UAB
Universitat Autònoma de Barcelona

Dipòsit digital de documents de la UAB

This is the **published version** of the bachelor thesis:

Martínez Batlle, Guillermo. *Models Predictius per a l'Anàlisi Esportiva i Generació d'Apostes basats en Estadística i Machine Learning*. Treball de Final de Grau (Universitat Autònoma de Barcelona), 2026 (Grau en Enginyeria de Dades)

This version is available at https://ddd.uab.cat/record/337326

under the terms of the BY-SA license.

Models Predictius per a l'Anàlisi Esportiva i Generació d'Apostes basats en Estadística i Machine Learning: BETIQ

Guillermo Martínez Batlle

Resum— La creixent accessibilitat de les apostes esportives en línia ha amplificat la desigualtat informacional entre els apostadors recreatius i els operadors de mercat, que disposen de models de preus sofisticats. En aquest context, les eines basades en dades poden contribuir a reduir aquesta asimetria. BETIQ és una plataforma d'analítica de futbol i intel·ligència per a apostes que ingereix dades de cinc lligues europees durant sis temporades (2020/21–2025/26), aproximadament 27.900 partits, a través de l'API de SportMonks. El sistema transforma les dades en brut en taules analítiques i entrena deu models d'aprenentatge automàtic: tres models d'equip (1X2, Over/Under 2,5 gols i BTTS) basats en XGBoost amb calibració isotònica per lliga, i set models de regressió per a estadístiques individuals de jugadors. El model 1X2 v2 assoleix una precisió del 52,9% i un log-loss calibrat de 0,942 al conjunt de validació. El ROI simulat roman negatiu ($-3{,}4\%$), indicant que el model no supera el marge del *bookmaker*, però l'arquitectura modular i la metodologia experimental establertes constitueixen una base sòlida per a futures millores. Totes les prediccions s'exposen a través d'una plataforma web construïda amb Next.js que inclou mercats, *value picks*, seguiment d'apostes i suport bilingüe.

Paraules clau— apostes esportives, aprenentatge automàtic, XGBoost, calibració probabilística, enginyeria de dades, futbol, predicció de resultats.

Abstract— The growing accessibility of online sports betting has widened the informational gap between recreational bettors and market operators, who rely on sophisticated pricing models. Data-driven tools can help reduce this asymmetry. BETIQ is a football analytics and betting intelligence platform that ingests data from five European leagues over six seasons (2020/21–2025/26), approximately 27,900 matches, via the SportMonks API. The system transforms raw data into analytical tables and trains ten machine learning models: three team-level models (1X2, Over/Under 2.5 goals, and BTTS) based on XGBoost with per-league isotonic calibration, and seven regression models for individual player statistics. The 1X2 v2 model achieves 52.9% accuracy and a calibrated log-loss of 0.942 on the validation set. Simulated ROI remains negative (-3.4%), indicating the model does not yet overcome the bookmaker's margin, but the modular architecture and experimental methodology provide a solid foundation for future improvements. All predictions are served through a Next.js web platform featuring markets, value picks, bet tracking, and bilingual support.

Keywords— sports betting, machine learning, XGBoost, probability calibration, data engineering, football, match prediction.

1 Introducció

Les apostes esportives constitueixen una indústria global que va superar els 90.000 milions de dòlars en ingressos l'any 2023, amb el futbol com a modalitat dominant [1]. Malgrat l'escala d'aquest mercat, la majoria dels apostadors recreatius operen amb un desavantatge

• E-mail de contacte: guillermo.martinezb@autonoma.cat
• Treball tutoritzat per: Natàlia Blasco Andreo (Departament d'Enginyeria de la Informació i de les Comunicacions)
• Curs 2025/26

informacional significatiu respecte als *bookmakers*. Un *bookmaker* (casa d'apostes) és una entitat que estableix i ofereix quotes —preus que reflecteixen la probabilitat implícita d'un resultat— sobre esdeveniments esportius, actuant com a contraparta de l'apostador i incorporant un marge de benefici (*vig* o *sobreround*) en cadascuna de les seves línies. Els *bookmakers* professionals disposen d'equips d'analistes quantitatius i models de preus sofisticats.

La tensió central és d'asimetria d'informació: el *bookmaker* fixa una quota que incorpora tant la seva estimació de probabilitat com un marge (el *sobreround* o *vig*). Per obtenir un retorn esperat positiu, l'apostador no tan sols ha de predir millor que l'atzar, sinó que ha de superar la probabilitat implícita del *bookmaker* de manera consistent. La hipòtesi del *Closing Line Value* (CLV) [2] estableix que les quotes de tancament constitueixen la millor estimació disponible de la probabilitat real, i que superar-les de manera consistent és el veritable indicador d'avantatge a llarg termini.

BETIQ es concep com una plataforma que tanca aquesta bretxa mitjançant tres eixos: (i) agregar i estructurar dades històriques de partits a nivell d'equip i jugador, (ii) entrenar models probabilístics calibrats per a múltiples mercats d'apostes, i (iii) proporcionar eines de seguiment i visualització per mesurar l'avantatge real. El sistema cobreix cinc lligues europees de primera divisió (Premier League, La Liga, Bundesliga, Serie A i Ligue 1) durant sis temporades (2020/21–2025/26), amb un volum total d'aproximadament 27.900 partits.

L'objectiu final no és maximitzar la precisió de classificació —un model que prediu sempre victòria local assoliria el 44% sense aprendre res— sinó obtenir **probabilitats ben calibrades** que permetin identificar apostes amb valor esperat positiu [12].

La resta del document s'organitza com segueix: la Secció 2 revisa l'estat de l'art, la Secció 3 presenta els objectius, la Secció 4 descriu la metodologia i la Secció 5 la proposta de solució. Les Seccions 6–10 cobreixen el desenvolupament tècnic. La Secció 11 presenta els resultats, la Secció 12 les conclusions i la Secció 13 les línies de treball futur.

