BACHELOR'S THESIS / ENGINEERING 2025

Estrelles del futur: Anàlisi i predicció de talent futbolístic

Gerard Cabot Agustin

upf. Universitat Pompeu Fabra Barcelona

Estrelles del futur: Anàlisi i predicció de talent futbolístic

TREBALL FI DE GRAU DE

Gerard Cabot Agustin

Directora: Dolors Sala Batlle

Grau en Enginyeria en Informàtica

Curs 2024-2025

Universitat Pompeu Fabra Barcelona

Escola d'Enginyeria

A la meva família, pel vostre suport i confiança.

Agraïments

En primer lloc, voldria expressar el meu agraïment a la meva tutora del treball, Dolors Sala Batlle. La seva guia, paciència i consells han estat fonamentals per a la direcció i el bon desenvolupament d'aquest projecte.

Així mateix, vull agrair a tots els professors que he tingut durant el Grau en Enginyeria en Informàtica. La seva dedicació i els coneixements transmesos han estat la base sobre la qual s'ha construït aquest projecte i la meva formació.

Finalment, estenc el meu agraïment a la meva família i amics. El seu suport durant aquest treball ha estat una contribució molt valuosa per poder-lo dur a terme.

Resum

Aquest treball presenta "Estrelles del Futur", una aplicació web desenvolupada per a l'anàlisi i predicció de talent en futbolistes joves. Mitjançant una arquitectura client-servidor i utilitzant les dades obertes de StatsBomb, el projecte ofereix eines de visualització interactives i un sistema de *scouting* basat en aprenentatge automàtic. La metodologia principal implementa un model per estimar el potencial màxim que un jugador assolirà en la seva carrera, una variable objectiva generada a través d'una mètrica heurística pròpia basada en la correlació de KPI. La contribució més innovadora és la capacitat que té l'usuari de construir i entrenar els seus propis models predictius, definint el talent segons la seva filosofia de joc. El model final, validat amb una metodologia temporal estricta, demostra una capacitat predictiva significativa i identifica qualitativament jugadors que posteriorment han assolit l'elit mundial, validant la viabilitat de l'eina.

Resumen

Este trabajo presenta "Estrelles del Futur", una aplicación web desarrollada para el análisis y la predicción de talento en futbolistas jóvenes. Mediante una arquitectura cliente-servidor y utilizando los datos abiertos de StatsBomb, el proyecto ofrece herramientas de visualización interactivas y un sistema de *scouting* basado en aprendizaje automático. La metodología principal implementa un modelo para estimar el potencial máximo que un jugador alcanzará en su carrera, una variable objetiva generada a través de una métrica heurística propia basada en la correlación de KPI. La contribución más innovadora es la capacidad del usuario para construir y entrenar sus propios modelos predictivos, definiendo el talento según su filosofía de juego. El modelo final, validado con una metodología temporal estricta, demuestra una capacidad predictiva significativa e identifica cualitativamente a jugadores que posteriormente han alcanzado la élite mundial, validando la viabilidad de la herramienta.

Abstract

This project introduces "Estrelles del Futur", a web application developed for the analysis and prediction of talent in young football players. Using a client-server architecture and leveraging StatsBomb's open data, the project provides interactive visualization tools and a machine learning-based scouting system. The core methodology implements a model to estimate the peak career potential a player will achieve, a target variable generated through a custom heuristic metric based on KPI correlation. The most innovative contribution is the user's ability to build and train their own predictive models, defining talent according to their game philosophy. The final model, validated with a strict methodology, demonstrates significant predictive power and qualitatively identifies players who later reached the world-class elite, thus validating the tool's viability.

Índex

Llista de figures

Llista de taules

Capítol 1

Introducció

1.1 Context i motivació

La passió pel futbol no comença quan pita l'àrbitre l'inici del partit i tampoc s'atura quan pita el final, és un sentiment que cadascú viu d'una manera o altra, però que és sempre present amb tu. Discussions amb amics i amigues, xerrades eternes i la tensió del mercat de fitxatges són algunes de les moltes característiques que tenen aquelles persones amb passió pel futbol. Jo soc una d'elles.

Per altra banda, des de petit he estat un amant de l'estadística en el món dels esports; sempre he estat atent a noves mètriques, com els *xG*, que són els gols esperats de cada xut (Statsbomb, 2022) i *xA*, les assistències esperades (StatsPerform, 2022), entre d'altres. Durant el meu grau, he descobert maneres de tractar amb dades que ni m'havia imaginat que eren possibles, i d'aquí surt la meva idea per aquest Treball de Fi de Grau.

En el futbol, la meva posició de treball somiada (a part de jugador, cosa que no es complirà), seria formar part de l'equip de *scouting*. La seva funció és identificar talents per incorporar al seu equip, i amb l'auge de la importància estadística en els esports he desenvolupat aquest projecte com una eina per aquests cercatalents.

1.2 Objectius del treball

L'objectiu principal d'aquest treball és desenvolupar una eina innovadora amb l'objectiu d'optimitzar i facilitar el procés d'identificació i avaluació de talents esportius, especialment en jugadors Sub-21. Els objectius específics són els següents:

- Dissenyar una interfície web interactiva: Crear una plataforma que mostri estadístiques clau i permeti als usuaris explorar el rendiment dels jugadors.
    - L'aplicació inclourà taules interactives per triar les dades a visualitzar.
- Desenvolupar algorismes predictius per posició: Implementar tres models, un per cada posició de joc (defensor, migcampista, atacant) basat en teoria matemàtica per identificar jugadors sub-21 amb alt potencial.
- Permetre als usuaris definir les seves pròpies mètriques per definir el model de potencial.
- Analitzar i estructurar les dades de StatsBomb (StatsBomb, https://statsbomb.com/es/, 2013): StatsBomb és una empresa reconeguda mundialment per l'anàlisi de dades d'esports i fan públiques un conjunt important de dades. En aquest treball es fa ús d'aquestes dades amb dos objectius:
    - Com a base de dades reals per la nostra plataforma i, per tant, per identificar els criteris que volem que tingui el programari (*software),* per testejar-lo.
    - Per analitzar aquestes dades amb el software i poder identificar els valors que els jugadors d'elit tenen en aquestes estadístiques i, per tant, definir bons indicadors d'identificació de talent.

Capítol 2

Estat de l'art

2.1 Introducció al *scouting*

Segons la Universitat Europea (Europea, 2023), el *scouting* futbolístic és un aspecte clau dins del futbol modern. Aquest procés combina l'observació directa de partits amb l'anàlisi de dades per identificar talents, analitzar rivals i buscar punts de millora del mateix equip.

Històricament, aquesta recerca ha depès de l'habilitat dels cercatalents i entrenadors per avaluar el rendiment i el potencial dels jugadors, però amb l'auge de l'anàlisi de dades aquest procés ha canviat. Avui en dia no només es fa servir aquestes opinions subjectives, sinó que també se li dona importància a l'estudi d'estadístiques i dades de cada jugador com a complement de l'opinió subjectiva.

El professor David Sumpter, en la seva anàlisi sobre l'aplicació de les dades al futbol, introdueix el concepte de *statistical scouting* (Sumpter, 2022). L'objectiu d'aquest enfocament és aprendre a utilitzar les dades per avaluar el potencial dels jugadors. Citant a Ian Graham, director d'investigació del Liverpool FC, Sumpter destaca que l'àrea on la ciència de dades pot tenir més impacte és el reclutament de jugadors. La raó és principalment econòmica: els fitxatges dels jugadors representen una part molt significativa de les despeses d'un club. Per tant, identificar joies amagades o jugadors infravalorats permet estalviar grans sumes de diners.

No obstant això, Sumpter adverteix que, tot i el poder de les dades, la cerca de talents "és més un art que una ciència" (Sumpter, 2022) i que simplement "fer un radar impressionant" (Sumpter, 2022) no és suficient. En aquest àmbit, un radar és una visualització gràfica, generalment de forma circular, que representa el rendiment d'un jugador en diverses mètriques clau. Tot i que és una eina potent per a la comparació ràpida, no captura tota la complexitat del joc.

Per això, Sumpter emfatitza la necessitat d'un "anar i venir" (Sumpter, 2022) entre l'anàlisi quantitativa i la qualitativa. L'anàlisi quantitativa es refereix a l'estudi de les estadístiques i mètriques numèriques (gols, passades, etc.). En canvi, l'anàlisi qualitativa implica l'observació directa del joc, l'avaluació del context de les accions i l'aplicació del coneixement expert d'un cercatalents. Un exemple clàssic que il·lustra aquesta dualitat és la famosa cita de Paolo Maldini: "Si he de fer una entrada, ja he comès un error" (Sumpter, 2022). Això demostra que un alt nombre d'entrades, una dada quantitativa, no sempre és positiu, ja que pot indicar un mal posicionament previ, una valoració qualitativa. Per tant, cal reconèixer les limitacions de les xifres i combinar-les sempre amb l'observació i el coneixement expert del joc.

2.2 L'evolució de les dades en el futbol: dades obertes com a catalitzador

La democratització de l'anàlisi de dades ha estat possible gràcies a la publicació de conjunts de dades obertes. Un treball pioner en aquest àmbit va ser el de Pappalardo (Pappalardo, et al., 2019), que van alliberar un gran volum de dades d'esdeveniments de partits de les principals lligues europees, provinents de Wyscout (Campodonico & Falzetti, 2004). Aquest conjunt de dades, amb informació espaciotemporal, va catalitzar una onada d'investigació acadèmica. Aquestes dades encara són disponibles al GitHub (koenvo, 2021).

Seguint aquesta línia, el present treball utilitza les dades obertes de StatsBomb (StatsBomb, open-data, 2022). Aquesta font de dades representa un pas endavant respecte a les anteriors, ja que ofereix un nivell de detall superior i mètriques enriquides, com els gols esperats.

2.3 Models acadèmics per a l'avaluació de jugadors

La disponibilitat de dades ha permès el desenvolupament de marcs teòrics i models computacionals cada cop més sofisticats.

Un punt d'inflexió va arribar amb Decroos et al. i la seva mètrica VAEP, *Valuing Actions by Estimating Probabilities* (Decroos, Van Haaren, Bransen, & Davis, 2019). Aquest model va canviar el paradigma en assignar un valor a cada acció amb pilota, superant les mètriques tradicionals centrades només en gols o assistències. El valor VAEP es calcula estimant com una acció modifica les probabilitats que l'equip marqui o concedeixi un gol en les següents jugades. Això permet quantificar la contribució total d'un jugador, tant ofensiva com defensiva, de manera molt més completa.

Paral·lelament, van sorgir els primers intents d'aplicar models d'aprenentatge automàtic directament al *scouting*. Barron et al. van utilitzar xarxes neuronals artificials (ANN) per predir la trajectòria futura dels jugadors, classificant-los en tres categories: els que baixarien de divisió, els que es mantindrien i els que pujarien a una lliga superior (Barron, Ball, Robins, & Sunderland, 2018). El seu model va assolir una precisió del 78,8% en el cas més clar (pujar vs. baixar), demostrant la viabilitat de les ANN. Tanmateix, el seu resultat és una classificació discreta menys matisada que la puntuació de potencial que busca aquest projecte.

Centrant-se específicament en talents joves, Sieghartsleitner et al. van investigar quins factors prediuen millor l'èxit futur. La seva conclusió clau va ser que el rendiment motor específic del futbol (SMP), com ara l'habilitat en el regat o el xut, és un predictor més fiable que el rendiment motor general (GMP), com la velocitat en esprints llargs o la capacitat de salt (Sieghartsleitner, Zibung, Zuber, & Charbonnet, 2019). Aquesta troballa és rellevant per l'enfocament en joves talents, però es diferencia d'aquest projecte en la font de dades: mentre ells van utilitzar dades de proves físiques, aquest treball es basa exclusivament en dades de rendiment obtingudes durant els partits.

En una línia més complexa, Simon Lacan va proposar un model d'apilament de xarxes neuronals per detectar talents (Lacan, 2024). Aquesta tècnica combina les prediccions de diverses xarxes per obtenir un resultat més robust, imitant el procés d'un cap de *scouting* que integra informes de diferents observadors. Tot i ser un enfocament potent, la seva naturalesa de caixa negra en dificulta la interpretabilitat, un problema que aquest treball busca mitigar amb la seva interfície interactiva.

Finalment, un dels projectes més avançats se centra en l'anàlisi de passades. Fernández et al. va desenvolupar un model d'*Expected Possession Value* (EPV) que no només valora l'inici i el final de la passada, sinó també el control de l'espai i el posicionament dels jugadors en tot moment (Fernández, Bornn, & Cervone, 2020). Aquest model ofereix una avaluació extremadament precisa, però requereix dades de seguiment que no estan disponibles públicament a StatsBomb Open Data, la qual cosa el fa inviable per a aquest projecte, però serveix com a referència del que és possible amb dades més completes.

2.4 Eines comercials i altres projectes

Més enllà del món acadèmic, existeixen eines comercials i projectes d'estudiants que aborden problemes similars. Un exemple destacat és Comparisonator (Zorlu, Batgun, & Yagiz, 2020), una plataforma d'anàlisi que utilitza intel·ligència artificial per avaluar el rendiment de cada jugador basant-se en més de 370 paràmetres, tenint en compte la seva posició i la dificultat de la lliga. El resultat és una puntuació de rendiment universal que s'actualitza partit a partit.

En l'àmbit acadèmic, projectes com el de *Football Player Analysis for Identifying Best Team using Machine Learning* (Ramnath & Priya, 2024) demostren l'aplicació de tècniques d'aprenentatge automàtic per a l'anàlisi de jugadors i la predicció de resultats. Un treball especialment rellevant és *Predicting the potential ability of football players in the Football Manager game* (Wijk, 2022), que

prediu l'atribut de *Potential Ability* de la base de dades del videojoc *Football Manager* (Football Manager, 1992).

Aquest projecte és un anàleg interessant, ja que utilitza una font de dades estructurada per predir un concepte abstracte de potencial; aquesta mètrica té un valor de 0 a 200 en el videojoc *Football Manager*, i és d'on he tret la inspiració de fer el càlcul en aquest rang.

2.5 Bretxes en l'estat de l'art actual

L'anàlisi de la literatura revela una tendència clara: la majoria dels models se centren a predir esdeveniments concrets a curt termini o en aplicar models de caixa negra que ofereixen una sortida binària i poc interpretable. Les eines comercials són potents, però propietàries i no permeten la personalització.

Aquest Treball de Fi de Grau aborda una bretxa específica en aquest panorama. En lloc de predir una estadística aïllada, l'objectiu és desenvolupar un valor de potencial global per a jugadors joves. No es tracta d'una predicció sobre una única mètrica, com els gols que marcarà un jugador la temporada que ve, sinó d'una estimació més abstracta del talent i la capacitat de desenvolupament futur, intentant capturar numèricament allò que un cercatalents anomena potencial.

La principal contribució d'aquest projecte és triple:

1. Enfocament en el potencial a llarg termini: Es desmarca dels models d'anàlisi de rendiment immediat.

2. Desenvolupament d'una eina interactiva: No es limita a ser un model estàtic, sinó una plataforma web amb interacció i decisió de l'usuari.

3. Personalització de models: Permet als usuaris definir les seves pròpies mètriques per construir models de potencial a mida. Aquesta funcionalitat democratitza l'anàlisi i reconeix que la definició de talent és subjectiva i depèn de la filosofia de joc, el club o les necessitats tàctiques.

Capítol 3

StatsBomb

En el món de l'anàlisi de dades esportives, la qualitat i el detall de la font de dades són fonamentals. Tot i que existeixen diversos proveïdors de dades de futbol com Wyscout o Opta (Cooney, 1996), aquest projecte es basa exclusivament en les dades de StatsBomb. Aquesta empresa, fundada per analistes, s'ha convertit en un referent en el sector per la profunditat i precisió de les seves dades.

3.1 Procés de recollida i generació de dades de StatsBomb

El procés de generació de dades de StatsBomb combina l'anàlisi manual de vídeo per experts amb la tecnologia (StatsBomb, StatsBomb Open Data Specification v1.1.pdf, 2019). Els aspectes més rellevants per a aquest projecte són:

- Anotació Humana Experta: La base són els enregistraments de vídeo oficials. Anotadors entrenats analitzen aquests vídeos, sovint fotograma a fotograma, per identificar i registrar amb precisió cada acció significativa, assegurant un alt nivell de detall.

- Catàleg d'Esdeveniments i Atributs: StatsBomb defineix un extens catàleg de més de 35 tipus d'esdeveniments. Cada esdeveniment s'enriqueix amb múltiples atributs generals (com *timestamp*, *player*, *location*) i específics del tipus (com *statsbomb*_xg per a un *Shot*), amb regles clares per garantir la consistència.

