Athlete profiling based on similar characteristics

AU2140066 -Harman Jani, AU2140077 -Kaival Shah, AU2140080 -Divykumar Patel, AU2140113 -Kathan Dave

***Abstract*—In today's time analyzing data has occupied a central role in competing sports like basketball, football, cricket and more, which help evaluate players at various levels, i.e., individual, team and conference levels. The project aims to utilize machine learning techniques to analyze a dataset, considering a wide range of characteristics, including sleep patterns, training details, cardiac rhythm patterns, emotional-mental state information, game scores, weekly readiness scores, and jump data (RSImod). The main goal of this project is to cluster the athletes based on their similar characteristics, thereby making distinct groups of athletes with similar features. This clustering will reveal the underlying patterns and relationships between the various features contributing to the player's performance. Further, the study will employ explainable AI (XAI) methods to explain the characteristics of each group of athletes. This, in turn, will provide insights into the features significantly associated with the cluster.**

Keywords **:** Gaussian Mixture Model, Principal Component Analysis, Unsupervised learning, Training, MICE, Basketball, Machine Learning, Sports Analytics.

I. Introduction

SPORTS Analytics, including data analytics in collegiate, professional, and Esports, has recently gained significant attention. Any insights regarding the athletics performance or the game statistics play a vital role in preventing injuries and ultimately determining the team's overall success rate and the player's overall training. This interdisciplinary research has brought many folks to investigate together a Division-1 basketball team's season where COVID-19 caused unprecedented disruptions. Using features such as daily sleep patterns, HRV, Recovery, Sleep Score, Hours in Bed, and several other features to cluster the athletes based on similar profiling, then explaining it using the XAI.

Explainable AI (XAI) is crucial for dissecting Division I basketball player data, revealing how factors like sleep and emotions affect readiness scores (RSImod). By clarifying model decisions, XAI facilitates actionable strategies for personalized training and recovery plans, optimizing performance through data-driven insights in sports.

II. Methodology

Data Imputation :

Sleep data and questionnaire responses were not available completely, as reasoned below :

1) Incomplete WHOOP strap data due to improper attachment to the wrist.
2) Incomplete surveys, as the athletes only sometimes completed the surveys.

The missing surveys can negatively impact the ML algorithms. Initially, there were about 100 features for both Season 2 (100) and Season 3 (105). We set a threshold in both seasons such that when the 'NA' values in that feature exceed 1000 (Season 3) and 1500 (Season 2). We particularly drop that column in that dataset. After doing the above process, 64 features were found in Season 3 and 36 in Season 2. We took standard features from both seasons. Then, we combined data on each athlete's standard features (23) from Season 2 and Season 3. Then, we calculated the average of each feature for all athletes from both seasons.

Athlete, RHR, HRV, Recovery, Sleep Score, Hours in Bed, Hours of Sleep, Sleep Need, Sleep Efficiency, Sleep Disturbances, Latency (min), Cycles, REM Sleep (hours), Deep Sleep (hours), Light Sleep (hours), Awake (hours), Sleep Debt (hours), Sleep Consistency, Respiratory Rate, Total Cycle Sleep Time (hours), Restorative Sleep (hours), EB (Energy Balance), Sprints

The data imputation techniques such as global mean, cluster mean, conditional local mean, and K-Nearest Neighbor (KNN) fail to reflect the uncertainty about imputed values. A multiple imputation approach, Multivariate Imputation by Chained Equation (MICE), was utilized. This method imputes each feature in the dataset sequentially, allowing the prior imputed data values to be used in the model to predict the following features. Correctly specified MICE can reduce bias and improve analyses for any proportion of missingness. While a MICE imputer is proven to be effective, we can only conclude that MICE is effective in some situations.

This process involves handling a dataset containing information about athletes, including their names and corresponding features. The features are adjusted to fit within a standardised scale ranging from 0 to 1, ensuring uniformity across the dataset. Finally, the modified dataset is saved for further analysis or usage.

GMM (Gaussian Mixture Model) is better than the k-means clustering algorithm because it handles more complex cluster structures and provides probabilistic cluster assignments. Unlike k-means, which assume clusters are spherical and equally sized, GMM allows for clusters of different shapes and sizes, making it more flexible in capturing the underlying data distribution.
Additionally, GMM provides a soft assignment of data points to clusters by assigning probabilities, which is more robust when dealing with data points on the boundaries between clusters. This probabilistic approach of GMM makes it more suitable for situations where data points may belong to

multiple clusters simultaneously, offering a more nuanced understanding of the data.

