TESIS DE MAESTRÍA

Inteligencia Artificial en el Fútbol: los datos salen a la cancha

Autor
Diego Adrián Castro

Director
Gustavo Denicolay

Tesis presentada en cumplimiento de los requisitos del título de ***Magíster en Explotación de Datos y Gestión del Conocimiento***

21 de Julio de 2024

Índice General

Capítulo 1

Introducción

1.1. Contexto

Durante los últimos años, el interés y la utilización de los datos ha crecido de manera sostenida en diferentes ámbitos, sobre todo en el corporativo. La reciente irrupción de los grandes modelos de lenguaje como *ChatGPT* ha acelerado este proceso y hoy la inteligencia artificial se hace presente en múltiples aspectos de nuestra vida cotidiana.

El deporte no ha sido la excepción y diferentes disciplinas han comenzado a analizar sus datos y a aplicarlos con distintos niveles de profundidad (Gong y Chen, 2023; Song et al., 2023; Melville et al., 2024). Si bien el fútbol ha empezado a utilizar estas técnicas relativamente tarde (Tuyls et al., 2021), el uso de *Football Analytics* ya está presente en los clubes más poderosos a nivel mundial (Cotton, 2022; Ntsoane, 2023). Sin embargo, esta tendencia no es exclusiva de las ligas de elite. Competiciones de diferentes latitudes y con presupuestos dispares, han incorporado este nuevo concepto y lo están utilizando para el análisis de rivales (Van Roy, Robberechts y Davis, 2021), táctica (Merhej et al., 2021) y también reclutamiento de nuevos jugadores (Pretto y De Caso, 2022).

A nivel sudamericano, Brasil lidera sin lugar a dudas esta corriente, contando cada vez con más equipos que incorporan la tecnología y la analítica en su proceso de toma de decisiones (StatsBomb, 2024; Palmeiras, 2024; O Tempo, 2021). Lamentablemente, esta adopción se encuentra en una etapa incipiente en nuestra Liga Profesional de Fútbol (LPF) de Argentina (Gantman, 2021; Bellizzi, 2020), donde aún no se han observado iniciativas firmes en ese sentido.

1.2. Objetivos

El presente trabajo se desarrolla sobre cuatro objetivos principales. En primer lugar, se realiza un profundo estudio del estado del arte de *Football Analytics*, abarcando desde la generación de datos hasta su aprovechamiento usando técnicas analíticas de distinta complejidad. En particular, se exploran variantes de los modelos estadísticos conocidos como de *Possession Value* (Sumpter, 2021; Rudd, 2011; Singh, 2018; Decroos et al., 2019; StatsPerform, 2019; StatsBomb, 2022c), los cuales buscan medir objetivamente el valor de cada acción de juego y así permitir determinar cuánto y de qué manera contribuye cada jugador dentro de un equipo. En segunda instancia, se presenta un modelo propio de *Possession Value*, bautizado **Peligro de Gol**, el cual logra capturar el concepto de *jugadas peligrosas* tal como lo haría un ojo experto. En tercer lugar, se utiliza una muestra de datos real (*Mundial de Qatar 2022*) y se analiza el recorrido del conjunto argentino mediante técnicas avanzadas, enriqueciendo este análisis con el modelo de desarrollo propio. Por último, se cierra

esta tesis con una propuesta sobre cómo podría estructurarse una *Oficina de Datos* en el Fútbol Argentino, definiendo necesidades, potenciales aplicaciones, costos y beneficios asociados, trazando una hoja de ruta realista que busque llevar las mejores prácticas de la industria a una disciplina en vías de desarrollo.

1.3. Contribuciones

Este trabajo busca mostrar el potencial de *Football Analytics* a través de su aplicación a un caso concreto de estudio y, en particular, destacar las capacidades de los modelos de *Possession Value* para detectar virtudes de jugadores que no se ven reflejadas por las típicas estadísticas empleadas en la actualidad. A su vez, se apunta a darle un marco formal al tratamiento de los datos en el mundo del fútbol, proveyendo una arquitectura que brinde una mirada integral y escalable, facilitando la interacción entre las distintas soluciones tecnológicas hoy disponibles y permitiendo su implementación en un club del fútbol argentino.

En concreto, el espíritu de este trabajo es realizar un aporte para que el Fútbol Argentino se acerque a estas técnicas y tecnologías, pudiendo contar con más herramientas para analizar rivales, detectar talento y encontrar **detalles que ayuden a ganar partidos**, lo que en definitiva desvela a todos los clubes y entrenadores.

1.4. Trabajos Relacionados

Desde hace varios años, en el fútbol mundial se ha tomado dimensión del valor y utilidad que tienen los datos aplicados a la toma de decisiones deportivas. La aplicación de técnicas de analítica avanzada puede realizarse en diferentes planos dentro del contexto de un club, abarcando desde aspectos tácticos del equipo, reclutamiento y detección de talento, gestión deportiva y hasta comportamiento de sus aficionados.

Instituciones educativas de distintas partes del mundo han mostrado interés en esta temática y hoy en día cuentan con grupos de investigación y cursos sobre estos conceptos. Entre las más destacadas se encuentra la *Universidad de Uppsala* (Suecia), la cual ha desarrollado un curso de *Modelado Matemático del Fútbol* dictado por **David Sumpter**, autor del libro *Soccermatics* (Sumpter, 2016) y referente mundial de *Football Analytics*. Otra entidad a resaltar es la *Universidad de KU Leuven* (Bélgica), la cual cuenta con un grupo de investigación muy activo en temas de analítica deportiva (más de 70 publicaciones desde 2011 a la fecha) y que ha decidido poner a disposición de la comunidad el código fuente de muchos de sus trabajos, colaborando fuertemente con el crecimiento de este campo en otras geografías.

Al mismo tiempo, han comenzado a desarrollarse cada vez más conferencias específicas alrededor de estas temáticas. La *MIT Sloan Sports Analytics Conference* es un evento anual que reúne a profesionales, académicos y líderes de la industria para explorar y discutir el papel de la analítica en el mundo del deporte. Fue fundada en 2006 por el emblemático *Daryl Morey*, conocido por su revolución analítica en los *Houston Rockets* de la *NBA*, y *Jessica Gelman*, quien trabajaba en los *New England Patriots* de la *NFL*. Por otro lado, algunas de las empresas más reconocidas dentro del mundo de recolección de datos deportivos (*StatsBomb* y *Stats Perform -Opta-*) organizan desde hace más de 5 años conferencias sobre temas relacionados al fútbol donde se presentan numerosos trabajos de investigación al respecto.

Como se mencionó anteriormente, el interés no se limita a lo que sucede dentro de un campo de juego. En ese sentido, trabajos como el desarrollado por Aichner,

2019 se centran en incrementar el *fan engagement* incorporando datos de rendimiento deportivo y combinándolos con información de las redes sociales con el fin de fomentar la participación de los hinchas antes, durante y después de los partidos.

En cuanto al marco organizativo de las áreas de analítica, trabajos como los de Altman, 2020, Left Field, 2023 y Sormaz, 2023 se han enfocado en describir productos, tendencias y proporcionar un contexto donde estas técnicas puedan aplicarse de manera exitosa.

Haciendo foco en el fútbol argentino, se ha relevado el desarrollo de interesantes trabajos donde se exploran diferentes temáticas como análisis de sentimiento de las redes sociales (Ferreyra, 2021), análisis estadístico del momento de ocurrencia de los goles (Brúgola, Durán y Farral, 2021), predicción de victorias (Tempone, 2017), análisis de futuras transferencias (Diament, 2021) o el uso de los datos para la gestión general de un club (Bernath, 2021). Ninguno de ellos se ha centrado en los modelos de *Possession Value* ni tampoco se ha orientado a realizar una propuesta concreta para un club del fútbol local.

1.5. Estructura General

Este trabajo se encuentra organizado de la siguiente manera:

- **Capítulos 2 y 3**: Estudio del Estado del Arte de *Football Analytics* (General y Modelos de *Possession Value*).
- **Capítulo 4**: Material a utilizar (conjunto de datos)
- **Capítulo 5**: Peligro de Gol, un nuevo modelo de *Possession Value*
- **Capítulo 6**: Caso de estudio del **Mundial de Qatar 2022**
- **Capítulo 7**: Propuesta de armado de *Oficina de Datos* para un club del Fútbol Argentino
- **Capítulo 8**: Conclusiones y futuras líneas de investigación

En todas las imágenes se hace mención a su fuente y, como convención, no tienen referencia aquellas que son de elaboración propia.

Capítulo 2

Estado del Arte (General)

2.1. Datos Disponibles en el Fútbol

Para comenzar a hablar de *Football Analytics*, el punto de partida obligatorio es indudablemente el de los **datos**. Existen diferentes categorías, las cuales se definen principalmente por su granularidad, disponibilidad, modalidad de captura y precio. Estos tipos de datos pueden clasificarse en:

- **Datos Generales/Planilla** (Matchsheet Data): información general de un partido como alineaciones, goles, sustituciones, etc.
- **Eventos con Pelota** (Ball Event Data): corresponde a la información *discreta* que surge de cada acción de juego con pelota como pases, conducciones, remates al arco, entre otras.
- **Datos 360** (Tracking/360 Data): relacionada a la información *continua* que captura el movimiento de los jugadores y la pelota.
- **Datos Biométricos** (Biometric Data): corresponde a mediciones fisiológicas de los jugadores como la frecuencia cardíaca, respiración o temperatura corporal.

2.1.1. Datos Generales/Planilla

Este tipo de información está relacionada a cada partido y está usualmente disponible para el público en general de manera gratuita. Se corresponde con lo que típicamente se carga en las planillas de los partidos (de ahí el nombre de *matchsheet data*). Suele comprender datos de alto nivel como las alineaciones iniciales de cada equipo, los cambios realizados, goles, asistencias, tarjetas amarillas y rojas. En algunas ocasiones, se brindan estadísticas generales sobre posesión, tiros al arco, entre otras métricas simples.

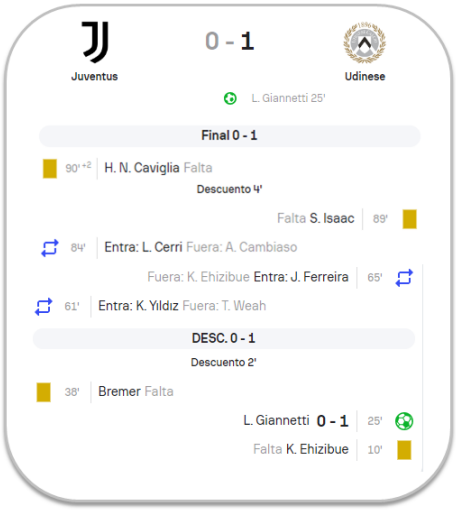


FIGURA 2.1: Datos Generales/Planilla - Matchsheet Data (Fuente: SofaScore)

2.1.2. Eventos con Pelota

Los eventos que realiza con pelota cada jugador durante un partido son capturados por este tipo de información. Incluye acciones tanto ofensivas (pases, centros, remates al arco o regates) como defensivas (quites, rechazos, atajadas o barridas). Estos datos son comercializados por proveedores específicos (*StatsBomb*, *Stats Perform*, *WyScout*, entre otros) y están disponibles para múltiples ligas de distintos países y categorías. La recolección de estos datos es **manual** y la llevan a cabo *annotators*, los cuales registran cada evento siguiendo la transmisión por televisión del partido. Típicamente, se asignan tres personas por partido (una para cada equipo, más un supervisor). Un caso excepcional fue el del *Mundial Qatar 2022*, donde la *FIFA* desplegó 25 analistas por encuentro (al menos un operador asignado a cada jugador). Esta estrategia les permitió pasar de entre 2.000 y 3.500 puntos de datos por partido a casi 15.000 (FIFA, 2021).

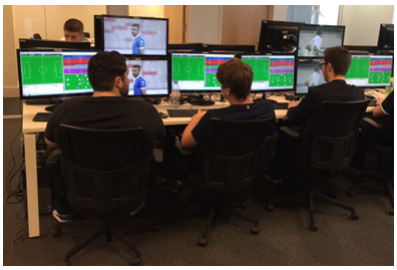


FIGURA 2.2: Carga manual de eventos realizada por *annotators* (Fuente: Stats Perform)

Si bien existen iniciativas como *SPADL* (Soccer Player Action Description Language, Decroos et al., 2019) y *Kloppy* (Kloppy, 2020) que buscan estandarizar esta información, cada proveedor tiene su propio formato y también su propia terminología. Generalmente, cada evento cuenta con los siguientes campos:

- Tipo de acción
- Jugador involucrado
- Equipo
- Timestamp (momento en que sucedió)

- Duración de la acción
- Ubicación origen y destino (coordenadas $< x, y >$)
- Parte del cuerpo con la que se ejecutó la acción
- Resultado de la acción

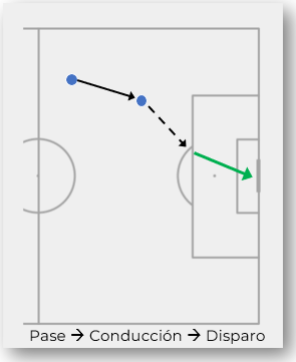


FIGURA 2.3: Eventos con Pelota (Ball Event Data)

2.1.3. Datos 360

Los eventos de *Tracking* o 360 corresponden a los datos más completos que se obtienen durante los partidos. Mediante cámaras especiales o desde la transmisión de los encuentros, se utilizan modelos de *Computer Vision* para capturar el movimiento y la posición de los jugadores y la pelota dentro del campo de juego. A partir de esta información, es posible calcular métricas como distancia recorrida, velocidad y aceleraciones de los jugadores y también se pueden determinar características colectivas como por ejemplo cambios de formaciones, ancho y largo del equipo, etc. Al igual que los eventos con pelota, estos datos son comercializados por empresas específicas como *SkillCorner*, *Metrica Sports*, *Stats Perform*, *StatsBomb*, entre otras.

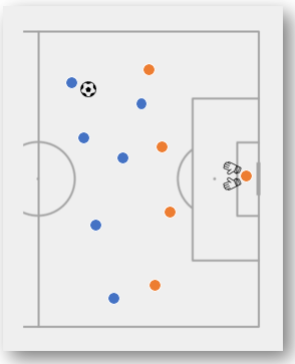


FIGURA 2.4: Datos 360

En cuanto a la precisión, las cámaras dedicadas son las que pueden tomar todo el campo de juego y brindar mayor exactitud. Sin embargo, requieren una considerable inversión y estarán disponibles solamente para los encuentros que se disputen en donde estén instaladas. Por el contrario, soluciones que se basen en videos (puede ser la transmisión del partido o una cámara táctica convencional instalada en el campo de entrenamiento) podrán capturar con un alto grado de detalle a los jugadores enfocados por la cámara y podrán inferir con un considerable margen de error a aquellos estén fuera de escena. Un punto importante a tener en cuenta es que mucha de esta información posee un control de calidad que se realiza de manera

manual, motivo por el cual los proveedores pueden demorar horas o días en brindar los datos finales.

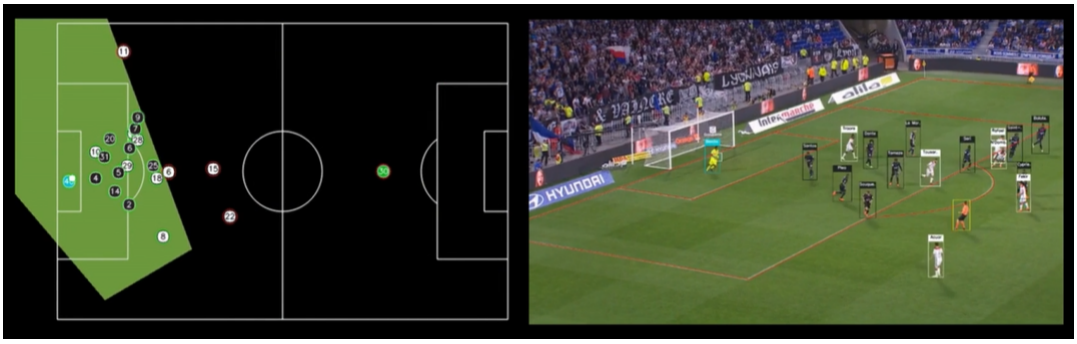


FIGURA 2.5: Captura de la posición de los jugadores que son alcanzados por la cámara y predicción para aquellos fuera de escena (Fuente: SkillCorner)

Generalmente, los eventos 360 contienen las coordenadas $< x, y >$ de los jugadores en escena y la pelota, se identifica al jugador que está realizando la acción y, para el resto de los participantes, se registra el equipo al que pertenece (sin individualizar al jugador) y si es o no el arquero.

2.1.4. Datos Biométricos

El monitoreo de los atletas no se limita solamente a su desempeño deportivo si no que también se analiza su rendimiento físico. Los datos biométricos son típicamente capturados por dispositivos *wearables* que permiten tomar mediciones fisiológicas del jugador como puede ser su ritmo cardíaco, temperatura corporal o respiración. Empresas como *Catapult*, *Oliver Sports* o *STATSports* proveen tecnología de estas características. A diferencia de otros tipos de datos que son únicamente capturados durante los partidos, este tipo de información también es registrada durante entrenamientos o períodos de descanso.

FIGURA 2.6: Una imagen ya clásica en el fútbol: jugadores luciendo *wearables* durante entrenamientos o partidos (Fuente: Leicester City)

2.1.5. Comparativa

Tal como se describió anteriormente, hay una gran variedad de datos que se generan alrededor del fútbol. A medida que se avanza en un mayor nivel de detalle, su costo irá aumentando y su disponibilidad irá disminuyendo. La disponibilidad en parte estará dada por la competición o liga en cuestión (algunos proveedores ya cubren alrededor de 100 torneos en más de 30 países) y también por el tipo de datos. Por ejemplo, la información generada por los *wearables* estará al alcance únicamente para el propio equipo, no así para los rivales.

Los *eventos con pelota* son los más frecuentes en la actualidad y cuentan con la granularidad suficiente como para hacer análisis relevantes. Gran parte de los trabajos que se vienen desarrollando en *Football Analytics* se basan en este tipo de datos.

A fin de cuentas, **los datos están y no hay excusas**. Contar o no con ellos dependerá del presupuesto que cada club esté dispuesto a invertir y del interés que posean tanto el cuerpo técnico como la directiva.

2.2. Proveedores de Datos

Actualmente existe un gran número de empresas dedicadas a la generación de datos en el fútbol. Estos proveedores no tienen como única finalidad brindar información a los clubes, sino que también ofrecen sus servicios a transmisiones deportivas, aplicaciones de celulares y, principalmente, a casas de juego o apuestas. Tal es la importancia que estos datos tienen en las casas de apuestas que muchos proveedores tienen verticales específicas que abarcan a este tan redituable negocio. En las secciones siguientes se describirá a los principales actores dentro del mundo de captura de datos del fútbol.

2.2.1. StatsBomb

StatsBomb nació como un blog de análisis de fútbol en 2013 cuando *Ted Knutson*, su fundador, comenzó a escribir con la intención de crear un lugar para centralizar interesantes análisis deportivos en Internet. Rápidamente se convirtió en un punto de referencia donde encontrar contenido de calidad basado en datos y pudo armar una sólida comunidad de analistas de todo el mundo.

Paulatinamente, clubes comenzaron a contactar a *StatsBomb* para recibir ayuda. El negocio arrancó como una consultoría para equipos, trabajando para consolidar una cultura basada en datos en organizaciones de ligas y competiciones de todo el mundo. Al tiempo, desarrollaron su primer producto *StatsBomb IQ* y pronto descubrieron que para seguir creciendo necesitaban crear sus propios datos.

A principios de 2022, *StatsBomb* se convirtió en un proveedor de datos multideportivos, abarcando también fútbol americano.

2.2.2. Stats Perform

La firma *Stats Perform* es producto de la fusión de dos empresas: *STATS*, fundada en 1981 y con orígenes en el baseball; y *Perform*, creada en 2007 y enfocada en generación de contenido deportivo. Antes, más precisamente en 2013, la empresa *Perform* había adquirido a otra reconocida compañía del mundo deportivo: *Opta*, la cual había iniciado su recorrido como proveedora oficial de estadísticas de la Premier League de Inglaterra.

En la actualidad, *Stats Perform* se posiciona como una empresa líder en tecnología y datos deportivos ofreciendo soluciones analíticas avanzadas y servicios de inteligencia artificial para organizaciones deportivas, medios de comunicación y también empresas de apuestas.

2.2.3. Wyscout

La historia de *Wyscout* se remonta a su fundación en 2004 en Génova, Italia, de la mano de *Matteo Campodonico* y *Simone Falzetti*. Su objetivo inicial era proporcionar una herramienta digital para que los clubes de fútbol pudieran tener una *base de datos global de jugadores* y así poder analizarlos de manera más eficiente.

La plataforma *Wyscout* permite a los clubes, entrenadores, agentes y gente del fútbol acceder a una amplia gama de datos estadísticos, videos de partidos, informes de rendimiento y análisis táctico. Con el tiempo, se ha convertido en una herramienta muy popular en el mundo del fútbol, utilizada por clubes de todas las ligas y niveles, desde clubes de elite hasta equipos de divisiones inferiores. La plataforma ha evolucionado continuamente, incorporando nuevas características y tecnologías para mejorar su utilidad y precisión en el análisis deportivo.

En 2019, *Wyscout* fue adquirido por la empresa estadounidense *Hudl*, la cual la ha integrado a su suite de productos.

2.2.4. FBRef

El sitio web *FBRef* fue creado por la empresa estadounidense *Sports Reference*, la cual gestiona múltiples sitios de estadísticas de diferentes deportes. Su misión principal es democratizar los datos para ayudar a que sus usuarios logren entender mejor los deportes que aman. Se caracterizan por cubrir una gran cantidad de competiciones de diferentes países y niveles y por brindar acceso **gratuito** a las mismas. La granularidad de la información dependerá en gran parte de la importancia de la liga en cuestión.

El hecho de brindar datos abiertos como realiza *FBRef* ha facilitado la aparición de librerías como *worldfootballR* (Zivkovic, 2021), la cual permite realizar web scraping y analizar datos de todas las ligas allí publicadas.

