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. Addressing missing values in datasets is important for maintaining the integrity and accuracy of machine learning models. Traditional imputation methods such as Multiple Imputation by Chained Equations (MICE) have more computational intensity and can have potential bias. Here we tried to make a data-driven imputation scheme for human-subject-based datasets. Our approach integrates clustering techniques and XGBoost regression to handle missing values effectively. Initially, missing values are imputed using an Iterative Imputer, followed by dividing the dataset into clusters using K-means clustering. XGBoost regression models are then trained on each cluster to capture intricate relationships between features and the target variable. The weighted averaging prediction function combines predictions from cluster-specific models, weighted by the similarity of new records to each cluster centroid. Further we have checked our model and got comparable performance to MICE.**

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

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 not 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. 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

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. 22 Features of the dataset are collected from WHOOP strap. The data is collected from 16 athletes for 22 weeks. It also contains few features from sleep, recovery, training, perception (survey), game performance and injury.

We were provided with 2 excel sheets for each season. First we convert Vertical Jump Season 2.xlsx (multiple sheets) into a format similar to Vertical Jump Season 3.csv. After that we merge the Vertical Jump Season 2.csv into Season 2 with Polar.csv file using date and athlete name as reference manually. We repeated same for season 3 data. The obtained dataset then were having alot of missing values. The dataset was having more than 50 percent missing values.

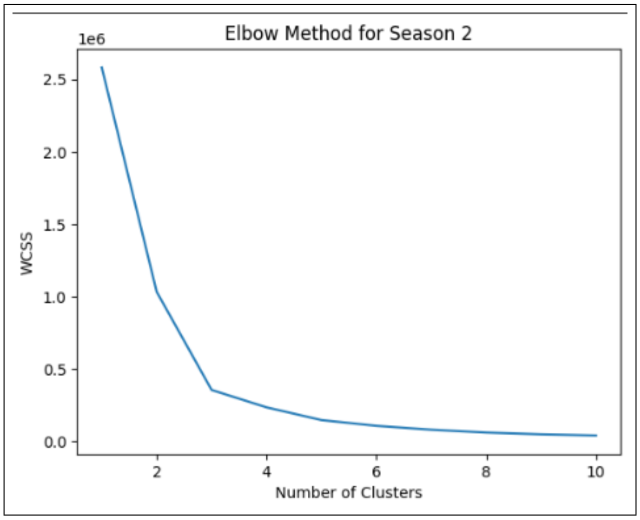
Handling missing values was a critical step in data preparation. The dataset was examined for null values, and only common rows for both seasons with null values less than 50 percent were preserved. Secondly, feature scaling was performed to ensure uniformity in the scale of features. Techniques like Min-Max scaling were utilized to standardize the features.

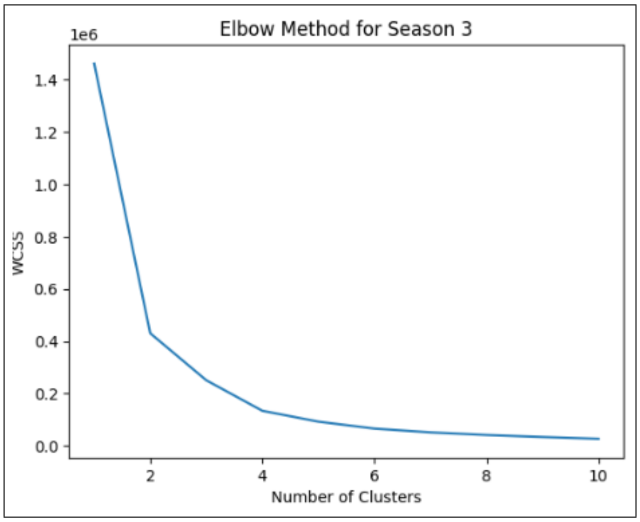
B. Feature Selection

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. *Elbow method*

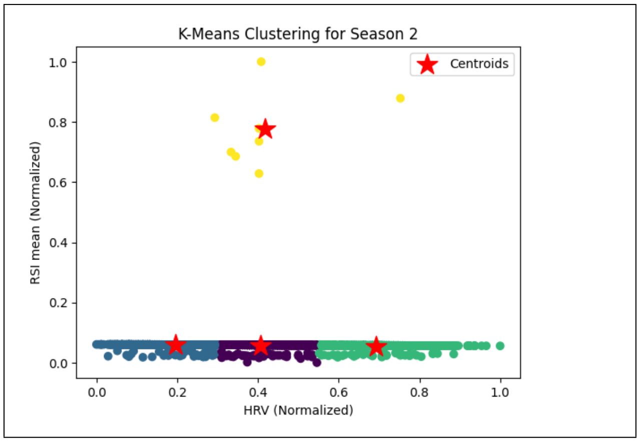
In this step, we applied the elbow method to find the best number of clusters (K) for K-means clustering on preprocessed data from Seasons 2 and 3. It calculates the within-cluster sum of squares (WCSS) for various K values and identifies the ”elbow” point, indicating the optimal balance between model complexity (number of clusters) and explanatory power.

