decision trees by constructing multiple decision trees and combining their predictions to enhance robustness and accuracy. It is an ensemble learning method where individual decision trees serve as the fundamental building blocks [1]. The multiple trees are generated using a technique called bootstrapping, in which random samples from the original training data are drawn with replacement, which allows the same data points to be selected multiple times. This process helps to create a diverse set of trees, which helps improve the model's overall performance and reduces the risk of overfitting [1]. We selected random forest regression for our NBA dataset because it naturally builds upon the strong results obtained from decision trees, as discussed earlier. By leveraging multiple trees, it enhances predictive performance, mitigates overfitting, and handles outliers more effectively by averaging predictions across trees rather than relying on a single one, making it more robust to noise [1]. This is particularly important given the presence of extreme values in elite player statistics, as mentioned earlier. A summary of the key details regarding all of the models can be found in Table 2

5 Conclusion and Future Work

This study aims to enhance sports analytics by utilizing artificial intelligence and machine learning to predict NBA player performance metrics such as points per game. We compiled a dataset from Basketball-Reference and official NBA statistics covering ten seasons, and employed various regression models-Multiple Linear Regression, K-Nearest Neighbors, Decision Tree, and Random Forest-within the proposed CONSE model framework, focusing on their capability to generalize and predict trends accurately. An extensive data analysis was key in identifying and excluding highly correlated variables that could affect model accuracy, an aspect that has not been widely emphasized in similar studies. Among the tested models, Random Forest Regression showed superior performance due to its ability to handle overfitting and statistical outliers. Our results demonstrate the potential of machine learning in accurately predicting and analyzing NBA player performance, providing valuable insights for analysts, coaches, and fans.

In future work, we aim to expand our predictive modeling by incorporating social media data we have gathered related to NBA players. This dataset includes features such as post impressions, engagement metrics, and sentiment trends throughout the season. We plan to analyze these patterns to identify potential correlations between a player's online presence and their on-court performance. By doing so, we seek to determine whether fluctuations in social media activity and public sentiment have a measurable impact on scoring behavior. Incorporating this off-court dimension could provide a more holistic view of performance

prediction and offer insights into the psychological and external factors that may influence player outcomes. Future research may also explore additional factors not considered in this study, such as player injuries, team dynamics, and advanced play-by-play data, which could significantly impact player performance. Furthermore, deep learning techniques or other advanced models could be investigated to determine whether they further enhance predictive accuracy and interoperability.

6 Figures and Tables

This section includes all tables and figures used in the paper. It consists of two tables (Tables 1 and 2) and three figures (Figures 1, 2, and 3).

Table 1. Cross-Validation scores for the four main models, including their average, R^2 and MAE metrics

Model	**Value 1**	**Value 2**	**Value 3**	**Value 4**	**Value 5**	**Average**	**R^2**	**MAE**
LR	0.8788	0.8653	0.8764	0.8768	0.8757	0.8746	0.8816	1.4813
KNNR	0.8018	0.8019	0.7885	0.7766	0.8045	0.7947	0.7815	1.9707
DTR	0.7902	0.8224	0.8259	0.8265	0.8148	0.8160	0.7974	1.8701
RFR	0.8862	0.8791	0.8779	0.8953	0.9033	0.8884	0.9011	1.3324

