Data Science Models for Football Scouting: The *Racing de Santander* Case Study

Diego Brunetti - Sebastián Ceria - Guillermo Durán
Manuel Durán - Andrés Farall - Nicolás Marucho - Pablo Mislej

33rd European Conference on Operational Research
30th June – 3rd July 2024
Copenhagen, Denmark

EURO 24 COPENHAGEN

Index

- The importance of data science in sports scouting
- Cooperation agreement between UBA and *Racing de Santander*
- Current process vs data-driven scouting
- Supervised learning model: understanding experts' thinking
- Example: finding the best substitute player
- Future enhancements and innovations

The importance of data science in sports scouting

Data assets are becoming critical for making informed decisions

Using big data for scouting in different sports is essential for detecting value where other teams do not

Leveraging the knowledge of expert scouts and expanding it across the market

"If we win, with our budget, with this team... we'll have changed the game. And that's what I want."
~ *Billy Beane. Moneyball*

Inspirational cases of its use:

- *Brentford – Premier League*
- *Oakland Athletics – MLB*

Cooperation agreement between the *University of Buenos Aires* and *Real Racing Club de Santander*.

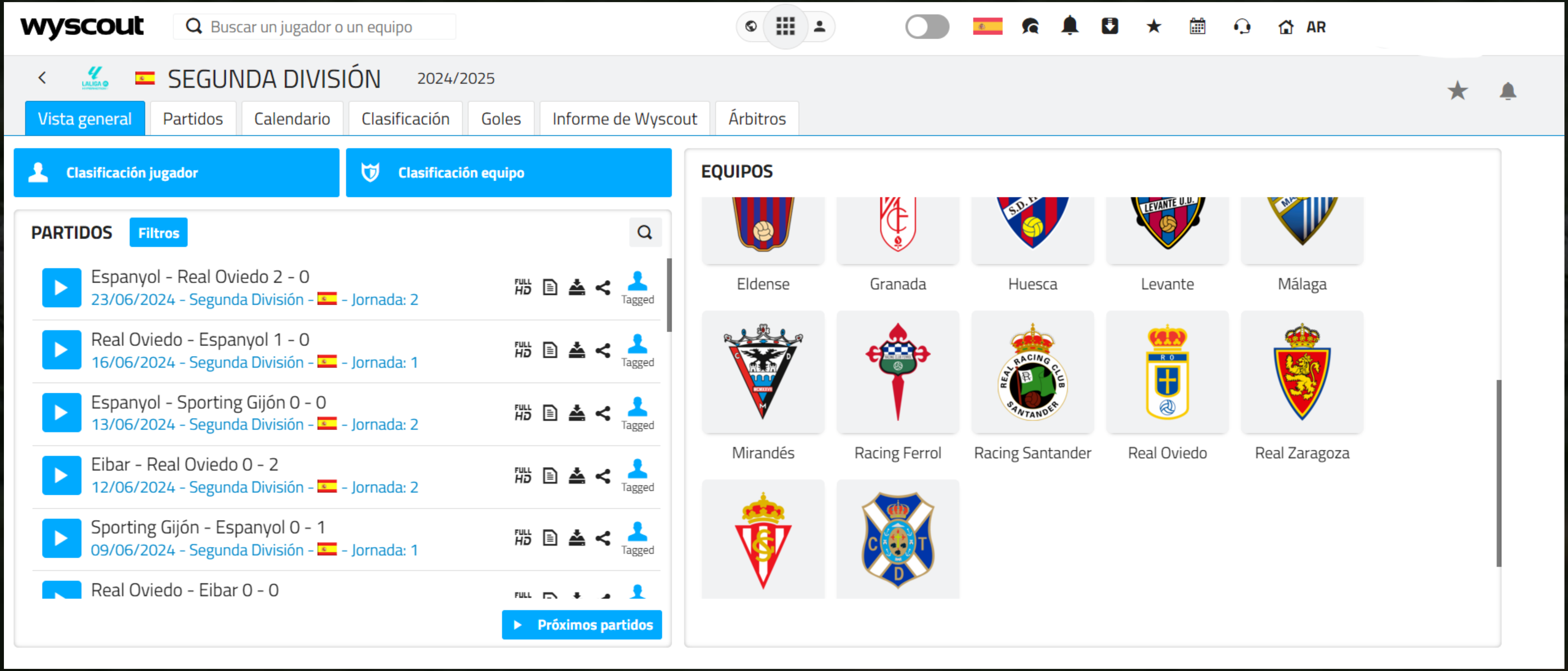
Data transformation at the football club

Sebastián Ceria, the owner of Racing Club de Santander who holds a PhD in mathematics, has spearheaded innovative projects such as the agreement between the Club and *the* ***"Instituto de Cálculo"*** at the **University of Buenos Aires** *for the promotion of research in applications of data science to football, through the development of statistical tools and machine-learning algorithms that support data-driven decisions.*

