Cas Kaggle: Regressió de la mètrica "Expected Goals"

Joan Tubert I Mascort / Eloi Mercader Morillas

1673326 / 1666675

Abstract—

A quest estudi aborda el desenvolupament d'un model de regressió per a estimar la probabilitat de gol en xuts de futbol, basant-se en la mètrica coneguda com Expected Goals (xG). Utilitzant una base de dades composta per més de 80.000 registres, s'analitzen diverses variables descriptives del context del xut, incloent la distància i l'angle respecte a la porteria, el tipus i la part del cos amb la qual s'ha realitzat el xut, així com la presència de la pilota en suspensió o d'una porteria buida. Addicionalment, s'incorpora informació derivada d'un freezeframe, que detalla la posició, l'angle i altres característiques dels jugadors en el moment del xut. En aquest treball hem desenvolupat dos models: un que quantifica la pressió defensiva mitjançant el càlcul d'un coeficient ideat per nosaltres i un altre que afegeix al model la distància i l'angle de cada jugador rival respecte al jugador que remata. Finalment, ens proposem que, donada una imatge d'un xut a porteria, el model sigui capaç de transcriure-la a les característiques del nostre model mitjançant tècniques de visió per computador, per després calcular la probabilitat de gol. Malgrat els esforços, aquesta part del projecte no ha aconseguit assolir els resultats esperats, però representa una base prometedora per a treballs futurs.

1. Introduction

La mètrica *expected goals*, relativament recent, ha guanyat rellevància tant en l'àmbit acadèmic com en el professional, gràcies a la seva capacitat per a quantificar les oportunitats de gol d'una manera objectiva i basada en dades. Recentment, empreses tecnològiques com Microsoft han desenvolupat els seus propis mètodes d'anàlisi d'*Expected Goals*, i la seva aplicació pràctica ja s'ha integrat a competicions d'alt nivell com La Lliga. Tot i els avenços recents, l'anàlisi detallada de la posició dels jugadors rivals i propis en relació amb el xut encara és una àrea d'investigació que presenta oportunitats per a millores significatives.

Amb aquest estudi, no només pretenem obtenir el resultat més precís, sinó també utilitzar aquest projecte com a porta d'entrada al camp del reconeixement d'imatges.

2. Metodologia

2.1. Data

El dataset utilitzat per a aquest estudi ha estat extret del repositori públic de StatsBomb, accessible a través del següent enllaç: https://github.com/statsbomb/open-data. Aquest repositori conté dades detallades de partits de futbol en format `.json`, agrupats sota un dataset anomenat *events*, que inclou tots els esdeveniments que poden ocórrer en un partit, com ara passades, xuts, gols o targetes vermelles.

Per al nostre treball, vam haver d'extreure únicament els xuts de tots els partits presents en aquest conjunt de dades. Aquest procés de filtratge es va dur a terme seleccionant exclusivament els esdeveniments identificats com a *shots* al camp `type` dels fitxers `.json`. A més, en aquest mateix pas es va aplicar *Feature Engineering* per a calcular i afegir al dataset la característica de l'angle del xut respecte a la porteria. Aquesta característica, crucial per al nostre model, es va calcular a partir de la posició del jugador que xuta agafant com a referència el centre de la porteria.

El resultat d'aquest procés es va guardar en un fitxer `shots_data.csv`.

2.2. Exploratory Data Analysis (EDA)

Durant l'anàlisi exploratòria de dades (*Exploratory Data Analysis*), hem investigat la distribució i les relacions entre les variables presents a la base de dades, identificant patrons significatius que poden influir en la predicció de la probabilitat de gol (*statsbomb_xg*).

La base de dades resultant després del filtratge i el *Feature Engineering* conté un total de 15 atributs i el target a predir que és statsbomb_xg. D'aquests, trobem variables numèriques, com la distància i l'angle respecte a la porteria, així com variables categòriques, com el tipus de xut o si la porteria estava buida. En total, tenim 7 atributs categòrics, 3 binaris i 5 numèrics.

El *target* de la nostra anàlisi, *statsbomb_xg*, és una variable numèrica contínua que representa la probabilitat de gol assignada per StatsBomb, amb valors que oscil·len entre 0 i 1. Aquesta variable és clau en la nostra investigació, ja que volem desenvolupar un model que sigui capaç d'estimar-la a partir de les altres característiques disponibles.

