Data-driven imputation scheme for human-subject-based dataset

Kashish Jethmalani, AU2140029, Srushti Thakar, AU2140117, Priyal Patel, AU2140204, Riya Patel, AU2140214

***Abstract*—We have to develop a model over the dataset to impute the missing values of the dataset. The dataset was collected during a pandemic-condensed season with unpredictable interruptions to the games and athletic training schedules. The dataset consists of several variables. The data was collected from the Division-1 Women's Basketball team at Sacred Heart University. There are 38 features. 22 Features of the dataset are collected from WHOOP strap. We are mainly focusing on sleep patterns, training details, cardiac rhythm patterns, emotional-mental state, game scores, readiness scores, and jump-data. The data is collected from 16 athletes for 22 weeks. Till now the use of traditional model like MICE is used. We are going to make a new model for the data imputation. We have done K-means clustering for clustering the data on the basis of the features.**

***Index Terms*—Data Preprocessing, Feature Selection, K-means Clustering and Cluster Analysis , Sub-dataset Extraction, Imputation.**

I. Introduction

Missing values in the dataset is a major consideration for any ML model. The missing values can change the decision of the model and can affect its accuracy and decision making power. To avoid making mistakes we try to do data imputation to fill all the missing values in the dataset. The filling of the empty spaces in the dataset is a difficult task as it can add bias in the model according to the values which have been filled in the dataset. We have to be very careful while doing the imputation part.

Traditionally the first approach which comes into the implementation is to rather delete the missing value part from the data set or to replace the missing value part from the mean of the dataset. But removing the missing value part is an appropriate approach as it can lead to data loss and important entries could be lost during this process. Replacing the missing value with the mean is also not correct as sometimes it can be biased as well as there might be the possibility that all the values might not be numbers or mean which is calculated is wrong. It can impact the model on a huge scale.

Here, MICE (Multiple Imputation by Chained Equations) comes into the picture. Ordinarily we utilize MICE for imputation of the dataset. But there are drawbacks related to it. MICE can be computationally expensive, particularly for huge datasets with many factors and a high proportion of missing values. Imputing numerous sets of values and iteratively upgrading them can increment the computational burden, making it illogical for exceptionally huge datasets or real-time applications. MICE depends on specifying appropriate models for imputing missing values for each variable. In case the chosen models are misspecified or improper for the information, it can lead to one-sided imputations and wrong results.MICE accept that lost information happens at arbitrary conditional on observed variables. If information is missing not at random (MNAR), meaning the likelihood of missingness depends on unobserved information, MICE may create biased estimates and invalidate statistical inferences.
And due to these reasons we need to develop new imputation techniques which can overcome these difficulties.

II. Methodology

A. Data Preprocessing

Handling missing values is a critical step in data preparation. The dataset was examined for null values, and two common strategies were employed to address them. Firstly, rows with missing values were either removed entirely or imputed using techniques such as mean, median, or predictive imputation. Secondly, feature scaling was performed to ensure uniformity in the scale of features. Techniques like Min-Max scaling or Standard scaling were utilized to standardize the features to have zero mean and unit variance.

B. Feature Selection

Identifying the most relevant features is essential for effective analysis. In this study, RSI and HRV were identified as the most significant features for clustering athletes based on their physiological data.

C. K-means Clustering

Determining the optimal number of clusters (K) is crucial for meaningful clustering. Techniques such as the elbow method or silhouette score were employed to decide on the appropriate K value. Subsequently, the K-means algorithm was applied to cluster the athletes based on their RSI and HRV features.

D. Cluster Analysis

Analyzing the formed clusters is essential for understanding their characteristics. The mean values of RSI and HRV within each cluster were analyzed to interpret the clusters' significance. Furthermore, the validity of the clusters was assessed through validation techniques, focusing on intra-cluster homogeneity and inter-cluster separation.

E. Sub-dataset Extraction

Assigning cluster labels to each data point facilitated the extraction of sub-datasets corresponding to each cluster. These sub-datasets enabled the exploration of similarities and differences in RSI and HRV among athletes within each cluster.

