Projecting WNBA Success from NCAA Performance: A Machine Learning Approach to Recognizing Overlooked NCAA Players

Adam Klaus
Master of Science in Data Science
DePaul University

Abstract:

This paper introduces a machine learning approach to identifying overlooked talent in NCAA women's basketball players who may succeed in the WNBA. Utilizing machine learning techniques, the study analyzes data from 564 athletes between 2002 and 2022, focusing on a key metric, Win Shares Per 48 Minutes (WS/48). Various models, including Random Forest, Logistic Regression, SVM, and XGBoost, were employed and evaluated using multiple metrics. The study also identifies gaps in existing research, particularly the lack of focus on women's basketball. The findings reveal significant correlations between college performance and professional success, offering a new perspective on talent recognition in women's basketball. The research contributes to the broader understanding of predictive analytics in sports and has implications for future WNBA expansion.

1 Introduction and Background

The Women's National Basketball Association (WNBA) currently includes a mere 12 teams, a number that has remained relatively stagnant over the past two decades. Reports, including "Cathy Engelbert: WNBA expansion remains '2-4 years out'" by E. Laase [1], hinted at an expansion in the near future, with Commissioner Engelbert voicing plans to introduce at least one more team by 2025. In the meantime, there is a significant limitation on the opportunities available to NCAA Women's basketball players aspiring to advance their careers professionally. The NCAA's study titled "Probability of Competing Beyond High School" offers a crucial frame of reference. The 2019 WNBA draft allowed for 36 slots, 31 were occupied by NCAA players. In comparison, the NBA draft with its 60 slots 52 NCAA Men's division were selected for those slots. When also accounting for international professional avenues, the transition rate from NCAA to professional playing is 6.9% for NCAA Women's players [8] and a significant 21% for NCAA Men's players [2] in the 2018-19 timeframe.

This research aims to leverage machine learning techniques to identify NCAA players likely to flourish in the WNBA. The prime objective of this research is as follows: defining WNBA success, constructing a predictive model for such success from college performance, highlighting the determinants to a player's professional trajectory, and applying the predictive model to showcase examples of overlooked talent.

Success in the WNBA can be quantified using the metric Win Shares Per 48 Minutes (WS/48). It offers an estimate of the number of wins contributed by a player per 48 minutes and is a reputable metric in sports analytics [4]. The target feature has been bifurcated into two categories: WS/48 > 0 and WS/48 <= 0. A positive WS/48 indicates the player positively impacted their teams ability to win and a negative value indicates the player negatively impacted their teams ability to win. Data will undergo exploration to check for basic statistics and missing values. Preprocessing will impute missing values using medians, removal of highly correlated features, and standardization. The data will then be split into training and test sets. Feature selection will be done through a Random Forest model on the full dataset to gauge feature importances [5]. The main modeling phase will encompass Random Forest, Logistic Regression, SVM, and XGBoost, with parameters determined through hyperparameter tuning. The evaluation metrics will include precision, accuracy, recall, f1-score and roc-auc as well as interpreting parameters used.

2 Literature Review

2.1 Existing Research on Predicting Sports Performance

The trajectory of athletes transitioning from college sports to professional basketball has received significant attention. Various studies, employing methodologies ranging from linear regression models to complex machine learning techniques like Random Forest and Support Vector Machines (SVM), have sought to predict NBA success from college performance [3], [4], [5], [6], [7].

"Predicting NBA Success from College Performance," took a holistic approach of gathering historical NBA data, NBA combine metrics, and college statistics. By leveraging machine learning, the study delineated the profound impact of athleticism and college performance on NBA success, using metrics such as Player Efficiency Rating (PER), Win Shares (WS), and Win Shares per 48 minutes (WS/48) [6]. Another research, "Predicting NCAA to NBA Performance from Scouting Reports," illuminated the richness of information embedded in scouting reports, affirming the promise of content analysis in player performance forecasting [7]. The study by Rodenberg & Kim titled "Precocity and Labor Market Outcomes in Professional Basketball" explored how player age and precocity correlate with career outcomes. Their insights revealed that players entering the NBA at younger ages tend to amass higher earnings and enjoy longer careers [8]. In "The Value of College Basketball Statistics in Predicting NBA Draft Success," the research spotlighted specific college statistics, beyond conventional indicators like height and weight, as more telling harbingers of NBA draft success [9]. Lastly, the study "From college to the pros: predicting the NBA amateur player draft" critiqued the predominant focus on scoring during drafts. Their insights drew attention to the undervalued significance of defense-related skills in the drafting process [10].

2.2 Gaps in Existing Research

While these studies have paved the way in predicting NBA success, they also highlight several gaps in our current understanding. For instance, the excessive emphasis on scoring in player

drafting, as discussed in "From college to the pros: predicting the NBA amateur player draft" [10], points to a need to consider a wider array of play styles, such as defense. The potential of content analysis, as highlighted in "Predicting NCAA to NBA Performance from Scouting Reports" [7], suggests that unconventional data sources could hold valuable predictive information.

Furthermore, although these studies provide significant insights into predicting NBA success, there's a scarcity of similar research focusing on the Women's National Basketball Association (WNBA). One of the few publications about the WNBA is [11], which focuses specifically on performance after returning from an ACL injury. Our study aims to bridge the gap by focusing on predicting the success of WNBA players based on their college performance, thus extending the reach of predictive analytics in Women's basketball.

3 Data Selection

3.1 Overview of the Dataset

The dataset offers a detailed account of the full college career performance of individuals who transitioned to the WNBA between the years 2002 and 2022. Of the initial data entries, 564 records pertained to athletes with NCAA careers with full career statistics publicly available. It is noteworthy that some of these records were excised due to the absence of pivotal data. A significant limitation of the dataset is its truncation at the year 2001, implying that college data prior to this year remains elusive.

3.2 Data Collection Methods

The data collection process was initiated with a web crawl of [3] to compile a comprehensive list of WNBA players. With this roster, a subsequent web scraping operation was executed, leveraging Google search to locate the appropriate link from [3] for each athlete. This facilitated the extraction of both per-game, total and advanced statistics for their college career.

3.3 Limitations and Restrictions

While the dataset is robust, certain limitations are evident that necessitate caution in the extrapolation of the findings derived from this dataset. Some player data was found to be incomplete. Additionally, both Her Hoop Stats [12] and Basketball Reference [3] databases offer complete data only until around the 2001-2002 season, limiting the historical depth of our dataset. Vital demographic details such as age, biometric data, and hometown are absent, potentially missing out on the explained variance; however, using this data would also warrant ethical considerations [13]. The data also spans a wide temporal spectrum, which might introduce variability due to evolving game strategies, training methodologies, or other temporal factors. Despite these limitations, the analysis still showcases the ability to understand college performance translating to professional success.

