



PROPUESTA PROYECTO
SISTEMAS RECOMENDADORES

Profesor: Denis Parra
Viernes, 09 de mayo, 2025

Javiera Belén López Massaro
Javiera Paz Azócar Oliva
Pablo Poblete Arraé
Gabriel Acevedo

1. Motivación

La formación de equipos colaborativos efectivos representa un desafío constante para diferentes sectores tanto empresariales como educativos. Tradicionalmente, la selección del personal capacitado se ha realizado de forma manual, considerando criterios subjetivos y poco confiables que pueden incluir cierto sesgo inconsciente por parte del tomador de decisiones, lo que conduce a desequilibrios en habilidades, baja cohesión y bajo rendimiento grupal.

El uso de herramientas que permitan automatizar este proceso ha llamado la atención de investigadores en diversos ámbitos, desde la psicología organizacional hasta el aprendizaje automático.

Este proyecto aborda el problema emergente de la recomendación de personas para equipos, donde, dado un equipo y una necesidad específica, se busca sugerir individuos adecuados para completarlo.

2. Problema

La formación de equipos de fútbol óptimos requiere equilibrar:

- **Rendimiento deportivo** (estadísticas como: goles, velocidad, pases).
- **Diversidad** (nacionalidad, edad, estilos de juego).

Brecha actual:

Consideramos que actualmente existe una brecha que es suficientemente relevante para justificar su estudio, dado que los modelos existentes (*Adila*, *Most Popular*) no priorizan diversidad ni se han probado en contextos deportivos.

3. Objetivos del proyecto

Durante el desarrollo del proyecto se implementarán distintos modelos base (baselines), entre ellos *Most Popular*, *Random* y *Adila*, con el propósito de generar recomendaciones para la conformación de equipos de fútbol. En este sentido, el proyecto tendrá un enfoque orientado a la replicabilidad y reproducibilidad de modelos previamente desarrollados. Inicialmente, se aplicará el modelo *Adila* sobre el dataset original (UMAP24); posteriormente, en una segunda etapa, se buscará replicar dicho modelo en un nuevo conjunto de datos, por ejemplo, uno basado en jugadores de fútbol y sus estadísticas (como velocidad, definición, minutos jugados, edad, nacionalidad, entre otros atributos).

Además, en una etapa avanzada, se propondrán mejoras sobre los métodos existentes, no solo orientadas a optimizar la eficiencia del equipo, sino también a fomentar la diversidad de los equipos generados. Esta diversidad será evaluada mediante métricas específicas,

considerando características como nacionalidad y edad de los jugadores, con el objetivo de abordar potenciales sesgos y promover la inclusión de perfiles variados.

- 3.1. Analizar y comparar métodos existentes para la recomendación de personas a equipos.
- 3.2. Implementar tres modelos de referencia: Random, Most Popular y al menos un modelo específico (como OpenNTF o Adila).
- 3.3. Evaluar el desempeño de los modelos utilizando métricas de ranking, diversidad y novedad.
- 3.4. Proponer mejoras o adaptaciones metodológicas sobre los modelos existentes, como ajustes en los algoritmos o incorporación de criterios de equidad.
- 3.5. Aplicar los modelos sobre el dataset de UMAP24 y explorar su aplicación a un nuevo dataset relacionado con equipos de fútbol, como por ejemplo: FC 25 Player Ratings Reveal o FIFA23 Official Dataset

A partir del análisis preliminar, visualizamos un posible objetivo de mediano a largo plazo, que por su relevancia creemos se podría analizar a nivel conceptual y de diseño y se refiere extender Adila a una versión Adila+ para optimizar diversidad. Esto podría ser el fundamento para una investigación más profunda, cuyas posibles hipótesis de investigación serían:

1. **H1:** La diversidad (nacionalidad/edad) mejora el rendimiento en equipos de fútbol.
2. **H2:** Adila+ supera a Adila en métricas de novedad sin perder precisión.

