Athlete performance in collegiate basketball: Predicting match line-up

Jash Parikh [AU2140108], Nevil Jobanputra [AU2140209], Paridhi Jain [AU2120226], Saahil Doshi [AU2140106]

***Abstract*—This report presents a comprehensive analysis to enhance athlete performance in collegiate basketball through predictive analytics and visualization techniques. Leveraging a diverse dataset encompassing sleep patterns, training details, emotional states, game scores, and more for Division I basketball players, our research focuses on forecasting match line-ups and supporting decision-making processes. The study involves merging datasets from two seasons, conducting correlation matrix analysis, imputing missing data using the Iterative Imputer with a Random Forest Regressor, standardizing the data, and analyzing feature importance using the RandomForestRegressor. To overcome the challenge of minority sampling, we employ SMOTE-ENN and SMOTE-Tomek techniques, and then fit an XGBClassifier for the classification of the inter-quantile range. Our findings demonstrate the ability to predict individual Athlete's RSI (Relative Strength Index) values with an accuracy of 97.47% and an F1-score of 0.98, and to predict RSI values for the next week with an accuracy of 88.23% and an F1-score of 0.88; we then leverage the same model to predict the next match lineup and utilize eXplainable AI (XAI) using SHAP for feature interpretation. Finally, we develop a visual dashboard using Streamlit for real-time visualization and comparison of individual performance metrics against team averages, aiming to provide coaches with actionable insights to optimize training strategies and elevate team performance in collegiate basketball.**

***Index Terms*—Collegiate Basketball, RSImod, Random Forest Regressor, XAI, Streamlit.**

I. Introduction

In the rapidly evolving landscape of collegiate basketball, leveraging data analytics has become increasingly vital for teams to gain a competitive edge.[1] By harnessing the power of predictive modeling and visualization techniques, coaches and analysts can extract actionable insights from multifaceted datasets, thereby enhancing athlete performance and refining strategic decision-making processes.

Our report focuses on analyzing a comprehensive dataset encompassing various facets of Division I basketball players' performance, including sleep patterns, training details, emotional states, game scores, and physiological indicators such as cardiac rhythm patterns and weekly readiness scores[2]. The ultimate goal is to develop predictive models for RSImod (a measure of player readiness derived from jump-data) and to provide interpretable explanations for these predictions using eXplainable AI (XAI).

This report outlines a structured approach to tackle the complexities associated with predicting match line-ups in collegiate basketball. By employing advanced data augmentation techniques, machine learning algorithms, and XAI methods, we aim to unearth underlying patterns and relationships within the dataset, thereby facilitating informed decision-making.

Through the development of predictive models and visualization dashboards, our research endeavors to modernize sports management practices by furnishing coaches with actionable insights derived from data-driven analytics.[3] By integrating predictive analytics and visualization tools, we seek to empower coaches to optimize training strategies, pinpoint areas for enhancement, and ultimately elevate athlete performance and team competitiveness.

II. Methodology

A. Data

In this project, we use the data obtained from the season 2 and season 3 performance of Division 1 Basketball players. The Dataset includes sleep patterns, training details, cardiac rhythm patterns, emotional-mental state information, game scores, weekly readiness scores, and jump-data (RSImod) of the atheletes.

B. Data Preparation

The initial step involved merging datasets from two seasons to create a comprehensive dataset for analysis. Subsequently, a correlation matrix was generated to explore the relationships between different variables in the dataset.

C. Data Preprocessing

Following data merging, missing values were addressed through an iterative imputation technique using a Random Forest Regressor as an estimator. A correlation matrix was then recalculated post-imputation to assess changes. To ensure consistency and comparability, the data underwent standardization to scale the features appropriately.

D. Feature Engineering

Feature importance was determined using a RandomForestRegressor model to identify key predictors of the RSImod variable. The results were visualized through bar graphs, providing insights into the relative importance of different features. Additionally, quartiles were created for RSI mapping, facilitating further analysis and interpretation.