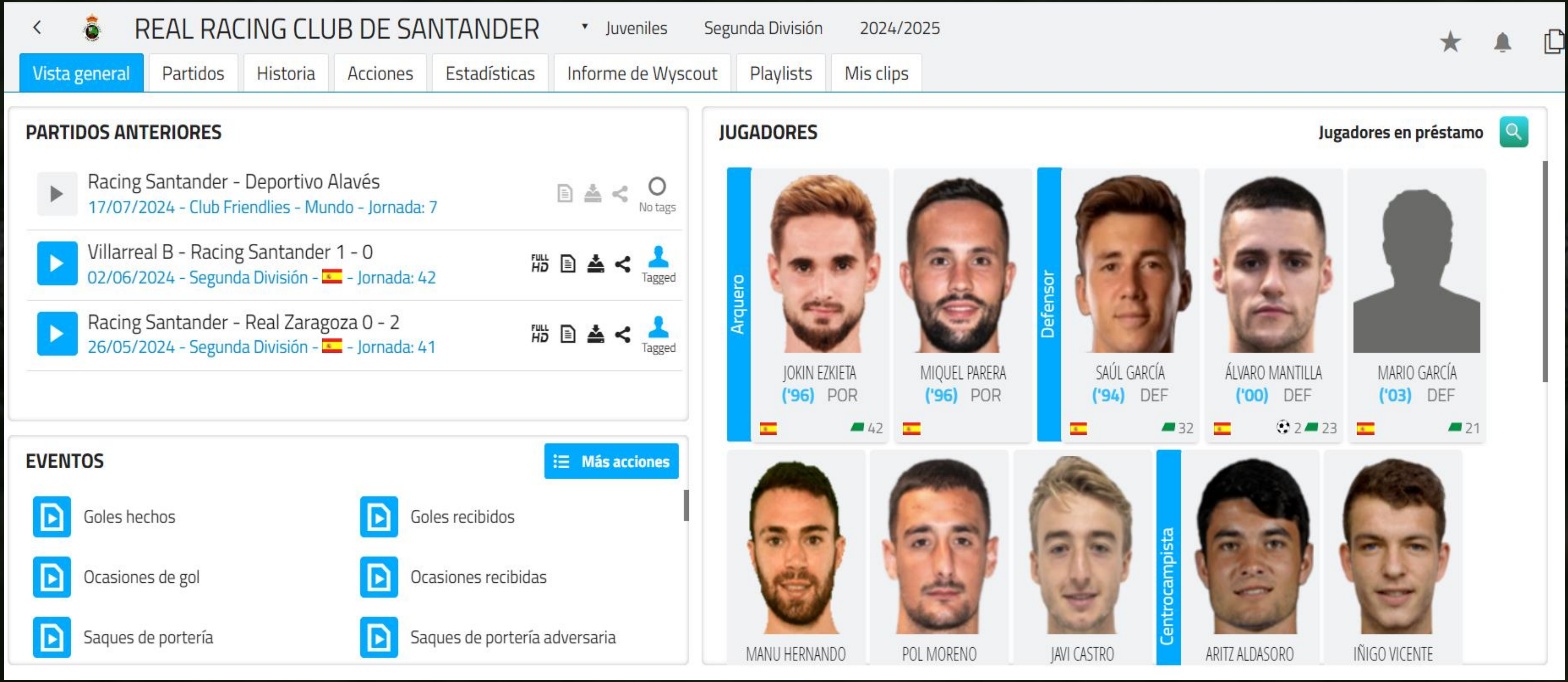
DATA PIPELINE

1. Construct a data lake connecting multiple data sources with the assistance of GlobalLogic Tech Company
2. Carry out exploratory data analysis and pre-processing of datasets
3. Train machine learning models and validate results.
4. Generate results and validate them with experts to support data-driven decisions

