



UNIVERSITAT DE
BARCELONA

Treball final de grau

GRAU D'ENGINYERIA
INFORMÀTICA

Facultat de Matemàtiques i Informàtica
Universitat de Barcelona

CIÈNCIA DE DADES PER A LA BUNDESLIGA D'HANDBOL

Autor: Iván Martínez Villa

Directora: Dra. Laura Igual Muñoz

Realitzat a: Departament de Matemàtiques i Informàtica

Barcelona, 15 de gener de 2025

Abstract

As it stands, handball is still regarded as a secondary sport in terms of data analysis. This paper serves as an investigation into what productive analyses one is able to generate considering the scarcity of data available within this sport. For this purpose, several questions have been asked and we will solve them by carrying out an exploratory analysis of two data sets from the German handball league, the Bundesliga. In this way, we will be able to extract the information embedded in the data and address the posed questions through graphs and data visualisations. One dataset contains the individual statistics of the players in the league; the other, the matchday results. This data will be used to run different models with the objective of predicting a team's victory in a match. The usage of data science and machine learning models has been crucial in the making of this study.

Resumen

Actualmente el balonmano sigue siendo un deporte secundario en cuanto al análisis de datos. Este trabajo es una investigación sobre qué análisis productivos podemos hacer con la escasez de datos con la que se trabaja en este deporte. Para ello se han realizado diversas preguntas que resolveremos realizando un análisis exploratorio de dos conjuntos de datos de la liga alemana de balonmano, la Bundesliga. De esta forma podremos extraer la información presente en los datos y responder mediante gráficos y visualizaciones de datos a las preguntas planteadas. Uno de los conjuntos de datos dispone de las estadísticas individuales de los jugadores de la liga; el otro, dispone de los resultados de las jornadas. Estos datos se usarán para realizar diferentes modelos para intentar predecir la victoria de un equipo en un partido. Este estudio ha sido realizado mediante ciencia de datos y modelos de aprendizaje automático.

Resum

Actualment l'handbol continua sent un esport secundari en quant a l'anàlisi de dades. Aquest treball és una recerca sobre quines anàlisis productives podem fer amb l'escassetat de dades amb la qual es treballa en aquest esport. Per a això, s'han

realitzat diverses preguntes que resoldrem realitzant una anàlisi exploratòria de dos conjunts de dades de la lliga alemanya d'handbol, la Bundesliga. D'aquesta manera podrem extreure la informació present a les dades i respondre mitjançant gràfics i visualitzacions de dades les preguntes plantejades. Un dels conjunts de dades disposa de les estadístiques individuals dels jugadors de la lliga; l'altre, dels resultats de les jornades. Aquestes dades s'usaran per a realitzar diferents models per a intentar predir la victòria d'un equip en un partit. Aquest estudi ha estat realitzat mitjançant ciència de dades i models d'aprenentatge automàtic.

Índex

1	Introducció	1
1.1	Característiques de l'handbol	2
1.2	Objectius i preguntes formulades	4
1.3	Planificació del projecte	5
1.4	Estructura del projecte	6
2	Obtenció i descripció de les dades	7
2.1	Conjunt de dades sobre jugadors de la Bundesliga 2022-2023	7
2.1.1	Índex de Rendiment a l'Handbol (HPI)	8
2.2	Conjunt de dades de resultats de la Bundesliga 2021-22, 22-23, 23-24	11
3	Anàlisi exploratòria de les dades	12
3.1	Anàlisi de les dades dels jugadors	12
3.2	Anàlisi de les dades de la classificació	18
4	Tècniques d'aprenentatge automàtic per predir la victòria/no victòria d'un partit	23
4.1	Regressió logística	23
4.2	Naive Bayes	24
4.3	Màquines de vectors de suport	24
4.4	Boscos aleatoris	25
4.5	XGBoost	26
4.6	Model escollit per a la predicció per gols	27
4.7	Model escollit per a la predicció per l'Índex de Rendiment d'Handbol (HPI)	28
4.8	Reducció de característiques	29
5	Experiments i resultats	33
5.1	Tècnica de validació	33

5.2	Mètriques d'avaluació	33
5.3	Model de gols	34
5.3.1	Resultats del model	34
5.4	Model de l'Índex de Rendiment d'Handbol (HPI)	36
5.4.1	Preparació de les dades	36
5.4.2	Construcció i resultats del model SVM	36
5.5	Comparativa i anàlisi dels resultats	38
6	Conclusions	40
6.1	Treball futur	41
7	Bibliografia	42

1 Introducció

L'handbol, encara que és un esport amb una rica història i una base de seguidors apassionats, no gaudeix de la mateixa visibilitat i popularitat que altres esports com el futbol o el bàsquet. Aquesta diferència en popularitat es tradueix en una notable disparitat en la quantitat i qualitat de dades disponibles per a anàlisis. Mentre que en el futbol i el bàsquet existeixen extenses bases de dades i eines avançades per a l'anàlisi del rendiment, l'handbol encara s'enfronta a desafiaments significatius en la recopilació i explotació de dades detallades.

La falta de dades extensives en l'handbol presenta tant un desafiament com una oportunitat. D'una banda, limita el desenvolupament d'estratègies basades en dades i la possibilitat d'optimitzar el rendiment a través de l'anàlisi detallada. D'altra banda, obre un camp fèril per a la recerca i l'aplicació de la ciència de dades, amb el potencial de tancar aquesta bretxa i oferir avantatges competitius als equips que adoptin un enfocament més analític.

Aquest treball es proposa abordar aquest buit mitjançant la cerca i recopilació, anàlisi i interpretació de dades específiques de l'handbol. A través de tècniques de ciència de dades, es buscarà demostrar com fins i tot un conjunt de dades limitat pot proporcionar valuosos coneixements per a millorar la presa de decisions tàctiques i avaluar el rendiment dels jugadors.

L'estudi es centrarà en la resolució de diverses preguntes com: quines estadístiques individuals són les més diferencials a l'hora de guanyar un partit? Quines posicions són les més determinants en un partit? Quins equips són els més regulars? Quants gols cal marcar en un partit per a poder afirmar que es guanyarà el partit?

També es desenvoluparan dos models per predir la victòria d'un partit que s'adaptin a l'escassetat de dades. Un dels models utilitzarà el número de gols de cada equip com a característiques i l'altre model descriurà el partit en base a informació de les alineacions de jugadors dels dos equips. S'utilitzaran mètodes d'aprenentatge automàtic per a maximitzar la informació obtinguda de les dades disponibles.

Per a poder realitzar aquest projecte, es treballa amb dades de la lliga alemanya d'handbol, anomenada Bundesliga. Per treure la màxima informació d'aquestes es farà una anàlisi exploratòria de les dades. Després s'usaran diferents tècniques de ciència de dades, com regressions logístiques i màquines de vectors de suport, per a elaborar 2 models per a predir la victòria d'un partit basant-nos en diferents

aspectes del joc, un basat en els gols de l'equip i l'altre en el rendiment individual dels jugadors alineats en un partit.

Un cop realitzats els models, s'analitzaran els resultats i es buscarà comparar els resultats obtinguts per a veure quin model és millor.

Aquest treball no sols pretén contribuir al desenvolupament de l'handbol com a esport, sinó també establir un precedent per a futurs estudis en el desenvolupament de models per predir la victòria d'un equip en un partit d'handbol en l'àmbit de la ciència de dades.

1.1 Característiques de l'handbol

L'handbol és un esport d'equip dinàmic que es juga en interiors, on dos equips de set jugadors cadascun intenten marcar gols llançant una pilota amb les mans cap a la porteria contrària. A continuació.

El camp d'handbol és rectangular, amb unes dimensions de 40 metres de llarg per 20 metres d'ample. Les principals àrees del camp són:

- Àrea de porteria: Un semicercle de 6 metres des de la porteria. Només el porter pot estar dins d'aquesta àrea.
- Línia de penal: Es troba a 7 metres de la porteria, des d'on es llancen els tirs penals.
- Línia de cop franc: Situada a 9 metres de la porteria, utilitzada per a reiniciar el joc després de faltes.

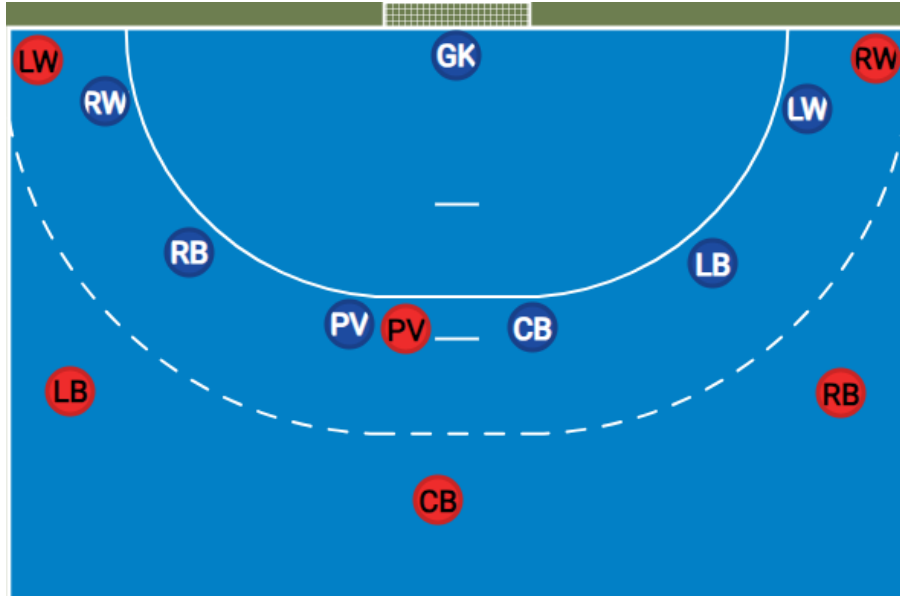


Figura 1: Posicions de l'handbol. Vermells ataquen, blaus defensen. Imatge creada utilitzant el software Tactical Board [5].

Cada equip està compostat per set jugadors en el camp, incloent-hi un porter. Aquestes posicions estan representades a la figura 1. Les posicions són:

- Porter (GK): L'únic jugador que pot tocar la pilota amb qualsevol part del cos dins de la seva àrea. La seva funció principal és evitar que l'equip contrari marqui gols.
- Extrems (W)(Esquerre i Dret): Juguen prop de les línies de banda. Són ràpids i àgils, especialitzats a realitzar llançaments des d'angles difícils. Es consideren posicions finalitzadores.
- Laterals (B)(Esquerre i Dret): Juguen en posicions intermèdies entre el centre i els extrems. Solen ser jugadors forts i amb habilitat per a llançar a llarga distància.
- Central (CB): És l'organitzador de l'atac, similar a un base en bàsquet o un migcampista al futbol. Distribueix el joc i pren decisions ràpides.
- Pivot (PV): Juga prop de la línia de sis metres, lluitant contra els defenses. La seva tasca és crear espais i oportunitats de gol. Es considera una posició finalitzadora.

En lligues professionals europees, com la Bundesliga (Alemanya), la mitjana de gols per partit se situa al voltant de 50 a 60 gols, amb alguns partits que superen aquesta xifra, depenent de l'estil de joc i les tàctiques emprades.

1.2 Objectius i preguntes formulades

Les preguntes inicials que es van formular per resoldre en aquest projecte són les següents:

- Quines estadístiques individuals de jugadors són les més diferencials a l'hora de guanyar un partit?
- Quines posicions són les més determinants per guanyar un partit?
- Quins equips són els més regulars de la lliga?
- Quants gols cal marcar en un partit per a poder afirmar que es guanyarà un partit?
- Podem predir quin equip guanyarà un partit coneixent les alineacions de jugadors dels dos equips participants?

