CSE 523 ML Project

Faculty mentor:

Prof. Mehul Raval

Project Number: 9

Project Title: Data-driven imputation scheme for human-subject-based dataset

Group name: Syntellect

Group member details:

ENROLLMENT NUMBER	**NAME**
Kashish Jethmalani	AU2140029
Srushti Thakar	AU2140117
Priyal Patel	AU2140204
Riya Patel	AU2140214

Progress report : (Week 6)

Date: 15-03-2024 to 21-03-2024

Handling Missing values

After preprocessing, the dataset had more than 50% null values. Handling missing values is a critical step in data preparation. The dataset was examined for null values, and specific strategies were employed to address them. We applied the following approach to ensure reliability.

- Removing columns (features) with null values greater than 50% from both datasets (season 2 and season 3).

- Removed all the columns that were not common in both datasets.
  Finally, we got 24 significant features.

- We kept the RSI feature (RSI mean, RSI cov, RSIsd) even though it has >50% null values as it was significant.

Scaling and Normalization

Further, we normalized the input data. The aim is to transform the dataset's features to a similar scale or range. We calculated normalized values for the input data by subtracting the minimum value from each element of the data frame and dividing the result by the range (maximum value - minimum value) of the data frame.

Principal component analysis (PCA)

Once the data was preprocessed, we applied principal component analysis to find the most significant features. (PCA) is a dimensionality reduction technique that identifies patterns and relationships in high-dimensional data by transforming it into a lower-dimensional space. It does this by maximizing the variance in data.The output variables were game score and athlete performance.

The principal component analysis follows the following steps:

1. PCA first centers the data by subtracting the mean from each feature.
2. Computes the covariance matrix of the centered data.
3. It performs eigenvalue decomposition (or singular value decomposition) on the covariance matrix to find its eigenvectors and corresponding eigenvalues.

4. The eigenvectors with the highest eigenvalues capture the most variance in the data and are chosen as the principal components.

After performing the PCA decomposition we found “RSI mean” and “HRV” as the most significant features in the athlete dataset.

Imputing values using MICE(Multiple Imputation by Chained Equation) and normalizing values:

We used MICE because it preserves Data Structure, avoids biases, improves Statistical Power, handles different types of Variables

Performing K-means Clustering:

- Determining the optimal number of clusters (K) is crucial for meaningful clustering.
- **The elbow method** was 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.

Elbow Method:

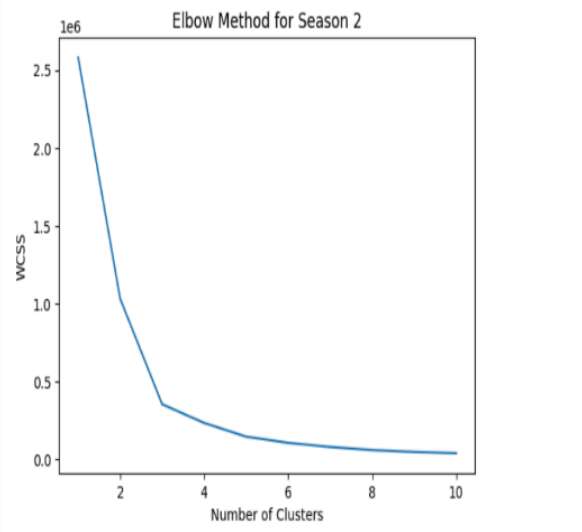
- Heuristic for determining the optimal number of clusters (K) in k-means clustering.
- Involves plotting within-cluster sum of squares (WCSS) against different values of K.
- Identifies the "elbow" point where the decrease in WCSS slows significantly.
- The elbow point signifies the optimal trade-off between model complexity (number of clusters) and the ability to explain variance in data.