Wyscout: League view

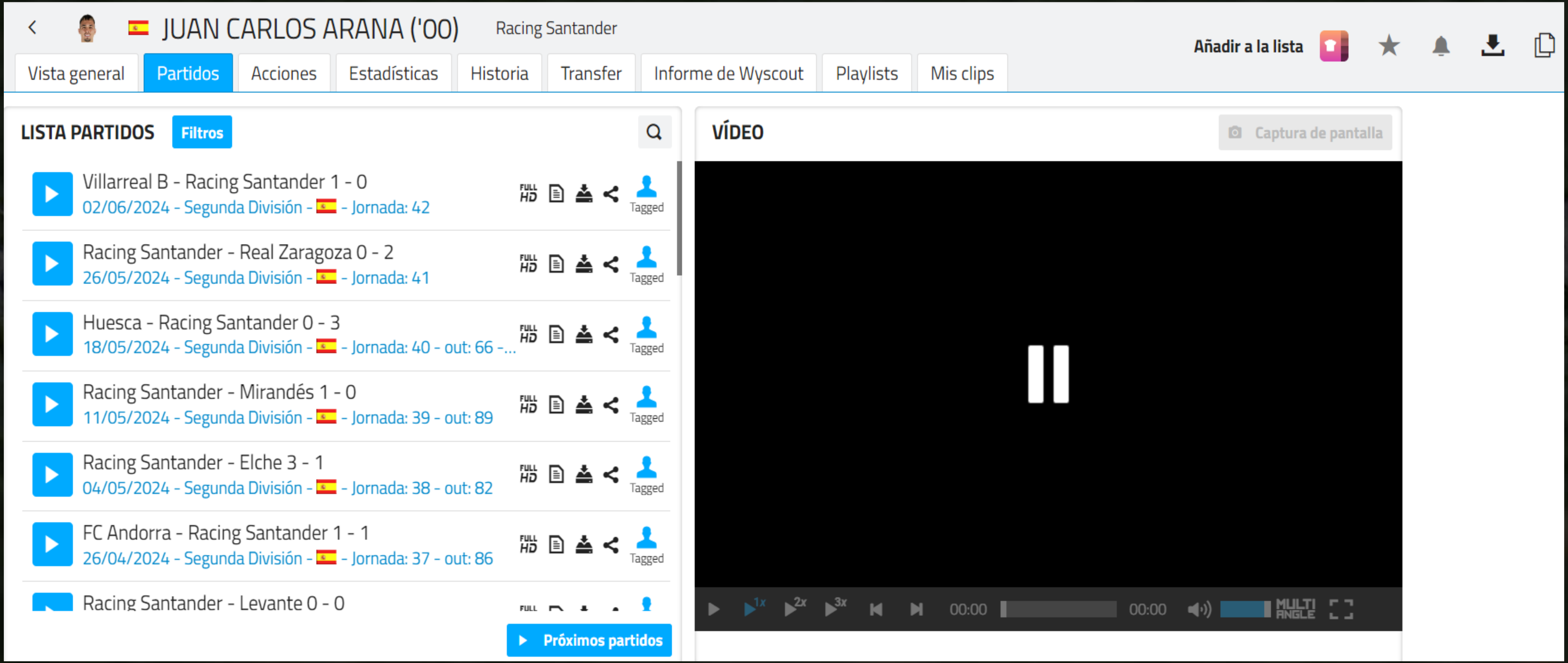


Wyscout: Team view

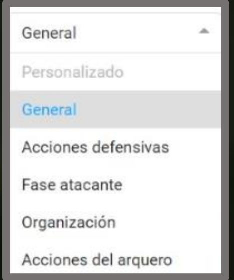


Wyscout: Player view

Wyscout: Player matches' view

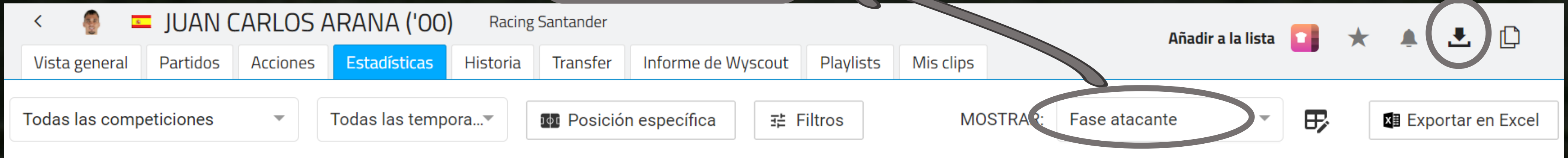


Wyscout: Player statictics view



Variables type

Export button



Partido	Posición específica	Minutos jugados	Goles	Asistencias	Tiros / logrados	xG	Asistencias a tiro	Centros / precisos	Regates / logrados	Duelos ofensivos / ganados	Toques en el área de penalti	Fuera de juego	Carreras en profundidad	Faltas recibidas
PROMEDIO / 90 TOTAL		6735	34	8	211 47.9%	28.51	42	36 30.6%	274 48.2%	665 38.6%	263	43	89	123
Villarreal B 1:0 Racing Santander Spain. Segunda División, 02.06.2024	CF	96	0	0	1/0 0%	0.04	2	1/0 0%	3/1 33%	13/3 23%	4	0	1	3
Racing Santander 0:2 Real Zaragoza Spain. Segunda División, 26.05.2024	CF	102	0	0	4/1 25%	0.48	1	0	1/0 0%	6/0 0%	6	1	0	1
Huesca 0:3 Racing Santander Spain. Segunda División, 18.05.2024	CF	72	1	0	2/2 100%	0.31	0	0	0	4/1 25%	2	0	1	1
Racing Santander 1:0 Mirandés Spain. Segunda División, 11.05.2024	CF	91	0	0	7/4 57%	0.75	0	0	8/4 50%	11/4 36%	8	0	1	2
Racing Santander 3:1 Elche Spain. Segunda División, 04.05.2024	CF	85	0	0	3/1 33%	0.41	1	0	5/3 60%	6/3 50%	3	1	1	1
FC Andorra 1:1 Racing Santander Spain. Segunda División, 26.04.2024	CF	87	0	0	6/2 33%	0.69	0	2/0 0%	4/4 100%	10/7 70%	4	1	2	0
Racing Santander 0:0 Levante	[illegible]	[illegible]	[illegible]	[illegible]	4/3	[illegible]	[illegible]	[illegible]	1/0	2/0	[illegible]	[illegible]	[illegible]	[illegible]

Challenges along the way

Key technical and cultural obstacles encountered during the project

Connecting databases

Detecting internal and external databases in the club and establishing connections between them for construction of a data lake.

Data cleaning and validation

Checking consistency and quality of data from different sources. Also, cleaning data fields and observations that are of interest to the experts.

