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*—In sports analytics, athlete profiling is important to optimize training plans and performance results. This research presents a framework for athlete profiling based on five essential features: Reactive Strength Index Mean (RSI Mean), Sleep Disturbances, Awake Hours, Deep Sleep Hours and Sleep Need. We have used the elbow method to identify the optimal number of possible clusters. We have used the k-means algorithm to cluster athletes into discrete groups based on similarities in their features. In our study, we used t-SNE for dimensionality reduction, factor analysis for feature engineering and K-means clustering to identify similar athlete groups, which will allow for individual training regiments. In conjunction with this we have evaluated the clustering using two key performance metrics silhouette score and Davies-Bouldin Index.**

***Index Terms*—k means clustering, factor analysis, feature engineering, silhouette score**

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 three primary modalities in the dataset are sleep and recovery patterns, training load statistics, and cognitive state. The data for the three modalities was collected from WHOOP straps for sleep and recovery patterns, Polar monitors for in-game statistics, and counter-movement jumps.

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. Previously, only two features were selected for clustering: RSI mean and HRV. We attained a silhouette score of 0.4757 in that scenario. In order to increase the score, so that the cluster(s) become clearly distinguished, we considered taking more features. We incorporated the new features in addition to the then two chosen features. We took the weekly average of those features as the values for RSI Mean was available on only one day. If, for certain features, no values existed for an entire week, then we used the MICE [6] imputation technique to impute the weekly average for that week. We chose features based on the correlation matrix, identifying traits highly correlated with the target variable and factor analysis. The final features that we selected were Sleep Disturbances, Awake Hours, Deep Sleep Hours, Sleep Need and RSI Mean.
We worked separately on the Season 2 and Season 3 datasets and then combined the processed datasets. The cleaned dataset comprised of the all the features which are common in both season 2 and season 3.

B. Factor Analysis

We worked with a trial and error approach, experimenting with different feature combinations to make use of the dataset's multi-modality and selecting the feature combinations that resulted in the highest silhouette score. The final features that were selected are Sleep Disturbances, Awake Hours, Deep Sleep Hours, Sleep Need and RSI Mean. Combinations of these features yielded better outcomes and more detailed insights. We kept Sleep Disturbances as individual feature which we have named Feature 1 and combined Awake Hours, Deep Sleep Hours, Sleep Need and RSI mean in the weights of 0.55, 0.1, 0.05 0.30 which we have named as Feature 2. This is how we used two features to get the Silhouette score

of 0.70442.