4. Análisis descriptivo de los datos

El dataset original (UMAP24) incluye representaciones matriciales dispersas de múltiples equipos. En particular, el **fteamsvecs** corresponde a un archivo en formato *pickle* que contiene un diccionario con tres matrices dispersas (*lil_matrix*):

- **id:** de tamaño $[n_equipos \times 1]$, identifica cada equipo.
- **skill:** de tamaño $[n_equipos \times n_habilidades]$, indica las habilidades presentes en cada equipo.
- **member:** de tamaño $[n_equipos \times n_expertos]$, indica qué expertos pertenecen a cada equipo.

En términos simples, cada fila de las matrices skill y member representa un equipo y muestra, mediante vectores de ocurrencia, qué habilidades y qué expertos están presentes.

Además del dataset original, se propone explorar un nuevo conjunto de datos relacionado con **jugadores de fútbol**, en el que se consideran tanto sus habilidades como sus estadísticas individuales. Las principales variables disponibles son:

- **RANK**: la posición en el ranking general (por nivel de rendimiento o popularidad).
- **PLAYER**: el nombre del jugador.
- **NAT** (Nationality): la nacionalidad del jugador.
- **TEAM**: el club o equipo en el que juega.
- **POS** (Position): la posición en el campo (por ejemplo, ST = delantero, CM = mediocampista central, GK = arquero).
- **OVR** (Overall): la valoración general del jugador, basada en su desempeño total.
- **PAC** (Pace): velocidad del jugador (aceleración + sprint).
- **SHO** (Shooting): capacidad para disparar y marcar goles.
- **PAS** (Passing): precisión y habilidad para realizar pases.
- **DRI** (Dribbling): habilidad para regatear y controlar el balón.
- **DEF** (Defense): capacidad defensiva (entradas, posicionamiento).
- **PHY** (Physical): atributos físicos (fuerza, resistencia, agresividad).
- **AGE** : edad del jugador

Este nuevo conjunto permitirá realizar análisis más detallados de diversidad y equidad, utilizando variables como **nacionalidad** y **edad** para detectar posibles sesgos. Las *labels* estarán definidas por el equipo de pertenencia (TEAM) y la posición en el campo (POS). Los datos serán extraídos del repositorio “FC 25 Player Ratings Reveal”.