Per resoldre aquestes preguntes es realitzarà un anàlisi exploratòria de dades d'handbol de la Bundesliga alemanya, així com l'elaboració de dos models predictius enfocats en la victòria dels partits. Els objectius específics d'aquest treball són els següents:

1. Dur a terme una anàlisi exploratòria de dades:
 - Examinar i descriure les principals característiques de les dades de la Bundesliga alemanya d'handbol, incloent-hi estadístiques d'equips i individuals.
 - Identificar patrons i tendències rellevants que puguin proporcionar una comprensió més profunda del rendiment en l'handbol.
2. Desenvolupar un model predictor de victòria d'un equip basat en el número de gols de l'equip:
 - Construir un model que predigui la possibilitat de victòria d'un equip, basant-se en la quantitat de gols anotats per aquest equip en el partit.

- Avaluar l'acompliment del model mitjançant mètriques de precisió i validació creuada, assegurant la seva capacitat per a proporcionar prediccions fiables.
- Desenvolupar un model predictiu de victòries centrat en mètriques individuals de les alineacions de jugadors:
 - Desenvolupar un model que predigui la possibilitat de victòria d'un equip basant-se en el rendiment individual dels jugadors. El rendiment individual es quantifica amb una mesura que és l'Índex de Rendiment a l'Handbol (HPI), que té en compte mètriques com gols anotats, assistències, i eficiència en llançaments.
 - Utilitzar tècniques d'aprenentatge automàtic per a transformar les variables utilitzades i ajustar el model en conseqüència.
 - Comparar i validar els models predictius:
 - Avaluar i comparar els resultats de tots dos models predictius, determinant la seva precisió.

1.3 Planificació del projecte

La figura 2 mostra una representació visual de com s'han estructurat els 4 mesos que ha durat la realització d'aquest projecte.

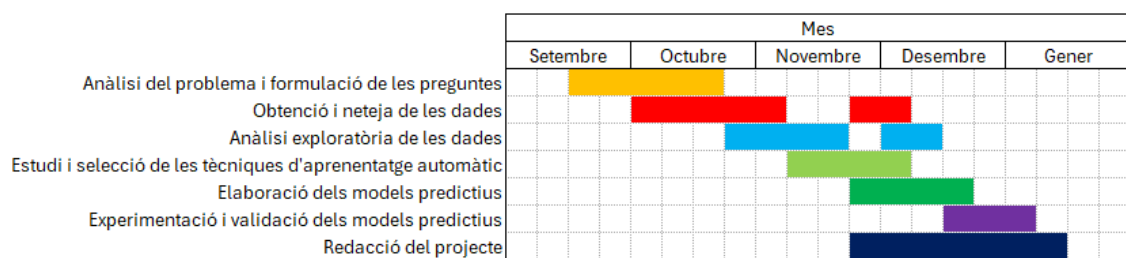


Figura 2: Diagrama de Gantt del projecte.

A l'inici del projecte es van formular moltes preguntes, però a mesura que es va anar analitzant el problema, es van anar reduint i enfocant en l'elaboració de dos models predictius.

Pel que concerneix les dades, primer es va obtenir i va analitzar un conjunt i, després de valorar-lo, es va decidir ampliar l'anàlisi amb un altre conjunt de dades diferent.

1.4 Estructura del projecte

Aquest projecte es dividirà en sis capítols. El Capítol 1 és aquesta introducció on s'ha presentat el problema i es fixen els objectius.

El capítol 2 serà l'obtenció de les diferents dades i la seva descripció per a poder entendre amb què treballarem.

El capítol 3 s'analitzaran i exploraran aquestes dades per a resoldre algunes de les preguntes formulades.

Al capítol 4 s'explicaran algunes tècniques d'aprenentatge automàtic i s'escolliran les més adients per aquest projecte.

El capítol 5 es duran a terme els experiments i s'obtindran els resultats mitjançant diferents models que utilitzen les tècniques escollides.

Finalment, el capítol 6 conclourà amb una comparació dels models i una reflexió sobre el que s'ha aconseguit i futures aplicacions de la ciència de dades a l'handbol.

El codi del projecte es pot consultar en aquest repositori de github:
<https://github.com/IvanMartinezVilla/DataScienceBundesliga>

2 Obtenció i descripció de les dades

La Bundesliga d'Handbol [2] és la principal lliga professional d'handbol a Alemanya i una de les més prestigioses a nivell mundial. Fundada en 1966, s'ha consolidat com un referent pel seu alt nivell competitiu, atraient a alguns dels millors jugadors del món.

La lliga està formada per 18 equips i uns 370 jugadors que competeixen en una temporada regular de 34 jornades. Cada equip juga contra tots els seus oponents en partits d'anada i tornada, sumant 2 punts per victòria i 1 per empat. Al final de la temporada, l'equip amb més punts és coronat campió, mentre que els dos últims classificats descendeixen a la segona divisió, i els dos millors equips d'aquesta divisió ascendeixen.

La lliga és llar d'estrelles internacionals, la qual cosa eleva el nivell competitiu i la converteix en un aparador global per al talent en l'handbol.

Per a començar a treballar en aquest projecte, necessitem 2 conjunts de dades, el de les dades individuals a la Bundesliga de la temporada 2022-2023, i un amb els resultats de les temporades 2021-22, 22-23, 23-24. En aquest capítol s'explicarà l'obtenció dels conjunts de dades amb els que s'ha treballat i es descriuran.

2.1 Conjunt de dades sobre jugadors de la Bundesliga 2022-2023

Les dades de la Bundesliga de la temporada 2022-2023 han estat extretes de Kaggle [7]. Aquestes dades recullen la informació i estadístiques de tots els jugadors durant la temporada 2022-2023. Aquest dataset té un total de 18 variables (columnes) i recullen les dades de 372 jugadors que van competir durant aquesta temporada. La taula 1 mostra que significa cada variable.

NAME	Nom del jugador.
CLUB	Club on el jugador ha jugat aquesta temporada.
POSITION	Posició del jugador. Hi ha 7 posicions diferents.
P	Número de partits jugats pel jugador aquesta temporada.
G	Número de gols que el jugador ha marcat aquesta temporada.
M	Número de llançaments que el jugador ha fallat aquesta temporada.
FG	Número de gols de camp que el jugador ha marcat aquesta temporada.
G.1	Número de gols de penal que el jugador ha marcat aquesta temporada.
%	Percentatge de tirs que el jugador ha marcat aquesta temporada.
AS	Número d'assistències que el jugador ha fet aquesta temporada.
TF	Número de faltes tècniques que el jugador ha fet aquesta temporada.
ST	Número de recuperacions de pilota que el jugador ha fet aquesta temporada.
BL	Número de llançaments que el jugador a blocat aquesta temporada.
YC	Número de targetes grogues que el jugador a rebut aquesta temporada.
2MIN	Número d'exclusions que el jugador ha rebut aquesta temporada.
RC	Número de targetes vermelles que el jugador ha rebut aquesta temporada.
BK	Número de targetes blaves que el jugador ha rebut aquesta temporada.
HPI	Índex de rendiment.

Taula 1: Variables descriptors d'un jugador.

Per facilitar l'anàlisi s'ha decidit afegir una variable que reculli la mitjana de gols que ha fet cada jugador per partit al llarg de la temporada. Aquesta s'anomena **GOALS_PER_GAME**.

2.1.1 Índex de Rendiment a l'Handbol (HPI)

L'índex de rendiment d'handbol (HPI) és una mètrica dissenyada per a avaluar de manera objectiva i quantitativa el rendiment dels jugadors d'handbol durant un partit. Desenvolupat per la Lliga Alemanya d'Handbol [8] en col·laboració amb experts en estadística i anàlisi de rendiment, el HPI assigna a cada jugador un valor numèric que reflecteix la seva contribució en el joc, cosa que permet comparacions transparents i basades en dades entre jugadors.

Cada jugador comença amb un valor base de 100 punts. Al llarg del partit, aquest valor s'ajusta sumant punts per accions positives i restant per accions negatives, segons la posició del jugador i la naturalesa de l'acció. Per exemple, un gol des

de la línia de 9 metres pot atorgar més punts que un des dels 6 metres, a causa de la menor probabilitat d'èxit des de distàncies majors. De manera similar, una parada del porter davant un llançament de 6 metres es valora més que una davant un llançament des de l'extrem.

Acció	Punts Èxit	Punts Fallada	Punts Assistència
Llançament de 6 metres centre	6	-7	4
Llançament de 6 metres lateral	7	-6	3
Llançament de 7 metres penal	6	-8	4
Llançament de 9 metres centre	10	-4	4
Llançament de 9 metres lateral	10	-4	0
Llançament de pivot	6	-7	4
Llançament de l'extrem	6	-7	4
Tir lliure directe	10	-1	0
Contraatac	5	-8	5
Desconegut	5	-8	5

Taula 2: Puntuació per a cada acció dels jugadors de camp.

Acció	Probabilitat d'aturada	Punts Aturada	Punts Gol
Llançament de 6 metres centre	20%	8	-2
Llançament de 6 metres lateral	25%	7	-3
Llançament de 7 metres penal	25%	8	-2
Llançament de 9 metres centre	31%	7	-3
Llançament de 9 metres lateral	30%	7	-3
Llançament de pivot	5%	9	-1
Llançament de l'extrem	34%	7	-3
Tir lliure directe	25%	5	-5
Contraatac	19%	8	-2
Desconegut	22%	8	-2

Taula 3: Puntuació per a cada acció dels porters.

Acció	Punts
Penalització exclusions	-3/-4/-6
Tarjeta vermella	-12
Tarjeta blava	-12
Penalització penal	-3
Blocatge	7
Robatori	8
Error tècnic	-6

Taula 4: Puntuació per altres accions dels jugadors.

Fins a la temporada 2021/22, el HPI es calculava exclusivament a partir de la suma de les accions positives i negatives d'un jugador, que es poden observar a les taules 2, 3 i 4. Per tant, el HPI estava “obert” per dalt i per baix. Des de l'inici de la temporada 2021/22, aquest valor s'ha convertit en el HPI actual mitjançant una fórmula.

$$HPI = (Punts\ reollits - Min1\% \text{ mitjà}) \cdot 50 / (Max1\% \text{ mitjà} - Min1\% \text{ mitjà}) \quad (2.1)$$

Calculant amb la fórmula 2.1, els valors del HPI se situen bàsicament dins d'una escala de 50-100. Cada posició (primera línia, pivot, extrem i porter) té el seu propi valor mínim i màxim, que podem observar a la Taula 5.

Mitjana 1%	Primera línia	Pivot	Extrem	Porter
MIN	58	69	62	58
MAX	179	169	161	203

Taula 5: Valors mínims i màxims del HPI.

En el futur, la mitjana de l'1% més feble o millor de tots els HPI aconseguits constituirà el marc per al càlcul. Això fa que l'índex sigui menys sensible als valors extrems a l'alça o a la baixa, la qual cosa millora la comparabilitat entre posicions.

2.2 Conjunt de dades de resultats de la Bundesliga 2021-22, 22-23, 23-24

A més a més de les dades dels jugadors explicats prèviament hem cercat dades sobre resultats de les temporades 2021-22, 22-23, 23-24 de la Bundesliga. Les dades han estat obtenides de la web CEROACERO [1], que és una web de resultats esportius. Les variables estan descrites en la Taula 6.

ROUND	Número de la jornada en la que es juga el partit.
DATE	Data en la que es juga el partit.
HOME	Nom de l'equip local.
RES	Resultat en gols del partit. Format HG-AG.
AWAY	Nom de l'equip visitant.