2 Estat de l'art

La predicció de resultats futbolístics amb mètodes estadístics té una rica trajectòria acadèmica. El model pioner de Maher [3] va proposar modelitzar els gols de cada equip com a variables de Poisson independents, parametritzades per la força atacant i defensiva de cada equip. Dixon i Coles [4] van estendre aquest enfocament introduint una correcció per a resultats de marcadors baixos (0–0, 1–0, 0–1, 1–1) i un factor de depreciació temporal, demostrant ineficiències explotables al mercat d'apostes britànic. Rue i Salvesen [13] van adoptar un marc bayesià amb paràmetres de força que varien en el temps, permetent l'actualització contínua a mesura que es disputen nous partits.

Des d'una perspectiva de regressió clàssica, Goddard [14] va demostrar que models de regressió ordenada amb variables com la forma recent, la classificació i el factor camp aconseguien rendiments predictius comparables als enfocaments de Poisson, però amb major interpretabilitat.

L'aplicació de tècniques d'aprenentatge automàtic ha guanyat força en l'última dècada. Hubáček et al. [5] van aplicar *gradient boosted trees* (XGBoost, LightGBM) sobre conjunts de característiques tabulars i van demostrar que aquests models superaven les xarxes neuronals en la predicció de resultats futbolístics. Grinsztajn et al. [6] van confirmar aquest resultat de manera més general: les característiques tabulars ben construïdes superen els models arquitectònicament més complexos en dades tabulars, en part perquè els resultats futbolístics presenten un elevat component aleatori que limita el sostre predictiu de qualsevol model.

Pel que fa als sistemes de rànquing, Hvattum i Arntzen [7] van demostrar que els rànquings Elo [15] capturen el senyal de qualitat d'equip que transcendeix una única temporada, proporcionant una línia base competitiva per a la predicció de resultats. El sistema Elo, originalment dissenyat per als escacs, assigna una puntuació contínua a cada equip que s'actualitza després de cada partit en funció del resultat i la diferència de puntuacions prèvia.

Un aspecte transversal de la literatura és la importància de la **calibració probabilística**: Constantinou [12] va demostrar que un model amb probabilitats ben calibrades és més útil per a l'apostador que un model amb alta precisió però probabilitats desviades, ja que la identificació de valor requereix estimacions fiables de la probabilitat real. Platt [10] va proposar el calibrador sigmoidal, i més recentment la calibració isotònica s'ha mostrat efectiva en contextos amb suficients dades de validació.

BETIQ es diferencia dels treballs anteriors en la **integració vertical** de tot el procés: des de la ingestió de dades en brut fins a la plataforma web d'usuari, passant per l'enginyeria de característiques, l'entrenament de models i el desplegament en temps real. A més, cobreix no només mercats de resultat d'equip sinó també set mercats d'estadístiques individuals de jugadors, un àmbit menys explorat a la literatura.

3 Objectius

1. Establir un pipeline d'ingestió robust que reculli dades de cinc lligues europees durant sis temporades a través de l'API de SportMonks.
2. Dissenyar una arquitectura de base de dades per capes que separi dades en brut i taules analítiques derivades.
3. Construir un model predictiu 1X2 (resultat del partit) amb XGBoost i calibració probabilística, i millorar-lo amb característiques Elo i calibració per lliga.
4. Estendre el sistema a mercats binaris (Over/Under 2,5, BTTS) i set mercats de jugadors individuals.
5. Desenvolupar una plataforma web funcional que exposi les prediccions als usuaris amb autenticació, mercats i seguiment d'apostes.
6. Desplegar un pipeline en temps real que generi prediccions automàtiques per a tots els partits programats.

4 Metodologia

El projecte segueix una **metodologia híbrida** que combina la metodologia en cascada [11] per a les fases estructurades (ingestió, disseny de BD, entrenament de models) amb principis àgils per al desenvolupament iteratiu del producte web.

La metodologia en cascada garanteix que cada fase es completi abans de passar a la següent, assegurant que les dades estiguin ben estructurades i preparades per a l'entrenament. Els principis àgils permeten iteracions curtes en la part de visualització i producte, adaptant-se als canvis i necessitats detectades durant el procés.

4.1 Planificació temporal

El treball s'ha organitzat en quatre fases principals al llarg del període febrer–juny de 2026:

- **Fase 0** (febrer–març): Fonamentació de dades — pipeline d'ingestió, disseny de la BD i construcció de les taules analítiques.
- **Fase 1** (març–abril): Primer model predictiu 1X2 — disseny de característiques, entrenament, experiments de calibració i simulació de ROI.
- **Fase 2** (maig–juny): Mercats addicionals (O/U 2,5, BTTS, 7 mercats de jugadors), producte web i pipeline en temps real.
- **Fase 3** (juny): Documentació final i preparació de la defensa.

4.2 Marc experimental

Per al desenvolupament dels models predictius, s'ha establert un **marc experimental estructurat** que garanteix la reproductibilitat: cada iteració es documenta amb (i) una hipòtesi explícita, (ii) un canvi controlat respecte a la versió anterior, i (iii) criteris d'èxit predefinits. Només si es compleixen tots els criteris, el canvi s'adopta. Aquest procés permet iterar de manera científicament rigorosa i evita decisions basades en intuïcions no validades.

4.3 Mètriques d'avaluació

L'avaluació dels models es basa en tres mètriques complementàries:

- **Log-loss:** $-\frac{1}{N}\sum_i \log \hat{p}_{y_i}$. Mesura la qualitat de les probabilitats estimades; valors inferiors indiquen millors probabilitats.
- **Brier score multiclasse:** $\frac{1}{N}\sum_i \sum_c (\hat{p}_{ic} - \mathbf{1}[y_i{=}c])^2$. Error quadràtic mitjà sobre les probabilitats estimades.
- **ROI simulat:** simulació d'apostes de *stake* fix al resultat amb major avantatge respecte a les quotes del *bookmaker*, en diferents llindars mínims d'avantatge.

