Athlete Statistics Visualization and Prediction

Deval Darji *AU2140060*, Krishn Patel *AU2140085*, Devanshi Shah *AU2140103*, Vedant Patel *AU2140178*

***Abstract*—The main objective of this project is to develop a machine learning-based approach to analyze the multi-modal data collected from Division I basketball players. The dataset includes their sleep pattern, training details, cardiac rhythm pattern, emotional-mental state information, game score, weekly readiness scores and jump data (RSImod). Our primary objective is to reduce this high dimensional data set consisting of around 40 features into five modalities namely: sleep, training, cardiac rhythm, jump, and cognitive. Feature clustering and dimensionality reduction techniques will be employed to achieve this goal. After that, machine learning models will be trained in order to predict a game score and that would be at last compared to the one calculated from the original set of features. In the end, an interactive dashboard will enable the athlete to visualize their generated scores and also help them improve their performances.**

***Index Terms*—Multi-modal data analysis, machine learning, feature clustering, dimensionality reduction, athlete performance monitoring, data visualization, basketball analytics.**

I. Introduction

USING multi-modal data sources has grown more common in the fields of sports science and athlete performance tracking. An athlete's condition and talents can be fully understood by integrating many data streams, including physiological signals, biomechanical measurements, and performance indicators. However, processing, interpreting, and making decisions are significantly hampered by the enormous complexity of such multi-modal information. This study centers on an extensive dataset that was gathered from Division I basketball players. The dataset includes a wide range of features, such as jump data (RSImod), game scores, weekly readiness scores, sleep patterns, training details, and cardiac rhythm patterns. The large dimensionality of the feature space (40 characteristics) can make analysis and interpretation more difficult, despite the great potential this data provides for understanding athletic performance and well-being.

This study suggests a machine learning-based method to divide the feature space into five interpretable modalities—cognitive, jump, cardiac rhythm, training, and sleep—in order to overcome this difficulty. The most pertinent information will be preserved while the high-dimensional data is converted into a more manageable and understandable representation through the use of feature clustering and model fit-based weighting approaches.

On the basis of the original dataset, machine learning models will also be created to forecast the scores for these five modalities. With the use of this predictive ability, modality-specific scores unique to each athlete will be generated, offering a clear and useful overview of their overall performance and well-being. A well-defined dashboard that offers multiple filtering options will be added to the proposed system to enable coaches, trainers, and athletes to view the original features. Furthermore, distinct graphs will be incorporated to exhibit the produced scores for each of the five modalities, enabling effortless comprehension and tracking of an athlete's advancement and preparedness.

Dataset: 17 Athletes of the Division-I women's basketball team at the Sacred Heart University, CT, USA were the subjects of this study . A total of 6224 records each with 113 features were collected from 17 athletes, 7 days a week for a period of 25 weeks.

II. Methodology

The approach that we will be using will be based on dimensionality reduction and score prediction in order to determine whether the player is fit or not. For cleaning and getting the data ready, we have taken the help of pandas library along with KNN for data imputation. The main principle of this method is that it uses similar data points in order to predict a value.

A. Distance Measurement

The first step is to figure out how far apart other data points in the dataset are from the data point with the missing value (lets call it $x_{missing}$). Although other distance metrics, such as Manhattan, Minkowski, or Hamming distance, can also be utilised, the Euclidean distance is a commonly used one. Calculating the Euclidean distance in a d-dimensional space between two points, x_i and x_j, is as follows:

$$d = \sqrt{\sum_{k=1}^{n} (X_{ik} - X_{jk})^2} \tag{1}$$

B. Finding Nearest Neightbours

Once the distances are computed, the method finds the k closest neighbours of $x_{missing}$, where k is a predetermined number. These are the k data points that are closest to $x_{missing}$ in terms of distance.

C. Imputations

Lastly, the values of the k nearest neighbours are added together to impute the missing value. If the attribute to be imputed is continuous, the aggregation is often performed using an average; if the attribute is categorical, the mode, or most frequent value, is used.

Mathematically, if V is the set containing the values of the k nearest neighbors for a missing continuous attribute, the imputed value, m, is given by:

$$m = \frac{1}{k} \sum_{v \in V} v \tag{2}$$

For a categorical attribute, the imputed value is the mode of the values within V.

This approach makes use of the fundamental presumption that output values are comparable for locations closer together in the feature space. When the assumption is correct, it can result in more accurate imputations than techniques that take the distribution of the data into account, including replacing missing values with the mean, median, or mode.

Now looking into feature reduction, we have used PCA for the same. It is an unsupervised machine learning technique that reduces the dimensionality of complex datasets while retaining essential information and it does it by identifying patterns in data. It first finds orthogonal directions along epoch data varies the most which are also known as principal components. Then following the steps of data normalization, covariance matrix calculation, eigenvalue decomposition, and projection onto principal components, this method transforms the high dimensional space into fewer features which helps for easy interpretation and analysis.

