Project Proposal: Automated Roster Optimization System

Group 4(Navya Sree Yellina, Francina Pali, Roschlynn Dsouza, Saaisathish Sankarabhattar Ayyappan)

1. Project Idea:

The primary objective of this project is to develop an automated system that can analyze a sports team's roster, identify areas of improvement based on age demographics, and recommend potential players to fill those gaps. The system will categorize players into age-based career phases: breakthrough (young and emerging talent), development (growing players with potential), peak (players at their highest performance levels), and twilight (players nearing retirement). The system would analyze age distributions and evaluate roster needs, for instance, identifying if a team lacks players in the "development" phase.

A recommendation engine will then suggest players who can fill these gaps based on historical performance, expected growth trajectory, and team needs. We plan to use machine learning techniques such as clustering and classification, potentially including regression models to predict player potential and collaborative filtering for recommendation. Additionally, we might use a reinforcement learning-based system to optimize suggestions over time based on feedback, such as team performance data after implementing recommendations.

2. Data Requirements:

To complete this project, we will need player data that includes:
- Demographics: Age, nationality, years of experience, etc.
- Performance Metrics: Historical and recent stats (e.g., goals, assists, defensive metrics).
- Injury History: Frequency and severity of injuries could impact recommendations for the "peak" and "twilight" phases.
- Contract and Transfer Data: Information on current team, contract length, and transfer market value.

Data sources could include publicly available sports datasets, sites like Kaggle, and APIs from sports platforms (e.g., ESPN API, Opta, or Transfermarkt). We may also explore scraping data from these sources if APIs are restricted. We will need to navigate access limitations, such as rate limits on APIs, and identify alternative sources if necessary.

3. Deliverable:

The final deliverable will be an interactive web or software application that allows team managers or analysts to upload a roster and obtain recommendations for players to add, based on roster gaps and potential synergies with existing players. This application will provide a visualization of the current age distribution across career phases and display suggested players, with details on their fit to the roster and expected contribution. Code for the machine learning

models and a report documenting our methods, challenges, and findings will also be included in the deliverables.

 4. Project Risks:

- Data Access Limitations: Obtaining detailed, up-to-date player data may be restricted by API access limits or proprietary constraints. We may need to rely on alternative or aggregated data sources, which could limit our model's effectiveness.
- Model Complexity: Balancing model complexity with interpretability will be essential, as team analysts need easily understood recommendations. Reinforcement learning or advanced recommendation systems may add complexity and training time.
- Time Constraints: Developing a full recommendation engine with both predictive and clustering models within the timeline may be challenging, especially if we encounter delays in data access or preprocessing.

 5. Division of Work:
- Data Collection & Preprocessing: Navya, who has experience with web scraping and API access, will lead data gathering and preprocessing.
- Model Development: Saaisathish Sankarabhattar Ayyappan, skilled in machine learning and predictive modeling, will develop the core recommendation and clustering models.
- Software Development & Integration: Roschlynn Dsouza will handle the software interface and integrate the machine learning model with the visualization component, ensuring a smooth user experience.
- Testing & Documentation: Francina Pali will focus on model testing, evaluation, and creating documentation to ensure transparency and reproducibility of the process.