5 Proposta de la Solució

Per donar resposta a la necessitat de reduir l'asimetria d'informació als mercats d'apostes esportives, es proposa la implementació de BETIQ: una plataforma integrada que combina enginyeria de dades, models predictius i un producte web en un únic sistema.

L'arquitectura del sistema s'organitza en quatre capes funcionals:

1. **Ingestió:** scripts Python que consumen l'API de SportMonks, paginen els resultats i escriuen a la base de dades amb *upsert* idempotent. El pipeline gestiona automàticament els límits de velocitat i les errors de xarxa.
2. **Emmagatzematge:** PostgreSQL allotjat a Supabase, que actua com a font única de veritat per a totes les dades en brut i derivades. L'API REST autogenerada de Supabase facilita l'accés des de múltiples clients.
3. **Analítica:** pipelines Python que construeixen les taules d'entrenament (`match_analytics`, `player_match_analytics`) a partir de les dades en brut, i models XGBoost que generen prediccions probabilístiques per a deu mercats d'apostes.
4. **Aplicació:** plataforma web Next.js per a l'usuari final (mercats, *value picks*, seguiment d'apostes) i tauler Streamlit per a exploració interna.

Aquesta arquitectura per capes permet que cada component evolucioni de manera independent: es pot millorar un model sense modificar la ingestió, o canviar la presentació web sense afectar els càlculs analítics.

6 Fonamentació de Dades

6.1 Font de dades i ingestió

SportMonks és un proveïdor comercial d'estadístiques de futbol que ofereix accés via REST+JSON a calendaris, resultats, estadístiques, esdeveniments, alineacions, àrbitres i quotes pre-partit. El pipeline d'ingestió, implementat en Python amb pandas, consumeix l'API amb gestió automàtica de límits de velocitat (3.000 peticions/hora) i *retry* exponencial.

El projecte cobreix **cinc lligues** (Premier League, La Liga, Bundesliga, Serie A, Ligue 1) $\times$ **sis temporades** (2020/21–2025/26) = 30 lots, amb un total aproximat de 27.900 partits disputats. Totes les escriptures utilitzen estratègia *upsert* sobre la clau primària, garantint idempotència.

Un aspecte clau del disseny és l'**endpoint de detall de partit**: retorna el *payload* complet d'un encontre (estadístiques, esdeveniments, alineació, àrbitre, quotes) en una única petició, reduint significativament les crides necessàries per temporada (≈380 per lliga). La gestió del límit de velocitat empra retards fixos (1,0 s entre crides) i un *backoff* exponencial en cas de respostes HTTP 429 (65s, 120s, 240s).

El pipeline d'ingestió segueix una estructura modular: per a cada entitat existeix un mòdul de recuperació (`fetch_X`) i un de transformació (`transform_X`), orquestrats per un mòdul de pipeline (`run_X`) que gestiona la lectura i escriptura a Supabase. Aquesta separació facilita el manteniment i l'extensió a noves fonts de dades.

6.2 Disseny de la base de dades

L'esquema, allotjat a PostgreSQL (Supabase, regió eu-west-3), s'organitza en una **arquitectura per capes** que separa les dades en brut de les taules analítiques derivades. Aquesta separació permet reconstruir les taules d'entrenament quan canvia el pipeline de característiques sense tornar a consultar l'API. A més, les taules en brut serveixen com a font única de veritat per a qualsevol anàlisi futura.

TAULA 1: TAULES PRINCIPALS DE LA BASE DE DADES.

Taula	Files	Descripció
fixtures	27,9K	Partits
fixtures_stats	795K	Estad. d'equip
fix_player_st.	5,9M	Estad. de jugadors
odds_prematch	2,4M	Quotes bet365
match_analytics	24,6K	Analítica d'equip
player_m._an.	429K	Analítica de jug.

La taula `match_analytics` adopta **dues files per partit** (perspectiva local i visitant), amb clau primària composta `(fixture_id, team_id)`. Inclou 104 columnes: identitat, estadístiques en brut (47), estadístiques per part (16) i enriquiment (24). La taula `player_match_analytics` conté 71 estadístiques individuals per jugador i per partit.

Decisions de disseny clau de la capa analítica:

- **Indicador `has_stats`:** quan `TRUE`, les estadístiques de recompte absents s'omplen amb zero; quan `FALSE`, es deixen com a `NULL` per no introduir valors artificials.
- **Percentatges nul·lables:** `ball_possession` i `pass_accuracy_pct` mai s'omplen amb zero, ja que un 0% de possessió és físicament impossible.
- **Context entre temporades:** les característiques de forma es calculen sobre tot l'historial disponible, no només la temporada actual, per donar context als partits d'inici de temporada.
- **Classificació computada:** la classificació de lliga es calcula dins del pipeline a partir dels resultats dels partits, no de l'API, garantint consistència.

Les taules addicionals (`fixture_lineups`, `fixture_events`, `fixture_referees`) proporcionen el context necessari per a l'enriquiment i permeten anàlisis complementàries com el perfil d'àrbitres i l'historial de substitucions.

7 ENGINYERIA DE CARACTERÍSTIQUES

7.1 Principi de no-fuga temporal

L'error més freqüent en la construcció de models de predicció esportiva és incloure informació post-partit en les variables d'entrenament [9]. Per evitar-ho, totes les característiques de forma es calculen desplacant un pas temporal (`shift(1)`) *abans* de calcular les finestres mòbils:

$$\text{forma}_{n,t} = \frac{1}{\min(n, t-1)} \sum_{k=1}^{\min(n,t-1)} r_{t-k} \qquad (1)$$

7.2 Grups de característiques

El model 1X2 v1 empra 53 característiques pre-partit distribuides en set grups (Taula 2). El model v2 afegeix quatre característiques Elo per un total de 57.