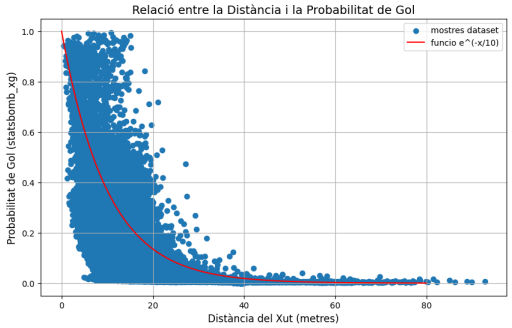

Figure 1. Relació entre la distància i la probabilitat de gol

Durant l'EDA, un dels patrons més destacats observats és la relació inversa entre la distància respecte a la porteria i la probabilitat de gol (*statsbomb_xg*). Com es mostra a la Figura 1, aquesta relació segueix una tendència exponencial decreixent, on xuts des de distàncies curtes tenen una probabilitat de gol significativament més alta. Per tant vam aplicar la següent transformació d'entrada a les dades:

$$\text{distància} = -\log(\text{distància}) \tag{1}$$

A més, hem analitzat específicament els penals, ja que representen un tipus de xut amb característiques úniques i una alta probabilitat de gol.

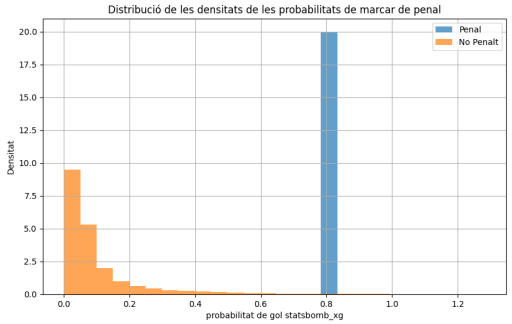

Figure 2. Distribució de les densitats de les probabilitats de marcar de penal.

Aquest gràfic mostra clarament que els penals són una excepció dins de la distribució general, amb una concentració elevada de probabilitats properes a 1.

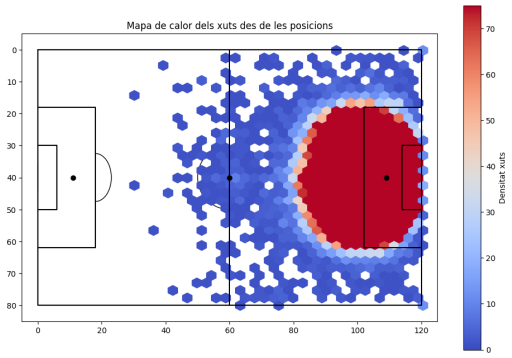

Figure 3. Mapa de calor de la zona dels xuts.

Aquest gràfic mostra, com era evident, que la majoria de remats a porteria que constitueixen el dataset s'han realitzat dins l'àrea.

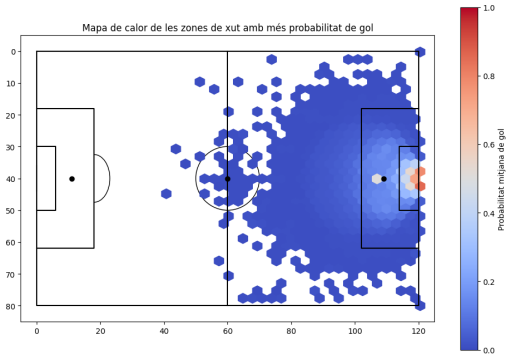

Figure 4. Mapa de calor de la probabilitat de gol per zones.

Com podem observar, com més aprop més alta és la probabilitat de gol. També podem notar que la zona del punt de penal té una probabilitat considerablement alta, tot i així és inferior a l'especificada a la Figure 2

2.3. Model Dual

Durant el desenvolupament del nostre estudi, ens hem adonat que la característica *freeze_frame* resulta altament carregosa per a l'anàlisi i modelització. Aquesta característica conté una gran quantitat d'informació detallada sobre cadascun dels jugadors rivals en el moment del xut, incloent dades com el nom del jugador, la posició en què juga o l'equip al qual pertany. No obstant això, l'única informació rellevant per al nostre model és la localització espacial dels jugadors rivals. A més, *freeze_frame* es presenta en un format complex, estructurat com a diccionaris de diccionaris, fet que complica el seu tractament directe.

