Anàlisi i avaluació del rendiment dels futbolistes

Martí Tuneu Font

Resum— Durant els últims anys l'anàlisi estadístic ha anat guanyant importància dins del món de l'esport, sent utilitzat per a la millora del rendiment dels esportistes i per a descobrir-ne que puguin encaixar amb determinat equip. En el futbol s'utilitzen moltes estadístiques variades que serveixen per analitzar els jugadors, però no hi ha cap estadística estandarditzada per a mesurar el rendiment general dels futbolistes durant un o més partits. En aquest treball es proposa el desenvolupament d'un mètode per a avaluar-ne el rendiment a partir d'una estadística ja força estesa, els *expected goals*, calculats a partir de la implementació d'una xarxa neuronal, i una equació per a la valoració que parteix del valor numèric d'aquests *expected goals* per a calcular el rendiment general dels futbolistes.

Paraules clau— xG (expected goals), regressió logística, xarxa neuronal, xA (expected assists)

Abstract—Over the last few years, statistical analysis has become increasingly important in the sports world, being employed both to improve the performance of sportsmen and sportswomen and to match new and emerging footballers with suitable teams. In football, many different statistics are used to analyse the players but there is no standardised statistic to measure the footballers' general performance during one or more games. In this work we propose the development of a method to evaluate such performance based on a statistic that is already widely used, the expected goals, predicted by a neural network model, and an evaluation equation that starts from the value of these expected goals to calculate the overall performance of football players.

Index Terms— xG (expected goals), logistic regression, neural networks, xA (expected assists)

1 Introducció - Context del treball

L'estadística i l'anàlisi estadístic s'han anat incorporant en la indústria de l'esport durant els últims anys amb molta rapidesa, principalment en esports com el bàsquet o el futbol americà, ja que per la pròpia topologia d'aquests esports resultava més fàcil generar anàlisis estadístics vàlids i rellevants. En el món del futbol la introducció d'aquestes anàlisis estadístics ha estat més lenta, però també s'ha acabat donant i avui en dia pràcticament tots els equips professionals els utilitzen tant per buscar jugadors que puguin encaixar en l'equip com per a millorar el rendiment dels futbolistes amb els quals ja conten.

Aquestes estadístiques generalment només aporten dades sobre certs moments i certes accions, que si bé poden ajudar, no donen una visió global del rendiment d'un futbolista durant el transcurs d'un partit.

A les ja tradicionals mesures com els gols, les assistències, les passades completades i d'altres mètriques simples en els últims anys se n'hi ha afegit algunes com les següents:

- xG[1]: Els *expected goals* es calculen com la probabilitat que un jugador marqui gol en un moment determinat, d'aquesta manera es mesura la qualitat de les oportunitats que ha tingut.
- xA[1]: Les *expected assists* es calculen com la probabilitat de que una passada acabi en una ocasió de gol.
- Àrea de defensa[1]: Àrea que defineix la zona en què un jugador s'ha implicat defensivament durant un partit.
- PPDA: *Passes Allowed per Defensive Action*, quantifica l'agressivitat de la pressió exercida per un equip quan no té la pilota. Calcula el nombre de passades fetes per l'equipo contrari abans que un equip faci una acció defensiva, ja sigui una falta, una intercepció o una lluita amb l'oponent.

Com podem veure, tant els xG, les xA i l'àrea de defensa, milloren la informació que en podem extreure respecte de les mètriques tradicionals, però tot i això continuen centrant-se en certs aspectes del joc i en certes funcions que no tots els futbolistes tenen per què desenvolupar.

A més, tot i que l'objectiu final d'un partit de futbol és aconseguir marcar gols, que es duen a terme xutant a porteria aquest tipus d'accions ocorren amb molta poca freqüència durant el transcurs d'un partit. De fet, la mit-

- *E-mail de contacte: marti.tuneu@e-campus.uab.cat*
- *Menció realitzada: Computació*
- *Treball tutoritzat per: Jordi Serra Ruiz (Departament de Ciències de la Computació)*
- *Curs 2020/21*

jana de xuts dels millors equips anglesos, la més alta d'Europa, és de tan sols 27,63 xuts per partit [2], una xifra molt petita si tenim en compte que el dataset de StatsBomb[3], empresa que es dedica a l'anàlisi futbolístic, que s'ha fet servir durant el desenvolupament d'aquest projecte registra més de 3400 esdeveniments per partit.

Per tant, i tot i tenir en compte que el més valuós durant un partit de futbol són els gols que es marquen es conclou que resulta necessari una mètrica que permeti analitzar el rendiment dels futbolistes de forma global durant un partit i aquest és l'objectiu final d'aquest projecte. A partir de tècniques d'aprenentatge computacional s'implementarà un sistema per a valorar el rendiment global dels jugadors durant un o més partits de futbol.

L'informe s'estructurarà de la següent manera. En la secció següent, la número 2, es presentaran els objectius que es pretenen complir. En la secció 3 es mostra i es descriu l'estat de l'art dins del camp de l'estadística esportiva. En la secció número 4 s'explica de forma detallada la metodologia i el desenvolupament del projecte i dels objectius proposats. La secció 5 mostra els resultats obtinguts i una discussió que acabarà en l'última secció on s'expliquen les conclusions extretes durant el projecte.

2 OBJECTIUS

El projecte té com a objectiu principal el desenvolupament d'una mètrica que permeti avaluar el rendiment global dels futbolistes durant un partit i durant tots els minuts jugats. Aquest rendiment s'avaluarà a partir dels anomenats events del dataset de StatsBomb, que s'utilitzarà com a base per aconseguir les puntuacions. Els events es descriuen com a cada acció, pot ser amb pilota o sense, que ocorre durant un partit, incloent-hi inicis i finals de cada part, les alineacions i també les substitucions de jugadors. Cada event porta associades certes característiques que resultaran útils per el desenvolupament del projecte.

Entre el punt inicial del projecte, els events del dataset, i l'objectiu final, la mètrica d'avaluació, hi ha objectius intermedis que es detallen a continuació.