Cultivating Data-Driven Culture

Overcoming resistance to change and promoting the benefits of data-driven decision-making. Encouraging experts to trust machine-learning model results.

Current scouting process

Scouting experts monitor players and manually categorize them using 20 different labels based on their unique characteristics (play-maker, goal-scorer, etc.), creating a database for future hirings.

Experts working all over Europe analyze players and manually categorize them under several labels.

A database is created and consolidated for future reference.

Players in the database are searched for possible replacements based on previous labeling.

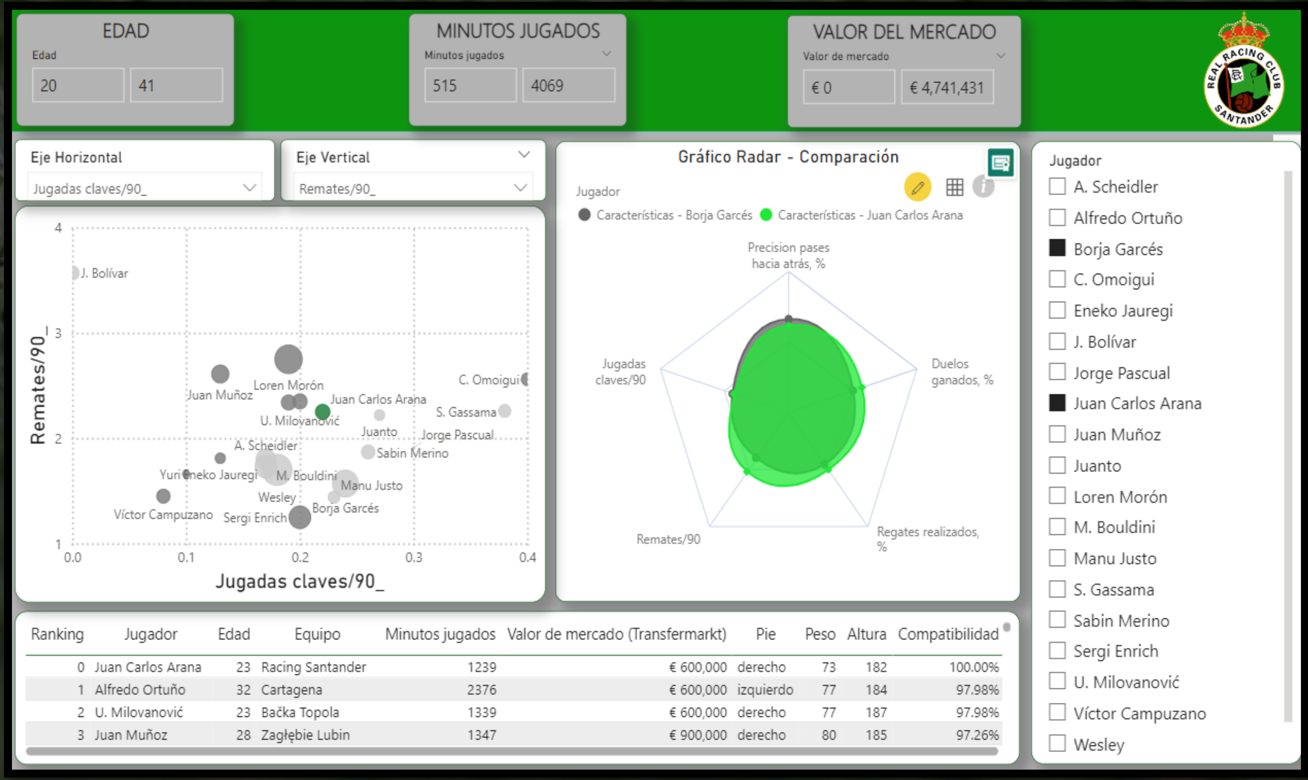
Scouting with data-driven decisions

*By leveraging data processing, **data analysis** and classic **machine learning models** can significantly enhance decision-making in player scouting. Some initial ideas have been developed to help professional scouts analyze players on the market. Small tools can have a huge impact.*

Data visualization and dashboard creation for comparison of players' attributes or in-game stats to better understand performance.

Calculation of Euclidean/Gower distances between players and adjustment of weights to emphasize variables that matter most to experts.

Use of K-Means clustering techniques to group players with similar characteristics as an aid to efficient scouting and comparison



Illustrative example – model results dashboard

Supervised learning model: understanding experts' thinking

*Using advanced **machine-learning multi-class models** to assist experts in **classifying** players from around the globe and improving the search for the best fit. Our models **learn to think like the experts**, identifying players who best match their criteria.*

HOW THE MODEL LEARN

x_i		y
Tackles *Interceptions* *Aerial duels*	→	***Defensive***
⋮		⋮
Assists *xA* *Dribbles*	→	***Play-Maker***
⋮		⋮
xG *Shot accuracy* *Offensive duels*	→	***Goal-Scorer***

Train on limited expert database and predict the most probable label for every player in the entire database