La importància de cada característica es mesura mitjançant el *gain* d'XGBoost: la reducció mitjana de la funció de pèrdua (*log-loss*) quan la característica s'utilitza com a variable de partició en els arbres. Formalment, per a una característica j, la importància és $I_j = \sum_{t \in T} \sum_{s \in S_j^t} \Delta\mathcal{L}(s)$, on T és el conjunt d'arbres, S_j^t les particions que utilitzen j a l'arbre t i $\Delta\mathcal{L}(s)$ la reducció de pèrdua associada. Les característiques de major importància al model final són `goal_diff_diff` (diferència en la diferència de gols acumulada, importància = 0,106), seguida de `points_diff` (0,062) i `position_diff` (0,035). La qualitat de temporada acumulada domina per davant de la forma recent, un resultat consistent amb la literatura [4, 5].

TAULA 2: GRUPS DE CARACTERÍSTIQUES (53+4 ELO).

Grup	Var.	Senyal capturat
Dies de descans	3	Fatiga
Classificació	6	Qualitat temporada
Forma (5 partits)	16	Momentum
Forma (10 partits)	6	Tendència mig termini
Forma en seu (5 p.)	7	Rendiment específic
H2H (5 partits)	4	Patrons directes
Perfil àrbitre	4	Biaix mesurable
Elo (v2)	4	Qualitat entre temp.
Identif. lliga	1	Estil de lliga
Total v2	**57**	

7.3 Gestió de valors nuls

Les dades H2H no estàn disponibles per a aproximadament el 16% dels enfrontaments (parelles d'equips que no s'han trobat en les sis temporades cobertes). XGBoost gestiona nativament els valors nuls [8], de manera que no cal imputació per al model principal. La regressió logística de referència empra imputació per la mediana. Les característiques d'àrbitre tenen una cobertura del 97,8%.

7.4 Sistema Elo

Les característiques de forma (mitjanes mòbils) es reinicien implícitament a l'inici de cada temporada per manca de dades prèvies. Això crea un *arrencada en fred* on els partits de les primeres jornades no disposen de context històric. Per resoldre-ho, el model v2 incorpora un sistema de rànquing Elo [7, 15] que manté una puntuació contínua per equip al llarg de totes les temporades, capturant així la qualitat relativa entre temporades.

El sistema Elo, originalment desenvolupat per Arpad Elo per classificar jugadors d'escacs [15], assigna a cada equip una puntuació que s'actualitza després de cada partit. Si un equip guanya a un rival més fort (puntuació Elo superior), guanya més punts que si guanya a un rival més feble. El sistema processa cronològicament 20.900 registres amb paràmetres $K = 20$ (velocitat d'adaptació), avantatge de casa $H = 100$ (punts addicionals per jugar a casa) i puntuació inicial $E_0 = 1500$. La probabilitat esperada de victòria local es calcula com:

$$E_H = \frac{1}{1 + 10^{(E_{\text{away}} - (E_{\text{home}} + H))/400}} \qquad (2)$$

8 MODELS PREDICTIUS

8.1 Model 1X2: resultat del partit

El mercat 1X2 és el de major liquiditat: per a cada partit, el *bookmaker* ofereix quotes per a la victòria local (1), l'empat (X) i la victòria visitant (2). Formalment, s'aprèn una funció $f(\mathbf{x}_i) = \hat{\mathbf{p}}_i = (\hat{p}_H, \hat{p}_D, \hat{p}_A) \in \Delta^2$ on $\mathbf{x}_i \in \mathbb{R}^{57}$ és el vector de característiques pre-partit.

El model principal és un **XGBoost** [8] amb els hiperparàmetres de la Taula 3. S'ha escollit XGBoost per la seva eficàcia documentada en dades tabulars [6], la gestió nativa de valors nuls (crítica pel 16% de parelles sense històric H2H) i la interpretabilitat mitjançant valors SHAP.

TAULA 3: HIPERPARÀMETRES XGBOOST (TOTS ELS MERCATS D'EQUIP).

Paràmetre	Valor
`max_depth`	4
`learning_rate`	0,05
`subsample`	0,8
`colsample_bytree`	0,8
`min_child_weight`	10
`gamma`	1,0
`reg_lambda`	2,0
`early_stopping_rounds`	30

Com a model de referència s'empra una regressió logística multinomial (`SimpleImputer` → `StandardScaler` → `LogisticRegression`). Si XGBoost no supera aquest model, les característiques no aporten senyal predictiu rellevant.

La divisió de dades es realitza per **temporades completes senceres**, mai de manera aleatòria, per respectar l'estructura temporal [9]: entrenament 2020/21–2023/24 (7.236 partits, 5 lligues), validació 2024/25 (∼1.750 partits) i test 2025/26 (∼1.200 partits). Dividir a meitat de temporada filtraria l'estructura d'inici (manca de forma) i final de temporada (pressió per títol o descens), introduïnt biaixos difícils de detectar.

L'entrenament s'atura a la iteració 73 (de 500 màximes), indicant que el model assoleix el seu sostre de capacitat relativament aviat amb les característiques actuals.

8.1.1 Calibració probabilística

Les probabilitats brutes d'XGBoost tendeixen a ser sobreconfiades. La calibració post-entrenament corregeix aquesta desviació ajustant un model auxiliar sobre el conjunt de validació (`CalibratedClassifierCV(cv="prefit")`). Dos mètodes es comparen:

- El calibrador **sigmoidal** (Platt scaling) [10]: mètode paramètric que ajusta una funció logística $P(y{=}1|f) = 1/(1+\exp(Af+B))$ amb només dos paràmetres (A, B). És robust amb poques dades de calibració però assumeix una relació sigmoidal entre la sortida del model i la probabilitat real.
- El calibrador **isotònic**: mètode no paramètric que ajusta una funció escalonada monòtona creixent minimitzant l'error quadràtic mitjà. Ofereix major flexibilitat però disposa de més graus de llibertat, cosa que el fa susceptible al sobreajustament quan les dades de calibració són limitades.