2.1 Extracció de les dades necessàries

Com ja s'ha indicat la base de la qual es partirà per dur a terme el projecte és el dataset d'events de StatsBomb[3]. Aquest inclou diferents lligues, d'ambdós sexes, i diferents temporades, que a la vegada inclouen diferents partits on s'han registrat events.

Aquests events inclouen molta informació irrellevant que es pot eliminar en alguns casos i en d'altres es pot simplificar. Per tant el primer objectiu és fer una correcta extracció de dades, quedant-nos finalment només amb aquelles útils per continuar amb el projecte i estandartitzant-les, de forma que tots els events tinguin les mateixes característiques.

2.2 Generació model d'xG

L'objectiu d'aconseguir una mètrica per a l'avaluació dels jugadors es basa en avaluar cada acció i puntuar-la respecte al seu impacte en les posteriors accions. Per puntuar aquestes accions es pretén seguir la següent lògica: en quina mesura una acció ens ha acostat a marcar un gol? Un aspecte bàsic per a solucionar aquesta qüestió és conèixer la probabilitat de marcar gol d'una determinada acció, el que es coneix com a xG, per tant un altre dels objectius a assolir és generar un model d'xG suficientment robust per dur a terme aquesta tasca.

2.2.1 Comparació de models d'xG

Partint de l'objectiu anterior, generar un model d'xG suficientment robust, sorgeix un nou objectiu, fer una comparativa entre dos dels models mes usats per a la predicció dels xG, la regressió logística i xarxes neuronals,

2.3 Valoració dels jugadors

En última instància i com a objectiu final tenim la valoració dels jugadors que es durà a terme a partir de la mètrica dels xG predits amb el model generat anteriorment. L'objectiu és valorar els jugadors per la seva contribució, sigui positiva o negativa, en el partit i estandarditzar la mesura pels 90 minuts que dura un partit, ja n'hagi jugat més o menys.

3 ESTAT DE L'ART

Com s'ha esmentat en la introducció d'aquest projecte, la majoria d'estadístiques que s'usen actualment avaluen aspectes concrets del joc. És per això que les primeres visualitzacions a aparèixer per analitzar el rendiment global d'un futbolista són els *player radars* [4]. Els *players radars* són eines per a la visualització de diferents estadístiques, que varien depenent de la posició del jugador a analitzar, de forma conjunta. En aquests radars es veuen els números que ha obtingut el jugador per a cada estadística dins d'un rang entre 0 i el màxim que hagi fet qualsevol altre jugador en aquell estadística. Normalment es mostren entre unes 10-12 estadístiques per jugador.

Tot i això també existeixen sistemes d'avaluació generals. El primer exemple, i segurament el més bàsic, és el *plus-minus rating*[5] que també s'utilitza en altres esports com el bàsquet, on resulta més útil pel propi tipus d'esport. La idea darrere d'aquest sistema és: quins canvis hi ha hagut al marcador mentre el jugador era al camp? La fórmula utilitzada és la següent:

$$(\sum G/m)\ X\ 90$$

Sent G els gols marcats o concedits, positiu quan parlem de gols marcats i negatiu quan són gols concedits i m el nombre de minuts jugats pel futbolista. En els casos més avançats en comptes d'utilitzar els gols marcats i concedits es fan servir els xG.

Un dels altres mètodes existents és l'utilitzat per

l'empresa EA Sports[6]. Aquest és una suma balancejada amb pesos de sis índexs d'aspectes diferents del joc, sentne aquesta l'equació resultant:

$$I = 100\,X\,(0.25I_1 + 0.375I_2 + 0.125I_3 + 0.125I_4 + 0.0625I_5 + 0.0625I_6)$$

Sent els índexs, respectivament: *model match outcome*, *point-sharing index*, *appearance index*, *goal-scoring index*, *assists index* i *clean-sheets index*. Es pot trobar més informació de cadascun d'aquests índexs en l'article referenciat.

Seguint amb més models, hi trobem el proposat per Sarah Rudd[7], que es basa en cadenes Markov i una segmentació del camp en set àrees. S'assigna a cada una d'aquestes àrees una probabilitat de marcar i el valor que li assignem a una acció en concret és la diferència en la possibilitat de marcar en aquella acció i la possibilitat de marcar en la següent.

Un dels altres models aparegut recentment és el VAEP [8] que es basa en les cadenes de possessió, el model extreu per cada estat del partit, sent un estat una cadena de tres accions, i avalua per cada un d'aquests estats quina és la probabilitat que després d'aquest estat hi hagi un gol, venint aquesta probabilitat donada per un model d'xG i assignant com a valor de la jugada aquests xG.

4 METODOLOGIA

En la següent secció s'explicarà el desenvolupament del projecte, amb les següents subseccions que s'enumeren a continuació: eines utilitzades, extracció de dades, generació dels models i finalment la generació de l'equació final per valorar els jugadors.

4.1 Eines utilitzades

Per a dur a terme el projecte s'ha utilitzat el dataset obert de StatsBomb, que s'organitza de la següent manera: un fitxer amb totes les competicions i les temporades en les quals hi ha partits, un fitxer amb tots els partits disponibles per cada temporada i cada competició i finalment per a cada partit dos fitxers, un amb l'alineació inicial i l'altre amb tots els events que han ocorregut durant aquest. El dataset conté dades de 7 competicions diferents, sent 878 el total de partits del dataset i contenint cada partit uns 3400 events registrats de mitjana. Tots els fitxers estan en format JSON.

Per a desenvolupar el projecte s'ha decidit implementar-lo en el llenguatge Python en la seva versió 3.8.8[9], sent aquest compatible amb les llibreries necessàries per poder completar le projecte i que s'esmenten a continuació: Pandas[10] , per a una correcta i més senzilla gestió de les dades a tractar, és a dir els events dels partits. JSON[11] per poder llegir els fitxers del datast de StatsBomb i exportar-los seguidament a dataframes de Pandas. SciPy[12] per poder fer de forma correcta el càlcul de distàncies entre events i porteria. Math [13], pel càlcul de forma correcta d'angles entre events i porteria. Per implementar els models probabilístics d'xG s'han utilitzat per una banda scikit-learn[14], per la regressió logística, i Tensorflow[15] i Keras[16] per la xarxa neuronal. També s'ha utilitzat la llibreria Joblib[17] com a complement per guardar el model de regressió logística.