2.2.5. SkillCorner

Fundada en 2016 por *Hugo Bordigoni* y *Charles Montmaneix*, la empresa francesa *SkillCorner* viene pisando cada vez más fuerte en el mundo del fútbol.

Su plataforma se basa en sofisticados modelos de *Computer Vision* que permiten realizar tracking de jugadores, pelota y árbitro utilizando una sola cámara de video (transmisión televisiva o cámara táctica). Luego, se extraen datos físicos y eventos donde participan los atletas para poder hacer análisis de rendimiento individual o colectivo. Una característica interesante de su producto es que no sólo identifica a los jugadores que están dentro del ángulo de la cámara, si no que interpola (con cierto margen de error) las trayectorias de aquellos que están fuera de cuadro, ofreciendo una mirada completa del partido.

En sus inicios, centraron su negocio en las casas de apuestas hasta que en 2018 captaron la atención de *Ian Graham*, líder del sector de analítica del Liverpool de Inglaterra. Hoy en día abarcan a deportes como el fútbol, básquet y fútbol americano y, desde su creación, ya han brindado servicios a más de 120 clubes y federaciones.

2.2.6. Metrica Sports

Fundada en Amsterdam en 2013, *Metrica Sports* es una empresa que combina tecnología y ciencia para ayudar a entender mejor el juego sobre varios deportes y competiciones. Nació como una idea de sus tres fundadores: los argentinos *Bruno Dagnino* y *Enzo Angilletta*, y el catalán *Rubén Saavedra Pascual*, todos apasionados por el fútbol.

Un diferencial que los define es que su foco está puesto en los clubes más que en las ligas o en los medios de comunicación. Su principal producto, el cual realiza tracking de jugadores y permite hacer anotaciones sobre los videos, es conocido como *Play*.

En sus comienzos, trabajaron con el Vittese de Holanda para luego dar el salto y colaborar con el Villarreal, el Barcelona y hasta con la selección estadounidense de fútbol. Hoy en día están presentes en muchos clubes de fútbol pero también abarcan otros deportes como hockey, rugby, básquet, fútbol americano y tenis.

2.3. Visualizaciones Utilizadas en el Fútbol

Una vez definido el proveedor de datos y, con la información en la mano, llega el momento de comenzar a explotarlos. Durante esta sección se analizan distintos tipos de visualizaciones que hoy son empleadas en el fútbol, más allá de las clásicas estadísticas que pueden verse durante las transmisiones deportivas.

2.3.1. Análisis de Pases

2.3.1.1. Mapa de Pases

Los *mapas de pases* consisten en graficar sobre un campo de juego el origen y destino de cada uno de los pases de un jugador o equipo, pudiendo diferenciar por tipo como centro, asistencia, progresivo, entre otros. Sirven para caracterizar zonas habituales y estilos de pases. En algunas ocasiones, la gran cantidad de acciones vuelve compleja la lectura de estos gráficos, con lo cual se suele complementar con algún tipo de *clustering* que permita agrupar y reducir las flechas a graficar.

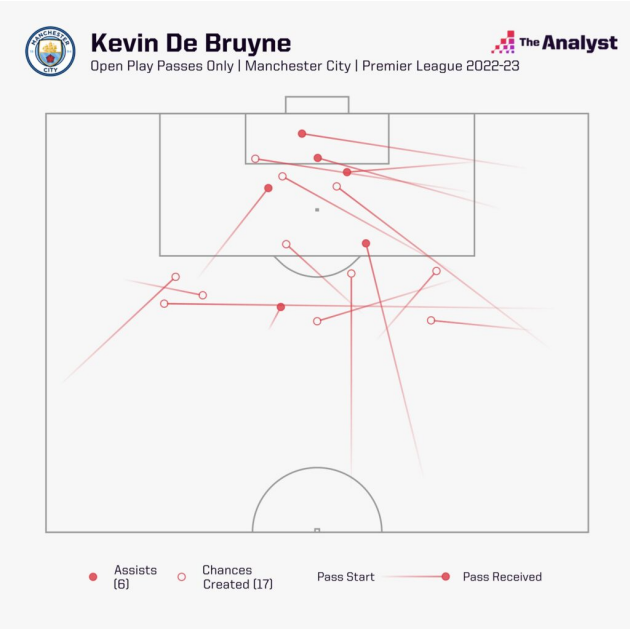


FIGURA 2.7: Mapa de pases de Kevin De Bruyne que terminaron en gol -asistencia- o en un disparo al arco -chance creada- (Fuente: The Analyst)

En los siguientes dos gráficos se puede comparar el estilo de pases en ataque de los dos laterales del Liverpool durante la temporada 2019-20 (*Alexander-Arnold* y *Robertson*).

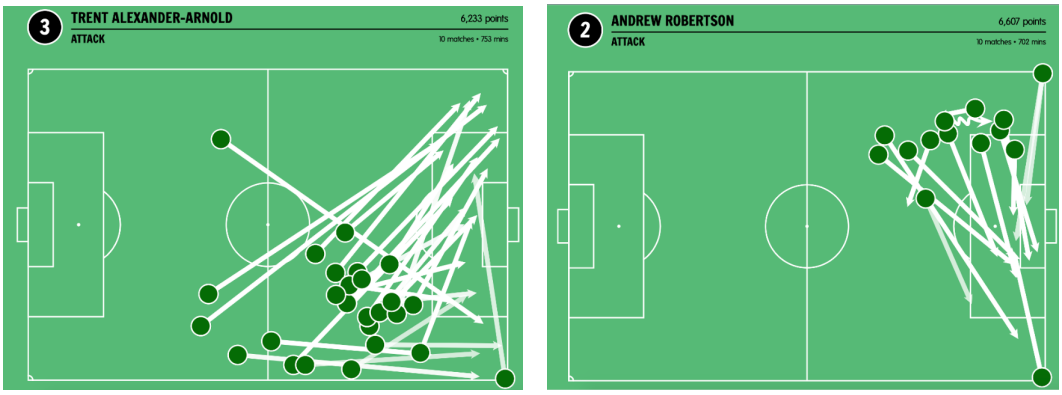


FIGURA 2.8: Mapa de pases de los laterales del Liverpool (Fuente: Soccermatics)

Mediante estas representaciones es posible observar que, por ejemplo, *Alexander-Arnold* participa en posiciones más centrales que *Robertson* y que ambos laterales tienen tendencia a lanzar pelotas cruzadas.

2.3.1.2. Mapa de Saques Laterales

Siguiendo la misma idea que la gráfica anterior, es posible centrar el análisis en los saques laterales de cada equipo.

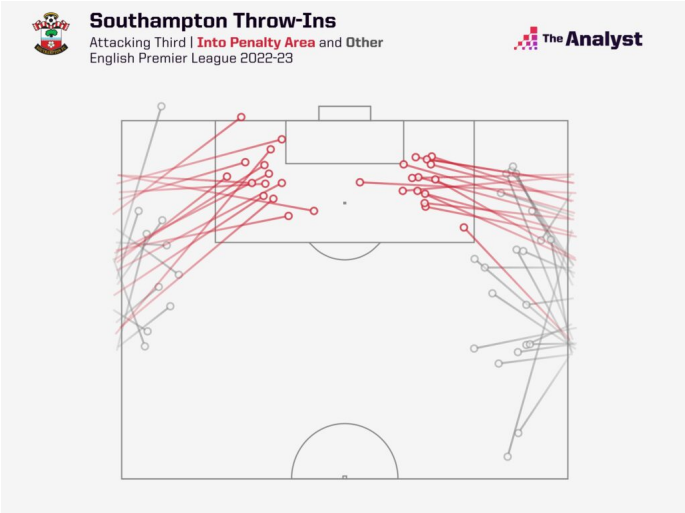


FIGURA 2.9: Mapa de saques laterales del Southampton de Inglaterra en el último tercio de la cancha (Fuente: The Analyst)

2.3.1.3. Mapa de Saques de Arco

Otra visualización interesante es la referida a los arqueros y sus saques. Los *mapas de saques de arco* representan las zonas por donde sale cada arquero. En el siguiente ejemplo se comparan dos estilos de juego de equipos distintos de la Premier League de Inglaterra (*Hugo Lloris* del Tottenham y *Nick Pope* del Newcastle).

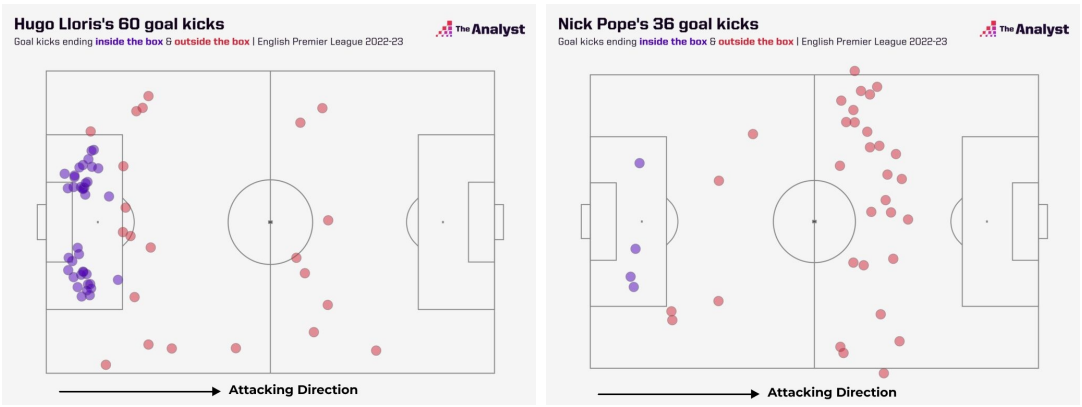


FIGURA 2.10: Mapa de saques de arco (violeta: saque con final dentro del área; rojo: saque con final fuera del área) (Fuente: The Analyst)

Aquí se puede observar una clara tendencia del Tottenham de buscar salidas cortas, en contraposición al Newcastle que prefiere saques largos para iniciar su juego desde el arco.

2.3.1.4. Red de Pases

Una *red de pases* se refiere a una representación visual que muestra los patrones de pases y las conexiones entre los jugadores de un equipo durante un partido. Proporciona información sobre cómo interactúan los jugadores entre sí en el campo, trazando los caminos de los pases entre ellos y la frecuencia entre los mismos.

Esta gráfica en el fondo está formada por un grafo, donde los nodos representan a los jugadores y las aristas denotan las conexiones entre ellos (tanto pases realizados como recibidos). La ubicación de cada jugador dentro de la cancha corresponde al

promedio de sus posiciones de sus pases realizados y recepciones. El tamaño del nodo está asociado a la cantidad de pases realizados por el jugador. Por último, el grosor de las aristas está vinculado al volumen de los pases intercambiados entre los jugadores.

Como detalle de implementación, las redes de pases sólo consideran pases exitosos y se realiza el análisis hasta la primera sustitución de un equipo. Además, a veces es necesario definir un umbral para las conexiones entre jugadores, mostrando sólo aquellas que superen los N pases.

Estas visualizaciones pueden revelar patrones de juego, como qué jugadores están más involucrados en la construcción de jugadas, qué áreas del campo se utilizan con frecuencia y cómo se distribuye la posesión entre el equipo.

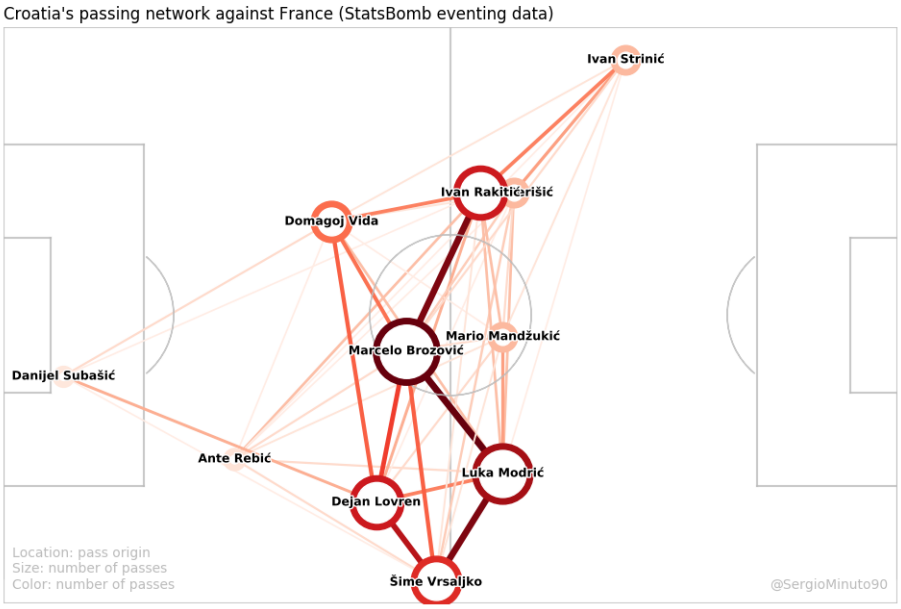


FIGURA 2.11: Red de pases de Croacia en la final de la Copa del Mundo 2018 vs. Francia (Fuente: Sergio Llana - Friends of Tracking)

Partiendo de este gráfico, es posible empezar a hablar del *índice de centralidad* de un equipo (en llano, la excesiva dependencia a uno o más jugadores). En la siguiente sección de métricas avanzadas se profundiza sobre este concepto.

2.3.1.5. Mapas de Calor de Pases

Este tipo de visualización muestra las áreas del campo desde las cuales se originan o se dirigen los pases durante un partido de fútbol. Las áreas con mayor densidad de pases se representan con colores más intensos, lo que permite identificar patrones de juego y áreas de enfoque para un equipo. Además, en algunas ocasiones se incluye una flecha que apunta al promedio o moda de la dirección adonde se dirigen los pases desde esa zona. La cantidad de pases no se toma en términos absolutos si no cada 90 minutos.

Pueden proporcionar información sobre las áreas de la cancha en las que un equipo o jugador está más activo en términos de posesión de balón y construcción de juego, así como revelar tendencias tácticas y estratégicas.

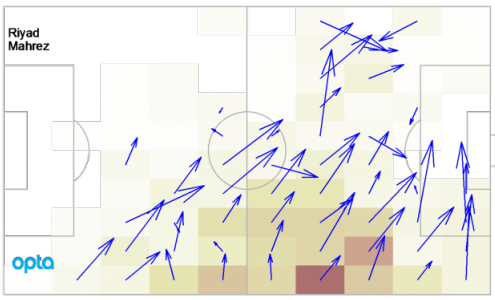


FIGURA 2.12: Mapa de calor de pases de Rihad Mahrez (Leicester) durante la temporada 2016/2017 (Fuente: Soccermatics)

En el gráfico anterior se puede observar que *Mahrez* genera mayor peligro desde el carril derecho jugando a perfil cambiado (su pierna hábil es la izquierda).

2.3.2. Análisis de Remates

2.3.2.1. Mapas de Remates

Estos mapas muestran la ubicación de los disparos realizados por un equipo durante uno o más partido. Cada punto en el mapa representa un intento de remate al arco, y el color o tamaño del punto puede indicar el resultado del disparo (por ejemplo, si fue un gol, un tiro desviado, bloqueado por un defensor, etc.). En algunas ocasiones (como en la gráfica de más adelante), el tamaño del punto es proporcional a la peligrosidad de la chance.

Representan una herramienta de análisis útil para evaluar la eficacia del ataque de un equipo, así como las áreas del campo desde las cuales tienen más éxito en crear oportunidades de gol. Al estudiar los patrones de remates en el mapa, los entrenadores y analistas pueden identificar áreas de fortaleza y debilidad en la ofensiva de su equipo o del rival, así como desarrollar estrategias para maximizar las posibilidades de convertir oportunidades en goles reales.

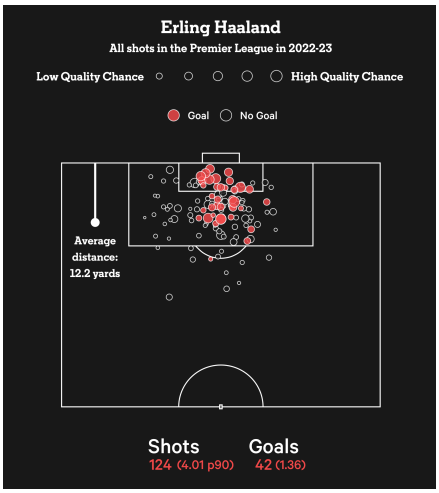


FIGURA 2.13: Mapa de Remates de Erling Haaland (Fuente: The Athletic)

2.3.2.2. Mapas de Calor de Remates

Así como existen los mapas de calor para pases, estos pueden realizarse para visualizar las zonas más frecuentes de disparos de un jugador o club.

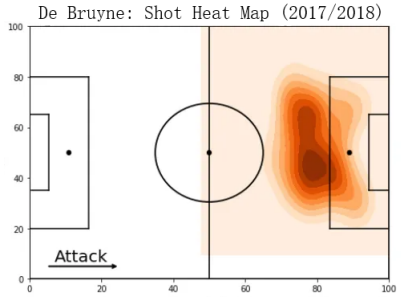


FIGURA 2.14: Mapa de Calor de Remates de Kevin De Bruyne (Fuente: Azziouni - Medium)

2.3.3. Análisis de Jugadores

2.3.3.1. Gráficos de Radar

Uno de los gráficos más utilizados para comparar jugadores es el de los *radar plots*, los cuales permiten analizar múltiples métricas al mismo tiempo. Se ha vuelto muy popular entre los amantes del deporte desde su utilización en videojuegos relacionados al fútbol. Existe una gran cantidad de variantes, pero se toma como ejemplo la utilizada por *Opta* (Whitmore, 2023a).

Los gráficos de radar muestran diferentes porciones para presentar la destreza relativa de cada jugador en cada una de las métricas. Típicamente no se utilizan más de 9 porciones para facilitar su lectura. La distancia de cada métrica con su centro representa el percentil del jugador comparado contra un conjunto de atletas (el cual variará según cada implementación). En el caso de *Opta*, toma en cuenta todos los jugadores de las ligas Big 5 (Inglaterra, España, Italia, Alemania y Francia) de los últimos 15 años.

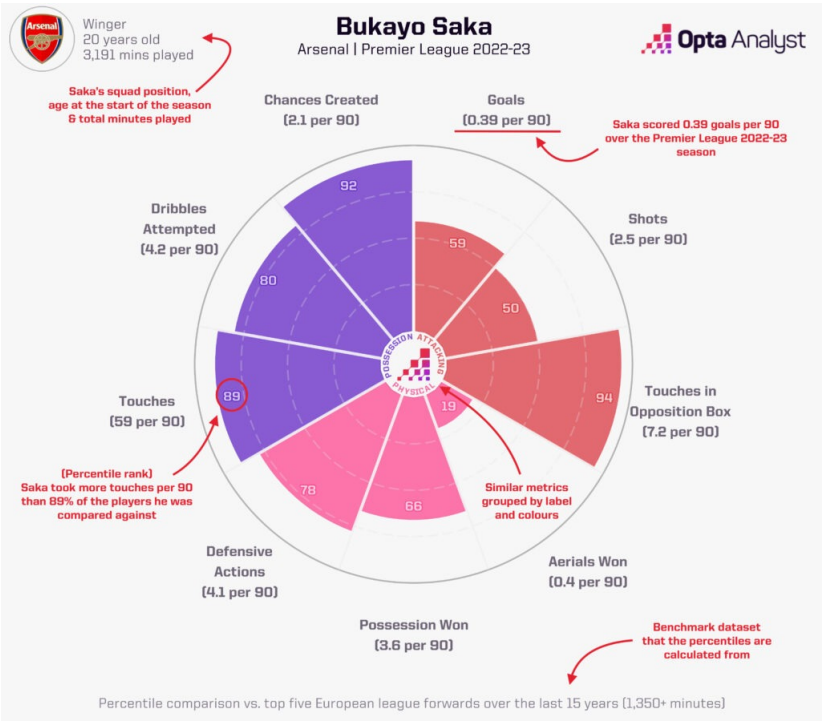


FIGURA 2.15: Gráfico de Radar de Bukayo Saka del Arsenal con sus respectivas explicaciones (Fuente: The Analyst)

En la gráfica anterior puede notarse que *Saka* tiene un rating de 92 en cuanto a chances creadas (2.1 cada 90 minutos), lo que significa que habilitó a sus compañeros

más que el 92 % de los delanteros en las mejores ligas europeas durante los últimos 15 años.

Como puede verse a continuación, los radares son muy útiles para comparar estilos de juego entre jugadores como *Erling Haaland* del Manchester City y *Gabriel Jesus* del Arsenal durante la Premier League 2022/23.

