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. 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. At 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. 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. 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.

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.

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

The initial step in our analysis involved cleaning the dataset to ensure data integrity. Using the Python library pandas, we performed data cleaning procedures, primarily focusing on removing rows containing entirely null values. This process ensured that the dataset did not have incomplete entries, thus enhancing the quality of subsequent analyses.

B. Null Value Imputations

After the data cleaning, we primarily focusing on the missing values in data set. Leveraging the precision of pandas functionality, we determined the frequency of null values across the dataset. Afterwards, we employed the highly accurate K-Nearest Neighbors (KNN) imputation technique, replacing missing values with those derived from neighbouring data points. This method ensured the highest level of accuracy in imputing missing values, thereby enhancing the completeness of the dataset.

C. Descriptive Statistics

We computed key descriptive statistics, including mean and standard deviation, for all features to gain deeper insights into the dataset. These statistics provided a comprehensive overview of each feature's central tendency and variability within the dataset. This type of analysis provides a better understanding of the distribution and variability of the data.

D. Correlation Analysis

Further processing of the dataset involved the calculation of a correlation matrix. This matrix quantified the relationships between all features, elucidating how variables were interconnected. By examining correlations, we identified patterns of association and dependency among variables, providing valuable insights into the underlying structure of the data.

Our results demonstrate the systematic approach for preparing and analysing the dataset, encompassing data cleaning, null value imputation, computation of descriptive statistics, and correlation analysis. These findings create the foundation for further analysis and interpretation, enabling a comprehensive dataset exploration.

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.

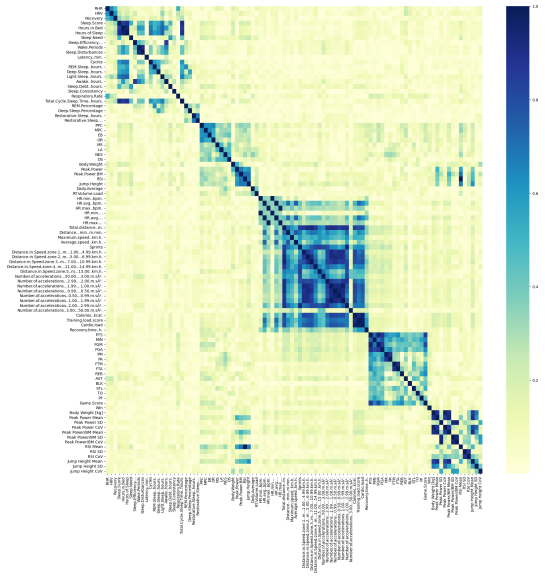

Fig. 1. Caption of the figure.

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 showcase that that score depends majorly on a few factors. The lesser amount of MSE and the closer the score of the model will showcase which one is the best fit for the dimensionality reduction which will at the end support our theory.

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.