4.2 Extracció de dades

La primera part necessària per a un correcte desenvolupament del projecte és l'extracció de dades a partir del dataset de StatsBomb. S'extreuen les dades de tots i cadascun dels partits que hi ha en el dataset. El dataset conté, per cada event, la següent informació:

- **Index**: Identificador únic de l'event.
- **Period**: Període durant el qual l'event ocorre.
- **Timestamp**: Temps del partit en el qual l'event ha ocorregut.
- **Minute**
- **Second**
- **Type**: Id i descripció del tipus d'event.
- **Possession**: Número que indica la cadena de possessió en aquell moment, s'increment cada vegada que la possessió canvia d'un equip a l'altre.
- **Possession_team**: L'equip en possessió de la pilota quan ocorre l'event.
- **Play_pattern**: Tipus de jugada (oberta, pilota parada, etc.)
- **Team**
- **Player**
- **Position**: (Posició del jugador: porter, lateral dret, etc.)
- **Location**: Posició en x,y del jugador dins del camp.
- **Outcome**: L'acció és exitosa o falla.

A més, alguns events contenen informació especial com la següent:

- **Duration**: Duració de la jugada.
- **Under_pressure**: La jugada s'ha fet amb la pressió d'un oponent.
- **Out**: Després de l'event la pilota surt del camp.
- **Related_events**: Events relacionats amb l'actual.
- **Tactics**: Alineació de l'equip. (*e.g.* 4-2-31)

Els objectius després d'aquesta extracció de dades són eliminar els events que no ens aportin informació rellevant sobre el que ha passat al terreny de joc, eliminar les dades de cada event que no ens resultin necessàries per al projecte i adaptar les restants a un format que resulti fàcilment tractable.

Primerament doncs s'han eliminat els tipus d'events que no aporten informació sobre el joc, per tant són els events que no corresponen a accions dutes a terme pels jugadors. S'eliminen, doncs, els següents events.

- Half Start
- Starting XI
- Dribbled Past
- Referee Ball-Drop
- Tactical Shift

- Half End
- Player On
- Player Off
- Substitution
- Offside
- Camera On
- Ball Receipt
- Bad Behaviour

Després de la neteja els events que poden ocórrer durant el partit són els que es mostren a l'Apèndix 1.

El següent tractament previ a l'adaptació de les dades és eliminar les columns que no siguin necessàries. S'eliminen doncs les següents columnes per tots els events:

- Timestamp
- Minute
- Second
- Possession
- Position
- Duration
- Under_pressure
- Out
- Related_events
- Tactics

Una vegada fets aquests tractaments previs es poden tractar els valors restants i adaptar-los a les accions que durem a terme després.

Per les següents tasques a completar del projecte es necessiten les següents variables per cada event:

- Game_id
- Possession_team
- Original_event_id
- Team_id
- Period
- Player_id
- Outcome
- Start_x
- Start_y
- Type_id
- Type_name
- Body_part
- Situation
- Distance
- Angle
- Prev_type_id

Les variables que coincideixen en nom amb les del dataset original són les mateixes. Se n'afegeixen de noves, com el **Game_id**, per identificar el partit en el qual ha ocorregut l'acció, el P**rev_type_id**, que com el seu nom indica és el tipus d'acció anterior a l'actual i servirà per a calcular els xG. Altres com la situation són transformacions de variables prèviament existents com el **Play_pattern** però simplificades, canviant el tipus original de diccionari a un simple nombre identificador del tipus de jugada, els tipus de jugada que poden ocórrer durant el partit es defineixen com: open play (jugada normal), free kick (falta directa o indirecta), corner i penalty. El tipus d'events, abans contingut en una sola variable, ara es desglossa en dues, l'identificador (**Type_id**) i el nom del tipus d'event (**Type_name**). El mateix passa amb l'anterior variable location que ara desglossada ara en les variables **start_x** i **start_y**. La variable **body_part**, abans continguda dins d'altres variables com el xut ara també es troba simplificada en forma d'un nombre identificador, havent-hi només dues possibilitats: peus o altres, que inclou el cap o altres parts com les cuixes. En última instància tenim les variables **Distance** i **Angle**. Com el seu propi nom indiquen són, respectivament, la distància entre el lloc des del qual ocorre l'event (*start_x*, *start_y*), fins al centre de la posició on es troba la porteria contrària, i l'angle d'obertura entre la posició des d'on ocorre l'event i els dos pals de la porteria. La distància es calcula com la distància euclidiana i l'angle a partir de trigonometria.

4.3 Generació de models

La següent etapa dins del projecte consisteix en generar un model probabilístic de càlcul d'xG suficientment fiable per poder dur a terme la següent iteració del projecte. Els dos models han estat entrenats en un ordinador portàtil amb processador Intel Core i5—7200U, 2700MHz i 8 GB de memòria RAM.

4.3.1 Consideracions prèvies

Dins d'aquesta secció es mostraran els resultats dels dos tipus de models. Cal especificar que per entrenar i fer les comprovacions sobre la fiabilitat dels dos models s'han seleccionat les accions de tipus xut extretes del dataset que dona com a resultat el mètode explicat en la subsecció anterior d'extracció de dades sobre el dataset de StatsBomb. El dataset final conté 22353 xuts a porteria, que s'han dividit, per l'entrenament i test dels dos models en 17882 xuts d'entrenament i 4471 xuts per testejar els resultats, és a dir, una proporció del 80%-20%.

De les variables que resulten per cada event de la extracció de dades duta a terme prèviament, hi trobem variables que serveixen per identificar l'event i per fer-lo més entenedor i d'altres que resulten útils per calcular els xG. Es faran servir com a característiques per a entrenar ambdós models les següents variables:

- Distance
- Angle
- Prev_type_id
- Body_part
- Situation

Els events de xut consten de més informació, per

exemple, la posició del porter sobre el terreny de joc en el moment del xut o la posició de tots els jugadors en el moment en què es xuta la pilota. Aquestes característiques podrien ser d'ajuda i es considera que millorarien els resultats a obtenir pels dos models, ja que es tracta de variables importants, una bona col·locació per part del porter és crucial per aturar un xut i la densitat de jugadors en l'angle entre el xut i els pals de la porteria també són característiques que dificulten la tasca de marcar un gol. Tot i així aquestes característiques no es poden utilitzar en els nostres models, ja que no comptem amb elles per tots els events i l'objectiu del projecte és usar els xG en totes les accions per avaluar-les, per tant es descarten aquestes variables.