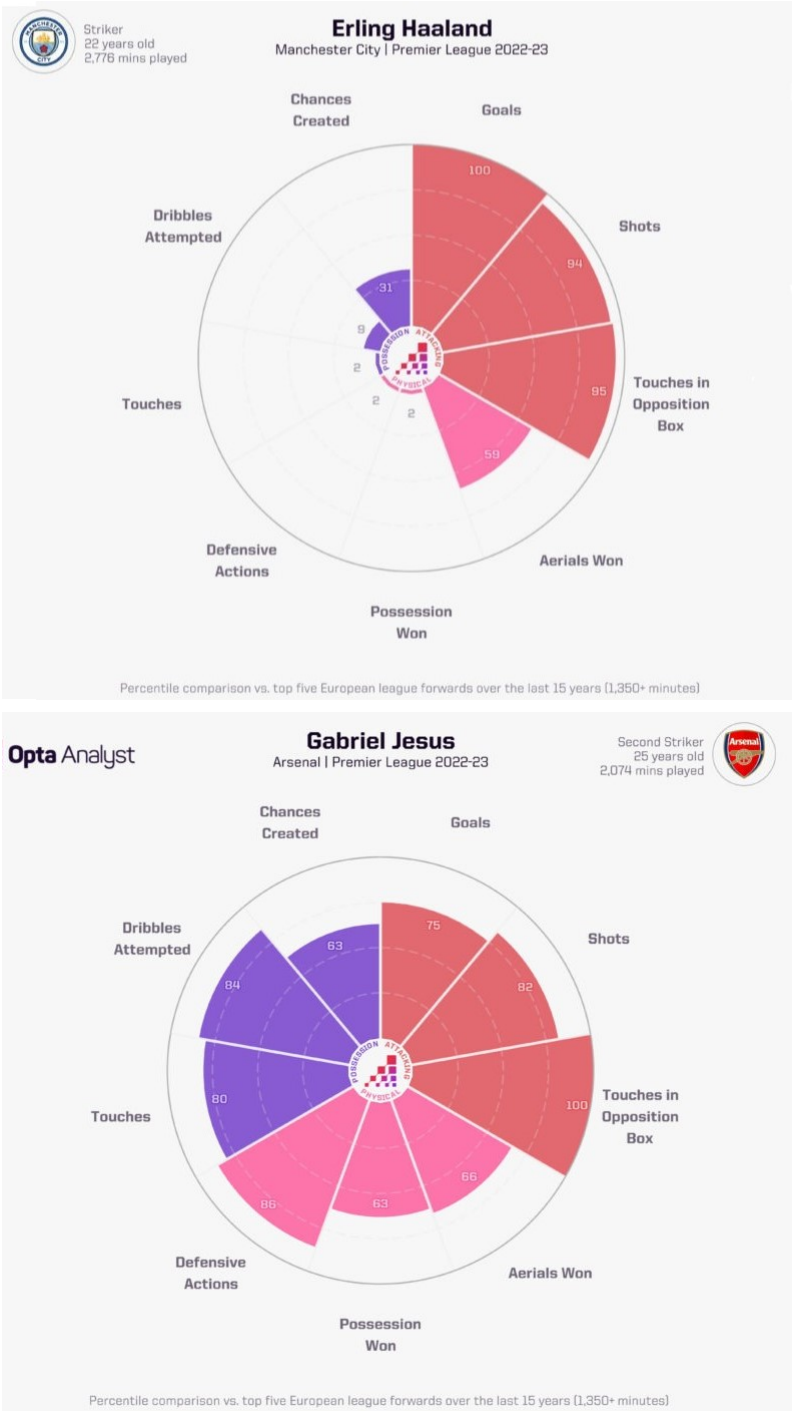


FIGURA 2.16: Gráfico comparativo entre Erling Haaland y Gabriel Jesús (Fuente: The Analyst)

A partir de lo anterior, es muy claro ver cómo *Haaland* tiene un rol de finalizador sin injerencia en el armado de juego o en la parte defensiva, lo que contrasta con las contribuciones más repartidas de *Gabriel Jesús*.

En cuanto a aplicaciones, estas representaciones pueden ser muy útiles para buscar reemplazo de jugadores de similares características, para análisis de rivales y también para medir las cualidades de un jugador en un contexto en particular.

2.4. Métricas Avanzadas

A lo largo de esta sección se analiza una serie de métricas avanzadas que pueden ayudar a entender aspectos del juego que no se observan a primera vista.

2.4.1. Índice de Centralidad

Cuando se estudiaron las redes de pases, se mencionó la existencia de un grafo de manera subyacente. El objetivo de esta métrica es medir el grado de *centralidad de un grafo* tal como lo describe Sumpter, 2022.

Dados N nodos (jugadores), w_{ij} las aristas entre los nodos i y j (representando la cantidad de pases del jugador j al i), se define el *índice de centralidad* (Grund, 2012) como:

$$C = \frac{\sum_{i=1}^{N}(P_* - P_i)}{(N-1)\sum_{i=1}^{N} P_i}$$

donde $P_i = \sum_{j=1}^{N} w_{ij}$ es la cantidad total de pases hacia el jugador i y $P_* = \text{máx}(P_i)$ representan la cantidad de pases recibidos máxima (el jugador que más pases recibió).

A continuación se presentan dos ejemplos de los casos extremos. El primero de ellos consiste en 5 jugadores donde todos los pases van a la misma persona.

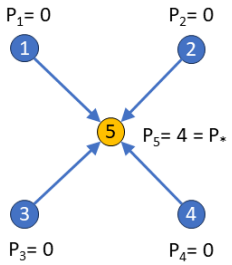


FIGURA 2.17: Índice de Centralidad Máxima

Considerando $N = 5$, la fórmula queda:

$$C = \frac{(4-0)+(4-0)+(4-0)+(4-0)+(4-4)}{(5-1)\times 4} = \frac{16}{16} = 1$$

Ahora, un ejemplo con 4 jugadores y dos pases cada uno.

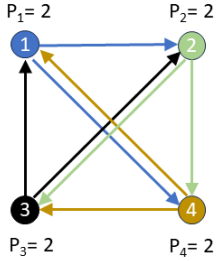


FIGURA 2.18: Índice de Centralidad Mínima

Considerando $N = 4$, la fórmula queda:

$$C = \frac{(2-2)+(2-2)+(2-2)+(2-2)}{(4-1)\times 2} = \frac{0}{8} = 0$$

De esta manera, es posible calcular esta métrica para un equipo a lo largo de la temporada para ver cómo va desarrollando su idea de juego, detectando puntos fuertes y debilidades y pudiendo definir estrategias para contrarrestarlos o aprovecharlos.

2.4.2. Gol Esperado

Los goles son los sucesos más importantes en un partido de fútbol pero son también los eventos más infrecuentes. En las grandes ligas, el promedio de gol por partido se sitúa entre 2.5 y 3 goles, mientras que la cantidad de tiros entre 20 y 30 (diez veces más). Sin embargo, no todos los remates son iguales. Los *goles esperados* surgen como una manera de medir la calidad de cada remate. No está claro quién fue el creador de esta métrica pero uno de los primeros en plantearlo fue *Sam Green* de Opta en 2012.

Actualmente, los *Expected Goals* (o usualmente **xG**) representan a una de las métricas más conocidas en el análisis del fútbol. En términos básicos, mide la probabilidad de que un tiro dado termine en gol. Cuando se dice que un disparo tuvo un xG de 0.1 (o 10 %), se refiere a que típicamente los remates de este tipo se convertirán en gol 10 de cada 100 veces. Se pueden considerar como la probabilidad de que en un día típico de fútbol de un jugador promedio, un disparo en particular desde esa ubicación resulte en gol.

2.4.2.1. Construcción y Factores Influyentes

Un modelo de xG utiliza información histórica de miles de tiros con características similares para estimar la probabilidad de gol en una escala de 0 a 1. No existe una única implementación de xG (cada proveedor de datos tiene sus particularidades), pero los factores que tradicionalmente han formado parte de la gran mayoría de los modelos son: distancia al arco, ángulo respecto al arco, parte del cuerpo con la que se realiza el tiro y tipo de asistencia o acción previa. Los porcentajes también difieren ligeramente entre mujeres y hombres, y también entre ligas distintas. Sin embargo, la regla más importante es que, cuanto más cerca del arco está un atacante y más del arco puede ver (ángulo), mejor será la oportunidad de marcar. La forma que tienen las probabilidades de anotar un gol se asemejan a círculos con sus costados aplastados (Sumpter, 2016), como los de la siguiente figura:

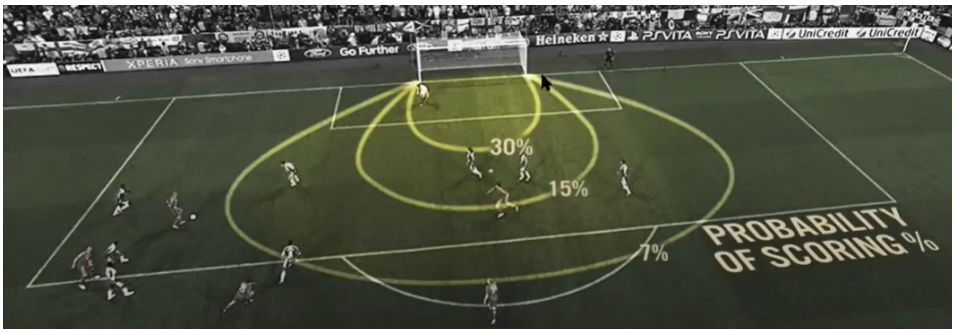


FIGURA 2.19: Las formas que mejor describen las probabilidades de marcar un gol (Fuente: Soccermatics)

Un caso particular es el de los penales, donde la gran mayoría de los modelos de goles esperados asignan un valor estático que puede variar entre 0.75 y 0.79 xG en línea con la tasa de conversión histórica.

2.4.2.2. Aplicaciones

Existen diferentes aplicaciones relacionadas a esta métrica. Por ejemplo, cuando un jugador o equipo marca más goles que los esperados durante un período sostenido, puede sugerir que tienen una habilidad para definir jugadas particularmente buena. En períodos más cortos, los datos podrían indicar que el jugador o equipo está atravesando una racha positiva cuando su confianza es alta. Entonces, es posible que dejen de anotar tantas oportunidades difíciles cuando su buena forma termine. Según describe *David Sumpter* en su curso *Soccermatics* (Sumpter, 2022), al principio de la temporada los goles esperados correlacionan bien con el rendimiento del equipo (más allá de su posición en la liga) pero, a partir de los 15 partidos, los goles reales son los que mejor describen su rendimiento. En cualquier caso, esta métrica no explica por qué las cosas van bien o mal, eso siempre precisará un análisis de mayor profundidad.

Otro ejemplo donde esta métrica puede resultar útil es para comparar los remates de jugadores distintos como hace *The Analyst* en el artículo Whitmore, 2023b.

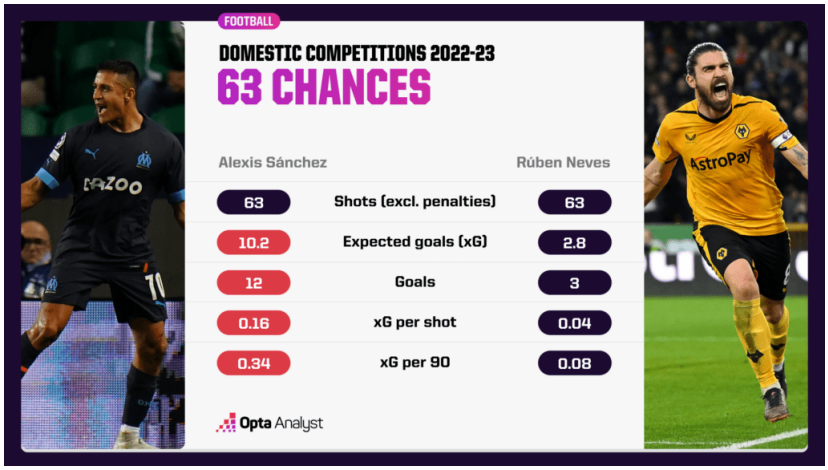


FIGURA 2.20: Comparativa de dos jugadores con igual cantidad de remates en la temporada (Fuente: The Analyst)

Al medir la calidad de las 63 oportunidades para cada jugador, el xG agrega contexto adicional a sus tiros que va más allá de las métricas tradicionales como remates al arco o distancia promedio, pudiendo comparar la calidad de las ocasiones que tuvo cada jugador.

De las oportunidades que tuvo *Alexis Sánchez*, esperaríamos que un jugador promedio anotara alrededor de 10 goles (10.2 xG). Por otro lado, de las de *Neves*, nuestra expectativa sería de tres goles (2.8 xG).

Observando los mapas de sus tiros a continuación, podemos entender inmediatamente por qué su producción goleadora fue tan diferente. Ambos jugadores rindieron ligeramente por encima de su xG, pero la calidad de sus 63 oportunidades fue muy distinta. *Sánchez* disparó desde posiciones mucho más cercanas al arco y *Neves* lo hizo desde posiciones más alejadas, de menor calidad.

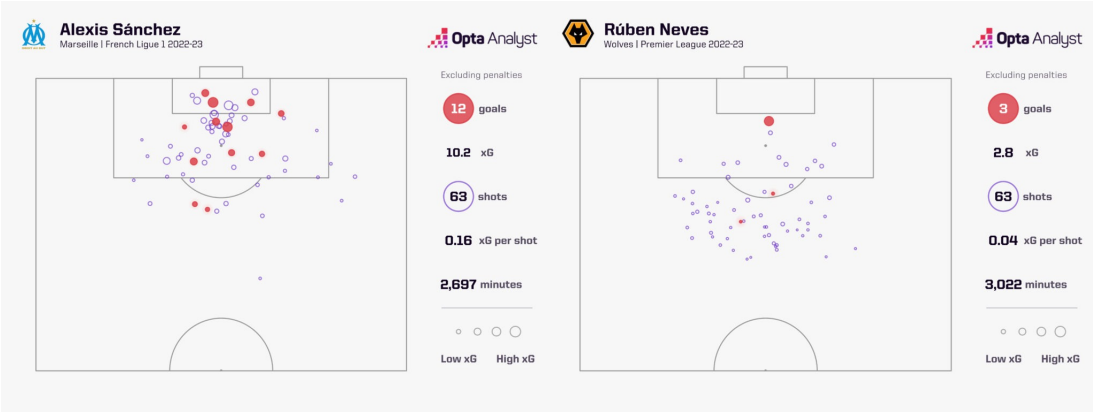


FIGURA 2.21: Mapa de tiros al arco de Sánchez y Neves (Fuente: The Analyst)

2.4.2.3. Errores Comunes

xG a Nivel de Partido

Las principales críticas a los goles esperados suelen aparecer en escenarios donde la métrica no se aplica correctamente. La más común se da a nivel de partido. Un equipo que tiene un total de xG más alto en un encuentro no necesariamente implica que debió haber ganado. El xG sólo mide la calidad de la oportunidad y no el resultado esperado del partido.

Rendimiento Superior al xG

Un jugador o equipo que ha estado superando su xG no tiene por qué rendir por debajo a futuro para *regresar a la media*. Este es un concepto conocido como la *falacia del apostador*, producto de las muestras pequeñas como sucede en el fútbol. Cuantos menos datos se tienen, mayor margen para la aleatoriedad existe. Cada partido, cada oportunidad y cada tiro son eventos independientes. El xG nos da una idea de la probabilidad de que un jugador o equipo anote un gol en una oportunidad específica, pero no nos dice nada sobre la probabilidad de que anoten en siguientes jugadas.

En el caso de jugadores que vienen superando su xG, si bien esperaríamos que vuelvan a anotar de acuerdo con su xG en tiros futuros, ya han de alguna forma *acumulado* este rendimiento superior. Si un jugador al comienzo de la temporada ya ha marcado cinco goles más que su total de xG, es probable que termine la temporada superando su comportamiento esperado por esos cinco goles. De la misma manera, si un lanzamiento de moneda sale cara diez veces seguidas, los lanzamientos futuros aún tienen la misma probabilidad de salir cara o ceca, pero las diez veces que la moneda salió cara ya han sucedido.

2.4.3. Peligrosidad/Defensa Esperada

Como se mencionó previamente, los goles y los remates al arco son eventos poco frecuentes en los partidos de fútbol si los comparamos con el resto de las acciones. Es por este motivo que se diseñaron métricas para evaluar al resto de los eventos y que apuntan a medir cómo una acción aumenta o reduce la probabilidad de que un equipo anote un gol. Éstas se basan en modelos estadísticos conocidos como de **Possession Value**, los cuales buscan medir de manera objetiva y cuantitativa el valor

de cada acción que ocurre en el campo de juego. De esta forma, los equipos y analistas pueden entender mejor el impacto de las acciones de sus jugadores y tomar decisiones basadas en datos para mejorar su rendimiento.

En este trabajo, se toma la misma postura que menciona *David Sumpter* en su curso *Soccermatics* (Sumpter, 2021): se utiliza el nombre *xT* (Peligrosidad Esperada) para referirse a cualquier métrica que hable de las acciones de ataque y *xD* (Defensa Esperada) para cualquier medida sobre las acciones de defensa. En la actualidad conviven muchas definiciones alternativas y, para poder compararlas entre sí, *Sumpter* adopta este atinado criterio.

A diferencia de los goles esperados (xG), los cuales sólo se centran en la probabilidad de que un tiro se convierta en gol, el xT tiene en cuenta todos los tipos de acciones de ataque, como pases, regates, centros y disparos, lo que la convierte en una métrica más completa del valor de la posesión del balón.

FIGURA 2.22: Un pase y su contribución ofensiva en base a la probabilidad de gol (Fuente: Statsbomb)

En el siguiente capítulo se analiza el tema con muchísima más profundidad, estudiando su historia y sus diferentes alternativas.

2.5. Football Analytics en Acción

El cierre de este capítulo se completa con una serie de historias donde la aplicación de *Football Analytics* reportó beneficios a los clubes o jugadores que lo emplearon.

2.5.1. Monchi y el impacto de la tecnología en el scouting

FIGURA 2.23: Monchi durante su estadía en el Sevilla (Fuente: Estadio Deportivo)

El Sevilla, de la mano de su director deportivo *Monchi*, fue precursor en España al usar la estadística y los datos para encontrar a los mejores futbolistas y aplicar una estrategia sencilla (y al mismo tiempo compleja): comprarlos baratos y venderlos caros.

Según cuenta la nota del diario *El Confidencial* (Villarreal, 2019), el Sevilla necesitaba un central y, para encontrarlo, *Monchi* y su equipo comenzaron a seleccionar entre una base de datos de entre 3.600 y 4.500 futbolistas, provenientes de 40 a 45 ligas en todo el mundo. Utilizaron criterios específicos, como el porcentaje de duelos aéreos ganados y la precisión en pases largos, para reducir la lista de candidatos a 800, luego a 200 y finalmente a uno. El elegido fue *Diego Carlos*, un central brasileño de 26 años, por quien pagaron 15 millones de euros al FC Nantes. Tres temporadas después, ese jugador fue vendido por 31 millones de euros al Aston Villa de Inglaterra. Esta metodología, impulsada por *Monchi*, ha sido fundamental en el éxito del Sevilla, llevándolo a la elite del fútbol mundial desde el año 2000.

Sevilla ha sido pionero en la implementación del *Football Analytics* poniendo en práctica un exitoso sistema que combina datos estadísticos con la experiencia humana para la evaluación y fichaje de jugadores, impulsando a otros clubes a seguir en la misma dirección.

2.5.2. Brighton y Brentford, fútbol y el mundo de las apuestas

FIGURA 2.24: Brighton y Brentford, dos nuevos protagonistas de la Premier League (Fuente: Antena 2)

Brighton y *Brentford* son dos equipos de la Premier League de Inglaterra que hace pocos años comenzaron a aparecer en el radar del público.

En el caso del Brighton, logró su ascenso a la primera división en 2017 luego de 34 años de ausencia. En la temporada 2022/2023 finalizaron en el sexto lugar. Por su parte, Brentford es un club que debutó en la primera categoría de Inglaterra en la temporada 2021/2022, luego de 74 años en divisiones menores. Entre ambos hay un factor en común: el juego. Brighton fue adquirido por *Tony Bloom* y Brentford, por *Matthew Benham*, ambos empresarios del mundo de las apuestas.

Estos clubes se percataron de que iba a ser imposible competir a nivel presupuestario contra los grandes equipos de Inglaterra como Manchester City o Chelsea. Entonces, decidieron invertir en poner a los datos en el centro de sus organizaciones para poder tomar decisiones más inteligentes (Analyst, 2023). La experiencia de sus dueños en el mundo de las apuestas facilitó la incorporación de modelos matemáticos y de *Machine Learning* en diversos aspectos de la vida del club.

Al igual que el mencionado caso del Sevilla en la sección anterior, estos dos clubes se han caracterizado por identificar tempranamente el talento, desarrollarlo y vender a los jugadores en un valor superior al invertido. Sin embargo, ambas organizaciones han diseñado una estrategia y un sistema que evita poner foco en el individuo, buscando personas que se ajusten a la filosofía del club y no a la inversa.

2.5.3. Klopp, su mala suerte y el desembarco en Liverpool

FIGURA 2.25: El entrenador Klopp junto al director de investigación Ian Graham (Fuente: Emol)

Indudablemente, *Liverpool* es uno de los clubes más emblemáticos de Inglaterra. Tuvo una época dorada en los años 70 y 80, ganando numerosos títulos. En 2007 fue comprado por empresarios estadounidenses con poco interés en el fútbol, lo que lo llevó a problemas financieros y deportivos. En 2010, *Fenway Sports Group*, un grupo estadounidense dueño de los *Boston Red Sox*, compró el Liverpool y decidió invertir fuertemente en análisis de datos para tomar decisiones.

Ian Graham es un doctor en Física teórica, egresado de la *Universidad de Cambridge* y transcurrió gran parte de su vida lejos del fútbol. Desde la temporada 2012/2013 se incorporó al Liverpool y fue su director de investigación por una década, hasta que decidió abrirse camino y crear su propia empresa de servicios analíticos deportivos (*Ludonautics*). Construyó un modelo para evaluar jugadores y predecir resultados usando datos. Aprovechando estos modelos, Liverpool incorporó a numerosos jugadores que han rendido a altísimo nivel (*Mohamed Salah, Mané, Roberto Firmino, Virgil van Dijk*, entre otros) y que, desde el punto de vista de un entrenador o un ojeador, no siempre eran las primeras alternativas a considerar.

En los artículos de *NY Times* (Schoenfeld, 2019) y *El País* (Álvarez, 2023) se cuenta la historia acerca de la incorporación del exitoso técnico alemán *Jürgen Klopp* al Liverpool. En el 2015, el club tenía en carpeta a varios entrenadores y uno de los apuntados era el alemán que venía dirigiendo al Borussia Dortmund de la Bundesliga alemana. Si bien *Klopp* había tenido buenas temporadas al mando del club alemán, en la última había cosechado magros resultados, finalizando en 7mo lugar y a 33 puntos del campeón Bayern Munich. A pesar de ello, luego de estudiar los datos de sus últimas temporadas, *Graham* recomendó su incorporación. Analizando a sus jugadores y al rendimiento colectivo, supo que el Borussia Dortmund había merecido ganar muchos partidos en los que la fortuna no los había acompañado.