Data Normalization: Normalize the data to have zero mean and unit variance:

$$X_{ij} = \frac{X_{ij} - \mu_j}{\sigma_j} \tag{3}$$

Compute Covariance Matrix: Calculate the covariance matrix of the normalized data:

$$Cov(X) = \frac{1}{m} X^T X \tag{4}$$

Eigenvalue Decomposition: Find eigenvectors V and eigenvalues A of the covariance matrix:

$$C = VAV^T \tag{5}$$

Select Principal Components: Sort eigenvectors by corresponding eigenvalues and select the top k eigenvectors to form matrix W.

Projection onto Principal Components: Project the data onto the subspace spanned by W:

$$Y = XW \tag{6}$$

By doing this, PCA reduces the data dimensionality while maintaining the maximum variance.

Then using Hierarchical Clustering we grouped data points into a hierarchy of clusters without requiring a predefined number of clusters. Where initially each data point is treated as a separate cluster, and pairwise distances between clusters are computed. Clusters are then iteratively merged with the help of linkage matrices. Lastly a dendogram is made and cut at appropriate levels to provide correct clusters

III. Results

Our dataset of 113 features and around 6224 data labels consisted of many anomalies, we tried to overcome it using these following techniques:

A. Data Cleaning

We began our analysis by cleaning the dataset using pandas in Python. We focused on removing rows with null values to ensure data integrity. This step eliminated incomplete entries, improving the quality of our subsequent analyses.

B. Null Value Imputations

Following data cleaning, our focus shifted to addressing missing values within the dataset. Utilizing pandas functionalities, we identified the frequency of null values across the dataset. Subsequently, we implemented the K-Nearest Neighbors (KNN) imputation technique, leveraging neighboring data points to replace missing values accurately. This approach significantly improved the completeness of the dataset by ensuring precise imputation of missing values.

C. Descriptive Statistics

We calculated essential descriptive statistics such as the mean and standard deviation for all features to delve deeper into the dataset. These metrics offered a comprehensive view of each feature's central tendency and variability, enhancing our understanding of the data's distribution and variability.

D. Correlation Analysis

Continuing with dataset processing, we computed a correlation matrix to quantify the relationships between all features. This matrix revealed how variables were interconnected, allowing us to discern patterns of association and dependency among them. This analysis provided valuable insights into the underlying structure of the data.

E. Feature Clustering

We employed hierarchical feature clustering, which is very helpful for examining intricate correlations between features or working with high-dimensional datasets. In a variety of machine learning and data analysis activities, it can help with feature selection, visualisation, and interpretation in addition to enabling the identification of significant feature groupings. This made it easier for us to compile a list of every feature and determine how closely connected each one was.

F. Principal Component Analysis (PCA)

Principal Component Analysis (PCA) was used for feature reduction following preprocessing. We were able to retain as much variety as feasible while transforming the original features into a lower-dimensional space thanks to this technique. We determined which principal components significantly contributed to the variability of the dataset by examining the variance explained by each of the main components.

G. Feature Reduction

To simplify the dataset and preserve its key information, feature reduction was done after PCA. We successfully decreased the dimensionality of the dataset by choosing the principal components with the highest variance explained. Processes for creating models and doing further studies were made easier by this simplified representation.

H. Performance Score Prediction

We then predicted performance scores using the smaller feature set that PCA had produced. We trained prediction models using machine learning methods, such as regression or classification models, to predict performance scores based on the chosen features. Performance measures like recall, accuracy, precision, and F1-score were calculated to assess how well the prediction models worked.

I. Results Summary

Our analysis's findings show how successful the preprocessing methods—such as data cleansing, null value imputation, and PCA-based feature reduction—were. Additionally, the predictive models that were created produced encouraging performance scores, suggesting that the dataset may be useful for tasks involving performance prediction.

These results provide valuable insights into the dataset's characteristics and its predictive capabilities, laying the groundwork for further exploration and utilization in machine learning applications.

IV. Discussions

The statistical component of the data showcases as to how all the labels in a feature vary and gives us more insights into the data .

Analysing the correlation matrix, we come to see that there are five major clusters that can be seen, the patches with the majority of blue gradient show that those features are the most intercorrelated, we will then apply four ML algorithms — PCA, XgBoost, Random Forest, and Factor Analysis to reduce these dimensionalities into five major modalities and will also delete those that are not correlated. Then we will form a score on the basis of those and then compare it with the score calculated with the help of all the features to show that the athlete's performance is mainly based on these five.

We first plotted the heat map of the correlation matrix of our original data post cleaning, imputation and normalization. Until this point we had 96 features. And we could clearly see that a lot of features were highly correlated and were redundant.