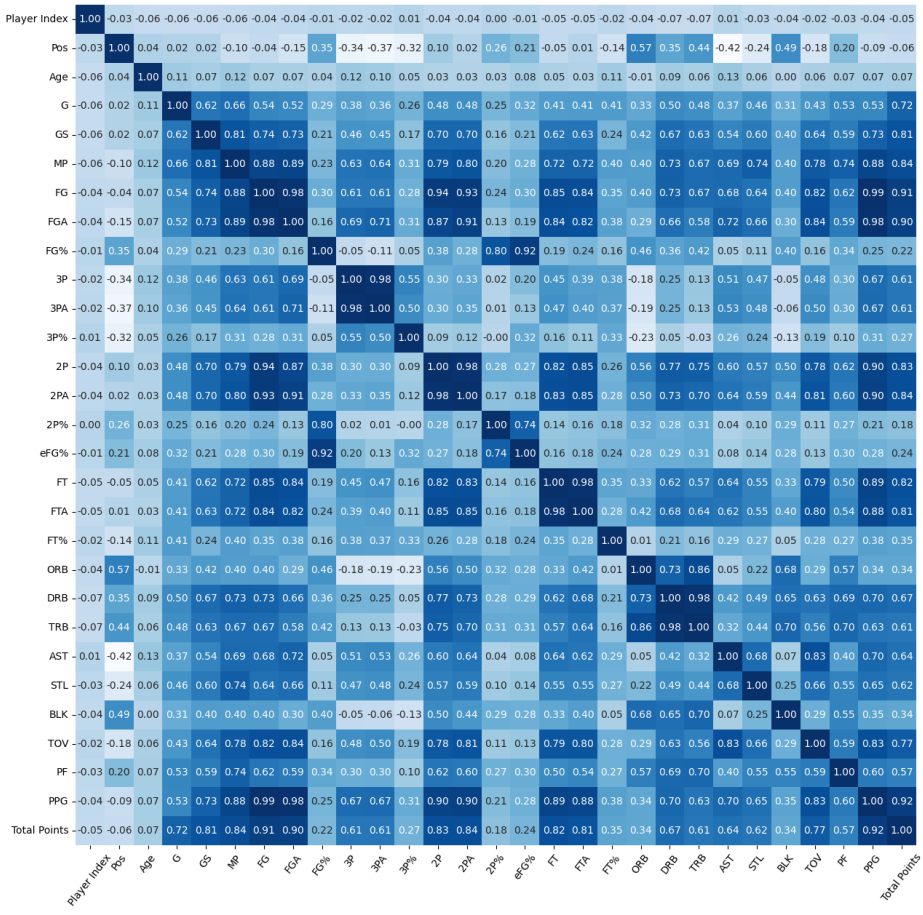

Fig. 1. Initial heat map with all features included

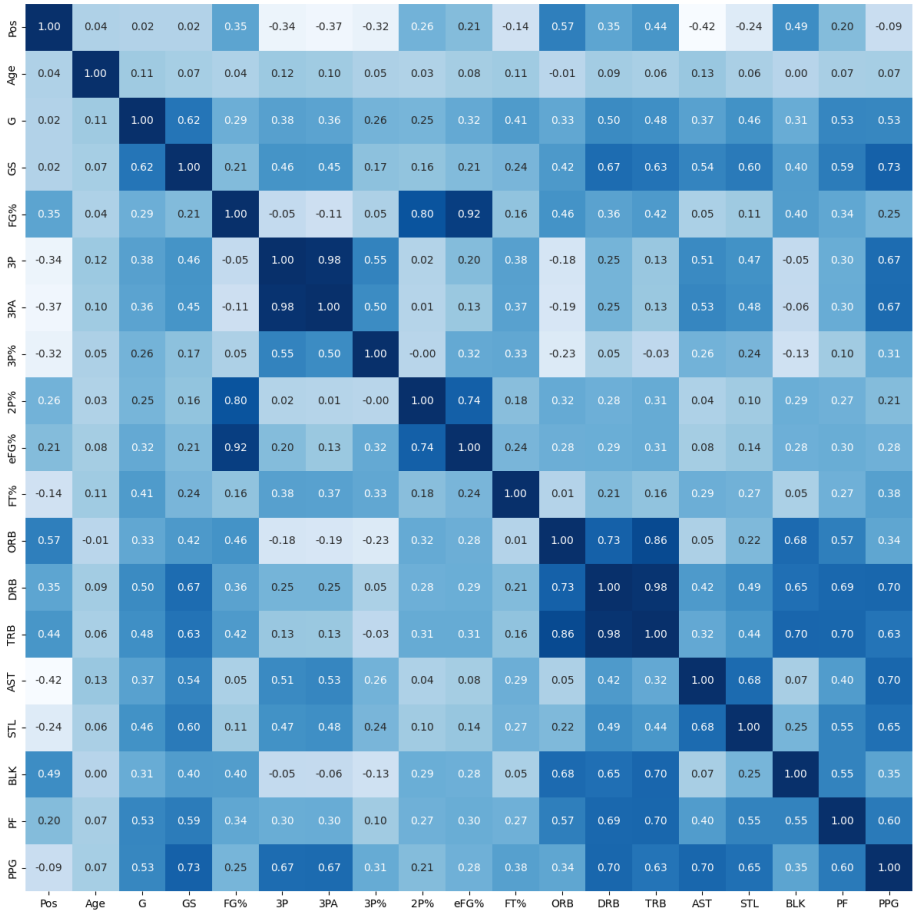

Fig. 2. Modified heat map after removing highly correlated features

Table 2. Summary of models and their reasoning

Model	Reasons/Benefits
MLR	– Simple and interpretable baseline model. – The domain of basketball relies on several interacting factors, which MLR can help capture. – Provides clear insights into feature importance, aiding performance analysis.
KNNR	– Non-parametric method which can help capture the complex relationships between player statistics. – Locality-based predictions that can help identify similar clusters of players more easily.
DTR	– Easy to interpret and visualize results to further understand the decision-making process. – Robust to outliers, making it a good fit for our dataset since there are certain elite players that exhibit statistical outliers.
RFR	– Builds on top of the strong results obtained from DTR. – Enhances predictive performance by leveraging ensemble learning techniques. – Mitigates overfitting and handles outliers effectively via the use of multiple trees.