Taula 6: Variables descriptors dels resultats.

Per poder treballar amb els resultats de les diferents temporades, s'ha decidit agrupar les diferents taules i afegir diferents variables per facilitar el treball futur amb el dataset. Per a fer-ho, treballarem en un Python Notebook, i amb la llibreria de Pandas [11], amb la qual podem modificar els datasets de manera eficient. El primer pas es juntar les diferents temporades en un mateix dataset. Un cop juntes, s'ha decidit eliminar la columna **DATE** i afegir una columna **SEASON** que conté només l'any d'inici de la temporada on es disputa el partit. Després d'agrupar els partits per temporades, els resultats segueixen en un format que no facilita l'anàlisi i els càlculs. Per solucionar aquest inconvenient, s'ha optat per separar la columna **RES** (veure Taula 6) en 5 columnes diferents, representades a la Taula 7.

HG	Número de gols marcats per l'equip local.
AG	Número de gols marcats per l'equip visitant.
RES	Resultat del partit. Pot ser H (victòria local), D (empat), A (victòria visitant).
W	Número de gols de l'equip guanyador. En cas d'empat, número de gols del local.
L	Número de gols de l'equip perdedor. En cas d'empat, número de gols del visitant.

Taula 7: Variables descriptors afegides al conjunt de resultats.

3 Anàlisi exploratòria de les dades

Per a poder definir més endavant diferents models de predicció amb les dades dels nostres datasets, ens interessa dur a terme un anàlisi d'aquestes dades per tal de veure quines són més o menys rellevants. El llenguatge emprat per fer l'anàlisi serà Python, més concretament Jupyter Notebook. Per a fer l'anàlisi de dades utilitzarem principalment les llibreries Pandas, Matplotlib [10] i Scikit-Learn [13].

3.1 Anàlisi de les dades dels jugadors

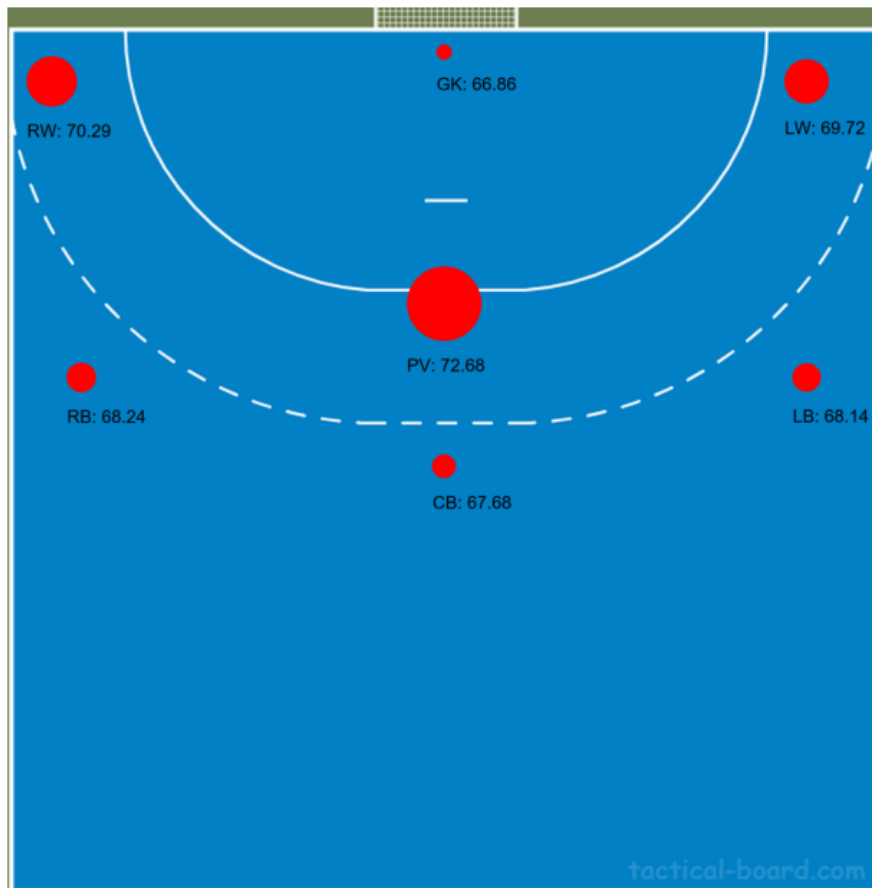


Figura 3: Il·lustració del HPI mitjà per posició.

La figura 3 mostra una il·lustració on es pot visualitzar el rendiment dels jugadors d'un equip d'handbol, basada en l'Índex de Rendiment (HPI). Les posicions dels jugadors estan distribuïdes en el camp, mostrant els HPI mitjà de cada posició: el pivot (PV) té el major HPI (72.68), mentre que el porter (GK) mostra el valor més

baix (66.86). Això suggereix que el pivot ha tingut un rendiment superior, la qual cosa podria influir en decisions tàctiques.

Els extrems dret (RW) i esquerre (LW) i els laterals (RB i LB) presenten valors de HPI bastant consistents entre les dues bandes, la qual cosa indica una estabilitat en aquestes posicions. No obstant això, el central (CB) mostra un HPI lleugerament inferior (67.68), la qual cosa podria assenyalar una àrea de millora en l'acompliment estratègic de l'equip.

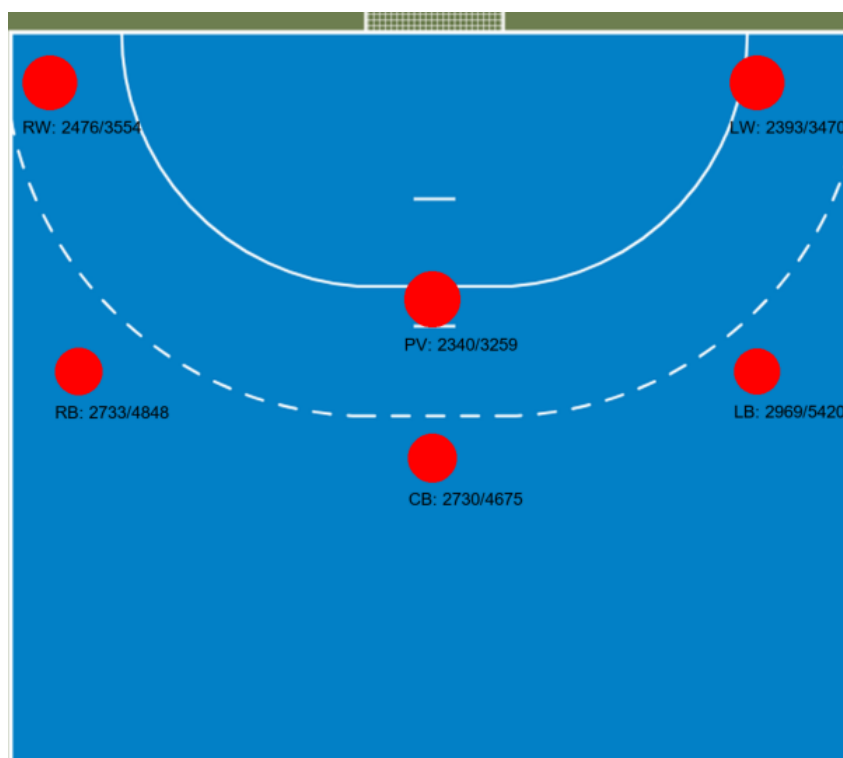


Figura 4: Representació dels llançaments per posició.

Abans hem vist com es calcula el HPI, donant prioritat als llançaments acabats en gol i als èxits en accions defensives. Aquests HPI s'entenen millor veient la figura 4. Aquesta visualització mostra com les zones finalitzadores (extrems i pivot) tenen un HPI més alt, ja que són les zones on hi ha més percentatge d'efectivitat de cara a porteria. En canvi, les zones on es mou el joc (laterals i central) i tenen llançaments més forçats, tenen valors inferiors pel fet que hi ha més pèrdues de pilota.

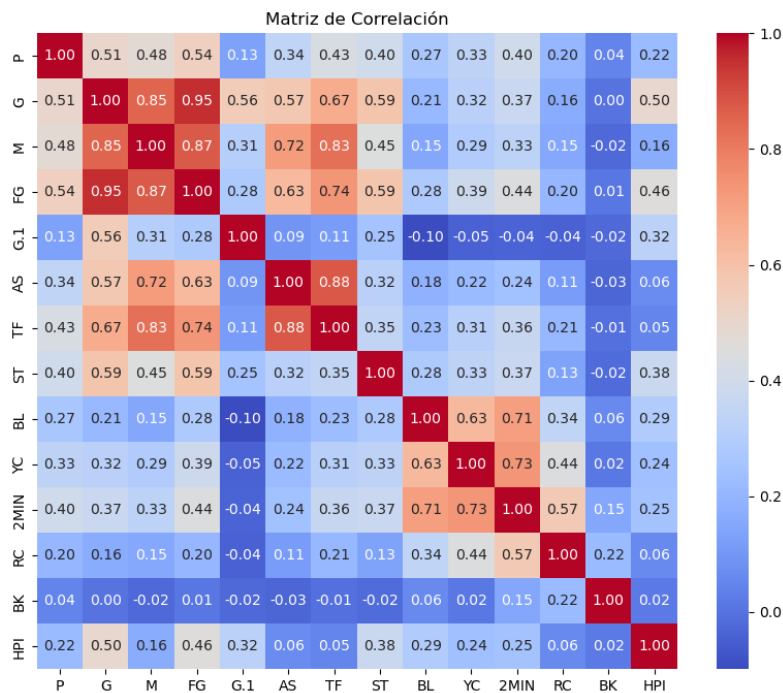


Figura 5: Matriu de correlació de les estadístiques individuals mostrades a la taula 1 de la secció 2.1.

La figura 5 ens mostra matriu la correlació [3] entre aspectes individuals dels jugadors durant el partit.

Veiem relacions clares entre gols i gols de camp, ja que la major part dels gols d'un partit són en situació de joc dinàmic.

També podem observar una clara relació entre les faltes comeses i els llançaments fallats. Això és degut a que durant les accions d'atac, si la defensa provoca faltes de manera constant, el ritme de la circulació de la pilota és torna més lent, provocant que l'atac hagi de precipitar-se per aconseguir un llançament i donant-li la possibilitat al porter d'aturar el tir.

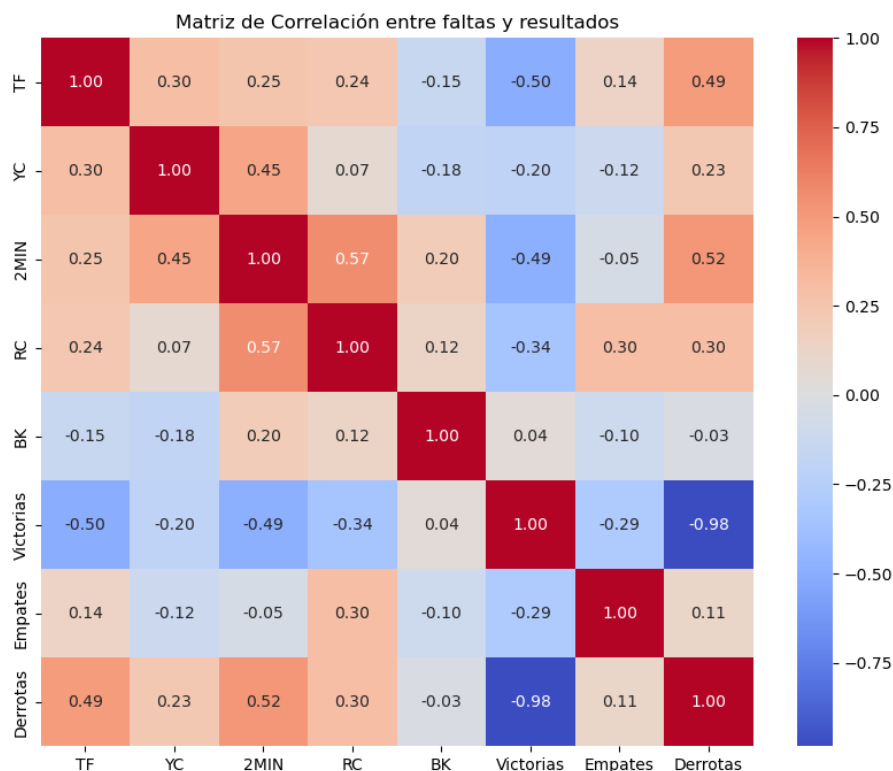


Figura 6: Matriu de correlació dels aspectes defensius i resultats.