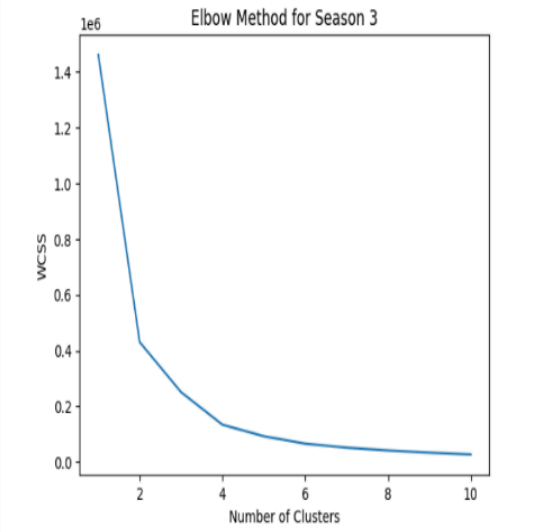
K-Means Clustering:

Unsupervised machine learning algorithm for partitioning data into K distinct clusters. Iteratively assigns data points to the nearest cluster centroid and updates centroids based on the mean of assigned points. Minimizes within-cluster sum of

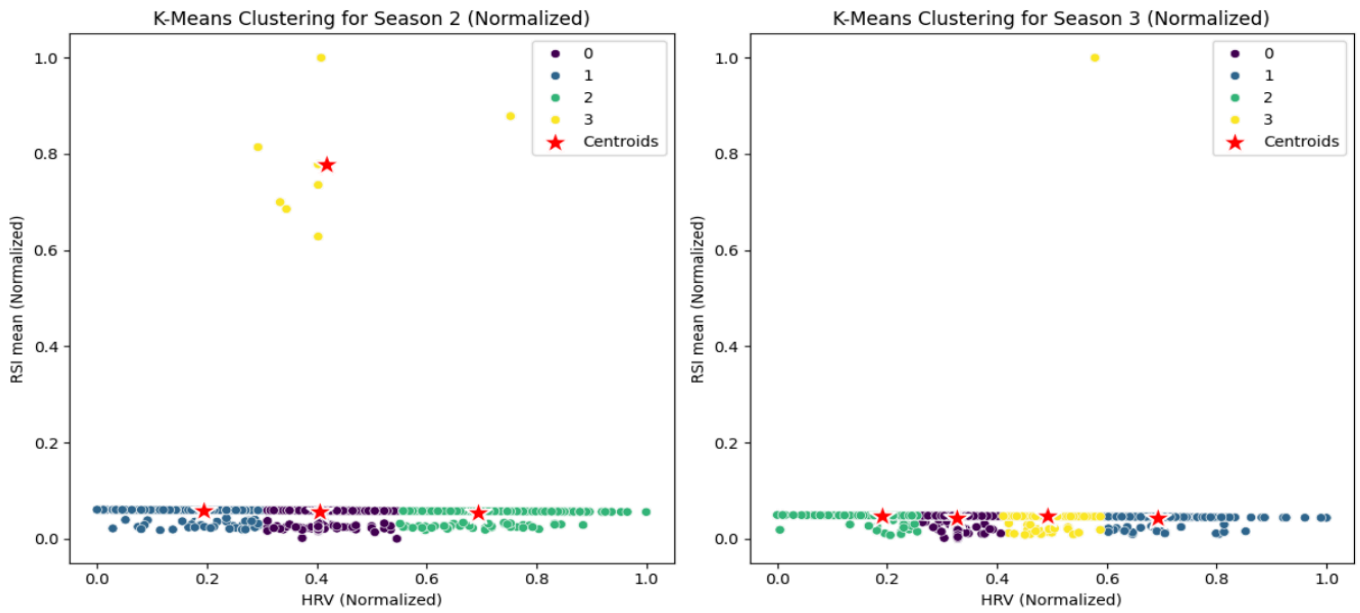
squares (WCSS) to find optimal cluster centroids.Continues iterating until centroids stabilize or a maximum number of iterations is reached.

Results of elbow method and k means clustering:





We choose k to be four as there is no significant decrease in WCSS after k=4.



Here, the clusters represent the following:

- Cluster 0: Data points belonging to this cluster have low HRV and RSI mean.
- Cluster 1: Data points in this cluster have moderate HRV and RSI mean.
- Cluster 2: Data points in this cluster have high HRV and RSI mean.
- Cluster 3: Data points in this cluster have low HRV and high RSI mean.

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.

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.

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.

Further process to be done:

XGBoost Regression:

- Apply the XGBoost regressor separately to each cluster formed in the previous step. XGBoost is a robust gradient-boosting algorithm commonly used for regression tasks.
- Train XGBoost regressor models on each cluster to predict some target variable of interest.