**CSE 523 Machine Learning**

**Section 1**

**Group: Decision Makers**

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

**Weekly Report**

**Week 5**

**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.

**Progress Summary:**

**Model Development**

**GMM model for clustering:**

* Reviewed the concepts of GMM from this source [3].
* GMM is an unsupervised learning technique that assumes that the data point belongs to different clusters and are the mixtures of a certain number of gaussian distributions. The GMM useful in our case where the clusters are not perfectly spherical. It will model the probability of a data point belonging to different clusters based on the features of the athlete.

**Data Analysis:**

* Analysed the provided dataset of Division-1 women's basketball players from Sacred Heart University.
* The dataset includes features related to performance (Game Score, Win), training and readiness (Week.Trimp.Total, Recovery Time), health and wellness (Sleep.Score, HRV, Sleep Patterns), and stress and emotional state (Emotional Balance).
* Focused on understanding the impact of these features (predictor variables) on the Response variable (RSImod). This analysis is crucial for using the data effectively in the GMM model.
* By understanding how the features influence RSImod, we can gain insights into the underlying structure of the data, which will help us determine the appropriate number of clusters and interpret the resulting groupings.

**Next Steps:**

**Data Imputation:**

* Generating the missing values using the MICE algorithm for the sleep patterns data
* Removing Outliners
* Normalization
* Coding the model with data provided
* Evaluate the clustering performance and analyse the characteristics of each identified cluster.
* Explore techniques for visualizing the clusters in a meaningful way.

**Appendix:**

1. A holistic approach to performance prediction in collegiate athletics: player, team, and conference perspectives. Taber, C. B., Sharma, S., Raval, M. S., Senbel, S., Keefe, A., Shah, J., Patterson, E., Nolan, J., Sertac Artan, N., & Kaya, T. (2024). https://www.nature.com/articles/s41598-024-51658-8
2. Impact of sleep and training on game performance and injury in division-1 women’s Basketball Amidst the Pandemic S Senbel, S Sharma, MS Raval, C Taber, J Nolan... - Ieee Access, 2022. <https://digitalcommons.sacredheart.edu/cgi/viewcontent.cgi?article=1180&context=computersci_fac>
3. <https://www.kaggle.com/code/vipulgandhi/gaussian-mixture-models-clustering-explained/notebook>