F. Interpretation and Conclusion

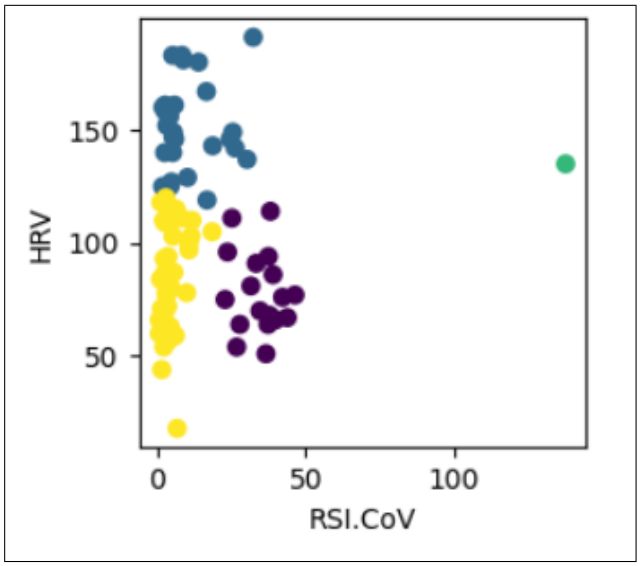
Interpreting the results of the cluster analysis and sub-dataset extraction is crucial for drawing meaningful conclusions. Patterns or trends observed within the clusters were identified, providing insights into athlete characteristics and behaviors.

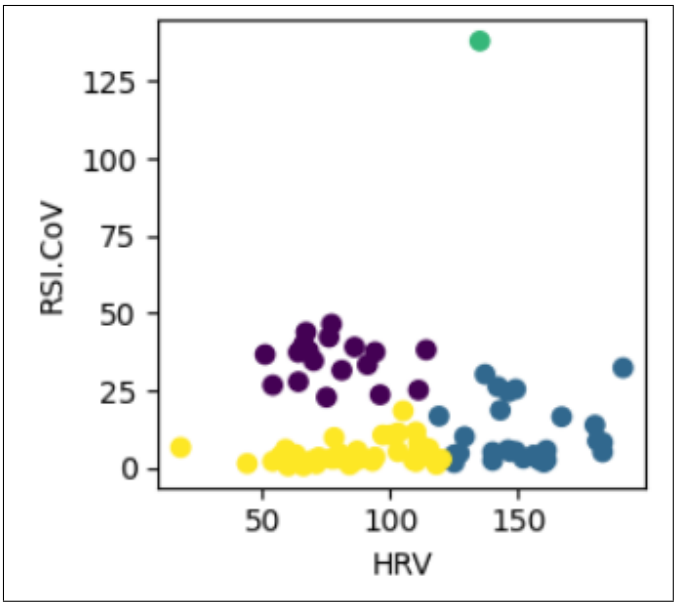
G. Iteration and Refinement

Iterating on the methodology may be necessary to address any limitations or areas for improvement identified during the analysis. Parameters such as feature selection, clustering algorithm, or the number of clusters were adjusted accordingly. Refinement of the analysis was based on feedback and additional insights gained from the initial analysis, ensuring the robustness and validity of the findings.

III. Results

The K-means clustering analysis was conducted on the season 3 athlete dataset. Clustering was done on a four-dimensional dataset with features ”RSI.Mean”, ”RSI.SD”, ”RSI.CoV”, ”HRV”. Our objective was to identify potential patterns or grouping of athletes based on the most significant features(HRV and RSI). We grouped the athletes into 4 clusters, as shown in fig. The figure shows two different dimensions of the same parameters.





Furthermore, clusters with higher HRV values may correspond to states of increased parasympathetic activity, whereas clusters with lower HRV values may indicate greater sympathetic dominance. Similarly, variations in RSI values within clusters may reflect differences in sympathetic nervous system modulation.

The K-means clustering analysis yielded distinct clusters based on HRV and RSI measurements. Each cluster exhibited characteristic patterns of HRV and RSI values, reflecting different physiological states or conditions.

IV. Discussion and Summary

The methodology involved handling missing values by removing or imputing them, and standardizing features through scaling. RSI and HRV were identified as key features for clustering athletes. The K-means algorithm was used to form clusters, with the optimal number of clusters determined using techniques like the elbow method. Cluster analysis included examining mean RSI and HRV values within each cluster. Sub-datasets were extracted for each cluster to explore similarities and differences among athletes. The results revealed four distinct clusters based on RSI and HRV, providing insights into athlete characteristics. Iteration and refinement of the methodology were conducted based on feedback and insights gained.

References

Impact of Sleep and Training on Game Performance and Injury in Division-1 Women’s Basketball Amidst the Pandemic. (2022). IEEE Journals Magazine — IEEE Xplore. https://ieeexplore.ieee.org/document/9690164

Taber, C. B., Sharma, S., Raval, M. S., Senbel, S., Keefe, A., Shah, J., Patterson, E., Nolan, J. K., Artan, N. S., Kaya, T. (2024, January 12). A holistic approach to performance prediction in collegiate athletics: player, team, and conference perspectives. Scientific Reports. https://doi.org/10.1038/s41598-024-51658-8