So we dropped all the features which were highly correlated (threshold > 0.7), to eradicate the redundancy in the data. By this we reduced our dataset from 96 features to 53 features

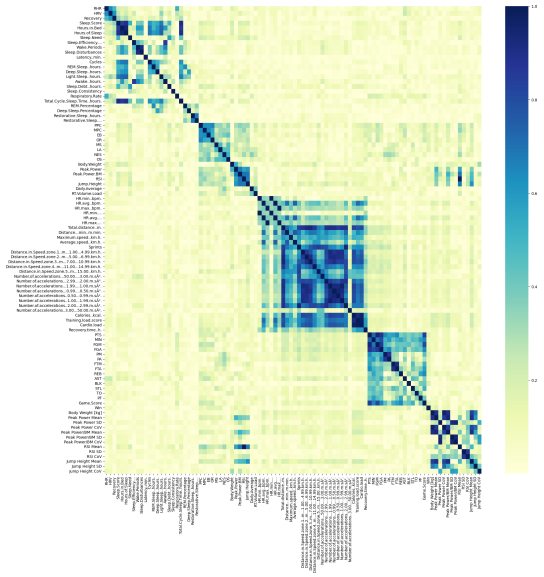

Fig. 1. Before removing highly correlated features

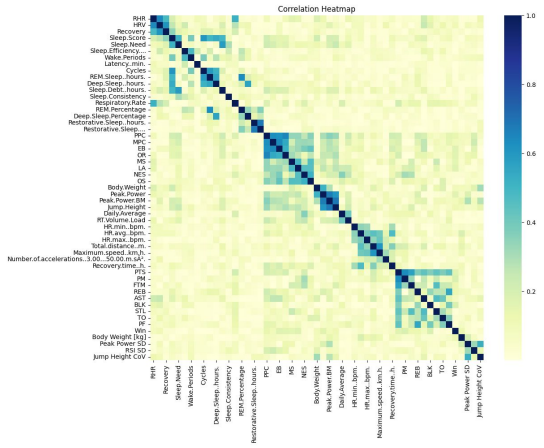

Fig. 2. After removing highly correlated feature

After this, we still needed to reduce the number of features into 5 modalities. So we used **Feature Clustering** for that. We tried different algorithms like KMeans, Hierarchical and GMM, but for our data, with the Hierarchical clustering, we could easily see which features are related and contribute to a single modality.

The hierarchical cluster tree gave us modalities according to their correlation with each other. After applying it on the dataset we got to know as to how many features were there in the dataset and also how closely or distantly they were correlated. Combining this along with the domain knowledge that we had, we started to apply PCA accordingly. We classified the features into the five modalities described to us and then applied PCA which gave us the aggregated scores.

After applying the PCA, we got the five principal components according to our requirements. Then we predicted a performance score using linear regression and

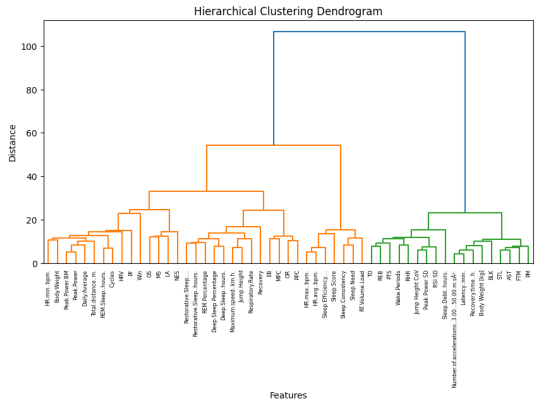

Fig. 3. Hierarchical Clustering

then used the same technique on our intermediate dataset of 52 features. It was found that those scores were somewhat similar which satisfies our efforts done so far.

Lastly taking the RMSE into account, the output received from the PCA dataset and the original dataset are very near in values which showcase that our algorithm is a good fit for the data and the slight difference encountered is due to a little of information loss.

V. Conclusion

Sports being a field that requires constant performance, it is necessary that on the basis of their lives, the coaches can know the performance ability of their athletes. This study attempts to quantify just that and also showcases that that score depends majorly on a few factors. The lesser amount of MSE and the closer the score to that of the original showcases which the model is best fit for our data and proves our theory. This will in the end help coaches to make error-free decisions. We have tried our best to deliver accurate results so that there is no hindrance in decision making.

References

[1] S. Senbel, S. Sharma, M. S. Raval, C. Taber, J. Nolan, et al., "Impact of sleep and training on game performance and injury in division-1 women's basketball amidst the pandemic," *IEEE Access*, vol. 10, pp. 12345-12354, 2022.

[2] C. B. Taber, S. Sharma, M. S. Raval, et al., "A holistic approach to performance prediction in collegiate athletics: player, team, and conference perspectives," *Sci. Rep.*, vol. 14, no. 1, article no. 1162, 2024. doi: 10.1038/s41598-024-51658-8.

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