4 Pre-Modeling

4.1 Exploratory Data Analysis

In order to develop a robust machine learning model, an in-depth exploration of the dataset was paramount. This entailed analyzing the basic statistics and understanding the distribution of key features. For instance, looking at the distribution of WNBA players' debut year in the league gives us an idea of the volume of players in our dataset entering the league during certain seasons. This displays that our dataset had a fairly consistent distribution of players to analyze across seasons.

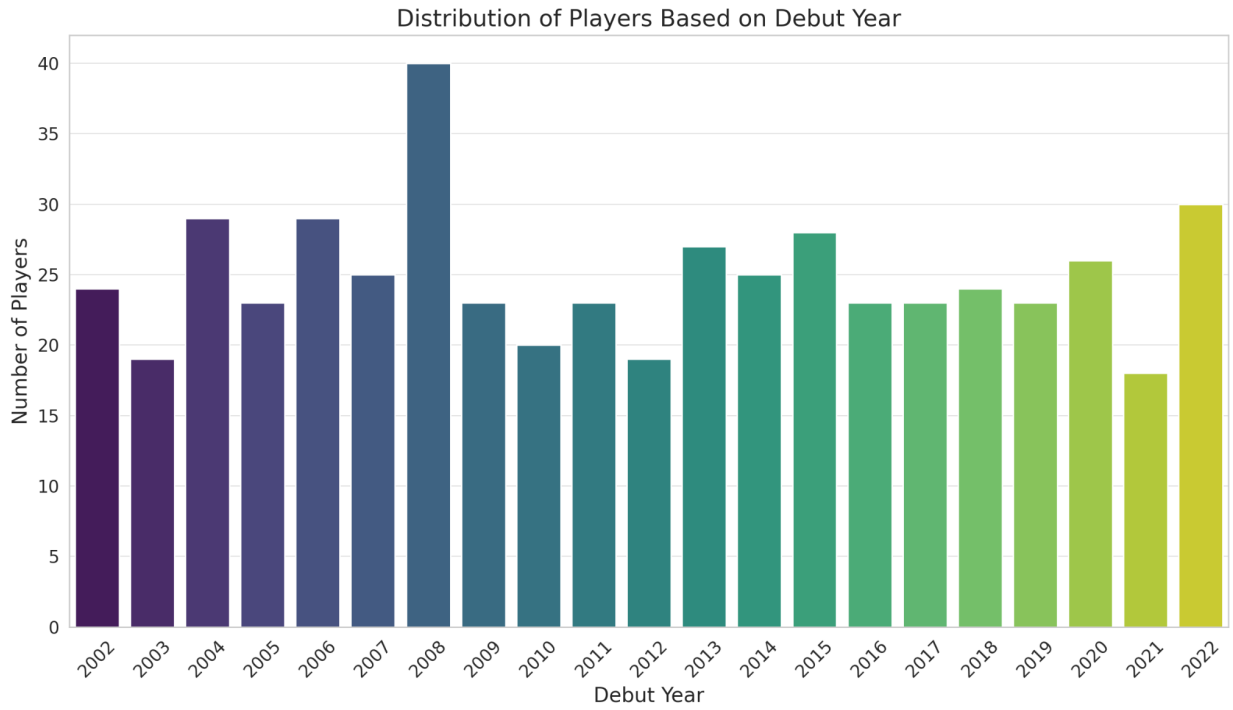

Fig. 1. Distribution of Players by Debut Year

Additionally, the dataset was analyzed for the most frequented colleges attended by WNBA players. This information is pivotal as certain colleges, due to their successful basketball programs, might produce a higher number of WNBA-ready athletes.

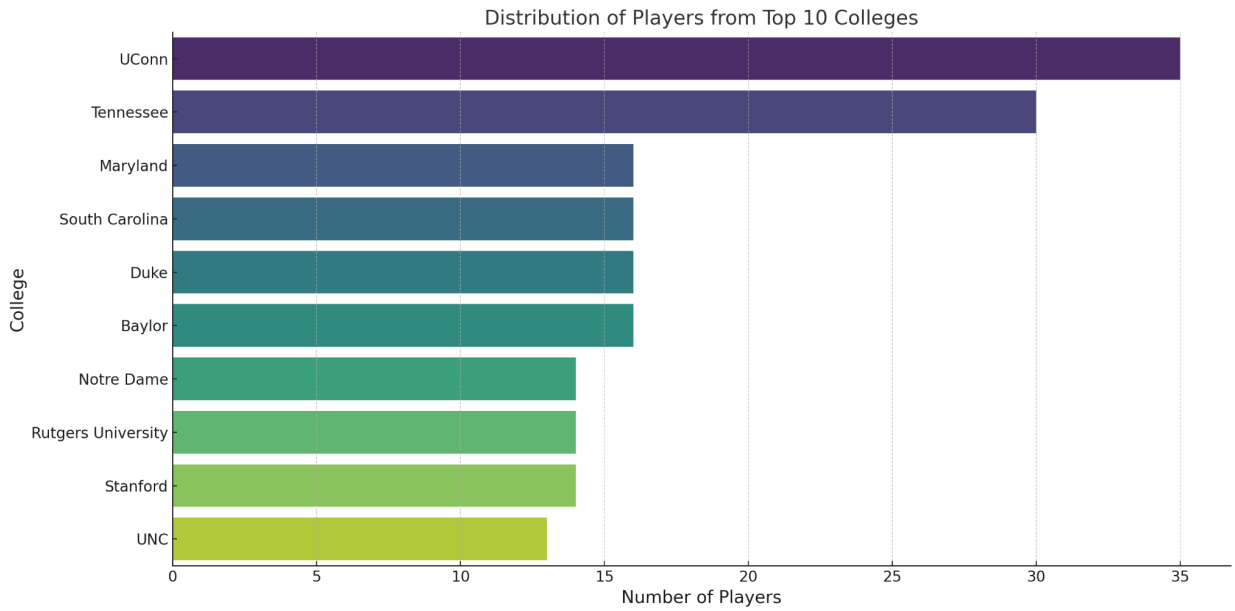

Fig. 2. Distribution of Players from Top 10 Colleges

Beyond individual colleges, the conferences to which these colleges belong can also be indicative of the quality of competition and, by extension, the readiness of athletes for professional play. The dataset contains a variety of conferences, and the attached chart highlights the most prominent ones.

