**CSE 523 Machine Learning**

**Section 1**

**Group: Decision Makers**

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

**Weekly Report**

**Week 6**

**Goal:** Derive feature importance to select the most prominent features for clustering the athletes based on similar characteristics

**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.

**Progress Summary :**

**Task**: Merging athlete data from multiple seasons into a single CSV file.

**Data Preparation:**

Dropping columns with null values exceeding a threshold.

Imputing remaining nulls using Multiple Imputation by Chained Equations (MICE).

**Data Merging:**

Averaging values across seasons for each athlete and normalizing them.

Feature Selection:

Using XGBoost and correlation analysis to determine feature importance.

Selecting seven features based on the results:

'RSI.Mean'

'HRV'

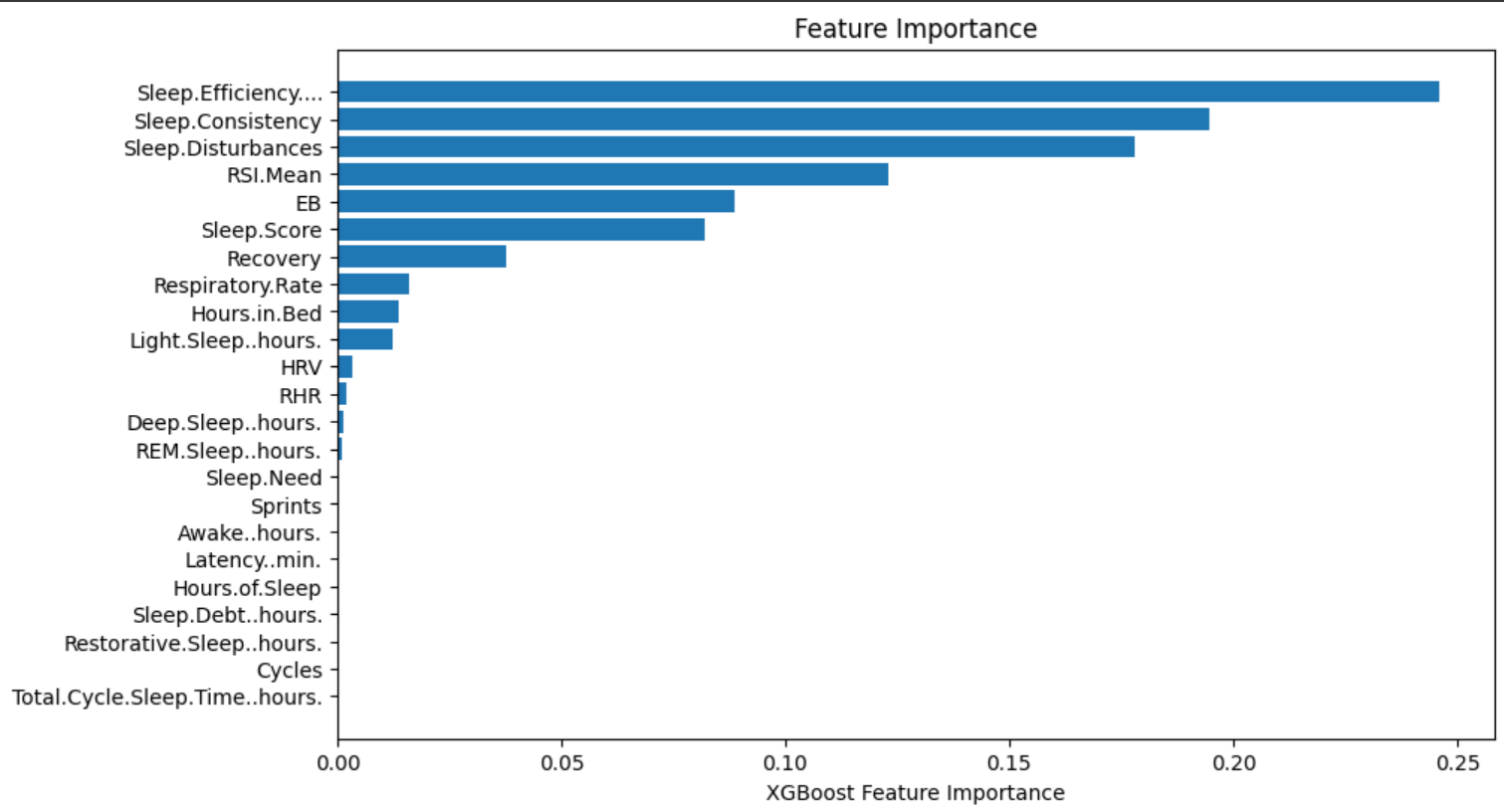
'Respiratory.Rate'