$\overrightarrow{P_{ij}}$ Every player *j* has his own vector of probabilities for each label *i*

$$\sum_{i=1}^{n} p_{ij} = 1 \quad \forall\, player\ j$$

Key takeaways from supervised ML model

*Enables the creation of a **database with complete player profiles** for consultation by experts, and **measures compatibility between players** using the probability vector*

ML multi-class model generates player profiles for entire dataset

PLAYER PROFILE 1

$\overrightarrow{P_{ij}}$

0.18	0.14	...	0.09

PLAYER PROFILE N

$\overrightarrow{P_{ij}}$

0.12	0.04	...	0.22

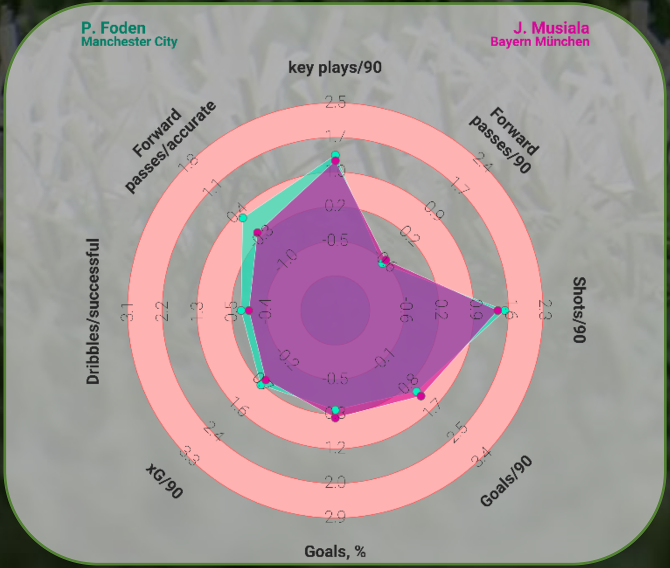
Distance between players using $\overrightarrow{P_{ij}}$ to measure compatibility from an expert perspective.

Expand player categorization database to open search.

Best substitute player using different approaches

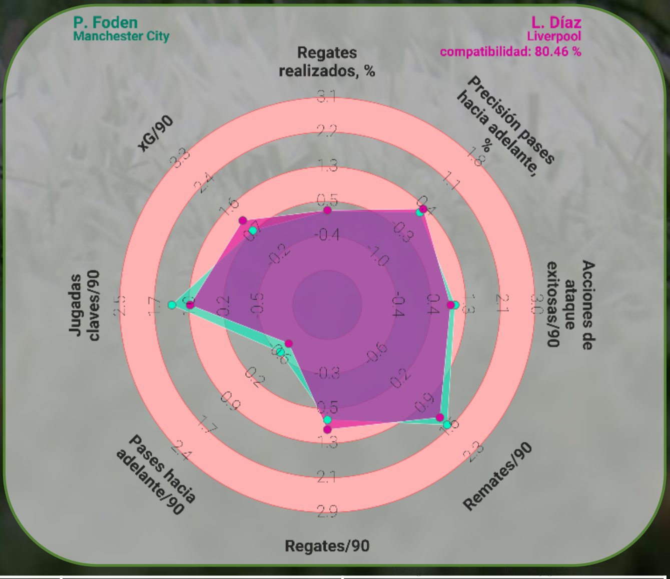
Suppose that Manchester City sells ***Phil Foden***, their young star and the most recent Premier League MVP, and must find the best replacement that matches his attributes as determined by the club's experts.

Supervised model approach



RANK.	PLAYER	TEAM	COMP.
1	Jamal Musiala	Bayern Munich	92%
2	Julian Brandt	Bor. Dortmund	91%
3	Leandro Trossard	Arsenal	86%

Original attributes distance approach



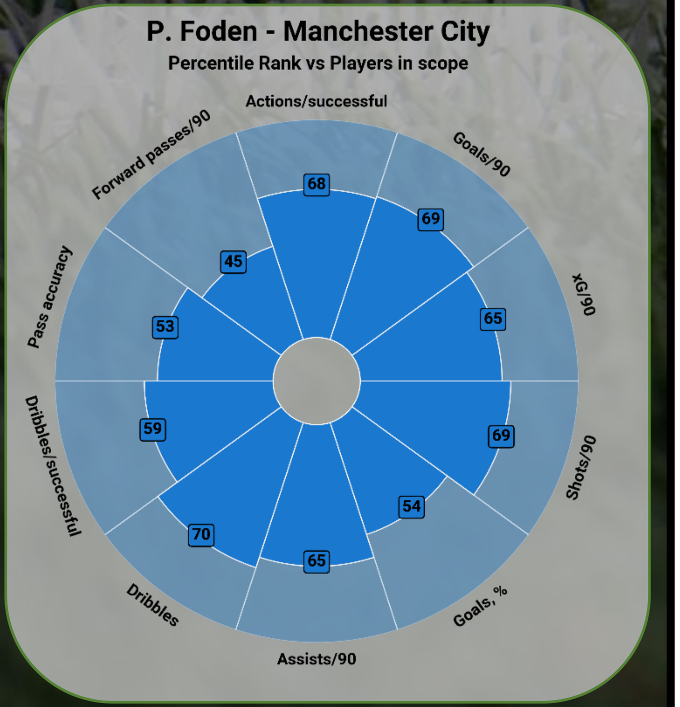
RANK.	PLAYER	TEAM	COMP.
1	Luis Diaz	Liverpool	81%
2	Steven Bergwjin	Ajax	78%
3	Rafa Silva	Benfica	76%