L'Experiment 1 va demostrar que el calibrador isotònic va **empitjorar** el log-loss al conjunt de test (1,014 vs. 1,001 sense calibrar), confirmant la hipòtesi de sobreajustament: el rang de probabilitats d'empat és estret (0,20–0,35) i el calibrador isotònic disposa de massa graus de llibertat. El calibrador sigmoidal, amb només dos paràmetres, va generalitzar millor: log-loss 0,997, i la concentració en apostes d'empat es va reduir del 56% al 29%.

El model v2 combina les característiques Elo amb un **calibrador isotònic independent per lliga**: s'entrena un calibrador per a cadascuna de les cinc lligues sobre el conjunt de validació (2024/25). Aquesta estratègia captura les diferències estructurals entre lligues (p. ex., la Bundesliga té una taxa de gols per partit superior a La Liga) i redueix el log-loss calibrat a 0,942.

8.2 Mercats binaris: O/U 2,5 i BTTS

Ambdós mercats es formulen com a classificació binària: $y_{\text{O/U}} = \mathbf{1}[\text{gols} > 2{,}5]$ i $y_{\text{BTTS}} = \mathbf{1}[\text{gols local} > 0 \wedge \text{gols visitant} > 0]$. La distribució de classes és equilibrada (48,6%/51,4% i 49,1%/50,9%, respectivament), de manera que no cal sobremostreig.

Comparteixen la mateixa infraestructura d'entrenament, hiperparàmetres i calibració per lliga. El model O/U 2,5 afegeix cinc característiques combinades (`elo_sum`, `combined_form5_goals`, etc.) per un total de 62 variables.

8.3 Models de jugadors

Els mercats de jugadors individuals (p. ex., "Pedri fa més de 45,5 passades") representen una oportunitat de valor: les quotes sovint reflecteixen la reputació més que el rendiment recent. BETIQ entrena **set models de regressió** (`XGBRegressor`, `max_depth=6`) que prediuen el valor esperat de: gols, assistències, tirs totals, tirs a porteria, passades, entrades i faltes.

Les característiques (∼25–30 per mercat) s'organitzen en quatre grups: (i) historial individual del jugador (mitjanes mòbils de l'estadística objectiu en finestres de 3, 5 i 10 partits, mitjana de carrera, desviació estàndard i taxa per minut), (ii) context estadístic (mitjana mòbil de 5 partits d'estadístiques correlacionades amb l'objectiu, p. ex. per a gols: tirs, tirs a porteria i grans oportunitats), (iii) context d'equip i rival (possessió, passades, tirs i entrades dels darrers 5 partits), i (iv) metadata (posició *one-hot*: DEF/MID/FWD, camp local/visitant, nombre acumulat de partits).

Les dades provenen de la taula `player_match_analytics` (∼429.000 files). Es filtren els suplents i els jugadors amb menys de tres partits per garantir un mínim d'historial. Totes les característiques temporals apliquen `shift(1)` per evitar fuga d'informació.

Per generar prediccions d'apostes Over/Under a partir del valor esperat continu $\hat{y}$, s'aplica una distribució de Poisson:

$$P(\text{Over } k{,}5) = 1 - \sum_{i=0}^{k} \frac{e^{-\hat{y}}\,\hat{y}^i}{i!} \tag{3}$$

8.4 Pipeline d'inferència unificat

El mòdul `predict.py` centralitza tota la lògica d'inferència en tres funcions principals:

- `predict_fixture_all(fixture_id)`: rep un identificador de partit i retorna les prediccions dels tres mercats d'equip (1X2, O/U 2,5, BTTS). Internament, carrega les característiques pre-partit des de `match_analytics`, aplica els models entrenats i la calibració per lliga corresponent.
- `predict_player_all(fixture_id)`: genera prediccions per als set mercats d'estadístiques individuals. Per a cada jugador de l'alineació, calcula les característiques temporals, aplica el model de regressió i converteix el valor esperat a probabilitats Over/Under mitjançant Poisson.
- `predict_upcoming()`: orquestra el processament per lots de tots els partits programats. Consulta la taula `upcoming_fixtures`, invoca les dues funcions anteriors per a cada partit i escriu els resultats a les taules `fixture_predictions` i `player_predictions` de Supabase.

El pipeline s'executa diàriament com a cron a Railway.

9 Plataforma Web

El tauler intern de Streamlit servia com a eina d'exploració, però no era apte per a un producte final. S'ha construït una plataforma web completa amb Next.js 16, React 18, Tailwind CSS i autenticació Supabase, desplegada a Vercel.

La plataforma consta de 21 rutes organitzades en cinc mòduls: (1) **Mercats**: navegació tipus *bookmaker* per totes les lligues, amb quotes i prediccions del model; (2) **Value Picks**: apostes identificades automàticament amb avantatge $\geq$3%; (3) **Detall de Partit**: cinc pestanyes (mercats, jugadors, estadístiques, H2H, cronologia); (4) **Jugadors**: prediccions individuals; (5) **Tracker**: gestió d'apostes amb vista Kanban i estadístiques de ROI.

La pila tecnològica de la plataforma web inclou Next.js 16 amb renderització del costat del servidor (SSR) per a un temps de càrrega ràpid, TypeScript per a tipat estàtic, Tailwind CSS per al disseny responsiu, i Supabase Auth per a l'autenticació basada en *cookies* de sessió.

El contingut de prediccions està protegit per autenticació: els usuaris no registrats veuen les quotes del *bookmaker* però no les probabilitats del model. La plataforma inclou suport bilingüe complet (anglès i castellà) amb cobertura de traduccions a totes les pàgines, pestanyes i etiquetes.

El mòdul de **Value Picks** és especialment rellevant: identifica automàticament les apostes amb un avantatge (*edge*) superior al 3% entre la probabilitat del model i la probabilitat implícita del *bookmaker*. L'*edge* es calcula com $e = \hat{p}_{\text{model}} - 1/\text{quota}$, i només les apostes amb $e \geq 0{,}03$ es mostren a l'usuari, juntament amb el desglossament de quota, probabilitat del model i valor esperat.