- Dades Espacials i Context Tàctic: La localització de cada esdeveniment es registra en una graella estandarditzada de 120x80 unitats. Aquesta és una normalització del terreny de joc a unes dimensions fixes de 120x80 iardes, on cada unitat representa una iarda. Encara que no s'especifica directament a la definició de les coordenades, la unitat de mesura es confirma a través d'altres atributs com *pass_length*, que la documentació de StatsBomb defineix explícitament en iardes. Aquest mètode permet comparar posicions de manera fiable entre estadis de mides lleugerament diferents. Per enriquir encara més l'anàlisi, s'inclouen fotogrames congelats (*Freeze frames)* per als xuts, que capturen la posició dels jugadors per proporcionar context tàctic.

- Enriquiment Analític: Les dades crues s'enriqueixen amb mètriques derivades de models propis, sent *statsbomb_xg* l'exemple més destacat. Aquesta mètrica dona un valor a cada xut segons la probabilitat de convertir-lo en gol en un rang del zero a l'u.

- Control de Qualitat i Versionat: Un procés de validació humana corregeix errors. StatsBomb no sobreescriu dades antigues, sinó que publica noves versions de l'especificació de dades quan s'afegeixen nous atributs o es milloren definicions. Per exemple, la versió 1.1 va introduir *Carry* com a tipus d'esdeveniment i va millorar definicions com *counterpress*. Cada partit a les dades inclou metadades que indiquen la versió de l'especificació utilitzada.

- Publicació i Accés Un cop generades i validades, StatsBomb distribueix les seves dades completes als clients professionals mitjançant les seves plataformes i API. De manera crucial per a la recerca i la comunitat, i fonamental per a aquest TFG, una selecció d'aquestes dades d'alta qualitat es publica al seu repositori de GitHub, StatsBomb Open Data.

3.2 Legalitat i protecció de dades

La publicació de dades de rendiment de futbolistes per part de StatsBomb s'emmarca dins d'una interpretació jurídica específica de la normativa de protecció de dades, principalment el Reglament General de Protecció de Dades (GDPR) de la Unió Europea.

La política de privacitat de StatsBomb, confirma que la base legal per al processament de les dades dels jugadors és el seu interès legítim. Concretament, estableixen que el seu interès legítim és "obtenir i utilitzar informació disponible públicament per a proporcionar un servei d'anàlisi de dades a la indústria de l'esport i els mitjans de comunicació" (StatsBomb, StatsBomb Privacy Policy, 2022). Aquesta declaració s'alinea directament amb el que permet l'Article 6 del GDPR (Parlament Europeu, 2016).

La validesa d'aquest interès se sosté en el fet que les dades provenen d'esdeveniments públics, com partits televisats, la qual cosa redueix l'expectativa raonable de privacitat del jugador sobre les seves accions professionals al camp. A més, la seva coneguda iniciativa de dades obertes es basa en aquest mateix principi, posant a disposició de la comunitat conjunta de dades per a la recerca i l'anàlisi, sempre sota un acord d'ús que l'usuari ha d'acceptar.

Finalment, les mètriques de rendiment publicades (passades, xuts, posició, etc.) no constitueixen "categories especials de dades personals" segons l'Article 9 del GDPR, ja que no inclouen informació delicada com dades de salut o biomètriques. Aquesta distinció és clau per a la legitimitat del tractament de dades sense necessitat d'un consentiment explícit per a cada dada generada en un partit.

3.3 Estructura de les dades

Les dades de StatsBomb s'organitzen en una estructura jeràrquica, que va des de les competicions fins als esdeveniments individuals de cada partit.

3.3.1 Dades de competicions

Aquestes dades proporcionen el context general, detallant informació sobre les lligues, temporades i els partits disputats. Són essencials per filtrar i entendre l'entorn de cada esdeveniment.

L'estructura completa d'aquestes dades es pot consultar a l'Apèndix B (vegeu Taula 5). Aquesta taula permet fer cerca de dades en funció del gènere, competició o temporada.

3.3.2 Dades de partits

Les dades de partits contenen informació sobre cada enfrontament individual, incloent-hi equips, resultats i metadades addicionals. Aquestes dades es vinculen a les alineacions i els esdeveniments per oferir una visió completa.

El detall d'aquests camps es troba a l'Apèndix B (vegeu Taula 6). Aquest nivell de detall permet anàlisis contextuals, com per exemple, estudiar el rendiment d'un jugador només en partits de fase de grups o analitzar si un àrbitre influeix en el nombre de faltes.

3.3.3 Dades d’alineacions

StatsBomb registra la informació de les alineacions dels jugadors, entrenadors i àrbitres implicats en cada partit. Les variables següents es recopilen a les alineacions de cada partit.

La descripció d'aquests camps es pot trobar a l'Apèndix B (vegeu Taula 7).

3.3.3.1 Informació dins de l'objecte d'alineació

La llista *lineup* conté les columnes que es poden veure a la Taula 8 de l'Apèndix B. Aquesta estructura és fonamental per al *scouting*, ja que permet rastrejar el rendiment d'un *player_id* específic al llarg de diferents temporades i competicions, independentment del club en el qual jugui.

3.3.4 Dades d'esdeveniments

El nucli de les dades de StatsBomb són els esdeveniments, que registren cada acció significativa que passa durant un partit. Un esdeveniment s'identifica per un conjunt de camps que descriuen què, qui, on, quan i com va succeir l'acció.

La definició completa dels camps d'esdeveniments es detalla a l'Apèndix B (vegeu Taula 9). Aquests atributs, combinats amb dades específiques de cada tipus d'esdeveniment, permeten una reconstrucció detallada del partit i anàlisis complexes, que són la base per a les visualitzacions i el model predictiu d'aquest projecte.

3.4 Volum i ús en el projecte

El repositori de StatsBomb Open Data ofereix un volum d'informació considerable, cobrint una gran varietat de competicions masculines i femenines d'arreu del món. No obstant això, per garantir la consistència i la rellevància de l'anàlisi, aquest projecte fa servir les dades de la Primera Divisió espanyola masculina, *LaLiga*.

Aquesta decisió es fonamenta principalment en el fet que, tal com es pot observar a la llista completa de competicions proporcionada, *LaLiga* és la competició amb el volum més gran de partits històrics disponibles al conjunt de dades obertes.

A l'Apèndix B (vegeu Taula 10) es pot consultar la llista completa de competicions i temporades que hi ha disponibles, juntament amb els partits totals.

Aquest conjunt de dades de partits de competició professional proporciona una base de dades real i rica, essencial per a les dues funcionalitats principals de l'aplicació. Per a la visualització i anàlisi, les dades d'esdeveniments permeten crear taules interactives i visualitzacions de mètriques, com mapes de passades o mapes de calor, que ajuden a explorar el rendiment dels jugadors.

La disponibilitat de múltiples temporades per a una mateixa lliga és especialment valuosa per observar l'evolució d'un jugador al llarg del temps, una funcionalitat clau implementada a la plataforma.

3.4.1 Emmagatzematge i organització de les dades

Per a la gestió eficient del considerable volum de dades que ofereix StatsBomb, es va dissenyar i implementar una arquitectura de fitxers local optimitzada per a les consultes específiques d'aquest projecte. Partint dels fitxers JSON originals, que contenen les dades completes de cada partit, es va executar un procés de transformació i agregació per reorganitzar la informació amb un enfocament centrat en el jugador.

Aquesta estructura organitza les dades jeràrquicament, creant primer una carpeta principal per a cada temporada analitzada. Posteriorment, dins de cada directori de temporada, es genera un fitxer individual per a cada futbolista que hi va participar. Cada un d'aquests fitxers de jugador consolida el llistat complet de tots els seus esdeveniments registrats al llarg d'aquella campanya.

Aquesta estratègia d'emmagatzematge, centrada en el jugador en lloc del partit, resulta fonamental per optimitzar el rendiment de l'aplicació. Permet carregar de manera gairebé instantània la totalitat de

dades d'un jugador específic per a una temporada concreta, evitant la necessitat de llegir, processar i filtrar repetidament els voluminosos fitxers de cada partit en cada consulta. Això redueix dràsticament la latència i la càrrega computacional de la plataforma web.

Capítol 4

Infraestructura i dispositius per a la captura de dades en el futbol

En el futbol professional modern, la qualitat i la precisió de les dades són essencials. Els dispositius i sistemes de captura asseguren que la informació utilitzada per a l'anàlisi sigui fiable i reflecteixi el veritable rendiment dels jugadors en competicions oficials. Sense aquestes eines, capturar el volum i la precisió de les dades necessàries per a l'anàlisi avançada —com la desenvolupada en aquest TFG— seria inviable. Aquest capítol explora les infraestructures i dispositius que permeten aquesta recollida de dades.

4.1 Infraestructures al camp

Les dades més rellevants per definir el talent i l'estil de joc d'un futbolista són aquelles que es generen directament al terreny de joc. Aquestes dades són la matèria primera de l'anàlisi tàctica i tècnica i, per tant, del *scouting* modern. Es capturen mitjançant infraestructures complexes instal·lades als estadis, principalment sistemes de càmeres.

4.1.1 Sistemes multicàmera per a l'anotació d'esdeveniments

La base per a la generació de dades d'esdeveniments, com les que utilitza StatsBomb i aquest projecte, és la gravació de partits des de múltiples angles. Als estadis professionals, s'instal·la una xarxa de càmeres estratègicament posicionades per garantir una cobertura completa.

Per a la temporada 2023/24, *LaLiga* va anunciar que les seves retransmissions tindrien un mínim de 18 càmeres per estadi, arribant a més de 32 en partits destacats (Broadcast, 2023). Aquesta configuració inclou:

- **Càmeres cinematogràfiques:** Situades molt a prop dels jugadors per oferir nous plans i apropar l'acció a l'espectador (SVG Europe, 2023).
- **Drons i càmeres aèries:** S'utilitzen nous models de drons per a vistes panoràmiques de més qualitat i es col·loca la càmera aèria a menys alçada per mostrar perspectives com el punt de vista del jugador en penals o faltes (SVG Europe, 2023).
- **Càmeres a les banquetes:** Per copsar les reaccions dels jugadors i el cos tècnic (SVG Europe, 2023).
- **Posicions tradicionals:** Es mantenen les càmeres clàssiques, com les de les línies de fons, darrere de les porteries i les càmeres tàctiques elevades per a la visió general del joc (SVG Europe, 2023).

Aquesta configuració multicàmera és fonamental perquè permet als analistes humans, i als sistemes automatitzats, revisar les jugades des de la millor perspectiva possible, reduir els punts cecs i registrar cada acció amb un alt grau de precisió i context, formant la base de les dades d'esdeveniments.

4.1.2 Sistemes de seguiment òptic

Més enllà de la gravació per a l'anàlisi manual, els sistemes de seguiment òptic representen la següent generació en la captura de dades. Es tracta d'una sèrie de càmeres d'alta resolució instal·lades a la part

superior de l'estadi que, mitjançant tècniques de visió per ordinador, rastregen automàticament la posició de tots els jugadors i de la pilota a una alta freqüència.

Empreses líders com Second Spectrum (Su, Maheswaran, & Chang, 2013) o ChyronHego (Mechner & Leonard, 1966) ofereixen aquesta tecnologia. El resultat són les anomenades dades de seguiment, un flux continu de coordenades que permeten anàlisis tàctiques molt més profundes sobre el posicionament, els desmarcatges, les distàncies entre jugadors i l'ocupació d'espais, aspectes que les dades d'esdeveniments no poden capturar completament.

4.2 Dispositius portàtils corporals: les dades del rendiment físic

Mentre que la infraestructura del camp captura el que fa un jugador, els dispositius corporals capturen el com ho fa des d'una perspectiva fisiològica i física. Aquestes eines són essencials per al *scouting*, ja que ajuden a avaluar si un jugador té el perfil físic adequat per a una lliga més exigent o per a un determinat estil de joc, a més de ser clau en la prevenció de lesions.

A la Taula 1 es detallen els dispositius portàtils més comuns, les mètriques clau que recullen i la seva aportació específica al procés de *scouting*.

Taula 1. Dispositius corporals portàtils i la seva aportació al scouting.

Dispositiu	Mètriques Clau	Aportació al *Scouting*	Proveïdors i Preu Orientatiu
Rastrejadors GPS	Distància total, velocitat màxima, acceleracions, mapes de calor.	Permet identificar jugadors amb una alta resistència o una gran capacitat d'explosió.	STATSports (Clark & O'Connor, 2008), Catapult (Holthouse & Griendt, 2006). Kits individuals: 200-400 €. Sistemes per a equips: diversos milers d'euros.
Monitors de freqüència cardíaca	Ritme cardíac (BPM), temps de recuperació.	Mesura la capacitat de recuperació després d'un esforç, un indicador clau de la condició física i el potencial de millora d'un jugador.	Garmin (Kao & Burrell, 1989), Polar (Säynäjäkangas, 1977). Preus variables, des de 50 € fins a 200 € per unitat.

4.3 Integració i importància de les dades per al *scouting*

La potència real per al *scouting* professional rau en la integració d'aquestes diferents fonts de dades.

D'una banda, les dades d'esdeveniments, obtingudes a partir dels sistemes de vídeo, són el cor d'aquest TFG i permeten avaluar les habilitats tècniques i la presa de decisions d'un jugador. Mètriques com les passades encertades o el rendiment sota pressió són fonamentals per al nostre model de potencial.

De l'altra banda, les dades físiques i fisiològiques, obtingudes dels dispositius portàtils, proporcionen el context sobre si un jugador té la capacitat atlètica per executar aquestes habilitats al més alt nivell. En conjunt, ofereixen una visió indispensable per a una avaluació completa del talent.

Capítol 5

Requisits

El desenvolupament de l'aplicació web "Estrelles del Futur" requereix establir una sèrie de requisits clars que garanteixin la viabilitat i l'enfocament del projecte. Aquest capítol defineix els requisits funcionals, no funcionals i les limitacions metodològiques que han guiat el disseny i la implementació del sistema.

5.1 Requisits funcionals

- R1: Àmbit d'usuari focalitzat en jugadors joves: El sistema està dissenyat per a l'anàlisi i *scouting* de talents en la categoria Sub-21. Això inclou un espectre ampli que va des del futbol base fins a les primeres etapes del futbol professional. Els usuaris objectiu són, per tant, analistes de dades, cercatalents de clubs, entrenadors i fins i tot els mateixos jugadors o les seves famílies que vulguin analitzar el seu rendiment.

- R2: Suport per a múltiples fonts de dades: Per ser útil en l'àmbit del futbol base on no hi ha dades públiques, el sistema ha de ser flexible en la captació de dades. Ha de permetre la ingesta de dades des de diverses fonts, incloent-hi la possibilitat d'una entrada manual de dades processades per un usuari a partir de l'anàlisi de vídeos de partits no oficials.

- R3: Adopció d'un model de dades estàndard: Totes les dades, independentment de la seva font, s'han d'estructurar internament seguint un model de dades consistent i reconegut. S'adopta el model de dades de StatsBomb com a estàndard de facto per la seva riquesa, detall i documentació pública. Això garanteix la coherència en l'anàlisi i la possibilitat de comparar jugadors de diferents contextos.

- R4: Arquitectura per a la gestió massiva de dades: El sistema ha d'estar dissenyat amb una arquitectura de dades que permeti gestionar i organitzar un volum creixent d'informació de manera eficient. Això implica una estructuració lògica de les dades per jugador, equip, competició i temporada, facilitant consultes complexes i l'entrenament de models predictius.

5.2 Requisits no funcionals

- R5: Plataforma i entorn d'execució: Inicialment, el sistema es desenvolupa i s'executa com una aplicació web en un entorn local. No obstant això, la seva arquitectura client-servidor ha de facilitar una futura migració a un servidor en línia. Per a l'usuari final, l'únic requisit ha de ser un navegador web modern, sense necessitat d'una alta capacitat de computació o memòria al dispositiu client.

- R6: Estratègia de protecció de dades: Per complir amb les normatives de protecció de dades, com el GDPR, i minimitzar els riscos de privacitat, no s'emmagatzemen els arxius de vídeo originals; en el seu lloc, es treballa amb les dades d'esdeveniments extretes d'aquests. Aquesta decisió redueix significativament els requisits d'emmagatzematge i la complexitat legal.

5.3 Restriccions metodològiques del projecte

- R7: Ús de dades públiques per al desenvolupament i validació: Tot i que el software està dissenyat per a l'àmbit Sub-21 (R1), per al desenvolupament i la validació dels models dins d'aquest projecte, s'utilitza exclusivament el conjunt de dades públiques StatsBomb Open

Data. Aquesta és una restricció pràctica i metodològica, imposada per la manca de disponibilitat pública de conjunts de dades de futbol base prou grans i detallats.