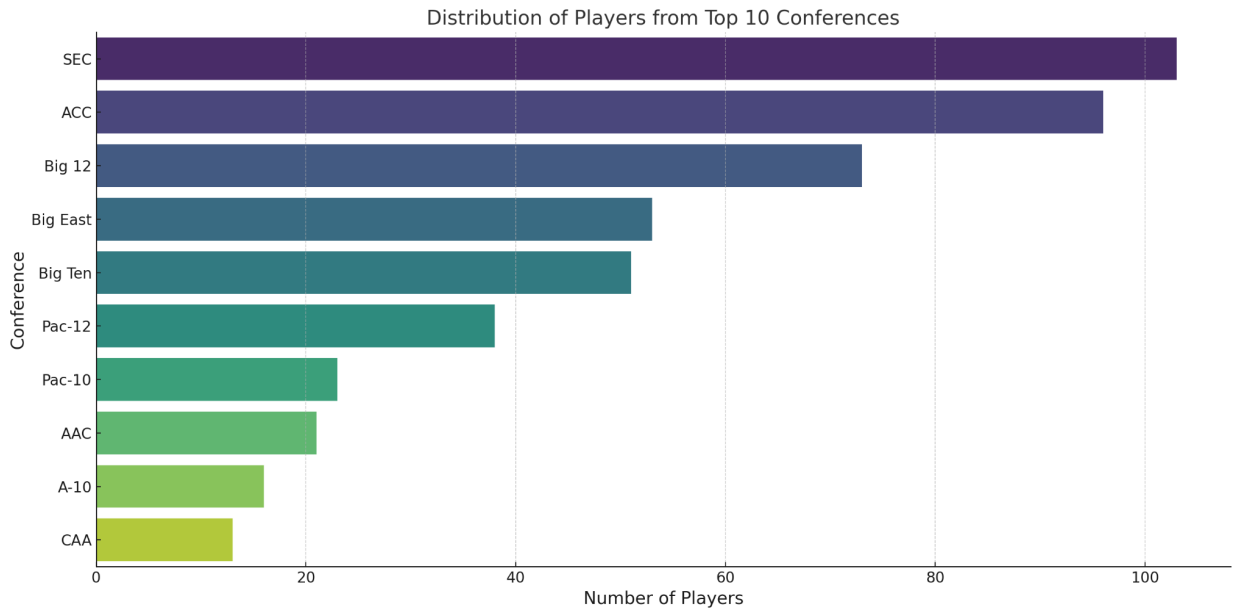

Fig. 3. Distribution of Players from Top 10 Conferences

A player's position can significantly influence their performance metrics. For instance, guards might have more assists, while centers might have more blocks or rebounds. Understanding the distribution of player positions within the dataset ensures a balanced representation and aids in interpreting model outcomes more contextually.

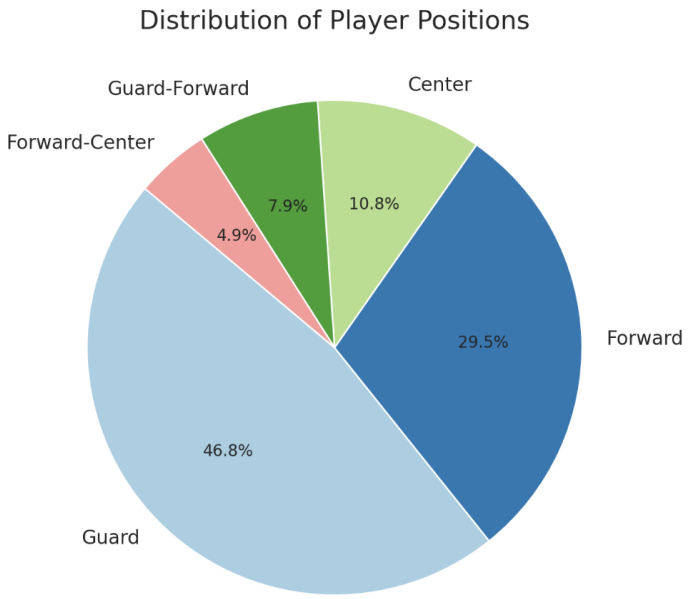

Fig. 4. Distribution of Player Positions

Win Shares Per 48 Minutes (WS/48) is a key metric in assessing a player's contribution to their team's victories and a main focus of this analysis. A distribution analysis of this metric provides a full view of player performance in the WNBA. The figure below illustrates this distribution, shedding light on the spread of performance metrics for the players in question. Of note we can see the distribution is relatively even and centers around .05 win shares per 48 minutes. This is crucial to consider to understand the class imbalance we phase when training the model because there are more WNBA players with positive WS/48 than there are negative.

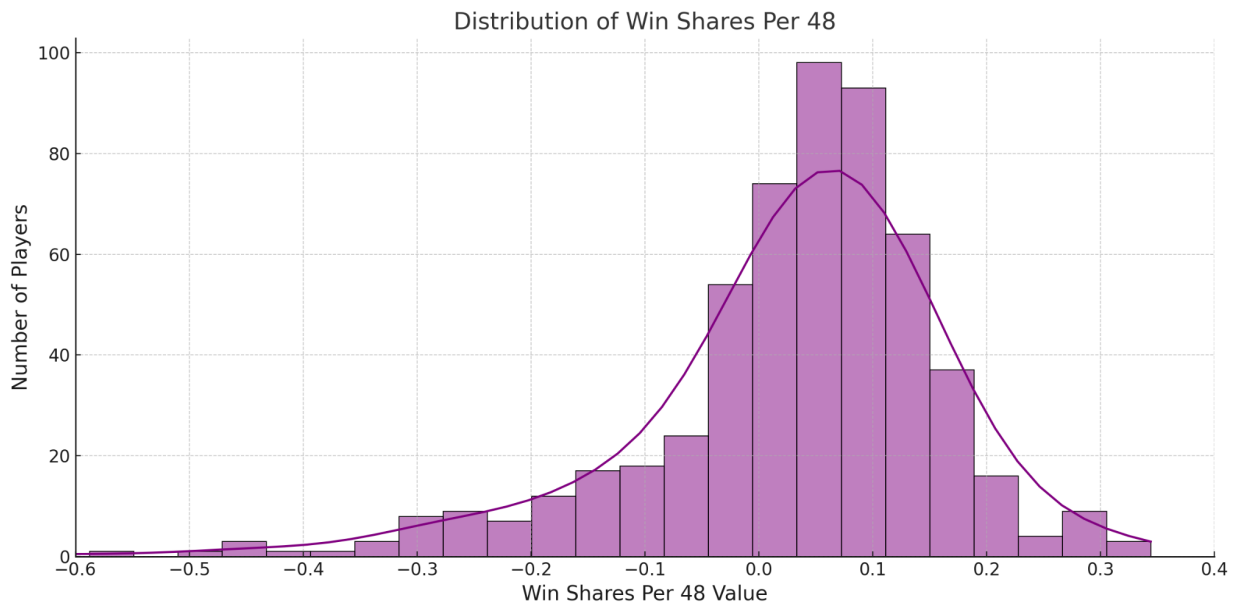

Fig. 4. Distribution of Win Shares Per 48

In sum, this exploratory analysis gives us perspective for subsequent data preprocessing and modeling. It ensures a comprehensive understanding of the dataset's nuances and hints at emerging patterns to guide the research towards more accurate and meaningful outcomes.

