Extending Adila to Professional Football Data: A Case Study with FIFA23

Javiera Paz Azócar Oliva
Pontificia Universidad Católica
javiera.azcar@uc.cl

Pablo Poblete Arrué
Pontificia Universidad Católica
ppobleta4@estudiante.uc.cl

Javiera Belén López Massaro
Pontificia Universidad Católica
javiera.lopezm@uc.cl

Gabriel Acevedo Osorio
Pontificia Universidad Católica
gacevedoo@estudiante.uc.cl

***Abstract*— Team building is a challenge that different industries in different areas are confronted with on a daily basis, making it an issue of great relevance. For this purpose, there are already certain models that seek to maximise the success or efficiency of the team, however, these do not address equity within the teams, an issue that has been on the rise in recent years, for example, with gender parity or with the inclusion of minority groups. In this paper, a dataset of football players with different roles and skills on the field was developed and teams were formed for certain clubs by applying different models such as item-KNN, random, Top-positions and Adila, the latter of which focuses on making teams with gender equity. Our main question throughout this research will be to evaluate how different models behave in contexts where equity and technical performance may conflict.**

Keywords*—*Team Formation, fairness

I. Introduction

Team formation is the process of selecting experts with appropriate skills to collaboratively solve complex tasks. The goal of *automated team formation* is to algorithmically construct teams that are not only effective in terms of utility but also satisfy desirable constraints, such as fairness.

While state-of-the-art neural team formation methods can efficiently analyze massive collections of experts to recommend high-utility teams [?], they often overlook fairness considerations in the team composition process. However, fairness is crucial in collaborative contexts: it fosters innovation, enhances team success, builds stronger communities, reduces conflict, and encourages creative thinking.

Adila is an extension of the OpeNTF framework focused on improving the fairness of neural team formation. In particular, Adila addresses *popularity bias* and other forms of unfairness by applying greedy re-ranking algorithms to the output of pre-trained neural team formation models.

To address fairness challenges, Adila introduces several re-ranking algorithms designed to mitigate popularity bias in the team selection stage. These re-ranking algorithms includes: Deterministic Greedy Rerankers (`det_greedy`, `det_cons`, `det_relaxed`) and Probabilistic Reranker: `fa*ir`

The formation of collaborative and effective teams is a critical challenge across various sectors, from business to education [5]. Traditionally, the selection of team members has been a manual process, often based on subjective criteria or unconscious biases. This non-automated approach frequently leads to imbalances in members' skills, low group cohesion, and ultimately suboptimal team performance [5]. This complexity is magnified in dynamic and high-performance environments such as professional sports, where tactical choice and team composition are crucial and costly tasks if performed inefficiently [3], in addition to having a direct impact on team success [2]. To overcome this inherent complexity and the limitations of conventional methods, research and industry have increasingly turned to the implementation of computational algorithms and automated tools, allowing these intelligent systems not only to minimize errors and time spent but also to open new avenues and opportunities to optimize team composition, improve performance, and maximize player potential in professional sports [5].

The goal of this project is to evaluate the impact of *fairness-aware reranking* on team recommendations generated from FIFA23 player data, focusing on fairness under the notion of Demographic Parity (DP). Starting from base predictions produced by a simple model, we applied the `det_cons` reranking algorithm from the Adila toolkit and assessed its effects using both utility metrics and fairness metrics.

Contributions

This project makes several key contributions toward the development of fairness-aware recommendation systems in the context of professional sports analytics:

1. **Creation of a New Dataset Based on FIFA23:**
   This dataset includes cleaned, deduplicated, and structured player records filtered by team and version. Each player was mapped to a unique identifier and associated with a professional club, enabling the definition of valid team configurations suitable for team recommendation tasks.
2. **Adila-Compatible Data Format Generation:** The FIFA23-derived dataset was successfully transformed into Adila-compatible serialized files: `teams.pkl`, `teamsvecs.pkl`, `indexes.pkl`.
   These outputs fully reproduce the expected data structures originally designed by Adila, despite the absence of complete documentation or code for their generation.
3. **Development of a Reproducible Pipeline:**
   A complete and reproducible pipeline that:
   - Ingests and processes **FIFA23** data.
   - Generates baseline predictions using a simple team selection model.
   - Applies fairness-aware reranking using Adila's `det_cons` method under the Demographic Parity (DP) notion.
   - Evaluates the outcomes using utility and fairness metrics.

