Athlete profiling based on similar characteristics

Dhairya Shah
Btech. CSE
School of Engineering and Applied Sciences
Ahmedabad, India
dhairya.s4@ahduni.edu.in

Aagam Shah
Btech. CSE
School of Engineering and Applied Sciences
Ahmedabad, India
aagam.s2@ahduni.edu.in

Aayushi Shah
Btech. CSE
School of Engineering and Applied Sciences
Ahmedabad, India
aayushi.s3@ahduni.edu.in

Aanal Dobariya
Btech. CSE
School of Engineering and Applied Sciences
Ahmedabad, India
aanal.d@ahduni.edu.in

***Abstract*—Athlete profiling is crucial for optimizing training plans and performance results. This research presents a framework for athlete profiling based on two essential features: Reactive Strength Index (RSI) and Heart Rate Variability (HRV). We use the elbow method to identify the optimal number of possible clusters. We use the k-means algorithm to divide athletes into discrete groups based on similarities in RSI and HRV. In our study, we use this strategy to identify similar athlete groups, which will allow for individual training regiments. In conjunction with this we can make use of predictive analysis based on these profiles to allow for coaches to make informed decisions on athlete development and enhancement of performance tactics.**

***Index Terms*—k means clustering, RSI mean, HRV**

I. Introduction

In this study, we look at a multi-modal dataset that includes several aspects of Division I basketball players' physiological and psychological characteristics. This dataset contains critical information such as sleep patterns, training specifics, heart rhythm patterns, emotional-mental state information, game scores, weekly readiness scores, and jump data represented by the Reactive Strength Index modified (RSImod) [2].

The goal of this study is to use machine learning techniques on this large dataset to categorise athletes into clusters based on comparable traits. We hope to identify complex patterns and connections in the data that typical analytic approaches may miss. Furthermore, we want to investigate the effectiveness of machine learning approaches in deriving relevant insights from complicated athlete datasets.

II. Methodology

A. Data Processing

The dataset included Division I basketball players' records from Seasons 2 and Season 3 and several physiological and psychological factors. Initially, the dataset had 115 features. However, two main features were chosen to concentrate on the profiling of athletes: Reactive Strength Index mean (RSI mean) and Heart Rate Variability (HRV) [1]. These features were selected because it was found that they have a more significant impact compared to other features in determining athlete readiness and performance. To assure data integrity and consistency, the features with little impact were neglected. The RSI mean data for all athletes was only available on Mondays, and the HRV values were available for all days. To make the data consistent, we took the weekly average of HRV values and entered it on Monday. We worked separately on the Season 2 and Season 3 datasets and then combined the processed datasets. The cleaned dataset is comprised of the RSI mean and weekly average HRV values of the athletes for all Mondays.

B. Feature Engineering

We discovered three key modalities in the dataset using correlation heat maps [1] and descriptive statistics, which provided valuable information regarding feature interdependencies. The correlation heat map indicated higher correlation between the three key modalities: sleep and recovery patterns, training load data, and cognitive state. Then, as part of the

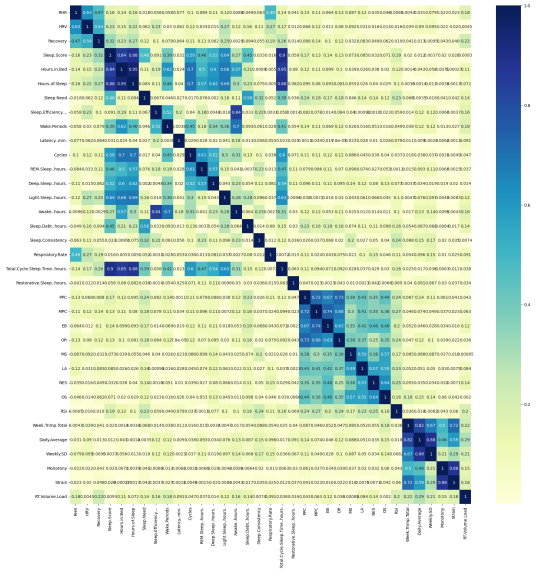

Fig. 1. Correlation Heat Map

exploratory data analysis (EDA), we created a descriptive statistics report to offer a better understanding of the data set variables and the relationships between them. The results showed the relative relevance of each feature in the dataset, with RSI mean and HRV being the most essential features. We selected these two key features to deal with since they were of utmost relevance throughout feature engineering for feature clustering.