4.2 Data Preprocessing

In machine learning, the quality and structure of the data often dictate the success of predictive models. To this end, preprocessing of the dataset was undertaken to ensure its readiness for modeling. Missing values were addressed by imputation of the median value for a given feature to ensure records had complete data, a method found to be effective in various studies, including the detection of liver disease [14]. For the benefit of algorithms sensitive to feature scales, the dataset underwent standardization, ensuring all attributes were rescaled to have a mean of 0 and a standard deviation of 1. Subsequently, to validate the efficacy of the models on unseen data, the dataset was partitioned into training and test subsets. This crucial step ensures models are robust and generalizable, preventing overfitting and enhancing the predictive accuracy on new data points. The importance of such preprocessing steps and the challenges of overfitting and generalization are well-documented in [14].

4.3 Feature Selection

Discerning which features hold the most predictive power is paramount for optimizing the performance of machine learning models. In this research, a comprehensive set of features were initially assessed for their predictive ability, including those generated during feature engineering. First, highly correlated features were identified and removed. This step is crucial as

closely interrelated features can induce multicollinearity, which can obscure the model's interpretability and occasionally hamper its performance.

To further refine the features, a Random Forest classifier was used. The Random Forest classifier proved instrumental in its inherent capability to rank features based on their predictive significance. The relative importance of each feature was visualized, and the Elbow Method was incorporated to determine the optimal number of features. By plotting the features against their importance scores in a descending manner, a point was identified—resembling an 'elbow'—where the marginal gain in importance significantly diminishes. Features above the 'elbow' point were deemed as having the most relevance, guiding our selection process. This approach ensured that the chosen features added substantial predictive value while maintaining model simplicity and efficiency.

5. Model Types

5.1 Supervised Classification Models

In this study, we use supervised classification to discern patterns within the data and make predictions about player success in the WNBA, as quantified by the metric Win Shares Per 48 Minutes (WS/48). Four popular classification models were used in training, described in Fig. 5.

Model Type	Description
Random Forest	An ensemble learning method that constructs multiple decision trees during training. Outputs the class that is the mode of the classes from individual trees.
Logistic Regression	A statistical model used for binary classification. Estimates the probability that a given instance belongs to a particular category.
Support Vector Machines (SVM)	A classification algorithm that seeks the best hyperplane separating classes in the input feature space.
XGBoost	An optimized gradient boosting library. Efficient, flexible, and portable. Uses gradient boosting framework to iteratively refine predictions based on errors of previous iterations.

Fig. 5. Model Types

To optimize the performance of our models, hyperparameter tuning was executed. This process involves adjusting various parameters to find the best model architecture. Both grid search and random search methodologies were explored to identify the optimal parameter configurations for each model. Upon completion of the model training and hyperparameter tuning processes, the best parameters for each model were extracted. These parameters detail the optimal configurations that output the best performance during the modeling phase.

We will be attentive to all of the metrics in Fig. 6 when evaluating the performance of different models. Recall, specifically, will be crucial to look at for practical application because we are interested in accurately identifying “True Positives” or players that would be successful in the WNBA. The performance of the models was assessed using various metrics to ensure a holistic understanding:

Metric	Description
Precision	Measures the number of true positive predictions among the total predicted positives.
Accuracy	Reflects the proportion of true results (both true positives and true negatives) in the total dataset.
Recall	Captures the ratio of true positive predictions to the actual positive observations.
F1-Score	The harmonic mean of precision and recall, providing a balance between the two metrics.
ROC-AUC	A performance measurement for classification problems at various thresholds settings. Reflects the model's capability to distinguish between the positive and negative classes.
Confusion Matrix	Used to understand the performance of a classification model. Reports the counts of the true positive, true negative, false positive, and false negative predictions of a classifier.

Fig. 6. Evaluation Metrics

5.2 Principal Component Analysis (PCA)

PCA was employed to gain a deeper understanding of the feature interactions and importance, a technique that has been effectively utilized in various applications [15]. In our case, PCA was used for understanding how features correlate with one another as opposed to dimensionality reduction. This approach aligns with methods used in other studies, where PCA has been leveraged not only for dimensionality reduction but also for understanding complex relationships between variables [15]. Additionally, it helps in visualizing the data in lower-dimensional space, providing insights into the underlying structure of the data [15].

6. Analysis

6.1 Binary Classification for WNBA Success

In the pursuit of predicting a player's positive Win Shares per 48 minutes (WS/48), we began constructing a predictive model. This started with an examination of the features, aiming to discern their importance within the model. Utilizing Random Forest, the top features were

isolated, each distinguished by its unique contribution to the model's predictive power, as illustrated in Figure 7.

An analytical technique known as the 'elbow method' was used to identify a significant drop-off in feature importance following 'adv_ast%'. The importance of feature selection in machine learning models is well-recognized, as it enhances the model's efficiency and accuracy [16] The table below shows the top features, each carefully identified for its substantial contribution to the model's ability to predict positive WS/48.