   This pipeline allows future research to be conducted without dependence on proprietary data or undocumented scripts.
4. **Metric Collection and Evaluation:**
   The results were aggregated across three cross-validation folds, visualized with comparative plots, and interpreted to identify potential trade-offs between fairness and utility.
5. **First Iteration Toward a Larger Research Agenda:**
   The present work constitutes a solid first iteration of a broader research agenda in fair team recommendation for professional football. It lays the foundation for future experimentation with advanced reranking models (e.g., BNN, Fa*ir), multiple fairness criteria, including intersectional attributes combining popularity, skill level, and demographic information.

We demonstrates that it is feasible to apply fairness-aware reranking pipelines to a new doimain like professional football, even starting from incomplete or undocumented codebases. The tools and methods developed here are a valuable foundation for continued research in the future.

II. Methods

The pipeline operates as follows for each `.pred` file (base model predictions): (1) Loads base team predictions from the file `fpred`, generated by an initial recommendation model. (2) Loads team-related information from `fteamsvecs`, such as vector-based team representations. (3) Loads data split definitions (training, validation, and test sets) from `fsplit`. (4) Applies a fairness-aware reranking algorithm (e.g., `det_cons`, a deterministic method with fairness constraints) to adjust the team recommendations. (5)Evaluates the results before and after reranking in terms of both utility and fairness metrics. [6].

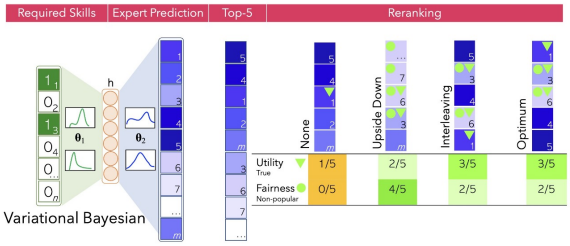

FIGURE I: ADILA PIPELINE

Baseline Predictions

Initial team recommendations were generated using the script `simple_predictor_folds_22.py`, which forms teams of 22 players for each of the three predefined folds (`f0`, `f1`, and `f2`). These predictions were evaluated using utility metrics such as `NDCG@k`, `MAP@k`, and others.

Fairness-Aware Reranking

To improve fairness, we applied the `det_cons` reranking algorithm, which seeks to maximize demographic parity across groups. Post-reranking evaluations included fairness metrics such as `exp` (exposure difference), `expu` (utility-aware exposure), `skew` (representation skew), and `ndkl` (Kullback–Leibler divergence from the fair distribution).

Sensitive Attribute Selection

Although Adila supports multiple sensitive attributes for fairness-aware analysis—including gender and popularity—this project focused exclusively on popularity as the sensitive attribute.

This decision was driven by the availability and quality of data in the **FIFA23** dataset, which allowed for reliable labeling of experts as *popular* or *non-popular* based on their historical frequency of team participation.

Consequently, all fairness metrics and reranking algorithms were applied within the scope of mitigating popularity bias, and no gender-related considerations were included in this phase of the study. This choice aligns with Adila's core objective of mitigating structural biases rooted in expert popularity.

Implementation Pipeline

The entire process was implemented using a reproducible pipeline that:

- **Generates splits and folds.**
- **Generates base predictions for each fold,**
- **Applies deterministic reranking based on popularity bias mitigation,**
- **Computes and exports both utility and fairness metrics,**
- **Visualizes the trade-offs between fairness and utility.**

This pipeline was built and validated to ensure compatibility with Adila's expected data structures and output formats, adapted specifically for the FIFA23 dataset.

A.. Original Dataset

Adila uses the file teams.pkl that organizes teams as structured lists of instances, containing player information, skills, and metadata relevant to use with Adila. This structure facilitates experimentation with team formation models and performance simulation in research environments.

