Athlete Profiling Based on Similar Characteristics

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

***Abstract*—In today's time, data analysis 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.**

***Index Terms*—Gaussian Mixture Model, Principal Component Analysis, Unsupervised learning, Training, MICE, Basketball, Machine Learning, Sports Analytics, Local Interpretable Model-agnostic Explanations (LIME), Explainable Artificial Intelligence (XAI).**

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.

Utilizing Explainable Artificial Intelligence (XAI) techniques can explain the reasoning behind RSImod predictions. By employing data-driven methodologies, we enable actionable insights that enhance athlete optimization. In our study, we applied LIME, a specific XAI technique, to identify player clusters. This approach allows for detailed explanations to be derived from the models, ensuring the development of data-informed strategies to maximize athletic performance.

II. Methodology

A. Data Preprocessing

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 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.

B. Clustering

Gaussian Mixture Model (GMM) was chosen over k-means clustering due to its ability to:

- Accommodate complex cluster structures.
- Offer probabilistic assignments of data points to clusters.

Unlike k-means, which assume spherical and equal-sized clusters, GMM adapts to varied shapes and sizes, enhancing its ability to capture the underlying data distribution. Additionally, GMM's soft assignment via probabilities is robust at cluster boundaries, which is ideal for scenarios where data points might belong to multiple clusters simultaneously. This probabilistic approach provides more nuanced insights into the data.

The silhouette score was used to assess clustering quality. This score measures how close data points are to their cluster compared to neighbouring clusters, with scores ranging from -1 to 1. Higher scores indicate better cluster separation.

C. XAI

Explainable AI techniques, such as SHAP (Shapely Additive explanations) and LIME (Local Interpretable Model-agnostic Explanations), provide explanations for the predictions made by machine learning models. SHAP utilizes game theory to assign feature importance, while LIME approximates local model behaviour.

Both techniques enhance understanding and trust in the predictions generated by the models. LIME's local, interpretable explanations are particularly suited for GMM due to its probabilistic cluster assignments. This allows us to gain insights into specific data points' clustering decisions and complement the complex nature of GMM.