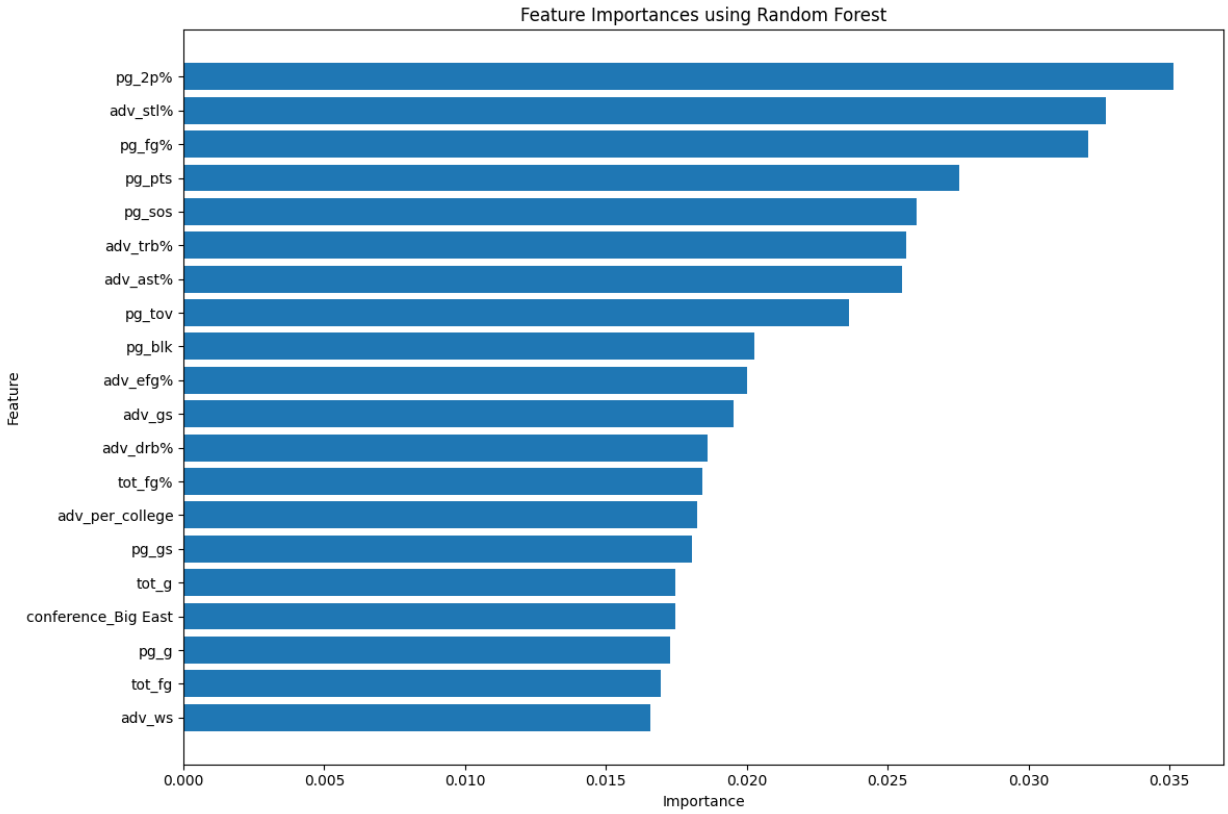

Fig. 7. Feature Importance using Random Forest

Feature Name	**Stat**	**Description**
pg_2p%	2-Point Field Goal Percentage	The percentage of successful 2-point field goals made out of attempts.
adv_stl%	Steal Percentage	The percentage of opponent possessions that end with a steal by the player.

pg_fg%	Field Goal Percentage	The percentage of successful field goals made out of attempts, including both 2-point and 3-point shots.
pg_pts	Points per Game	The average number of points scored by the player per game.
pg_sos	Strength of Schedule	A measure of the difficulty of the opponents a team has faced or will face in the future.
adv_trb%	Total Rebound Percentage	The percentage of available rebounds grabbed by the player while on the court.
adv_ast%	Assist Percentage	The percentage of teammate field goals that the player assisted while on the court.
pg_tov	Turnovers per Game	The average number of times the player loses possession of the ball to the opposing team per game.

Fig. 8. Top Important Features

Among various algorithms, Random Forest was selected as the primary modeling technique. Its ensemble nature, which combines multiple decision trees, enhances the model's robustness and reduces the risk of overfitting. Additionally, Random Forest's ability to capture complex, non-linear relationships between features and the target variable, along with insights into feature importance, made it a suitable choice [5]. Comparative analysis further justified this selection. Random Forest achieved the highest accuracy of 0.726, indicating strong overall performance. With a precision of 0.750, recall of 0.896, F1 Score of 0.817, and ROC AUC of 0.629, the model demonstrated a balanced ability to identify true positives without excessively misclassifying negatives, reflecting its discriminative power.

Model	Accuracy	Precision	Recall	F1 Score	ROC AUC	Best Parameters
Random Forest	0.726	0.750	0.896	0.817	0.629	{'n_estimators': 200, 'min_samples_split': 2, 'min_samples_leaf': 4, 'max_depth': 10, 'bootstrap': True}
Logistic Regression	0.637	0.725	0.753	0.739	0.571	{'solver': 'saga', 'penalty': 'l1', 'C': 10}
SVM	0.655	0.698	0.870	0.775	0.532	{'kernel': 'rbf', 'gamma': 'scale', 'C': 1}
XGBoost	0.655	0.726	0.792	0.758	0.577	{'subsample': 1, 'n_estimators': 50, 'max_depth': 5, 'learning_rate': 0.5, 'colsample_bytree': 0.8}

Fig. 9. Evaluation of Models

The optimal parameters for Random Forest were found to be {'n_estimators': 200, 'min_samples_split': 2, 'min_samples_leaf': 4, 'max_depth': 10, 'bootstrap': True}. These parameters play a crucial role in the model's performance. The ensemble of 200 trees reduces variance without overcomplicating the model. Parameters like min_samples_split and min_samples_leaf control the tree's growth, preventing overfitting by ensuring that the leaves have a minimum number of samples. A maximum depth of 10 ensures that the trees are deep enough to capture complex patterns but not so deep that they memorize the training data. Bootstrapping introduces randomness into the model, enhancing generalization. Collectively, these parameters contribute to a model that is complex enough to capture underlying patterns

but restrained enough to generalize well to unseen data. The confusion matrix for the best model is presented in Fig 10.

	Predicted WS/48 < 0	Predicted WS/48 >= 0
Predicted WS/48 < 0	14	22
Predicted WS/48 >= 0	7	70

Fig. 10. Random Forest Confusion Matrix

This matrix provides insights into the model's ability to correctly classify players based on their predicted Win Shares Per 48 Minutes (WS/48). It highlights the trade-offs between sensitivity and specificity, reflecting the model's performance in differentiating successful and less successful players.

The process of training and assessing the best model involved a systematic exploration of feature importance, model selection, parameter tuning, and evaluation. The Random Forest model's success underscores the value of ensemble methods and careful feature selection in predicting WNBA success from NCAA performance. This analysis shows that college performance is predictive of success in the WNBA.

6.2 Multi-Class Classification: Labeling Performance

In this section, the WNBA win shares are categorized into three distinct performance buckets: High, Medium, and Low. This categorization is achieved through even binning, allowing for a more nuanced analysis of player performance. A bar chart shown in Fig. 11 visually represents the distribution of players across these three categories, providing insights into the overall landscape of performance levels in the WNBA.

