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 8)
Date: 31-03-2024 to 06-04-2024

Our project demonstrates a comprehensive approach to impute missing values, cluster data, and predict missing values for basketball players in different seasons, with an evaluation of the imputation performance using MICE and our custom method.

After changing the scale of the graph, we applied K means ++ on the datasets
Now from the obtained results, we will make 4 different datasets based on the clusters obtained. After making 4 datasets we will train a model on all the datasets.

Our major updates include:

Missing Value Imputation: Imputed missing values using the Iterative Imputer (MICE) for the selected features ('HRV' and 'RSI mean').

Model Training: Trained XGBoost regression models on each sub-dataset created after clustering for both seasons to predict the target variable ('RSI mean').

Prediction and Weighted Averaging: Defined functions to calculate the similarity scores with each cluster centroid, predict using weighted averaging and calculate the final RSI score for a new record provided as an example.

Evaluation: Obtained original missing values and identified rows being imputed.
Then we imputed missing values using MICE (SimpleImputer with 'mean' strategy).
Further, we displayed and compared the imputed values obtained from our method and MICE.
Also, we evaluated the performance of both methods using Mean Absolute Error (MAE), Root Mean Squared Error (RMSE), and R-squared (R^2) metrics.

After that we trained XGBoost regression models on each sub-dataset created after clustering for both seasons to predict the target variable ('RSI mean').

Prediction and Weighted Averaging: We defined functions to calculate the similarity scores with each cluster centroid, predict using weighted averaging and calculate the final RSI score for a new record provided as an example.

We then saved the predicted values to CSV files for further analysis.

We also prepared the code that demonstrates a pipeline for imputing missing values, clustering, training models, and predicting RSI scores for basketball players in different seasons based on their physiological and performance metrics.

```
Final RSI for Season 2: 0.5092125019948113
Final RSI for Season 3: 0.5970152808742315
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
```