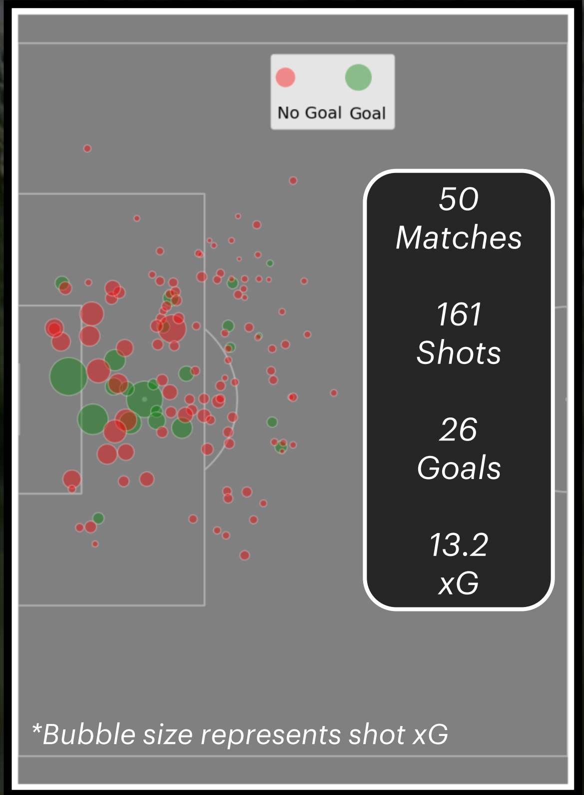
Comparisons with best substitute player

Phil Foden

TALENT	PLAY-MAKER	POWERFUL
34.88 %	26.93 %	21.79 %

Shots season 23/24

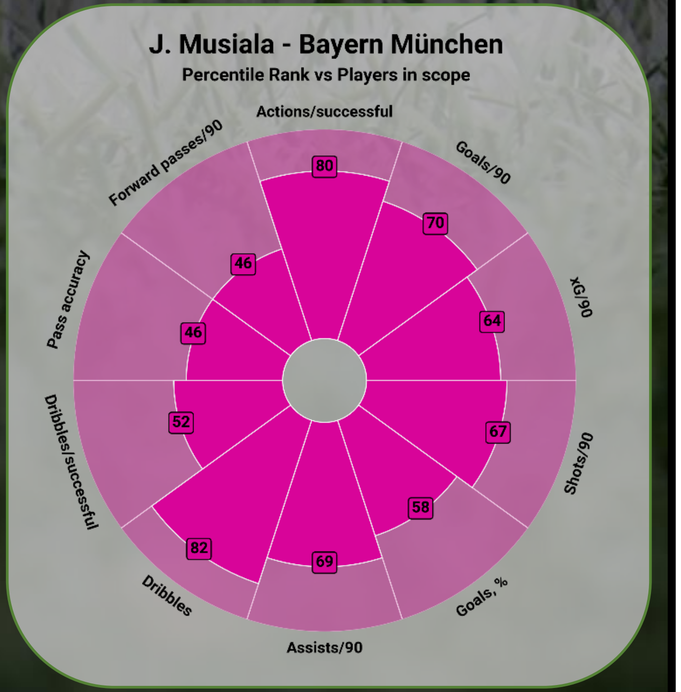


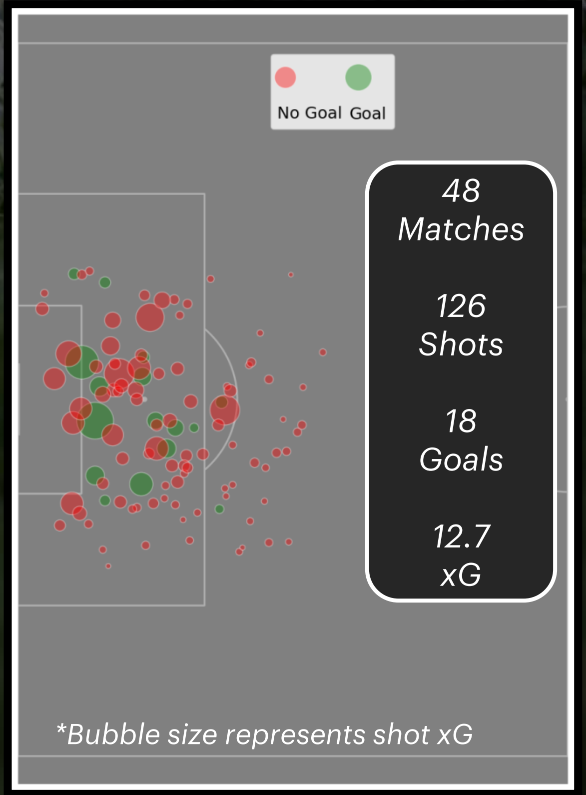


Jamal Musiala

TALENT	PLAY-MAKER	POWERFUL
34.58 %	19.54 %	19.76 %

Shots season 23/24





Future enhancements and innovations

Evaluate the performance and potential of prospective players to be scouted, and take into account the expectations of the experts

Evaluate game tactics and enhance team performance by searching for players who best fit the team's playing style.