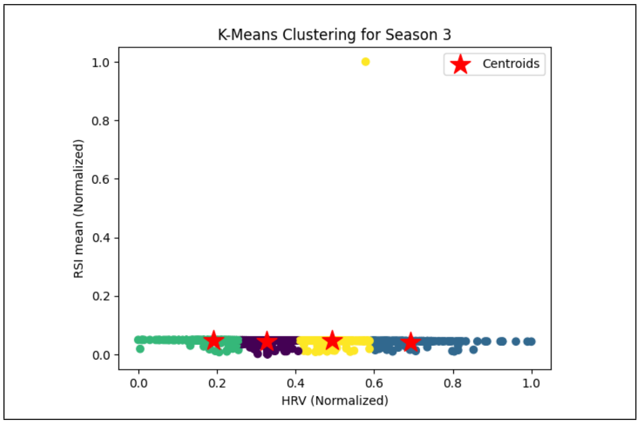




D. *K-means Clustering*

Next, we applied the K-means clustering algorithm with a value of K=4, determined from the elbow method, to partition the data into four distinct clusters. This approach ensures that the clustering is based on the optimal trade-off between model complexity and the ability to explain variance in the data.

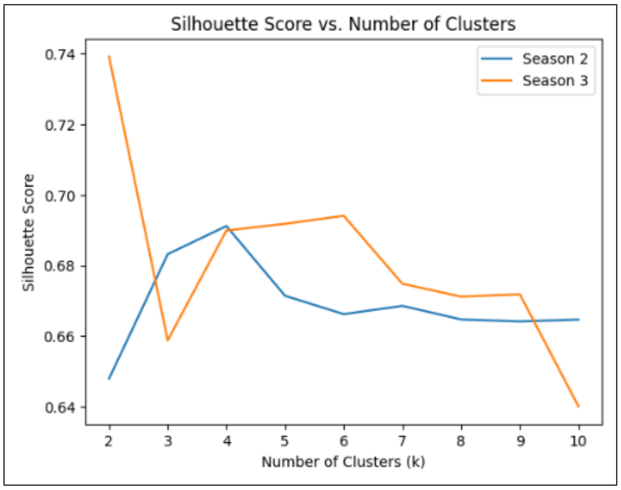




E. *Cluster Analysis and its Validation*

We analyzed the clusters by calculating the mean values of RSI and HRV within each cluster to understand their significance.

For the validation of clusters, we applied silhouette scores for both Season 2 and 3 in K-means clustering. Silhouette scores measure the similarity of each object to its own cluster compared to other clusters, ranging from -1 to 1. Higher scores indicate better cluster fit. The plotted scores help justify the choice of 4 clusters as optimal for the datasets.



F. Sub-dataset Extraction

This step involved assigning cluster labels to each data point, allowing for the extraction of sub-datasets corresponding to each cluster. These sub-datasets were then used to explore similarities and differences in RSI and HRV among athletes within each cluster.

G. XGboost models on each subdataset

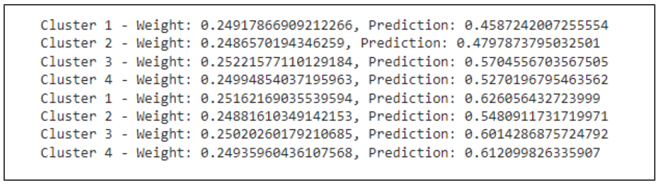
This step involved training XGBoost models on each sub-dataset extracted from the clustered data for Season 2 and Season 3. Each model was trained to predict the cluster label based on the features in the dataset. The models were trained and evaluated separately for each cluster to analyze the characteristics and differences among athletes within each cluster.

H. Similarity Calculation

In the similarity calculation step, a function is defined to quantify the similarity between a new record and each cluster centroid based on Euclidean distance. Euclidean distance measures the straight-line distance between two points in Euclidean space. By computing the Euclidean distance between the features of a new record and the centroid of each cluster, we obtain a measure of how similar the new record is to each cluster. The inverse of these distances is then calculated to obtain similarity scores, where smaller distances correspond to higher similarity scores. This step enables us to identify which cluster a new record is most similar to, thereby facilitating the subsequent prediction process.