E. Handling Imbalance

To address class imbalance in the dataset, minority sampling techniques such as Synthetic Minority Over-sampling Technique (SMOTEENN) and resampling were employed, ensuring a more balanced representation of data across different classes.

F. Model Development

A machine learning model, specifically an XGBClassifier, was selected and trained to classify players based on their RSI quartiles. This model was then used to predict RSI values for the upcoming week, providing valuable insights into player readiness and performance. Furthermore, the model was utilized to forecast the next match lineup, aiding coaches in strategic decision-making.

G. eXplainable AI (XAI)

To interpret model predictions and understand the influence of different features on outcomes, SHAP (Shapley Additive exPlanations) analysis was conducted. This provided valuable insights into the underlying mechanisms driving the model's predictions, enhancing transparency and interpretability.

H. Visualization

Graphical representations were generated to visualize the results of XAI analysis and model predictions. Additionally, a dashboard was developed using Streamlit, offering an interactive platform for visualizing and exploring the data, as well as facilitating feedback propagation and decision-making.

III. Results

A. Data Imputation and Preprocessing

The iterative imputation process effectively addressed missing values in the dataset, enhancing data completeness and reliability. Post-imputation, the correlation matrix revealed changes in the relationships between variables, reflecting the impact of imputation on data integrity.

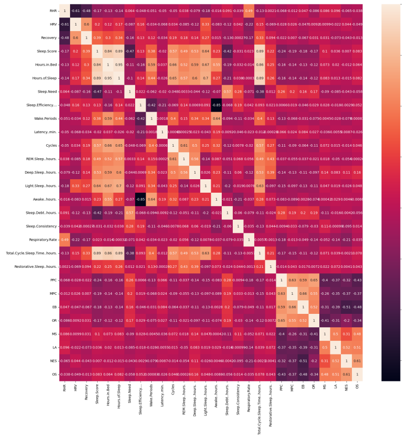


Fig. 1. Correlation Matrix before Imputation

Following the imputation of more than half of the total data points, measures were taken to preserve the correlation structure of the dataset. This approach ensured the imputed values maintained the inherent feature relationships, upholding the integrity and fidelity of the original data distribution.

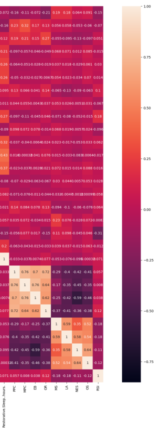


Fig. 2. Correlation Matrix after Imputation

B. Feature Importance Analysis

Analysis of feature importance using a RandomForestRegressor model highlighted significant predictors of the RSImod variable. Features such as sleep patterns, training details, and cardiac rhythm patterns emerged as key determinants of player readiness and performance. Targeted imputation is performed on the identified high-impact features to increase the efficiency and accuracy of the data imputation process.

Some of the main features include Respiratory Rate, HRV, Sleep need, MPC, RHR, Recovery,etc.

C. Model Performance

The XGBClassifier model demonstrated high accuracy in classifying players based on their RSI quartiles, achieving excellent performance in predicting RSI values for the upcoming week. Furthermore, the model accurately forecasted the next match lineup, providing valuable insights for strategic decision-making.

Our results showcase the model's remarkable predictive capabilities, demonstrating the ability to forecast individual athletes' RSI (Relative Strength Index) values with an exceptional accuracy of 97.47% and an F1-score of 0.98. Furthermore, the model exhibited strong performance in predicting RSI values for the upcoming week, achieving an accuracy of 88.23% and an F1-score of 0.88.

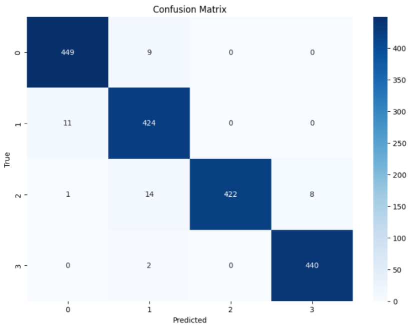


Fig. 3. Confusion Matrix for classification of players

D. eXplainable AI (XAI) Insights

SHAP analysis provided interpretable insights into the underlying mechanisms driving model predictions. By visualizing the impact of different features on outcomes, SHAP analysis enhanced transparency and understanding, enabling teams to make informed decisions.