- R8: Reproductibilitat del treball acadèmic: L'ús de dades obertes (R7) i d'eines de codi obert garanteix que la investigació i els resultats presentats en aquest TFG siguin transparents i reproduïbles per altres investigadors, augmentant-ne el valor acadèmic i la verificabilitat.

Capítol 6

Teoria Matemàtica per a Models Predictius

La predicció del potencial en l'esport, i particularment en el futbol, és una tasca complexa que implica analitzar una gran quantitat de variables per estimar el desenvolupament futur d'un jugador. Els models predictius en aquest context busquen aprendre patrons a partir de dades històriques de rendiment per fer inferències sobre el potencial no observat, una mètrica que no es pot mesurar directament.

Una decisió metodològica clau en aquest projecte ha estat la de crear models predictius independents per a cada posició de camp (atacant, migcampista i defensor). El motiu és que les mètriques i els comportaments que defineixen l'èxit varien significativament entre rols. Un model especialitzat per posició té el potencial de capturar aquests matisos amb més precisió que un model generalista.

Aquest capítol explora diverses tècniques d'aprenentatge automàtic, justifica l'elecció del model per a aquest projecte i detalla la metodologia implementada per a la seva construcció.

6.1 Models predictius possibles

S'han considerat diversos algorismes d'aprenentatge automàtic per a la tasca de predicció de potencial. A continuació, es descriuen breument, justificant l'elecció final.

6.1.1 Regressió lineal

La regressió lineal és un dels models estadístics més fonamentals. El seu objectiu és modelar la relació entre una variable dependent (l'objectiu a predir) i una o més variables independents (les característiques) mitjançant l'ajust d'una equació lineal a les dades observades (Bishop, 2006). Per exemple, en el context del futbol, es podria intentar predir el valor de mercat d'un jugador basant-se en el nombre de gols i assistències que ha fet en una temporada.

Tot i la seva simplicitat i alta interpretabilitat, la seva principal limitació és l'assumpció de linearitat. Aquesta assumpció rarament es compleix en fenòmens complexos com el rendiment esportiu, on les interaccions entre variables solen ser no lineals i multifactorials.

6.1.2 Arbres de decisió

Un arbre de decisió és un model no paramètric que funciona dividint les dades en subconjunts cada cop més petits basant-se en una sèrie de regles condicionals aplicades als atributs, creant una estructura similar a un arbre (Hastie, Tibshirani, & Friedman, 2009). En l'àmbit del futbol, un arbre de decisió podria classificar un jugador com a talent d'alt potencial o talent de baix potencial seguint un camí de decisions com: "*Si gols_p90* > 0.5 i *pass_completion_rate* > 85%, llavors és un talent d'alt potencial". Són models molt visuals i fàcils d'interpretar.

El seu principal desavantatge és la tendència al sobreajust. Un arbre de decisió pot créixer fins a ser molt profund i complex, memoritzant el soroll i les particularitats de les dades d'entrenament en lloc d'aprendre els patrons generals. Com a resultat, generalitza malament a dades noves que no ha vist mai (Shalev-Shwartz & Ben-David, 2014). Aquest projecte, que treballa amb un nombre limitat de temporades per jugador, seria especialment vulnerable a aquest problema. Per la seva inestabilitat, un arbre de decisió individual es descarta com a model principal, encara que la seva lògica és la base per a models més robustos.

6.1.3 Random Forest

Random Forest és un algorisme d'aprenentatge basat en el concepte d'aprenentatge per conjunts, que millora significativament la robustesa dels arbres de decisió individuals (Breiman, 2001). El seu funcionament es basa a construir un gran nombre d'arbres de decisió durant l'entrenament. Cada arbre es construeix sobre una mostra aleatòria de les dades i considerant només un subconjunt aleatori de les característiques a cada node. Per fer una predicció, es recullen els resultats de tots els arbres del bosc i el resultat final és la mitjana (per a regressió) o la classe més votada (per a classificació).

Aquest procés redueix dràsticament el sobreajust, ja que els errors d'un arbre individual són compensats per la resta. Per exemple, per predir el potencial d'un jugador, construiríem centenars d'arbres, cadascun entrenat amb un subconjunt diferent de jugadors i les seves estadístiques. La predicció final del potencial seria la mitjana de les prediccions de tots els arbres.

És un candidat molt potent per a aquest projecte, ja que gestiona bé dades d'alta dimensionalitat, captura relacions no lineals i és menys propens al sobreajust. La seva eficàcia en l'anàlisi esportiva ha estat demostrada en diversos estudis. Per exemple, al projecte *Predicting Team Success in the Indian Premier League Cricket 2024 Season Using Random Forest* Analysis (S, S, Y., & Boby, 2024) van utilitzar un model de *Random Forest* per predir el rendiment de jugadors de criquet basant-se en dades històriques, destacant la seva capacitat per identificar les característiques més influents. Aquesta capacitat per ponderar la importància de les mètriques el fa especialment rellevant per al *scouting* de talent.

6.1.4 Xarxes neuronals artificials

Les Xarxes Neuronals Artificials (ANN) són models inspirats en l'estructura del cervell humà, formats per capes de nodes (neurones) interconnectats (Goodfellow-et-al-2016, 2016). Cada connexió té un pes associat, i cada neurona aplica una funció d'activació no lineal a la suma ponderada de les seves entrades. L'aprenentatge es produeix ajustant aquests pesos, generalment mitjançant l'algorisme de retropropagació, per minimitzar una funció de pèrdua.

En el context del futbol, una ANN podria rebre com a entrada un vector amb totes les estadístiques d'un jugador, com els gols, passades, intercepcions, etc., i, a través de les seves capes ocultes, aprendre a combinar aquestes característiques de manera complexa per produir una puntuació de potencial. De fet, l'article de *Barron* (Barron, Ball, Robins, & Sunderland, 2018), ja referenciat a l'estat de l'art, va aplicar amb èxit una ANN per classificar jugadors en categories de fitxar o no fitxar, demostrant la seva viabilitat en l'àmbit del reclutament professional.

Tot i la seva gran potència per modelar funcions altament no lineals, les ANN presenten dos desavantatges clau per a aquest projecte. Primer, requereixen grans quantitats de dades per entrenar-se eficaçment i evitar el sobreajust, un requisit que difícilment es compleix amb el conjunt de dades obert de StatsBomb, on el nombre de temporades per jugador és limitat. Segon, sovint són considerades caixes negres a causa de la dificultat per interpretar els milers de pesos i les decisions internes, la qual cosa dificulta l'extracció de coneixement accionable.

6.1.5 Regressió logística

La regressió logística és un model estadístic utilitzat per a tasques de classificació binària (Hastie, Tibshirani, & Friedman, 2009). Modela la probabilitat que una instància pertanyi a una classe particular mitjançant la funció logística, que transforma una combinació lineal de les característiques d'entrada en un valor entre 0 i 1.

En un context futbolístic, es podria utilitzar per predir la probabilitat que un xut acabi en gol basant-se en característiques com la distància a la porteria, l'angle del xut i el nombre de defensors propers. De fet, aquest és el principi darrere de mètriques com els gols esperats, mencionats prèviament.

Encara que és útil per a classificacions binàries, el seu ús en aquest projecte és limitat, ja que assumeix una relació lineal entre les característiques i el resultat, una limitació compartida amb la regressió lineal.

Per altra banda, l'objectiu del projecte és generar una puntuació de potencial contínua que permeti una ordenació i comparació matisada entre jugadors, no una classificació binària. Per tant, es descarta en favor de models de regressió més flexibles.

6.1.6 Extreme Gradient Boosting

Extreme Gradient Boosting *(XGBoost)* és una implementació optimitzada i escalable de l'algorisme de *gradient boosting* (Chen & Guestrin, 2016). El *gradient boosting* és, com Random Forest, una tècnica d'aprenentatge per conjunts que utilitza arbres de decisió. Tanmateix, en lloc de construir arbres de manera independent i paral·lela, ho fa de forma seqüencial.

El procés comença amb un model simple. Després, es calcula l'error que comet aquest model en les seves prediccions. A continuació, es construeix un segon arbre que, en lloc de predir el variable objectiu original, s'entrena per predir aquests errors. La predicció del conjunt esdevé la suma de les prediccions del primer i el segon arbre. Aquest procés es repeteix iterativament: cada nou arbre s'entrena per corregir els errors residuals del conjunt d'arbres anterior. D'aquesta manera, el model es va refinant gradualment, centrant-se en les instàncies més difícils de predir.

XGBoost millora aquest concepte amb característiques que li atorguen un rendiment superior en molts casos, especialment en competicions de ciència de dades sobre dades tabulars:

- Regularització: Incorpora termes de regularització L1 (Lasso) i L2 (Ridge) directament en la funció objectiva de l'entrenament. Això penalitza la complexitat dels arbres, ajudant a prevenir el sobreajust i millorant la generalització del model.

- Gestió de valors nuls: Pot gestionar valors absents a les dades de manera automàtica, aprenent el camí òptim per a les instàncies amb dades que poden faltar durant l'entrenament.

- Optimització i paral·lelització: Està dissenyat per ser computacionalment eficient, amb implementacions que aprofiten el processament en paral·lel i una gestió eficient de la memòria.

6.1.7 Anàlisi comparativa i elecció del model

Després d'avaluar les diferents alternatives, XGBoost s’ha escollit com l'algorisme principal per al desenvolupament del model predictiu de talent. Aquesta decisió es fonamenta en la seva capacitat demostrada per gestionar dades complexes i no lineals, i la seva robustesa contra el sobreajust gràcies a la regularització integrada i la seva flexibilitat.

A diferència dels models lineals, pot capturar les interaccions complexes inherents al rendiment esportiu. A diferència de les xarxes neuronals, no requereix volums de dades massius i és més tractable per al conjunt de dades disponible. I, tot i que Random Forest és una alternativa molt sòlida, l'enfocament seqüencial i de correcció d'errors de XGBoost sovint li proporciona un avantatge en termes de precisió predictiva, la qual cosa el posiciona com una opció excel·lent per a l'objectiu d'aquest projecte.

6.2 Implementació del model

Aquesta secció detalla la implementació pràctica del model, des de la configuració de l'entorn fins a la generació de la variable objectiu.

6.2.1 Configuració del model

El model XGBoost s'implementa utilitzant la llibreria xgboost de Python. Per optimitzar el seu rendiment i evitar la selecció manual de paràmetres, el projecte fa servir eines de l'ecosistema Scikit-learn:

- Optimització d'hiperparàmetres: S'usa RandomizedSearchCV de Scikit-learn (Cournapeau & Brucher, 2007) per explorar de manera eficient un espai ampli d'hiperparàmetres i trobar la combinació que maximitza el rendiment del model, evitant la selecció manual i propensa a errors.

- Escalat de característiques: Abans de l'entrenament, totes les característiques numèriques d'entrada són estandarditzades amb StandardScaler de Scikit-learn. Aquest procés transforma cada característica perquè tingui una mitjana de 0 i una desviació estàndard d'1. Encara que els models basats en arbres com XGBoost no són tan sensibles a l'escala com altres algorismes, l'estandarditzación millora l'estabilitat i la velocitat de convergència durant l'optimització.

- Validació creuada per grups: Per evitar la fuita de dades entre les diferents temporades d'un mateix jugador, la validació creuada es realitza amb GroupKFold. Aquesta tècnica assegura que totes les instàncies d'un mateix jugador romanguin juntes en el mateix subconjunt. Això proporciona una avaluació molt més realista i fiable del rendiment del model en jugadors completament nous, que és l'objectiu final.

La implementació precisa d'aquests principis, especialment la validació creuada i l'escalat, defineix la robustesa de l'avaluació del model. El projecte explora dues estratègies diferents per a aquest propòsit. Per mesurar la fiabilitat del model, es van implementar dues metodologies d'avaluació, reflectides en dos fitxers d'entrenament diferents:

1. **Aproximació 1: Validació creuada per grups (implementada a trainer_v2.py)**

    - Mètode: Aquesta estratègia utilitza GroupKFold de Scikit-learn sobre tot el conjunt de dades de jugadors Sub-21. Les dades es barregen i es divideixen en diversos plecs, on cada plec serveix alternativament com a conjunt de test.

    - Avantatge: És un mètode estàndard i robust que aprofita totes les dades per a l'entrenament i l'avaluació.

    - Limitació: Tot i agrupar per jugador, no simula un escenari temporal on es prediu el futur a partir del passat.

2. **Aproximació 2: Validació temporal estricta (implementada a trainer_v2_15_16.py)**

    - Mètode: Aquesta és una estratègia de validació temporal molt més rigorosa que simula un cas d'ús real de predicció. Es defineix una temporada completa, la 2015/2016, com un conjunt de test cec.

    - Entrenament: El model s'entrena exclusivament amb les dades de totes les temporades Sub-21 anteriors a la temporada de test (2015/2016). D'aquesta manera, el model només aprèn de

patrons del passat per predir el futur, evitant qualsevol fuita de dades de temporades posteriors.

- Avaluació: El rendiment final del model (MAE, R^2, etc.) es mesura únicament sobre les dades de la temporada 2015/2016, que el model no ha vist mai.

- Avantatge: Aquesta metodologia proporciona l'estimació més fiable del rendiment del model en un entorn de producció. Replica fidelment la tasca de predir el futur a partir de dades històriques, garantint que l'avaluació no estigui contaminada per informació futura.

6.2.2 Generació de la variable objectiu: "Potencial de pic de carrera"

Un dels reptes més significatius en la predicció del talent és que el potencial no és una etiqueta directament present a les dades. Cal crear una variable objectiva heurística que el representi. Aquest projecte implementa un procés sofisticat per generar una variable anomenada *peak_potential_target*.

L'objectiu no és predir el rendiment de la temporada següent, sinó estimar el nivell de rendiment màxim que un jugador jove assolirà en el futur.

El procés es pot descompondre en els següents passos:

1. Definició de l'univers de rendiment i KPI (*Key Performance Indicator*)

Per contextualitzar i normalitzar les mètriques de manera robusta, primer es processen totes les temporades de tots els jugadors disponibles al conjunt de dades, no només els menors de vint-i-un anys. Per a cada jugador-temporada, s'extreu un ampli vector de característiques base mitjançant la funció *extract_season_features*.

Aquesta funció, desenvolupada específicament per a aquest projecte, és l'encarregada de processar el conjunt de dades d'esdeveniments d'un jugador per a una única temporada. La seva tasca consisteix a agregar totes les accions registrades per calcular un conjunt exhaustiu de mètriques de rendiment. Aquestes mètriques inclouen des de recomptes totals i ràtios d'eficiència fins a estadístiques avançades normalitzades per 90 minuts de joc (*p90*), consolidant així el rendiment d'una temporada sencera en un únic vector de dades.

A continuació, es defineixen dos conjunts de KPI per a cada posició (Atacant, Migcampista, Defensor). Aquests conjunts, definits a *trainer_v2.py*, són la base per a la creació de la mètrica de rendiment.

- KPI d'Impacte (*COMPOSITE_IMPACT_KPIS*): A la Taula 2 es presenta un conjunt reduït de mètriques considerades representatives d'un rendiment d'alt nivell per a cada posició. La selecció i el format d'aquestes mètriques es basa en la seva relació directa amb accions decisives, normalitzades per temps de joc per permetre una comparació justa:

    - Atacants: Se centra en la producció de gols i la generació de perill. Les mètriques seleccionades són els gols esperats per 90 minuts (*sum_xg_p90_sqrt_*) i els gols reals per 90 minuts (*goals_p90_sqrt_*). Ambdues mètriques s'han transformat amb una arrel quadrada per estabilitzar la seva variància i reduir l'impacte de valors atípics. S'hi afegeixen els driblatges completats per 90 minuts (*dribbles_completed_p90*), que representen la capacitat de desequilibri individual.

    - Migcampistes: Es valora el control del joc i la contribució en les dues fases. Les mètriques són les intercepcions per 90 minuts (*interceptions_p90*), que reflecteixen la

intel·ligència tàctica defensiva; les assistències de gol per 90 minuts (*goal_assists_p90*), que mesuren la visió i el producte final; i les conduccions totals per 90 minuts (*carries_total_p90*), que indiquen la capacitat de progressar amb la pilota.

- Defensors: Es prioritzen les accions defensives fonamentals, totes normalitzades per 90 minuts per reflectir la seva taxa d'activitat. S'inclouen les entrades guanyades per 90 minuts (*tackles_won_p90*), els duels aeris guanyats per 90 minuts (*aerial_duels_won_p90*) i les intercepcions per 90 minuts (*interceptions_p90*) com a indicadors directes i objectius de l'èxit en les tasques defensives clau.