Centrant-nos més en les accions defensives i l'impacte que tenen en el resultat del partit, en la figura 6 podem observar una clara relació entre les diferents penalitzacions progressives ja que, a mesura que vas fent faltes durant un partit, es van acumulant les diferents amonestacions, la progressió és, YC (tarjeta groga) $\rightarrow$ 2MIN (suspensió de 2 minuts) i 2MIN $\rightarrow$ RC (tarjeta vermella).

També veiem que un equip que rep diverses 2MIN tendeix a perdre més partits, ja que passa més temps amb un jugador de menys al terreny de joc.

Per aquest motiu, els equips que tenen jugadors que fan una gran quantitat de faltes, acostumen a perdre els partits, ja que juguen en inferioritat numèrica durant diferents períodes del joc.

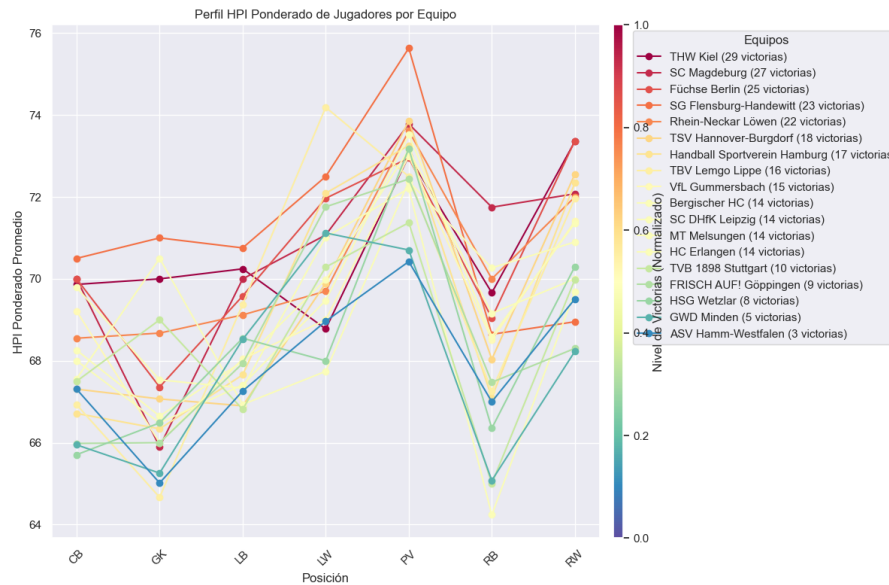


Figura 7: HPI per posició dels equips.

En la figura 7 es mostra un gràfic dels valors del HPI per posició. En aquest gràfic veiem que els equips amb més victòries destaquen per un rendiment més homogeni en totes les posicions. Per exemple, el THW Kiel (29 victòries) mostra un perfil alt i estable, especialment en posicions clau com el central (CB) i l'extrem dret (RW). Altres equips com el SC Magdeburg (27 victòries) i el Füchse Berlin (25 victòries) també sobresurten, mantenint línies superiors en la majoria de les posicions.

En la meitat de la taula, equips com el SG Flensburg-Handewitt (23 victòries) i el Rhein-Neckar Löwen (22 victòries) presenten perfils competitius, tot i que no tan destacats com els líders. Els seus valors són més homogènies entre posicions, la qual cosa indica un enfocament equilibrat en el rendiment global de l'equip.

En canvi, els equips amb menys victòries, com l'ASV Hamm-Westfalen (3 victòries), mostren perfils baixos, especialment en posicions fonamentals com el CB i el RW. Aquesta falta de rendiment en posicions clau podria ser un factor determinant en la seva posició baixa a la classificació.

Pel que fa a les posicions, hi ha diferències significatives en els valors de HPI. Posicions com l'extrem esquerra (LW), l'extrem dret (RW) i el pivot (PV) solen tenir pics en els equips líders, mentre que el central (CB) destaca per tenir valors generalment més baixos, fins i tot en equips forts.

Les posicions de primera línia (laterals i central) són les que tens el HPI més baix. Pot deure's al fet que aquestes posicions són en les quals més vegades passa

la pilota i es poden produir més errors. Al seu torn, la posició del pivot (PV) és la que major HPI té, ja que una vegada arriba una pilota a aquesta posició, la major part de les vegades, acaba en gol o penal, que es recompensa amb més punts en el càlcul del HPI.

Finalment, hi ha una clara correlació visual entre el nombre de victòries i el perfil del HPI. Els equips amb més victòries presenten línies més altes i consistents en gairebé totes les posicions, mentre que els equips amb menys victòries tenen línies més baixes i, sovint, més irregulars.

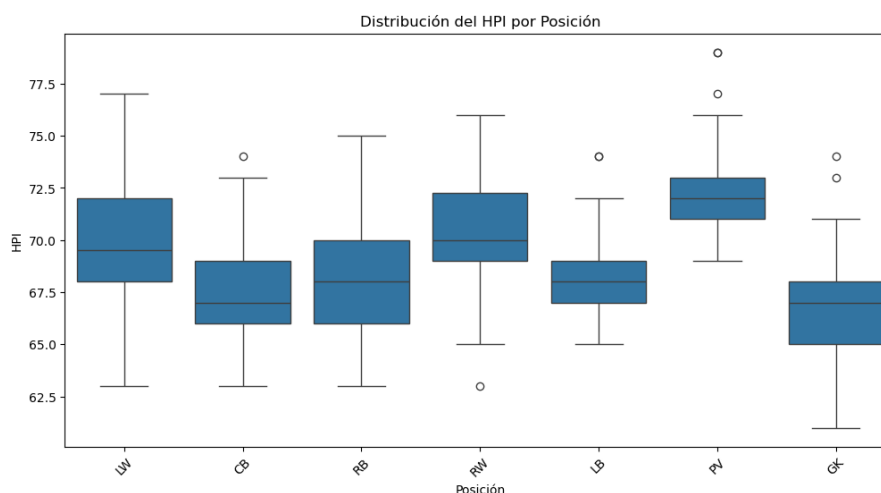


Figura 8: Distribució del HPI per posició.

En la figura 8 es mostra la distribució del HPI de les diferents posicions, considerant tots els equips de la lliga. Podem veure com les posicions que són finalitzadores (LW, RW, PV) són les que tenen els HPI més alts, ja que són les que creen més situacions valuoses (llançaments acabats en gol, penals).

Per altra banda podem veure com les posicions de primera línia (CB, RB, LB) tenen un HPI més baix degut a que és la zona de joc en la que més es mou la pilota i pot haver-hi més errors de passada i pèrdues de pilota.

Per últim tenim el porter, que té un càlcul diferent, on compten les parades i els gols que rep. Per aquest motiu, és normal que els porters tinguin un HPI més baix que la resta, ja que és difícil que un porter faci més d'un 50% de parades durant una temporada.

3.2 Anàlisi de les dades de la classificació

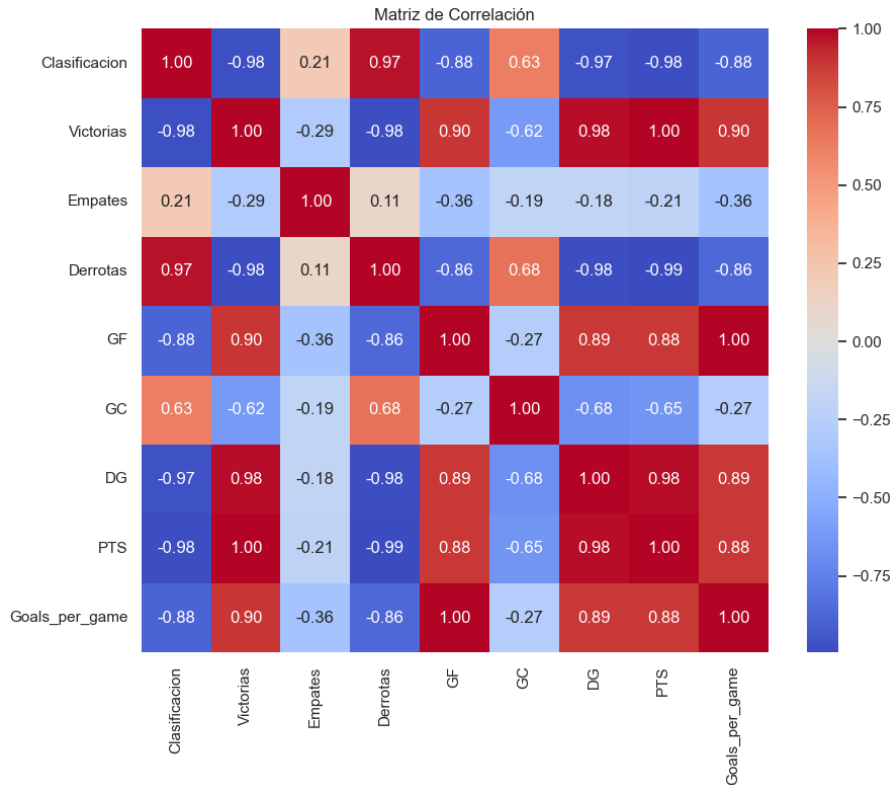


Figura 9: Matriu de correlació estadístiques dels equips.

La figura 9 ens mostra la matriu correlació que hi ha entre els diferents atributs d'un equip a la competició. Podem observar coses que són clares, si guanyes molt o perds poc, és molt probable que estiguis en una bona posició a la classificació.

També veiem que els equips que aconsegueixen marcar molts gols i deixen al rival amb pocs gols, es més probable que al llarg de la temporada aconsegueixin assolir una molt bona posició a la classificació.

Per tant podem deduir que és més important tenir un bon balanç entre els gols a favor i en contra que només anar a marcar els màxims possibles.

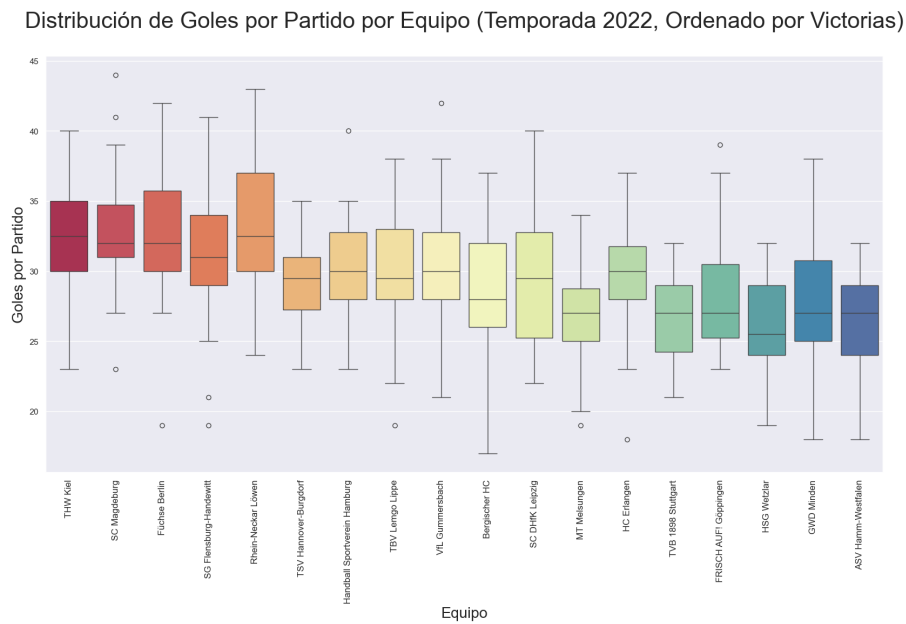


Figura 10: Distribució de gols dels equips. Els equips estan ordenats per quantitat de victòries d'esquerra a dreta.