5. Planificación Midterm:

Proyecto de Recomendación

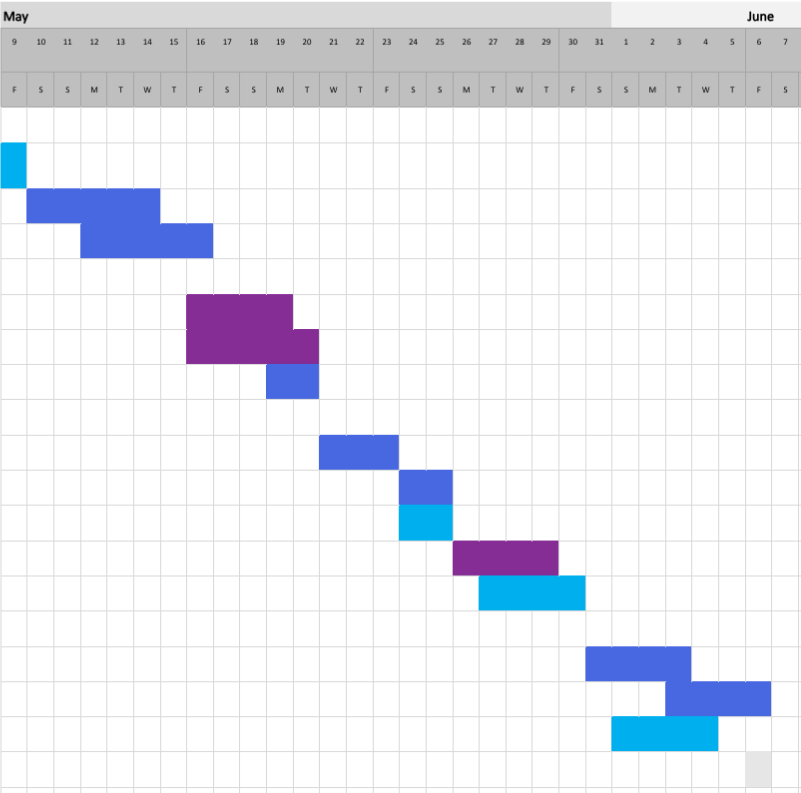
Grupo Recomendadores

Legenda: Según lo previsto, Dificultad baja, Dificultad media, Dificultad alta, Sin asignar

Fecha de inicio del proyecto:

Incremento de desplazamiento:

Descripción del hito	Inicio	Días
Generar entorno de ejecución adaptado		
Consultar a ayudante forma de ejecutar un entorno que está obsoleto y con inconsistencias	5/9/2025	1
Revisar dependencias y estudiar funciones del repositorio	5/10/2025	5
Factorizar funciones y adaptar código	5/12/2025	5
Probar método 1		
Generar el código necesario	5/16/2025	4
Entrenar modelo	5/16/2025	5
Generar métricas ranking, diversidad, novedad	5/19/2025	2
Correr modelos con nuevo dataset		
Generar el dataset de equipos de fútbol	5/21/2025	3
Probar modelo random y Most Popular	5/24/2025	2
Generar las métricas de evaluación	5/24/2025	2
Probar modelo #####	5/26/2025	4
Generar las métricas de evaluación	5/27/2025	4
Creación de la documentación		
Describir problemas identificado	5/31/2025	4
Describir analisis preliminar de los resultado	6/3/2025	4
Revisión del plan propuesto en etapa anterior y justificación de ajustes.	6/1/2025	4
Entregar tarea	6/6/2025	1



6. Viabilidad Técnica del Proyecto

La viabilidad técnica de este proyecto se sustenta en dos pilares fundamentales. Por una parte, la existencia del dataset “FC 25 Player Ratings Reveal” y los los datasets disponibles en Kaggle, proporciona una base sólida, estructurada y actualizada para el análisis. Por otra parte, el desarrollo e implementación de modelos de referencia simulados (sección 6.2) demuestra la factibilidad de llevar a cabo el enfoque propuesto.

A continuación, se presenta la justificación del uso del dataset FC 25 Player Ratings Reveal y la descripción de los modelos de referencia implementados como prueba de concepto.

6.1. Disponibilidad del Dataset “FC 25 Player Ratings Reveal”

'FC 25 Player Ratings Reveal' es un evento anual organizado por EA Sports en el que se presentan oficialmente las valoraciones de los jugadores en el videojuego EA Sports FC 25, sucesor de la serie FIFA. Estas valoraciones reflejan el rendimiento y las habilidades individuales de los futbolistas, y son ampliamente utilizadas en modos de juego como Ultimate Team y Carrera.

Las estadísticas disponibles incluyen:

- Valoración general (OVR): medida global del nivel del jugador.
- Atributos detallados: ritmo (PAC), disparo (SHO), pase (PAS), regate (DRI), defensa (DEF), físico (PHY).
- Datos adicionales: nombre, nacionalidad, edad, equipo y posición.

Este dataset resulta altamente relevante para el presente proyecto por las siguientes razones:

- Ofrece un conjunto de datos estructurado y actualizado basado en evaluaciones profesionales.
- Permite analizar rendimiento individual y colectivo en la formación de equipos.
- Facilita estudios de diversidad en nacionalidad y edad.
- Sirve como base realista para la implementación de modelos de recomendación enfocados en balancear eficiencia y diversidad.
- Posibilita la identificación de posibles sesgos, contribuyendo al análisis justo e inclusivo en el ámbito deportivo.

Su incorporación respalda la viabilidad técnica del proyecto y proporciona una fuente sólida para validar los modelos propuestos.