Per abordar aquest problema i optimitzar el modelatge, hem dissenyat dos enfocaments alternatius que tractarem simultàniament al llarg del treball i guardarem en les bases de dades df i $df2$:

1. Substitució de *freeze_frame* per un coeficient de pressió. Hem ideat un coeficient que reemplaça la característica *freeze_frame*, modelitzant la pressió exercida pels jugadors rivals sobre el rematador. Aquest coeficient es basa en la distància entre el punt de xut i cadascun dels jugadors rivals, considerant un pes exponencial que decreix amb la distància. D'aquesta manera, les distàncies grans tenen molt menys pes que les distàncies curtes. Per a regular la funció, hem incorporat una constant arbitrària d_0, que pot ser ajustada posteriorment com un hiperparàmetre del model.

La funció dissenyada per calcular aquest coeficient és la següent:

$$\text{Pressió} = \sum_{i=1}^{11} \frac{e^{-\frac{d_i}{d_0}}}{11}$$

On:

- d_i és la distància del jugador rival i al rematador.
- d_0 és un paràmetre de referència utilitzat per regular la funció.

Aquest enfocament permet obtenir un valor comprès entre 0 i 1:

- Si la pressió fos màxima, és a dir, si els 11 jugadors rivals estiguessin a 0 metres del rematador, el coeficient seria 1.
- Si la pressió fos nul·la, amb els jugadors rivals a una distància infinita del rematador, el coeficient seria 0. Farem referència a aquest primer model amb el terme $coeficient$.

2. Ampliació del model amb noves característiques. Com a alternativa, hem proposat afegir 22 noves característiques al model per substituir *freeze_frame*. Aquestes característiques inclouen:

- Les posicions en coordenades de tots els 11 jugadors rivals en el moment del xut.
- Els angles que formen cadascun dels rivals amb el jugador rematador, considerant el porter com a punt de referència.

Aquest enfocament permet al model capturar més detalladament la informació espacial de la situació de joc. Per altra banda es farà referència a aquest model amb el concepte ampliació.

2.4. Preprocessing

El preprocessament (*Preprocessing*) de les dades va ser una fase clau per a garantir la qualitat i consistència del dataset final. Aquesta etapa va incloure diverses tasques:

- **Gestió de Nans:**
  Durant el preprocessament de dades, un dels problemes en el model d'$ampliaci\approx$ va ser la presència de valors *NaN* en les característiques derivades de la variable *freeze_frame*. Aquesta característica, que descriu la posició dels jugadors rivals en el moment del xut, no sempre conté informació completa sobre els 11 jugadors rivals. Això es deu al fet que, en moltes situacions, la càmera del partit no registra certs jugadors que es troben molt lluny de la zona d'acció. Per solucionar aquest problema, vam implementar un tractament dels valors *NaN* basat en els següents passos:
  1. Vam determinar un llindar per despreciar els jugadors que es trobessin a una distància superior a aquest. És a dir, els jugadors rivals que es trobessin més enllà de dos metres del rematador en la **direcció contrària** a la porteria es van ignorar.
  2. Per a tractar els valors *NaN* (ja sigui per la no existència d'un jugador registrat o per la seva posició més enllà de la distància rellevant), es van assignar valors nuls (és a dir 0) a les característiques corresponents:
- **Transformació de variables:** Es van normalitzar i escalar les variables numèriques per a optimitzar el rendiment dels models de regressió.
- **Selecció de característiques (*Feature Selection*):** Durant aquest procés, es van identificar i eliminar diverses característiques que no aportaven informació significativa al model o podien provocar problemes de colinealitat. Les característiques eliminades van ser les següents:
  1. `player`: Aquesta característica descriu el nom del jugador que realitza el xut. Igual que en altres models, com el desenvolupat per Microsoft, hem decidit no valorar aquesta característica.
  2. `event_id`: Aquesta característica té una correlació molt baixa amb el *target* i, per tant, no és representativa per al model.
  3. `team`: Igual que amb la característica `player`, no considerarem l'equip al que pertany el rematador.
  4. `period`: Aquesta característica mostra correlació amb la característica `minute`, fet que podria introduir problemes de colinealitat. Per tant, s'ha eliminat.