Un dels objectius del projecte era provar dos tipus diferents de models de predicció per comparar-ne els resultats entre si i acabar escollint-ne el millor. Les propostes plantejades pels dos models diferents són la regressió logística i un model de xarxes neuronals. S'han triat aquests models ja que són dos dels models més populars a l'hora d'implementar models per la predicció d'xG però el projecte és aplicable a qualsevol altre tipus de model.

4.3.2 Regressió logística

El primer dels dos models a aplicar és el de regressió logística. Com ja s'ha explicat anteriorment es fa servir la regressió logística implementada en la llibreria de Scikit-learn.

Aquesta implementació té en compte diferents paràmetres que a continuació es detallaran:

- Penalty: Defineix la tècnica de regularització a usar, scikit permet fer servir L1, L2 i Elastic-Net o cap. El nostre model usa L2.
- C: La inversa de la pressió de la regularització, en el nostre cas es demostra millor una regularització més relaxada i li donem a C un valor de 10.
- Max_iter: Permet especificar un nombre màxim d'iteracions si no s'ha convergit. S'eleven el nombre d'iteracions fins a 6000.
- Solver: Algorisme que s'utilitzarà per a minimitzar la funció de cost. Per el nostre model s'utilitza l'algorisme LBFGS[19], que és una versió amb memòria limitada de l'algorisme de Broyden-Fletcher-Goldfarb-Shanno.
- Class weight: Permet balancejar les classes de sortida (0,1) amb pesos. Després de diverses proves sense balancejar les classes i diferents tests amb diferents pesos els millors resultats han resultat amb els següents pesos: 0: 3, 1: 2.5.

La resta de paràmetres com *dual*, *tol*, o *fit_intercept* s'han deixat amb els valors per defecte de la implementació de scikit-learn.

Una vegada entrenat el model amb els 17882 xuts d'entrenament s'han recollit les següents mètriques, que es poden veure en la Taula 1, per calcular-ne la fiabilitat i comparar-lo amb el model de xarxes neuronals, i que ens donen resultats com una *accuracy* per sobre del 87% i una precisió que es troba per sobre del 55%:

	Regressió logística
Accuracy	87.53%
Precision	55.627%
Log Loss	0.326
Brier Score	0.097

Taula 1. Resultats del model de regressió logística en les mètriques: Accuracy, precision, log los i Brier score.

Una vegada entrenat i generat el model s'ha guardat en format joblib amb l'ajuda de la llibreria que porta el mateix nom.

4.3.3 Xarxes neuronals

El segon i últim dels dos models a implementar és el de xarxes neuronals. En aquest cas hem utilitzat les llibreries de TensorFlow i de Keras ja que són de gran ajuda per desenvolupar models de xarxes neuronals de forma senzilla.

S'ha implementat un model de xarxes neuronals seqüencial format per diverses capes que es descriuran a continuació. Una vegada passades per totes les capes el model retorna un valor probabilístic.

En la Figura 1 es mostra una representació de les capes i de l'estructura de la xarxa neuronal implementada pel projecte.

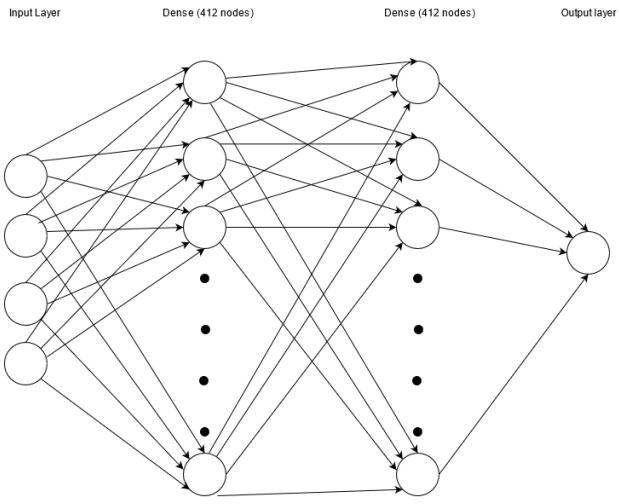

Figura 1. Representació de la xarxa neuronal.

Com es pot veure en l'estructura el model amb dues capes Dense amb activació Relu, la més habitual i la que ha donat més bon resultat tot i haver provat Swish[20], i

una última capa Dense amb un output de 1 i amb una funció d'activació *sigmoid*.

A continuació es fa una breu explicació del tipus de capa triada i les dues funcions d'activació escollides.

- **Dense**: La capa implementa la següent operació, on *dot* és el producte escalar entre input, el nombre d'entrades que rep de la capa anterior, *kernel* és la matriu de pesos creada per la pròpia capa i *bias* és el vector de regularització, en aquest cas les capes implementades no contaven amb cap vector d'aquest tipus:

$$Output = Activació(dot(input, kernel) + bias)$$

- **ReLu**: És la funció d'activació més comuna en xarxes neuronals i la funció que aplica és la següent:

$$f(x) = \max(0, x)$$

- **Sigmoid**: Com ReLu també és una de les funcions d'activació més comunes, en aquest cas quan el valor a predir és una probabilitat, ja que la funció només existeix entre 0 i 1. La funció que aplica és la següent:

$$f(x) = \frac{1}{1 + e^{-t}}$$

Com hem fet amb el model de regressió logística, una vegada entrenat amb els 17882 xuts d'entrenament es recullen les següents mètriques, que es poden veure en la Taula 2. Obtenim mesures similars en tres dels aspecte, *accuracy* (per sobre del 87%), *log loss* i *Brier Score*, però la precisió presenta una millora significativa, elevant-se fins més del 94%.