La figura 10 és un diagrama de caixa que representa la distribució de gols per partit dels equips, mostrant com els jugadors de diferents equips contribueixen ofensivament en cada partit. En observar el gràfic, es destaca que els jugadors d'equips com THW Kiel i SC Magdeburg tendeixen a tenir una major mitjana de gols per partit. Això suggereix que aquests equips no sols compten amb jugadors ofensivament destacats, sinó també amb un acompliment més uniforme entre els seus integrants.

L'amplitud interquartílica (IQR), que abasta des del primer quartil (Q1) fins al tercer quartil (Q3), indica la variabilitat en la quantitat de gols anotats per jugador en cada equip. Per exemple, en equips com MT Melsungen i FRISCH AUF! Göppingen, s'observa una major dispersió en la quantitat de gols, la qual cosa suggereix una diferència significativa entre els jugadors més i menys efectius en termes d'anotació. Per contra, equips com TSV Hannover-Burgdorf presenten una distribució més concentrada, indicant que els seus jugadors tendeixen a contribuir de manera més equitativa en el marcador.

A més, el gràfic mostra la presència de valors atípics en alguns equips, com en VfL Gummersbach, on uns certs jugadors aconsegueixen un rendiment excepcional, marcant un nombre significativament major de gols en alguns partits. Aquests

valors atípics poden ser indicatius de jugadors estrella que, en determinats partits, sobresurten notablement per sobre de la mitjana de l'equip.

En general, els equips al principi del gràfic, com THW Kiel i SC Magdeburg, semblen beneficiar-se d'una major contribució de gols per part dels seus jugadors, la qual cosa està associat amb un major èxit en termes de victòries. En contrast, equips com GWD Minden i ASV Hamm-Westfalen, situats al final del gràfic, mostren una menor mitjana de gols per jugador, la qual cosa podria reflectir una menor efectivitat ofensiva a nivell individual.

En conclusió, aquesta anàlisi subratlla la importància de tenir jugadors consistents i efectius en l'anotació per a l'èxit de l'equip. Equips amb una menor dispersió i majors contribucions per jugador tendeixen a estar més ben posicionats en la competició.

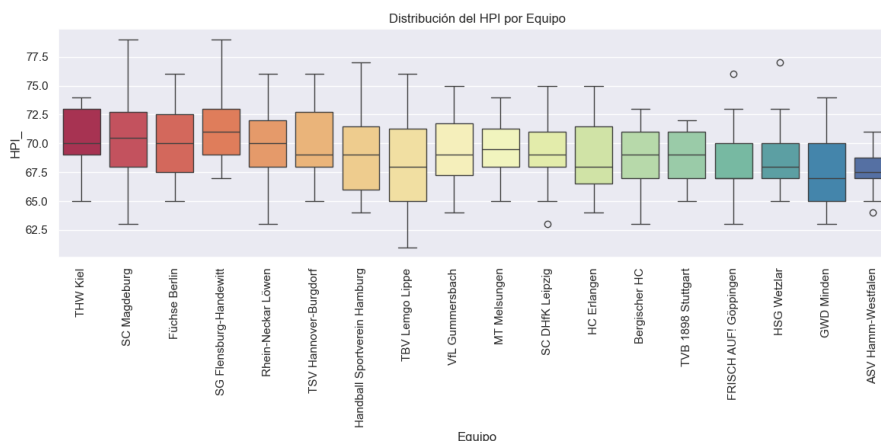


Figura 11: Distribució del HPI per equip. Els equips estan ordenats per número de victòries d'esquerra a dreta.

La figura 11 mostra un gràfic de caixa on podem veure la distribució del Handball Performance Index (HPI) dels jugadors en cada equip, oferint una visió detallada del rendiment individual dins de cada conjunt.

Els equips THW Kiel i SC Magdeburg destaquen per tenir un HPI mitjana més alta, la qual cosa suggereix que els seus jugadors mantenen un acompliment superior de manera constant. La baixa dispersió en aquests equips indica una uniformitat en el nivell de rendiment dels seus jugadors, la qual cosa és clau per al seu èxit.

En contrast, equips com MT Melsungen i FRISCH AUF! Göttingen presenten una major variabilitat en el HPI, la qual cosa reflecteix diferències significatives en

el rendiment dels seus jugadors. Aquesta dispersió pot assenyalar la necessitat de millorar l'acompliment dels jugadors amb índexs més baixos per a augmentar la consistència de l'equip.

D'altra banda, GWD Minden i ASV Hamm-Westfalen es troben en l'extrem inferior del gràfic, amb HPI més baixos en mitjana, la qual cosa podria indicar un menor rendiment general. Malgrat això, la consistència en alguns equips com TVB 1898 Stuttgart mostra un nivell de rendiment homogeni, encara que amb un HPI inferior als líders.

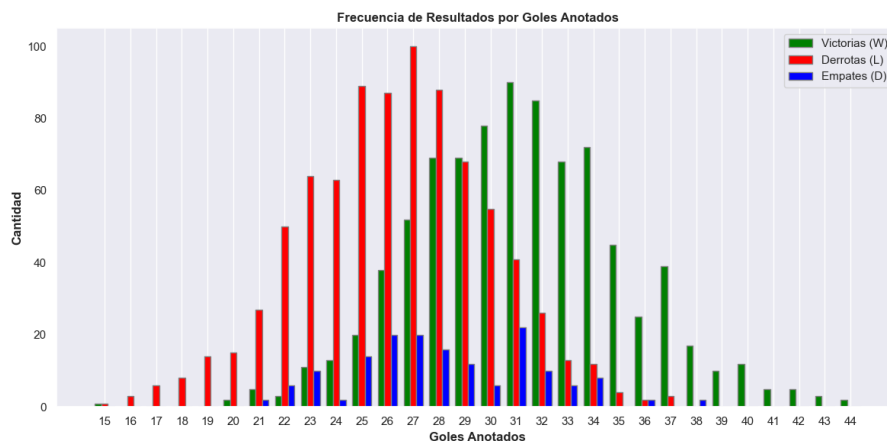


Figura 12: Freqüència de resultats per gols anotats.

El gràfic de barres de la figura 12 mostra la freqüència de resultats en funció de la quantitat de gols anotats pels equips en els partits, diferenciant entre victòries, derrotes i empats.

Les victòries (en verd) són més freqüents a mesura que els equips van anotant més gols, situant-se la mitjana en 31, la qual cosa indica que anotar per sobre d'aquesta marca augmenta significativament les probabilitats de guanyar.

D'altra banda, les derrotes (en vermell) predominen quan els equips anoten menys gols, aconseguint el seu punt màxim als 27 gols, la qual cosa suggereix que no aconseguir una xifra alta de gols incrementa la probabilitat de perdre.

Els empats (en blau), encara que menys comuns, es distribueixen principalment entre els 25 i 28 gols, tot i que hi ha un pic a 31 gols, reflectint que els partits tendeixen a finalitzar igualats quan tots dos equips anoten en un rang de gols moderat.

En conjunt, el gràfic ressalta que una major capacitat ofensiva i defensiva, re-

flectida en una major quantitat de gols anotats i menys encaixats, està directament associada amb un major nombre de victòries, mentre que anotar menys gols es tradueix en un major risc de derrota.

4 Tècniques d'aprenentatge automàtic per predir la victòria/no victòria d'un partit

Abans de començar a experimentar amb els models de predicció cal analitzar diferents models d'entrenament per a veure quin s'ajusta millor als nostres problemes. S'analitzaran breument els algorismes més coneguts i emprats en la ciència de dades, es veuran els avantatges i inconvenients de cadascun d'ells i a partir d'això s'escolliran els models, els quals s'estudiaran amb més detall i amb els quals intentarem resoldre les nostres preguntes. Aquesta selecció es fa degut al teorema “No Free Lunch” [17]. Aquest teorema ens diu que no hi ha un algorisme de “Machine Learning” que funcioni millor per a tots els problemes. És especialment notable per a l'aprenentatge supervisat, que és el que busquem nosaltres.

S'han estudiat els avantatges i inconvenients dels següents algorismes: Regressió logística, Naive Bayes, Màquines de vectors de suport, Boscos aleatoris i XGBoost.

4.1 Regressió logística

La regressió logística [15] és un algorisme de classificació que s'empra per trobar la probabilitat d'un esdeveniment. Se sol emprar en problemes que tenen resultats binaris, com és en un dels nostres casos (Victòria, No Victòria). Els punts forts i dèbils més importants són els següents:

Avantatges:

- Fàcil d'implementar i interpretar, a m'es de tenir alta eficiència.
- Funciona bé quan el conjunt de dades és linealment separable. Aquest concepte significa que podem “dibuixar una línia recta” entre les variables que són, en el nostre cas “Victòria” i “No Victòria”.
- Poca possibilitat de “Over-fitting” en conjunts de dades amb poques dimensions.

Inconvenients:

- Assumeix la linealitat entre la variable dependent i les independents, és a dir entre els “Gols” i la resta.

- No s’ha d’emprar si el nombre d’observacions (files o valors) és menor que el nombre de característiques (variables o columnes). Per nosaltres, no és el cas.
- Només pot ser emprada per predir funcions discretes. Per la predicció de dades contínues no és convenient. Tampoc és el nostre cas.

4.2 Naive Bayes

Naive Bayes [14] és un classificador basat en el teorema de Bayes. Els punts forts i dèbils són els següents:

Avantatges:

- Dona resultats molt bons si les variables del conjunt de dades son independents.
- Fàcil d’implementar ja que només es calcula la probabilitat.
- Funciona bé amb datasets de dimensions altes, és a dir, si “Gols” pot tenir molts valors.

Inconvenients:

- Si les variables no són independents, no funciona gaire bé.
- S’ha de fer un procés de “Smoothing” si la probabilitat d’un resultat és zero.
- Pot tenir “Vanishing value”. Això vol dir que un valor molt petit “desapareix” per el fet de ser tan petit, quan s’hauria de tenir en compte.

4.3 Màquines de vectors de suport

Aquest model s’utilitza per a classificació i regressió [16]. El seu objectiu principal és trobar l’hiperplà òptim que maximitza la separació entre les diferents classes en l’espai de característiques. Les dades que estan més prop d’aquest hiperplà es denominen vectors de suport, i són fonamentals per a definir la posició i orientació d’aquest.

Avantatges:

- Eficax en espais d'alta dimensionalitat: funciona bé en situacions on el nombre de dimensions és major que el nombre de mostres.
- Eficiència amb un nombre limitat de mostres: És particularment útil quan es disposa d'un conjunt de dades reduït, però ben estructurat.
- Flexibilitat a través de kernels: permet l'ús de diferents funcions kernel (lineals, polinòmiques, RBF, etc.), la qual cosa li permet modelar relacions no lineals entre les característiques.
- Bona generalització: Tendeix a tenir un alt poder de generalització a causa del seu enfocament a maximitzar el marge entre les classes.