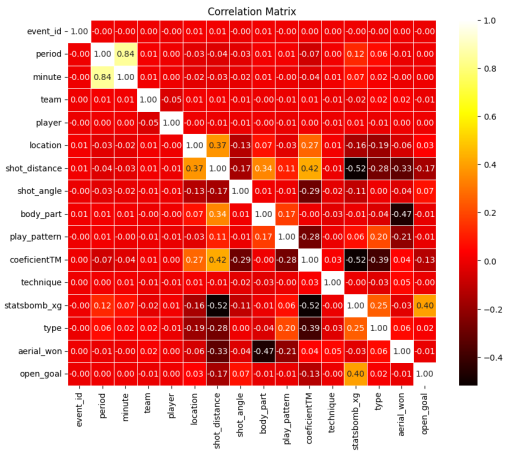

Figure 5. Matriu de correlació.

Aquí podem observar la matriu de correlació entre les diverses característiques del dataset en el cas d'haver utilitzat un coeficient de pressió. Un clar indici de la seva importància és el valor de 0,52 de correlació amb els statsbomb_xg. Això és molt significatiu, ja que, per predir els expected goals, aquesta característica és la més rellevant juntament amb la distància del xut.
Després d'aquestes dues, les característiques més essencials per a dur a terme la regressió són open_goal i type. Aquestes ens

indiquen si el xut es realitza a porteria buida i el tipus de jugada que acaba en xut (jugada oberta, penal, etc.). Per lògica, podríem esperar la rellevància d'aquests estadístics, però és especialment interessant observar com l'alt valor del coeficient de pressió evidencia que es tracta d'una característica clau. Alhora, també demostra que aquest coeficient està prou ben definit per exemplificar de manera senzilla tota la informació que teníem en el freeze_frame.

3. SELECCIÓ DE MODEL

3.1. Regressió lineal i Polynomial Features de grau 2

3.1.1. Resultats

Table 1. Comparació dels models.

Model	R^2 Train	R^2 Test
RL (coeficient)	0.5823	0.5827
PF (coeficient)	0.8318	0.8388
RL (ampliació)	0.5659	0.5672
PF (ampliació)	0.8318	0.8368

Nota: RL representa La Lineal Regression i PF l'aplicació de Polynomial Features

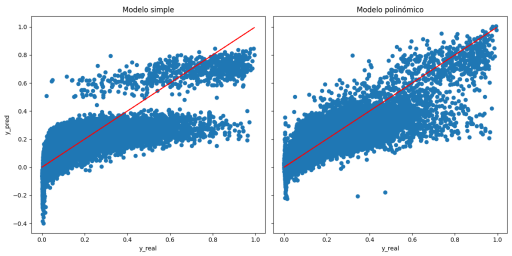

Figure 6. predicció vs valor real de la Regressió Lineal vs Polynomial Features

Com podem veure en el subplot anterior, la distribució dels punts de Polynomial Features s'ajusta molt més a la recta y=x, és a dir, al cas ideal. Això demostra la no linealitat d'algunes característiques i que, com era d'esperar, el model lineal és massa simple.

3.2. RANSAC

3.2.1. Intuïció

Donat que el nostre dataset és molt extens, amb dades extretes de múltiples partits de futbol, és probable que hi hagi presència d'*outliers* en algunes de les característiques.
Per validar aquesta hipòtesi, hem realitzat una anàlisi exploratòria de les dades mitjançant un *boxplot*.

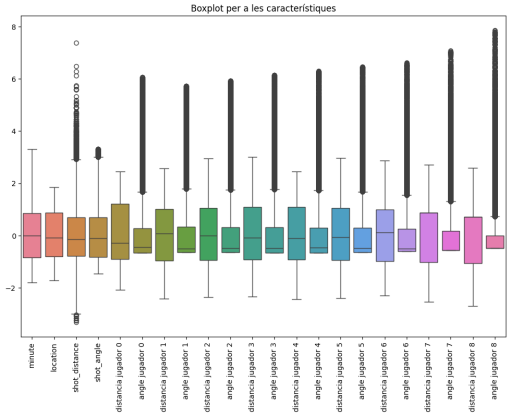

Figure 7. Boxplot per visualitzar outliers.