'Sleep.Efficiency....'

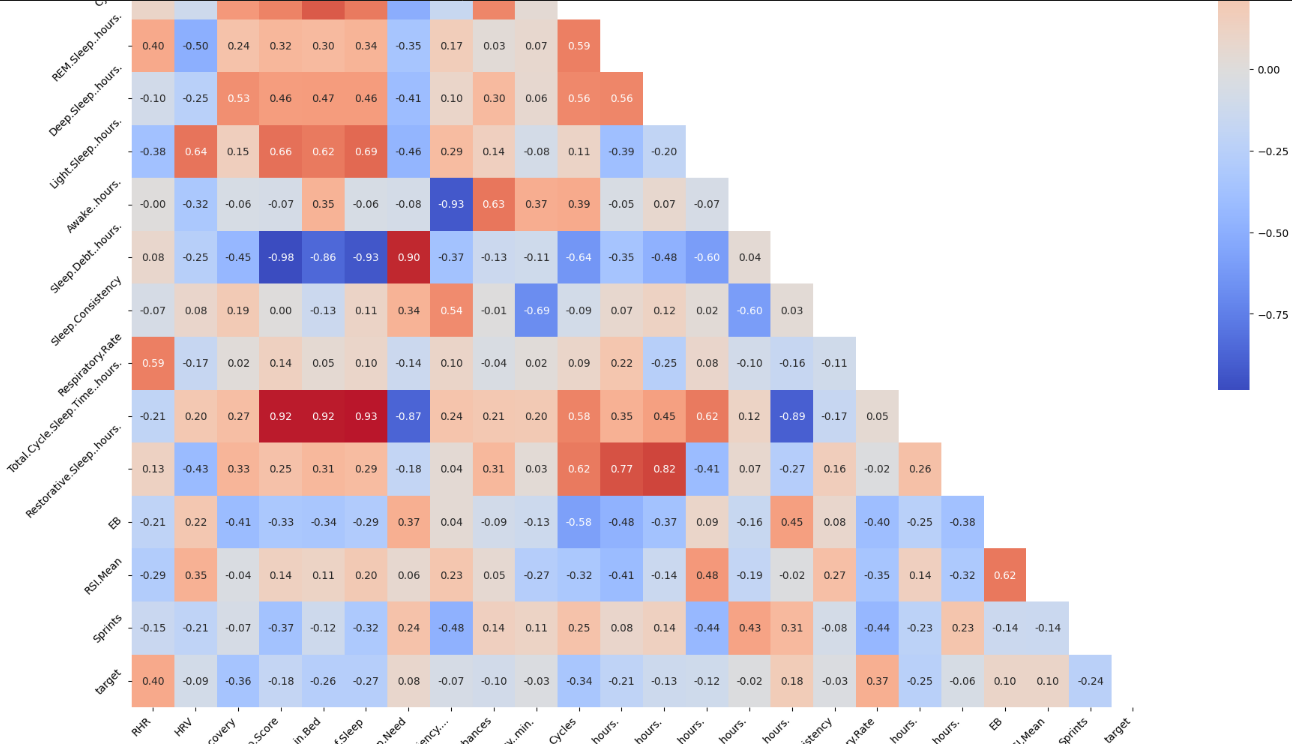
'Sleep.Consistency'

'Sleep.Disturbances'

'Recovery'



Feature Importance using XGBoost



Correlation Matrix