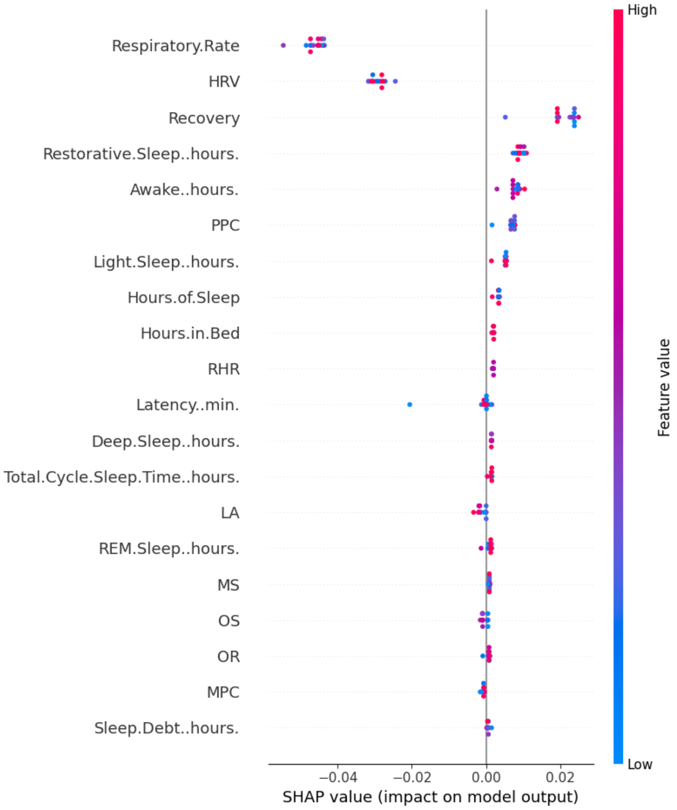
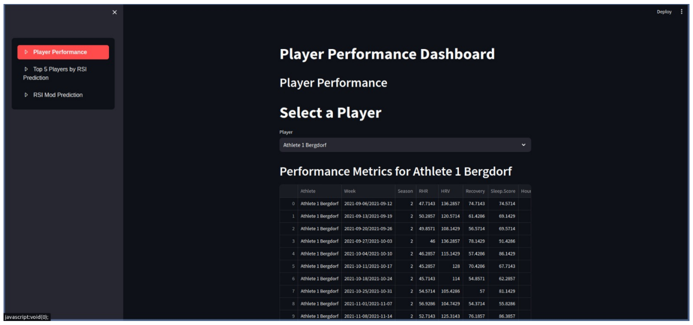


Fig. 4. SHAP value: impact of features

E. Dashboard Visualization

We developed Streamlit dashboard to offer an intuitive platform for visualizing and exploring the data, facilitating interactive analysis and feedback propagation. Through graphical representations and interactive features, the dashboard can empower coaches and athletes to gain actionable insights and optimize performance strategies.