Com podem veure en aquest Boxplot tenim una quantitat considerable d'outliers als angles dels rivals. Els outliers es mostren com a punts situats fora de les "caixes" (els límits de la caixa són el quartil 1 i quartil 3, i les línies de la caixa mostren el rang interquartil). Això justifica l'ús del model *RANSAC* (*Random Sample Consensus*), que és robust davant d'outliers. Aquest model selecciona iterativament subconjunts de dades que tenen un límit màxim d'error permès quantificat per un llindar, és a dir, treballa amb els inliers.

3.2.2. Resultats

Table 2. Comparació dels models.

Model	R^2 Train	R^2 Test
RANSAC (coeficient)	0.8314	0.8380
RANSAC (ampliació)	0.8291	0.8345

Nota: Taula de resultats model RANSAC

3.2.3. Resultats

3.3. Regressió logística

3.3.1. Funcionament

En aquest cas, la regressió logística s'utilitza per modelar la probabilitat que un xut es consideri un gol, basant-se en diverses característiques del dataset. Per a això, es transforma la variable dependent (statsbomb_xg) en una variable binària: si la probabilitat és superior al 50% (threshold = 0.50), es classifica com a gol (1), i en cas contrari com a no-gol (0).

Un cop ajustat, el model calcula les probabilitats predites que un xut sigui un gol tant per al conjunt d'entrenament com per al de prova.

3.3.2. Resultats

Table 3. Comparació dels models.

Model	R^2 Train	R^2 Test
RL (coeficient)	0.68449	0.69987
RL (ampliació)	0.67472	0.68035

Nota: RL representa La Logistic Regression

3.3.3. Interpretació de Resultats

Com podem observar, en comparació amb altres resultats obtinguts, aquests són força mediocres. De fet, només són lleugerament millors que els de la regressió lineal, que és el model més simple. Això es pot concloure pel fet que l'associació de valors reals a binaris en el context futbolístic no té gaire sentit. Per tant, la regressió logística no és la millor opció per modelar la mètrica en qüestió.

3.4. Cerca d'hiperparàmetres amb models avançats

Per obtenir els millors resultats dels models considerats, hem seguit una metodologia en dues fases per optimitzar els hiperparàmetres. Aquesta estratègia permet assegurar-nos que cada model s'ajusti de manera òptima a les característiques del dataset. Els models emprats són: Decision Trees, Random Forest i XG Boost

3.4.1. Selecció del millor valor de k

En primer lloc, vam realitzar una cerca per determinar el millor valor de k per a la validació creuada (*cross-validation*) de cada model. Es van provar diferents valors de k, incloent 2, 3, 5, 7, 10 i 20.

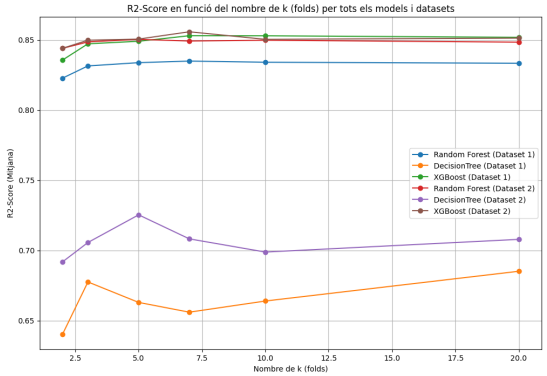

Figure 8. R^2 en funció del nombre de k per diferents models i datasets (ampliació i coeficient).

La Figura 8 mostra la relació entre k i el R^2 per als models estudiats, il·lustrant com els valors òptims de k poden variar segons el model i el dataset.

3.4.2. Cerca d'hiperparàmetres amb GridSearchCV

Un cop determinat el millor valor de k per a cada model, es va realitzar una cerca exhaustiva d'hiperparàmetres mitjançant *GridSearchCV* utilitzant només un 20% del dataset per tal de reduir el cost computacional.