III. RESULTS

A. Relevant Links

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

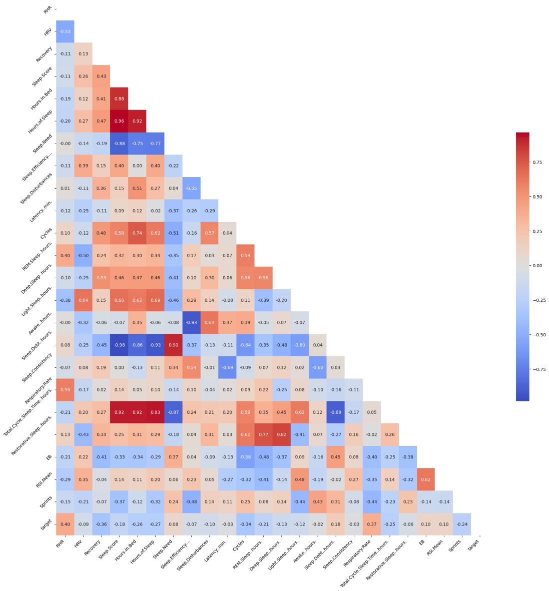


Fig. 1. XGBoost Feature Importance

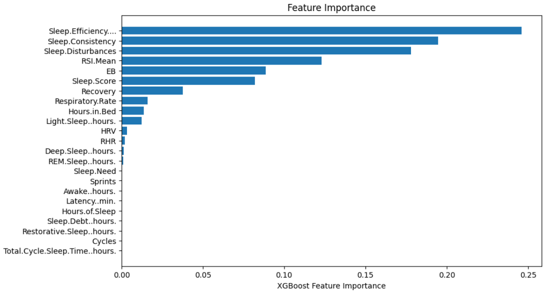


Fig. 2. XGBoost Feature Importance

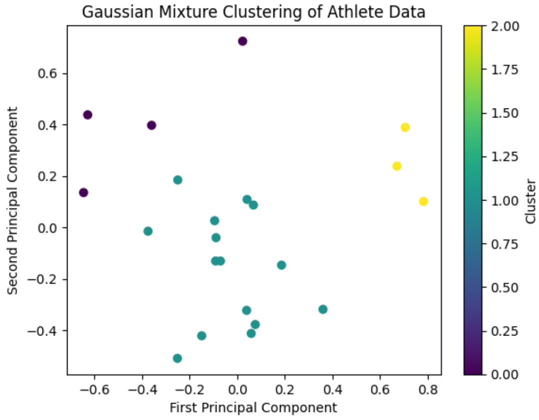


Fig. 3. Cluster

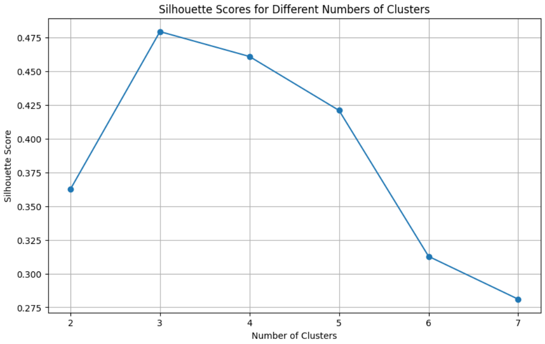


Fig. 4. Silhouette Score After PCA

Clustering	Silhouette
With PCA	3 clusters: 0.47938317319534834
Without PCA	7 clusters: 0.2932979230893263

TABLE I
SILHOUETTE SCORES MAX OBTAINED

After applying PCA preprocessing, our data achieved a remarkable silhouette score of 0.47 for 3 clusters, indicating excellent cluster cohesion and separation. In contrast, the silhouette score without PCA was notably lower at 0.29.

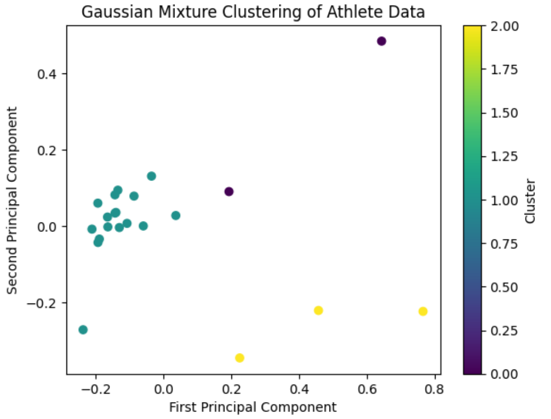
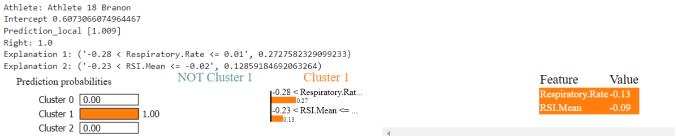


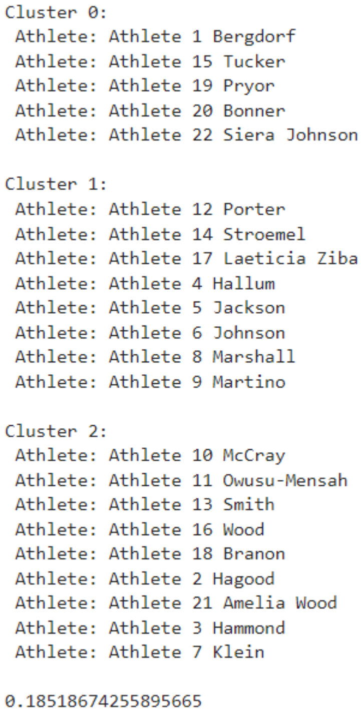
Fig. 5. Atheletes Clustered With PCA of selected two features



```
Athlete: Athlete 18 Branon
Intercept 0.6073066074964467
Prediction_local [1.009]
Right: 1.0
Explanation 1: ('-0.28 < Respiratory.Rate <= 0.01', 0.2727582329099233)
Explanation 2: ('-0.23 < RSI.Mean <= -0.02', 0.12859184692063264)
```