El mòdul de **Tracker** permet als usuaris registrar les seves apostes (singles i combinades), classificar-les per estat (Pendent/En joc/Guanyada/Perduda) i visualitzar estadístiques de rendiment: ROI acumulat, bènefici net, nombre d'apostes i ràtio de guany.

10 Pipeline en Temps Real

Tres serveis cron desplegats a Railway garanteixen l'actualització contínua de les dades:

TAULA 4: Serveis en temps real (Railway).

Script	Funció	Freqüència
`fetch_upcoming`	Sincronitzar partits pròxims	Diari 7h UTC
`fetch_odds_snapshots`	Quotes en temps real	Cada 3 hores
`fetch_completed`	Resultats completats	*/30 12–23 UTC

Després de la sincronització de partits, el pipeline d'inferència (`predict_upcoming.py`) genera prediccions per a tots els mercats i les escriu a les taules `fixture_predictions` i `player_predictions` de Supabase, on la plataforma web les consumeix en temps real.

El servei de quotes (`fetch_odds_snapshots`) recopila les quotes de bet365 cada tres hores per a tots els partits amb *kick-off* dins de les pròximes 72 hores. Aquestes sèries temporals de quotes s'emmagatzemen a la taula `odds_snapshots` ($\sim$12,5 milions de files), permetent analitzar el moviment de les línies i calcular la probabilitat implícita del *bookmaker* en el moment més recent.

El servei de resultats (`fetch_completed_stats`) s'executa cada 30 minuts durant els horàris de competició (12:00–23:00 UTC, dilluns a dissabte) i ingereix els resultats dels partits completats, actualitzant les taules `match_analytics` i `player_match_analytics` per mantenir les característiques de forma al dia.

11 Resultats

Per facilitar la interpretació dels resultats, es defineixen a continuació les mètriques i els conceptes clau:

- **Precisió (*accuracy*):** percentatge de partits correctament classificats. Per al mercat 1X2, el *baseline* trivial (predir sempre victòria local) assoleix un 44,1%; qualsevol model útil ha de superar-lo significativament.
- **Log-loss:** mesura la qualitat de les probabilitats estimades; valors inferiors indiquen millors prediccions. Un predictor aleatori uniforme obté $\log(3) \approx 1{,}099$ per a tres classes. Un model que prediu les taxes base (44%/26%/30%) obté $\approx$1,05. Valors per sota d'1,0 indiquen calibració efectiva.
- **Brier score:** error quadràtic mitjà de les probabilitats; 0 és perfecte, 0,667 és aleatori per a tres classes.
- **ROI (*Return on Investment*):** benefici net dividit pel total apostat, expressat en percentatge. Un ROI del 0% indica que es recupera exactament el capital invertit; valors negatius impliquen pèrdues. Per obtenir ROI positiu, el model ha de superar no només la probabilitat real sinó també el marge del *bookmaker* ($\sim$5–8%).
- **Llindar d'avantatge:** avantatge mínim (*edge*) requerit per col·locar una aposta. L'*edge* es calcula com

$e = \hat{p}_{\text{model}} - 1/\text{quota}$. Un llindar del 3% significa que només s'aposta quan el model estima almenys 3 punts percentuals més de probabilitat que el *bookmaker*.

11.1 Models d'equip

La Taula 5 resumeix els resultats dels tres mercats d'equip al conjunt de validació (2024/25, cinc lligues).

TAULA 5: RESULTATS DELS MODELS D'EQUIP (VALIDACIÓ 2024/25).

Mercat	Tipus	Car.	Acc.	LL cal.	Brier
1X2 v2	3 cl.	57	52,9%	0,942	0,188
O/U 2,5 v3	Bin.	62	55,7%	0,661	0,235
BTTS v2	Bin.	53	—	—	—

Nota: BTTS desplegat amb calibració per lliga; mètriques no registrades al moment de redacció.

La Taula 6 mostra l'evolució del model 1X2 al llarg de les iteracions experimentals.

TAULA 6: EVOLUCIÓ DEL MODEL 1X2.

Versió	Car.	Acc.	LL cal.	Brier
Reg. Logística	53	50,6%	1,048	—
v1 (sense cal.)	53	51,9%	1,001	0,596
v1 + sigmoid	53	50,3%	0,997	0,593
v2 + Elo + per lliga	**57**	**52,9%**	**0,942**	**0,188**

El model v2 millora en totes les mètriques. La reducció de log-loss ($0{,}980 \rightarrow 0{,}942$, $-3{,}9\%$) indica que les probabilitats calibrades per lliga s'aproximen significativament millor a les freqüències observades.

11.2 Simulació de ROI

La Taula 7 mostra la simulació de ROI per al model v1 amb calibrador sigmoidal en cinc llindars d'avantatge mínim.

TAULA 7: SIMULACIÓ DE ROI (STAKE FIX, MODEL SIGMOIDAL).

Llindar	N_{raw}	ROI_{raw}	N_{sig}	ROI_{sig}
$\geq 1\%$	245	$-8{,}0\%$	242	$-8{,}1\%$
$\geq 2\%$	233	$-6{,}2\%$	227	$-7{,}2\%$
$\geq 3\%$	204	$-5{,}7\%$	196	$-3{,}4\%$
$\geq 4\%$	181	$-8{,}5\%$	167	$-3{,}2\%$
$\geq 5\%$	158	$-12{,}9\%$	149	$-6{,}4\%$

El calibrador sigmoidal millora consistentment el ROI als llindars del 3% i 4%, però el ROI romàn negatiu en tots els escenaris. No obstant això, la Taula 8 mostra que la distribució d'apostes és més equilibrada: la concentració en empats baixa del 56% (model sense calibrar) al 29%, i la taxa d'encert en victòria local arriba al 49,2%.

La taxa d'encert en victòria visitant (24,7%) és inferior a la taxa base (30,6%), suggerint que el model subvalora els equips visitants amb bon rendiment ofensiu. Aquesta observació identifica la millora de característiques com el proper vector d'investigació.

TAULA 8: DISTRIBUCIÓ D'APOSTES (SIGMOIDAL, LLINDAR $\geq 3\%$).