A las tres semanas de haber arribado al Liverpool, *Klopp* se reunió con *Graham*, quien acudió a la cita con todos sus papeles de trabajo y con la esperanza de convencerlo acerca del valor de sus análisis a un entrenador que hasta ese entonces no los había utilizado. *Graham* le comentó las conclusiones sobre la última temporada del alemán con el Borussia Dortmund. Según sus estudios, la temporada no había finalizado como debía ser, pero no por culpa del entrenador. *Klopp* solamente había estado a la cabeza de uno de los equipos con menos suerte en los años recientes. Le contó lo que había observado, sin mirar apenas una imagen de sus partidos. *Graham* relató: "Creo que Jürgen no se había dado cuenta del todo de cuánta mala suerte habían tenido, así que psicológicamente fue una reunión importante. Los datos son ese factor de calma que dicen que lo que importa es la trayectoria a largo plazo, no si ganamos o perdimos ese partido en concreto". Esta reunión de alguna forma convenció a *Klopp* de darle una oportunidad al enfoque de *Graham*.

Durante su estadía en el Liverpool, el alemán *Klopp* logró cosechar 8 títulos (incluyendo una *Champions League*).

2.5.4. Renegociación de De Bruyne con Manchester City

FIGURA 2.26: De Bruyne en la renovación de su contrato con el Manchester City (Fuente: Publimetro)

Kevin De Bruyne es un mediocampista de origen belga, considerado uno de los mejores en su posición a nivel mundial. Es parte del primer equipo del Manchester City desde el año 2015 y juega en la selección de Bélgica desde 2010. Su estadía en el Manchester City ha sido realmente exitosa, consiguiendo una gran cantidad de títulos incluida la Champions League en 2023.

Corría el año 2021 y habían comenzado las tratativas por la renovación del contrato de *De Bruyne*, una de las estrellas de la institución. El jugador, sin representante en ese entonces, decidió apelar a una estrategia novedosa. Contrató los servicios de una empresa especializada en consultoría deportiva, *Analytics FC* (FC, 2021), para calcular su valor y mostrar su injerencia dentro del elenco inglés.

Fue la primera vez que un jugador solicitaba un análisis detallado de su rendimiento para fundamentar sus pretensiones salariales. El análisis comparó el desempeño de *De Bruyne* con otros jugadores top de Europa, considerando tanto sus contribuciones en la cancha como su valor financiero relativo. Se determinó, usando un modelo propio creado por *Analytics FC*, que *De Bruyne* era el mejor generador de oportunidades de gol en Europa. Además, el análisis comparó su salario con el de otros jugadores atacantes de primer nivel, demostrando que *De Bruyne* producía más que algunos de ellos pese a ganar menos.

Este informe detallado ayudó a convencer al Manchester City de la valía de *De Bruyne*, lo que derivó en un nuevo contrato por cinco años y £104 millones. Esta situación captó la atención de los medios y podría impulsar a otros futbolistas a tomar control de sus propios datos.

Capítulo 3

Estado del Arte (Modelos de Possession Value)

3.1. Introducción

Es indudable que los goles representan el evento más importante de un partido de fútbol. Más allá de las preferencias futbolísticas, estilos de juego o estrategias, los goles, en definitiva, son las acciones que terminan definiendo quién gana un encuentro o un campeonato. Sin embargo, muchas veces no reflejan o explican lo que sucede en un campo de juego. Hasta este punto se han explorado distintos enfoques (desde visualizaciones a métricas) que buscan dar una mirada más profunda sobre lo que aconteció durante un partido. En particular, se analizó la medida de *xG*, la cual hace referencia a los remates y su probabilidad de anotar un gol. Sin embargo, sólo el 1 % de las acciones de un partido son tiros (entre 20 y 30 vs. hasta 3.500 eventos). Entonces, ¿cómo medir la injerencia del 99 % restante de las acciones? Aquí es donde entran en escena los modelos de **Possession Value**.

Comencemos tomando como ejemplo dos momentos puntuales del último Mundial 2022. El primero corresponde al minuto 86 del partido *Argentina vs. México* y el segundo al minuto 122 de la final *Argentina vs. Francia*.

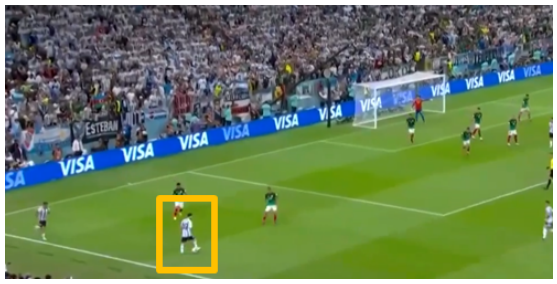


FIGURA 3.1: ARG vs. MEX, minuto 86, Messi con la pelota cerca del córner (Fuente (fotos): FIFA+)

FIGURA 3.2: ARG vs. FRA, minuto 122, Konaté recibiendo un rechazo de la defensa argentina (Fuente (fotos): FIFA+)

Ahora observemos cómo continúa la secuencia de acciones en cada caso.

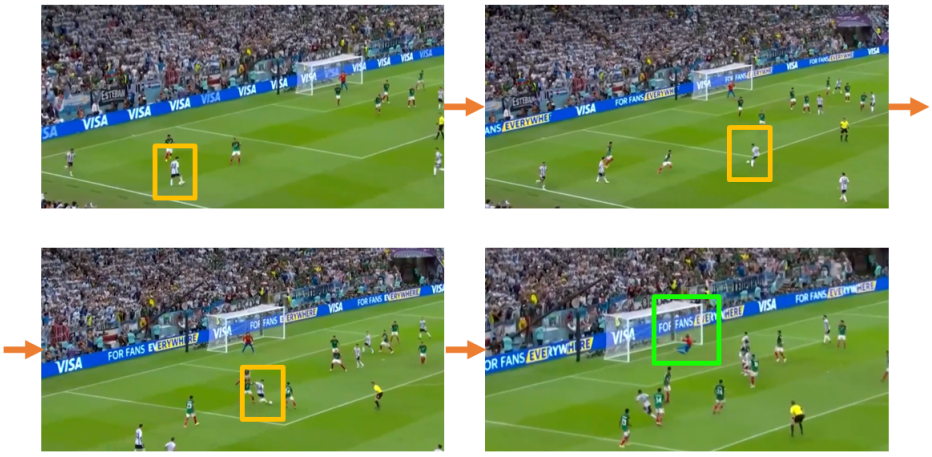


FIGURA 3.3: Pase de Messi, amague de Enzo Fernández, remate y gol de Argentina (Fuente (fotos): FIFA+)

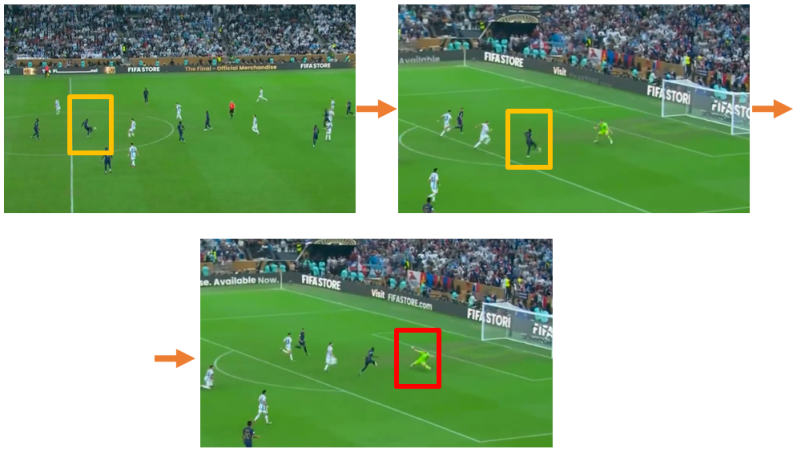


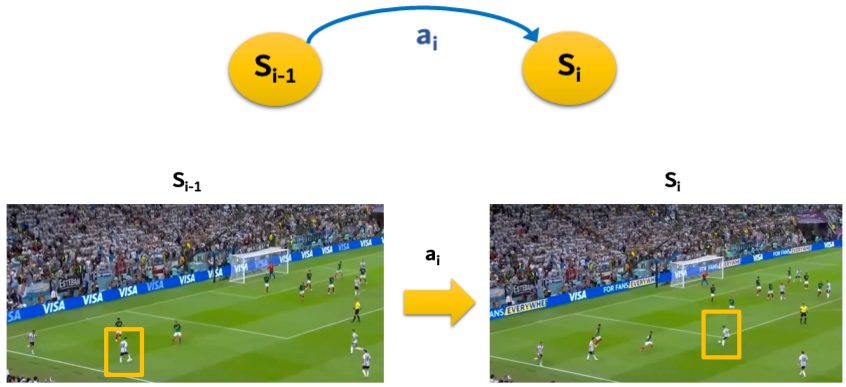
FIGURA 3.4: Pase largo de Konaté, remate de Kolo Muani y monumental atajada de Dibu Martínez (Fuente (fotos): FIFA+)

Las estadísticas que suelen aparecer en transmisiones deportivas o medios de comunicación considerarían la primera acción (pase de *Messi* a *Fernández*) como una asistencia, mientras que el nombre de *Konaté* ni siquiera sería parte de las crónicas deportivas. Sin embargo, más allá del desenlace de cada jugada, la peligrosidad generada por el pase de *Messi* parece ser significativamente menor que la provocada por el pase/pelota larga de *Konaté*.

En este contexto es donde cobran relevancia los modelos de *Possession Value*, los cuales buscan establecer la probabilidad de anotar un gol a partir de una posesión. Son modelos estadísticos que sirven para medir *objetivamente* el valor de cada acción en un partido de fútbol, permitiendo ponderar *cuánto* contribuye cada jugador al rendimiento del equipo y *cómo* hace este aporte. Las principales dificultades son dos: por un lado, poder determinar el efecto de cada acción de juego sabiendo que su resultado puede no ser inmediato; por otro lado, poder juzgar a un evento independientemente del desenlace que tenga la jugada. En este punto es donde el *Machine Learning* tomará protagonismo y, junto a un conjunto de definiciones y decisiones de implementación, permitirá evaluar positiva o negativamente cada acción de juego.

Si bien en la actualidad conviven diferentes modelos de *Possession Value*, cada uno con distintas características y definiciones, todos tienen una base común que se puede resumir en los siguientes puntos:

- Un partido es modelado como una secuencia de estados $[S_1, \ldots, S_i, \ldots]$ y una secuencia de acciones de juego $[a_1, \ldots, a_i, \ldots]$, donde cada una de ellas genera una transición desde el estado S_{i-1} al estado S_i.



- Las acciones a considerar no se limitan a los pases y los remates, si no que se tienen en cuenta otros tipos de eventos como conducciones, despejes, centros, etc.
- La valorización busca ser independiente del resultado final real de la posesión
- Trabajan principalmente sobre *ball event data* debido a su disponibilidad y alto nivel de detalle

Como menciona *David Sumpter* en su curso *Soccermatics* (Sumpter, 2021), hoy conviven muchas métricas similares pero con nombres muy distintos. Por este motivo, en este trabajo se sigue su clasificación propuesta:

- **xT** *(Expected Threat / Peligrosidad Esperada)*: para acciones de ataque como pases, conducciones, remates, etc.
- **xD** *(Expected Defense / Defensa Esperada)*: para acciones defensivas como bloqueos, barridas, intercepciones, etc.

Existen diferentes abordajes para la problemática de evaluar acciones pero se estudiaron dos grandes grupos: los **basados en posición** y los **basados en acciones**. El posicional asigna un valor a cada zona de un campo de juego, mientras que el basado en acciones considera la zona y muchas más características de la jugada que describen el evento.

En las siguientes secciones se presentan diferentes implementaciones/ideas de modelos de *Possession Value*.

3.2. Modelos Basados en Posición

3.2.1. Primer acercamiento a xT: Modelo de Markov de Sarah Rudd

En el año 2011, la primera en plantear el concepto de *xT* fue *Sarah Rudd*. Ella no lo bautizó con ese nombre, pero fue quien sentó las bases del modelo matemático detrás de esta idea. Era una época donde estaba naciendo el *Football Analytics*

y la presentación que ella realizó en *New England Symposium on Statistics in Sports* de Harvard (*"A Framework for Tactical Analysis and Individual Offensive Production Assessment in Soccer Using Markov Chains"*, Rudd, 2011) fue una verdadera revolución. Para dimensionar su impacto, en 2012 fue contratada por el Arsenal de Inglaterra y estuvo a cargo de su área de *Analytics* hasta 2021, cuando decidió abrir su propia consultora deportiva.

El planteo consistió en modelar un partido como una secuencia de estados $S_1, \ldots, S_n$ y a las acciones como las transiciones que llevan al partido de un estado a otro (usando *ball event data*). Cada estado está determinado por el equipo atacante que tiene la pelota y por la distribución en la cancha del equipo que defiende. Se definieron 39 estados:

- 2 estados terminales (Gol y Pérdida de Posesión)
- 7 de pelota detenida (penales, tiros libres, laterales y córners -variantes cortos y largos para estos últimos tres-)
- 30 restantes definidos por la zona de la cancha y el estado de la defensa

A partir de lo anterior, se construye una matriz de transiciones, donde se calcula la probabilidad de moverse de un estado S_a a un estado S_b para todas las combinaciones de estado posibles. Para los estados terminales, la probabilidad de moverse a otro estado es 0 y la de quedarse, 1.

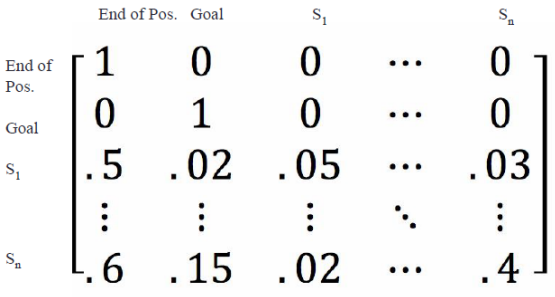


FIGURA 3.5: Matriz de transición de estados (Fuente: Rudd, 2011)

La idea se basa en los *modelos de cadenas de Markov*, donde se asume que el cambio entre estados es independiente de los estados anteriores y que depende únicamente del estado actual.

Lo novedoso que plantea *Rudd* es que las probabilidades de gol de cada estado pueden usarse para darle crédito o valor a los jugadores involucrados durante una posesión de ataque, no solamente a aquel que la finaliza o a quien da la asistencia. Además, propone valorar positivamente acciones que pueden no haber terminado en gol (por ejemplo, remates desviados o atajados por el arquero). En definitiva, define el valor de una transición $S_a \rightarrow S_b$ en base a la mejora en la probabilidad de gol entre esos dos estados de juego y lo hace mediante la siguiente ecuación:

$$Value_{a \rightarrow b} = P(Goal)_{S_b} - P(Goal)_{S_a}$$

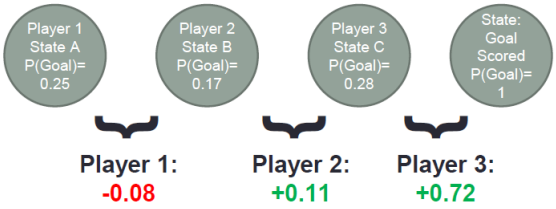


FIGURA 3.6: Cómo valorizar las transiciones de estado en base a su diferencia de probabilidad de gol (Fuente: Rudd, 2011)

Entre sus puntos débiles, aparece el supuesto de *estados sin memoria* de los modelos de Markov ya que parecería no ajustarse a la mayoría de las jugadas en el fútbol. Además, otra limitación de esta técnica es que no considera el tipo de acción para la transición entre estados. Es decir, podría asumir que un pase o una conducción desde cierta zona hasta otra tienen el mismo efecto cuando éste podría variar según el tipo de evento. De todas maneras, fue una propuesta original en ese entonces y sirvió como punto de partida para otras ideas más avanzadas.

3.2.2. xT Basado en Posición de Karun Singh

El primero en introducir el nombre de *Expected Threat* fue *Karun Singh*, quien en 2018 publicó en su blog (Singh, 2018) una versión que se basaba exclusivamente en la posición de la acción dentro del campo de juego. Dentro de la categorización que seguiremos en este trabajo, la versión propuesta por *Singh* se considera como *xT Basado en Posición*. Como dato de color, aparece nuevamente el nombre del Arsenal de Inglaterra ya que tiene a *Singh* como científico de datos desde 2022.

La idea planteada seguía los siguientes lineamientos:

- Valorizar cada acción individual de un jugador (pases, regates, remates, etc.)
- Operar sobre *ball event data* debido a restricciones de disponibilidad de información
- Considerar el valor de las acciones independientemente del resultado final de la posesión
- Premiar no sólo acciones que llevaban inmediatamente a una jugada de remate al arco si no también a aquellas que podían generar un buen disparo unas acciones más adelante

Su objetivo principal era asignar un *valor de peligrosidad* a cada zona de la cancha y definió las siguientes características para cada zona (x, y):

- **Probabilidad de Movimiento** $m_{x,y}$: cuando un jugador está en posesión en la zona (x, y), con qué frecuencia mueve la pelota (pase o conducción) como siguiente acción
- **Probabilidad de Remate** $s_{x,y}$: cuando un jugador está en posesión en la zona (x, y), con qué frecuencia dispara al arco como siguiente acción
- **Matriz de Transición de Movimientos** $T_{x,y}$: cuando un jugador está en posesión en la zona (x, y), todas las probabilidades de que la mueva al resto de las zonas como siguiente acción

- **Probabilidad de Gol** $g_{x,y}$: con qué frecuencia los remates desde la zona (x, y) son goles (esencialmente xG)

Entonces, asumiendo que un jugador cuando tiene la pelota cuenta con dos opciones (rematar o mover el balón) y que la cancha está dividida en $N \times M$ zonas, define a la peligrosidad de una zona (x, y) como:

$$xT_{x,y} = (s_{x,y} \times g_{x,y}) + (m_{x,y} \times \sum_{w=1}^{N} \sum_{z=1}^{M} T_{x,y \to w,z} \times xT_{w,z})$$

Es decir, la probabilidad de gol ponderada por la probabilidad de rematar sumada a la peligrosidad de cada zona, ponderada primero por la probabilidad de mover a esa zona y luego por la probabilidad de mover la pelota.

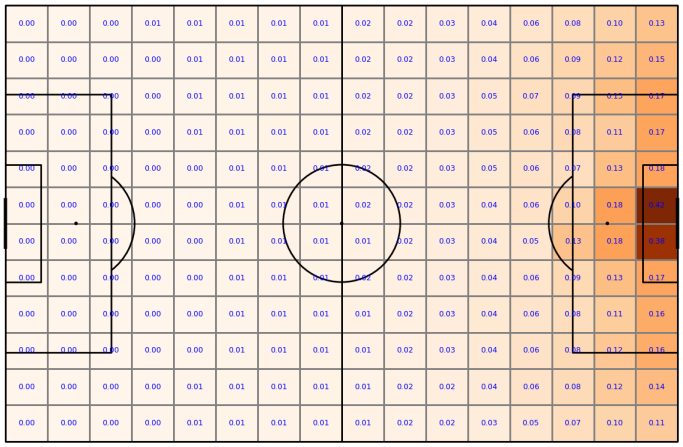


FIGURA 3.7: A cada zona de la cancha se le asigna un valor asociado a su peligrosidad (Fuente: Soccermatics)

Puede observarse que la definición de xT es cíclica, con lo cual se propone un algoritmo iterativo para converger a una solución. Se inicia con $xT_{x,y} = 0$ para todas las zonas y se evalúa la fórmula iterativamente hasta la convergencia. Específicamente, $xT_{x,y}$ en la iteración n puede interpretarse como la probabilidad de anotar un gol dentro de las siguientes n acciones desde la zona (x, y).

En esta propuesta, un estado de juego está representado solamente por la posición de la pelota y el valor de una acción quedará determinado por el destino que tenga ese balón. Análogamente al enfoque propuesto por *Rudd*, cada estado no tiene memoria (no importan los eventos previos) y el crédito asignado a una acción estará dado por la diferencia en términos de probabilidad de gol. En este caso, una acción que mueva la pelota desde la zona (x, y) a la (w, z) tendrá un valor de $xT_{w,z} - xT_{x,y}$.

Una de las críticas a este enfoque es que tener en cuenta sólo la posición parecería no capturar la dinámica completa del juego. Por ejemplo, un pase atrás podría reducir la probabilidad de gol en lo inmediato pero terminar abriendo un espacio que genere mayor peligro. Además, todas las zonas tienen el mismo tamaño y, especialmente en lugares cercanos al arco contrario, movimientos pequeños que no implican un cambio de cuadrícula podrían ser ignorados por este modelo.

3.3. Modelos Basados en Acciones

Con el objetivo de atacar los puntos débiles de las otras alternativas es que surgen los *Modelos Basados en Acciones*, los cuales tienen en cuenta las características de los eventos que son parte de la posesión. Es crucial introducir el concepto de **cadena de posesión** ya que todas estas implementaciones se apoyan en esta idea. Una *cadena de posesión* es una secuencia de acciones del mismo equipo que solamente se ve interrumpida por un gol, offside, falta, pelota fuera de la cancha o si el equipo contrario realiza dos acciones con pelota consecutivas.

A diferencia de la última alternativa analizada, aquí se tienen en cuenta otras variables más allá de las zonas donde se origina y termina el evento, pudiendo capturar mejor el contexto y el tipo de juego que se viene desarrollando.

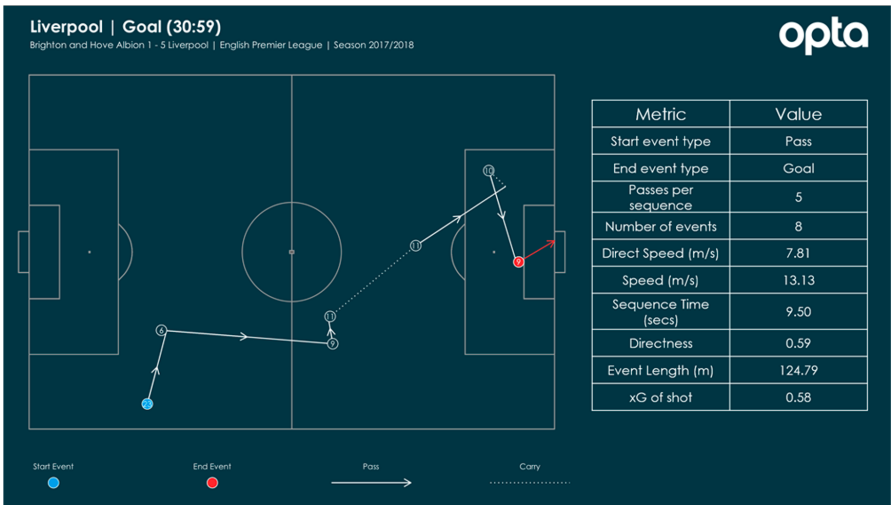


FIGURA 3.8: Un ejemplo de cadena de pases (Fuente: Soccermatics)

Es importante destacar que muchas de las variantes que se presentan a continuación son herramientas desarrolladas por empresas privadas, con lo cual solamente se conocen lineamientos generales pero no los detalles de implementación de cada una de ellas.