Feature	Value
Respiratory.Rate	-0.13
RSI.Mean	-0.09

Fig. 6. LIME With PCA



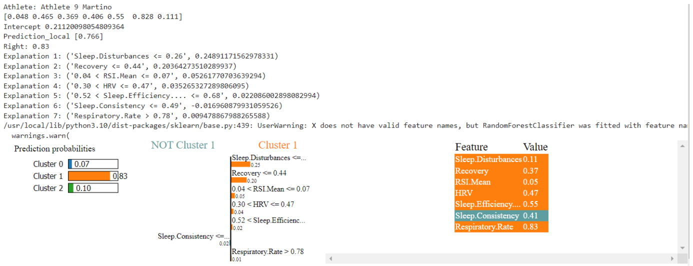
```
Cluster 0:
 Athlete: Athlete 1 Bergdorf
 Athlete: Athlete 15 Tucker
 Athlete: Athlete 19 Pryor
 Athlete: Athlete 20 Bonner
 Athlete: Athlete 22 Siera Johnson

Cluster 1:
 Athlete: Athlete 12 Porter
 Athlete: Athlete 14 Stroemel
 Athlete: Athlete 17 Laeticia Ziba
 Athlete: Athlete 4 Hallum
 Athlete: Athlete 5 Jackson
 Athlete: Athlete 6 Johnson
 Athlete: Athlete 8 Marshall
 Athlete: Athlete 9 Martino

Cluster 2:
 Athlete: Athlete 10 McCray
 Athlete: Athlete 11 Owusu-Mensah
 Athlete: Athlete 13 Smith
 Athlete: Athlete 16 Wood
 Athlete: Athlete 18 Branon
 Athlete: Athlete 2 Hagood
 Athlete: Athlete 21 Amelia Wood
 Athlete: Athlete 3 Hammond
 Athlete: Athlete 7 Klein

0.18518674255895665
```

Fig. 7. Atheletes Clustered Without PCA



```
Athlete: Athlete 9 Martino
[0.048 0.465 0.369 0.406 0.55  0.828 0.111]
Intercept 0.21120098054809364
Prediction_local [0.766]
Right: 0.83
Explanation 1: ('Sleep.Disturbances <= 0.26', 0.24891171562978331)
Explanation 2: ('Recovery <= 0.44', 0.20364273510289937)
Explanation 3: ('0.04 < RSI.Mean <= 0.07', 0.05261770703639294)
Explanation 4: ('0.30 < HRV <= 0.47', 0.035265327289806095)
Explanation 5: ('0.52 < Sleep.Efficiency.... <= 0.68', 0.022086002898082994)
Explanation 6: ('Sleep.Consistency <= 0.49', -0.016960879931059526)
Explanation 7: ('Respiratory.Rate > 0.78', 0.009478867988265588)
/usr/local/lib/python3.10/dist-packages/sklearn/base.py:439: UserWarning: X does not have valid feature names, but RandomForestClassifier was fitted with feature na
  warnings.warn(
```

Feature	Value
Sleep.Disturbances	0.11
Recovery	0.37
RSI.Mean	0.05
HRV	0.47
Sleep.Efficiency....	0.55
Sleep.Consistency	0.41
Respiratory.Rate	0.83

Fig. 8. LIME without PCA

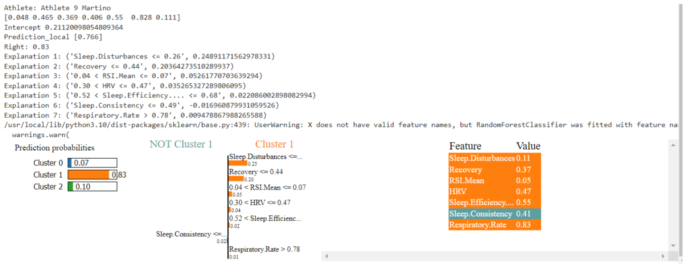
IV. DISCUSSION

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.

features=[’RSI.Mean’, ’HRV’, ’Recovery’, ’Sleep.Consistency’, ’Sleep.Efficiency....’, ’Respiratory.Rate’, ’Sleep.Disturbances’]

Clusters were formed both with and without PCA, yielding optimal results with PCA. Additionally, experimentation with feature combinations, like [’RSI.Mean’, ’Sleep.Consistency’], produced a high silhouette score of 0.80. However, this combination lacks meaningful relevance for clustering objectives. While achieving a high silhouette score, it lacks interpretability and practical significance, making it unsuitable for effective clustering purposes.