Taula 2: KPI d'impacte per defecte per posició.

Posició	**KPI d'Impacte per Defecte**
Atacant	*sum_xg_p90_sqrt_, goals_p90_sqrt_, dribbles_completed_p90*
Migcampista	*interceptions_p90, goal_assists_p90, carries_total_p90*
Defensor	*tackles_won_p90, interceptions_p90, aerial_duels_won_p90*

- KPI de Definició d'Objectiu (*KPI_DEFINITIONS_FOR_WEIGHT_DERIVATION*): Un conjunt més ampli de mètriques que, ponderades, formaran la puntuació de rendiment final de la temporada. La seva importància relativa es deriva de la seva correlació amb els KPI d'Impacte. La llista completa d'aquestes mètriques es pot consultar a l'Apèndix C (vegeu Taula 11).

2. Creació d'una mètrica heurística de rendiment per temporada

Per a cada temporada de cada jugador, es calcula una puntuació de rendiment (de 0 a 200). El mètode per construir aquesta puntuació s'inspira en pràctiques establertes en l'anàlisi de dades esportives, on l'objectiu és crear una mètrica composta que reflecteixi la qualitat del rendiment, especialment quan aquesta no és directament observable. L'estratègia general es basa a definir primer un indicador de rendiment d'elit i després utilitzar la correlació per ponderar un conjunt més ampli de mètriques. Aquest enfocament troba suport tant en la literatura acadèmica com en la pràctica professional.

En l'àmbit acadèmic, estudis com el de (Schlenger, Wunderlich, Raabe, & Memmert, 2023) demostren la validesa de fer servir anàlisis de correlació per identificar quins indicadors de rendiment (KPI) s'associen més fortament amb l'èxit d'un equip, proporcionant una base metodològica per a la ponderació de mètriques. En l'àmbit professional, un exemple anàleg és el desenvolupament de la mètrica *Passer Score* per part de l'equip de Next Gen Stats de la NFL (Lander Analytics, 2022)**,** on es va crear una puntuació composta per a *quarterbacks* ponderant diverses estadístiques segons la seva correlació amb la probabilitat de guanyar partits.

Per a cada temporada de cada jugador, es calcula una puntuació de rendiment (de 0 a 200) seguint aquests subpassos:

a) Derivació de pesos per correlació

El nucli d'aquesta metodologia rau a determinar objectivament el pes o la importància de cada KPI del model. En lloc d'assignar pesos subjectius, s'utilitza una tècnica basada en la correlació. Primer, es construeix un *composite_impact_score*, una puntuació que actua com a representació numèrica del rendiment d'elit. Aquesta puntuació es calcula sumant els valors normalitzats (entre 0 i 1) d'un petit conjunt de KPI d'Impacte prèviament seleccionats per a cada posició.

Un cop tenim aquest indicador d'elit, es calcula la importància (pes) de cada KPI (KPI_k) del conjunt ampli de "Definició d'Objectiu" mitjançant la següent fórmula:

Equació conceptual per al pes d'un KPI:

$w\_k = |\rho(KPI\_k, S\_impact)| / \Sigma |\rho(KPI\_i, S\_impact)|$

On:

- w_k és el pes final normalitzat del KPI *k*.
- $\rho(KPI\_k, S\_impact)$ és el coeficient de correlació de Pearson entre els valors del KPI *k* i la puntuació d'impacte (*S_impact*) a través de totes les observacions de jugador per temporada.
- El numerador |...| representa el valor absolut de la correlació, assegurant que tant correlacions positives com negatives contribueixin al pes.
- El denominador Σ|...| és la suma de totes les correlacions absolutes, un pas de normalització que garanteix que la suma de tots els pesos sigui 1.

Aquest mètode assegura que les mètriques que tenen una relació estadística més forta amb el rendiment d'elit (sigui positiva o negativa) rebin una major importància en el càlcul final del rendiment.

b) Càlcul de la raw_composite_score

Amb els pesos ja determinats, la puntuació de rendiment per a un jugador *j* en una temporada concreta es calcula com una suma ponderada de tots els seus KPI de definició d'objectiu, normalitzada finalment a una escala de 0 a 200 per facilitar-ne la interpretació i comparació.

Equació conceptual de la puntuació de rendiment:

$PuntuacióRendiment\_j = MinMax\_200 (\Sigma (w\_k * MinMax\_1(KPI\_k,j)))$

On:

- *PuntuacióRendiment_j* és la puntuació final per al jugador-temporada *j*.
- *MinMax_200* és una funció de normalització Min-Max que escala el resultat final a un rang de 0 a 200, considerant totes les puntuacions dins del mateix grup posicional.
- Σ representa la suma sobre tots els KPI de definició d'objectiu.
- w_k és el pes del KPI *k*, calculat en el pas anterior.
- *MinMax_1(KPI_k,j)* és el valor del KPI *k* per al jugador *j*, prèviament normalitzat a una escala de 0 a 1 dins del seu grup posicional.

Com a consideració pràctica per garantir la fiabilitat estadística, s'aplica una regla sobre els minuts jugats (*MIN_90S_PLAYED_FOR_P90_STATS = 3.0*). Si un jugador no assoleix aquest llindar de temps de joc en una temporada, els seus KPI normalitzats per 90 minuts s'estableixen a 0 per evitar que mètriques de jugadors amb pocs minuts distorsionin l'agregat.

3. Identificació del potencial de pic de carrera (peak_potential_target)

Un cop s'ha calculat una puntuació de rendiment (0-200) per a totes les temporades de la carrera de cada jugador, s'identifica el valor màxim assolit. Aquesta puntuació màxima es designa com el seu *peak_potential_target*. Aquesta és la veritable variable objectiva que el model aprendrà a predir.

Exemple conceptual:

- Suposem que, un cop aplicada la fórmula anterior, obtenim les següents Puntuacions de Rendiment calculades per a un jugador al llarg de la seva carrera:
    - Temporada 1 (19 anys): 110
    - Temporada 2 (20 anys): 125
    - Temporada 3 (21 anys): 140
    - Temporada 4 (22 anys): 165
    - Temporada 5 (23 anys): 155
- El seu peak_potential_target és 165. El model intentarà predir aquest valor.

4. Creació del conjunt de dades d'entrenament

Finalment, es construeix el conjunt de dades per a l'entrenament del model.

- Instàncies (Files): Són exclusivament les temporades en què els jugadors tenien 21 anys o menys. Cada temporada Sub-21 és una oportunitat perquè el model aprengui.
- Característiques (X): El vector d'entrada per a cada instància Sub-21 no només inclou les estadístiques d'aquella temporada, sinó que s'enriqueix amb informació contextual sobre la trajectòria del jugador. Conceptualment, està format per:
    1. Mètriques de la temporada actual: El rendiment del jugador en l'any que s'està avaluant (p. ex., als 19 anys).
    2. Agregats històrics: Resums estadístics de totes les temporades anteriors, com la mitjana, el valor màxim i la suma acumulada de cada KPI. Això proporciona al model una línia base del nivell històric del jugador.
    3. Indicadors d'evolució: Característiques que capturen la progressió del jugador, com la diferència de rendiment entre la temporada actual i l'anterior o la tendència general del seu rendiment al llarg del temps, calculada com el pendent d'una regressió lineal sobre les seves mètriques històriques.
- Variable objectiu (Y): L'etiqueta a predir per a cada instància Sub-21 és la mateixa: el *peak_potential_target* del jugador, calculat en el pas anterior.

La tasca predictiva del model, per tant, esdevé: donades les estadístiques completes i l'evolució d'un jugador de 21 anys o menys, predir quina serà la puntuació màxima de rendiment que assolirà al llarg de tota la seva carrera.

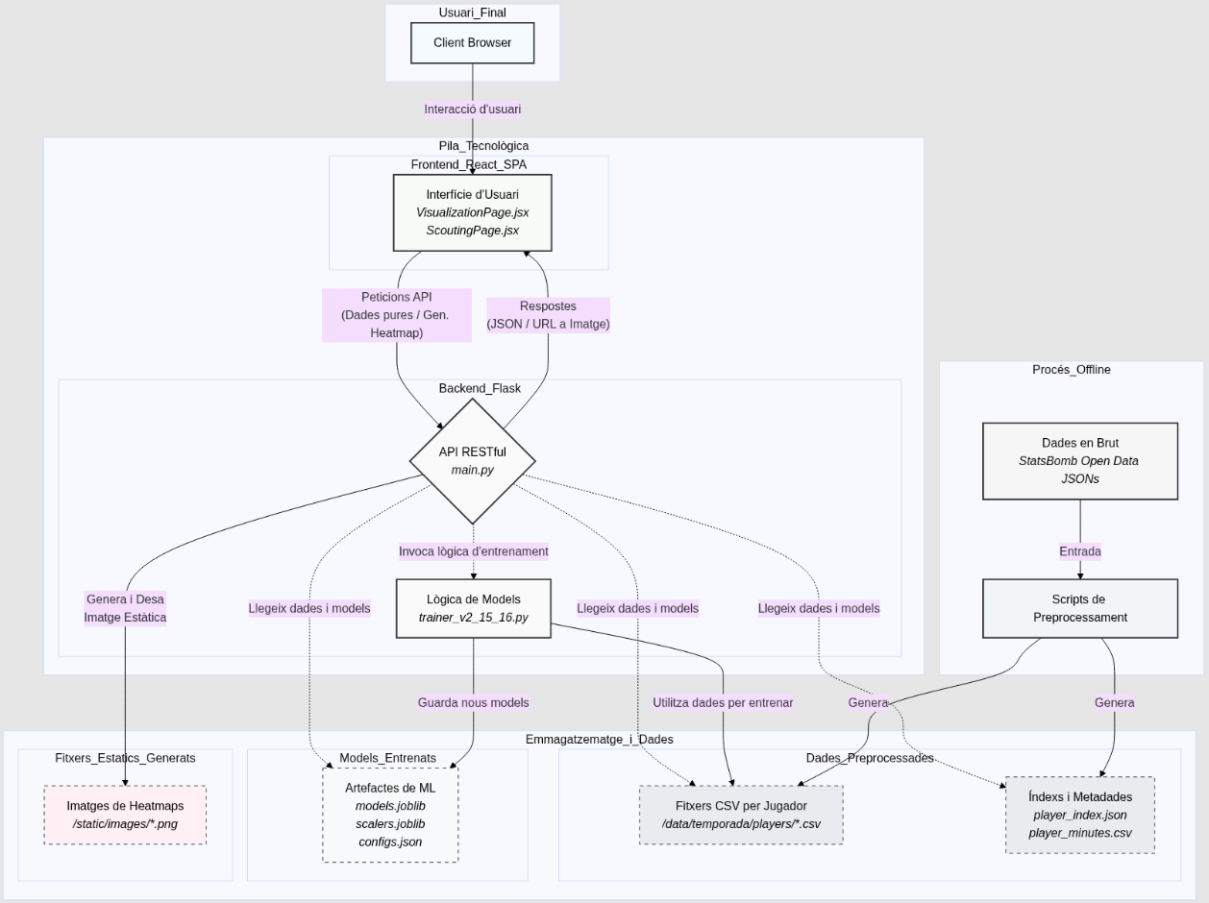
Capítol 7

El desenvolupament d'una aplicació robusta i escalable com "Estrelles del Futur" exigeix una base arquitectònica sòlida i ben definida. Aquest capítol descriu l'estructura fonamental del sistema, des de les decisions de disseny d'alt nivell fins als detalls tècnics concrets, connectant la pila tecnològica seleccionada amb la seva implementació pràctica.

Disseny i arquitectura del sistema

Per donar resposta als requisits funcionals i no funcionals establerts en el capítol 5, es va dissenyar una arquitectura de sistema basada en el patró client-servidor. Aquesta elecció estratègica permet una separació clara de responsabilitats entre la interfície d'usuari (client) i el processament de dades (servidor), facilitant el desenvolupament modular, el manteniment i una futura escalabilitat. La il·lustració 1 mostra els components principals del sistema i les seves interaccions.

Il·lustració 1. Components i organització del sistema



7.1 Arquitectura del servidor

El nucli del servidor es basa en l'ecosistema de Python per a la ciència de dades. La seva arquitectura es pot descompondre en mòduls lògics que gestionen diferents etapes del cicle de vida de les dades i els models.

7.1.1 Pila Tecnològica del *back-end*

El nucli del servidor es basa en l'ecosistema de Python (Rossum, 1991) per a la ciència de dades. La pila tecnològica es compon del *framework* Flask (Ronacher, 2010) per al servidor web, la llibreria Pandas per a la manipulació de dades, Scikit-learn i XGBoost per a la implementació dels models, i Matplotlib (Hunter, 2003) juntament amb Mplsoccer (Tracking, 2020) per a la generació de visualitzacions gràfiques.

7.1.2 Arquitectura de dades i mòduls de preprocessament

Un dels reptes principals del projecte és la gestió eficient del volum de dades de StatsBomb. Les dades originals s'ofereixen en fitxers JSON per partit, un format que resulta ineficient per a consultes centrades en un jugador concret al llarg de múltiples partits o temporades. Per resoldre això, es va executar un procés de preprocessament fora de línia que transforma i organitza les dades en una estructura optimitzada per al rendiment de l'aplicació.

L'estructura de dades final, guardada a la carpeta data, segueix un esquema jeràrquic centrat en el jugador:

data/[TEMPORADA]/players/[ID_JUGADOR]_[TEMPORADA].csv

Per exemple, tots els esdeveniments del jugador amb ID 6391 durant la temporada 2004_2005 es consoliden en un únic fitxer.

CSV: data/2004_2005/players/6391_2004_2005.csv.

Aquesta arquitectura de dades és fonamental, ja que permet a l'aplicació carregar de manera gairebé instantània tots els esdeveniments d'un jugador per a una temporada concreta, evitant la necessitat de llegir i filtrar repetidament els voluminosos fitxers originals de cada partit en cada petició.

Per accelerar encara més l'aplicació i evitar càlculs repetitius, es van desenvolupar diversos codis auxiliars per generar fitxers d'índex que actuen com una memòria de metadades precalculades. Per a la generació d'aquests scripts, que s'encarreguen d'extreure informació específica com els minuts jugats o la data de naixement, es va utilitzar assistència d'eines d'intel·ligència artificial generativa (Gemini 2.5, 2025), la qual cosa va agilitzar significativament el procés d'implementació de l'arquitectura de dades. Els fitxers d'índex més importants són:

- *player_index.json*: Actua com un índex mestre que relaciona el nom de cada jugador amb el seu *player_id*, les temporades disponibles i la data de naixement. Això accelera la càrrega inicial de l'aplicació, ja que evita haver d'escanejar tots els fitxers de dades per construir la llista de jugadors.

- *player_season_minutes_with_names.csv*: Conté el total de minuts jugats per cada jugador en cada temporada, una dada crucial que es calcula una sola vegada i es reutilitza constantment per a la normalització fiable de mètriques per 90 minuts.

7.1.3 Motor de l'aplicació i API RESTful

El motor de l'aplicació s'exposa a través d'una API RESTful definida a main.py. Aquesta API ofereix serveis per a les dues grans funcionalitats de l'aplicació, seguint dos patrons principals:

1. *Endpoints* de dades pures: Rutes com /players o /player_events retornen dades en format JSON que el client (React) s'encarrega de renderitzar. Això permet interfícies d'usuari dinàmiques i interactives.

2. *Endpoints* de renderització al servidor: Per a visualitzacions computacionalment més complexes, com els mapes de calor, el sistema adopta una estratègia de renderització al costat del servidor. Rutes com /position_heatmap generen la imatge PNG al servidor i retornen només el seu URL al client. Aquest patró redueix significativament la càrrega computacional sobre el navegador de l'usuari.

7.2 Arquitectura del client

La interfície d'usuari està construïda com una aplicació de pàgina única utilitzant React (Walke, 2013), la qual cosa proporciona una experiència d'usuari fluida i dinàmica.

7.2.1 Pila tecnològica del *front-end*

Es basa en React com a llibreria principal, usant react-router-dom per al moviment entre pàgines, axios per a la comunicació asíncrona amb l'API de Flask, i react-konva per al dibuix dinàmic.

7.2.2 Arquitectura de components i flux de dades

L'estructura s'organitza de manera jeràrquica. El component principal, App.jsx, defineix l'esquelet de l'aplicació, incloent-hi la capçalera i la navegació. Aquest, al seu torn, carrega els components de pàgina pertinents: VisualizationPage.jsx per a l'anàlisi visual i ScoutingPage.jsx per a la funcionalitat de *scouting*. Aquests components de pàgina orquestren un conjunt de components de visualització més específics.