3.3.1. Possession Value Framework (Opta)

Uno de los proveedores de datos más importantes del mercado (*Stats Perform/Opta*) se interesó por este tipo de modelos y desarrolló uno propio en el año 2019 (StatsPerform, 2019).

Utilizando *ball event data* y basándose en la idea de premiar/castigar a los jugadores según aumenten/disminuyan la probabilidad de gol de su equipo, desarrollaron su propia versión centrada en las cadenas de posesión, denominándolo *Possession Value (PV)*.

Inicialmente, su modelo basaba su estimación usando hasta cinco eventos previos dentro de la misma posesión y los comparaba con datos históricos para intentar determinar la probabilidad de que se marque un gol en esa posesión. Además, no sólo se concentraba en otorgar crédito a las jugadas positivas si no que *castigaba* a aquellas donde se perdía la pelota y las ponderaba según la probabilidad de gol del rival (limitando su *castigo* a 0.025, el valor promedio de posesión).

En implementaciones posteriores (StatsPerform, 2020), en pos de ganar en interpretabilidad, comenzó a utilizar un enfoque basado en tiempo (en lugar del fin de la

posesión): la probabilidad de gol de una acción estará dada por la probabilidad de que el equipo anote en los próximos *10 segundos*. Adicionalmente, removieron la penalidad arbitraria por perder la pelota, permitiendo premiar a jugadores que ceden posesión en zonas de peligro. El espíritu detrás de este último cambio es que una pérdida de balón en una zona peligrosa para el rival puede, en el corto plazo, volver a generar peligro para el equipo atacante (por ejemplo, un centro rechazado por un defensor puede dejar la pelota en una zona comprometedora para la defensa).

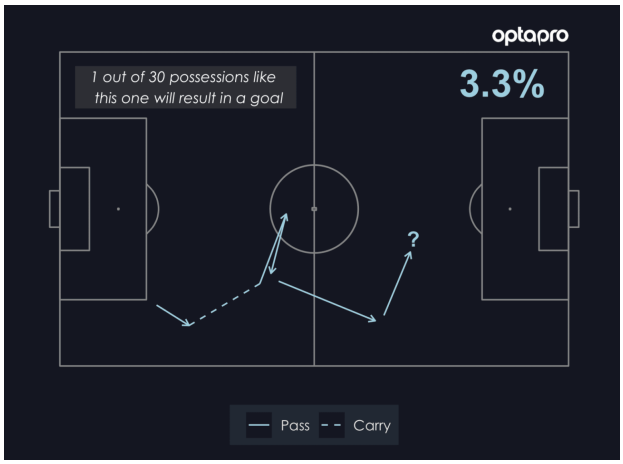


FIGURA 3.9: Un ejemplo de posesión y su PV asociado (Fuente: Stats Perform)

3.3.2. VAEP (Universidad de KU Leuven)

En el año 2019, un grupo de investigadores del *DTAI Sports Analytics Lab* de la universidad belga de *KU Leuven* presentó su propio modelo para valorizar las acciones de los jugadores bajo el nombre de **VAEP** (Decroos et al., 2019). Al ser un desarrollo académico, sus publicaciones brindan un alto nivel de detalle y hasta incluso han compartido con la comunidad parte del código que implementa su modelo.

Su postulado se basa en que todas las acciones de un partido persiguen dos objetivos básicos:

1. Incrementar las chances de anotar un gol
2. Reducir las chances de recibir un gol

Plantean un lenguaje estándar para representar las acciones de un partido (*SPADL*) y así poder trabajar de manera uniforme con diferentes proveedores de datos. Cada evento es representado por:

- **Tiempo de Inicio**
- **Tiempo de Fin**
- **Coordenadas (x,y) de Inicio**
- **Coordenadas (x,y) de Fin**
- **Jugador**
- **Equipo**
- **Tipo de Acción**: 21 acciones posibles (remate, pase, conducción, etc.)
- **Parte del Cuerpo**: 4 valores posibles (pie, cabeza, otro o ninguna)
- **Resultado de la Acción**: 6 valores posibles (éxito o falla las más comunes)

Dados los estados de juego S_i y el equipo x, definen a $P_{scores}(S_i, x)$ y a $P_{concedes}(S_i, x)$ como la probabilidad de que el equipo x anote o conceda un gol en el futuro cercano. Al igual que en otros enfoques, el crédito de una jugada estará dado por el cambio de probabilidad (en este caso de las dos probabilidades utilizadas). Suponiendo que la acción a_i es aquella que genera el cambio de estado de S_{i-1} y S_i, se definen:

$$\Delta P_{scores}(a_i, x) = P_{scores}(S_i, x) - P_{scores}(S_{i-1}, x)$$

$$\Delta P_{concedes}(a_i, x) = P_{concedes}(S_i, x) - P_{concedes}(S_{i-1}, x)$$

La primera ecuación hace referencia al *valor ofensivo* de la jugada mientras que la negación de la segunda al *valor defensivo* (se usa en negativo debido a que se busca reducir la probabilidad de conceder un gol).

Entonces, se llega a la definición del *VAEP value* como:

$$V(a_i, x) = \Delta P_{scores}(a_i, x) + (-\Delta P_{concedes}(a_i, x))$$

De esta manera, el cálculo de las probabilidades de anotar y de conceder se consideran problemas de clasificación binaria, los cuales pueden ser resueltos con clasificadores probabilísticos permitiendo aplicar múltiples algoritmos de *Machine Learning* (en particular, los autores utilizan *XGBoost* y *CatBoost*).

Su implementación cuenta con dos parámetros que son definidos por el usuario:

1. Cuántas acciones de historia se tienen en cuenta (n)
2. Cuántas acciones se miran hacia adelante para considerar la probabilidad de gol (k)

Los valores que sugieren para estas variables son $n = 3$ y $k = 10$.

A la hora de entrenar los modelos, se utiliza la información de las acciones según su lenguaje *SPADL*, características derivadas como el ángulo y la distancia al arco y variables de contexto como los goles anotados por cada equipo y su diferencia.

Si bien el método permite asignar un valor a cada acción, en su trabajo proponen una manera de agregar información y proveer un *rating* por jugador, acotado a cierta ventana de tiempo que podría ser un partido o un torneo, según lo que se busque analizar. Debido a que a mayor tiempo en cancha más acciones podrá realizar un jugador, la propuesta es hacer los cálculos cada 90 minutos de juego. Dada la ventana temporal T y el jugador p, se calcula el *rating* como:

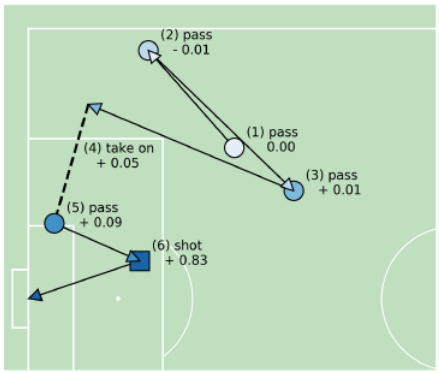
$$rating(p) = \frac{90}{m} \sum_{a_i \in A_p^T} V(a_i)$$

Donde m es la cantidad de minutos jugados por el jugador p, A_p^T son las acciones del jugador p en la ventana temporal T, y $V(a_i)$ es el VAEP de la acción a_i.

Para evaluar el entrenamiento de sus modelos utilizan tres métricas: Brier Score, Logarithmic Loss y Curva ROC.

Una de las críticas que se ha formulado sobre este enfoque está relacionada a su *interpretabilidad*. La complejidad de la representación de estados, la combinación de modelos (probabilidad de anotar y conceder) sumado a que busca abarcar a todas las acciones de juego (defensivas y ofensivas) pueden volver complejo explicar el por qué de ciertas asignaciones de valor. Además, al buscar estandarizar la representación entre diferentes proveedores de datos, deben trabajar exclusivamente

con información presente en todos y pueden perderse detalles capturados por los proveedores más avanzados.



	TIME	PLAYER	ACTION	P_{scores}	VALUE
1	92m4s	S. Busquets	pass	0.03	0.00
2	92m6s	L. Messi	pass	0.02	- 0.01
3	92m8s	S. Busquets	pass	0.03	+ 0.01
4	92m11s	L. Messi	take on	0.08	+ 0.05
5	92m12s	L. Messi	pass	0.17	+ 0.09
6	92m14s	A. Vidal	shot	1.00	+ 0.83

FIGURA 3.10: Una jugada del Barcelona con sus respectivos valores VAEP (Fuente: Decroos et al., 2019)

3.3.3. On-Ball Value (StatsBomb)

StatsBomb, otro de los proveedores de datos más importantes de la industria, no podía quedarse atrás de estos desarrollos y en 2021 presentó su propio modelo de *xT Basado en Acciones* bajo el nombre de **On-Ball Value** (StatsBomb, 2022c)

A diferencia de los modelos anteriores, esta variante no se entrena utilizando los goles efectivos si no el xG. Su implementación evalúa todas las acciones que ocurren en el campo y además considera tanto los goles a favor y goles en contra para medir con precisión el riesgo/recompensa de cada acción.

En cuanto a las características que utiliza el modelo, han optado por incluir:

- **Ubicación**: Coordenadas X e Y, distancia y ángulo hacia el arco
- **Contexto de la acción**: pelota parada, acción de juego abierto, etc.
- **Presión del rival**: Si el evento se realizó bajo presión de un jugador del equipo contrario
- **Parte del cuerpo utilizada**: Cabeza, pie, etc.

Intencionalmente, han decidido excluir características relacionadas al historial de posesión. Si bien esta información podría ser útil para inferir la probable estructura defensiva del rival, conocer disponibilidad y ubicación de compañeros de equipo o del oponente, en la práctica muchas se correlacionan fuertemente con otros factores como el estilo de juego y el poderío del equipo. Por ejemplo, otros modelos sobrevaloran los pases realizados en cadenas de posesión más largas debido a que los equipos más fuertes suelen tener secuencias más extensas que los equipos más débiles.

Adicionalmente, mencionan que decidieron no otorgar crédito por las recepciones de balón, asumiendo que quien recibe luego será evaluado por la siguiente acción que realice.

Como se destacó previamente, al ser una versión comercial se desconocen detalles finos sobre su implementación (por ejemplo, si usa o no datos de tracking).

3.3.4. Otras Implementaciones

Las descriptas anteriormente son algunas de las variantes que se utilizan en la actualidad, pero no son las únicas. El mismo *David Sumpter* desarrolló junto a su

empresa *Twelve* su propia versión utilizando datos de tracking; sin embargo, no hay información pública con detalles al respecto. Otra alternativa es la de *goals added* desarrollada por el grupo *American Soccer Analysis* (Kullowatz, 2020).

3.4. Resumen

Características / Modelos	**Cadenas de Markov (Rudd)**	**xT (Singh)**	**PV (Stats Perform)**	**VAEP**	**On-Ball Value**
Tipo Modelo	xT Basado en Posición	xT Basado en Posición	xT/xD Basado en Acciones	xT/xD Basado en Acciones	xT/xD Basado en Acciones
Tipo de Acciones Consideradas	Ofensivas	Ofensivas	Ofensivas + Defensivas	Ofensivas + Defensivas	Ofensivas + Defensivas
Tipo de Datos	Ball Event Data	Ball Event Data	Ball Event Data	Ball Event Data	No informado
Representación de Estados de Juego	Zona determinada por posición (x,y) origen + estado de la defensa	Zona determinada por posición (x,y) origen	(x,y) + 5 jugadas anteriores de la cadena de posesión	Últimas 3 acciones de juego y además: (x,y) + distancia y ángulo al arco; tipo de acciones previas y parte del cuerpo	Sin historial de posesión, solamente: (x,y) + distancia y ángulo al arco; contexto de la acción (pelota parada, juego abierto, etc); presión del rival y parte del cuerpo
Entrenado sobre	Goles reales	Goles reales	Gol en los próximos 10 segundos	Gol en las próximas 10 acciones	xG
Comentarios	No se brindan detalles sobre las características del "estado de la defensa"	Divide a la cancha en $N \times M$ zonas	No se brindan detalles de las características usadas en el entrenamiento.	-	No se brindan detalles acerca de cómo se usa el xG para entrenar el modelo

Capítulo 4

Materiales

En el presente capítulo se detallan los materiales (conjuntos de datos) que sirvieron de insumo para el entrenamiento de los modelos y los análisis que se llevaron a cabo en este trabajo.

4.1. StatsBomb Open Data

El proveedor de datos *StatsBomb*, fiel a su compromiso de fomentar el crecimiento de la comunidad de *Football Analytics*, ha decidido disponibilizar gratuitamente datos de ciertas ligas para uso público en proyectos de investigación y para aquellos con un interés genuino en el análisis del fútbol.

La información es proporcionada en formato abierto *JSON*, exportados desde la API de datos de *StatsBomb*.

Las competiciones más importantes que han publicado a la fecha son las siguientes:

- Ligas **Big Five** (Inglaterra, España, Italia, Alemania y Francia) – Temporada Completa 2015/16
- Eurocopa 2020
- Copa África 2023
- Mundiales 2018 y 2022

Además, han puesto a disposición todos los partidos de *Lionel Messi* desde sus inicios en el Barcelona hasta su reciente incursión en la *MLS* (StatsBomb, 2021b).

Por cada uno de los partidos de esas competiciones, se cuenta con:

- Eventos con pelota
- Eventos tácticos (alineación inicial de cada equipo y cambios de formación durante el partido)
- Relación entre los eventos (brinda la posibilidad de encadenarlos entre sí)
- Ubicación de los jugadores (cada uno con su nombre) para cada evento de *remate*

En solamente dos torneos (Eurocopa 2020 y Mundial 2022) se dispone de *eventos 360*.

4.2. Competiciones Incluidas

Los torneos con los que se trabajó son los descriptos en la siguiente tabla:

Torneo	Temporada	# Equipos	# Partidos
Bundesliga	2015/2016	18	306
La Liga	2015/2016	20	380
Ligue 1	2015/2016	20	377
Premier League	2015/2016	20	380
Serie A	2015/2016	20	380
Eurocopa	2020	24	51
Mundial	2018	32	64
Mundial	2022	32	64
Copa de África	2023	24	52
Temporadas Incompletas (Lionel Messi)	Múltiples	71	551
Total			2605

4.2.1. Bundesliga 2015/2016

FIGURA 4.1: Fuente: FC Bayern Munich

Otra temporada exitosa para el Bayern Munich dirigido por *Pep Guardiola* en ese entonces, ganando su tercera liga consecutiva bajo su mando. Fue seguido de cerca por el Borussia Dortmund, quien había sufrido la partida de *Jürgen Klopp* pero estaba bajo el mando de *Thomas Tuchel* (StatsBomb, 2023b).

4.2.2. La Liga 2015/2016

FIGURA 4.2: Fuente: El País

Barcelona logró conquistar el torneo luego de una dramática definición contra Real Madrid y Atlético Madrid. Era la época de oro del tridente MSN (*Messi*, *Suárez* y *Neymar*) con *Luis Enrique* comandando desde el banco de suplentes (StatsBomb, 2023c).

4.2.3. Ligue 1 2015/2016

FIGURA 4.3: Fuente: Depor

PSG ganó cómodamente la liga (38 puntos arriba del 2do, Lyon). Comandado por *Laurent Blanc* y liderado dentro del campo por *Ibrahimovic*, *Cavani* y *Di María*. Incluye el debut de *Mbappé* en Mónaco a los 16 años de edad (StatsBomb, 2023d). Como particularidad del conjunto de datos, *StatsBomb* aclara no fue posible recolectar datos de 3 partidos (Bastia vs AC Gazélec Ajaccio, 22/11/2015; Saint-Etienne vs Paris Saint-Germain, 31/01/2016; Troyes vs Bordeaux, 30/04/2016).

4.2.4. Premier League 2015/2016

FIGURA 4.4: Fuente: Depor

Inédito campeonato del Leicester City de *Claudio Ranieri*, con jugadores como *Jamie Vardy*, *Riyad Mahrez* y *N'Golo Kanté* a un nivel superlativo. Además, fue la primera temporada del alemán *Jürgen Klopp* al mando del Liverpool (StatsBomb, 2023e).

4.2.5. Serie A 2015/2016

FIGURA 4.5: Fuente: Soccerzz

Fue la temporada en la que Juventus conquistó su quinto título consecutivo en Italia. Respaldado por la muralla defensiva de *Bonnucci*, *Chielini* y *Barzagli* más *Buffon*, el equilibrio de un fantástico *Pogba* y el talento de *Paulo Dybala*. Tuvo como competidor al Napoli de *Maurizio Sarri*, impulsado por los goles del *Pipita Higuaín*, máximo artillero de la temporada con 36 goles (StatsBomb, 2023f).

4.2.6. Eurocopa 2020

FIGURA 4.6: Fuente: Olé

Originalmente se iba a disputar en 2020, pero el torneo se pospuso hasta 2021 debido a la pandemia de COVID-19. La Italia de *Roberto Mancini* se impuso a Inglaterra en la final por penales en el estadio de Wembley. La competición Incluye *eventos 360* (StatsBomb, 2021a).

4.2.7. Copa África 2023

FIGURA 4.7: Fuente: Clarín

Competencia desarrollada en Costa de Marfil, que tuvo como campeón al equipo local. Los marfileños llegaron al certamen como uno de los candidatos y, si bien sufrieron para clasificarse (despidieron a su entrenador en medio del torneo), lograron imponerse en las definiciones mano a mano y terminaron venciendo a Nigeria en la final (StatsBomb, 2023a).

4.2.8. Mundial 2018

FIGURA 4.8: Fuente: La Nación

La Copa del Mundo 2018 se disputó en territorio ruso y tuvo a Francia como campeón, luego de vencer a la revelación, Croacia. Argentina tuvo un papel para el olvido, clasificando sobre la hora a segunda ronda y quedando fuera en 8vos de final en manos de Francia (StatsBomb, 2018).

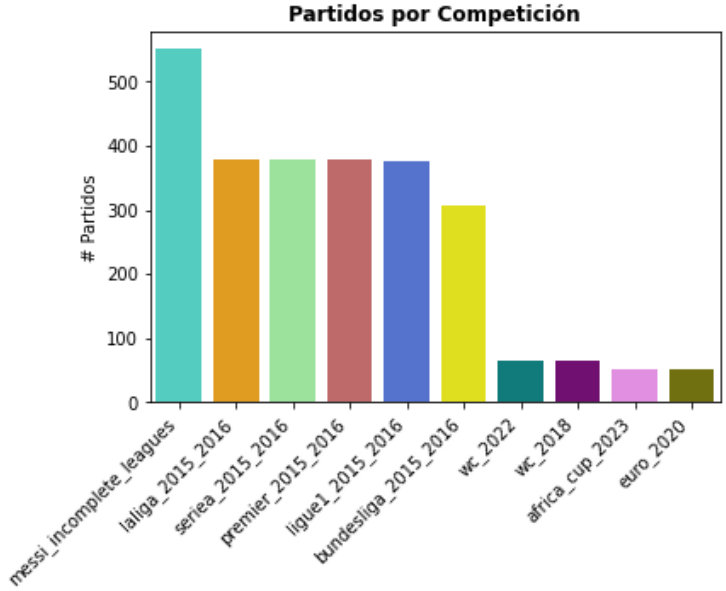
4.2.9. Mundial 2022

FIGURA 4.9: Fuente: Infobae

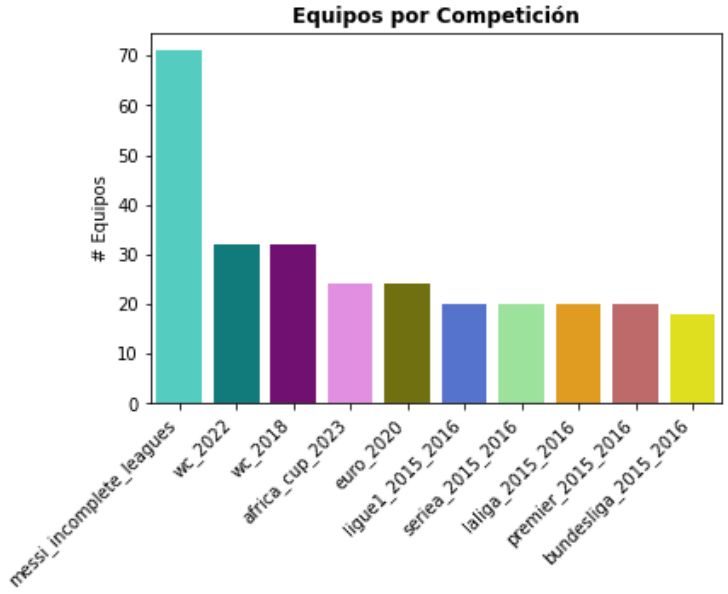
El torneo que volvió a consagrar a Argentina como campeona después de su conquista en México 86', cortando una sequía de 36 años. En tierras qataríes *Lionel Messi* pudo romper el maleficio y logró conseguir el último gran título que le faltaba a su vitrina, disipando ya de manera completa los cuestionamientos alrededor de su desempeño con la selección nacional. Comandados por el debutante *Lionel Scaloni* y apoyados en las brillantes actuaciones de *Dibu Martínez*, *Julián Álvarez* y *Lionel Messi*, lograron imponerse en una infartante final con Francia. La competición incluye *eventos 360* (StatsBomb, 2022b).