Inconvenients:

- Sensibilitat a les dades d'entrenament: pot ser sensible als outliers, ja que l'hiperplà depèn dels vectors de suport, que podrien veure's afectats per valors atípics.
- Complexitat computacional: L'eficiència computacional disminueix en grans conjunts de dades, especialment en termes de temps d'entrenament i ús de memòria.
- Elecció del kernel i paràmetres: Seleccionar el kernel adequat i ajustar els paràmetres pot ser complicat i requereix un enfocament acurat.
- Escalabilitat: Encara que és eficaç en espais d'alta dimensionalitat, el seu rendiment no escala bé amb grans quantitats de dades a causa de la seva elevada complexitat computacional.

4.4 Boscos aleatoris

Aquest algorisme és un model d'aprenentatge que empra múltiples arbres de decisió per fer les prediccions [4]. Pot ser emprat per classificació i regressió. El que fa és crear un gran nombre d'arbres amb diferents mostres aleatòries de les dades. S'entrenen tots aquests arbres i s'agrega el resultat de tots per obtenir el resultat final. Els avantatges i inconvenients són els següents:

Avantatges:

- Funciona bé per relacions no lineals i complexes entre les característiques i la variable “Gols”.
- Pot manejar dades numèriques i categòriques sense problemes, valors nuls, i no necessita escalar els valors.
- Dona informació sobre la importància de cada característica.
- Funciona amb conjunts de dades grans i amb altes dimensions.

Inconvenients:

- La interpretabilitat dels resultats és molt baixa.
- Pot tenir alts costos computacionals.
- Sensible a dades amb soroll.

4.5 XGBoost

Aquest és un algorisme que implementa arbres de decisió potenciats amb el gradient, dissenyat per tenir bona velocitat i rendiment que tracta d’obtenir un resultat a través de combinar resultats d’un conjunt de models més simples [12]. Els avantatges i inconvenients són els següents:

Avantatges:

- Disposa d’una regularització que impedeix l’overfitting.
- Treballa bé amb valors que manca.
- Paral·lelització, bon rendiment.

Inconvenients:

- Si els paràmetres no estan ben ajustats, sí que és possible l’overfitting.
- Per conjunts de dades grans pot trigar molt.

4.6 Model escollit per a la predicció per gols

La regressió logística, il·lustrada a la figura 13, es defineix com a un algorisme d'aprenentatge automàtic supervisat que aconsegueix tasques de classificació binària a través de predir la probabilitat d'un esdeveniment [15]. El model analitza la relació entre una o més variables independents i classifica dades a classes discretes.

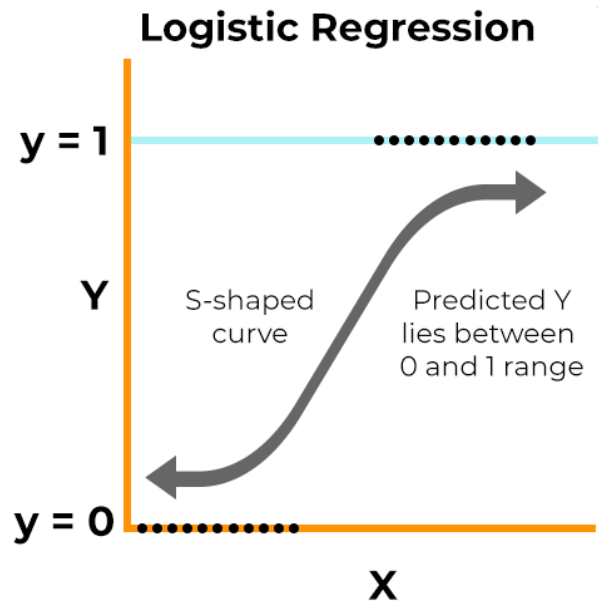


Figura 13: Regressió logística. Imatge extreta de la web Spiceworks [15].

Aquest algorisme emprava una funció logística anomenada “sigmoid function” per a mapar prediccions i les seves probabilitats. Aquesta funció es visualitza com una corba en forma de S que converteix qualsevol valor real a un rang d'entre 0 i 1.

Si el resultat de la funció és major que un llindar predefinit, el model prediu que la instància pertany a una classe. Si el resultat és més petit que el llindar, la instància no pertany a aquesta classe. Aquesta funció està representada a l'equació 4.1.

$$f(X) = \frac{1}{1 + e^{-(x)}} \quad (4.1)$$

On la e es refereix a la base de logaritmes naturals i el valor x al valor numèric que es vol transformar al rang de 0 a 1.

L'equació 4.2 representa la regressió logística.

$$y = \frac{e^{(b_0+b_1x)}}{1 + e^{(b_0+b_1x)}} \quad (4.2)$$

On x és el valor que es rep, y la predicció, b_0 un valor de calibració i b_1 el coeficient per x .

En el nostre cas només necessitem els gols anotats per l'equip guanyador i no guanyador (W i L) i saber si guanya o no el partit (Res). Veure la Taula 7.

4.7 Model escollit per a la predicció per l'Índex de Rendiment d'Handbol (HPI)

El Support Vector Machine (SVM) és un algorisme d'aprenentatge automàtic supervisat que s'utilitza principalment per a tasques de classificació binària i regressió. SVM funciona trobant l'hiperplà òptim que separa les dades en dues classes diferents amb el màxim marge possible entre elles.

Aquest hiperplà es defineix per un subconjunt de punts de dades anomenats vectors de suport, que determinen la posició de l'hiperplà. La funció de decisió de SVM es pot expressar seguint l'equació 4.3.

$$f(x) = \sum_{i=1}^n \alpha_i y_i K(x_i, x) + b \quad (4.3)$$

On:

- x_i són els vectors de suport.
- y_i són les etiquetes de classe corresponents.
- α_i són els coeficients determinats durant l'entrenament.
- $K(x_i, x)$ és la funció de nucli (kernel) que transforma les dades a un espai de característiques de més dimensions.
- b és un valor de desplaçament.

SVM pot treballar amb dades que no són linealment separables utilitzant funcions de nucli per a projectar les dades a un espai de dimensions superiors on es puguin separar linealment. Algunes de les funcions de nucli més populars són:

- Nucli lineal: usat quan les dades són linealment separables.
- Nucli polinòmic: permet separar dades amb fronteres de decisió no lineals.
- Nucli RBF (Radial Basis Function): és el més fet servir per a problemes de classificació no lineals.

En SVM, l'objectiu és maximitzar el marge entre les dues classes. La funció de decisió es pot expressar seguint l'equació 4.4.

$$y = \text{sign} \left(\sum_{i=1}^n \alpha_i y_i K(x_i, x) + b \right) \quad (4.4)$$

On la funció de signe determina a quina classe pertany una instància en funció de la seva posició relativa a l'hiperplà òptim.

En resum, l'algorisme SVM és una eina potent per a la classificació de dades, especialment en situacions en què les classes no són linealment separables, gràcies a l'ús de funcions de nucli que permeten projectar les dades a espais de més dimensions. Aquesta capacitat d'adaptació el fa especialment útil per a la predicció de l'índex de rendiment HPI (Handball Performance Index).

Les característiques que s'han utilitzat per descriure el partit han estat els valors HPI dels jugadors alineats en un partit. Veure Taules 1 i 6.

Jornades: Inclouen informació dels partits, com els equips locals ("Home") i visitants ("Away"), i el resultat del partit. Vectors d'equip: Contenen el HPI per a diferents posicions dels jugadors de cada equip. Per a garantir la coherència en les dades, es van eliminar els espais en blanc dels noms dels equips i es van unir els vectors HPI corresponents als equips locals i visitants en cada partit. Aquesta fusió va permetre crear un conjunt de dades enriquit que vincula el rendiment individual dels jugadors amb el resultat del partit.

4.8 Reducció de característiques

L'Anàlisi de Components Principals (PCA) [9] és una tècnica de reducció de dimensionalitat que s'utilitza àmpliament en el camp de la ciència de dades. El seu principal objectiu és transformar un conjunt de variables possiblement correlacionades en un conjunt més petit de variables no correlacionades anomenades components principals, sense perdre massa informació en el procés. PCA és especialment útil

quan es treballa amb dades d'alta dimensionalitat, on pot ser difícil visualitzar o processar les dades.

Aquesta tècnica es basa en conceptes d'àlgebra lineal i estadística. El procés comença calculant la matriu de covariància de les dades per a determinar com les diferents variables estan correlacionades entre si. A partir d'aquesta matriu, es calculen els autovalors i autovectors, que representen la magnitud i la direcció de la variància en les dades. Els components principals són els vectors propis ordenats en funció dels autovalors, que indiquen quanta variància de les dades originals pot explicar cada component.

Passos del PCA:

- Estandardització de les dades: Atès que PCA és sensible a l'escala de les variables, és comuna estandaritzar les dades perquè cada variable tingui una mitjana de 0 i una desviació estàndard de 1.
- Càlcul de la matriu de covariància: Determina la relació entre les variables.
- Obtenció dels autovalors i autovectors: Els autovectors determinen la direcció dels components principals, i els autovalors la seva importància o magnitud.
- Selecció de components principals: Es trien els components amb major variància explicada, la qual cosa permet reduir la dimensionalitat mantenint la major part de la informació.
- Transformació de les dades: Les dades originals es projecten en l'espai dels components principals seleccionats.

La figura 14 ens mostra la gràfica de la varianza acumulada del PCA.

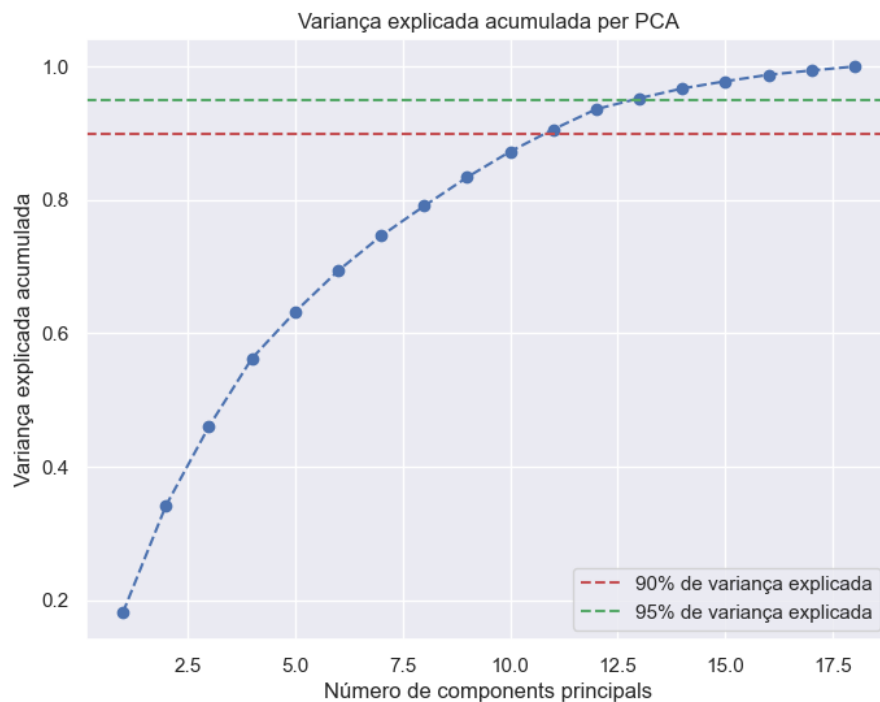


Figura 14: Variança explicada acumulada PCA.