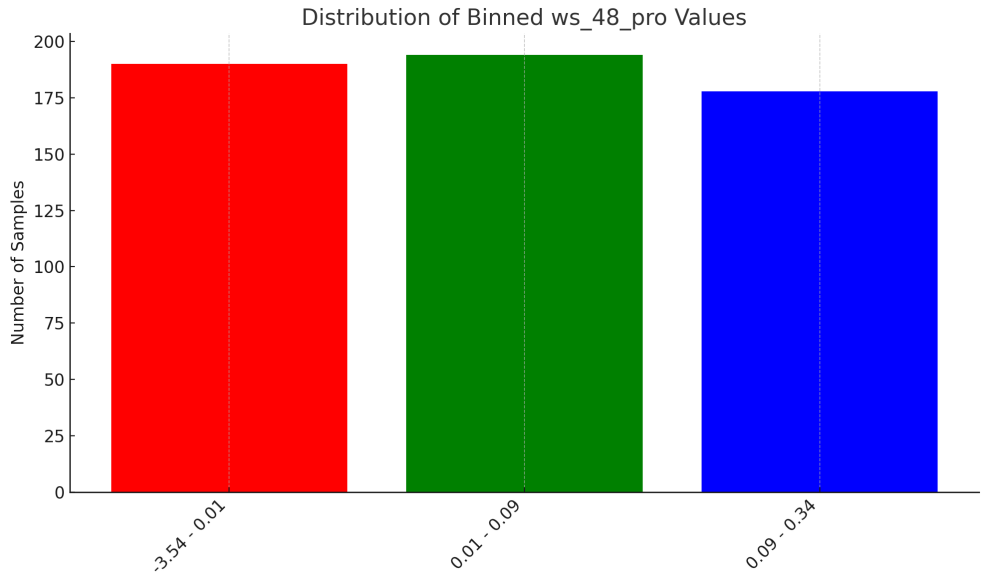

Fig. 11. Distribution of Binned WS/48

Several models were evaluated for their ability to classify players into the High, Medium, and Low performance buckets. As a note, the top features identified in the binary classification were carried over to this multiclass classification task.

Model	Accuracy	Precision	Recall	F1 Score
Logistic Regression	0.496	0.496	0.501	0.492
SVM	0.487	0.484	0.488	0.483
Random Forest	0.398	0.383	0.406	0.392
XGBoost	0.434	0.429	0.439	0.432

Fig. 12. Evaluation of Multi-Class Classification Models

The models varied in their performance, with none showing a clear dominance in all metrics. That being said, the Logistic Regression model edged out the other models slightly, especially in the recall. This variation highlights the complexity of the task and the challenges associated with accurately classifying players into distinct performance categories. Fig. 14 shows the confusion matrix of the Logistic Regression model.This matrix provides insights into the model's ability to correctly classify players into the Low, Medium, and High performance categories. It highlights the trade-offs between sensitivity and specificity, reflecting the model's performance in differentiating various levels of success.

	Predicted Low	Predicted Medium	Predicted High
Actual Low	25	5	8
Actual Medium	16	13	11
Actual High	6	5	24

Fig. 14. Multi-Class Confusion Matrix

The multiclass classification of WNBA win shares into High, Medium, and Low performance buckets provided valuable insights into player performance levels. While the models exhibited varying degrees of success, the exercise underscores the potential of machine learning in nuanced player evaluation and offers a foundation for further exploration and refinement.

6.3 Principal Component Analysis:

In the context of predicting WNBA performance from college statistics, PCA serves as an essential tool for reducing dimensionality, enabling the isolation of key underlying factors that might influence a player's professional success. By transforming the high-dimensional college data into a smaller set of orthogonal components that capture the most significant variance, PCA provides a more manageable and interpretable structure, assisting in uncovering the complex relationships between college performance metrics and subsequent achievements in the WNBA.

This chart illustrates how each principal component contributes to the total variance within the dataset, offering insight into the key dimensions that influence the transition from college to professional basketball. The individual bars depict the percentage of variance explained by each component, while the step line marks the cumulative explained variance. By pinpointing the "elbow" in the chart, where additional components provide diminishing returns in explained variance, the analysis helps us understand which components are the most important. From the figure below we can see that the first two components encapsulate a significant portion of the variance.

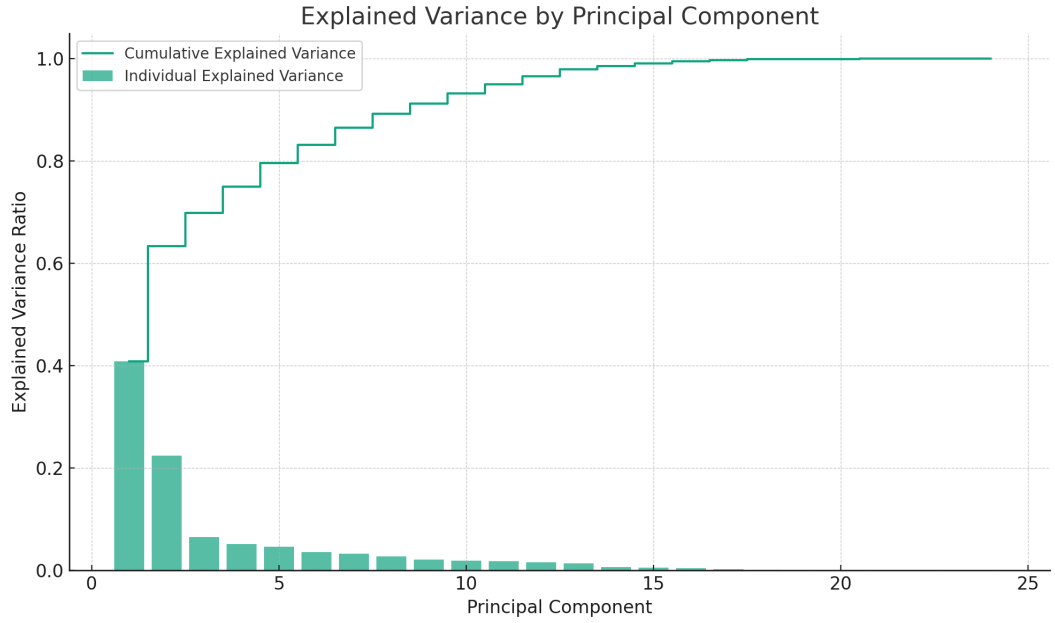

Fig. 13. PCA Explained Variance

Further digging into the loadings of each component helps us understand the attributes that may be most pertinent in forecasting success at the professional level in the WNBA as well as how the features interact together.