	Xarxes neuronals
Accuracy	87.45%
Precision	94.667%
Log Loss	0.325
Brier Score	0.096

Taula 2. Resultats del model de xarxes neuronals en les mètriques: Accuracy, precision, log loss i Brier score.

Una vegada entrenat i general el model s'ha guardat en format .h5 per a poder-lo recuperar després.

4.3.4 Comparació de models

Amb els dos models ja generats i conegudes les seves respectives mètriques se n'ha de fer l'elecció, tenint en compte els seus resultats. A part de comparar les taules 1 i 2, també es conta amb la mètrica ROC-AUC, que es mostra en la figura 2, i on es pot veure la corba ROC i el valor AUC (area under the curve).

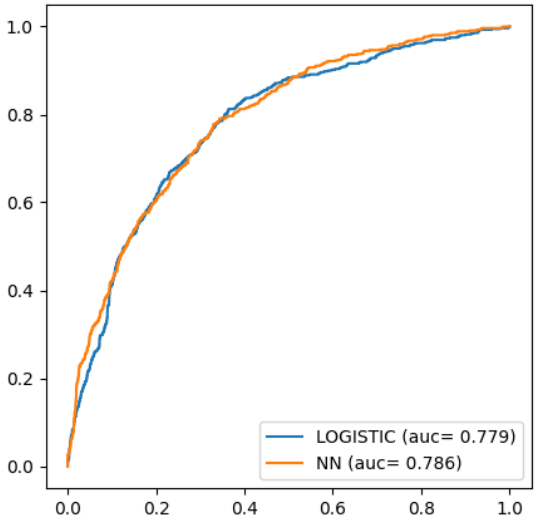

Figura 2. Corba ROC i valor AUC.

Com podem comprovar tant en les mètriques que es mostren en les taules 1 i 2 com en la figura 2, els dos models presenten un rendiment similar en tots els aspectes excepte en la precisió. El model de regressió logística mostra uns resultats decebedors en aquest aspecte mentre que el de xarxes neuronals té un 94.75%.

Tenint en compte que en els altres aspectes els models són pràcticament similars es decideix optar pel model de xarxes neuronals, ja que té una millor precisió i també presenta una lleugera superioritat en l'AUC.

Per comprovar que el nostre model és fiable, i no només que és millor respecte del de regressió logística, el comparem amb altres models. Respecte l'accuracy els models que es troben, duts a terme per Eoin O'Brien, oscil·len en valors entre el 87% i el 88% [21], exactament en els termes en què es mouen els nostres dos models. En la mètrica *log loss*, la pèrdua logística, es fa una comparació exclusivament amb un model de xarxes neuronals, proposat pel portal openGoal, i els valors estan entre 0.33 i 0.29 [22], sempre depenent de la quantitat de mostres d'entrenament que s'agafin. El nostre model té un resultat de 0.325 en aquest camp, és a dir que estaria entre els pitjors resultats però tot i així es mouria dintre dels valors raonables. L'AUC també té comparacions amb altres models, en aquest cas de regressió logística i amb Gradient Boost Classifier, els models comparats donen resultats d'entre el 75% i el 82% [23], per tant com passa en les altres mètriques els nostres models es troben dins dels valors raonables. Per últim podem comparar els models amb els estudis duts a terme per el DTAI Sports Analytics Lab de la Universitat de Leuven on es poden veure els resultats en funció de la Brier Score i AUC per models de regressió logística i de XGBoost [24]. En Brier Score els models tenen una puntuació entre 0.80 i 0.81 i per tant rendeixen significativament millor i en AUC entre el 76% i el 77%, és a

dir, amb resultats similars als nostres models.

Amb totes aquestes dades i fetes les comparacions podem concloure que el model escollit, més enllà de ser el millor d'entre els dos entrenats, és suficientment robust, amb les dades amb les quals es contava, com per continuar endavant amb el projecte.

4.4 Equació per la valoració de jugadors

L'objectiu final del projecte era crear una mètrica per la valoració del rendiment general dels futbolistes durant un partit. La tesi principal darrere la proposta per la valoració dels jugadors és la següent, l'objectiu principal d'un equip de futbol és marcar gols i no concedir-ne, per tant el rendiment d'un fubolista serà positiu sempre i quan acosti el seu equip a aconseguir un gol i/o redueixi les possibilitats de l'equip contrari de marcar-ne. D'aquí la importància de comptar amb un model suficientment robust de càlcul d'xG per a poder controlar en tot moment les possibilitats d'ambdós equips de marcar un gol.

Amb aquesta premissa, el sistema de valoració que es proposa és el càlcul de la millora de les possibilitats de marcar i de concedir per a cada acció que un jugador duu a terme. És a dir, les possibilitats de marcar, i les de concedir, que té un equip abans que el jugador en qüestió dugui a terme una acció i les possibilitats de marcar,i de concedir, una vegada el jugador ha fet l'acció. El problema principal és establir fins a quin punt una acció influeix en les següents, no hi ha estudis que validin que una jugada pot impactar fins la *n* jugada següent, tot i així sabem que l'impacte d'una jugada és limitat. És per això que es decideix establir l'abast en les dues jugades següents. Es pren aquesta decisió ja que calcular la diferència entre les possibilitats de marcar i concedir de la jugada actual i posterior seria un recorregut massa curt en una jugada, i no tindria en compte conceptes com el tercer home/dona, i com ja hem dit, sabem que l'impacte d'una acció és limitat, per tant no volem establir un nombre *n* de jugades posteriors massa elevat.

Es descarta també introduir biaixos en la valoració, com per exemple si l'equip que controla la pilota en el moment de l'acció va per sobre en el marcador, ja que es considera que fer-ho seria sobrevalorar el fet que un equip aconseguís allunyar el contrari de la possibilitat de marcar gols. A més de no tenir en compte el tipus de competició del que es tracta, a vegades els gols fora de casa valen doble, el goal-average entre els equips o el goal-average general en competicions regulars.

Una vegada aclarits tots aquests aspectes podem veure com es calcula la valoració dels jugadors.