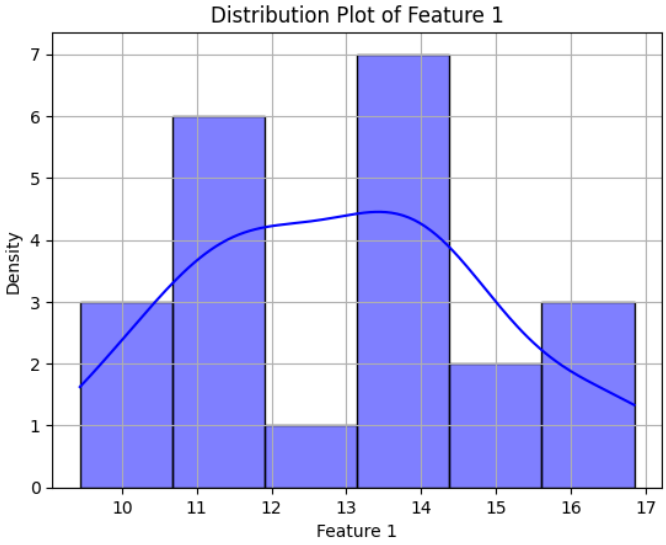

Fig. 1. Distribution of Feature 1

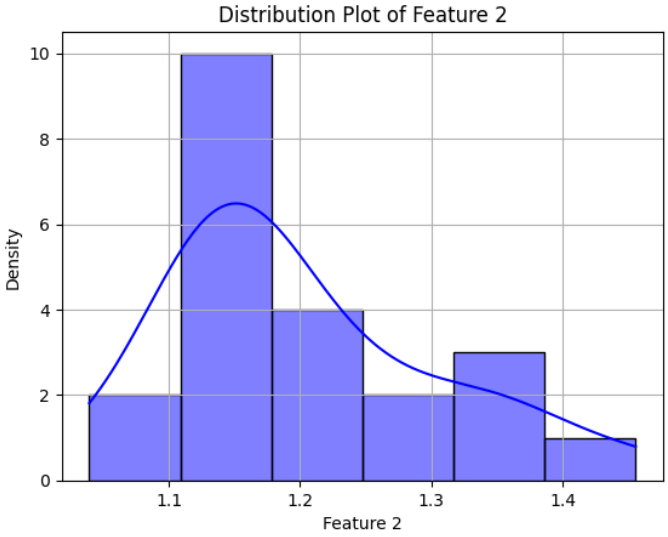

Fig. 2. Distribution of Feature 2

C. Feature Engineering

The dataset had large amount of features or dimensions, therefore the amount of data necessary to produce a statistically meaningful result grows exponentially. As the number of dimensions expands, so does the number of potential feature combinations, making it computationally challenging to get a representative sample of the data and more difficult to accomplish tasks like clustering or classification. We applied the Principle Component Analysis (PCA) [5] approach to reduce the dimensionality of the dataset by finding a new set of variables, smaller than the original set of variables, retaining most of the sample's information that is useful for the classification of data.

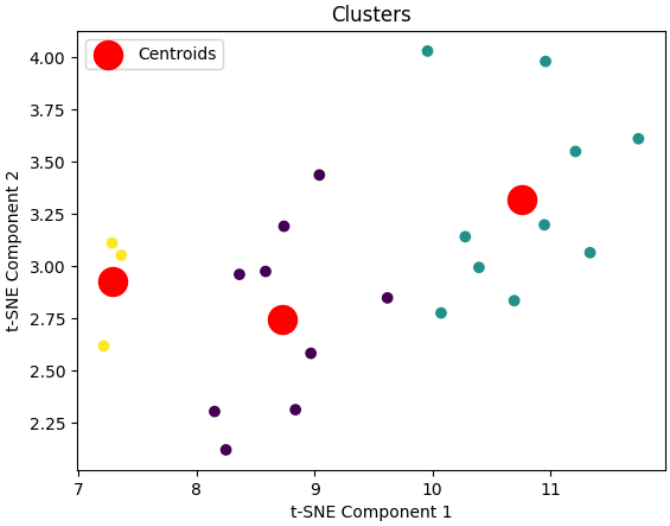

Fig. 3. t-SNE dimensionality reduction and KMeans clustering

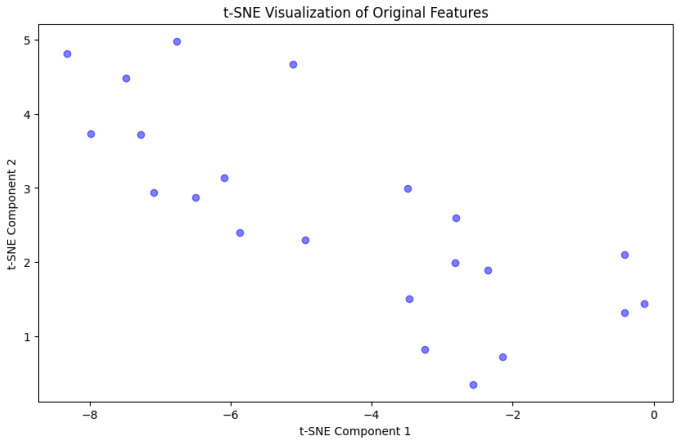

Fig. 4. t-SNE visualization with original features

Then, as part of the 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. Feature 1 consists of sleep disturbances and Feature 2 consists of Awake Hours, Deep Sleep Hours,Sleep Need and RSI Mean. While feature 1 doesn't appear to be a Gaussian distribution, feature 2 appears to be a right skewed distribution.