The primary component, PC1, predominantly showcases positive associations with three-point metrics: tot_3pa (total three-point attempts), tot_3p (total three-pointers made), and tot_3p% (three-point percentage). In contrast, there are significant negative associations with tot_fg (total field goals), tot_pts (total points), and tot_2pa (total two-point attempts). This suggests that PC1 emphasizes the distinction between players' three-point shooting proficiency and their overall scoring capability.

For PC2, there are marked positive associations with defensive and efficiency metrics such as tot_blk (total blocks), tot_fg% (field goal percentage), and tot_orb (total offensive rebounds). On the other side, negative correlations are observed with tot_3pa, tot_3p, and tot_ast (total assists). This component seems to encapsulate a contrast between players skilled in defensive attributes and those who are more offensively oriented, especially with long-range shooting and playmaking.

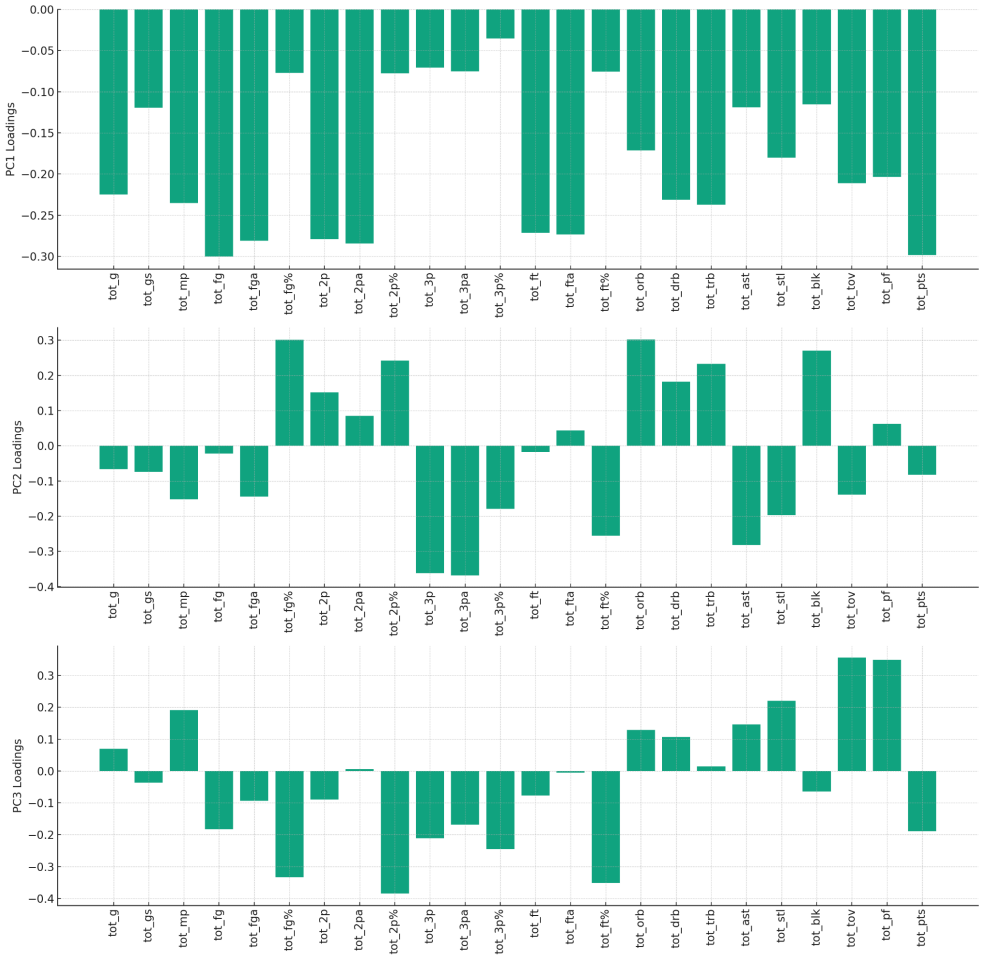

Fig. 14. PCA Loadings

7. Real-World Application: A Case Study of Predictions versus Actual Outcomes

The model's application was specifically directed at players who were seniors during the 2021-2022 season and did not get drafted into the WNBA. This focus allowed for a targeted analysis of individuals at a critical juncture in their basketball careers. The predicted probability generated by the model represents a quantifiable likelihood of success in the WNBA, based on a player's NCAA performance. This probability serves as a guiding metric, identifying players who may have been overlooked but possess the potential to succeed at the professional level. For a comparison, Fig. 15 displays a snapshot of the players who had the highest predicted probability and their current performance in other professional leagues around the world found on [19].

Player Name	College	Predicted Probability	Current Team	Points per Game	Rebounds per Game	Assists per Game
Que Morrison	University of Georgia	99.43%	Lucca (Italy-Serie A1)	12.5	4.7	3.6
Meral Abdelgawad	Western Kentucky University	95.77%	Al Ahly (Egypt)	11.1	4.1	3.1
Hannah Nihill	Drexel University	95.70%	N/A			
Cierra Hooks	Ohio University	95.60%	Nyon (Switzerland-SBL)	14.5	8.5	4.3
Juliunn Redmond	Texas A&M International	95.55%	Fribourg (Switzerland-SBL)	17.7	5.3	3.4
Marie Hunter	University of Nevada, Las Vegas	95.28%	Riva (Switzerland)			
Summer Menke	Sacramento State University	95.12%	N/A			
Georgia Dale	Deakin University	95.03%	Northside W. (Australia-NBL One)	6.5	4	2.8
Lauren Heard	Texas Christian University	94.68%	Rockhampton C. (Australia-NBL One)	16.4	6.5	7
Mariah McCully	Wichita State University	94.34%	N/A			

Fig. 15. Predicted Probability of WNBA Success in the Class of 2022

These players have demonstrated varying degrees of success, with some excelling in European and Australian leagues. The diversity of their current teams and performance metrics reflects the multifaceted nature of basketball talent and the opportunities available beyond the WNBA. The application of the model to this specific cohort underscores the potential of predictive analytics in recognizing overlooked talent, particularly from lesser-known basketball schools. It offers a tangible demonstration of how data-driven insights can guide a case for expansion in the WNBA.

8. Ethical Considerations

8.1 Ethics of Data Collection:

The data for this study was collected from publicly available sources: HerHoopStats [12] and Basketball-Reference.com [3]. As these are open-source data providers that aggregate and provide statistics for public use, collecting data from these sources is ethically sound.

The data collected and used in this study is non-sensitive and non-personal. It relates to the professional careers of WNBA players, which is public knowledge and widely reported.