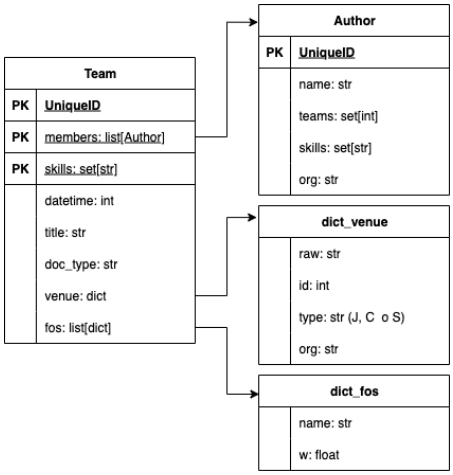

Figure II: Diagram of the data in teams.pkl

Statistic	Value
#Publications (teams)	4,877,383
#Unique Authors (members)	5,022,955
#Unique Fields of Study	89,504
Avg. #Authors per Publication	3.06
Avg. #FOS per Publication	8.57
Avg. #Publications per Author	2.97
Avg. #FOS per Author	16.73
#Publications without FOS	?
#Publications with Single Author	768,956
#Publications with Single Skill	5,569

Table I: Overall statistics of the Adila dataset

B.. New Dataset

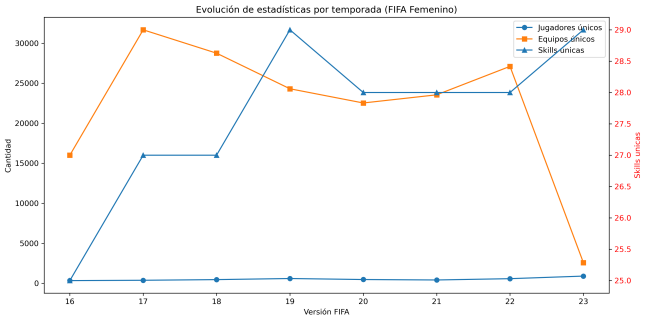

Figure III: Seasonal trends for female players

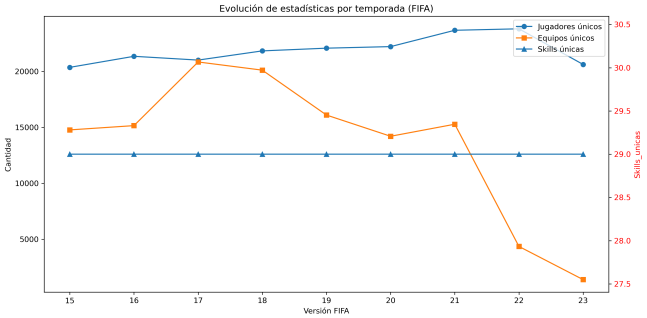

Figure IV: Seasonal trends for male players

Despite the high number of unique players per season, the number of represented teams drops sharply in recent versions. This reinforces the methodological decision to work only with the latest versions (FIFA 22–23), which present a lower data volume but a more consolidated structure.

FIFA Version	#Teams	#Players	#Skills
15	14,774	20,361	29
16	15,156	21,347	29
17	20,842	21,010	29
18	20,111	21,831	29
19	16,101	22,074	29
20	14,206	22,214	29
21	15,281	23,672	29
22	4,365	23,803	29
23	1,403	20,621	29

Table II: Statistics Male Players

FIFA Version	#Teams	#Players	#Skills
16	16,003	336	25
17	31,676	377	27
18	28,756	460	27
19	24,309	597	29
20	22,533	470	28
21	23,552	415	28
22	27,107	577	28
23	2,584	899	29

Table III: Statistics Female Players

III. Results

A.. Ground Truth

The ground truth was to take players from the four semi-finalist teams of the Champions League 2023,

because if these teams have made good progress by achieving good positions, it means that their team is performing well and being effective.

B.. Baselines

Top-Position Model

This approach defines a base model that selects the best player per position through a skills analysis. For each position on the field, the five technical characteristics with the highest average among all players occupying that position are calculated, considering a set of attributes such as passing, finishing, speed, among others. These five skills represent the most relevant competencies for that specific role.

Subsequently, for each player occupying that position, the average of their values in the five selected skills is calculated. The player with the highest average is considered the best in that position and is therefore included in the ideal team (or dream team).

iKNN Baseline

This baseline leverages an instance-based learning approach using the k-Nearest Neighbors (kNN) algorithm. The model begins by identifying the most valuable player from a given club as the seed player. A set of relevant features—including physical, technical, and market attributes such as overall rating, potential, pace, passing, dribbling, and market value—is normalized and used to compute similarity between players.

The kNN algorithm retrieves the top k players in the dataset who are most similar to the seed player in terms of these attributes. From this pool of similar players, a complete team is assembled by selecting individuals who match a predefined tactical formation (e.g., 1-4-4-2), ensuring role-specific constraints are satisfied. This method reflects a content-based similarity approach, simulating the selection of a team based on resemblance to a club's top performer.