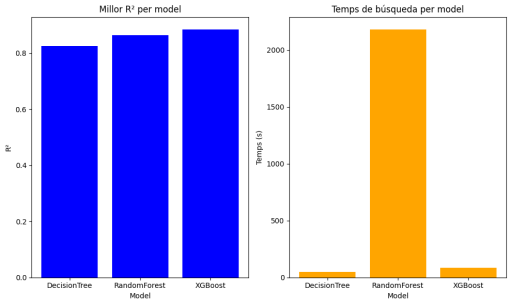

Figure 9. Subplot dels R scores i els temps dels 3 models amb *coeficient*

3.5. XGBoost

3.5.1. Funcionament

El model XGBoost (Extreme Gradient Boosting) és un algorisme basat en arbres de decisió que utilitza tècniques de gradient boosting per obtenir prediccions precises. Funciona construint successivament una sèrie d'arbres de decisió, on cada arbre nou corregeix els errors del model anterior. Els pesos dels exemples mal predits s'ajusten perquè el model enfocat als errors millori.

3.5.2. Resultats

Table 4. Comparació dels resultats (R^2 i MSE).

Model	R^2 Train	R^2 Test	MSE Train	MSE Test
XGB (coeficient)	0.9133	0.8848	0.001935	0.002568
XGB (ampliació)	0.9221	0.8942	0.001738	0.002357

Nota: XGB representa el XGBoost.

3.5.3. Anàlisi de resultats

Com podem observar, no hi ha dubte que XGBoost és el millor model per a la regressió de la mètrica Expected Goals. A més, els resultats obtinguts superen àmpliament les nostres expectatives inicials. També hem inclòs en aquest cas el valor de la mètrica MSE (Mean Squared Error), la qual presenta valors baixos, no només per la bona generalització del model sinó també perquè es calcula a partir dels quadrats de les diferències entre dues probabilitats (valors entre 0 i 1).

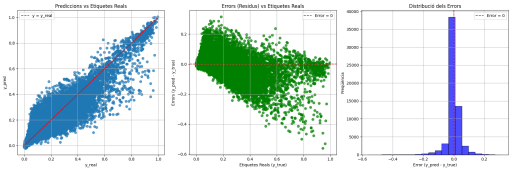

Figure 10. Subplot d'analisi del model.

4. Reconeixement d'imatges

Un cop seleccionat el model de regressió, ens vam proposar una extensió del treball que consisteix en utilitzar tècniques de visió per computador per tal de permetre al nostre model processar imatges directament. L'objectiu és que el model sigui capaç d'interpretar les característiques d'una imatge corresponent a un remat a porteria, extreure'n les dades del model i predir la probabilitat de gol de manera automàtica.

4.1. Detecció d'objectes

Per a la detecció d'objectes, hem utilitzat el model YOLOv8 (acrònim de *You Only Look Once*) i hem realitzat l'entrenament amb la seva versió small per a combinar eficiència i precisió.

4.1.1. Funcionament de YOLOv8:

YOLOv8 segueix el paradigma de "mirar només una vegada" (*You Only Look Once*), en el qual una única passada pel model permet identificar múltiples objectes dins d'una imatge. Aquesta arquitectura divideix una imatge en una graella i assigna a cada cel·la responsabilitat per predir si conté un objecte. Per a cada objecte detectat, el model genera:

- Una predicció de la **caixa delimitadora** (*bounding box*) que envolta l'objecte.
- Una classificació amb la **categoria** de l'objecte (per exemple, jugador, pilota o porteria).
- Un **nivell de confiança** associat a cada predicció.

4.1.2. Dataset utilitzat

Per entrenar el nostre model YOLOv8, hem utilitzat un dataset públic obtingut de la plataforma Roboflow, específicament el dataset disponible a https://universe.roboflow.com/soccer-m74r8/soccer-detectiin. Aquest dataset està pensat per a la detecció d'objectes en partits de futbol i conté un total de 2.549 imatges, dividides en tres conjunts: **Train Set:** 1.782 imatges (70%), **Validation Set:** 515 imatges (20%) i **Test Set:** 252 imatges (10%)..
Cada imatge del dataset ve acompanyada d'una carpeta de *labels* amb les etiquetes corresponents. Aquestes etiquetes inclouen informació sobre les caixes delimitadores (*bounding boxes*) i les classes dels objectes presents en la imatge. Les classes del dataset són: *person*, *sports_ball* i *porter*.