D. Clustering

The K-means clustering [3] technique was used to categorize athletes based on Feature 1 that consists of sleep disturbances and Feature 2 consists of Awake Hours, Deep Sleep Hours,Sleep Need and RSI Mean. 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. Precisely we used k-means++ to perform initializations for k-centroids. Earlier, we performed clustering with only two feature: RSI mean and HRV which gave the silhouette score of 0.4757. After adding more features by doing factor analysis, we used five features for clustering and converted them into two features (one fully and we used the factors of other four to make one complete feature, the silhouette score increased to 0.7044 with more distinguishable clusters.

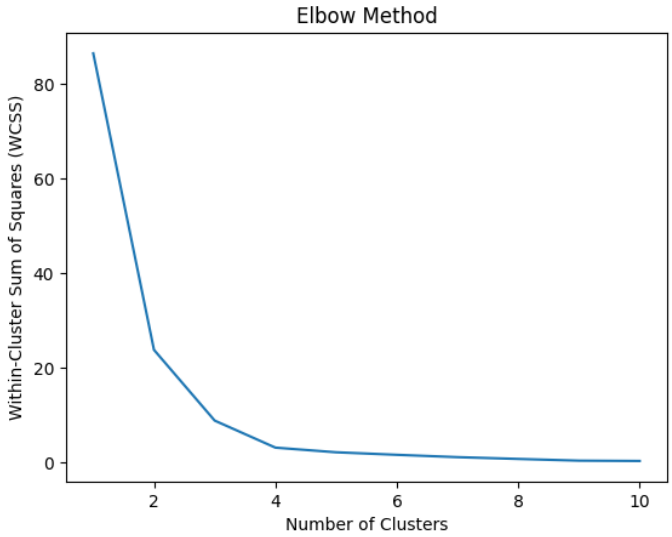

Fig. 5. Elbow method: Optimal number of clusters = 3

E. Interpretation

In our analysis, we found collegiate athletes who sleep less than the recommended hours (6-7 instead of 8 hours) has led to poor performance compared to the athletes who sleep optimal time. We figured out that sleep disturbance impacted the performance of athletes significantly. Apart from that, other sleep features like awake hours, deep sleep hours and sleep need also had an impact on their performance. Reactive Strength Index mean (RSI mean) which is a feature of training load statistics also impacted athlete's performance. Based on the K-means algorithm's clustering results each athlete was allocated to a specific cluster.

From Fig 6., we can see the relation between Feature 1 (Sleep Disturbance) and Feature 2 (combination of Awake Hours, Deep Sleep Hours, Sleep Need and RSI Mean). The three different clusters are colored green, purple, and yellow. We can say that the athletes belonging to the yellow cluster would ideally perform excellent because they have had enough deep sleep hours and have a high value of RSI Mean. RSI Mean (Reactive Strength Index Mean) describes the individual's capability to quickly change from an eccentric muscular contraction to a concentric one. In a sport like basketball, if the athlete can do that efficiently, then he/she is less likely to get injured i.e. more the value of RSI Mean, the less is the chances of the athlete to get injured. Inspite of athlete's sleep getting disturbed, the athletes belonging to the yellow cluster are able to have good deep sleep hours and have high value of RSI. Therefore, we can say athletes belonging to the yellow cluster can be considered as excellent. Now, based on the above argument, when we compare between athletes belonging to green and purple cluster, its clear that athletes belonging to the purple cluster are better than athletes belonging to the green cluster. Athletes belonging to the purple cluster instead of more sleep disturbance are able to have almost same RSI Mean and Deep Sleep Hours compared to athletes belonging to green cluster. So, we can say that athletes belonging to purple cluster are good and subsequently athletes belonging to green cluster as average.