Avantatges:

- Reducció de Dimensionalitat: PCA permet reduir la quantitat de variables en un conjunt de dades, la qual cosa facilita la visualització i pot millorar l'eficiència dels algorismes d'aprenentatge.
- Eliminació de Redundància: En enfocar en la variància explicada, PCA elimina la redundància entre variables correlacionades.
- Millora d'Acompliment: En alguns casos, els models d'aprenentatge automàtic poden beneficiar-se d'una representació més compacta i menys sorollosa de les dades.

Limitacions:

- Interpretabilitat: Els components principals són combinacions lineals de les variables originals, la qual cosa pot dificultar la seva interpretació.
- Pèrdua d'Informació: Depenent del nombre de components seleccionats, pot haver-hi una pèrdua d'informació significativa.

- Assumeix Linealitat: PCA assumeix relacions lineals entre les variables, per la qual cosa pot no capturar bé relacions no lineals.

5 Experiments i resultats

Un cop preparades i analitzades les dades i els models d'entrenament, ja estem preparats per a començar amb els experiments i a crear i validar diferents models d'aprenentatge automàtic que ens ajudi a saber si un equip guanyarà o no un partit d'handbol depenent de diferents característiques.

5.1 Tècnica de validació

A l'hora de l'entrenament s'ha tingut en compte com separarem aquestes dades en conjunts d'entrenament i testatge. S'ha decidit emprar la tècnica “K-fold cross validation” [6]. “K-fold cross validation” és un mètode emprat per separar les dades d'una manera que totes les dades siguin emprades per entrenar i testejar. El que fa és primerament barrejar les dades aleatòriament. Després separa la taula de dades en k grups. Un cop separades, entrena i testeja les dades k vegades, emprant un grup com a testatge cada cop. En acabar suma tots els resultats obtinguts als tests, i fa una mitjana per obtenir a la precisió.

En el nostre cas, hem emprat un valor de $k=5$. Ara que s'han escollit les tècniques que s'empraran per a l'entrenament de les dades, és hora de posar la feina feta en pràctica i començar amb els experiments, veient quin rendiment podem treure dels nostres models.

5.2 Mètriques d'avaluació

En aquest treball, s'utilitzen diverses mètriques d'avaluació per analitzar el rendiment dels models. A continuació, es descriuen breument:

- **Precisió:** Mesura la proporció de prediccions positives que són correctes. Es calcula com:

$$Precisió = \frac{TP}{TP + FP} \quad (5.1)$$

on TP són els vertaders positius i FP els falsos positius. La desviació fa referència a la variabilitat de la precisió entre diferents proves o classes.

- **Especificitat:** Avalua la capacitat del model per identificar correctament els

vertaders negatius. La seva fórmula és:

$$Especificitat = \frac{TN}{TN + FP} \quad (5.2)$$

on són els vertaders negatius i FP els falsos positius. Aquesta mètrica és crucial quan els falsos positius són costosos.

- **Sensibilitat:** Indica la capacitat del model per identificar tots els casos positius. Es calcula com:

$$Sensibilitat = \frac{TP}{TP + FN} \quad (5.3)$$

on FN són els falsos negatius.

- **Puntuació F1:** És la mitjana harmònica entre la precisió i la sensibilitat, proporcionant un equilibri entre ambdues mètriques. Es calcula com:

$$Puntuació F1 = 2 \cdot \frac{Precisió \cdot Sensibilitat}{Precisió + Sensibilitat} \quad (5.4)$$

És especialment útil en problemes amb classes desbalancejades.

- **Prova de Permutació:** És una tècnica de validació estadística que avalua la significància del rendiment del model. Consisteix a permutar aleatòriament les etiquetes de les mostres i recalculat les mètriques del model. Això permet determinar si el rendiment observat és significativament millor que l'esperat per atzar.

5.3 Model de gols

A continuació es mostrarà el procediment que s'ha dut a terme per a realitzar el model predictiu que ens diu si un equip guanyarà o no un partit basat en els gols que anotarà. L'objectiu d'aquest experiment és trobar el model que millors resultats ens proporcioni, és a dir, el que tingui un millor rendiment.

5.3.1 Resultats del model

L'anàlisi del model de regressió logística per a la predicció de victòries en funció del nombre de gols marcats mostra resultats prometedors. En primer lloc, la figura 15

mostra la matriu de confusió del model de gols. Aquesta mostra que el model classifica correctament 662 derrotes i 661 victòries, amb 256 falses victòries i 257 falses derrotes, la qual cosa indica un acompliment equilibrat entre totes dues classes.

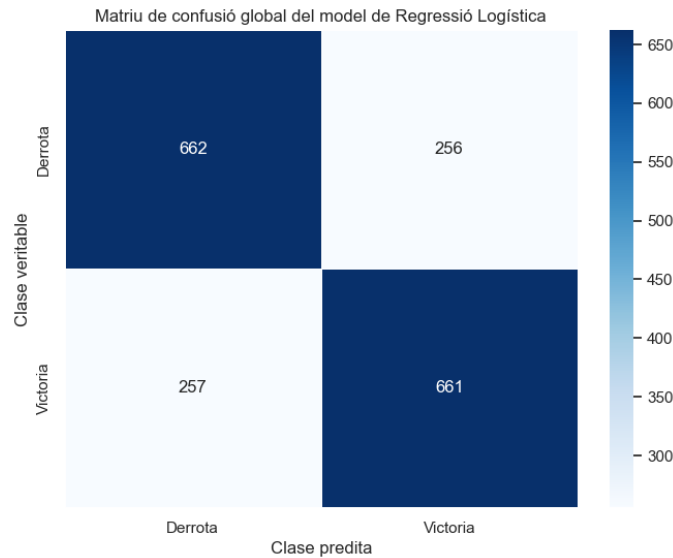


Figura 15: Matriu de confusió del model de gols.

El gràfic de probabilitat de victòria en funció dels gols marcats, mostrat a la figura 16, confirma la relació no lineal capturada pel model. La corba de regressió logística, en forma de “S”, mostra que a mesura que augmenta el nombre de gols, també el fa la probabilitat de victòria, estabilitzant-se prop d’1 (o 100%) per a valors alts de gols marcats. Les prediccions del model s’alineen raonablement bé amb les dades reals, la qual cosa reflecteix un bon ajust del model.

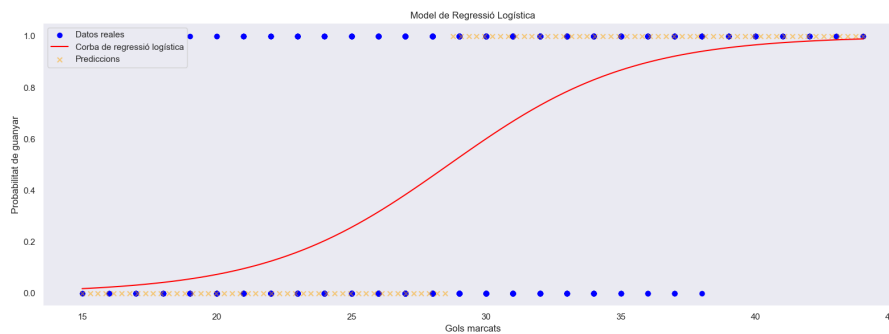


Figura 16: Regressió logística del model de gols.

Els resultats de l’avaluació del model mitjançant testeig de permutacions sostenen la seva fiabilitat. El Original Score de 0.713, comparat amb una mitjana

de Permutation Scores de 0.501 ± 0.013 , evidència que el model supera significativament les expectatives aleatòries. El P-value de 0.001 indica que la probabilitat d'obtenir aquest resultat per atzar és extremadament baixa, confirmant la robustesa del model.

En conjunt, aquests resultats evidencien que el model de regressió logística és una eina valuosa per a predir victòries en funció dels gols marcats, proporcionant una base sòlida per a futures millores.

5.4 Model de l'Índex de Rendiment d'Handbol (HPI)

En aquest apartat es descriu el desenvolupament d'un model predictiu dissenyat per a anticipar el resultat de partits d'handbol, utilitzant com a principals característiques els índexs de rendiment dels jugadors (HPI). A través d'un enfocament sistemàtic, es realitza la preparació de dades, construcció del model, i avaluació del seu acompliment, detallant cada pas a continuació.

5.4.1 Preparació de les dades

Les dades es van obtenir de dues fonts principals:

Jornades: Inclouen informació dels partits, com els equips locals ("Home") i visitants ("Away"), i el resultat del partit.

Vectors d'equip: Contenen el HPI per a diferents posicions dels jugadors de cada equip.

Per a garantir la coherència en les dades, es van eliminar els espais en blanc dels noms dels equips i es van unir els vectors HPI corresponents als equips locals i visitants en cada partit. Aquesta fusió va permetre crear un conjunt de dades enriquit que vincula el rendiment individual dels jugadors amb el resultat del partit.

5.4.2 Construcció i resultats del model SVM

Es va utilitzar una Màquina de Vectors de Suport (SVM) amb un nucli radial (RBF) per a la predicció. La SVM es va configurar amb un paràmetre de regularització C d'1.0 i un coeficient de nucli gamma ajustat automàticament a scale. Aquest model és adequat per a problemes de classificació binària i es va seleccionar per la seva capacitat per a manejar relacions no lineals entre les característiques.

Es va avaluar el model per a obtenir la precisió en cada partició, i es van calcular la mitjana i la desviació estàndard de la precisió global.

Les prediccions generades van permetre construir una matriu de confusió total, mostrant els veritables positius, veritables negatius, falsos positius i falsos negatius. A més, es va generar un informe de classificació que detalla mètriques clau com la precisió, recall, f1-score i el suport per a cada classe (“No victòria a casa” i “Victòria a casa”).

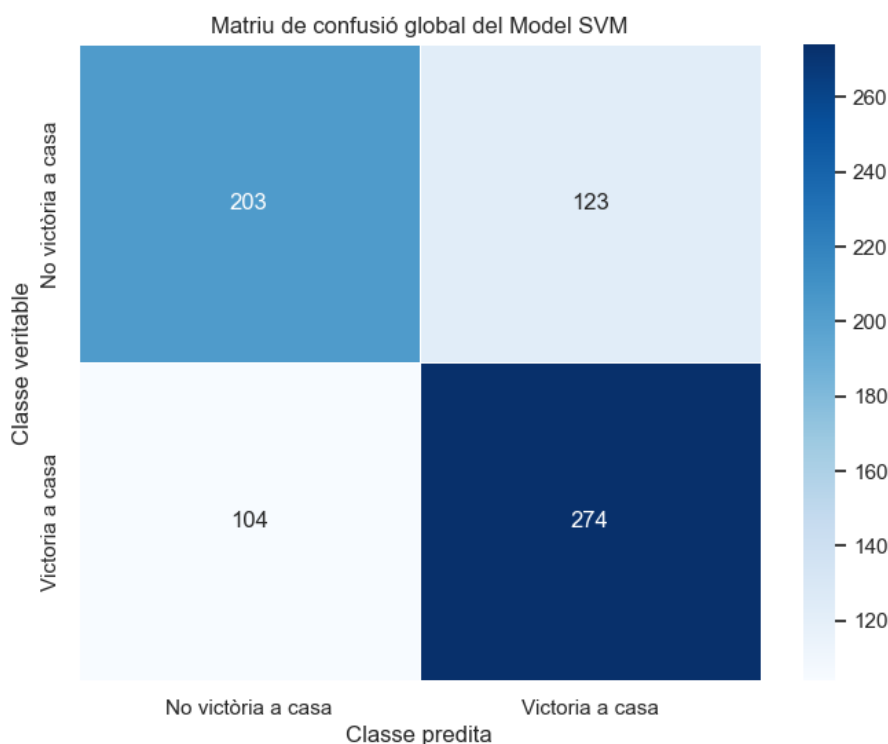


Figura 17: Matriu de confusió del model de partits.