4.1.3. Entrenament del model:

```python
from ultralytics import YOLO
import matplotlib.pyplot as plt

model = YOLO('yolov8s.pt')


model.train(
    data="C:/Users/eloim/OneDrive/Escritorio
    /3 er carrera/apc/cas kaggle/open-data-
    master/open-data-master/Cas-kaggle-/
    soccer detection dataset/data.yaml",
    epochs=40,
    imgsz=640,
    batch=8,
    name='goal_detection'
)

model.val(data='data.yaml', test=True)
```

Code 1. funció càlcul coeficient

on **data.yaml** és un arxiu que contenia les rutes a una reducció del dataset ($train_reduced$, $test_reduced$, $valid_reduced$) i el el nombre d'**epochs** indica quantes vegades el model processa tot el conjunt d'entrenament per ajustar els seus pesos i minimitzar l'error.

4.1.4. Resultats preliminars:

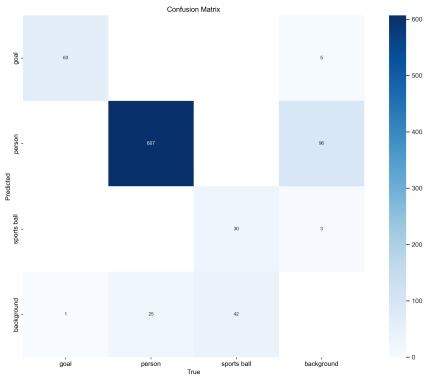

Figure 11. Matriu de confusió.

Aquesta matriu de confusió ens mostra la capacitat del model d'etiquetatge per associar els valors reals de porteria, persona i pilota amb les prediccions obtingudes. Els valors classificats com a background indiquen que l'objecte ha estat ignorat o que s'ha predit un objecte on no n'hi havia. És destacable que, en cap cas, el model confon objectes entre si, sinó que els únics errors estan associats al background. Aquests errors són comprensibles i segueixen certa lògica. Per exemple, hi ha una quantitat significativa de background classificat com a persona, cosa que podem atribuir a la presència de públic, àrbitres... Pel que fa a la manca de detecció, destaquen especialment els casos de pilotes no identificades, un fet previsible, ja que es tracta de l'objecte més petit.

En conclusió, els resultats són tan esperables com satisfactoris. El model ofereix una eina fiable per identificar objectes en diferents entorns i circumstàncies, cometent pocs errors "greus".

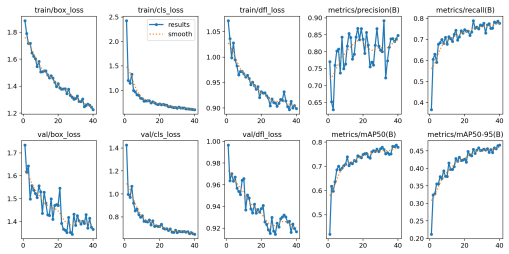

Figure 12. Resultats model

4.1.5. Exemple d'ús:

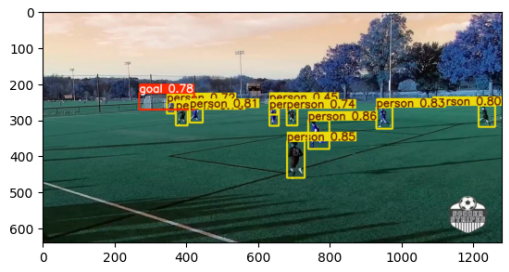

Figure 13. Resultat de l'algoritme entrenat amb els milllors pesos.

5. Treball futur

Després d'haver aconseguit detectar correctament els jugadors, la pilota i les porteries utilitzant YOLOv8, un dels objectius del nostre treball futur consisteix a transformar les coordenades de les imatges detectades a les dimensions del terreny de joc. Aquesta transformació és essencial per poder integrar de manera efectiva la informació visual obtinguda amb el nostre model de regressió.

5.1. Transformació de coordenades

Per dur a terme aquesta transformació, vam explorar diversos mètodes basats en la geometria projectiva i les relacions entre punts clau de la porteria detectats a les imatges i al camp de joc.

Homografia: Inicialment, vam intentar calcular la matriu d'homografia, que és la transformació projectiva necessària per relacionar l'espai imatge amb l'espai físic del terreny de joc. Aquesta matriu requereix disposar d'almenys quatre punts en correspondència entre ambdós espais. Tot i això, amb les dades proporcionades per YOLOv8 només disposàvem de tres punts clau de la porteria (el pal esquerre, el centre, i el pal dret), fet que va impossibilitar la implementació correcta d'aquest mètode.