6.2. Diseño de Modelos de Referencia

Para validar la viabilidad del enfoque, se implementaron tres modelos de referencia utilizando un conjunto de datos simulado con 30 jugadores de fútbol, cada uno con atributos como equipo, nacionalidad, edad y posición. Los modelos implementados son los siguientes:

- **Modelo Random:** selecciona jugadores al azar desde el conjunto total.

- **Modelo Most Popular:** selecciona jugadores del equipo con mayor cantidad de representantes.
- **Modelo Diversity-Oriented:** prioriza perfiles con diversidad en nacionalidad y edad.

Cada uno de estos modelos puede ser adaptado y extendido al conjunto de datos real del proyecto, y sirve como línea base para comparaciones futuras. El código fuente de estos modelos se entrega como anexo en el archivo '**modelos_baseline_simulados.py**'.

Modelo 1: Random

Este modelo selecciona jugadores de forma aleatoria desde el conjunto de candidatos disponibles. No considera criterios de desempeño ni diversidad. Sirve como línea base simple para comparación.

Entrada: conjunto de jugadores.

Proceso: selección aleatoria sin reemplazo.

Salida: lista de k jugadores seleccionados.

Modelo 2: Most Popular

Este modelo selecciona jugadores del equipo con mayor cantidad de representantes en el conjunto. Se asume que equipos con más jugadores son indicadores de popularidad o desempeño grupal.

Entrada: conjunto de jugadores con atributo TEAM.

Proceso: contar jugadores por equipo, elegir el equipo más frecuente, seleccionar los primeros k jugadores de ese equipo.

Salida: lista de k jugadores del equipo más frecuente.

Modelo 3: Diversity-Oriented

Este modelo prioriza jugadores cuya inclusión aporta mayor diversidad, considerando nacionalidad y edad.

Entrada: conjunto de jugadores con atributos NAT y AGE.

Proceso: calcular una puntuación combinada basada en diversidad de nacionalidades y edades; ordenar y seleccionar los k más altos.

Salida: lista de k jugadores diversos.

Referencias

- Chiarello, F., Spada, I., Barandoni, S., Giordano, V., & Fantoni, G. (2023). *A data driven tool to support design team composition measuring skills diversity*. Proceedings of the Design Society, 3, 91–100. <https://doi.org/10.1017/pds.2023.10>
- Fani, H. (2024). *Tutorial on Collaborative Team Recommendation for Skilled Users: Objectives, Techniques, and New Perspectives*. Fani's Lab!, School of Computer Science, University of Windsor. Presentado en ACM UMAP 2024. Recuperado de SlideShare.
- Fani Lab. (2024). *OpeNTF Tutorial – UMAP24 Dataset*. Recuperado de <https://fani-lab.github.io/OpeNTF/tutorial/umap24/>
- Geyik, S. C., Ambler, S., & Kenthapadi, K. (2019). *Fairness-Aware Ranking in Search & Recommendation Systems with Application to LinkedIn Talent Search*. LinkedIn Corporation. Recuperado de <https://arxiv.org/abs/1905.01989>
- Ge, Y., Liu, S., Fu, Z., Tan, J., Li, Z., Xu, S., Li, Y., Xian, Y., & Zhang, Y. (2024). *A survey on trustworthy recommender systems*. ACM Transactions on Recommender Systems, 3(2), Article 13, 1–68. <https://doi.org/10.1145/3652891>
- Herlocker, J. L., Konstan, J. A., Terveen, L. G., & Riedl, J. T. (2004). *Evaluating collaborative filtering recommender systems*. ACM Transactions on Information Systems, 22(1), 5–53. <https://doi.org/10.1145/963770.963772>
- Morrison-Smith, S., & Ruiz, J. (2020). *Challenges and barriers in virtual teams: A literature review*. SN Applied Sciences, 2(1096). <https://doi.org/10.1007/s42452-020-2801-5>
- 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>