Un exemple clar d'aquesta arquitectura es troba en la implementació de les visualitzacions. El component ShotMap.jsx il·lustra la renderització del costat del client: rep una matriu de dades de xuts en format JSON i utilitza react-konva per dibuixar dinàmicament cada acció, cosa que permet que hi hagi interactivitat. En canvi, components com PositionHeatmap.jsx demostren la renderització del costat del servidor, ja que només reben un URL d'una imatge generada i la seva única responsabilitat és mostrar-la.

7.2.3 Gestió de l'estat

Finalment, la gestió de l'estat de l'aplicació —com el jugador seleccionat, els filtres aplicats o els estats de càrrega— es realitza de manera local dins de cada component mitjançant les funcions natives de React useState, useEffect i useMemo, una estratègia eficient i prou potent per a l'abast d'aquest projecte.

Capítol 8

Interfície d'usuari i demostració del sistema

Aquest capítol té com a objectiu presentar la implementació pràctica de l'arquitectura i els models descrits prèviament a través de la interfície d'usuari de l'aplicació web "Estrelles del Futur". Es demostrarà com les decisions de disseny es tradueixen en una eina funcional i interactiva, capaç de complir els requisits establerts. Es navegarà per les dues seccions principals de l'aplicació: l'anàlisi visual de jugadors i el mòdul de cerca de talents.

8.1 Visió general de l'aplicació

Com es pot observar a la il·lustració 2 de l'Apèndix E, l'aplicació rep a l'usuari amb una interfície neta i centrada en la navegació principal. La capçalera mostra el nom del projecte, "Estrelles del Futur", i una barra de navegació clara que permet accedir a les dues grans àrees funcionals: "Anàlisi del jugador" i "Buscador de talents". Aquesta estructura, implementada amb react-router-dom, permet una experiència d'usuari fluida i sense recàrregues de pàgina, guiant l'usuari de manera intuïtiva cap a la funcionalitat desitjada.

8.2 Pàgina d'anàlisi de jugadors

Aquesta secció està dissenyada per a l'exploració detallada del rendiment d'un jugador individual. Un cop l'usuari selecciona un jugador (per exemple, Lionel Messi) i una temporada específica, la interfície carrega dinàmicament un conjunt de visualitzacions de dades per a una anàlisi exhaustiva.

Per comprendre la capacitat de creació i distribució d'un jugador, l'aplicació ofereix un mapa de passades complet, renderitzat al client amb react-konva, que mostra l'origen i el destí de cada passada, diferenciant per color entre passades completades i incompletes (vegeu il·lustració 3, Apèndix E). Complementàriament, un segon gràfic agrega aquestes dades per zones del camp, mostrant el percentatge d'encert (vegeu il·lustració 4, Apèndix E). Això permet a l'analista identificar no només els patrons de passada d'un jugador, sinó també la seva eficàcia en les àrees més decisives del terreny de joc.

Un dels components centrals d'aquesta secció és el mapa de xuts interactiu (il·lustració 5, Apèndix E). Aquesta visualització representa cada xut com un hexàgon sobre una meitat del camp, la mida del qual és directament proporcional al seu valor de gols esperats (*xG*), que permet identificar ràpidament la qualitat de l'ocasió. L'element interactiu clau és que, en passar el cursor sobre qualsevol xut, una informació emergent mostra el seu valor *xG* exacte.

Per analitzar patrons de comportament, l'aplicació genera mapes de calor al servidor, com el mapa de posicionament general i el de pressions defensives, que ofereixen una visió ràpida del radi d'acció i el compromís defensiu d'un jugador (il·lustració 6, Apèndix E). A més, l'eina compta amb visualitzacions especialitzades per a porters, oferint mètriques com el percentatge d'aturades (il·lustració 7) i la distribució de les seves accions (il·lustració 8).

Finalment, la pàgina permet a l'usuari triar qualsevol mètrica agregada per visualitzar-la de dues maneres. Primer, com un valor únic per a la temporada seleccionada (il·lustració 9, Apèndix E). Segon, i més potent, pot generar un gràfic de tendència per a qualsevol mètrica al llarg de totes les temporades registrades del jugador. La il·lustració 10 de l'Apèndix E mostra, a tall d'exemple, l'evolució del rendiment respecte a els gols esperats de Lionel Messi, una mètrica que mesura la seva capacitat de finalització per sobre o per sota de l'expectativa estadística.

8.3 Pàgina de cerca de talents

Aquesta és la secció on s'aplica el model predictiu per avaluar el potencial dels jugadors joves. Consta de dues funcionalitats principals: la predicció amb models existents i la creació de models personalitzats.

La funcionalitat principal permet a l'usuari obtenir una predicció del potencial d'un jugador de manera ràpida i intuïtiva. El flux d'usuari comença amb la selecció d'un jugador d'una llista filtrada que només mostra futbolistes Sub-21. Posteriorment, es tria una temporada específica d'aquest jugador per a la qual es vol generar la predicció. A continuació, l'usuari pot seleccionar el model a utilitzar, sigui el model per defecte o un dels models personalitzats que hagi creat prèviament.

En prémer el botó "Predir puntuació de potencial", l'aplicació realitza una crida a l'*endpoint scouting_predict* del servidor. El *back-end* carrega els artefactes del model seleccionat, construeix el vector de característiques per al jugador i la temporada especificats, i retorna la predicció. El resultat es presenta al *front-end* en una targeta clara i concisa, com es pot veure a la il·lustració 11 de l'Apèndix E, que mostra la puntuació de potencial predita en una escala de 0 a 200, acompanyada de dades contextuals clau com l'edat i la posició.

La contribució més innovadora d'aquesta pàgina és la capacitat que atorga a l'usuari per definir la seva pròpia definició de potencial i crear un model predictiu a mida. La interfície guia l'usuari a través d'un formulari estructurat, que es mostra a la il·lustració 12 de l'Apèndix E. El procés es divideix en tres passos:

1. Configuració inicial: L'usuari assigna un nom al seu model i selecciona el grup posicional sobre el qual es construirà.

2. Definició de rendiment d'elit: L'usuari selecciona un conjunt de KPI d'Impacte. Aquests KPI defineixen què significa, segons el seu criteri, un rendiment d'alt nivell per a aquella posició.

3. Definició de l'heurística de potencial: Es tria un conjunt més ampli de KPI de definició, que formaran la base de la puntuació que el model aprendrà a predir. Per a usuaris avançats, existeix un pas opcional per seleccionar manualment les característiques d'entrada que alimentaran el model XGBoost; si no es fa, el sistema les tria automàticament.

Per facilitar l'ús, la interfície inclou informació contextual que explica els conceptes tècnics. Un cop configurat, l'usuari pot iniciar la construcció del model al *back-end* amb un sol clic, que executarà el procés d'entrenament complet.

Capítol 9

Resultats i avaluació

Aquest capítol presenta els resultats obtinguts pel model predictiu i avalua tant quantitativa com qualitativa del sistema desenvolupat, analitzant-ne els punts forts i les limitacions.

9.1 Avaluació de la funcionalitat de visualització

La primera gran funcionalitat d'"Estrelles del Futur" és la seva capacitat com a eina d'anàlisi exploratòria. Aquest mòdul s'ha implementat amb èxit, complint els requisits de disseny d'una interfície web interactiva. L'aplicació permet als usuaris seleccionar un jugador i una temporada i, a través de crides a l'API del *back-end*, renderitza un conjunt de visualitzacions complexes.

Per exemple, l'anàlisi de la distribució de passades es materialitza en un mapa interactiu que mostra l'origen i destí de cada passada, diferenciant-ne l'èxit. De la mateixa manera, el mapa de xuts, construït amb react-konva, representa cada xut amb una mida proporcional al seu valor xG i un color que n'indica el resultat, oferint una anàlisi visual immediata de l'eficàcia del jugador. La capacitat de l'eina per adaptar-se a diferents posicions queda demostrada amb les visualitzacions específiques per a porters, que mostren mètriques com el percentatge d'aturades o la distribució de les seves accions.

Finalment, la funcionalitat d'anàlisi de l'evolució del jugador, que genera gràfics de tendència de qualsevol mètrica al llarg de múltiples temporades, s'ha implementat correctament, que permet als usuaris identificar patrons de desenvolupament a llarg termini. En conjunt, la secció de visualització constitueix una eina robusta i funcional per si mateixa.

9.2 Avaluació quantitativa del model predictiu

Per avaluar la precisió i la robustesa del model XGBoost, es van comparar les dues estratègies d'entrenament i validació detallades al capítol 6.2.1. A continuació, es presenten els resultats obtinguts per a cada aproximació.

9.2.1 Resultats de l'estratègia 1: Validació creuada per grups

En aquesta primera aproximació (*trainer_v2.py*), el model es va entrenar i avaluar utilitzant una validació creuada sobre totes les dades Sub-21. Aquesta tècnica garanteix que totes les temporades d'un mateix jugador pertanyin al mateix subconjunt, per evitar la filtració de dades de jugadors. Els resultats obtinguts van ser els següents:

Taula 3. Resultats de la validació creuada per grups.

Posició	R^2	MAE	RMSE
Atacant	0.372	31.93	48.75
Migcampista	0.348	46.14	57.56
Defensor	0.220	51.367	63.556

Anàlisi: Els resultats mostren una capacitat predictiva modesta amb aquesta estratègia de validació barrejada. El R^2 és relativament baix per a totes les posicions, indicant que una part significativa de la

variabilitat en el potencial de pic de carrera no és capturada pel model. Els errors absoluts (MAE) i quadràtics mitjans (RMSE) són considerables en l'escala 0-200. S'observa que el rendiment és pitjor per al grup de defensors, amb el menor R^2 i els majors errors. Això podria suggerir que les mètriques derivades exclusivament de les dades d'esdeveniments són menys predictives per a les posicions defensives en comparació amb altres rols.

9.2.2 Resultats de la Validació Temporal Estricta (Test sobre la temporada 2015/2016)

Aquesta segona aproximació implementa una estratègia de validació fora del temps per simular un escenari d'ús més realista. El model s'entrena utilitzant totes les dades de temporades disponibles prèvies la temporada 2015/2016. Aquesta temporada es reserva íntegrament com a conjunt de test cec. Aquesta temporada específica es va triar perquè, tal com s'ha documentat prèviament, StatsBomb Open Data ofereix un volum i riquesa de dades particularment alts per a aquesta temporada, proporcionant un conjunt de test d'alta qualitat per avaluar el rendiment del model en condicions òptimes de dades. Els resultats de l'avaluació sobre aquest conjunt de test dedicat són els següents:

Taula 4. Resultats de la validació temporal estricta.

Posició	R^2	MAE	RMSE	Variació R^2 vs. Estratègia 1
Atacant	0.342	16.75	22.41	**-8,1%**
Migcampista	0.290	38.57	46.98	**-16,7%**
Defensor	0.189	47.51	53.89	**-14,1%**

Anàlisi: Aquests resultats són reveladors. Com era d'esperar d'una metodologia de validació més estricta, els valors de R^2 són més modestos que els obtinguts en la primera estratègia. La columna "Variació % R^2" mostra aquesta disminució, la qual cosa valida l'enfocament: en evitar qualsevol fuita d'informació del futur, obtenim una mesura molt més honesta i fiable del rendiment real del model en un escenari de predicció real.

Aquests resultats validen la metodologia de validació temporal com l'eina correcta per a tasques predictives d'aquesta naturalesa. Tot i que els números són menys optimistes que els de la validació creuada, reflecteixen la veritable dificultat de predir el talent sense 'contaminació' de dades futures i, per tant, proporcionen una base molt més realista i honesta sobre la qual avaluar i millorar el model."

9.3 Avaluació qualitativa del model predictiu

Més enllà de les mètriques quantitatives, l'avaluació clau d'un model de *scouting* és respondre a la pregunta: és capaç d'identificar talent real? Per respondre-la, es va executar el model final (trainer_v2_15_16.py) per predir el potencial de tots els jugadors Sub-21 de la temporada 2015/2016. Aquesta anàlisi qualitativa dels resultats demostra la validesa pràctica del model.

9.3.1 Anàlisi de cas d'estudi: Temporada 2015/2016

Primer, s'analitzen les prediccions generades específicament per a la temporada 2015/2016, que va servir com a conjunt de test. Aquest exercici permet avaluar el comportament del model en un entorn concret i observar quins perfils de jugadors va identificar com a potencials talents en aquell precís moment.

En examinar les prediccions generades (vegeu les Taules 13, 14 i 15 de l'Apèndix D pels millors potencials per cada posició), el model demostra una notable capacitat per identificar jugadors que posteriorment van tenir carreres d'elit.

- Atacants: En la posició d'atacant, el model destaca jugadors com Iñaki Williams i Santi Mina. Menció especial mereixen Isaac Success i Antonio Sanabria, que, tot i no haver assolit el nivell de superestrella, eren considerats en aquell moment com dos dels talents joves amb més potencial de La Lliga.

- Migcampistes: El model atorga puntuacions altes a un grup de jugadors que es van convertir en estrelles internacionals. Entre els més destacats trobem a Fabián Ruiz i Mateo Kovačić. Aquests jugadors són identificats correctament com a talents de gran projecció.

- Defensors: Tot i tenir un R^2 més baix, el model identifica un grup de defensors d'altíssim nivell, com ara Aymeric Laporte, José María Giménez, João Cancelo i José Luis Gayà. Això suggereix que, tot i que el model té dificultats per generalitzar numèricament, el senyal de talent que captura és potent per identificar els casos més evidents.

9.3.2 Resum de les prediccions del top 10

Per realitzar una avaluació més completa i construir un rànquing definitiu de talents, es va executar el model final sobre totes les temporades Sub-21 disponibles. Per a cada jugador, es va identificar la seva temporada amb la predicció de potencial més alta. Aquest valor representa el "pic de potencial" que el model li atribueix.

Aquest enfocament permet comparar el talent inherent dels jugadors, independentment de si la seva millor temporada Sub-21 va ser la 2007-2008 o la 2018-2019. Els resultats d'aquesta anàlisi es resumeixen en les taules de Top 10 per posició que es troben a l'apèndix (Taules 16, 17 i 18), on es compara el rendiment dels dos models.

L'anàlisi d'aquestes taules comparatives confirma la validesa de la metodologia del Model 2. Per exemple, en la categoria d'atacants, el Model 2 identifica inequívocament a Lionel Messi com el talent generacional absolut, mentre que en la de defensors destaca figures com Gerard Piqué i Sergio Ramos. Aquestes taules demostren que el model metodològicament correcte no només és més fiable, sinó qualitativament superior, validant que la mètrica *peak_potential_target* captura un senyal de talent real.

9.4 Avaluació del compliment de requisits

En avaluar el projecte final enfront dels objectius inicials, es constata un compliment satisfactori de tots els requisits establerts. L'ús exclusiu de dades obertes de StatsBomb ha estat un pilar fonamental que garanteix la reproductibilitat i l'accessibilitat del treball. El focus en futbolistes professionals Sub-21 s'ha implementat rigorosament, tant en la interfície com en la lògica de predicció. L'arquitectura tecnològica planificada, basada en Python/Flask per al *back-end* i React per al *front-end*, s'ha materialitzat completament, donant com a resultat una aplicació web accessible des de qualsevol navegador modern.

Respecte al requisit de la capacitat predictiva, el resultat és mixt. S'ha demostrat una capacitat predictiva significativa per a atacants, modesta per a migcampistes i insuficient per a defensors. No obstant això, el projecte ha complert l'objectiu inicial en desenvolupar una eina que no només integra un model, sinó que permet als usuaris construir i entrenar els seus propis. Aquesta funcionalitat, combinada amb la interfície intuïtiva, converteix el projecte en una potent plataforma d'investigació, complint així l'esperit innovador del treball.

Capítol 10

Conclusions i treball futur

10.1 Disponibilitat i guia d'execució del projecte

Per garantir la total transparència i reproductibilitat d'aquest treball, el projecte "Estrelles del Futur" està disponible públicament en un repositori de GitHub:

https://github.com/gerardcabot/React-Flask

Totes les instruccions detallades per a la instal·lació, configuració i execució de l'aplicació es troben al fitxer README.md a l'arrel del repositori. A continuació, es presenta un resum del procés per a la seva posada en marxa en un entorn local.

1. Requisits: L'entorn de desenvolupament requereix Python 3.11 i Node.js.

2. Estructura de dades: A causa de la seva gran mida, el conjunt de dades del projecte no està inclòs en el repositori de GitHub. És necessari descarregar-lo manualment des de l'enllaç proporcionat al README.md i descomprimir la carpeta data a l'arrel del projecte.