4.3. Análisis Exploratorio

Como se describió previamente, el conjunto de datos consiste de **2.605** partidos, distribuidos de la siguiente manera:



A su vez, las competiciones contienen distinta cantidad de equipos como se muestra a continuación:

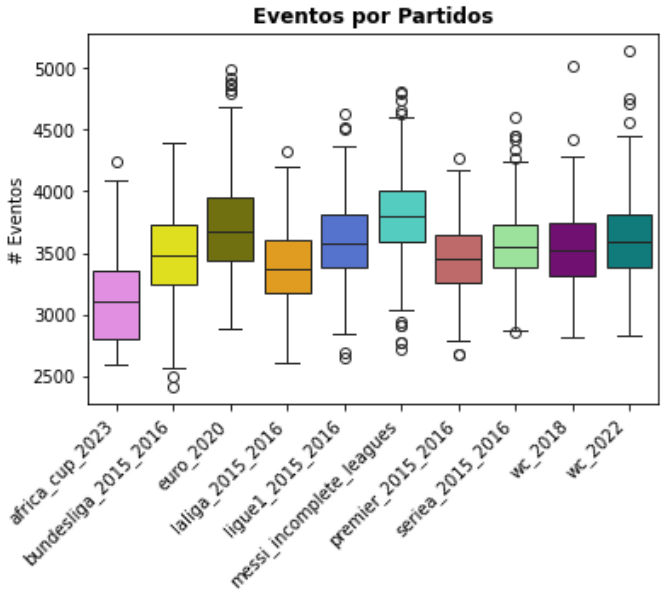


Los datos provistos por *StatsBomb* se dividen en cinco tipos (StatsBomb, 2022a):

- Eventos
- Eventos Tácticos
- Relaciones de Eventos
- Fotos de Disparos (*Freezes*)
- Eventos 360

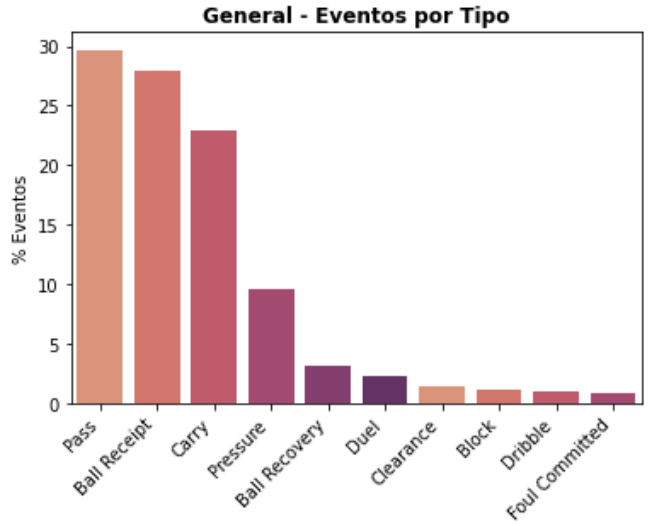
4.3.1. Eventos

El conjunto de datos posee un total de **9.300.640** eventos, que se agrupan en **35** tipos distintos, mayormente acciones con balón. Por partido, se dispone de alrededor de **3.500** eventos como lo muestra el siguiente boxplot:

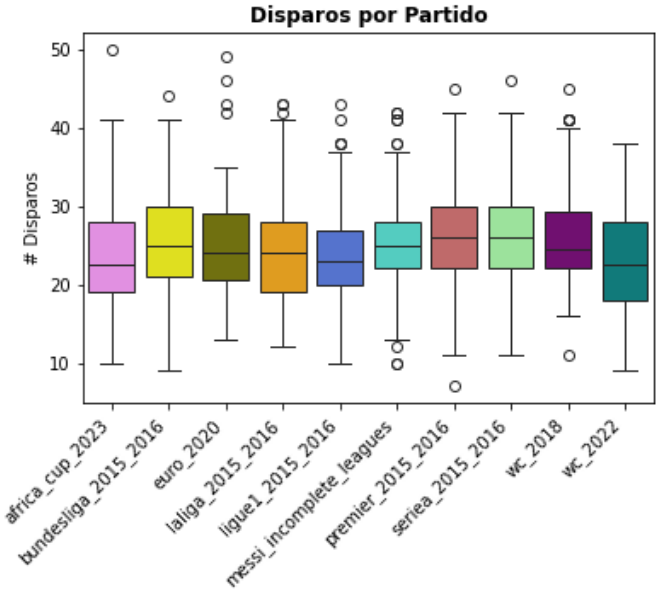


Puede observarse que la cantidad de acciones parece ser bastante pareja por competición, siendo la última Copa de África la que menos eventos por partido registra.

En términos generales, los pases, las recepciones de pelota y las conducciones (*Carry*) se llevan casi el **75 %** de las acciones. Si se analiza por competición, puede notarse idéntico comportamiento.



Por su parte, una de las acciones más importantes (y no tan frecuente) como son los disparos, también presentan un comportamiento similar por competición con valores entre 20 y 30 tiros por partido.



StatsBomb ofrece un muy alto nivel de detalle, disponiendo de **90** campos por evento, aunque muchos de ellos solamente están presentes para determinado tipo de acción. A continuación se presenta un diccionario de datos con la descripción de cada uno de ellos.

#	Campo	Tipo	Descripción
		Generales	
1	id	object	ID único
2	index	int64	Índice intrapartido
3	period	int64	Período (1=PT, 2=ST, 3=Suplementario 1, 4=Suplementario 2, 5=Penales)
4	timestamp	object	Tiempo del evento, con milisegundos incluidos (Se registra el timestamp por período)
5	minute	int64	Minutos (Resetea en 45 en PT, 90 y 120 en suplementarios)
6	second	int64	Segundos
7	possession	int64	Identificador de posesión (única) que tiene un equipo durante el partido

8	possession_team_id	int64	Id del equipo de la posesión
9	possession_team_name	object	Nombre del equipo de la posesión
10	duration	float64	Duración en segundos del evento
11	match_id	int64	Id de partido
12	team_id	int64	Id del equipo
13	team_name	object	Nombre del equipo
14	player_id	int64	Id del jugador
15	player_name	object	Nombre del jugador
16	position_id	int64	Id de la posición
17	position_name	object	Nombre de la posición del jugador involucrado
18	off_camera	boolean	True si el evento ocurrió cuando la cámara no estaba enfocando (evento con alta probabilidad de error)
19	type_id	int64	Id de tipo de evento (Son 35 tipos de evento distintos)
20	type_name	object	Nombre de tipo de evento
21	play_pattern_id	int64	Id de tipo de jugada
22	play_pattern_name	object	Nombre de tipo de jugada (De córner, de saque de arco, normal, etc)
23	sub_type_id	int64	Id de subtipo de evento (El 90 % no tiene subtipo)
24	sub_type_name	object	Nombre de subtipo de evento
25	under_pressure	boolean	Indica si el evento con pelota fue realizado bajo presión
26	counterpress	boolean	Indica si fue una acción de presión dentro de los 5 segundos posteriores a una pérdida de balón en juego abierto.
27	x	float64	Posición x, de 0 a 120
28	y	float64	Posición y, de 0 a 80
29	end_x	float64	Posición x, de 0 a 120 (solamente en Pass, Carry, Goal Keeper y Shot)
30	end_y	float64	Posición y, de 0 a 80 (solamente en Pass, Carry, Goal Keeper y Shot)
31	end_z	float64	Posición z, de 0 a 8 (solamente para algunos tiros)
32	body_part_id	int64	Id de parte del cuerpo (disponible para Clearance, Pass, Shot & Goal Keeper)
33	body_part_name	object	Parte del cuerpo
34	out	boolean	Indica si la pelota salió fuera del campo
35	outcome_id	int64	Id de resultado del evento
36	outcome_name	object	Resultado del evento (varía según el tipo de evento)
37	technique_id	int64	Id de técnica utilizada para el evento (disponible principalmente para Shots)
38	technique_name	object	Nombre de técnica utilizada para el evento
39	aerial_won	object	Indica si el evento fue realizado ganando un duelo aéreo (disponible solamente para Passes, Shots, Clearances y Miscontrols)
		Pase	
40	pass_recipient_id	int64	Id del receptor del pase
41	pass_recipient_name	object	Nombre del receptor del pase (puede estar nulo en pases incompletos)
42	pass_length	float64	Longitud en yardas (De 0 a 121)
43	pass_height_id	int64	Id de altura de pase
44	pass_height_name	object	Altura de pase. Ground (a ras del piso), Low (con pico debajo de la altura de los hombros), High (con pico encima de la altura de los hombros)
45	pass_assisted_shot_id	object	Id del evento de tiro asociado al pase. Permite encadenar jugadas
46	pass_shot_assist	boolean	Indica si el pase fue previo a un disparo (no gol)
47	pass_goal_assist	boolean	Indica si el pase fue previo a un gol
48	pass_switch	boolean	Indica si fue un cambio de frente (pase de más de 40 yardas de ancho)
49	pass_cross	boolean	Indica si fue un centro
50	pass_cut_back	boolean	Indica si fue un pase atrás (ofensivo, desde un costado con destino cercano al punto de penal)
51	pass_deflected	boolean	Indica si fue desviado el pase (puede ocurrir en pases completos o incompletos)
52	pass_no_touch	boolean	Indica si fue un pase sin tocar la pelota
53	pass_miscommunication	boolean	Indica si hubo falta de entendimiento entre los jugadores (el pase fue correcto pero el receptor malinterpretó el pase o el receptor se movió y el ejecutante lo mandó a la posición previa)
54	pass_backheel	boolean	Indica si fue pase de taco
55	pass_angle	float64	En radianes (0 es recto, de -PI (antihorario) a PI (horario))
		Disparo	
56	shot_statsbomb_xg	float64	Statsbomb Expected Goal (de 0 a 1)
57	shot_first_time	boolean	Indica si el disparo fue de primera
58	shot_one_on_one	boolean	Indica si el disparo fue en un mano a mano
59	shot_open_goal	boolean	Indica si el disparo fue con el arco libre
60	shot_redirect	boolean	Indica si el disparo fue un desvío
61	shot_deflected	boolean	Indica si el disparo fue desviado
62	shot_follows_dribble	boolean	Indica si el disparo siguió un regateo
63	shot_key_pass_id	object	Id del evento de la asistencia
		Regates	
64	dribble_no_touch	boolean	Indica si fue un regateo sin tocar la pelota
65	dribble_overrun	boolean	Indica si el regateo terminó con posesión de otro jugador
66	dribble_nutmeg	boolean	Indica si el regateo fue "de caño"
		Falta Cometida	
67	foul_committed_penalty	boolean	Indica si la falta derivó en un penal
68	foul_committed_card_id	int64	Indica si la falta derivó en una tarjeta y en cuál
69	foul_committed_card_name	object	Indica si la falta derivó en una tarjeta y en cuál
70	foul_committed_offensive	boolean	Indica si la falta se cometió teniendo posesión del balón
71	foul_committed_advantage	boolean	Indica si el árbitro dio ventaja
		Pelota Recuperada	
72	ball_recovery_recovery_failure	boolean	Indica si se pierde la pelota intentando recuperarla
73	ball_recovery_offensive	boolean	Indica si la pelota fue recuperada en fase ofensiva
		Falta Ganada	
74	foul_won_advantage	boolean	Indica si la falta no se terminó cobrando por dar ventaja
75	foul_won_defensive	boolean	Indica si la falta se ganó sin posesión del balón
76	foul_won_penalty	boolean	Indica si se consiguió un penal
		Sustitución	
77	substitution_replacement_id	int64	Id del jugador reemplazado
78	substitution_replacement_name	object	Nombre del jugador reemplazado
		Bloqueo	
79	block_deflection	boolean	Indica si el bloqueo fue un desvío que no terminó de alterar la trayectoria original
80	block_save_block	boolean	Indica si bloqueó un disparo

81	block_offensive	boolean	Indica si fue un bloqueo ofensivo
		Mal Comportamiento	
82	bad_behaviour_card_id	int64	Especifica el id de la tarjeta correspondiente
83	bad_behaviour_card_name	object	Especifica la tarjeta correspondiente
		Arquero	
84	goalkeeper_position_id	int64	Indica cómo estába el arquero antes de recibir el disparo
85	goalkeeper_position_name	object	Indica cómo estába el arquero antes de recibir el disparo
		Otros	
86	player_off_permanent	boolean	Indica si el jugador no participa más del juego (lesionado y sin cambios disponibles)
87	injury_stoppage_in_chain	boolean	Indica si la pelota estaba en posesión del jugador lesionado antes del parate por lesión
88	half_start_late_video_start	boolean	Indica si el feed de video comenzó tarde (y se perdió parte del partido)
89	half_end_early_video_end	boolean	Indica si el feed de video finalizó antes de completar el tiempo
90	tactics_formation	object	Especifica la formación usada cuando comienza el partido o cuando hay un cambio durante el mismo

4.3.2. Eventos Tácticos

En cada partido, se registra la alineación inicial de cada uno de los equipos y la misma queda almacenada como un *evento táctico*. Además, cuando *StatsBomb* detecta un cambio de táctica durante el partido, también puede generarse un nuevo evento táctico mostrando la nueva disposición de los jugadores.

Cada evento táctico tiene asociados hasta 11 registros con los siguientes campos:

#	Campo	Tipo	Descripción
1	jersey_number	object	Número de camiseta del jugador
2	match_id	int64	Id del partido
3	id	object	Id del evento táctico
4	player_id	int64	Id del jugador
5	position_id	int64	Id de la posición
6	position_name	object	Nombre de la posición
7	event_tactics_id	int64	Id interno del evento táctico (va de 1 a 11)

4.3.3. Relaciones de Eventos

El conjunto de datos brinda la posibilidad de encadenar eventos utilizando las *relaciones de eventos*. Son representados de manera simétrica, es decir que si existe la relación $A \rightarrow B$, también aparecerá la $B \rightarrow A$. Además, un mismo evento podría estar relacionado a N acciones distintas.

#	Campo	Tipo	Descripción
1	match_id	int64	Id de partido
2	id	object	Id del evento (un mismo evento puede tener múltiples relaciones)
3	index	int64	Índice del evento dentro del partido
4	type_name	object	Tipo de evento
5	id_related	object	Id del evento relacionado
6	index_related	int64	Índice del evento relacionado dentro del partido
7	type_name_related	object	Tipo de evento relacionado

4.3.4. Fotos de Remates (Freezes)

Por cada disparo existente, se cuenta con una *foto* o *freeze* que muestra la posición y nombre de cada uno de los jugadores que estaban cercanos a la jugada. En líneas generales, cada remate contiene entre 12 y 13 jugadores.

#	Campo	Tipo	Descripción
1	teammate	boolean	Indica si es compañero de equipo de quien ejecuta el disparo
2	match_id	int64	Id de partido
3	id	object	Id del evento de disparo
4	x	float64	Ubicación del jugador
5	y	float64	Ubicación del jugador
6	player_id	int64	Id del jugador
7	player_name	object	Nombre del jugador
8	position_id	int64	Id de la posición
9	position_name	object	Nombre de la posición
10	event_freeze_id	int64	Id interno de cada freeze

4.3.5. Eventos 360

Los *eventos 360* son los últimos que *StatsBomb* ha incorporado, los más costosos de obtener y los que terminan de brindar una perspectiva completa de lo que está sucediendo en el campo de juego. Lamentablemente, se han disponibilizado muy pocos partidos con este tipo de eventos (sólo dos competiciones).

De manera análoga a las *fotos de remate*, cada evento 360 tendrá una *foto del evento*. Sin embargo, sólo estará identificado el protagonista de la acción y del resto únicamente se conocerá el equipo o si es el arquero. En cada *evento 360* hay alrededor de 16 jugadores.

#	Campo	Tipo	Descripción
1	teammate	boolean	Indica si es compañero de equipo de quien es protagonista
2	actor	boolean	Indica si es el protagonista del evento
3	keeper	boolean	Indica si es el arquero
4	match_id	int64	Id de partido
5	id	object	Id del evento
6	x	float64	Ubicación del jugador
7	y	float64	Ubicación del jugador

Vale la pena destacar que no todos los eventos del partido tendrán su correspondiente *evento 360* y una acción podría no tener ningún protagonista asignado (*actor=TRUE*). Además, no todos los jugadores de campo estarán visibles si no que aparecerán los que son captados por la cámara de televisión (esta información se extrae con *Computer Vision* a partir de la transmisión deportiva). Por este motivo, cada evento 360 tendrá un *shape* que describirá el área cubierta por la cámara.

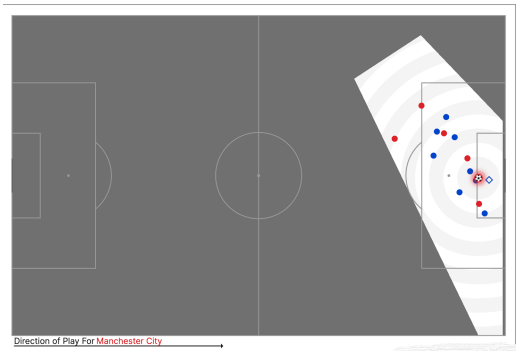


FIGURA 4.10: Fuente: Statsbomb

#	Campo	Tipo	Descripción
1	match_id	int64	Id del partido
2	id	object	Id del evento
3	visible_area	object	Shape de la toma de la cámara

Capítulo 5

Peligro de Gol, un nuevo modelo de Possession Value

5.1. Introducción

Luego de haber estudiado en profundidad las distintas propuestas de modelos de *Possession Value* existentes en la actualidad, a lo largo de este capítulo se presenta una implementación propia buscando tomar las fortalezas de las alternativas analizadas y poniendo un especial foco en la **interpretabilidad** y **consistencia** de los resultados.

Se exploran diferentes variantes a la hora del entrenamiento y se comparan resultados contra el modelo *VAEP*, ya que es el único que ofrece una implementación abierta a la comunidad.

5.2. Características

5.2.1. Tipo de Información Utilizada

Al igual que la mayoría de las alternativas analizadas, nos basamos en *ball event data* por cuestiones de disponibilidad de información. En la actualidad, solamente es posible conseguir un puñado de partidos con información de *tracking* de manera gratuita. Si bien los datos de *tracking* brindan información importante sobre el contexto de la jugada, se busca compensar esta situación usando acciones previas. De las opciones estudiadas, la única que no usa el historial de posesión es *Statsbomb*, aunque es muy probable que aproveche los datos de *tracking* que ellos mismos generan.

Por otro lado, se ha decidido trabajar directamente sobre la representación de *Statsbomb*, pudiendo aprovechar cada una de las características que provee esta compañía para cada uno de los eventos.

5.2.2. Tipo de Modelo

La propuesta se enmarca dentro de los *modelos basados en acciones*. El supuesto de *estados sin memoria* en el que se sustenta la idea de los *modelos basados en posición* no es suficiente para capturar la dinámica de un partido de fútbol. Los algoritmos de *Machine Learning* más populares en la actualidad y que se usan para entrenar el *modelo basado en acciones* permiten trabajar con un conjunto de *features* muy rico y encontrar patrones de gran complejidad.

5.2.3. Identificación de Posesiones

Se utiliza el criterio de posesiones definido por *Statsbomb*, según el cual una posesión denota un período en el que la pelota está en juego y un solo equipo tiene el control de la misma. Una nueva posesión se inicia después de que un equipo demuestra que ha tomado el control del balón. Cada acción posee un campo (*possession*) que representa un número de posesión único intrapartido.

5.2.4. Interpretabilidad y Consistencia

Dos conceptos fundamentales que deben atravesar a cualquier modelo aplicado al fútbol son el de la **interpretabilidad** y la **consistencia**. Si sus resultados no pueden ser comunicados de manera simple a los diferentes actores presentes en el deporte y si no tienen una conexión con conceptos futbolísticos básicos, su adopción nunca llegará a buen puerto. Más allá del análisis cuantitativo, se exploran distintos tipos de acciones de manera visual para corroborar la razonabilidad del modelo.

5.2.5. Foco en Acciones de Ataque

En este trabajo nos enfocamos en un modelo de xT, asignando valor únicamente a acciones ofensivas. Este es el motivo por el cual el nombre elegido es **Peligro de Gol**, ya que las acciones de ataque son el centro y se mide el valor de una acción en términos de cuánto *peligro de gol* termina generando.

5.2.6. Target de Entrenamiento

Como la mayoría de los modelos, se utilizan los goles reales como variable *target* de entrenamiento y se ha optado por emplear una ventana temporal en términos de segundos y no de cantidad de acciones. De todas maneras, ambos enfoques son equivalentes ya que es posible pasar de uno a otro considerando que el promedio de duración de cada acción es de aproximadamente 1,5 segundos.

5.3. Modelado

5.3.1. Definiciones Generales

- Cada partido entre los equipos E_1 y E_2 cuenta con una secuencia de N acciones $[a_1, \ldots, a_i, \ldots, a_N]$.
- Las acciones pueden agruparse en n posesiones $p_j = [a_{j_1}, \ldots, a_{j_m}]$ donde:
  - La posesión corresponde a un único equipo E_k, que es quien está atacando.
  - La acción a_{j_m} representa el final de la posesión, pudiendo darse por múltiples factores como pase fallido, pelota fuera de la cancha, final del período, falta, entre otros.
- Cada acción posee f características, las cuales pueden dividirse en los siguientes grupos:
  - **Contexto de la acción**
    - Timestamp

    - (x, y) origen
    - Si hay presión del rival (valor *booleano*)
    - Resultado parcial del partido
  - **Ejecución de la acción**
    - Tipo de acción (remate, pase, conducción, etc.)
    - Parte del cuerpo usada (pie, cabeza, otro)
    - Técnica utilizada (normal, de volea, globo, etc.)
    - Resultado de la acción (ok, balón fuera, offside, etc.)
    - Características propias del tipo de acción (para pases: el tipo de pase, altura; para remates: de primera, mano a mano, arco libre, etc.)
  - **Variables Derivadas**
    - Distancia al arco
    - Ángulo hacia el arco
    - Diferencia de gol
    - Goles a favor

- El partido puede verse como una secuencia de $N+1$ estados $[S_0, \ldots, S_i, \ldots, S_N]$ donde la acción a_i es aquella que provoca la transición de $S_{i-1} \rightarrow S_i$.
- Cada estado S_i tiene asociada una probabilidad de gol para el equipo atacante E_k, que corresponde a la probabilidad de que E_k anote en los próximos s segundos:

$$P_{\text{gol } E_k}(S_i) = \text{probabilidad de que } E_k \text{ anote un gol en } s \text{ segundos}$$

5.3.2. Planteo del Problema

A partir de las definiciones anteriores, se busca estimar la probabilidad de gol de cada estado de juego. Se lo trata como un problema de **clasificación**, pero se entrena un **clasificador probabilístico** ya que el interés está puesto en conocer la probabilidad de gol más que la clase discreta (gol/no gol).