C. Clustering

The K-means clustering [**?**] technique was used to categorize athletes based on their RSI mean and HRV characteristics. This approach separates the dataset into 'k' clusters, with 'k' calculated using the elbow technique. The elbow method was used to discover the ideal number of clusters for the dataset by identifying the inflexion point in the within-cluster sum of squares (WCSS) curve.

D. Interpretation

Based on the K-means algorithm's clustering results each athlete was allocated to a specific cluster. This included determining the most relevant feature values within each cluster and their corresponding effects on athletes' weekly readiness ratings, with a special emphasis on RSImod.

III. Results

Table I summarizes the clustering findings for athlete profiling using KMeans clustering, Gaussian mixture model (GMM), and hierarchical clustering. A silhouette score was generated by each clustering technique. This number indicates the quality of the clustering, with higher values denoting better-defined clusters. The silhouette scores of the three algorithms were as follows: KMeans clustering scored 0.4757, quite close behind GMM (0.4732) and hierarchical clustering (0.45262). According to these findings, KMeans clustering might be the best method for athlete profiling given the attributes and dataset available.

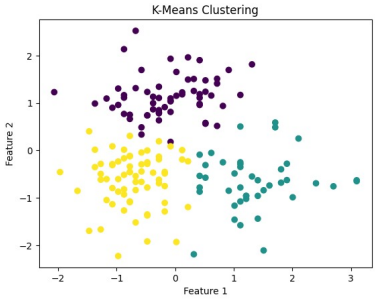

Fig. 2. KMeans Clustering

A. Key Performance Metrics

The Silhouette Score quantifies the performance of clustering techniques. That is Silhouette Coefficient or score is a metric used to calculate the goodness of a clustering technique. Its value ranges from -1 to 1. Values near to 1 mean that the clusters are clearly distinguished and values close to -1 mean that the clusters overlap each other. Whereas values close to 0 mean clusters are indifferent, or we can say that the distance between clusters is not significant. Below, the score is shown for each clustering technique. Due to more similarity-based importance in the features, KMeans outperformed both Hierarchical and GMM clustering.

TABLE I
SILHOUETTE SCORES FOR CLUSTERING ALGORITHMS

Clustering Algorithm	Silhouette Score
Heirarchical	0.4526
GMM	0.4732
KMeans	0.4757

IV. Discussion

After successfully assigning athletes to certain clusters based on the findings of K-means clustering, we aim to analyse each group's XAI-explained features using XAI [2] techniques, the most significant feature values, and their effect on weekly readiness scores (RSImod) [2]. The results will subsequently help us understand athletes' preparedness and performance depending on cluster characteristics.

V. Conclusion

We presented a thorough framework for athlete profiling based on important physiological and psychological features that include 3 significant modalities (sleep and recovery patterns, Training load statistics, and Cognitive state information), that in turn includes heart rate variability (HRV) and the mean of the reactive strength index (RSI mean). Using feature engineering, data processing, and clustering approaches like K-means, an athlete's RSI mean, and HRV were used to group them into different groups. The efficiency of these clustering in defining distinct athlete groupings is demonstrated by the silhouette scores derived from different clustering algorithms. This profiling method provides insightful information for customized training plans and tactics for improving their performance. The study shows how machine learning approaches can be used to maximize performance results and optimize training plans by utilizing multimodal athlete data.

References

[1] Senbel, S., Sharma, S., Raval, M. S., Taber, C., Nolan, J., Artan, N. S., ... Kaya, T. (2022). Impact of sleep and training on game performance and injury in division-1 women's Basketball Amidst the Pandemic. Ieee Access, 10, 15516-15527.

[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).

[3] K means Clustering: GfG. (2023, December 21). K means clustering - introduction. GeeksforGeeks.

[4] Silhouette Coefficient: Validating Clustering Techniques. (May 26, 2020). Retrieved from https://towardsdatascience.com/silhouette-coefficient-validating-clustering-techniques-e976bb81d10c