Evaluate advanced game metrics (xG, xA, xT, etc.) for *Racing de Santander* and its opponents to optimize playing style on a match-by-match basis.

Machine-learning models for player care, injury prevention and cost savings associated with injury recovery periods.

Data Science Models for Football Scouting: The *Racing de Santander* Case Study

Thank you!

33rd European Conference on
Operational Research
30th June – 3rd July 2024
Copenhagen, Denmark

EURO 24
COPENHAGEN

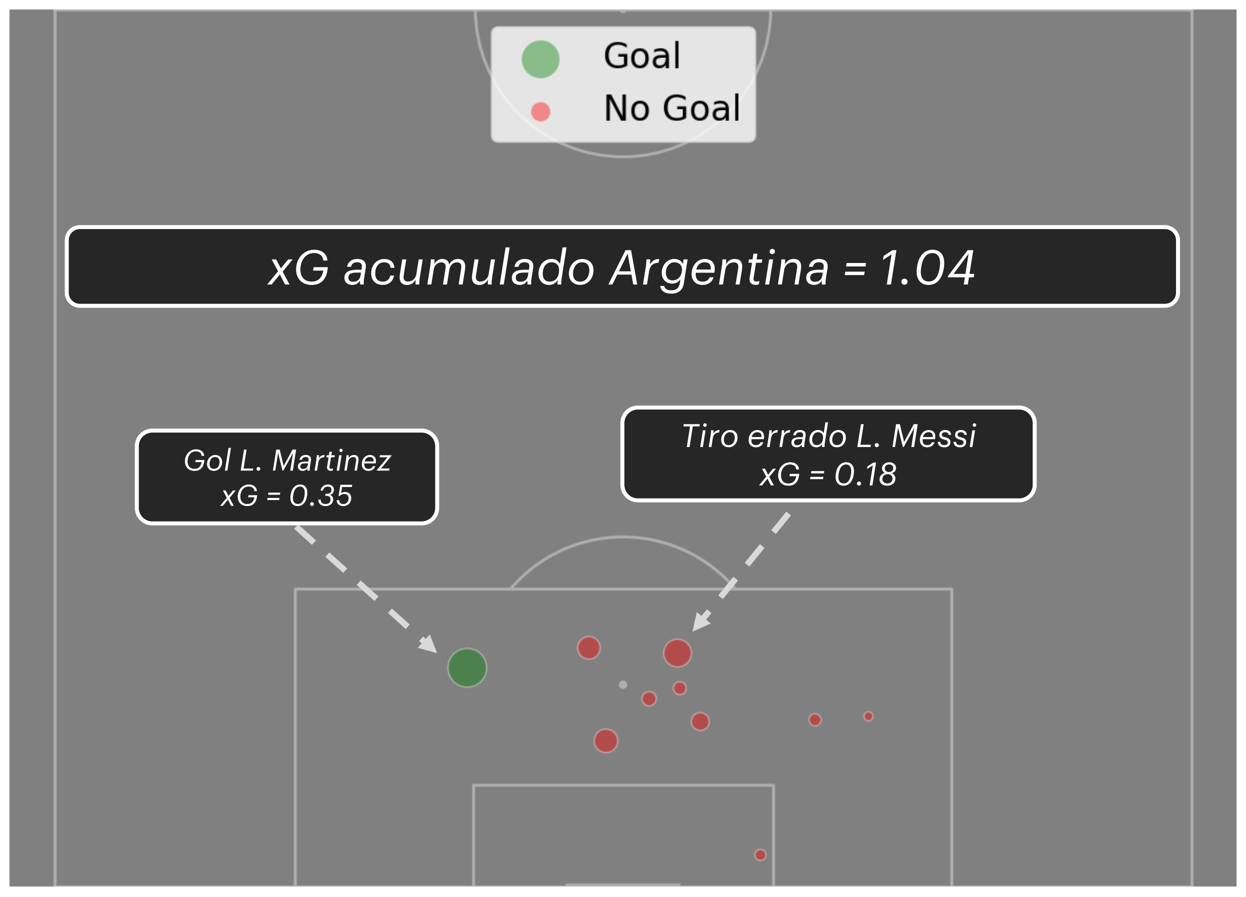
¿Qué es el concepto de xG y como usarlo para scouting?

Probabilidad (de 0 a 1) de que cualquier tiro termine en gol.