3. Execució del servidor: Calen dos terminals:

    - Terminal 1 (*Back-end*): S'ha d'activar un entorn virtual de Python i instal·lar les dependències (pip install -r requirements.txt). Un cop configurat, s'inicia el servidor Flask amb la comanda python main.py. El *back-end* estarà actiu a http://localhost:5000.

    - Terminal 2 (*Front-end*): S'han d'instal·lar les dependències de Node.js (npm install) i iniciar el servidor de desenvolupament de React amb npm run dev. El *front-end* serà accessible des del navegador a http://localhost:5173.

Aquesta configuració permet a qualsevol usuari, com el tribunal d'avaluació, clonar, configurar i provar l'aplicació completa en el seu propi ordinador.

10.2 Conclusions

El projecte ha culminat amb èxit en la creació d'una aplicació web completament funcional que integra eines d'anàlisi de dades i d'aprenentatge automàtic per a l'avaluació de talent futbolístic. Les principals conclusions que es poden extreure d'aquest treball són les següents:

1. Desenvolupament d'una eina d'anàlisi funcional i robusta: S'ha dissenyat i implementat una arquitectura client-servidor sòlida, amb un *back-end* en Python (Flask) i un *front-end* en React. Aquesta estructura ha demostrat ser eficaç per gestionar un flux de dades complex, des de la càrrega i processament de dades d'esdeveniments fins a la renderització d'informació a la interfície d'usuari. La plataforma compleix amb els requisits funcionals i no funcionals, oferint una experiència d'usuari interactiva i fluida.

2. Validació de l'anàlisi heurística del potencial: El mètode dissenyat per generar una variable objectiva sintètica ha demostrat ser eficaç. L'avaluació qualitativa dels models predictius, fins i tot amb les seves limitacions, ha mostrat que el sistema és capaç d'identificar de manera consistent jugadors que han assolit l'elit mundial. Això valida que la metodologia de

ponderació de mètriques i la definició del "pic de carrera" capturen un senyal de talent real i significatiu.

3. Impacte de la qualitat de les dades i l'estratègia de validació: L'experimentació amb dues estratègies d'entrenament diferents ha revelat una conclusió crítica: la precisió del model predictiu és altament dependent de la consistència i qualitat de les dades d'entrada. El Model 2, entrenat amb una estratègia de retenció, va superar dràsticament el Model 1, subratllant la importància de fer avaluacions robustes que simulin escenaris del món real. Aquesta troballa és fonamental per a qualsevol aplicació pràctica d'aprenentatge automàtic en l'anàlisi esportiva.

4. Democratització del s*couting* basat en dades: La contribució més innovadora del projecte és la funcionalitat que permet als usuaris construir i entrenar els seus propis models predictius. En lloc d'oferir una "caixa negra", l'eina atorga a l'usuari el control sobre la definició de talent, i això li permet seleccionar els KPI que millor s'ajustin a la seva filosofia. Això no només resol parcialment el problema de la interpretabilitat, sinó que transforma la plataforma en un laboratori d'investigació per a analistes i aficionats.

En definitiva, Estrelles del Futur demostra que és factible desenvolupar una eina potent i accessible a partir de dades obertes, combinant una metodologia rigorosa d'aprenentatge automàtic amb un disseny d'interfície centrat en l'usuari.

10.3 Línies de treball futur

Malgrat l'èxit en l'assoliment dels objectius, "Estrelles del Futur" és una plataforma amb un enorme potencial de creixement. Les futures línies de treball són nombroses i podrien augmentar-ne significativament l'impacte, la precisió i la rellevància en el món de l'anàlisi esportiva.

Una de les expansions més naturals seria l'ampliació de les fonts de dades. La integració de dades de seguiment, si esdevenen accessibles, revolucionaria la capacitat analítica de l'eina. Aquestes dades permetrien calcular mètriques físiques avançades, com la distància recorreguda a alta intensitat, i indicadors tàctics complexos, com el control de l'espai, proporcionant una visió holística del rendiment que podria millorar dràsticament la precisió dels models. Aquest enriquiment es podria complementar amb dades contextuals, com l'historial de lesions o el valor de mercat, per modelar la trajectòria d'un jugador de manera més fidedigna.

Paral·lelament a l'enriquiment de les dades, l'arquitectura predictiva ofereix un ampli marge d'evolució. Un pas fonamental seria la inclusió dels porters dins del marc predictiu, desenvolupant un model específic per a ells amb KPI propis que mesurin la seva qualitat sota pressió i la seva contribució al joc. S'obre també la porta a experimentar amb un model generalista que no distingeixi per posicions, on la definició de talent recaigui exclusivament en els indicadors que l'usuari consideri clau, desafiant la hipòtesi inicial i atorgant una flexibilitat total.

Aquestes millores anirien acompanyades d'una evolució de la interfície d'usuari. La creació de perfils amb persistència, que permetin desar models personalitzats, es podria complementar amb un sistema d'avaluació més formal: quan un usuari entrenés un model, l'aplicació podria retornar un informe amb mètriques de rendiment clau, oferint un *feedback* quantitatiu sobre la seva qualitat. A més, es podrien desenvolupar noves eines com un quadre de control per a la comparació directa de jugadors o un motor de cerca de talents amb perfils estadístics similars, una funcionalitat de gran valor per a la identificació de joies amagades.

Finalment, abans de fer el salt definitiu, una validació qualitativa amb experts del sector — analistes, cercatalents o entrenadors— aportaria un segell de qualitat indispensable, contrastant les prediccions del model amb el coneixement del món real. Això prepararia el terreny per al desplegament de l'aplicació en un servidor al núvol, el pas definitiu per fer-la accessible públicament.

Aleshores, "Estrelles del Futur" podria passar de ser un projecte acadèmic a una eina viva i col·laborativa, usada per una comunitat global d'analistes i aficionats, complint plenament la seva missió de democratitzar l'anàlisi avançada del talent en el futbol.

Bibliografia

Barron, D., Ball, G., Robins, M., & Sunderland, C. (2018). *Artificial neural networks and player recruitment in professional soccer.* Recollit de https://www.researchgate.net/publication/328647803_Artificial_neural_networks_and_player_recruitment_in_professional_soccer

Bishop, C. (2006). *Pattern Recognition and Machine Learning.*

Breiman, L. (2001). *Random Forests*. Recollit de https://www.stat.berkeley.edu/~breiman/randomforest2001.pdf

Broadcast. (2023). *LaLiga reveals range of new camera angles for 2023/24*. Recollit de https://www.broadcastnow.co.uk/tech-innovation/laliga-reveals-range-of-new-camera-angles-for-2023/24/5184540.article#:~:text=The%20new%20cameras%20and%20angles,team%20benches%20to%20see%20player

Campodonico, M., & Falzetti, S. (2004). *Wyscout*. Recollit de https://wyscout.hudl.com

Chen, T., & Guestrin, C. (2016). *XGBoost: A Scalable Tree Boosting System.* Recollit de https://dl.acm.org/doi/10.1145/2939672.2939785

Clark, A., & O'Connor, S. (2008). *StatSports*. Recollit de https://statsports.com/

Cooney, A. (1996). *Opta*. Recollit de https://optaplayerstats.statsperform.com

Cournapeau, D., & Brucher, M. (2007). *scikit-learn*. Recollit de https://scikit-learn.org/stable/index.html

Decroos, T., Van Haaren, J., Bransen, L., & Davis, J. (2019). *Actions Speak Louder than Goals:*. Recollit de https://arxiv.org/pdf/1802.07127

Europea, U. (2023). *Scout en fútbol: como serlo y funciones*. Recollit de https://universidadeuropea.com/blog/scouting-futbol/

Fernández, J., Bornn, L., & Cervone, D. (2020). *A framework for the fine-grained evaluation of the instantaneous expected value of soccer possessions*. Recollit de https://arxiv.org/pdf/2011.09426

Football Manager. (1992). *Football Manager*. Recollit de https://www.footballmanager.com

Gemini 2.5. (2025). *https://aistudio.google.com*.

Goodfellow, I., Bengio, Y., & Courville, A. (2016). *Goodfellow-et-al-2016.* Recollit de Deep Learning: https://www.deeplearningbook.org/

Hastie, T., Tibshirani, R., & Friedman, J. (2009). *The Elements of Statistical Learning.*

Holthouse, S., & Griendt, I. v. (2006). *Catapult*. Recollit de https://www.catapult.com

Hunter, J. D. (2003). *Matplotlib*. Recollit de https://matplotlib.org/

Kao, D. M., & Burrell, G. (1989). *Garmin*. Recollit de https://www.garmin.com/es-ES/

koenvo. (2021). *wyscout-soccer-match-event-dataset*. Recollit de https://github.com/koenvo/wyscout-soccer-match-event-dataset

Lacan, S. (2024). *Stacking-based deep neural network for player scouting in football*. Recollit de https://arxiv.org/pdf/2403.08835

Lander Analytics. (2022). *Mike Band - The Many Ways to Create a Composite Score in Sports & Beyond*. Recollit de https://www.youtube.com/watch?v=iI4Let_eID8

Mechner, F., & Leonard, E. (1966). *ChyronHego*. Recollit de https://chyronhego.com/

Pappalardo, L., Cintia, P., Rossi, A., Massucco, E., Ferragina, P., Pedreschi, D., & Giannotti, F. (2019). *A public data set of spatio-temporal match events in soccer competitions.*

Parlament Europeu. (2016). *REGLAMENT (UE) 2016/679 DEL PARLAMENT EUROPEU I DEL CONSELL*. Recollit de https://apdcat.gencat.cat/web/.content/03-documentacio/Reglament_general_de_proteccio_de_dades/3132.pdf

Ramnath, A., & Priya, R. (2024). *Football Player Analysis for Identifying Best*. Recollit de https://iciset.in/Paper1245.pdf

Ronacher, A. (2010). *Flask*. Recollit de https://flask.palletsprojects.com/en/stable/

Rossum, G. v. (1991). *Python* . Recollit de https://www.python.org/

S, S., S, N., Y., L. P., & Boby, F. A. (2024). *Predicting Team Success in the Indian Premier League Cricket 2024 Season Using Random Forest Analysis.* Recollit de https://tmfv.com.ua/journal/article/view/2619

Säynäjäkangas, S. (1977). *Polar*. Recollit de https://www.polar.com/es/
Schlenger, J., Wunderlich, F., Raabe, D., & Memmert, D. (2023). *Systematic Analysis of Position-Data-based Key Performance Indicators*. Recollit de https://www.researchgate.net/publication/371659352_Systematic_Analysis_of_Position-Data-based_Key_Performance_Indicators
Shalev-Shwartz, S., & Ben-David, S. (2014). *Understanding machine learning.*
Sieghartsleitner, R., Zibung, M., Zuber, C., & Charbonnet, B. (2019). *Talent selection in youth football: Specific rather than general motor performance predicts future player status of football talents.* Recollit de https://www.researchgate.net/publication/337447059_Talent_selection_in_youth_football_Specific_rather_than_general_motor_performance_predicts_future_player_status_of_football_talents
StatsBomb. (2013). *https://statsbomb.com/es/*. Recollit de https://statsbomb.com/es/
StatsBomb. (2019). *StatsBomb Open Data Specification v1.1.pdf*. Recollit de https://github.com/statsbomb/open-data/blob/master/doc/StatsBomb%20Open%20Data%20Specification%20v1.1.pdf
StatsBomb. (2022). *open-data*. Recollit de https://github.com/statsbomb/open-data
StatsBomb. (2022). *StatsBomb Privacy Policy*. Recollit de https://statsbomb.com/privacy-policy/
Statsbomb. (2022). *What Are Expected Goals (xG)?* Recollit de https://statsbomb.com/soccer-metrics/expected-goals-xg-explained/
StatsPerform. (2022). *Opta Event Definitions*. Recollit de https://www.statsperform.com/opta-event-definitions/#:~:text=Expected%20assists%20(xA)%20measures%20the,point%20and%20length%20of%20pass.
Su, J., Maheswaran, R., & Chang, Y.-H. (2013). *Second Spectrum*. Recollit de https://www.secondspectrum.com
Sumpter, D. (2022). *Statistical Scouting*. Recollit de https://www.youtube.com/watch?v=MKHj2aZp8ms
SVG Europe. (2023). *LaLiga brings new season fresh camera angles and graphics*. Recollit de https://www.svgeurope.org/blog/headlines/laliga-brings-new-season-fresh-camera-angles-and-graphics/
Tracking, F. o. (2020). *Mplsoccer*. Recollit de https://mplsoccer.readthedocs.io/en/latest/
Walke, J. (2013). *React*. Recollit de https://es.react.dev/
Wijk, W. V. (2022). *Predicting the potential ability of football players in the Football Manager game.* Recollit de https://arno.uvt.nl/show.cgi?fid=161854
Zorlu, U. B., Batgun, T., & Yagiz, F. S. (2020). *Comparisonator*. Recollit de https://comparisonator.com/es

Apèndix

Apèndix A: Declaració sobre l'ús d'intel·ligència artificial generativa

En l'elaboració d'aquest Treball de Fi de Grau, s'ha fet ús de l'eina d'intel·ligència artificial generativa Gemini (Gemini 2.5, 2025) com a eina d'assistència per a tasques específiques de desenvolupament i redacció.

L'ús de l'eina es pot desglossar en les següents àrees:

1. Generació de codi per a scripts auxiliars: L'eina es va utilitzar per a la generació de la base de tres scripts de Python, la funció dels quals és auxiliar a la lògica principal del projecte. Aquests scripts, disponibles al repositori públic del projecte a https://github.com/gerardcabot/React-Flask/tree/main/server-flask/GenAI%20codes, van ser posteriorment revisats, adaptats i validats per l'autor. Aquesta aproximació va permetre accelerar el desenvolupament de components complementaris del sistema.

2. Assistència en la programació i refactorització de codi: Durant el desenvolupament del programari principal, es va emprar l'eina com un assistent de programació per suggerir millores d'estil, optimitzar el rendiment de funcions específiques i ajudar en la depuració d'errors. Totes les propostes generades van ser avaluades críticament, modificades i integrades manualment per l'autor per assegurar-ne la coherència i el correcte funcionament dins l'arquitectura del projecte.

3. Suport en la redacció i millora del text: S'ha utilitzat l'eina per a la millora de la qualitat de la redacció d'aquesta memòria. Les tasques concretes han estat la correcció gramatical i ortogràfica, la reformulació de frases per millorar-ne la claredat i la precisió tècnica, i la proposta de sinònims. En cap cas es va usar per a la generació d'idees, arguments o contingut original. La responsabilitat sobre l'estructura, el contingut i la validesa de la informació presentada recau íntegrament en l'autor.

Apèndix B: Estructura de dades de StatsBomb

Aquest apèndix detalla l'estructura dels principals fitxers de dades utilitzats en el projecte, provinents de les dades obertes de StatsBomb. Cada taula descriu les columnes, el tipus de dada i un exemple per facilitar-ne la comprensió.

Taula 5.Descripció dels camps de dades de competicions (competitions.json).

	Columna	Tipus	Descripció	Valors Exemple
1	competition_id	Integer	Identificador únic de la competició	2
2	season_id	Integer	Identificador únic de la temporada	1
3	competition_name	String	Nom de la competició	"Premier League"
4	season_name	String	Nom de la temporada	"2018/2019"
5	gender	String	Gènere de la competició	"Male"

Taula 6. Descripció dels camps de dades de partits (matches.json).

	Columna	Tipus	Descripció	Valors Exemple
1	match_id	Integer	Identificador únic del partit	12345
2	match_date	Date	Data del partit	"2018-03-10"
3	kick_off	Time	Hora d'inici del partit	"20:00:00"
4	home_team	Object (id/name)	Identificador i nom de l'equip local	1 / "Arsenal"
5	away_team	Object (id/name)	Identificador i nom de l'equip visitant	1 / "Chelsea"
6	home_team_gender	String	Gènere de l'equip local	"male"
7	away_team_gender	String	Gènere de l'equip visitant	"male"
8	home_team_group	String	Grup/conferència de l'equip local	"Group A"
9	away_team_group	String	Grup/conferència de l'equip visitant	"Group B"
10	home_team_country	Object (id/name)	País d'origen de l'equip local	68 / "England"
11	away_team_country	Object (id/name)	País d'origen de l'equip visitant	68 / "England"
12	home_team_managers	Data Frame	Informació de l'entrenador local (id, nom, etc.)	471 / "Óscar Pareja"
13	away_team_managers	Data Frame	Informació de l'entrenador visitant (id, nom, etc.)	472 / "Mikel Arteta"
14	home_score	Integer	Gols finals de l'equip local	2
15	away_score	Integer	Gols finals de l'equip visitant	1
16	match_status	String	Estat del partit (disponible, programat, etc.)	"available"
17	match_week	Integer	Setmana de la competició	25
18	competition_stage	Object (id/name)	Fase de la competició	1 / "Regular Season"

	Columna	Tipus	Descripció	Valors Exemple
19	stadium	Object	Nom i país de l'estadi	"Emirates Stadium" / "England"
20	referee	Object	Nom i país de l'àrbitre	"Anthony Taylor" / "England"
21	last_updated	DateTime	Data i hora de l'última actualització	"2018-08-08T15:44:27"
22	metadata	Object	Informació sobre versions de dades (data_version)	"1.1.0"

Taula 7. Descripció dels Camps Principals d'Alineacions (lineups.json).