Guidance and consideration was also taken from papers on public data ethics and big data such as [18] and [19].

Consent for data collection is implicitly granted as the data is publicly available and intended for public consumption. The data was used purely for the purpose of this research study: to explore factors contributing to a successful WNBA career and to build predictive models. The study focused on player performance metrics and did not involve any personal or sensitive information about the players. Thus, the research respects the privacy and dignity of the players.

8.2 Potential Ethical Issues and Mitigation:
One potential ethical issue in using this data might be the risk of overgeneralization. It's important to remember that while statistics can provide insights into trends and patterns, they do not define an individual player's success or potential. Players can improve over time, and many factors contributing to their success cannot be captured in statistics.

To mitigate this, the results of the study were presented in a way that emphasizes their predictive, not definitive, nature. The models can suggest trends and inform decisions, but they should not be used as the sole factor in determining a player's potential or value. Lastly, it is crucial to use these findings to support and enhance Women's basketball, not to limit opportunities for players.

9. Conclusion and Future Work

This research paper has embarked on a comprehensive exploration of predicting player success in the WNBA, employing a multifaceted approach that encompasses both binary and multiclass classification. Through meticulous analysis and modeling, the study has illuminated the underlying factors that contribute to understanding predictors of success in the WNBA, a subject lacking attention in the field of sports analytics.

The binary classification model, leveraging Random Forest, successfully pinpointed key features such as 2-point field goal percentage, steal percentage, and assist percentage, among others [8]. These features were instrumental in predicting WNBA success from NCAA performance, with the model demonstrating robustness and generalization. The multiclass classification extended the analysis by categorizing WNBA win shares into High, Medium, and Low performance buckets, providing nuanced insights into the challenges and complexities of accurately classifying players into distinct performance categories.

The application of Principal Component Analysis (PCA) played a pivotal role in the study. By transforming the original features into a set of uncorrelated principal components, PCA facilitated a more efficient and interpretable analysis, enhancing our ability to understand the types of features that capture the essential attributes contributing to player success. Furthermore, the application of the model to specific players offered practical insights, showcasing the depth of talent that exists in the NCAA.

In conclusion, this research contributes valuable insights to women's basketball, demonstrating the potential of machine learning techniques in player evaluation and talent projection. The findings offer a solid foundation for further exploration, practical application, opening new avenues for research and innovation in the field of sports analytics.

Further research could explore other machine learning algorithms and feature selection methods, or apply the models to other basketball leagues, like international professional women's leagues, to see if the results are consistent across professional leagues. Future research could also explore other aspects that could impact a player's success at the professional level, such as scouting reports, team performance and coaching impact. Incorporating these aspects could provide a more holistic view of player productivity. It would also be interesting to analyze the performance of the models over time. For instance, do they perform better or worse when predicting success for players who started their careers in different years or eras. Finally, this paper specifically looked at WNBA players who made the transition from the NCAA to the WNBA to find the factors that differentiated their success, however additional research could be done to explore how this would apply to the entire population of NCAA Women's basketball players.

References

1. E. Laase, "Cathy Engelbert: WNBA expansion remains '2-4 years out'," Just Women's Sports, 2023. [Online]. Available: https://justwomenssports.com/reads/wnba-expansion-cathy-engelbert-timeline-portland/
2. NCAA Research, "Estimated Probability of Competing in College Athletics and Professional Athletics," NCAA, 2020. [Online]. Available: https://www.ncaa.org/sports/2015/3/6/estimated-probability-of-competing-in-professional-athletics.aspx. [Accessed: Aug. 6, 2023].
3. "Men's and Women's College Basketball Statistics and History," Available: https://www.sports-reference.com/cbb/.
4. N. Y. Li, N. J. Lemme, S. DeFroda, E. Nunez, D. Hartnett, and B. Owens, "Performance After Operative Versus Nonoperative Management of Shoulder Instability in the National Basketball Association," *Journal of Orthopaedic Surgery and Research*, Dec. 2019, doi: 10.1177/2325967119889331.
5. Breiman, L. (2001). "Random Forests." Machine Learning, 45(1), 5-32.
6. Kannan, S., et al., "Predicting NBA Success from College Performance," Journal of Sports Analytics, 2018.
7. Maymin, A., "Predicting NCAA to NBA Performance from Scouting Reports," Journal of Quantitative Analysis in Sports, 2021.
8. Rodenberg, R. & Kim, A., "Precocity and Labor Market Outcomes in Professional Basketball," Journal of Sports Economics, 2011.
9. Alamar, B., "The Value of College Basketball Statistics in Predicting NBA Draft Success," Journal of Quantitative Analysis in Sports, 2014.
10. David J. Berri , Stacey L. Brook, Aju J. Fenn, "From college to the pros: predicting the NBA amateur player draft," Journal of Sports Analytics, 2011.

11. Return to Play (RTP) and Performance in the Women's National Basketball Association (WNBA) Following Anterior Cruciate Ligament (ACL) Reconstruction (194) by J. Tramer et al. (2021)
12. "Her Hoop Stats," Available: https://www.herhoopstats.com.
13. K. Karkazis and J. R. Fishman, “Tracking u.s. professional athletes: The ethics of biometric technologies.,” American Journal of Bioethics, vol. 17, no. 1, pp. 45 – 60, 2017. 9.
14. A. Q. Md, S. Kulkarni, C. J. Joshua, T. Vaichole, S. Mohan, and C. Iwendi, "Enhanced Preprocessing Approach Using Ensemble Machine Learning Algorithms for Detecting Liver Disease," Biomedicines, vol. 11, no. 2, p. 581, Feb. 2023. [Online]. Available: https://dx.doi.org/10.3390/biomedicines11020581. [Accessed: Aug. 13, 2023].
15. I. Fayad et al., "Canopy Height Estimation in French Guiana with LiDAR ICESat/GLAS Data Using Principal Component Analysis and Random Forest Regressions," 2014.
16. P. Ghosh et al., "Efficient Prediction of Cardiovascular Disease Using Machine Learning Algorithms With Relief and LASSO Feature Selection Techniques," IEEE Access, [Online]. Available: PDF. [Accessed: 2022-01-21].
17. EuroBasket, "Player Statistics and Profiles," EuroBasket. [Online]. Available: https://basketball.eurobasket.com/player/.
18. M. Zimmer, ““but the data is already public”: on the ethics of research in facebook,” Ethics and Information Technology, no. 12(4), pp. 313–325, 2010. 10.
19. R. Herschel and V. Miori, “Ethics & big data,” Technology in Society, vol. 49, pp. 31 – 36, 2017.