Resultat	N	Encert	Edge mitj.
Vict. local	59	49,2%	0,090
Empat	56	30,4%	0,070
Vict. visitant	81	24,7%	0,084
Total	**196**	35,7%	0,082

11.3 Models de jugadors

La Taula 9 resumeix els set mercats de jugadors amb l'estadística objectiu i les característiques de context principals.

TAULA 9: SET MERCATS DE JUGADORS.

Mercat	Objectiu	Context principal
Gols	`goals`	Tirs, tirs a porteria, grans oport.
Assistències	`assists`	Passades clau, centrades
Tirs totals	`shots`	Tirs a porteria, gols, regats
Tirs a port.	`shots_on`	Tirs totals, gols, tocs
Passades	`passes`	Passades precises, tocs, minuts
Entrades	`tackles`	Intercepcions, duels
Faltes	`fouls`	Entrades, duels, grogues

Cada model té entre 25 i 30 característiques i es desplega en producció generant $\sim$309 prediccions de jugadors actives per jornada. La conversió de valor esperat a probabilitat Over/Under mitjançant Poisson permet comparar directament amb les quotes del *bookmaker*.

11.4 Discussió

Els resultats demostren que el sistema BETIQ és capaç de generar probabilitats calibrades que superen els models de referència trivials (predicció constant, regressió logística). El model 1X2 v2 assoleix una precisió del 52,9%, 8,8 punts percentuals per sobre del baseline trivial (predir sempre victòria local, 44,1%). El model O/U 2,5 és especialment prometedor amb un 55,7% de precisió sobre un baseline del 51,4%.

No obstant això, el ROI negatiu indica que superar el marge del *bookmaker* ($\sim$5–8%) requereix un salt qualitatiu en la precisió de les probabilitats. La diferència entre el ROI del model ($-3{,}4\%$) i el marge del *bookmaker* suggereix que el model recupera una part significativa del *vig*, però no la totalitat.

11.5 Estat global del sistema

La Taula 10 resumeix l'estat complet del sistema BETIQ.

12 CONCLUSIONS

Aquest treball ha presentat el disseny, la implementació i l'avaluació de BETIQ, una plataforma d'analítica de futbol i intel·ligència per a apostes.

TAULA 10: ESTAT GLOBAL DEL SISTEMA (JUNY 2026).

Component	Estat	Detall
Pipeline d'ingestió	Complet	30/30 temporades
Taules analítiques	Complet	24,6K + 429K files
Model 1X2 v2	Producció	57 car., 5 calib.
Model O/U 2,5	Producció	62 car., 5 calib.
Model BTTS	Producció	53 car., 5 calib.
Models jugadors	Producció	7 models
Elo ratings	Complet	20,9K registres
Pipeline infer.	Desplegat	Cron diari
Plataforma web	Funcional	21 rutes, i18n
Pipeline real-time	Desplegat	3 serveis Railway

12.1 Principals aportacions

- Un pipeline d'ingestió idempotent per a 27.900 partits de cinc lligues europees durant sis temporades, amb taules analítiques de 24.600 files (equip) i 429.000 files (jugador).
- Un model 1X2 amb XGBoost, Elo i calibració per lliga que assoleix un log-loss calibrat de 0,942 i una precisió del 52,9%.
- Un model O/U 2,5 gols amb log-loss calibrat de 0,661 i precisió del 55,7%.
- Set models de regressió per a estadístiques individuals de jugadors amb conversió Poisson.
- Una plataforma web completa amb autenticació, mercats, *value picks* i seguiment d'apostes.
- Un pipeline en temps real amb tres serveis cron que generen prediccions diàries automàtiques.

12.2 Limitacions

El ROI simulat del model 1X2 és negatiu ($-3{,}4\%$ al llindar del 3%), indicant que la millora de la calibració és necessària però no suficient per obtenir benefici. Les mètriques del model BTTS no es van registrar formalment al `model_registry`, cosa que limita la comparabilitat. A més, l'avaluació dels models de jugadors és qualitativa: no es va dur a terme una simulació de ROI sistemàtica per als set mercats individuals per manca de quotes històriques suficients.

Malgrat el ROI negatiu, el valor principal del treball rau en l'arquitectura i la metodologia desenvolupades. El pipeline d'ingestió, les taules analítiques, el marc experimental estructurat i la plataforma web constitueixen una infraestructura completa i reutilitzable que permet iterar ràpidament sobre nous models i característiques. El fet que el model recuperi una part significativa del marge del *bookmaker* ($-3{,}4\%$ vs. un marge típic del 5–8%) suggereix que millores incrementals en les característiques o la calibració podrien apropar el sistema a la rendibilitat.

13 Treball Futur

Les línies d'investigació futures inclouen:

1. **Millora de característiques:** incorporació de gols esperats (*xG*), xuts a porteria i estadístiques de pressió defensiva. La taxa d'encert en victòria visitant (24,7% vs. base 30,6%) indica que el model necessita millors indicadors de qualitat ofensiva visitant.
2. **Optimització d'hiperparàmetres:** cerca grid o random sobre el conjunt de validació. El model v1 empra paràmetres per defecte raonables però no optimitzats.
3. **Llindars diferenciats per classe:** l'avantatge mitjà de les apostes en victòria local (0,090) és superior al de l'empat (0,070), indicant que llindars per classe podrien millorar la selecció.
4. ***Player embeddings*:** codificar la composició de l'equip com a vector de característiques, rellevant quan jugadors clau estan lesionats o sancionats.
5. **Extensió a competicions internacionals:** adaptar el sistema per a tornejos com la Copa del Món, amb dades de seleccions nacionals i característiques específiques (rànquing FIFA, format de torneig).

Agraïments

L'autor agraeix a la tutora Natàlia Blasco Andreo la seva orientació i disponibilitat durant tot el desenvolupament d'aquest treball.

Referències

[1] Statista Research Department, "Sports Betting — Worldwide," *Statista Market Insights*, 2024.