$\boldsymbol{X} =$ características de los estados de juego S_i

$$\boldsymbol{Y} = \begin{cases} 1 & \text{si el equipo atacante anota un gol en los próximos } s \text{ segundos} \\ 0 & \text{caso contrario} \end{cases}$$

$\boldsymbol{F} =$ clasificador probabilístico

Por otro lado, el problema es netamente **desbalanceado**, ya que los estados con gol asociado representan sólo el 1 % del total.

5.3.3. Métricas a Utilizar

En cuanto a los criterios de evaluación, se aprovechan las siguientes medidas:

- **Brier Score** (Brier, 1950)
- **Brier Skill Score** (Ford, 2000)

- **Logarithmic Loss** (Vovk, 2015)
- **Curva ROC** (Fawcett, 2006)

La primera de las métricas se define como:

$$\text{Brier Score}(y, \hat{y}) = \frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2,$$

donde n es la cantidad de predicciones, y es el vector de valores reales (0 y 1) e $\hat{y}$ es el vector de probabilidades predichas. Esencialmente, representa el error cuadrado promedio entre el valor real y la probabilidad inferida, pudiendo tomar valores entre 0 (predicción perfecta) y 1 (falla absoluta en la predicción).

Para enriquecer la comparación de *Brier Score*, se puede calcular la métrica *Brier Skill Score*, la cual refleja el desempeño relativo de un modelo F con respecto a otro modelo de referencia ref en términos de *Brier Score* (BS).

$$\text{Brier Skill Score} = 1 - \frac{BS_F}{BS_{ref}}$$

Si los modelos comparados tienen igual BS, se obtendrá un valor de 0. Por el contrario, si el modelo F predice mejor que el de referencia, se obtendrán valores positivos; caso contrario, valores negativos. Cuanto más se acerque a uno, mejor será el clasificador. Típicamente, como modelo de referencia se toma a aquel que siempre devuelve la *probabilidad a priori* (enfoque utilizado en el trabajo actual).

Otra de las métricas de interés se define mediante la siguiente fórmula:

$$\text{Logarithmic Loss}(y, \hat{y}) = -\frac{1}{n}\sum_{i=1}^{n}[y_i \ln(\hat{y}_i) + (1 - y_i)\ln(1 - \hat{y}_i)]$$

Esta medida es utilizada para cuantificar qué tan cerca está la probabilidad predicha del valor binario real. Para una predicción perfecta, se obtiene un valor de 0 y, cuanto más divergencia haya, mayor será el valor final (pudiendo tomar valores superiores a 1).

Finalmente, la *Curva ROC* es útil a la hora de evaluar el desempeño de modelos de clasificación binaria. Ayuda a visualizar el balance entre dos métricas clave de rendimiento como la *sensibilidad* (Tasa de Verdaderos Positivos) y *especificidad* (Tasa de Verdaderos Negativos). En particular, se prestará atención a la AUC (Área Bajo la Curva), según la cual un valor más grande de AUC indica un mejor desempeño, con un AUC de 1 representando a un clasificador perfecto.

5.4. Flujo de Datos

En esta sección se describen los pasos que se siguieron para la construcción del modelo de *Peligro de Gol*.

1. **Simplificación y Filtrado de Acciones**: reducción de dimensionalidad en varios sentidos (filas, columnas y cardinalidad de cada una de las variables categóricas).
2. **Ingeniería de Características (Feature Engineering)**: se enriquecen las acciones con nuevos campos (derivados o con nuevas características).

3. **Construcción de Estados de Juego**: a partir de las acciones, se generan los estados de juego que servirán para entrenar el modelo.
4. **Generación de Etiquetas**: tomando los goles anotados, a cada estado se lo etiqueta con 1 o 0 dependiendo de la ocurrencia de un gol dentro de los s segundos de juego.
5. **Entrenamiento del Modelo**: utilizando a los estados con sus respectivas etiquetas, se entrena un modelo probabilístico que busca predecir la probabilidad de gol de nuevos estados de juego.
6. **Aplicación del Modelo**: sobre un conjunto de acciones que no fue utilizado para el entrenamiento, se aplica el modelo de *Peligro de Gol* y se determina la peligrosidad generada en base a la diferencia de probabilidad de gol entre su estado anterior y posterior.
7. **Análisis de Resultados**: se exploran los resultados para analizar su consistencia e interpretabilidad.

5.4.1. Simplificación y Filtrado de Acciones

Este paso tiene como objetivo hacer más sencilla la información presente en cada acción y también a limitarse a aquellos eventos que son de interés para el entrenamiento. Esencialmente consiste en:

- **Reducción de cantidad de columnas**: sobreviven solamente aquellas que se refieren a características ofensivas del evento. Se pasa de 90 a 30 variables.
- **Reducción de cantidad de filas**: se descartan acciones que corresponden a eventos tácticos como cambios de posiciones o sustituciones y también a acciones de tipo defensivo.
- **Reducción de cardinalidad de variables categóricas**: se simplifican variables categóricas pudiendo unificar aquellas con similar significado o agrupando en una etiqueta genérica a aquellas de ocurrencia infrecuente. Un ejemplo es el caso de *body_part* donde se simplifica a *Left Foot*, *Right Foot*, *Head* y *Other*.

5.4.2. Ingeniería de Características (Feature Engineering)

En esta etapa se enriquece cada una de las acciones con nuevas características, procedimiento conocido como *feature engineering* (FE). Durante el entrenamiento se analiza la utilidad de cada una de ellas de cara al modelo final. A continuación, se detallan las nuevas variables generadas, agrupadas por tipo.

Características de Ubicación

En primer lugar, se agregan variables asociadas a las distancias entre el inicio de la acción y el centro del arco rival.

- *dx_goal* - Distancia al centro del arco en coordenada x: $arco\_x - x$
- *dy_goal* - Distancia al centro del arco en coordenada y: $arco\_y - y$
- *total_distance_goal* - Distancia al centro del arco en coordenadas (x, y): $\|(arco\_x, arco\_y) - (x, y)\|$
  (según el sistema de coordenadas de *Statsbomb*, el centro del arco está en las coordenadas $(120, 40)$)

Adicionalmente, se incorpora una variable relacionada al ángulo que se genera entre la posición inicial y el arco rival.

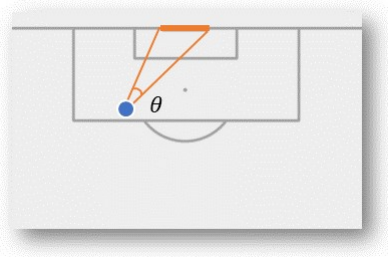


FIGURA 5.1: Ángulo θ que se genera entre el vector que forma el arco y la posición de inicio de la jugada

Considerando que los triángulos que se forman no siempre son rectángulos, para obtener el ángulo se utiliza el *Teorema de los Cosenos* y se define la característica como:

$$\theta = angle\_to\_goal = \arccos\left(\frac{a^2 + b^2 - c^2}{2ab}\right),$$

donde a es la distancia al primer palo, b es la distancia al segundo palo y c es la longitud del arco (8 para *Statsbomb*).

Características de Juego

Teniendo en cuenta el desarrollo mismo del partido, se generan variables como las que siguen:

- *possession_at_risk* - Posesión en Riesgo: valor *booleano* que indica si la acción realizada no es del equipo atacante, con lo cual está poniendo en peligro la posesión en curso (recordar que en el medio de la posesión de un equipo puede haber acciones aisladas del rival, sin que éste tome control efectivo de la misma).
- *attacking_players_in_pitch* y *defending_players_in_pitch* - Cantidad de Jugadores en Campo: a partir de las expulsiones que ocurren en el partido, se calcula la cantidad de jugadores de cada uno de los equipos
- *attacking_goals_count*, *defending_goals_count* y *goal_difference* - Goles y Diferencia de Gol: se cuentan los goles convertidos por cada equipo y la diferencia de gol parcial.

Características Básicas

Finalmente, se unifican dos variables que refieren al tiempo de juego, dando lugar a:

$$minute\_seconds = minutes + \frac{seconds}{60}$$

5.4.3. Construcción de Estados de Juego

A la hora de construir los estados de juego, se toma cada acción a_i y se genera el estado S_{i-1}. Es decir que a partir de una acción se construye únicamente su estado anterior. Dicho de otra forma, el estado conoce todo lo previo a la ejecución propia de la acción, pero no está al tanto de la decisión que terminó tomando el jugador (si

dio un pase, lanzó un centro o pateó al arco, por ejemplo). Además, puede contar con información de k acciones anteriores.

En concreto, el estado S_{i-1} se construye con:

- Información de a_i:
  - Minutos y segundos (*variable continua*)
  - Período (*variable numérica discreta*)
  - (x, y) (*variables continuas*)
  - Distancia y ángulo al arco (*variables continuas*)
  - Si el jugador está presionado por el rival (*variable booleana*)
  - Si la posesión está en riesgo (*variable booleana*)
  - Cantidad de jugadores en cancha atacando y defendiendo (*variables numéricas discretas*)
  - Goles a favor, en contra y diferencia de gol (*variables numéricas discretas*)
- Información de $a_j, j = \{i-1, \ldots, i-k\}$:
  - (x,y) (*variables continuas*)
  - Minutos y segundos (*variable continua*)
  - Tipo de acción (*variable categórica*)
  - Subtipo de acción (*variable categórica*)
  - Parte del cuerpo (*variable categórica*)
  - Técnica utilizada (*variable categórica*)
  - Resultado de acción (*variable categórica*)
  - Pase - Centro (*variable booleana*)
  - Pase - Centro Atrás (*variable booleana*)
  - Pase - Cambio de Frente (*variable booleana*)
  - Pase - Altura (*variable categórica*)
  - Remate - De Primera (*variable booleana*)
  - Remate - Mano a Mano (*variable booleana*)
  - Remate - Arco Libre (*variable booleana*)
  - Distancia y ángulo entre acción y el arco (*variables continuas*)
  - Si el jugador está presionado por el rival (*variable booleana*)
  - Si la posesión está en riesgo (*variable booleana*)
  - Distancia entre acción a_j y a_i (*variable continua*)
  - Tiempo entre acción a_j y a_i (*variable continua*)
  - Cantidad de jugadores en cancha atacando y defendiendo (*variables numéricas discretas*)
  - Goles a favor, en contra y diferencia de gol (*variables numéricas discretas*)

5.4.4. Generación de Etiquetas

En cuanto a la variable *target* de entrenamiento, a cada estado se le asigna la clase *gol/no gol* usando los minutos y segundos del estado y observando la ocurrencia de un gol en los siguientes s segundos, siempre dentro del mismo período (primer o segundo tiempo del partido normal o prórroga).

Para definir el valor de s, se usó la mediana de duración de las posesiones, la cual se ubica en los **10 segundos**. Este número coincide con el utilizado por *Stats Perform* en su modelo de *Possession Value*.

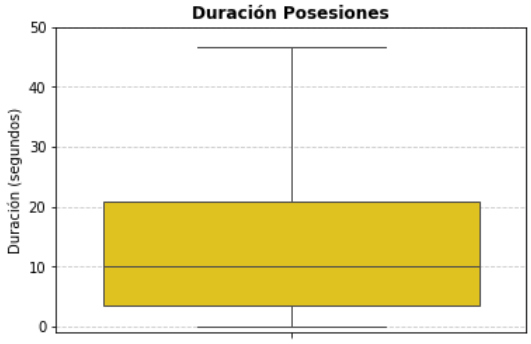


FIGURA 5.2: Distribución de duración de posesiones en el dataset (sin outliers)

Duración Posesión (segundos)

Mín	Máx	25 %	Mediana	75 %	Media	Std. Dev.
0.0	332.40	3.60	10.20	21.00	15.55	17.41

5.4.5. Entrenamiento del Modelo

Para esta etapa, se realiza el entrenamiento del modelo usando **LightGBM** (Ke et al., 2017), uno de los algoritmos de *boosting* basados en árboles más populares en la actualidad. Este tipo de modelos se caracteriza por su gran velocidad de entrenamiento, su uso eficiente de memoria y su capacidad para manejar datos faltantes directamente durante el proceso de entrenamiento, situación que se presenta en nuestro conjunto de datos.

La estrategia de entrenamiento consiste en:

1. División del conjunto de datos en entrenamiento y testeo
   - **Entrenamiento**: Ligas Big Five Temporada 2015/16 + Eurocopa 2020 + Temporadas Incompletas Lionel Messi (2425 partidos, 93 %)
   - **Testeo**: Mundial 2018, Mundial 2022 y Copa África 2023 (180 partidos, 7 %)
2. Tuneo y selección de hiperparámetros usando la técnica de **Optimización Bayesiana** y **cross-validation** con 5 *folds* estratificados.
3. Validación de resultados contra conjunto de testeo

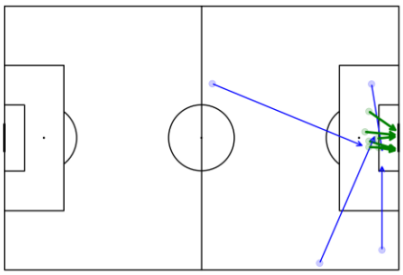
5.4.6. Aplicación del Modelo

Sobre el conjunto de acciones que no fue utilizado para el entrenamiento, se aplica el modelo determinando el *Peligro de Gol* (**pdg**) generado por cada acción a_i según la siguiente fórmula:

$$pdg(a_i) = \Delta P_{gol}(a_i) = P_{gol}(S_i) - P_{gol}(S_{i-1})$$

5.4.7. Análisis de Resultados

Con la peligrosidad de cada acción ya calculada, se analizan los valores obtenidos por tipo de acción y se exploran visualmente eventos particulares para ver qué tan confiables son los resultados obtenidos.



5.5. Experimentación

Con el fin de dar con un modelo robusto, se llevaron a cabo una serie de experimentos, cuyos principales lineamientos se describen a continuación:

- Distintos conjuntos de variables:
  - Variables originales (sin *feature-engineering*) y una acción anterior
  - Variables originales + *feature-engineering* y una acción anterior
  - Variables originales + *feature-engineering* y dos acciones anteriores
  - Variables originales + *feature-engineering* y tres acciones anteriores
  - Selección de variables que aparecen como más influyentes
- Diferentes estrategias de entrenamiento:
  - *LightGBM* con hiperparámetros fijos
  - *LightGBM* con *Optimización Bayesiana* con rango de hiperparámetros y *cross-validation* con 5 *folds* estratificados

Para el *LightGBM* con hiperparámetros fijos, se utilizaron los *default*, a excepción de $max\_depth = 3$.

Para la *Optimización Bayesiana*, se realizaron *25 iteraciones* y se usaron los siguientes rangos[1]:

Hiperparámetro	Valores
learning_rate	$[0.01, 0.3]$
num_leaves	$[5, 1000]$
n_estimators	$[5, 250]$
colsample_bytree	$[0.2, 1]$
min_child_samples	$[0, 2000]$
reg_alpha	$[0, 1000]$
reg_lambda	$[0, 1000]$
max_bin	63, 255 o 1023

[1]Para el modelo final, la cota superior de *n_estimators* se incrementará a 1000

Se inspecciona la importancia de variables usando **SHAP Values** (Lundberg y Lee, 2017), buscando entender las distintas características que guían el modelo.

Split Conjunto de Datos

# Registros Train	4.614.359	93 %
# Registros Test	336.459	7 %
# Registros Total	4.950.818	100 %

Se presentan los principales experimentos realizados, donde se muestra el valor de cada métrica en el conjunto de testeo:

Exp.	Entrenamiento	Variables	# Variables	Brier Score	Brier Skill Score	Logloss	AUC
1	LGBM Default	Básicas + 1 acción anterior (sin FE)	21	0.00636	0.03641	0.03330	0.84171
2	LGBM Default	1 acción atrás + FE Full	45	0.00625	0.05259	0.03210	0.86238
3	LGBM Default	2 acción atrás + FE Full	75	0.00624	0.05443	0.03201	0.86399
4	LGBM Default	3 acción atrás + FE Full	105	0.00623	0.05548	0.03195	0.86523
5	LGBM Opt. Bay.	Básicas + 1 acción anterior (sin FE)	21	0.00630	0.04610	0.03290	0.84497
6	LGBM Opt. Bay.	1 acción atrás + FE Full	45	0.00623	0.05625	0.03192	0.86386
7	LGBM Opt. Bay.	2 acción atrás + FE Full	75	0.00621	0.05910	0.03175	0.86678
8	LGBM Opt. Bay.	3 acción atrás + FE Full	105	0.00621	0.05969	0.03167	0.86850

Es posible observar que:

- La incorporación de variables producto del *feature-engineering* arroja resultados positivos, mejorando todas las métricas bajo análisis
- El hecho de agregar más acciones de historia produce mejoras pero no tan significativas como el *feature-engineering*
- Como era de esperarse, el *tuning* de hiperparámetros lleva a mejores resultados que utilizando los estándar

Los gráficos que siguen muestran los valores *SHAP* de los últimos 4 modelos, con las 20 variables más influyentes:

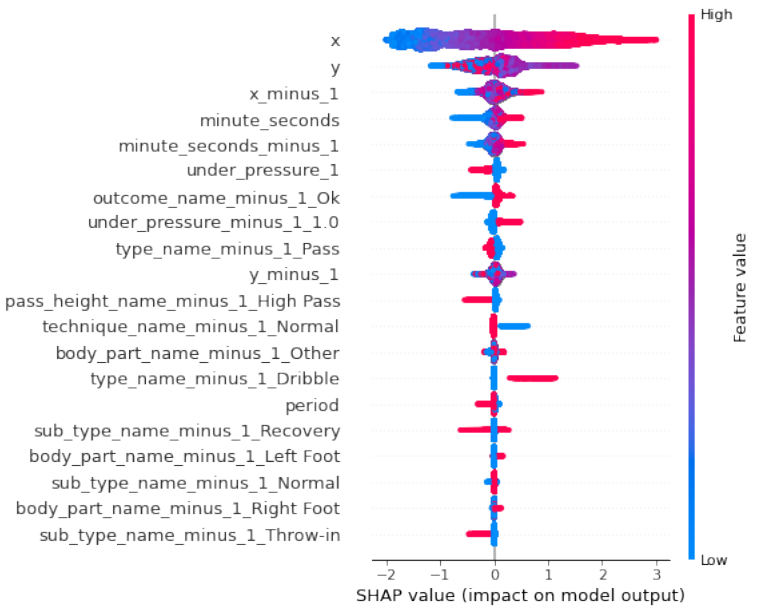


FIGURA 5.3: SHAP del experimento n°5 (1 acción anterior sin FE)

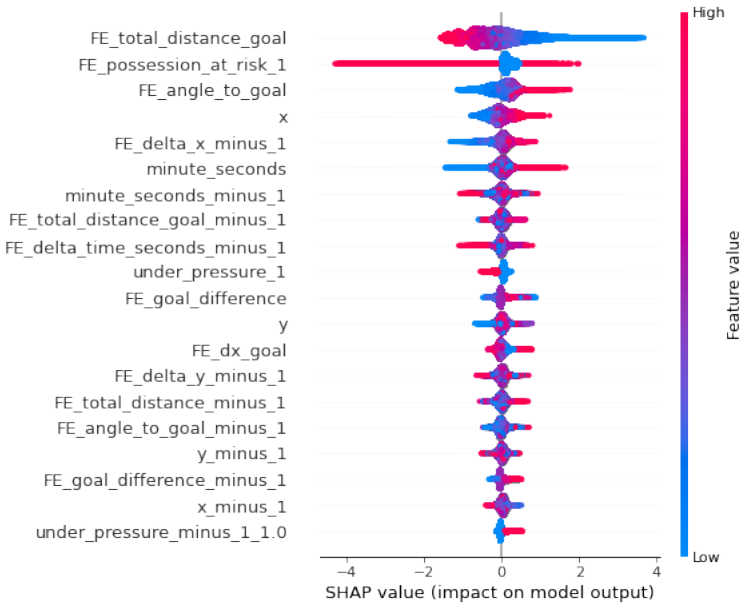


FIGURA 5.4: SHAP del experimento n°6 (1 acción anterior con FE)

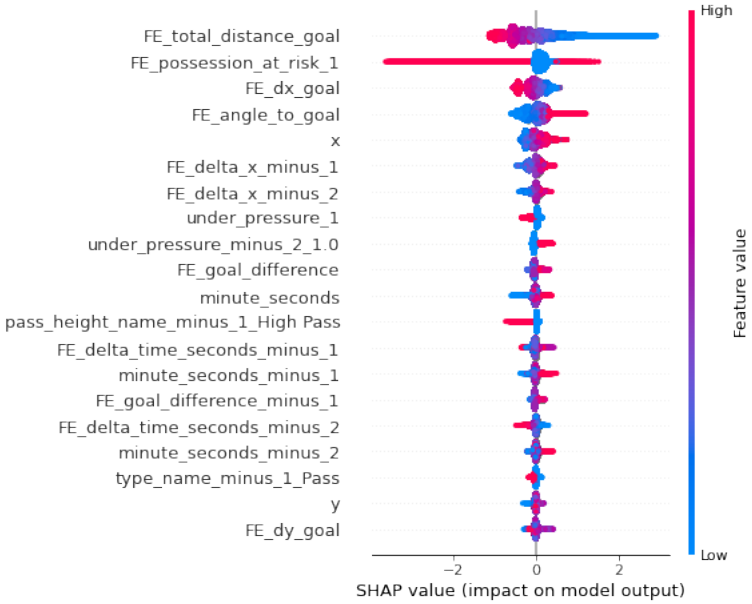


FIGURA 5.5: SHAP del experimento n°7 (2 acciones anteriores con FE)