Random Baseline

The random baseline provides a control condition by assembling a team through random sampling. A fixed number of players are randomly selected from the global player pool, without considering club affiliation, skill, or position. From this set, the team is built by selecting players that fulfill the required tactical formation (e.g., 1-4-4-2), ensuring the necessary number of players per position (goalkeepers, defenders, midfielders, and strikers). This baseline serves to establish a performance floor, against which more sophisticated methods can be compared.

C.. Evaluation of Utility and Fairness Metrics

To evaluate the effectiveness of the proposed fairness-aware reranking method (`det_cons`), we assessed both utility-based and fairness-based metrics before and after reranking. Table IV reports the results for standard utility metrics used in ranking evaluation, while Table V summarizes the fairness metrics.

Utility Metrics. As shown in Table IV, the reranking process preserved or slightly improved utility. The overall ranking performance, measured by AUCROC, remained stable (0.4992 before vs. 0.5000 after). Mean Average Precision (MAP@k) showed slight increases across all cutoffs, indicating that relevant players were retained or slightly promoted in the top-k positions. Notably, Normalized Discounted Cumulative Gain (NDCG@k) values improved consistently, with `ndcg.cut_50` increasing from 0.0008 to 0.0021. This suggests that the most relevant players were ranked earlier after reranking, enhancing the quality of the generated lineups.

Fairness Metrics. Table V highlights a significant improvement in fairness. The normalized discounted KL divergence (NDKL), a sensitive measure of distributional disparity, decreased from 1.2255 to 0.0675, indicating a more equitable allocation of exposure. Similarly, the skew in exposure was reduced from -8.83 to -4.97. These results confirm that the reranking strategy substantially mitigated exposure bias without sacrificing overall team quality.

Conclusion. The fairness-aware reranking method succeeded in improving exposure fairness while preserving, and in some cases enhancing, utility metrics such as NDCG and MAP. This demonstrates the effectiveness of the method in balancing player performance and equity considerations in team selection scenarios.

Metric	Before	After
aucroc	0.4992	0.5000
map_cut_10	0.0000	0.0002
map_cut_100	0.0001	0.0004
map_cut_2	0.0000	0.0002
map_cut_20	0.0000	0.0003
map_cut_5	0.0000	0.0002
map_cut_50	0.0001	0.0004
ndcg_cut_10	0.0000	0.0016
ndcg_cut_100	0.0023	0.0038
ndcg_cut_2	0.0000	0.0030
ndcg_cut_20	0.0001	0.0015
ndcg_cut_5	0.0000	0.0024
ndcg_cut_50	0.0008	0.0021

TABLE IV: AVERAGE UTILITY METRICS BEFORE AND AFTER RERANKING.

Metric	Before	After
exp	0.4939	0.4839
expu	5.1051	NaN
ndkl	1.2255	0.0675
skew	-8.8324	-4.9713

TABLE V: AVERAGE FAIRNESS METRICS BEFORE AND AFTER RERANKING.

D.. Comparison of Metrics with Baselines

Model	Avg. nDCG
Top-Position	1.0000
iKNN	0.3832
Random	0.0000
Adila (Before)	0.0008
Adila (After)	0.0021

Table VI: Average nDCG across models.

The Top-Position model reaches the maximum possible average nDCG score (1.0000), reflecting a perfect match with the ground truth player rankings. The iKNN model offers a meaningful compromise, with a respectable average nDCG of 0.3832, indicating decent alignment with ideal rankings.

Conversely, both the Random baseline and Adila methods (before: 0.0008; after: 0.0021) perform extremely poorly in terms of ranking quality. Despite a slight improvement in Adila after reranking, the utility remains negligible when compared to the Top-Position or iKNN baselines.

Model	Avg. Precision
Top-Position	0.0833
iKNN	0.0943
Random	0.0000
Adila (Before)	0.0000
Adila (After)	0.0002

Table VII: Average Precision across models.

The precision results show that the iKNN model achieves the highest average precision (0.0943), narrowly outperforming the Top-Position model (0.0833). This indicates that iKNN is slightly more effective at identifying relevant players in the top selections across teams. Both models significantly outperform the Random baseline and the Adila methods, which report near-zero precision—even after reranking (0.0002).

This contrast suggests that while fairness-aware methods like Adila might improve equity, they do so at the expense of utility, as they fail to prioritize relevant players effectively.