	Columna	Tipus	Descripció	Valors Exemple
1	team_id	Integer	L'identificador únic de cada equip	5079
2	team_name	Integer	Nom de l'equip	"FC Barcelona"
3	lineup	Array	Una llista de jugadors a la plantilla d'aquest equip. Vegeu a continuació els detalls de la llista de jugadors.	Array

Taula 8. Detall de la subtaula d'alineacions (lineup).

	Columna	Tipus	Descripció	Valors Exemple
1	player_id	Integer	Identificador únic del jugador	5079
2	player_name	String	Nom del jugador	"Zlatan Ibrahimovic"
3	player_nickname	String	Sobrenom del jugador	"Zlatan"
4	jersey_number	Integer	Número de samarreta del jugador	9
5	country	Object (id/name)	Nacionalitat del jugador	211 / "Sweden"

Taula 9. Descripció dels camps comuns de dades d'esdeveniments.

	Columna	Tipus	Descripció	Valors Exemple
1	id	UUID	Identificador únic de l'esdeveniment	"0052d1b5-e2b0-4629-bbea-c18c884ab103"
2	index	Integer	Ordre seqüencial de l'esdeveniment	1
3	period	Integer	Període del partit (meitats, pròrroga, etc.)	1 (1a meitat)
4	timestamp	Timestamp	Moment exacte de l'esdeveniment	"00:00:06.293"
5	minute	Integer	Minut del rellotge al moment de l'esdeveniment	40
6	second	Integer	Segon del timestamp	15
7	type	Object (id/name)	Tipus d'esdeveniment (Pass, Shot, etc.)	30 / "Pass"
8	possession	Integer	Número de possessió única en el partit	5

	Columna	Tipus	Descripció	Valors Exemple
9	possession_team	Object (id/name)	Equip que inicia la possessió	1 / "Arsenal"
10	play_pattern	Object (id/name)	Patró de joc (ex., From Corner)	1 / "Regular Play"
11	team	Object (id/name)	Equip relacionat amb l'esdeveniment	1 / "Arsenal"
12	player	Object (id/name)	Jugador relacionat amb l'esdeveniment	5079 / "Zlatan Ibrahimovic"
13	position	Object (id/name)	Posició del jugador en l'esdeveniment	1 / "Goalkeeper"
14	location	Array [x, y]	Coordenades de l'esdeveniment	[60, 40]
15	duration	Decimal	Durada en segons de l'esdeveniment	1.5
16	under_pressure	Boolean	Si l'acció es realitza sota pressió	TRUE
17	off_camera	Boolean	Si l'esdeveniment ocorre fora de càmera	FALSE
18	out	Boolean	Si el resultat és la pilota fora de joc	TRUE
19	related_events	Array [UUID]	Identificadors d'esdeveniments relacionats	["2b7d06c7-9bcb-4bbf-a6e5-08e54e1303ac"]

Taula 10. Llista de competicions i temporades disponibles a les dades.

	Competició	País	Temporades	Partits totals
1	Bundesliga	Germany	2015/2016, 2023/2024	340
2	African Cup of Nations	Africa	2023	52
3	Champions League	Europe	1970/1971, 2004/2005 - 2018/2019	19
4	Copa America	South America	2024	32
5	Copa del Rey	Spain	1977/1978, 1982/1983, 1983/1984	3
6	FA Women's Super League	England	2018/2019, 2019/2020, 2020/2021	326
7	FIFA U20 World Cup	International	1979	1
8	FIFA World Cup	International	1958, 1962, 1970, 1974, 1986, 1990, 2018, 2022	147
9	Indian Super League	India	2021/2022	115
10	La Liga	Spain	1973/1974, 2004/2005 - 2020/2021	915
11	Liga Profesional	Argentina	1981, 1997/1998	2
12	Ligue 1	France	2015/2016, 2021/2022, 2022/2023	435
13	Major League Soccer	United States of America	2023	6
14	North American League	North and Central America	1977	1

	Competició	País	Temporades	Partits totals
15	NWSL	United States of America	2018	36
16	Premier League	England	2003/2004, 2015/2016	418
17	Serie A	Italy	1986/1987, 2015/2016	381
18	UEFA Euro	Europe	2020, 2024	102
19	UEFA Europa League	Europe	1988/1989	3
20	UEFA Women's Euro	Europe	2022	31
21	Women's World Cup	International	2019, 2023	116

Apèndix C: Llista completa de KPI per a la definició d'objectiu

Aquesta taula detalla el conjunt complet de mètriques que es poden utilitzar per construir la puntuació de rendiment d'un jugador. Aquestes són les mètriques que reben un pes derivat de la seva correlació amb els KPIs d'Impacte.

Taula 11. KPI de definició d'objectiu agrupats per posició.

Posició	**KPI**
Atacant	goals, goals_p90, goals_p90_sqrt_
	sum_xg, sum_xg_p90, sum_xg_p90_sqrt_
	shots_total, shots_total_p90
	shots_on_target, shots_on_target_p90
	conversion_rate_excl_xg_kpi
	goal_assists, goal_assists_p90
	dribbles_completed, dribbles_completed_p90
	carries_total, carries_total_p90
	pressures, pressures_p90
	turnovers_p90_inv_kpi_base
	aerial_duels_won, aerial_duels_won_p90
Migcampista	successful_passes, successful_passes_p90
	pass_completion_rate_kpi
	goal_assists, goal_assists_p90
	carries_total, carries_total_p90
	dribbles_completed, dribbles_completed_p90
	interceptions, interceptions_p90
	tackles_won, tackles_won_p90
	ball_recoveries, ball_recoveries_p90
	pressures, pressures_p90
	turnovers_p90_inv_kpi_base
Defensor	tackles_won, tackles_won_p90
	tackle_win_rate_kpi
	interceptions, interceptions_p90
	clearances, clearances_p90
	blocks_total, blocks_total_p90
	aerial_duels_won, aerial_duels_won_p90
	aerial_duel_win_rate_kpi
	successful_passes, successful_passes_p90
	pass_completion_rate_kpi
	carries_total, carries_total_p90
	turnovers_p90_inv_kpi_base

Apèndix D: Rànquings de potencial predits

Aquest apèndix presenta els rànquings Top 10 de jugadors Sub-21 per a cada posició, generats a partir dels dos models predictius desenvolupats. El "Model 1" correspon a trainer_v2.py (Validació Creuada) i el "Model 2" correspon a trainer_v2_15_16.py (Validació Temporal Estricta), que és el model final i metodològicament més robust. Les dades s'obtenen executant cada model sobre totes les temporades Sub-21 disponibles per identificar el pic de potencial de cada jugador.

Taula 12. Top 10 atacants Sub-21 per potencial predit (Temporada 2015-2016).

Rànquing	Nom del Jugador	Edat	Potencial
1	Isaac Ajayi Success	19	78,61
2	Arnaldo Antonio Sanabria Ayala	19	63,77
3	Armindo Tué Na Bangna	21	63,12
4	Iñaki Williams Arthuer	21	60,42
5	Santiago Mina Lorenzo	20	59,41
6	Carlos Castro García	20	57,43
7	Adalberto Peñaranda Maestre	18	56,13
8	Mikel Oyarzabal Ugarte	18	55,62
9	Moisés Gómez Bordonado	21	54,79
10	Sandro Ramírez Castillo	20	51,76

Taula 13. Top 10 migcampistes Sub-21 per potencial predit (Temporada 2015-2016).

Rànquing	Nom del Jugador	Edat	Potencial Predit
1	Josip Radošević	21	101,72
2	Joan Jordán Moreno	21	101,72
3	Fabián Ruiz Peña	19	100,67
4	Matías Nahuel Leiva Esquivel	19	100,67
5	Mateo Kovačić	21	100,00
6	Jefferson Andrés Lerma Solís	21	99,80
7	Álvaro Medrán Just	21	99,80
8	Pablo Fornals Malla	19	99,80
9	Ángel Martín Correa	20	98,75
10	Danilo Barbosa da Silva	19	98,75

Taula 14. Top 10 defensors Sub-21 per potencial predit (Temporada 2015-2016).

Rànquing	Nom del Jugador	Edat	Potencial Predit
1	Rubén Duarte Sánchez	20	101,56
2	Jonathan Castro Otto	21	101,09
3	Aymeric Laporte	21	100,46
4	João Pedro Cavaco Cancelo	21	98,39
5	Eric Bertrand Bailly	21	96,74
6	José María Giménez de Vargas	20	95,59
7	Emiliano Buendía	19	95,16
8	Emiliano Daniel Velázquez Maldonado	21	94,58
9	José Luis Gayà Peña	20	93,11
10	Federico Ricca Rostagnol	21	88,41

Taula 15. Comparativa de rànquings de migcampistes (Model 1 vs. Model 2).

Rànquing	Model 1	Potencial	Model 2	Potencial
1	Andrés Iniesta Luján (2005/06)	156,22	Jorge Resurrección Merodio (2013/14)	125,90
2	Andrés Iniesta Luján (2004/05)	155,71	Jorge Resurrección Merodio (2012/13)	125,90
3	Thiago Alcântara do Nascimento (2011/12)	146,66	Jorge Resurrección Merodio (2011/12)	120,70
4	Sergio Busquets i Burgos (2009/10)	146,03	Andrés Iniesta Luján (2005/06)	117,46
5	Thiago Alcântara do Nascimento (2012/13)	144,80	Sergio Busquets i Burgos (2008/09)	116,83
6	Thiago Alcântara do Nascimento (2010/11)	142,60	Sergio Busquets i Burgos (2009/10)	116,40
7	Sergio Busquets i Burgos (2008/09)	141,42	Thiago Alcântara do Nascimento (2012/13)	116,33
8	Ignacio Camacho Barnola (2011/12)	140,30	Andrés Iniesta Luján (2004/05)	116,05
9	José Luis García del Pozo (2012/13)	138,53	Thiago Alcântara do Nascimento (2012/13)	115,69

10	Gabriel Fernández Arenas (2004/05)	135,63	Thiago Alcântara do Nascimento (2012/13)	114,42

Taula 16. Comparativa de rànquings d'atacants (Model 1 vs. Model 2).

Rànquing	**Model 1 (Validació Creuada)**	**Potencial**	**Model 2**	**Potencial**
1	Neymar da Silva Santos Junior (2013/14)	105,49	Lionel Andrés Messi Cuccittini (2005/06)	110,60
2	Antoine Griezmann (2012/13)	97,27	Lionel Andrés Messi Cuccittini (2006/07)	110,58
3	Antoine Griezmann (2011/12)	92,86	Lionel Andrés Messi Cuccittini (2007/08)	110,55
4	Lionel Andrés Messi Cuccittini (2008/09)	90,34	Lionel Andrés Messi Cuccittini (2008/09)	110,55
5	Antoine Griezmann (2010/11)	86,80	Lionel Andrés Messi Cuccittini (2004/05)	99,17
6	Bojan Krkíc Pérez (2007/08)	81,76	Neymar da Silva Santos Junior (2013/14)	90,97
7	Bojan Krkíc Pérez (2008/09)	81,72	Ousmane Dembélé (2018/19)	81,26
8	Lionel Andrés Messi Cuccittini (2006/07)	80,54	Isaac Ajayi Success (2015/16)	78,61
9	Bojan Krkíc Pérez (2010/11)	78,05	Ousmane Dembélé (2017/18)	73,07
10	Bojan Krkíc Pérez (2009/10)	77,65	Antoine Griezmann (2012/13)	68,65

Taula 17. Comparativa de rànquings de defensors (Model 1 vs. Model 2).

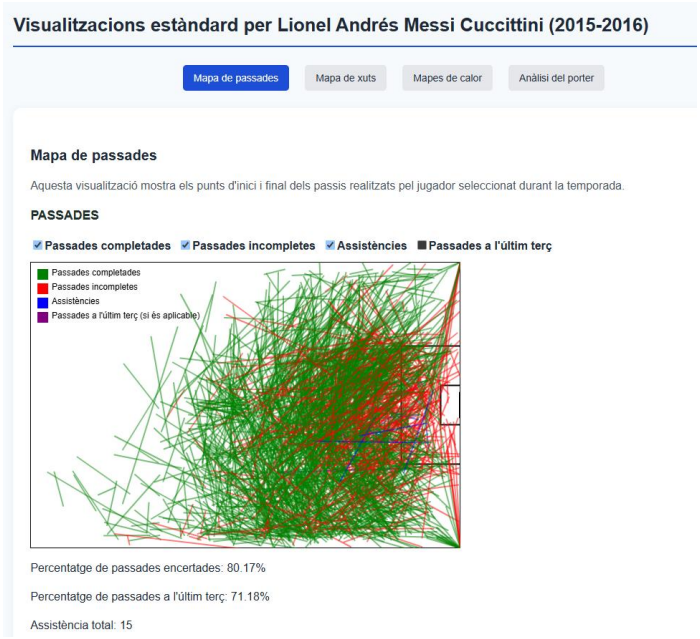
Rànquing	**Model 1 (Validació Creuada)**	**Potencial**	**Model 2**	**Potencial**
1	Jordi Alba Ramos (2010/11)	143,99	Gerard Piqué Bernabéu (2008/09)	114,46
2	Gerard Piqué Bernabéu (2008/09)	141,72	Jordi Alba Ramos (2010/11)	114,18
3	Sergio Ramos García (2007/08)	141,13	Sergio Ramos García (2006/07)	113,89
4	Marcelo Vieira da Silva Júnior (2008/09)	137,69	Víctor Ruíz Torre (2010/11)	110,91

5	Sergio Ramos García (2006/07)	135,92	Igor Zubeldia Elorza (2018/19)	109,43
6	Raphaël Varane (2013/14)	132,56	Marcelo Vieira da Silva Júnior (2008/09)	104,22
7	Aymeric Laporte (2015/16)	127,42	Aymeric Laporte (2014/15)	103,40
8	Marcelo Vieira da Silva Júnior (2008/09)	126,64	Jonathan Castro Otto (2014/15)	101,69
9	Aymeric Laporte (2014/15)	125,04	Rubén Duarte Sánchez (2015/16)	101,56
10	Raphaël Varane (2013/14)	124,63	Jonathan Castro Otto (2015/16)	101,09

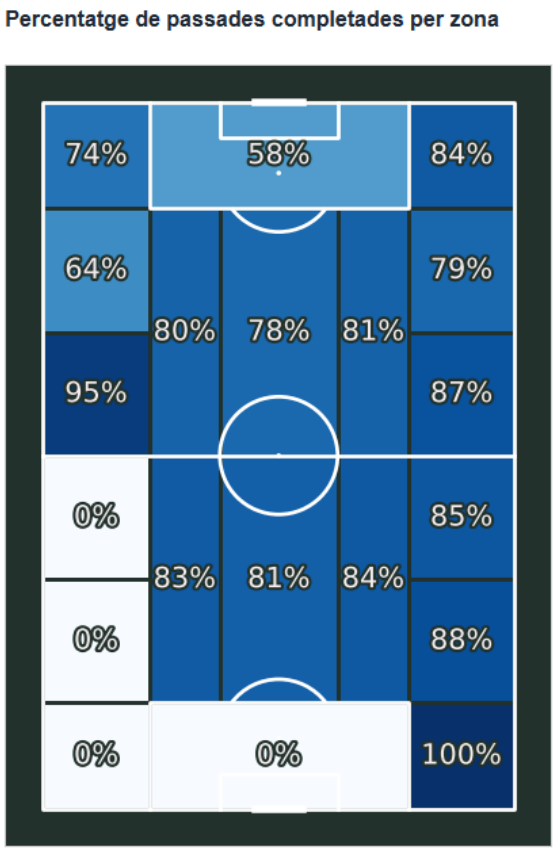
Apèndix E: Il·lustracions de la interfície d'usuari

Il·lustració 2. Pantalla de navegació principal d'"Estrelles del Futur".