Player Performance Dashboard

Player Performance

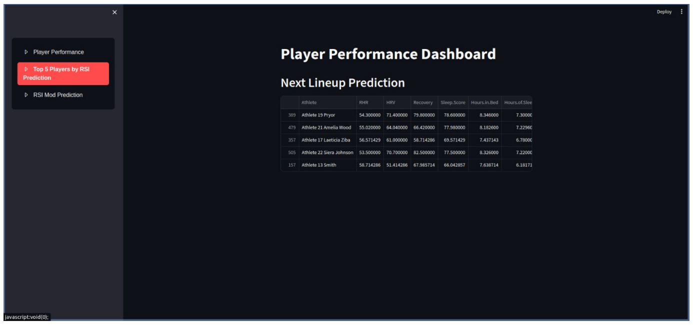
Select a Player

Player: Athlete 1 Bergdorf

Performance Metrics for Athlete 1 Bergdorf

	Athlete	Week	Season	RHR	HRV	Recovery	Sleep.Score	Hour
0	Athlete 1 Bergdorf	2021-09-06/2021-09-12	2	47.7143	136.2857	74.7143	74.5714	
1	Athlete 1 Bergdorf	2021-09-13/2021-09-19	2	50.2857	120.5714	61.4286	69.1429	
2	Athlete 1 Bergdorf	2021-09-20/2021-09-26	2	49.8571	108.1429	56.5714	69.5714	
3	Athlete 1 Bergdorf	2021-09-27/2021-10-03	2	46	136.2857	78.1429	91.4286	
4	Athlete 1 Bergdorf	2021-10-04/2021-10-10	2	46.2857	115.1429	57.4286	86.1429	
5	Athlete 1 Bergdorf	2021-10-11/2021-10-17	2	45.2857	128	70.4286	67.7143	
6	Athlete 1 Bergdorf	2021-10-18/2021-10-24	2	45.7143	114	54.8571	62.2857	
7	Athlete 1 Bergdorf	2021-10-25/2021-10-31	2	54.5714	105.4286	57	81.1429	
8	Athlete 1 Bergdorf	2021-11-01/2021-11-07	2	56.9286	104.7429	54.3714	55.8286	
9	Athlete 1 Bergdorf	2021-11-08/2021-11-14	2	52.7143	125.3143	76.1857	86.3857	

Fig. 5. Dashboard: Player Performance

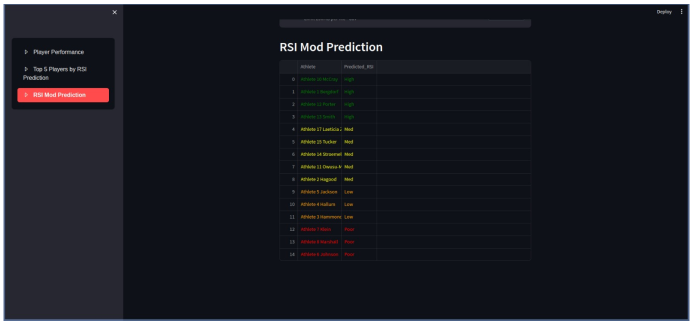


Player Performance Dashboard

Next Lineup Prediction

	Athlete	RHR	HRV	Recovery	Sleep.Score	Hours.in.Bed	Hours.of.Slee
389	Athlete 19 Pryor	54.300000	71.400000	79.800000	78.600000	8.346000	7.30000
479	Athlete 21 Amelia Wood	55.020000	64.040000	66.420000	77.980000	8.182600	7.22960
357	Athlete 17 Laeticia Ziba	56.571429	61.000000	58.714286	69.571429	7.437143	6.78000
505	Athlete 22 Siera Johnson	53.500000	70.700000	82.500000	77.500000	8.326000	7.22000
157	Athlete 13 Smith	58.714286	51.414286	67.985714	66.042857	7.638714	6.18171

Fig. 6. Lineup Prediction



RSI Mod Prediction

	Athlete	Predicted_RSI
0	Athlete 10 McCray	High
1	Athlete 1 Bergdorf	High
2	Athlete 12 Porter	High
3	Athlete 13 Smith	High
4	Athlete 17 Laeticia Z	Med
5	Athlete 15 Tucker	Med
6	Athlete 14 Stroemel	Med
7	Athlete 11 Owusu-M	Med
8	Athlete 2 Hagood	Med
9	Athlete 5 Jackson	Low
10	Athlete 4 Hallum	Low
11	Athlete 3 Hammonc	Low
12	Athlete 7 Klein	Poor
13	Athlete 8 Marshall	Poor
14	Athlete 6 Johnson	Poor

Fig. 7. RSIMod Prediction

IV. Discussion

The findings of this study present a significant advancement in the field of sports analytics, particularly in collegiate basketball. By leveraging predictive analytics and visualization techniques, we have demonstrated the potential to enhance athlete performance and inform strategic decision-making processes. The accuracy of our predictive models, particularly in forecasting RSI values and match line-ups, underscores the value of integrating data-driven insights into sports management practices.