Per cada acció calcularem, primerament, un valor d'atac, que anomenarem *attacking value*, i un valor defensiu que anomenarem *deffensive value*, a partir de les probabilitats de marcar, *score_prob,* i de concedir un gol, *concedí_prob,* respectivament. A partir d'aquests dos valors es calcula el *total value* que és la suma d'ambdós valors.

El fluxe del càlcul de la valoració és el següent:

1. Predicció de les probabilitats de marcar i de concedir per cada acció. Calculem únicament les probabilitats d'un equip ja que les de l'altre equip seran les inverses.

2. Calculem per cada acció *l'attacking value* (AV) amb la següent equació, on S és la possibilitat de marcar un gol i *n* és l'acció actual:

$$AV = ((S_{n+1} - S_n) + (S_{n+2} - S_{n+1})) \ / \ 2$$

3. Calculem per cada acció el *deffensive value* (DV) amb la següent equació, on C és la possibilitat de concedir un gol i *n* és l'acció actual:

$$DV = ((C_{n+1} - C_{pn}) + (C_{n+2} - C_{n+1})) \ / \ 2$$

4. Calculem el *total_value*(TV):

$$TV = AV + DV$$

Amb el *total_value* calculat ara falta normalitzar totes les valoracions perquè siguin comparables, ja que òbviament hi ha jugadors amb més minuts que d'altres i aquests poden tenir més oportunitats per millorar o empitjorar la valoració o es poden donar altres anomalies. Per fer-ho es calcula el que anomenem *rating*, que consisteix en extrapolar la valoració obtinguda a 90 minuts, la duració estàndard d'un partit de futbol. Ho fem amb la següent equació:

$$rating = (90/minutes\_played) * TV$$

Tal i com s'intueix, *minutes_played* són els minuts jugats pel futbolista i TV el *total_value.* Una vegada calculat el *rating* podem donar per finalitzada la tasca del càlcul de la valoració dels jugadors.

4.5 Valoració porters

Els porters s'han mesurat de forma diferent, si bé aquesta equació resulta útil per els jugadors de camp en els porters comporta certs problemes. El principal problema trobat és que sobrevalora les accions de porters i més en concret dels porters que juguen en equips denominats petits i amb tendència a defensar en camp propi. Els porters d'aquests equips es torben en situacions on els jugadors rivals xuten a porteria en moltes ocasions però des de posicions de poc perill i on el xut sovint no es dirigeix ni a porteria degut a la quantitat de jugadors que es troben entre la porteria i el llançador. Aquestes situacions faciliten que el porter no intervingui en la jugada i per tant no el pugui perjudicar el fet que la pilota pugui acabar dins la porteria i en canvi, al sacar de porteria i ser equips que busquen el desplaçament en llarg, millorin molt les probabilitats de marcar un gol i disminueixin les de concedir. Per tant, els porters reben puntuacions molt elevades per jugades on moltes vegades ni tan sols han intervingut, més enllà de sacar de porteria. Per això es decideix que per els porters s'utiltizarà una medició diferent. Cal remarcar que aquesta distinció només és possible en la posició del porter i que si qualsevol altra posició

es trobés amb sobrevaloracions d'aquest tipus no es faria aquest canvi de medició. Amb els porters es creu encertat és una posició especial dins del terreny de joc que duu a terme una funció molt limitada i cenyida a un aspecte en concret, no concedir gols i que a més la seva mobilitat dins del camp també es reduïda ja que no acostumem a veure porters fora de la seva àrea. Amb aquests motius ens centrem doncs en la funció especial del porter, no encaixar gols i valorarem la seva aportació com la diferència entre els xG rebuts i els gols rebuts realment.

$$TV = xG_{rebuts} - G_{rebuts}$$

4.6 Aplicatiu

Per ajudar en l'obtenció de les puntuacions dels jugadors s'ha creat un aplicatiu que en retorna les puntuacions depenent de diversos criteris que en seleccioni l'usuari. Tenint ja guardats en arxius *.pkl* per cada partit els valors de cada acció, l'aplicatiu duu a terme l'identificació de cada acció amb el seu jugador i el càlcul del rating.

Així, quan l'usuari executa l'arxiu *valuing_players* es desplega un menú que li permet escollir entre les diferents competicions disponibles o entre totes, si s'ha seleccionat aquesta última opció s'ha d'elegir entre seleccionar els jugadors masculins o femenins i en qualsevol cas triar si es volen veure els jugadors de camp o els porters. L'últim pas és triar quin és el nombre de jugadors que es volen veure, introduint el nombre per consola o '*all*' si es volen veure tots.

5 RESULTATS

En aquesta secció es mostren com a resultats diferents taules amb les diferents valoracions dels jugadors depenent dels paràmetres d'entrada que se li explicitin a l'aplicatiu que s'ha creat per extreure resultats.

En primer lloc es mostra una taula amb els 10 jugadors masculins amb la màxima valoració i que han jugat més de 900 minuts (10 partits) en total.

Jugador	Rating	Posició
Lionel Andrés Messi	2.822	Right Wing
Ronaldo de Assis	2.501	Left Wing
Thierry Henry	2.023	Left Wing
Samuel Eto'o	1.923	Right Wing
Neymar da Silva	1.903	Left Wing
Zlatan Ibrahimovic	1.761	Center Forward
Luis Alberto Suárez	1.704	Center Forward
Sergio Leonel Agüero	1.697	Center Forward
Cristian Tello	1.676	Center Defensive Midfield
Robert Pirès	1.641	Left Center Midfield

Taula 3. Resultats del 10 jugadors masculins amb màxima valoració i que han jugat més de 900 minuts en totes les competicions.

La taula que es mostra a continuació mostra les 10 jugadores femenines amb la màxima valoració i que han jugat més de 900 minuts (10 partits) en total.

Jugadora	*Rating*	Position
Lynn Williams	3.074	Right Wing
Vivianne Miedema	2.823	Center Forward
Jordan Nobbs	2.56	Right Center Midfield
Megan Rapinoe	2.104	Left Attacking Midfield
Lindsey Horan	2.104	Left Center Midfield
Chloe Kelly	2.057	Left Wing
Martha Thomas	1.953	Center Forward
Fara Williams	1.863	Left Center Midfield
Débora Cristiane	1.734	Left Center Midfield
Léa Le Garrec	1.651	Left Center Midfield

Taula 4. Resultats de les 10 jugadores femenines amb màxima valoració i que han jugat més de 900 minuts en totes les competicions.