Initially, SHAP was explored for Explainable Artificial Intelligence (XAI), but its tree-based nature limited its effectiveness in determinations. Consequently, LIME was adopted for XAI. In analyzing PCA, LIME revealed that RSI.MEAN and Respiratory Rate significantly influenced athlete performance in PCA clustering. Conversely, without PCA, all seven features were utilized for clustering. LIME delineated clustering boundaries, offering insights into how clustering was conducted and feature impacts on clustering outcomes.



```
Athlete: Athlete 9 Martino
[0.048 0.465 0.369 0.406 0.55  0.828 0.111]
Intercept 0.21120098054809364
Prediction_local [0.766]
Right: 0.83
Explanation 1: ('Sleep.Disturbances <= 0.26', 0.24891171562978331)
Explanation 2: ('Recovery <= 0.44', 0.20364273510289937)
Explanation 3: ('0.04 < RSI.Mean <= 0.07', 0.05261770703639294)
Explanation 4: ('0.30 < HRV <= 0.47', 0.035265327289806095)
Explanation 5: ('0.52 < Sleep.Efficiency.... <= 0.68', 0.022086002898082994)
Explanation 6: ('Sleep.Consistency <= 0.49', -0.016960879931059526)
Explanation 7: ('Respiratory.Rate > 0.78', 0.009478867988265588)
/usr/local/lib/python3.10/dist-packages/sklearn/base.py:439: UserWarning: X does not have valid feature names, but RandomForestClassifier was fitted with feature na
  warnings.warn(
```

Feature	Value
Sleep.Disturbances	0.11
Recovery	0.37
RSI.Mean	0.05
HRV	0.47
Sleep.Efficiency....	0.55
Sleep.Consistency	0.41
Respiratory.Rate	0.83

Fig. 9. LIME example

This example showcases how athletes are clustered based on specific criteria, like Recovery ¡= 0.44, assigning them to Cluster 1. It clarifies the reasoning behind each athlete's cluster placement, shedding light on the decision-making process. This insight not only defines the clustering parameters but also highlights the influence of individual metrics, such as Recovery, on clustering outcomes, enhancing understanding of athlete performance patterns and aiding in informed decision-making.

V. Conclusion

This project successfully explored the feasibility of using machine learning to classify basketball athletes based on various performance-related features. The methodology addressed several challenges:

Incomplete Data: Missing sleep data and questionnaire responses were effectively handled using Multiple Imputation by Chained Equations (MICE). This approach preserves the inherent relationships within the data and reduces bias compared to simpler imputation techniques.

- Clustering Choice: Gaussian Mixture Models (GMM) were chosen over k-means clustering due to their ability to handle complex cluster structures, such as those potentially seen in athlete performance data. Additionally, GMM's probabilistic assignments provide more nuanced insights into cluster memberships compared to k-means' hard assignments.
- Interpretability: While SHAP was initially explored for Explainable AI (XAI), its limitations with GMM led to the adoption of LIME. LIME's local interpretability provided valuable insights into how features like Recovery Time and Respiratory Rate influenced athlete clustering, particularly when using Principal Component Analysis (PCA) for dimensionality reduction.

These initial findings demonstrate the potential of using machine learning for athlete classification. Future work will involve:

- Enhanced XAI Analysis: Employing advanced XAI techniques to better grasp athlete group characteristics and key factors driving cluster formation, moving beyond basic feature importance to reveal intricate feature interactions.
- Performance Forecasting: Utilizing athlete clusters to develop machine-learning models for predicting performance metrics and assessing injury risks based on specific cluster traits.
- Tailored Training and Recovery: Translating insights into actionable strategies by creating personalized training and recovery plans tailored to individual athlete groups to optimize 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://www.nature.com/articles/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*.

4) Li, L. (2021, December 10). Principal component analysis for dimensionality reduction. Medium. https://towardsdatascience.com/principal-component-analysis-for-dimensionality-reduction-115a3d157bad

5) How LIME works, Understanding in 5 steps, Openlayer. (n.d.). https://www.openlayer.com/blog/post/understanding-lime-in-5-steps