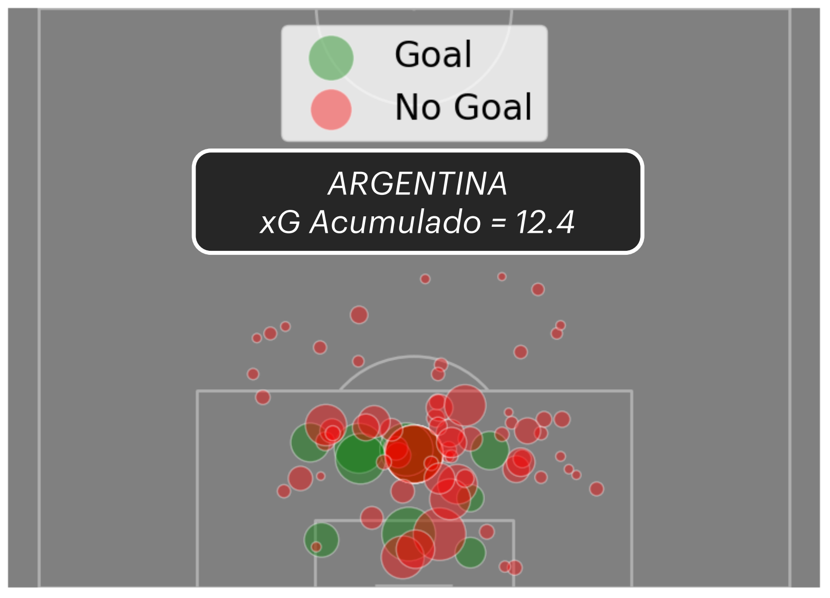
Se entrenan modelos de Machine Learning (clásicos, redes neuronales, etc.) con cientos de miles de tiros históricos y se le asigna la probabilidad.

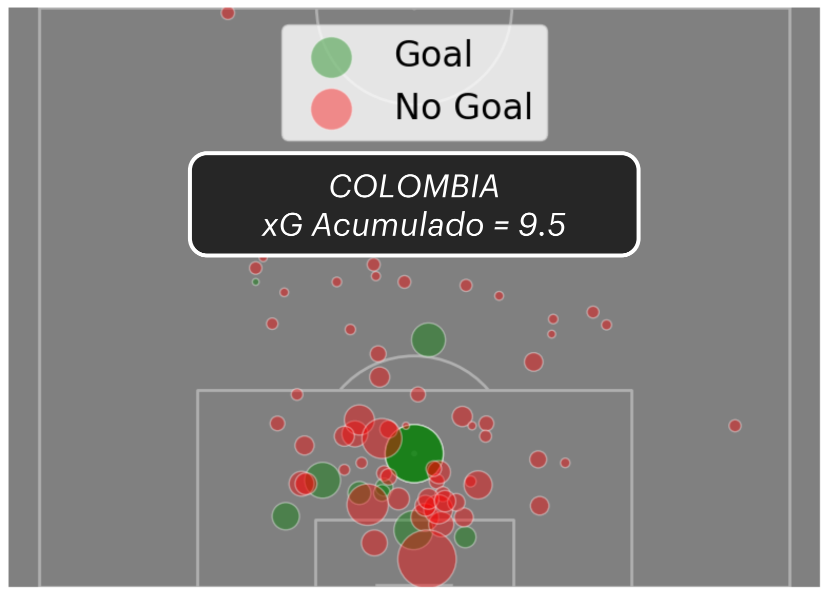
Se usan features como distancia al arco, ángulo de tiro, posición, tipo de disparo, etc.

Veamoslo con un ejemplo cercano en el tiempo (y que nos hace felices recordar)...



Usemos de ejemplo la Copa América...





Pais	xG Acumulado	xG en contra	Dif. xG	DG Real
Argentina	**12.4**	**5.0**	**7.4**	**8**
Colombia	**9.5**	**3.6**	**5.9**	**9**
Uruguay	**9.3**	**4.8**	**4.5**	**7**
Brazil	6.4	3.2	3.2	3
Mexico	4.9	2.1	2.8	0
USA	4.5	2.5	2.0	0
Ecuador	5.5	4.4	1.0	1
Venezuela	5.7	5.9	-0.2	5
Canada	**8.4**	**9.1**	**-0.6**	**-3**
Paraguay	2.9	4.9	-2.1	-5
Panama	3.1	5.6	-2.5	-4
Chile	1.7	4.6	-3.0	-1
Peru	1.7	5.2	-3.5	-3
Jamaica	1.8	5.4	-3.5	-6
Costa Rica	0.6	5.8	-5.2	-2
Bolivia	1.0	7.1	-6.2	-9

Pero entonces como lo uso para Scouting…

Veamos el top 10 de xG acumulado por jugador en la Copa América y Eurocopa

Jugador	xG Acumulado	Goles
L. Martinez	3.26	5
Ricardo Pepi	**2.72**	**0**
Darwin Nuñez	2.62	2
Salomón Rondón	2.44	3
Lucas Paqueta	2.29	1
Tani Oluwaseyi	**2.25**	**0**
Julián Alvarez	2.0	2
Lionel Messi	1.95	1
Kendry Paez	**1.76**	**1**
Jonathan David	**1.63**	**2**

Jugador	xG Acumulado	Goles
Kai Havertz	4.12	2
C. Ronaldo	3.60	0
Kylian Mbappe	2.92	1
Harry Kane	2.87	3
G. Mikautadze	**2.25**	**3**
Breel Embolo	2.12	2
A. Griezmann	1.97	0
Memphis Depay	1.96	1
Donyell Malen	1.88	2
Lamine Yamal	1.83	1