[2] J. Buchdahl, *Squares and Sharps, Suckers and Sharks: The Science, Psychology and Philosophy of Gambling*, Oldcastle Books, 2016.

[3] M. J. Maher, "Modelling Association Football Scores," *Statistica Neerlandica*, vol. 36, núm. 3, pp. 109–118, 1982.

[4] M. J. Dixon i S. G. Coles, "Modelling Association Football Scores and Inefficiencies in the Football Betting Market," *J. Royal Statistical Society: Series C*, vol. 46, núm. 2, pp. 265–280, 1997.

[5] O. Hubáček, G. Šourek i F. Železný, "Exploiting Sports-betting Market Using Machine Learning," *International Journal of Forecasting*, vol. 35, núm. 2, pp. 783–796, 2019.

[6] L. Grinsztajn, E. Oyallon i G. Varoquaux, "Why Tree-based Models Still Outperform Deep Learning on Tabular Data," *NeurIPS*, vol. 35, 2022.

[7] L. M. Hvattum i H. Arntzen, "Using ELO Ratings for Match Result Prediction in Association Football," *Int. Journal of Forecasting*, vol. 26, núm. 3, pp. 460–470, 2010.

[8] T. Chen i C. Guestrin, "XGBoost: A Scalable Tree Boosting System," *Proc. ACM SIGKDD*, pp. 785–794, 2016.

[9] S. Kaufman, S. Rosset, C. Perlich i O. Stitelman, "Leakage in Data Mining: Formulation, Detection, and Avoidance," *ACM Trans. Knowledge Discovery from Data*, vol. 6, núm. 4, 2012.

[10] J. Platt, "Probabilistic Outputs for Support Vector Machines," *Advances in Large Margin Classifiers*, MIT Press, pp. 61–74, 1999.

[11] W. W. Royce, "Managing the Development of Large Software Systems," *Proc. IEEE WESCON*, pp. 1–9, 1970.

[12] A. C. Constantinou, "Dolores: A Model that Predicts Football Match Outcomes from All Over the World," *Machine Learning*, vol. 108, pp. 49–75, 2019.

[13] H. Rue i O. Salvesen, "Prediction and Retrospective Analysis of Soccer Matches in a League," *J. Royal Statistical Society: Series D (The Statistician)*, vol. 49, núm. 3, pp. 399–418, 2000.

[14] J. Goddard, "Regression Models for Forecasting Goals and Match Results in Association Football," *International Journal of Forecasting*, vol. 21, núm. 2, pp. 331–340, 2005.

[15] A. E. Elo, *The Rating of Chessplayers, Past and Present*, Arco Publishing, Nova York, 1978.

APÈNDIX

Les taules següents complementen el cos principal del document. La Taula 11 defineix els termes clau del domini de les apostes esportives. La Taula 12 detalla la pila tecnològica del projecte. La Taula 13 recull els identificadors de temporada de SportMonks necessaris per reproduir la ingestió de dades (Secció 6). La Taula 14 mostra la planificació temporal del projecte.

A.1 Glossari de termes d'apostes

A.2 Pila tecnològica

A.3 Identificadors de temporada

A.4 Planificació temporal (Gantt)

TAULA 11: GLOSSARI DE TERMES CLAU DEL DOMINI D'APOSTES ESPORTIVES.

Terme	Definició
Bookmaker	Casa d'apostes que fixa quotes i actua com a contraparta de l'apostador.
Quota (*odds*)	Preu ofert pel *bookmaker*; reflecteix la probabilitat implícita d'un resultat. Una quota de 2,50 implica $p = 1/2{,}50 = 40\%$.
Vig / sobreround	Marge de benefici del *bookmaker*: la suma de probabilitats implícites supera el 100% (típicament 105–108%).
Edge (avantatge)	Diferència entre la probabilitat estimada pel model i la implícita del *bookmaker*: $e = \hat{p} - 1/\text{quota}$.
Value bet	Aposta on l'*edge* és positiu: el model estima més probabilitat que el *bookmaker*.
CLV	*Closing Line Value*: diferència entre la quota obtinguda i la quota de tancament. Superar-la consistentment indica avantatge a llarg termini [2].
Stake	Import apostat en una selecció.
ROI	*Return on Investment*: benefici net dividit pel total apostat ($\times 100\%$). Un ROI del 0% indica recuperació exacta del capital.
1X2	Mercat principal: victòria local (1), empat (X) o victòria visitant (2).
O/U 2,5	Mercat Over/Under: si es marquen més o menys de 2,5 gols al partit.
BTTS	*Both Teams To Score*: si ambdues equips marquen almenys un gol.

TAULA 12: PILA TECNOLÒGICA DEL PROJECTE BETIQ.

Tecnologia	Ús
Python + pandas	ETL, enginyeria de característiques
Supabase (PG)	Base de dades PostgreSQL gestionada
SportMonks API	Font de dades de futbol
XGBoost	Models predictius
scikit-learn	Calibració probabilística
Next.js + React	Plataforma web (SSR)
Tailwind CSS	Disseny responsiu
Supabase Auth	Autenticació
Railway	Pipeline en temps real (cron)
Vercel	Desplegament web continu
Streamlit	Tauler intern d'exploració

TAULA 13: IDENTIFICADORS SPORTMONKS PER LLIGA I TEMPORADA.

Lliga	20/21	21/22	22/23	23/24	24/25	25/26
Prem. Lea.	17420	18378	19734	21646	23614	25583
La Liga	17480	18462	19799	21694	23621	25659
Bundesl.	17361	18444	19744	21795	23744	25646
Serie A	17488	18576	19806	21818	23746	25533
Ligue 1	17160	18441	19745	21779	23643	25651

TAULA 14: DIAGRAMA DE GANTT DEL PROJECTE BETIQ.

Fase	Feb	Mar	Abr	Mai	Jun
Fase 0: Dades					
Fase 1: Model 1X2					
Fase 2: Mercats + Web					
Fase 3: Docum.					