Com podem comprovar en les dues taules la majoria de jugadors amb millors puntuacions pertanyen a posicions d'atac, el que resulta comprensible, per això a continuació mostrem també dues taules amb els 10 millors jugadors i jugadores respectivament, excloent les posicions de Center Forward, Right Wing i Left Wing, les posicions atacants, per comprovar com es comporta el sistema de puntuació amb jugadors que no acostumen a marcar tants gols.

Jugador	Rating	Posició
Cristian Tello	1.676	Center Defensive Midfield
Robert Pirès	1.641	Left Center Midfield
Gareth Bale	1.531	Center Attacking Midfield
Philippe Coutinho	1.413	Left Center Midfield
Jesús Navas	1.274	Right Back
Éver Banega	1.24	Left Center Midfield

Anderson Luís	1.167	Center Attacking Midfield
Toni Kroos	1.014	Left Center Midfield
Francesc Fàbregas	1.004	Left Center Midfield
Daniel Parejo	0.999	Left Center Midfield

Taula 5. Resultats dels 10 jugadors masculins amb més 900 minuts, màxima valoració i que no juguen en posicions atacants en totes les competicions.

Jugador	Rating	Posició
Jordan Nobbs	2.560	Right Center Midfield
Megan Rapinoe	2.104	Left Attacking Midfield
Lindsey Horan	2.104	Left Center Midfield
Fara Williams	1.862	Left Center Midfield
Débora Cristiane	1.733	Left Center Midfield
Léa Le Garrec	1.651	Left Center Midfield
Jill Roord	1.495	Left Center Midfield
Katie Zelem	1.385	Left Center Midfield
So-yun Ji	1.327	Center Attacking Midfield
Kim Little	1.321	Right Center Midfield

Taula 6. Resultats de les 10 jugadores femenines amb més de 900 minuts, màxima valoració i que no juguen en posicions atacants, en totes les competicions disponibles.

Com s'ha comentat en l'apartat anterior, els porters es valoren de forma diferent, a continuació n'exposem les taules amb els porters d'ambdós sexes, com en els resultats anteriors.

Jugador	Rating	Posició
Thibaut Courtois	0.548	Goalkeeper
Marc-André Ter Stegen	0.415	Goalkeeper
Idriss Carlos Kameni	0.413	Goalkeeper
Antonio Rodríguez	0.313	Goalkeeper
César Sánchez	0.193	Goalkeeper
Jens Lehmann	0.186	Goalkeeper
Claudio Andrés Bravo	0.152	Goalkeeper
Víctor Valdés	0.089	Goalkeeper
Keylor Navas Gamboa	0.085	Goalkeeper
Andrés Palop	0.054	Goalkeeper

Taula 7. Resultats dels 10 porters masculins amb més 900 minuts i màxima valoració en totes les competicions .

Jugador	Rating	Posició
Sari van Veenendaal	0.878	Goalkeeper
Katelyn Rowland	0.771	Goalkeeper
Sophie Baggaley	0.699	Goalkeeper
Courtney Brosnan	0.575	Goalkeeper
Kirstie Levell	0.548	Goalkeeper
Rut Hedvig Lindahl	0.523	Goalkeeper
Anne Moorhouse	0.441	Goalkeeper
Mary Alexandra Earps	0.403	Goalkeeper
Ellie Roebuck	0.387	Goalkeeper
Pauline Peyraud Magnin	0.211	Goalkeeper

Taula 8. Resultats de les 10 porteres femenines amb més 900 minuts i màxima valoració en totes les competicions .

6 CONCLUSIONS

Una vegada aconseguits i examinats els resultats mostrats en l'apartat anterior en podem extreure algunes conclusions. Al contrari que els resultats dels xG que podíem extreure'n mètriques i comparar-les amb altres models de predicció amb aquests resultats no podem ferne una comparació numèrica directa amb altres models.

Analitzant l'objectiu principal del projecte, valorar el rendiment dels futbolistes durant un o diversos partits, i si mirem les taules, tant masculina com femenina, amb els 10 millors jugadors, trobem que la majoria són jugadors de posicions atacants. Aquest és un bon símptoma ja que els jugadors de posicions atacants són els que més gols fan i els que no generen directament un nombre tan gran de gols, juguen en posicions on la seva funció principal és assistir aquells als quals la seva funció sí que és marcarlos. Així doncs podem comprovar com en els primers llocs dels rànquings hi trobem futbolistes com Leo Messi, Thierry Henry, Vivianne Miedema o Megan Rapinoe, tots ells considerats com a jugadors d'èlit durant les seves respectives carreres, si és que ja han acabat. A més, en l'apartat de jugadors masculins trobem que 7 dels 10 millors jugadors juguen o han jugat en algun moment en el FC Barcelona durant els anys on aquest club ha aconseguit grans resultats esportius, el que també resulta un bon símptoma de que el mètode proposat funciona bé.

També s'ha volgut comprovar quin és el comportament del sistema de valoració en jugadors que no són de posicions atacants, amb la intenció de comprovar que les puntuacions, si bé valoren més els atacants per les qüestions comentades anteriorment, no difereixen molt entre jugadors i no hi ha una excessiva sobrepuntuació en els juga-

dors que acostumen a marcar gols. En aquest aspecte el sistema es comporta de forma satisfactòria, com podem comprovar en les taules 4 i 5, on es mostra com molts i moltes de les futbolistes que apareixien en les primeres taules desapareixen però tot i així les diferències de puntuació resulten acceptables i les diferències de magnituds no són excessivament grans com per fer-ne impossible la comparació.

Respecte els porters, tot i haver hagut d'usar un sistema de medició diferents respecte el dels jugadors de camp, els resultats també resulten acceptables i es demostra que el canvi proposat de medició és adequat ja que els porters que ocupen les parts superiors dels rànquings són considerats dels millors en la seva posició.