Canvi de base: Posteriorment, vam explorar una aproximació basada en la matriu de canvi de base, utilitzant els tres punts detectats per YOLOv8 per a la porteria (extrem esquerre, centre i extrem dret de la caixa delimitadora) i els tres punts corresponents al terreny de joc real. Tot i això, els resultats obtinguts amb aquest mètode van ser incoherents, ja que sovint els punts calculats apareixien fora del camp o en ubicacions incorrectes.

5.2. Línies futures de treball

Actualment, ens hem quedat en aquest punt del treball, amb la necessitat de trobar una metodologia que permeti calcular la matriu d'homografia amb la informació disponible. Una possible línia de treball futur seria identificar parelles de punts que formen rectes paral·leles a l'espai imatge, però que, a l'infinit, convergeixin en un punt comú. Aquestes rectes paral·leles i els seus punts de fuga podrien constituir la informació addicional necessària per completar la matriu d'homografia requerida.

6. Conclusions

En aquest estudi, hem analitzat l'impacte de diferents aproximacions per a predir la probabilitat de gol (*Expected Goals, xG*), comparant els models basats en l'ampliació de característiques i en el coeficient de pressió defensiva. Un dels resultats més destacats ha estat la similitud en el rendiment d'aquests dos enfocaments, amb diferències pràcticament nul·les en les mètriques de R^2. Tenint en compte que el model del $coeficient$ no inclou els angles de cada rival respecte el remtador,obtenim dues possibles interpretacions:

- En primer lloc, podria indicar que l'angle que forma el defensor amb el rematador no té una influència significativa en la probabilitat de gol, tot i que aquesta conclusió pugui semblar una paradoxa.
- En segon lloc, és possible que el model basat en el coeficient hagi capturat de manera més efectiva les relacions entre les distàncies dels jugadors rivals i la probabilitat de gol, ajustant-se millor a les dades que no pas el model que considera les característiques de forma independent.

Un aspecte rellevant del nostre estudi és que no hem necessitat aplicar tècniques de regularització per evitar l'*overfitting*, ja que no hem detectat indicis d'aquesta problemàtica durant el desenvolupament del model. Hem conclós que això es deu a diverses raons:

En el cas del model basat en el coeficient, l'absència d'*overfitting* és perfectament lògica, ja que aquest modelitza un fenomen tan complex com la pressió defensiva que rep un jugador mitjançant una fórmula senzilla i compacta. Aquesta simplificació permet condensar gran part de la informació rellevant en una sola característica, reduint significativament la dimensionalitat del problema. Aquesta reducció fa que el model sigui menys susceptible a sobreajustaments, ja que evita treballar amb un nombre elevat de variables que podrien capturar patrons específics però irrellevants.

Per contra, és sorprenent que el model d'ampliació, tot i incorporar totes les distàncies i angles que formen els defensors amb el rematador, tampoc presenti *overfitting*. Una possible explicació per aquest fenomen resideix en la qualitat i extensió del dataset. Amb més de 87.000 xuts, el model disposa d'una quantitat de dades suficient per aprendre patrons generals sense dependre d'exemples concrets. Finalment, donat que els resultats d'ambdós models són tan similars, com que el model del coeficient és notablement més ràpid, ens decantaríem per aquest model en cas d'haver de triar.

7. Referències

- Dataset statsbomb_xg: https://github.com/statsbomb/open-data
- Dataset amb les imatges: https://universe.roboflow.com/soccer-m74r8/soccer-detectiin
- Informació del model Yolov8: https://docs.ultralytics.com/es
- Informació del model XGBoost: https://xgboost.readthedocs.io/en/stable/
- Informació matrius de confusió: https://datascience.recursos.uoc.edu/matrius-de-confusio/
- Informació matrius de confusió: https://foro.rinconmatematico.com/index.php?topic=110196.0
- Altres enllaços d'interès:
  - https://dspace.uib.es/xmlui/bitstream/handle/11201/162407/tfm_2021-22_MUSI_jfr165_5308.pdf?sequence=1&isAllowed=y
  - https://dspace.uib.es/xmlui/bitstream/handle/11201/162407/tfm_2021-22_MUSI_jfr165_5308.pdf?sequence=1&isAllowed=y
- I finalment hem usat el Caronte amb la informació del curs per a repassar continguts.