CSE623 Machine Learning

Project 9: Athlete Profiling NCAA

Group 10

Anant Kedia	AU2220040
Maanit Shah	AU2220043
Samarth Bankar	AU2220001
Ayush Chaurasia	AU2220032
Viraj Bhatia	AU2220038

15th April, 2025

Problem Statement

The objective of this project is to analyze the NEC and MAAC conference game data (including all players and teams) for over the past four years and perform athlete profiling. This framework will be helpful for the recruiters to match player against their team specifications and select accordingly.

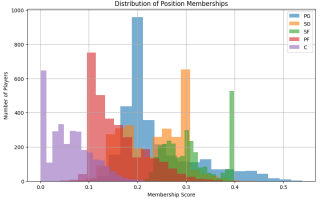
Instructor's Feedback

- **Exploring modelling player synergies with graphs.**
  Play-by-play data scraped from ESPN. However, it did not provide the desired degree of detail.
- **Improving win prediction model accuracy.**
  The ensemble learning win prediction model proposed is able to provide competent predictions with an accuracy of 71.4%.

Approaches

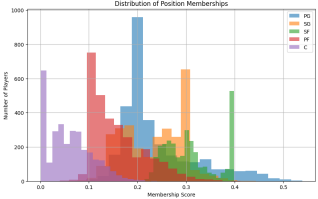
Fuzzy C-means Clustering

We used Fuzzy C-Means (FCM) to cluster athletes based on key performance metrics (FG%, PTS, 3P%, AST, REB, STL, BLK) and assigned soft membership scores to five basketball positions (PG, SG, SF, PF, C).



Approaches

Fuzzy C-means Clustering
We used Fuzzy C-Means (FCM) to cluster athletes based on key performance metrics (FG%, PTS, 3P%, AST, REB, STL, BLK) and assigned soft membership scores to five basketball positions (PG, SG, SF, PF, C).



Spectral Clustering with Feature Selection
We used Spectral Clustering with Feature Selection (SC-FS) to aim for finding graph-based clusters in individual athletes and teams. The algorithm considers all subsets of parameters and ranks the clusterings based on their Silhouette Scores.

Approaches

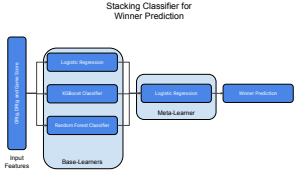
Win Prediction Using Ensemble Learning We used offensive efficiency ($ORtg$), defensive efficiency ($DRtg$) and game score for prediction using a stacking classifier with a random forest classifier, logistic regression and XGBoost as base learners and logistic regression as the meta learner. Here we calculate $ORtg$ and $DRtg$ as,

$$ORtg = (PTS/Possessions) * 100 \tag{1}$$

$$DRtg = (OpponentPTS/Possessions) * 100 \tag{2}$$

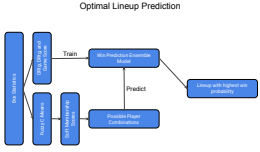
Where possessions are computed using the equation,

$$Possessions = FGA - OREB + TO + (0.4 * FTA) \tag{3}$$



Approaches

Optimal Line-Up Prediciton All possible valid combinations of five players, one from each position (PG, SG, SF, PF, C), are generated for a team. For each line-up, average *ORtg* and Game Score were calculated. Using the trained ensemble model, the win probability against the specified opponent is predicted for each line-up. The line-up with the highest predicted win probability is selected as the optimal line-up.



Results

Spectral Clustering with Feature Selection

We found that every clustering taking 3 or more features into account produced negative silhouette scores. The highest scores were produced by single parameter clusterings. Finally, the few 2-feature clusters with positive silhouette scores used features which does not allow much for interpretation about the athletes and teams.

Results

Spectral Clustering with Feature Selection
We found that every clustering taking 3 or more features into account produced negative silhouette scores. The highest scores were produced by single parameter clusterings. Finally, the few 2-feature clusters with positive silhouette scores used features which does not allow much for interpretation about the athletes and teams.

Win Prediction Using Ensemble Learning
The ensemble model achieved an accuracy of 71.43% and obtained an ROC AUC score of 0.778.

Results

Spectral Clustering with Feature Selection
We found that every clustering taking 3 or more features into account produced negative silhouette scores. The highest scores were produced by single parameter clusterings. Finally, the few 2-feature clusters with positive silhouette scores used features which does not allow much for interpretation about the athletes and teams.

Win Prediction Using Ensemble Learning
The ensemble model achieved an accuracy of 71.43% and obtained an ROC AUC score of 0.778.

Optimal Line-Up Prediciton
When determining the optimal line-up for the Sacred Heart Pioneers against the Arizona State Sun Devils, these were the players selected by the model for each position.

Optimal Lineup:

	Athlete	Position
28320	Sajada Bonner	PG
28317	Nalyce Dudley	SG
11665	Ann Porter	SF
28323	Faith Pappas	PF
28322	Kelsey Wood	C

Future Work

- **Better Clustering Approaches**
  There is a scope to use other clustering approaches suited for high dimensional data such as biclustering, subspace clustering, and domain knowledge based feature selection for dimensionality reduction.

Future Work

- **Better Clustering Approaches**
  There is a scope to use other clustering approaches suited for high dimensional data such as biclustering, subspace clustering, and domain knowledge based feature selection for dimensionality reduction.
- **Incorporating Additional Match Data**
  Analyzing spatio-temporal data from player tracking systems would provide a more granular understanding of player movements, interactions, and tactical patterns, which can be caputred effectively by graph-based approaches.

Future Work

- **Better Clustering Approaches**
  There is a scope to use other clustering approaches suited for high dimensional data such as biclustering, subspace clustering, and domain knowledge based feature selection for dimensionality reduction.
- **Incorporating Additional Match Data**
  Analyzing spatio-temporal data from player tracking systems would provide a more granular understanding of player movements, interactions, and tactical patterns, which can be caputred effectively by graph-based approaches.
- **Better Fuzzy Positioning**
  Fuzzy position profiling approach could be refined by incorporating statistics beyond the standard box score. The position weight matrix can be learned from data rather than prior knowledge.

References

- "The Northeast conference." [Online]. Available: https://northeastconference.org/sports/2024/8/22/gen_aboutnec_2425.aspx
- "The MAAC." [Online]. Available: https://maacsports.com/sports/2017/6/20/GEN_0620175706.aspx
- T. Liu, Y. Lu, B. Zhu, and H. Zhao, "Clustering high-dimensional data via feature selection," *Biometrics*, vol. 79, no. 2, pp. 940–950, 2023.
- NBAstuffer, "Offensive efficiency in basketball explained," 2017. [Online]. Available: https://www.nbastuffer.com/analytics101/offensive-efficiency/
- S. Džeroski and B. Ženko, "Is combining classifiers with stacking better than selecting the best one?" *Machine learning*, vol. 54, pp. 255–273, 2004.
- Scikit-learn, "Stackingclassifier — scikit-learn 1.6.1 documentation," 2025. [Online]. Available: https://scikit-learn.org/stable/modules/generated/sklearn.ensemble.StackingClassifier.html