III. Results

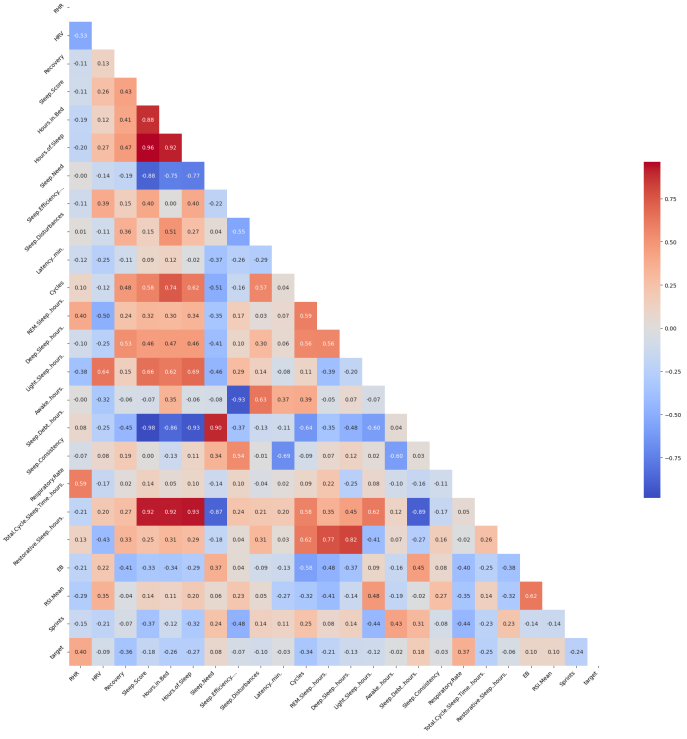

Fig. 1. correlation heatmap of 23 feature

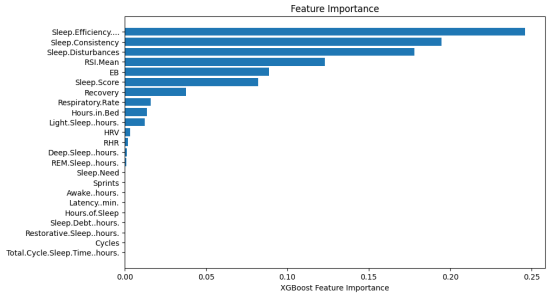

Fig. 2. XGBoost Feature Importance

Top 10 highly correlated feature pairs:

	Feature 1	Feature 2
24	Sleep.Score	Sleep.Debt..hours.
26	Hours.of.Sleep	Total.Cycle.Sleep.Time..hours.
28	Sleep.Score	Total.Cycle.Sleep.Time..hours.
30	Hours.of.Sleep	Sleep.Score
32	Total.Cycle.Sleep.Time..hours.	Sleep.Debt..hours.
34	Awake..hours.	Sleep.Efficiency....
36	Hours.of.Sleep	Hours.in.Bed
38	Sleep.Debt..hours.	Hours.of.Sleep
40	Hours.in.Bed	Total.Cycle.Sleep.Time..hours.
42	Hours.in.Bed	Sleep.Score

Fig. 3. 10 most highly correlated features

Code implementation results in a function to identify highly correlated features. Generates a correlation matrix heatmap using a seaborn heatmap, the mask, and the correct aspect ratio. Also top 10 most highly correlated feature pairs are derived

IV. Discussion

Relevant Links

1) Code implementation - colab link
2) Final Data CSV - link

Data Normalization - Data Normalization has been done using MinMax Scaler and all the features have been normalized in the range of 0 to 1

Feature Selection - After imputing the dataset using MICE and removing the features that have scarce data, using XGBoost, feature importance has been generated. Features of significant importance have been selected, and other features have been dropped. The prominent features would be used further to cluster athletes based on similar characteristics.

After clustering, XAI would be done.

V. Conclusion

This project investigated the feasibility of using machine learning techniques to cluster basketball athletes based on various performance-related features. Data imputation using MICE addressed missing values in sleep and data. GMM clustering was chosen due to its ability to handle complex cluster structures and provide soft cluster assignments. Explainable AI (XAI) techniques will be applied in future work to interpret the characteristics of each player group and the features most influential in cluster formation. This approach can reveal valuable insights into athlete performance and guide the development of personalized training and recovery plans, ultimately optimizing athletic performance.

VI. References

1) 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.*

2) S. Senbel et al., "Impact of Sleep and Training on Game Performance and Injury in Division-1 Women's Basketball Amidst the Pandemic," in IEEE Access, vol. 10, pp. 15516-15527, 2022, *.doi: 10.1109/ACCESS.2022.3145368.*

3) H. Wan, H. Wang, B. Scotney and J. Liu, "A Novel Gaussian Mixture Model for Classification," 2019 IEEE International Conference on Systems, Man and Cybernetics (SMC), Bari, Italy, 2019, pp. 3298-3303,*doi: 10.1109/SMC.2019.8914215.*