However, it is important to acknowledge the limitations of our study. The accuracy of our models is contingent upon the quality and comprehensiveness of the dataset used. Future research could benefit from incorporating additional variables and datasets to improve model performance and generalizability. Additionally, while our models have demonstrated promising results, further validation and testing in different contexts and seasons are necessary to assess their robustness and applicability across various teams and players.

The integration of predictive analytics and visualization tools into sports management practices represents a significant step forward[5]. By providing actionable insights derived from data-driven analytics, our research aims to modernize sports management practices. This approach not only enhances athlete performance and team competitiveness but also fosters a culture of continuous improvement and innovation within the sport.

Our study highlights the potential of predictive analytics and visualization techniques to revolutionize sports management practices in collegiate basketball. Through the development of sophisticated predictive models, the use of eXplainable AI for transparency, and the creation of user-friendly dashboards, we have demonstrated the value of data-driven insights in enhancing athlete performance and informing strategic decision-making processes. As we continue to explore the capabilities of data analytics in sports, it is imperative to address the challenges and limitations of our current models and methodologies to further advance this field.

V. Conclusion

We have applied techniques to enhance athlete performance management in collegiate basketball. Through the utilization of diverse datasets and sophisticated machine learning models, coaches can predict player readiness indicators, optimize lineup strategies, and make data-driven decisions. The use of eXplainable AI (XAI) through SHAP analysis has provided valuable insights into the underlying mechanisms driving our model's predictions. This transparency is crucial for building trust and understanding among stakeholders, including coaches, players, and analysts. The development of a user-friendly dashboard using Streamlit further facilitates the accessibility and usability of our data-driven insights, empowering coaches and athletes to make informed decisions in real-time.

Analysing player performance is capable of producing a revolution in sports management and training systems. It can be used to analyze data from previous performances and therefore help coaches make an informed choice on team strategy and player selection. Machine Learning can be used effectively in monitoring the on-field progress of players in real-time, supplying the coaches with a very useful information on the players' strengths and weaknesses.

VI. References

[1] The Rise of Predictive Analytics in Professional Sports. Daily Press, 2023. website: https://www.dailypress.net/sponsored-content/2023/04/the-rise-of-predictive-analytics-in-professional-sports/.

[2] Taber, C.B., Sharma, S., Raval, M.S. et al. A holistic approach to performance prediction in collegiate athletics: player, team, and conference perspectives. Sci Rep 14, 1162 (2024). https://doi.org/10.1038/s41598-024-51658-8

[3] Wang, Yuanchen Liu, Weibo Liu, Xiaohui. (2022). Explainable AI techniques with application to NBA gameplay prediction. Neurocomputing. 483. 59-71. 10.1016/j.neucom.2022.01.098.

[4] S. U. Sharma, S. Divakaran, T. Kaya and M. Raval, ”A Hybrid Approach for Interpretable Game Performance Prediction in Basketball,” 2022 International Joint Conference on Neural Networks (IJCNN), Padua, Italy, 2022, pp. 01-08, doi: 10.1109/IJCNN55064.2022.9892583.

[5] Claudino, J.G., Capanema, D.d., de Souza, T.V., Serrão, J.C., Machado Pereira, A.C., Nassis, G.P., et al. (2019). Current Approaches to the Use of Artificial Intelligence for Injury Risk Assessment and Performance Prediction in Team Sports: a Systematic Review. Sports Medicine - Open, 5, Article number: 28. https://doi.org/10.1186/s40798-019-0202-3.