**CSE 523 Machine Learning**

**Section 1**

**Group: Decision Makers**

**Project Number 6: Athlete profiling based on similar characteristics**

**Weekly Report**

**Week 7**

**Goal:** Develop a model to cluster athletes based on similar characteristics to help coaches improve player performance.

**Progress till now:**

**Week 1:** Conducted literature survey on athlete performance prediction, including player, team, and conference levels.

· Reviewed paper [1]: "A holistic approach to performance prediction in collegiate athletics..."

· Identified relevant methods like XG Classifier, Random Forest, and MICE (data imputation).

**Week 2:** Continued literature review, focusing on sleep and training data's influence on performance and injury.

· Reviewed paper [2]: "Impact of sleep and training on game performance and injury..."

· Found similar methods used in Week 1 (MICE, SMOTE, XGB, Random Forest).

**Week 3:** Explored a dataset of 16 Division-1 female basketball players, including:

· RSI Mod

· Workload data

· Subjective questionnaire data

· Sleep data

**Week 4:** Discussed methods for handling missing data (null cells) in the dataset.

· Defined a potential model approach: Gaussian Mixture Models (GMM) for clustering.

· This approach considers the probability of a data point belonging to different clusters

**Week 5 :**

* Developed GMM for clustering Sacred Heart University's Division-1 women's basketball player data, analyzing features' impact on RSImod.
* Next steps: impute missing data using MICE, remove outliers, normalize, code the model, evaluate clustering performance, and explore visualization techniques.

**Week 6:**

* We were given the dataset consisting of CSV files with athletes' data over different seasons.
* The main task was to merge the data and create a final CSV that did not consist of null values and had a uniform number of features for all the athletes.
* We dropped some columns that had null values higher than a certain threshold, after that we imputed the remaining columns using MICE.
* Then to merge the values across the seasons for the athletes, we averaged them.
* From the final csv which had the values of 23 features for 21 athletes, then to extract feature importance we applied XGB and CORR.
* Based on the results, we selected 7 features namely 'RSI.Mean', 'HRV', 'Respiratory.Rate', 'Sleep.Efficiency....', 'Sleep.Consistency', 'Sleep.Disturbances', 'Recovery'.

**Progress Summary :**

* First look for the relevant features for purpose of Clustering:

**features = ['RSI.Mean', 'Respiratory.Rate', 'HRV', 'Sleep.Efficiency....', 'Sleep.Consistency', 'Sleep.Disturbances', 'Recovery']**

* Then we computed , GMM for the model with three clusters , but clustering was not done in an appropriate manner .

for n\_clusters in range\_n\_clusters:

# Apply Gaussian Mixture Clustering with the current number of clusters

gmm = GaussianMixture(n\_components=n\_clusters, random\_state=42) gmm.fit(data[features])

* So then we try to implement PCA to apply dimensionality reduction , to plot it in a 2D graph.
* But the PCA relevantly showed characteristics of two features only , thus XAI is very difficult to interpret.
* So we try to compute the Silhouette Score and BIC on the corresponding seven features .

silhouette\_avg = silhouette\_score(data\_pca, data['cluster'])

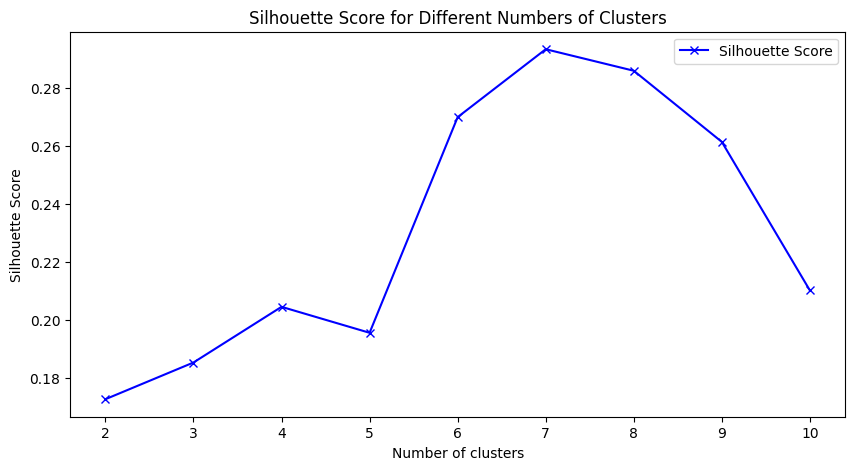
print("Silhouette Score:", silhouette\_avg)

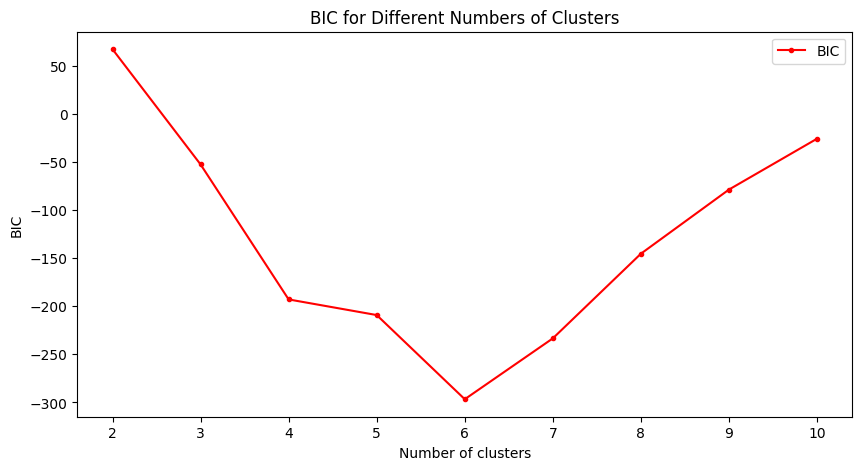
# Compute the BIC

bic = gmm.bic(data[features])

bic\_values.append(bic

* Results of Silhouette Score and BIC





* After it we tried to understand the over plotting of features , using two few relevant feature importance.

**The best combination of features is: ('RSI.Mean', 'Sleep.Consistency').**

**The best combination of features is: ('RSI.Mean', 'Sleep.Consistency')**

**The best silhouette score is: 0.6036496705253964**

**The best number of clusters is: 2**

* Main Problem : Curse of Dimensionality : The curse of dimensionality refers to challenges arising when analyzing high-dimensional data, where the volume of data increases exponentially with each additional dimension. This leads to sparsity, increased computational complexity, and difficulty in visualizing and interpreting data. Techniques like dimensionality reduction mitigate these issues for more effective analysis.