III. Results

Using different feature combinations, we found the best silhouette score for Sleep disturbances and Awake hours (out of all 2 feature combinations), Sleep Disturbances, Deep Sleep hours, and Awake hours (out of all 3 feature combinations), and Sleep Disturbances, Sleep need, Awake hours and RSI Mean (out of all 4 feature combinations). Following that we did a factor analysis on these important features and created two new features with weights to the original features as follows:
feature1Weights: ('Sleep Disturbances': 1)
feature2Weights: ('Awake hours':0.55, 'Deep Sleep hours':0.1, 'Sleep Need':0.05, 'RSI Mean': 0.30)

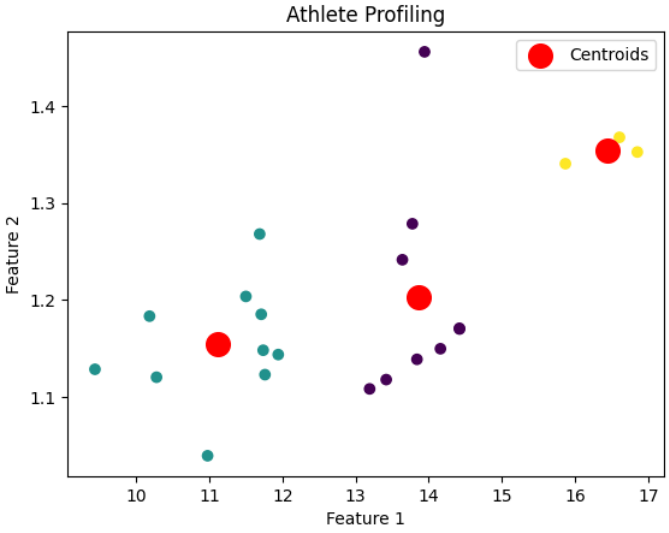

Fig. 6. K-Means Clustering Results

With the help of dimensionality reduction technique t-SNE, feature engineering using factor analysis and KMeans clustering we achieved the silhouette score: 0.7044 and Davies-Bouldin Index: 0.3595.

A. Key Performance Metrics

Table I summarizes the clustering findings for athlete profiling using DBSCAN, KMeans and Ensemble clustering with different features and their factors.

The Silhouette Score [4] 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.

The Davies-Bouldin Index measures the average similarity between each cluster's points, relative to the distance between clusters. Lower values indicate better clustering, with 0 being the best, and higher values indicating worse clustering. We found the DB Index for Athlete profiling to be 0.3595.

TABLE I
SILHOUETTE SCORES FOR CLUSTERING ALGORITHMS

Clustering Algorithm	Silhouette Score
DBSCAN	0.3883
KMeans	0.7044
Ensemble Clustering	0.7044

We are getting same Silhouette score for K-means and Ensemble Clustering, this implies that the structure of the data is such that the algorithms find similar groupings. It indicates that the final dataset we have created is well suited for these clustering algorithms, indicating that the clusters are distinct and well separated.

IV. DISCUSSION

After successfully assigning athletes to certain clusters based on the findings of K-means clustering, we analysed each group's XAI-explained features using explainable AI [2] techniques, the most significant feature values, and their effect on weekly readiness scores (RSImod) [2]. The results helped 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 sleep disturbances, awake hours, deep sleep hours, sleep need 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, sleep statistics like sleep disturbances,awake hours,sleep need and deep sleep hours 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.

VI. FUTURE WORK

Due to unfortunate circumstances (pandemic), we had quite a bit of missing data. We hope to get more data in the next season. If we get an adequate amount of data, we would be imputing very few values compared to the present scenario. We are able to achieve a Silhouette score of approximately 0.7 even when there are very much null values. If we get a decent amount of data, we can further increase the Silhouette score and subsequently classify the athletes in a more clear and distinguishable manner. We mostly focused on sleep parameters and training load statistics. In future, we aim to consider the cognitive features as well and see how the classification gets impacted in such a multi modal (all three modals) scenario.

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

[5] GfG, "Principal Component Analysis(PCA)," GeeksforGeeks, Dec. 06, 2023. https://www.geeksforgeeks.org/principal-component-analysis-pca/

[6] Azur, M. J., Stuart, E. A., Frangakis, C., Leaf, P. J. (2011). Multiple imputation by chained equations: what is it and how does it work?. International journal of methods in psychiatric research, 20(1), 40–49. https://doi.org/10.1002/mpr.329