La conclusió final doncs és que amb les dades i les eines amb les quals s'ha contat tant a l'inici com en el transcurs del projecte s'ha aconseguit un sistema suficientment robust com per ser vàlid a l'hora de valorar el rendiment dels futbolistes, acomplint d'aquesta manera el que era l'objectiu principal del projecte. El fet d'aconseguir l'objectiu principal ve donat perquè, com s'ha vist en apartats anteriors, s'han complert també els objectius intermitjos, aconseguint primer una bona extracció de dades, necessària per al següent objectiu, generar un model robust per el càlcul dels xG, que tal i com s'esmenta en l'apartat 4.3.4 es considera també com a objectiu complert. A partir d'aquest sistema de valoració es podrien implementar millores si es contés amb més dades, com per exemple les de tracking, que mostren en tot moment el recorregut de cada futbolista pel terreny de joc i ajudaria no només a millorar la precisió del model d'xG, sinó que també podria ser útil per millorar el sistema de valoració. Per exemple, no només valorant com ha millorat un jugador les possibilitats del seu equip d'anotar un gol sinó que es podrien comprovar les altres opcions a l'abast del jugador i es podria comparar el resultat de la decisió presa pel jugador amb els possibles resultats de les altres opcions al seu abast.

En resum, s'ha complert l'objectiu proposat a l'inici del projecte, generant un sistema de valoració dels jugadors vàlid per si sol i que també pot resultar com a base per a sistemes més sofisticats.

AGRAIMENTS

Vull agrair a la meva família i a la meva parella el suport que m'han aportat en tot moment no només durant el transcurs d'aquest projecte sinó durant tots aquests anys de formació en l'enginyeria informàtica.

També vull agrair a en Jordi Serra la seva ajuda com a tutor d'aquest projecte durant tot el seu desenvolupament.

Moltes gràcies.

BIBLIOGRAFIA

[1] StatsPerform (2020, May. 27). *Advanced mètrics* [Online]. Availabe: https://www.statsperform.com/opta-analytics/ (Accedit Maig 2021)

[2] M. Van der Werf (2020, Oct. 18) *An overview of Advanced metrics in Football Analyisis*. [Online] Available: https://medium.com/@max_vander_werf/an-overview-of-advanced-metrics-in-football-analysis-4e75fd82bef8 (Accedit Juny 2021)

[3] B. McAleer. (2013, May. 2) *Match Focus: Most Shots in a Game Across Europe.* [Online] Available: https://it.whoscored.com/Articles/2uQ77vcYBUyn3oTbK1CYZQ/Show/Match-Focus-Does-A-Higher-Shot-Count-Result-In-More-Goals (Accedit Maig 2021)

[4] StatsBomb. (2018, June. 5) *StatsBomb: Open Data*. [Online] Available: https://github.com/statsbomb/open-data

[5] T. Knutson (2018, Aug. 3) *New Data, New StatsBomb radars*. [Online] Available: https://statsbomb.com/2018/08/new-data-new-statsbomb-radars/

[6] T. Kharrat, IG McHale, JL Peña. (2020) "Plus-minus player ratings for soccer", *European Journal of Operational Research,* [Online]. Available: https://arxiv.org/pdf/1706.04943.pdf

[7] IG McHale, PA Scarf, DE. Folker (2012, April). "On the development of a soccer player performance rating system for the English Premier League". *Informs Journal on Applied Analytics* [Online] vol. 42, Issue 4, pp. 329-420. Available: https://pubsonline.informs.org/doi/abs/10.1287/inte.1110.0589

[8] S. Rudd. 2011. *A framework for tactical analysis and individual offensive production assessment in soccer using Markov chains*. New England Symposium on Statistics in Sports [video] Available: https://www.metacafe.com/watch/7337475/2011_nessis_talk_by_sarah_rudd/

[9] T. DeCroos, J. Van Harren, L. Bransen, J. Davis, "Actions Speak Louder than Goals: Valuing Players Actions in Soccer" in KDD'19: Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining, July 2019, pp. 1851 – 1861. [Online] Available: https://arxiv.org/pdf/1802.07127.pdf

[10] Python 3.8 (https://docs.python.org/3.8/) (accedit Juny 2021)

[11] Pandas (https://pandas.pydata.org/) (accedit Juny 2021)

[12] JSON (https://docs.python.org/3/library/json.html) (accedit Juny 2021)

[13] SciPy (https://www.scipy.org/docs.html) (accedit Juny 2021)

[14] Math (https://docs.python.org/3/library/math.html) (accedit Juny 2021)

[15] SciKit – Learn (https://scikit-learn.org/stable/) (accedit Juny 2021

[16] TensorFlow (https://www.tensorflow.org/) (accedit Juny 2021)

[17] Keras (https://keras.io/) (accedit Juny 2021)

[18] Joblib (https://joblib.readthedocs.io/en/latest/) (accedit Juny 2021)

[19] D.C. Liu, J. Nocedal. , 1989. "On the limited memory BFGS method for large scale optimization". *Mathematical Programming* vol 42. pp. 502-528.

[20] P. Ramachandran, B. Zoph, QV. Le, 2017. "Searching for Activation Function". [Online] Available: https://arxiv.org/pdf/1710.05941.pdf

[21] E. O'Brien, "Comparing expected goals models". [Online] Available: https://eoin-obrien.com/2020/05/12/comparingexpected-goals-models/

[22] openGoal. (2020, July). "*Evaluating the performance of Convolutional Neuran Network based xG Models*". [Online] Available: https://www.opengoalapp.com/xg-with-cnns-full-study .

[23] Gradient Boosting Classifier (https://scikit-le-arn.org/stable/modules/generated/sklearn.ensemble.GradientBoostingClassifier.html)

[24] J. Davis, P. Robberechts, (2020, May. 12) *How Data availability affects the ability to learn good xG models*. [Online] Available: *https://dtai.cs.kuleuven.be/sports/blog/how-data-availability-affects-the-ability-to-learn-good-xg-models*

APÈNDIX

A1. TIPUS D'ACCIONS

- Dribble
- Clearance
- Dispossessed
- Interception
- Tackle
- Save
- Shot
- Foul
- Pass
- Own Goal