I. Weighted Averaging Prediction

The weighted averaging prediction function leverages the similarity scores calculated in the previous step to predict the target variable for a new record. This function combines predictions from XGBoost models trained on sub-datasets corresponding to each cluster, weighted by the similarity scores of the new record to each cluster centroid. By assigning higher weights to predictions from clusters with higher similarity scores, the model adapts to the specific characteristics of the new record, leading to more accurate predictions. This approach enables us to incorporate information from multiple clusters in a weighted manner, capturing the nuances of the dataset and improving prediction performance.The weights and predictions calculated for season 2 and season 3 are as follows:



```
Cluster 1 - Weight: 0.24917866909212266, Prediction: 0.4587242007255554
Cluster 2 - Weight: 0.2486570194346259, Prediction: 0.4797873795032501
Cluster 3 - Weight: 0.25221577110129184, Prediction: 0.5704556703567505
Cluster 4 - Weight: 0.24994854037195963, Prediction: 0.5270196795463562
Cluster 1 - Weight: 0.25162169035539594, Prediction: 0.626056432723999
Cluster 2 - Weight: 0.24881610349142153, Prediction: 0.5480911731719971
Cluster 3 - Weight: 0.25020260179210685, Prediction: 0.6014286875724792
Cluster 4 - Weight: 0.24935960436107568, Prediction: 0.612099826335907
```

J. Final RSI Calculation

The final RSI (Relative Sleep Index) for a new record is computed using the weighted averaging prediction function for both season 2 and season 3. By applying the weighted averaging prediction function to each season's XGBoost models and their respective cluster centroids, we obtain predicted RSI values for the new record in each season. These predicted RSI values are then compared and potentially combined to derive a final RSI estimate. This final RSI calculation accounts for differences between seasons and leverages information from both to generate a more robust estimate of the new record's RSI, enhancing the reliability of the prediction.

K. Comparison with MICE

The approach described above offers several advantages over standard imputation techniques like Multiple Imputation by Chained Equations (MICE). While MICE imputes missing values based on the observed data and iteratively updates estimates, our approach incorporates cluster-based similarity and weighted averaging to generate imputed values. This allows our method to capture the underlying structure of the data more effectively, particularly in the context of human-driven datasets where patterns may be complex and variable. Additionally, by leveraging XGBoost models trained on cluster-specific sub-datasets, our approach can adapt to non-linear relationships and interactions between features, potentially leading to more accurate imputations compared to traditional methods. Furthermore, the use of cluster centroids enables us to incorporate information from similar records in a principled manner, enhancing the robustness and generalizability of our imputation model. Overall, our approach represents a novel and effective way to handle missing data in human-driven datasets, offering improvements over standard techniques like MICE in terms of accuracy and interpretability.

III. Results

Here, we have performed our method on the dataset of season 2 and season 3. On performing we are getting results as close as MICE. On seeing the results in detail we found that our method gives slightly better results than MICE. Furthermore the use of our algorithm can be tested on other datasets with less amount of data missing to test it.

As our method is cluster based it can offer advantages in terms of feature engineering, dimensionality reduction, interpretability, robustness to missing data, and model flexibility.

IV. Discussion and Summary

The new imputation process introduced in the study represents a significant departure from conventional methods by integrating advanced techniques like clustering and XGBoost regression. Initially, missing values within the dataset are addressed through an IterativeImputer, which employs iterative modeling to estimate missing values based on existing data patterns. Following this, the dataset is partitioned into distinct clusters using K-means clustering, enabling the identification of subsets of data points with similar characteristics.

Subsequently, XGBoost regression models are trained on each cluster to capture the intricate relationships between features and the target variable. These models provide accurate predictions within each cluster, allowing for more nuanced analysis and prediction. Importantly, the weighted averaging prediction function is employed to generate predictions for new records by combining predictions from cluster-specific models, with weights determined by the similarity of the new record to each cluster centroid. This adaptive approach ensures that predictions reflect the underlying structure of the data, leading to improved accuracy and reliability. Overall, the amalgamation of clustering and XGBoost regression in the imputation process offers a robust and versatile solution for handling missing data in human-driven datasets, facilitating more insightful analysis and decision-making.

In conclusion, the development of a new imputation technique for handling missing values in datasets, presents advantages over traditional methods like MICE (Multiple Imputation by Chained Equations). Further research can explore the applicability of our algorithm to other datasets and domains, potentially uncovering additional benefits and further validating its effectiveness in addressing missing data challenges.

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