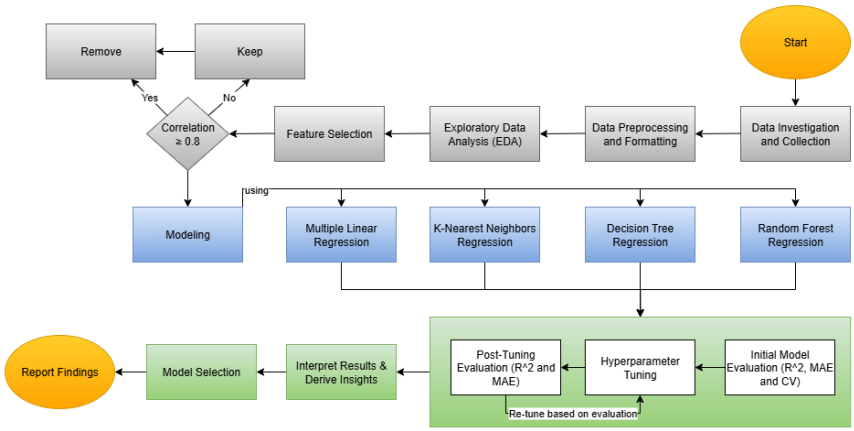

Fig. 3. Framework diagram outlining the Correlation-Optimized NBA Scoring Estimator (CONSE) model pipeline, including data processing, modeling, and evaluation stages

References

1. Becker, T., Rousseau, A.J., Geubbelmans, M., Burzykowski, T., Valkenborg, D.: Decision trees and random forests. American Journal of Orthodontics and Dentofacial Orthopedics 164(6), 894–897 (2023)
2. Chicco, D., Warrens, M.J., Jurman, G.: The coefficient of determination r-squared is more informative than smape, mae, mape, mse and rmse in regression analysis evaluation. Peerj computer science 7, e623 (2021)
3. Ellis, R.P., Mookim, P.G.: K-fold cross-validation is superior to split sample validation for risk adjustment models. Department of Economics, Boston University., Boston, Amerika Serikat 270 (2013)
4. Few, S., Edge, P.: Practical rules for using color in charts. Perceptual Edge, Visual Business Intelligence Newsletter (2008)
5. Gehlenborg, N., Wong, B.: Heat maps. Nature Methods 9(3), 213 (2012)
6. Gogtay, N.J., Thatte, U.M.: Principles of correlation analysis. Journal of the Association of Physicians of India 65(3), 78–81 (2017)
7. Houde, M.: Predicting the Outcome of NBA Games. Ph.D. thesis, Bryant University (2021)
8. Manner, H.: Modeling and forecasting the outcomes of nba basketball games. Journal of Quantitative Analysis in Sports 12(1) (2016), https://doi.org/10.1515/jqas-2015-0088
9. NBA Media Ventures, LLC: Official nba stats (2025), https://www.nba.com/stats
10. Pathak, S., Mishra, I., Swetapadma, A.: An assessment of decision tree based classification and regression algorithms. In: 2018 3rd International Conference on Inventive Computation Technologies (ICICT). pp. 92–95. IEEE (2018)
11. Schneider, P., Xhafa, F.: Anomaly detection and complex event processing over iot data streams: with application to EHealth and patient data monitoring. Academic Press (2022)
12. Senaviratna, N., Cooray, T.: Detecting multicollinearity of binary logistic regression model: An analysis of motorcycle accidents in sri lanka. International Multidisciplinary Research Journal (2019)
13. Sports Reference LLC: Basketball statistics and history (2025), https://www.basketball-reference.com/
14. Stephanos, D.K., Husari, G., Bennett, B.T., Stephanos, E.: Machine learning predictive analytics for player movement prediction in nba: applications, opportunities, and challenges. In: Proceedings of the 2021 ACM Southeast Conference. pp. 2–8 (2021)
15. Timbers, T., Campbell, T., Lee, M., Ostblom, J., Heagy, L.: Data Science: A First Introduction with Python. CRC Press (2024)
16. Wang, J.: Predictive analysis of nba game outcomes through machine learning. In: Proceedings of the 6th International Conference on Machine Learning and Machine Intelligence. pp. 46–55 (2023)