IV. Discussion

Throughout the development of this project, we encountered several critical challenges related to the reproducibility and completeness of the Adila framework and its integration with the FIFA23 dataset. Moreover, this project required significant effort not only in implementing reranking experiments, but also in reconstructing a working pipeline from incomplete artifacts. The work demanded persistence, experimentation, and creative problem solving. The final results—functional reranking experiments using Demographic Parity over FIFA23 data—stand as a testament to the feasibility of reusing fairness-aware ranking frameworks with new datasets, despite initial setbacks.

Below, we describe the main obstacles faced and the steps taken to overcome them.

A.. Incomplete Reproducibility of Adila

While the Adila repository provides a foundation for fairness-aware reranking research, it lacks critical components necessary for a complete and seamless pipeline. Notably, baseline model implementations such as the Bayesian Neural Network (BNN) referenced in the Adila original publication are not publicly available. Additionally, essential Python modules, such as `cmn`, are missing from the repository, which initially prevented deserialization of the original `teams.pkl` file.

B.. Reverse Engineering of the Dataset Format

Due to the absence of documentation and key serialization modules, we initially attempted to reverse-engineer the structure of the data files, including `teams.pkl`, `teamsvecs.pkl`, and `indexes.pkl`. This led to several incorrect assumptions about the internal structure of the objects, resulting in failed compatibility with the reranking code.

A turning point came after a detailed analysis of the related `OpeNTF` repository, which we identified as the likely source of Adila's internal data format design. By importing the `cmn` module from OpeNTF, we were able to successfully deserialize the original Adila dataset. This allowed us to verify and reconstruct the correct object hierarchy and serialization logic needed for compatibility with reranking scripts.

C.. Data Construction from FIFA23

Once the internal structure was understood, we designed a new pipeline to construct Adila-compatible datasets using cleaned and deduplicated data from FIFA23. This involved:

- Identifying and filtering duplicate rows across player entries.
- Defining consistent team identifiers using `player_id` and `club_team_id`.
- Generating vector representations for team composition and skill matrices aligned with Adila's expectations.
- Creating prediction files with the correct `.pred` format required by the reranking module.

D.. Debugging Through Trial and Error

Many parts of the Adila pipeline are undocumented or ambiguous. For example, the role of files such as `labels.csv` and `stats.pkl` had to be inferred manually through extensive inspection of intermediate outputs and by tracing function behavior. Numerous errors (e.g., serialization mismatches, shape inconsistencies, missing fields) were identified and resolved via incremental testing, logging, and manual validation of outputs.

V. State of the Art

The current literature has mainly focused on the team formation process from an automated perspective, seeking to improve team composition and performance using various computational methodologies and techniques [5]. Effective team formation requires selecting players based on specific attributes while fostering cohesive interactions. Studies identify five key attributes: Knowledge/Technical Expertise (57% of studies, quantifiable skills such as technique or speed), Personality Traits (20%, evaluated using tools like Belbin or Myers-Briggs for group cohesion), Communication/Collaborative Skills (16%, leadership and motivation), Learning (7%, capacity to share knowledge), and Internal Roles (recent studies on team dynamics). Around 60% of the studies combine 2–3 attributes simultaneously, highlighting the need for objective data collection [5]. In contexts such as the Fantasy Premier League (FPL), additional attributes such as player form, Fixture Difficulty Rating, Bonus Points System (BPS), and the ICT Index (influence, creativity, threat) are considered to predict performance [4]. Coaches' evaluations, based on 18 qualitative and quantitative criteria over long observation periods, are crucial for player selection and team formation, complementing training data.

To improve team formation, researchers have explored computational models and methods. Among these models, Search and Optimization (53%) using genetic, heuristic, and greedy algorithms stand out; Statistics and Mathematics (27%), using regression models and distance measures; and Data Mining (7%) with clustering, association, and classification techniques [5]. Other techniques include fuzzy logic, Bayesian networks, and grey decision theory [5].