La figura 17 ens mostra matriu de confusió del model SVM. Podem observar l’acompliment del model en predir si un equip local guanyarà o no en un partit d’handbol. Dels casos analitzats, 203 corresponen a veritables negatius, és a dir, partits on no va haver-hi victòria local i el model va predir correctament aquesta categoria. D’altra banda, es van registrar 123 falsos positius, on el model va predir una victòria local que no va ocórrer, i 104 falsos negatius, on el model no va predir una victòria que sí que va tenir lloc. Finalment, 274 casos van ser veritables positius, on l’equip local va guanyar i el model el va identificar correctament.

5.5 Comparativa i anàlisi dels resultats

	Precisió (Desviació)	Sensibilitat	Especificitat	Puntiacó F1	Puntuacions de permutació	Nº Partits del Conjunt
Model HPI PCA 100%	0.6776 (± 0.0296)	0.7249	0.6227	0.6767	0.506 ± 0.023	704
Model HPI PCA 95%	0.6818 (± 0.0333)	0.7328	0.6227	0.6808	0.503 ± 0.0235	704
Model HPI PCA 90%	0.6676 (± 0.0232)	0.7090	0.6160	0.6670	0.504 ± 0.0233	704
Model Gols	0.7206 (± 0.0273)	0.7200	0.7211	0.7180	0.501 ± 0.013	1836

Taula 8: Mètriques dels models predictius.

S’han avaluat dos enfocaments per a predir la victòria d’un equip en un partit d’handbol. Les dades recollides han estat agrupades en la taula 8 amb les mètriques dels models.

Models basats en HPI: Utilitzen l’Índex de Rendiment dels Jugadors (HPI) per a predir la victòria, organitzats en tres variants segons la reducció dimensional mitjançant PCA (100%, 95% i 90%).

Model basat en gols (Model Gols): Prediu el guanyador basant-se en els gols que s’espera que marqui un equip.

Pel que fa als tests de permutacions, tant el de HPI 18 com el de gols 19, podem confirmar que els dos són estadísticament significatius i el seu acompliment no es deu a l’atzar, la qual cosa recolza la seva vàlida predicció des d’un punt de vista estadístic.

El model de gols es destaca amb les millors mètriques predictives generals, aconseguint una precisió del 72.06% ($\pm 2.73\%$), una sensibilitat del 72.00%, i una puntuació F1 de 0.7180. Aquestes mètriques superen a les dels models HPI, la precisió dels quals es manté al voltant del 66.76%-68.18% i la seva puntuació F1 entre 0.6670 i 0.6808.

No obstant això, les tres variants del model HPI mostren resultats molt consistents entre si, amb diferències mínimes en les mètriques. Entre elles, el model HPI PCA 95% té el millor equilibri, destacant-se amb la precisió més alta (68.18%) i una puntuació de permutació superior (0.503 ± 0.0235), la qual cosa ho fa més robust.

Encara que el model de gols mostra un rendiment superior, té una limitació crítica: depèn d’una dada especulativa, els gols esperats, que si bé es una dada que es pot analitzar i intentar predir, no pot conèixer-se de manera precisa abans del partit. Això ho converteix en un model teòricament fort, però d’aplicació limitada en escenaris reals.

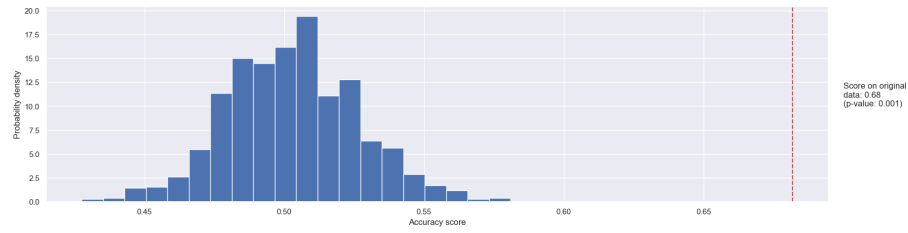


Figura 18: Test de permutació del model HPI amb PCA del 95%.

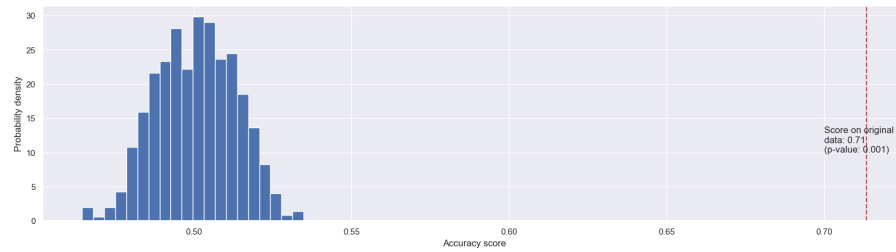


Figura 19: Test de permutació del model de gols.

En contrast, els models HPI es basen en dades accessibles i fiables, com el rendiment històric dels jugadors (HPI), la qual cosa els fa més pràctics per a prediccions abans d'un partit. Malgrat tenir mètriques lleugerament inferiors, la seva viabilitat en contextos reals és notablement major.

6 Conclusions

L'objectiu principal d'aquest treball ha estat fer una anàlisi de les dades disponibles de la Bundesliga i realitzar dos models per predir la victòria d'un equip. Mitjançant l'estudi de dos conjunts de dades, el de les dades individuals dels jugadors i el dels resultats dels partits de les temporades 21-22, 22-23 i 23-24 la Bundesliga, s'han definit les característiques importants d'un partit.

Per a fer això primer s'han recollit totes aquestes dades. S'ha fet una anàlisi exploratòria per tractar d'obtenir els factors més importants a l'hora de guanyar un partit amb l'objectiu de trobar quines característiques són més rellevants a l'hora de definir un model de predicció. Per a fer això, s'ha fet un estudi de les correlacions i s'han visualitzat diferents gràfics de característiques. D'aquesta anàlisi se'n poden treure les conclusions que, efectivament, hi ha característiques que semblen influir en el resultat final d'un partit d'handbol. Les variables que més rellevància tenen són l'Índex de Rendiment d'Handbol (HPI) juntament amb la quantitat de gols que pot marcar un equip. De les dades que es disposen de la Bundesliga, aquestes variables referents a rendiment i gols influeixen en el resultat final.

Finalment, s'han fet els experiments necessaris i els resultats reafirmen el què s'ha extret de l'anàlisi de les dades. Per a fer això, primer s'han escollit dos algorismes d'aprenentatge automàtic d'entre 5 estudiats. S'ha escollit la Regressió Logística, principalment degut a la interpretabilitat dels seus resultats i que és una predicció binària, on "Victòria" és del tipus sí/no. Per altra banda, s'ha escollit l'algorisme de Màquines de Vectors de Suport ja que és eficaç en espais d'alta dimensionalitat com és una alineació d'un partit. El més interessant que s'ha extret d'aquests experiments, és el següent:

- El model de predicció de gols aconsegueix un 72.06% d'encert per als partits.
- El model de predicció del HPI aconsegueix un 68.18% d'encert per als partits.

Vol dir això, que el model de gols és millor que el model del HPI? No, això es deu al fet que el model de predicció de gols depèn dels gols anotats, que no podem saber de forma precisa abans del partit.

Per tant, com a conclusió final d'aquest projecte, es pot extreure que: utilitzant la mesura de rendiment individual dels jugadors alineats en un partit, es pot predir quasi amb la mateixa precisió la victòria d'un equip que sabent els gols que aquest ficarà.

Per finalitzar, des d'aquest projecte, la proposta que es fa és la de dur a terme una anàlisi de les diferents lligues europees, i d'aquesta manera poder comparar les característiques més importants per predir la victòria a les diferents lligues.

6.1 Treball futur

Aquest projecte es pot ampliar en un futur de diverses maneres. Algunes de les tasques que es podrien realitzar són:

- Analitzar les dades d'altres lligues, i veure si en aquests casos l'HPI segueix sent la millor característica per predir la victòria.
- Estudiar la possibilitat d'aplicar diferents algorismes d'aprenentatge automàtic per veure si la precisió dels models millora.
- Treballar amb conjunts de dades de competicions internacionals per a comparar l'acompliment dels equips de les diferents lligues nacionals.
- Treballar amb conjunts de dades més rics, que incloguin dades dels partits per parcials de temps, i així poder avaluar si val la pena afegir més característiques per descriure els jugadors que participen al partit.

7 Bibliografia

- [1] CEROACERO. <https://www.ceroacero.es/edicion/liqui-moly-handball-bundesliga-2022-23/166339>. Accessed: 2024-10-30.
- [2] Beginners guide to handball bundesliga. <https://www.eaglebet.com/beginners-guide-to-handball-bundesliga/>, 2021. Accessed: 2025-01-11.
- [3] Correlation. jmp statistical discovery. https://www.jmp.com/en_au/statistics-knowledge-portal/what-is-correlation.html, 2024. Accessed: 2024-10-08.
- [4] What are the advantages and disadvantages of random forest? <https://aiml.com/what-are-the-advantages-and-disadvantages-of-random-forest/>, 2024. Accessed: 2024-12-04.
- [5] Handball tactical board. <https://tactical-board.com/sp/handball-half>, 2025. Accessed: 2025-01-10.
- [6] Jason Brownlee. A gentle introduction to k-fold cross-validation. <https://machinelearningmastery.com/k-fold-cross-validation/>, 2022. Accessed: 2024-12-03.
- [7] Nemanja Gojkovic. Handball bundesliga stats 2022/23. <https://www.kaggle.com/datasets/nemanjagojkovic/handball-bundesliga-stats-202223?resource=download>, 2023.
- [8] Bundesliga Handbol. Bundesliga handbol. <https://www.daikin-hbl.de/de/hbl/statistiken/handball-performance-index-die-saisonwerte-im-uberblick/berechnung-des-hpi>, 2025. Accessed: 2025-01-05.
- [9] Santi Seguí Laura Igual. *Introduction to Data Science A Python Approach to Concepts, Techniques and Applications*. Springer Nature, 2024.
- [10] Matplotlib Development Team. Matplotlib API Reference. <https://matplotlib.org/stable/api/index.html>, 2024. Accessed: 2024-12-20.

- [11] Pandas Development Team. Pandas API Reference. <https://pandas.pydata.org/docs/reference/index.html#api>, 2024. Accessed: 2024-12-20.
- [12] V. Parasaniya. Xgboost. machine learning book. https://vatsalparsaniya.com/ML_Knowledge/XGBoost/Readme.html, 2020. Accessed: 2024-12-04.
- [13] Scikit-learn Development Team. Scikit-learn API Reference. <https://scikit-learn.org/stable/api/index.html>, 2024. Accessed: 2024-12-20.
- [14] A. Soni. Pros and cons of naive bayes classifier. <https://medium.com/@anuuz.soni/pros-and-cons-of-naive-bayes-classifier-40b67249ae8>, 2020. Accessed: 2024-12-04.
- [15] Spiceworks. What is logistic regression? <https://www.spiceworks.com/tech/artificial-intelligence/articles/what-is-logistic-regression/>, 2025. Accessed: 2024-12-07.
- [16] Carnegie Mellon University. Support vector machines. <https://www.stat.cmu.edu/~larry/=sml/SVM.pdf>, 2012. Accessed: 2024-12-07.
- [17] D.H. Wolpert and W.G. Macready. No free lunch theorems for optimization. *IEEE Transactions on Evolutionary Computation*, 1(1):67–82, 1997.