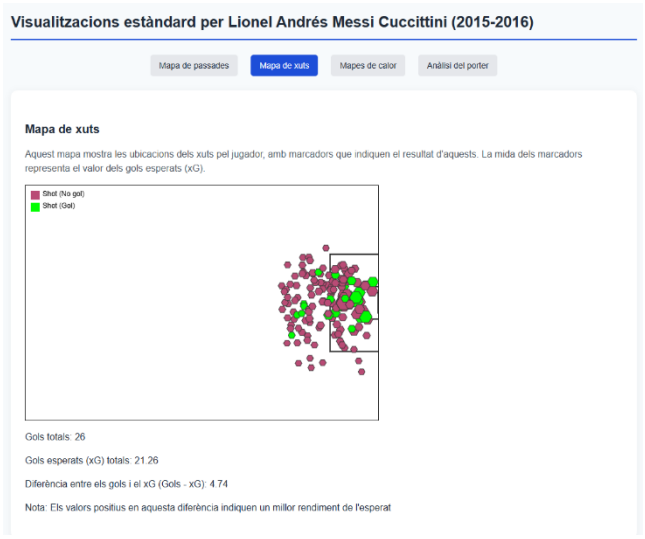
Il·lustració 3. Mapa de passades de Lionel Messi (2015-2016).



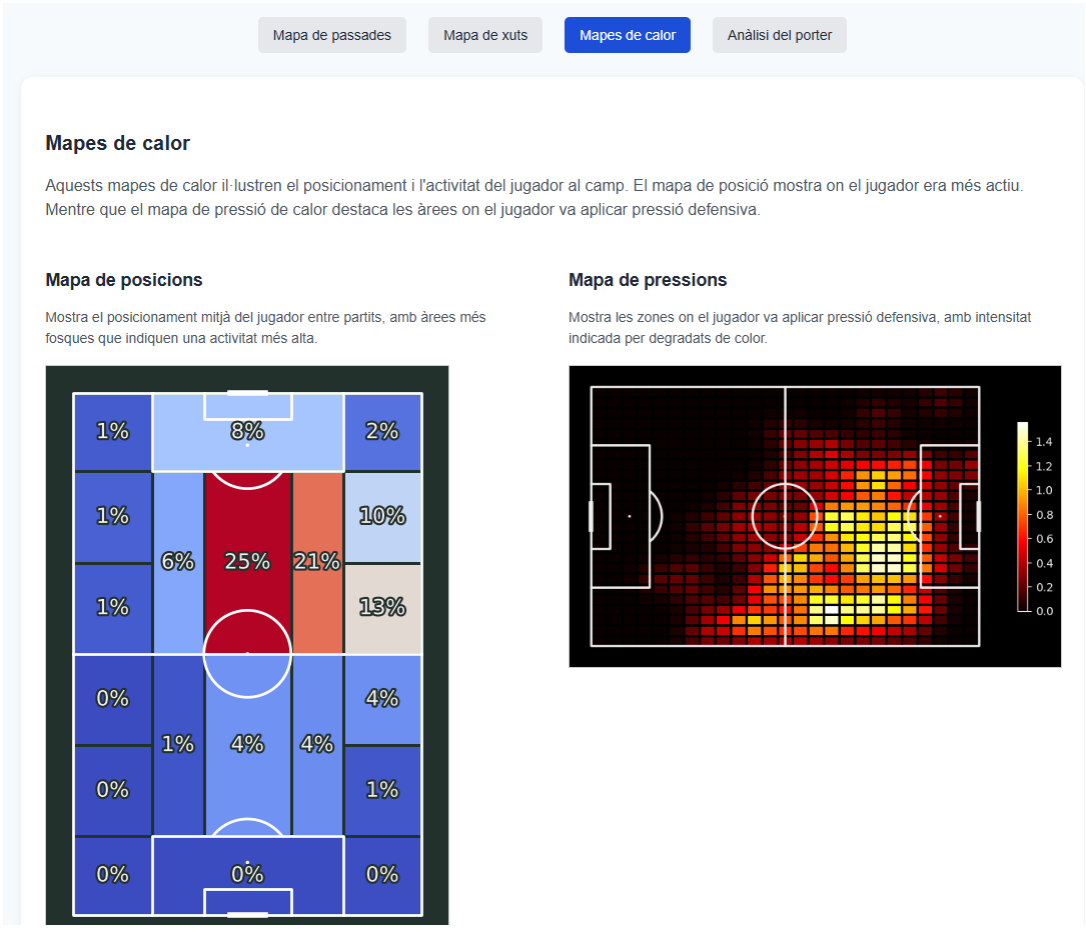
Il·lustració 4. Percentatge d'encert en passades per zona.



Il·lustració 5. Mapa de xuts interactiu, on la mida de l'hexàgon representa el valor xG.



Il·lustració 6. Mapes de calor de posició (esquerra) i pressió (dreta).



Il·lustració 7. Anàlisi d'accions del porter Vicente Guaita (2015-2016).

Aquesta secció proporciona una anàlisi detallada del rendiment del porter, incloent mètriques de tir, estadístiques de pas i altres accions clau

Anàlisi del porter: Vicente Guaita Panadero (2015-2016)

Rendiment d'aturades

Tirs a porteria rebuts: 157

Aturades: 99

Gols rebuts: 58

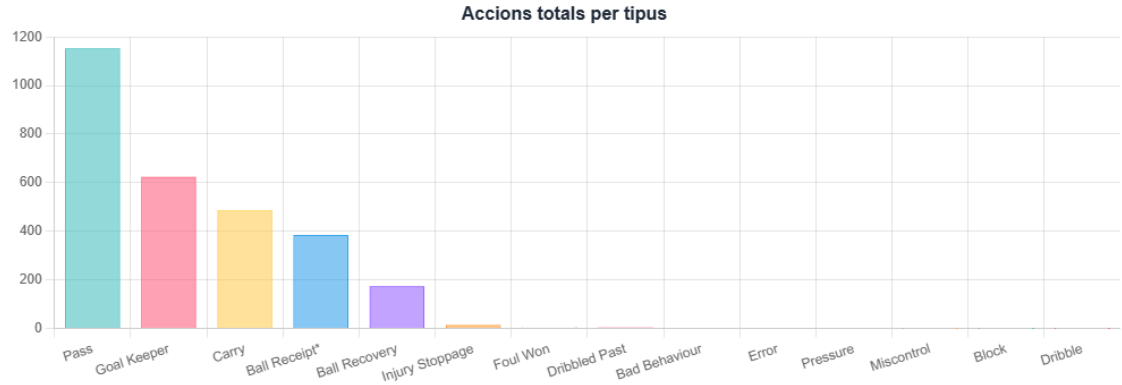
Percentatge d'aturades: 63.06%

Resum de passades

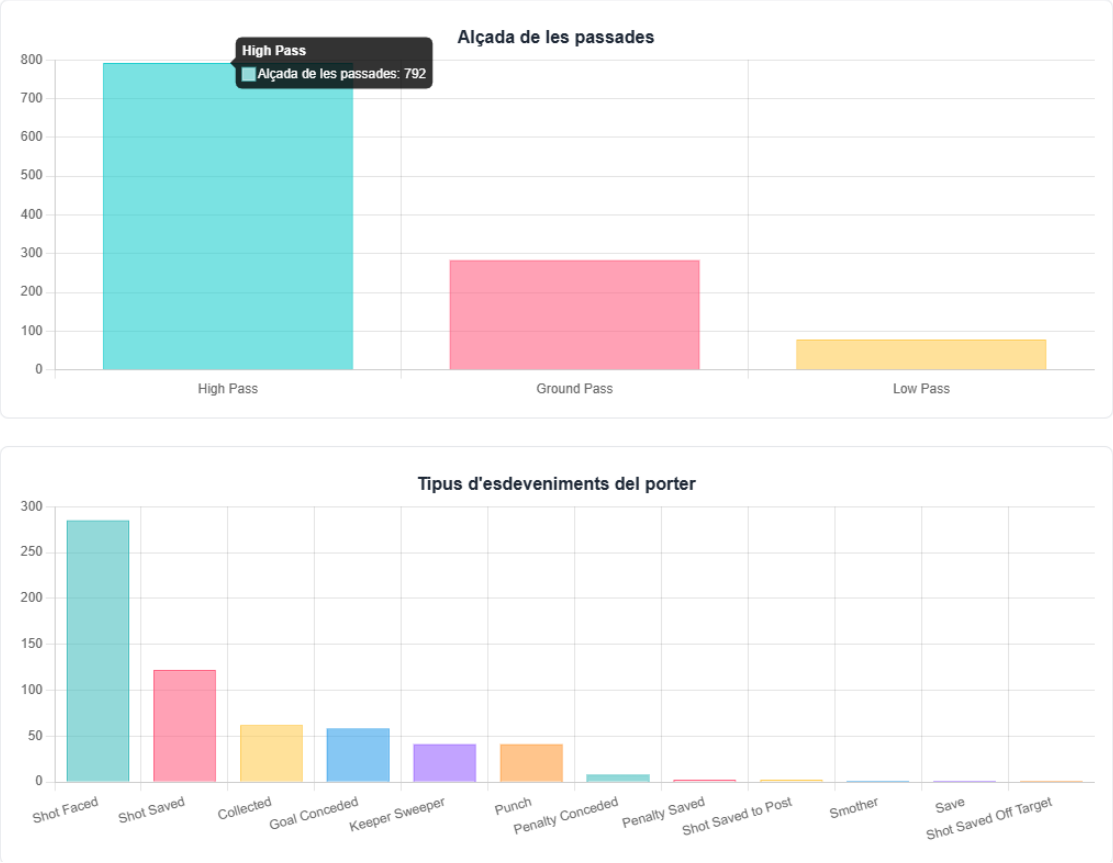
Passades totals: 1153

Passades completades: 609

Precisió de passada: 52.82%

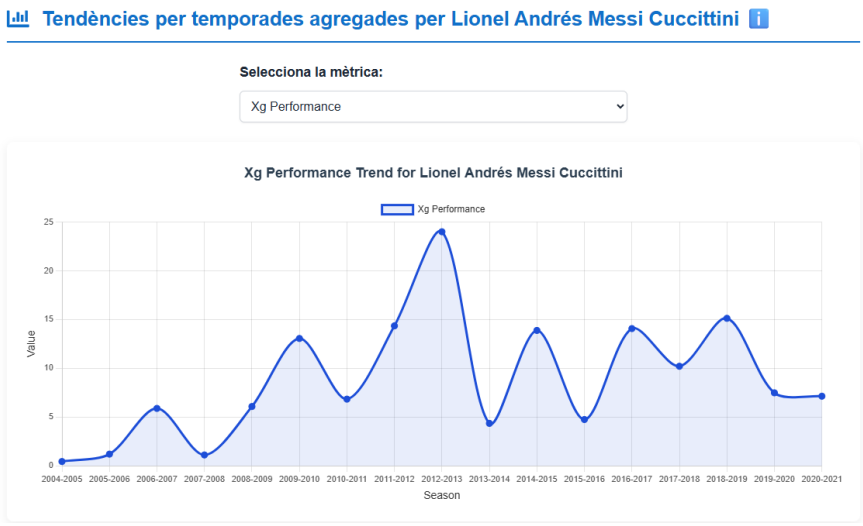


Il·lustració 8. Tipus d'esdeveniments i alçada de les passades del porter.



Il·lustració 9. Visualització d'una mètrica agregada individual (Total de Centrades).

Il·lustració 10. Gràfic de tendència de la mètrica al llarg de les temporades.



Il·lustració 11. Targeta de resultat de la predicció de potencial a la pàgina de scouting

Il·lustració 12. Interfície per a la Construcció de Models de Potencial Personalitzats.

Apèndix F: Fragments de codi clau

Aquest apèndix presenta dos fragments de codi essencials de l'script trainer_v2_15_16.py, que il·lustren la lògica central desenvolupada per a la generació de la variable objectiu i la construcció de les característiques del model d'aprenentatge automàtic.

F.1 Funció de generació de la mètrica heurística de rendiment

La funció generate_potential_target és el nucli de la creació de la variable objectiu. Itera sobre cada grup posicional, calcula una puntuació de rendiment composta (raw_composite_score) per a cada jugador-temporada basant-se en els pesos derivats de la correlació, i finalment normalitza aquestes puntuacions a una escala de 0 a 200.

```
def generate_potential_target(df_all_player_seasons, derived_kpi_weights_config):
    df = df_all_player_seasons.copy()
    df['raw_composite_score'] = 0.0
    df['potential_target'] = 0.0
    for position_group, weights in derived_kpi_weights_config.items():
        pos_mask = (df['general_position_identifier'] == position_group)
        if pos_mask.sum() == 0:
            continue

        # Normalitzar els pesos per a la posició actual
        current_total_weight = sum(w for w in weights.values() if isinstance(w, (int, float)))
        if current_total_weight == 0:
            current_total_weight = 1.0

        position_composite_score_series = pd.Series(0.0, index=df[pos_mask].index)
        df_pos_group_subset = df[pos_mask].copy()

        for kpi_col_name, weight_value in weights.items():
            actual_weight = safe_division(weight_value, current_total_weight)
            # Normalitzar cada KPI a una escala de 0 a 1 dins del seu grup
            kpi_data = df_pos_group_subset[kpi_col_name].copy()
            min_val, max_val = kpi_data.min(), kpi_data.max()

            norm_value_series = pd.Series(0.0, index=kpi_data.index)
            if max_val > min_val:
                norm_value_series = (kpi_data - min_val) / (max_val - min_val)

            # Penalitzar mètriques p90 si es basen en pocs minuts
            if "_p90" in kpi_col_name:
                low_mins_mask = df_pos_group_subset['num_90s_played'] <
MIN_90S_PLAYED_FOR_P90_STATS
                norm_value_series[low_mins_mask] = 0.0

            position_composite_score_series += norm_value_series.fillna(0.0) * actual_weight

        df.loc[pos_mask, 'raw_composite_score'] = position_composite_score_series

        # Escalar la puntuació final a un rang de 0 a 200 per posició
        min_raw_score = df.loc[pos_mask, 'raw_composite_score'].min()
        max_raw_score = df.loc[pos_mask, 'raw_composite_score'].max()
        if max_raw_score > min_raw_score:
```

```
            scaled_potential = ((df.loc[pos_mask, 'raw_composite_score'] - min_raw_score) /
(max_raw_score - min_raw_score)) * 200.0
            df.loc[pos_mask, 'potential_target'] = scaled_potential.clip(0, 200).round(2)

    return df[['player_id_identifier', 'target_season_identifier', 'potential_target']]
```

F.2 Funció de construcció de característiques per al model

La funció trainer_construct_ml_features_for_player_season és responsable del *feature engineering*. Per a una temporada Sub-21 donada, combina les mètriques d'aquella temporada (current_...) amb agregats i tendències de les temporades anteriors del jugador (hist_..., growth_...) per crear el vector d'entrada final que alimentarà el model XGBoost.

```
def trainer_construct_ml_features_for_player_season(
    current_season_base_features_row: pd.Series,
    historical_base_features_df: pd.DataFrame,
    all_base_metric_names: list
):
    instance_ml_features = pd.Series(dtype='float64')

    # 1. Característiques de la temporada actual
    for base_fname in all_base_metric_names:
        instance_ml_features[f'current_{base_fname}'] =
current_season_base_features_row.get(base_fname, 0.0)

    # 2. Característiques d'interacció i polinòmiques (exemple simplificat)
    g_p90s = instance_ml_features.get('current_goals_p90_sqrt_', 0.0)
    cr_kpi = instance_ml_features.get('current_conversion_rate_excl_xg_kpi', 0.0)
    instance_ml_features['current_inter_goals_x_conversion'] = g_p90s * cr_kpi

    # 3. Característiques històriques i d'evolució
    instance_ml_features['num_hist_seasons'] = 0.0
    if historical_base_features_df is not None and not historical_base_features_df.empty:
        df_history = historical_base_features_df.copy()
        instance_ml_features['num_hist_seasons'] = float(len(df_history))

        for col_name in all_base_metric_names:
            if col_name in df_history.columns:
                hist_values = pd.to_numeric(df_history[col_name], errors='coerce').fillna(0.0)
                if hist_values.empty: continue

                # Agregats històrics
                instance_ml_features[f'hist_avg_{col_name}'] = hist_values.mean()
                instance_ml_features[f'hist_max_{col_name}'] = hist_values.max()

                # Indicadors d'evolució
                last_season_val = hist_values.iloc[-1]
                current_val = instance_ml_features.get(f'current_{col_name}', 0.0)
                instance_ml_features[f'growth_{col_name}'] = current_val - last_season_val

                # Càlcul de tendència (pendent de regressió lineal)
                if len(df_history) >= 2:
                    x_h_valid = pd.to_numeric(df_history['season_numeric'], errors='coerce')
                    y_h_valid = hist_values
```

```
                slope, _ = np.polyfit(x_h_valid, y_h_valid, 1)
                instance_ml_features[f'hist_trend_{col_name}'] = slope

    return instance_ml_features
```