Machine Learning (ML) has also been used in sports for player selection and team formation. One study applied seven ML algorithms to a youth football team, where Random Forest (RF) performed best, achieving a 93.93% reliability in player selection and 89.36% in team formation when compared to actual match line-ups [1]. Other tested algorithms include Multilayer Perceptron (MLP), Support Vector Machines (SVM), Decision Trees (DT), Logistic Regression, Naive Bayes, and CART [1]. Deep Neural Networks (DNNs) have been employed to evaluate the suitability of each player for different positions [3]. The Hungarian Algorithm is used alongside DNNs to find the maximum weighted match between players and designated positions, optimizing team line-ups [3]. Additionally, the Distinct Representatives System (SDR) concept allows coaches to select a group of candidates for each position, offering a less rigid and more realistic solution that leverages the full potential of the squad [3]. At the tactical analysis level, systems have been developed to classify and visualize team formations (e.g., 4-4-2) in match situations based on player positional data [2]. This enables the generation of a visual formation summary (VFS) and its comparison with a set of predefined formation templates, calculating a formation similarity (FSIM) using the Hungarian Algorithm for tactical classification [2].

Data is crucial for ML techniques. It includes "Hit/it" training data obtained from devices measuring control, technique, reaction time, speed, and agility, which are considered as valuable as match data for player selection [1]. Additionally, Sofifa/FIFA data provides player scores across various skills, contract information, age, height, weight, and preferred foot, offering a source with a large number of players and features [3].

Despite advancements, current research faces significant challenges. Most solutions are tested with hypothetical or simplified cases, limiting their effectiveness in complex real-world environments [5]. The specification of attributes is often vague, without identifying specific details, which can increase subjectivity and reduce reliability in attribute evaluation, as well as limited or no access to source code or pseudocode, hindering the reproducibility and reuse of findings [5].

VI. Conclusion

In this project we have worked along two main lines. The first has consisted of adapting Adila, a team creation system originally designed for expert contexts, to the sports environment, specifically for the formation of football teams. The second line, and main contribution of the work, has been the elaboration of a dataset of football players, which lays the foundations for future applications and research in this field.
This new dataset is particularly interesting for the possibility of assessing the applicability of different models in a context where tensions may arise between technical performance and potential biases, such as gender. Once the dataset was constructed, we focused on analysing the behaviour of different models with respect to fairness, taking player nationality as a study attribute. This allowed us to assess whether the models select players equally across countries or whether they tend to concentrate on certain nationalities.
This assessment is of great relevance, as different levels of sporting support and development between countries can significantly influence player representation, making fairness a key factor for inclusion in team formation. So, the next steps may be to:

- Explore fairness in other attributes (age, position).

Authors' Contributions

All authors participated in the writing of the manuscript. All authors read and approved the final version of the manuscript.

Data Availability

Here is the GitHub repository with the dataset and the baselines used in the paper: https://github.com/JaviPeace/Proyecto-Recomendadores.git

References

[1] ABİDİN, D. (2021). A case study on player selection and team formation in football with machine learning. *Turkish Journal of Electrical Engineering and Computer Sciences, 29*(3), 1672–1691. https://doi.org/10.3906/elk-2005-27

[2] Kotian, I. (2021). FIFA19 Player and Team Analysis and Value Predict . https://github.com/Ishan-Kotian/FIFA19-Player-and-Team-Analysis-and-Value-Predict

[3] Müller-Budack, E., Theiner, J., Rein, R., & Ewerth, R. (2019). "Does 4-4-2 exist?" – An Analytics Approach to Understand and Classify Football Team Formations in Single Match Situations. *arXiv*. http://arxiv.org/abs/1910.00412

[4] Nouraie, M., Eslahchi, C., & Baca, A. (2023). Intelligent team formation and player selection: a data-driven approach for football coaches. *Applied Intelligence, 53*(24), 30250–30265. https://doi.org/10.1007/s10489-023-05150-x

[5] Rajesh, V., Arjun, P., Jagtap, K. R., Suneera, C. M., & Prakash, J. (2022). Player Recommendation System for Fantasy Premier League using Machine Learning. In *2022 19th International Joint Conference on Computer Science and Software Engineering, JCSSE 2022* (pp. 1–6). IEEE. https://doi.org/10.1109/JCSSE54890.2022.9836260

[6] Stavrou, G., Adamidis, P., Papathanasiou, J., & Tarabanis, K. (2023). Team Formation: A Systematic Literature Review. *International Journal of Business Science and Applied Management, 18*(2), 17–34. https://doi.org/10.69864/ijbsam.18-2.174

[7] Loghmani H., Fani H., Rueda G., Paul E., Lakshmi Y. & Moasses R. (2022). *Adila: Fairness-Aware Team Formation.* WorkShop Collaborative Team Recommendations for Skilled Users. https://github.com/fani-lab/Adila.git