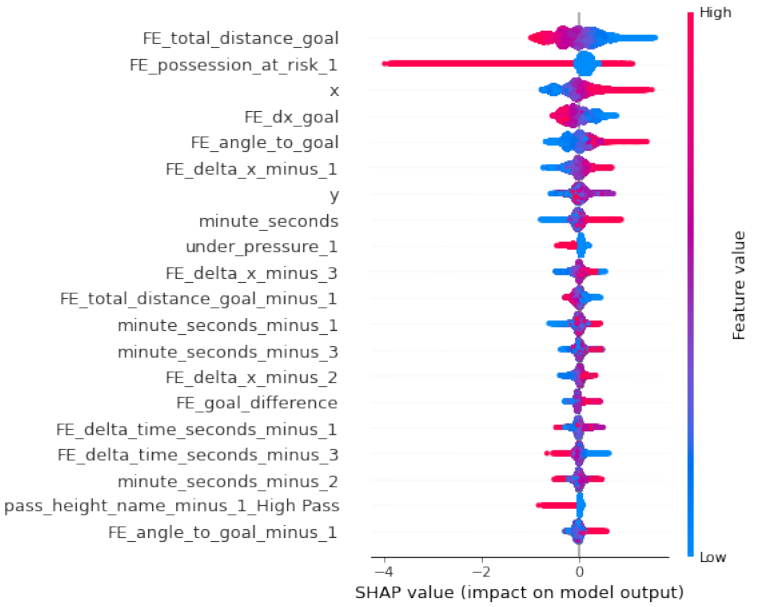


FIGURA 5.6: SHAP del experimento n°8 (3 acciones anteriores con FE)

Es posible apreciar que se destacan las siguientes variables:

- *Distancia al arco*: lógicamente, cuanto menor distancia, más tendencia a aumentar la probabilidad de gol
- *Posesión en riesgo*: influyente pero sin un patrón claro
- *Ángulo al arco*: cuanto mayor ángulo (más visibilidad del arco), mayor tendencia a subir la probabilidad de gol
- *Minutos y segundos*: siendo consistente con uno de los patrones observables en el fútbol, a mayor tiempo, mayor tendencia a subir la probabilidad
- *Bajo presión rival*: de existir presencia del contrario, tiende a reducirse la probabilidad de gol
- *Pase Previo Alto*: si la acción anterior es de este tipo, se observa tendencia a reducir la probabilidad de gol (la jugada tiene un mayor grado de dificultad que si viene a ras del piso)
- *Delta de Tiempo y en coordenada x*: aparecen como importantes para acciones anteriores (1, 2 y 3)

Entonces, a partir de estos resultados, se construye un modelo seleccionando 61 características:

- Información de acción anterior completa
- Variables de distancia y ángulo al arco, delta de coordenada x y deltas de tiempo de las *3 acciones anteriores*
- Tipo y subtipo de acción, parte del cuerpo y resultado de la acción de las *3 anteriores*
- Se remueven variables asociadas a cantidad de goles anotados y la diferencia de gol para evitar el riesgo de correlación con el poderío del equipo (punto mencionado por *Statsbomb* en su modelo *On-Ball Value*)

Exp.	Entrenamiento	Variables	# Variables	Brier Score	Brier Skill Score	Logloss	AUC
9	LGBM Opt. Bay.	Seleccionadas, hasta 3 acciones atrás	61	0.00622	0.05758	0.03176	0.86789

Los hiperparámetros finales producto de la *Optimización Bayesiana* fueron:

Hiperparámetro	Valores
learning_rate	0,1
num_leaves	812
n_estimators	201
colsample_bytree	0,6
min_child_samples	127
reg_alpha	1,24
reg_lambda	475,68
max_bin	255
metric	logloss
objective	binary

Por último, las variables que resultaron más influyentes en el modelo final son:

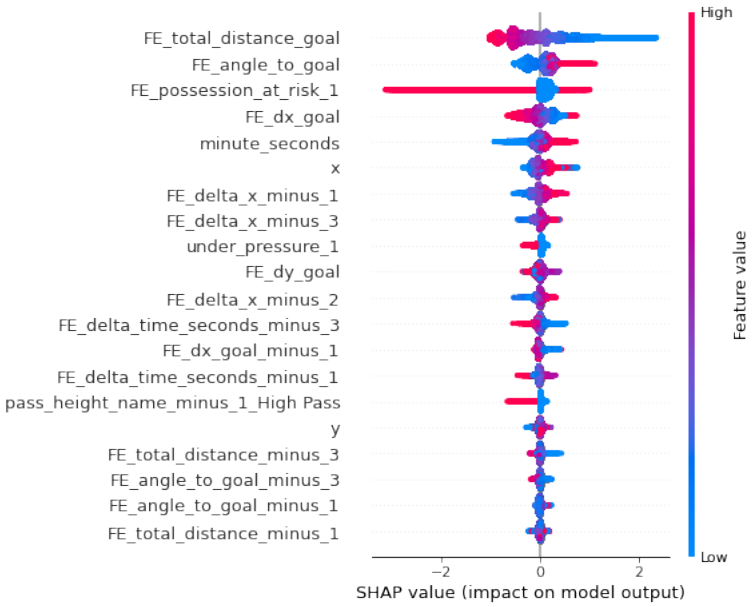


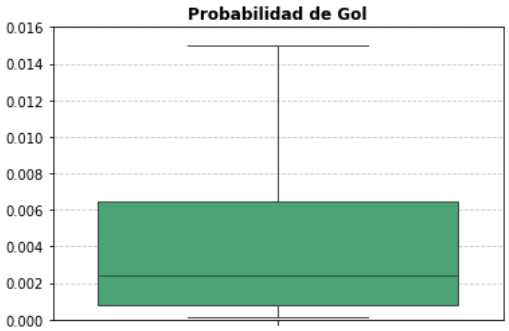
FIGURA 5.7: SHAP del experimento n°9 (selección de features, hasta 3 acciones anteriores con FE)

5.6. Análisis de Resultados

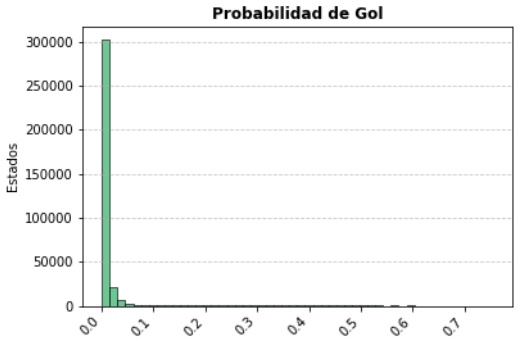
Con la intención de verificar la consistencia de los modelos entrenados, antes de calcular el *peligro de gol*, se analiza cómo se distribuyen las probabilidades de gol de cada estado según su valor y también según su ubicación (x, y):

Probabilidad de Gol de Estados

Mín	Máx	25 %	Mediana	75 %	Media	Std. Dev.
0.0	0.75	0.00	0.00	0.01	0.01	0.02



(A) Boxplot de probabilidad de gol (sin outliers)



(B) Histograma de probabilidad de gol

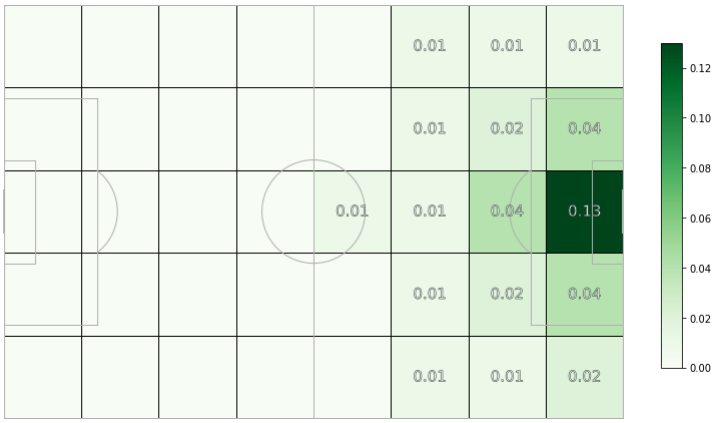


FIGURA 5.9: Promedio de probabilidad de gol por zona de la cancha

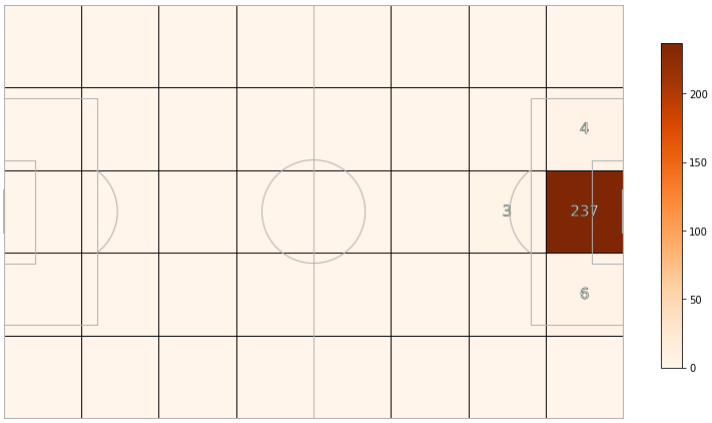
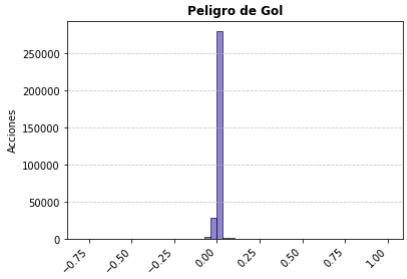


FIGURA 5.10: Zonas donde se ubican el top 250 de las jugadas con mayor probabilidad de gol

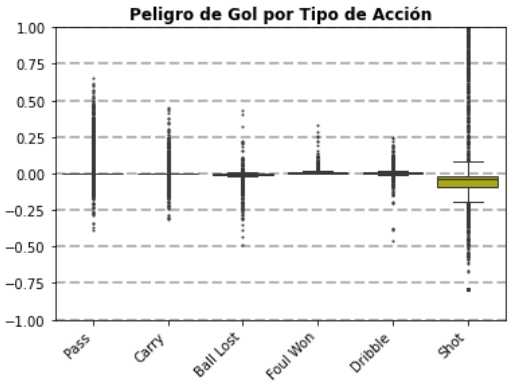
Como era de esperarse, la probabilidad de gol se concentra en la zona central del área y muestra una distribución razonable. Además, el top de los 250 estados más peligrosos se ubica en posiciones coherentes.

En cuanto al modelo propuesto de **Peligro de Gol**, se estudia su distribución en términos generales y por tipo de acción.

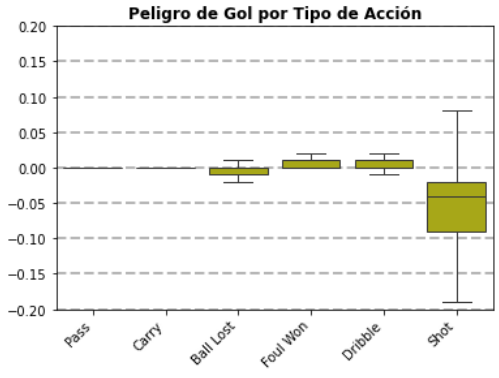


Distribución Peligro de Gol

Menor a 0	Igual a 0	Mayor a 0
10.55 %	78.78 %	10.67 %



(A) Boxplot de Peligro de Gol (con outliers)



(B) Boxplot de Peligro de Gol (sin outliers)

Estadísticas Descriptivas *pdg* por tipo

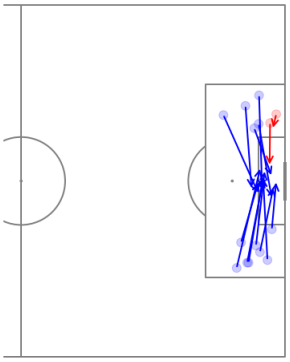
Tipo Acción	Mín	Máx	25 %	Mediana	75 %	Media	Std. Dev.
Pass	-0.39	0.65	0.00	0.00	0.00	0.00	0.02
Carry	-0.31	0.45	0.00	0.00	0.00	0.00	0.01
Ball Lost	-0.49	0.43	-0.01	-0.01	0.00	-0.01	0.03
Foul Won	0.00	0.33	0.00	0.00	0.01	0.01	0.02
Dribble	-0.46	0.25	0.00	0.00	0.01	0.00	0.03
Shot	-0.79	1	-0.09	-0.04	-0.02	0.00	0.28

Se puede apreciar una gran cantidad de acciones con aporte nulo (valoración 0) y una distribución pareja entre las positivas y negativas. Al hacer el corte por tipo de acción, se logran observar los siguientes detalles:

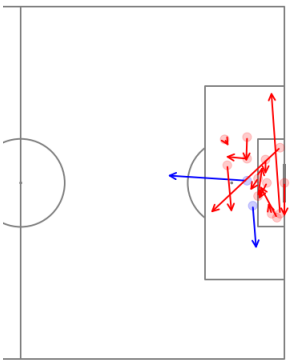
- Las pelotas perdidas tienden a tener valoración negativa
- Por el contrario, los foules recibidos generalmente se valorizan positivamente
- La mayor dispersión se hace presente en los remates al arco

A continuación se analizan visualmente[2] las 15 jugadas con mayor y menor *pdg* por tipo de acción, buscando verificar la consistencia del modelo.

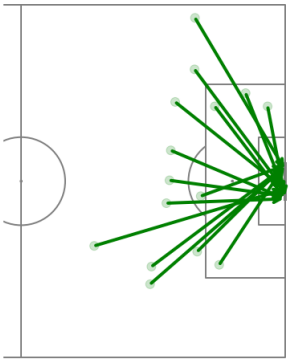
[2]El color se basa en el campo *outcome_name*, siendo *Ok*=Azul, *Fail*=Rojo y *Goal*=Verde



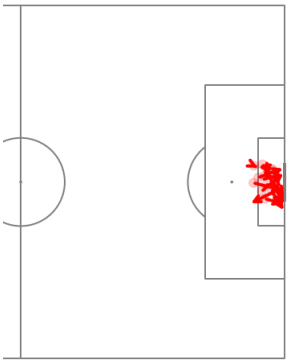
(A) Pases con mayor Peligro de Gol (top 15)



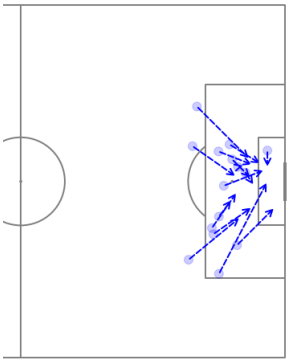
(B) Pases con menor Peligro de Gol (bottom 15)



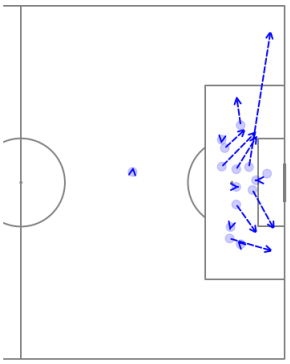
(A) Remates con mayor Peligro de Gol (top 15)



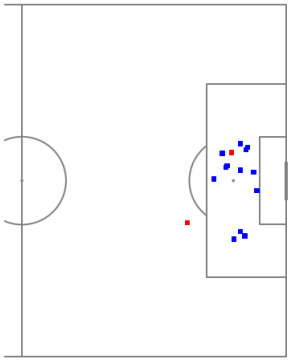
(B) Remates con menor Peligro de Gol (bottom 15)



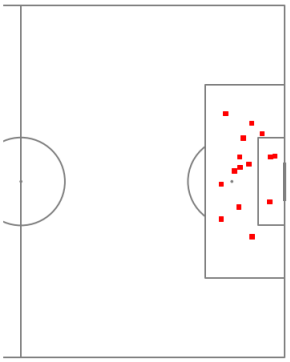
(A) Conducciones con mayor Peligro de Gol (top 15)



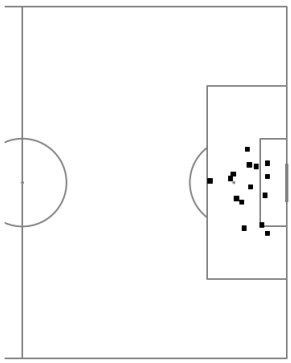
(B) Conducciones con menor Peligro de Gol (bottom 15)



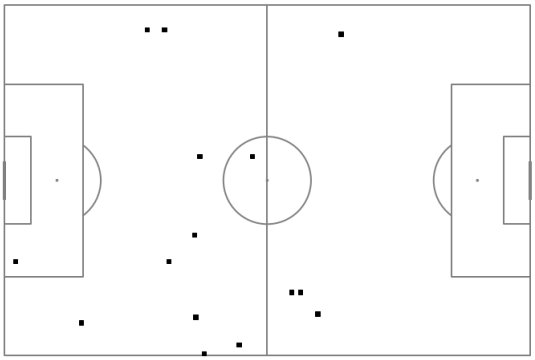
(A) Regates con mayor Peligro de Gol (top 15)



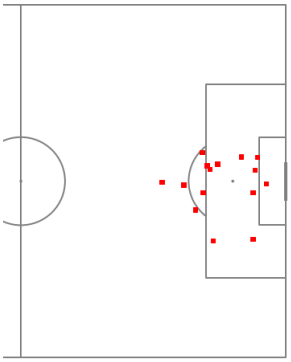
(B) Regates con menor Peligro de Gol (bottom 15)



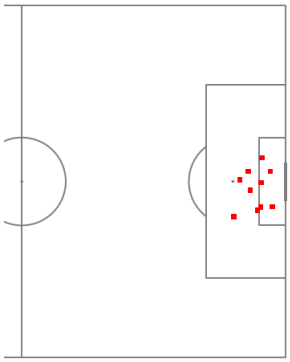
(A) Foules Ganados con mayor Peligro de Gol (top 15)



(B) Foules Ganados con menor Peligro de Gol (bottom 15)



(A) Pérdidas con mayor Peligro de Gol (top 15)



(B) Pérdidas con menor Peligro de Gol (bottom 15)

En todos los casos, es posible notar la razonabilidad de las jugadas resaltadas (tanto positiva como negativamente), mostrando una consistencia y coherencia con lo que se esperaría de un modelo de estas características.

Finalmente, se toma como ejemplo el Mundial 2022 y se calculan los jugadores más relevantes según *pdg*[3]:

#	Jugador	Equipo	pdg	Goles
1	Kylian Mbappé	Francia	5.16	9
2	Lionel Messi	Argentina	2.97	9
3	Bukayo Saka	Inglaterra	2.55	3
4	Julián Álvarez	Argentina	2.51	4
5	Cody Gakpo	Países Bajos	2.38	3
6	Mohammed Kudus	Ghana	1.94	2
7	Wout Weghorst	Países Bajos	1.90	3
8	Vincent Aboubakar	Camerún	1.83	2
9	Vinícius Júnior	Brasil	1.80	1
10	Rafael Leão	Portugal	1.72	2

Los nombres que surgen parecen razonables, con varios de los destacados del torneo en los primeros lugares. Sin embargo, parece haber una tendencia a ponderar a aquellos que anotaron muchos goles en la competición. Un análisis de correlación confirma esta hipótesis, obteniéndose un coeficiente de **0.68** entre los goles convertidos y *pdg*, indicador de correlación fuerte. Si bien tiene sentido que los goles se traduzcan en una valoración positiva, estos modelos persiguen como objetivo justamente resaltar situaciones que no surgen de las estadísticas clásicas. Teniendo en

[3]Se agrega una columna de goles que incluye los penales de partidos con definición por esta vía

cuenta esta situación es que se propone un esquema de pesos para cada tipo de acción, el cual permitiría atenuar el impacto de los goles convertidos en el análisis (lo llamaremos **pdg_ponderado**).

$$pdg\_ponderado_{Total} = \sum_{t \, \epsilon \, tipo\_accion} w_t \times pdg_t$$

Moderando el impacto de los remates ($w_{Shot} = 0{,}5$ y dejando el resto en 1), se obtiene una distribución más pareja del efecto de cada acción.

Estadísticas Descriptivas *pdg_ponderado* **por tipo**

Tipo Acción	Mín	Máx	25 %	Mediana	75 %	Media	Std. Dev.
Pass	-0.39	0.65	0.00	0.00	0.00	0.00	0.02
Carry	-0.31	0.45	0.00	0.00	0.00	0.00	0.01
Ball Lost	-0.49	0.43	-0.01	-0.01	0.00	-0.01	0.03
Foul Won	0.00	0.33	0.00	0.00	0.00	0.01	0.02
Dribble	-0.46	0.25	0.00	0.00	0.01	0.00	0.03
Shot	-0.40	0.5	-0.04	-0.02	-0.01	0.00	0.14

Al calcular el top 10 de jugadores luego de esta adaptación, se obtiene:

#	Jugador	Equipo	pdg_ponderado	Goles
1	Kylian Mbappé	Francia	2.93	9
2	Lionel Messi	Argentina	2.47	9
3	Julián Álvarez	Argentina	1.67	4
4	Bukayo Saka	Inglaterra	1.54	3
5	Antoine Griezmann	Francia	1.47	0
6	Mateo Kovačić	Croacia	1.43	0
7	Theo Hernández	Francia	1.37	1
8	Vinícius Júnior	Brasil	1.36	1
9	Kevin De Bruyne	Bélgica	1.35	0
10	Mohammed Kudus	Ghana	1.29	2

Es posible notar que comienzan a aparecer jugadores que no convirtieron goles en el listado, consistente con la reducción en la correlación con los goles anotados (disminuyendo a **0.57**).

5.7. Comparación vs. VAEP

Con el objetivo de comparar el rendimiento del modelo presentado contra otras alternativas disponibles en la actualidad, se tomó la versión publicada por el grupo *DTAI* de la *Universidad KU Leuven* de Bélgica[4]. Se realizó el entrenamiento estándar usando n=3 (acciones anteriores a utilizar) y k=10 (cantidad de acciones para considerar un gol), con el mismo split en entrenamiento y testeo utilizado en la última sección.

Se comienza comparando los resultados en cuanto a métricas de evaluación de los modelos:

Modelo	# Columnas	Brier Score	Logloss	AUC
VAEP - Scores	64	0.00882	0.04542	0.82483
VAEP - Concedes	64	0.00239	0.01431	0.83727
Peligro de Gol - Hasta 3 feature atrás	61	0.00622	0.03176	0.86789

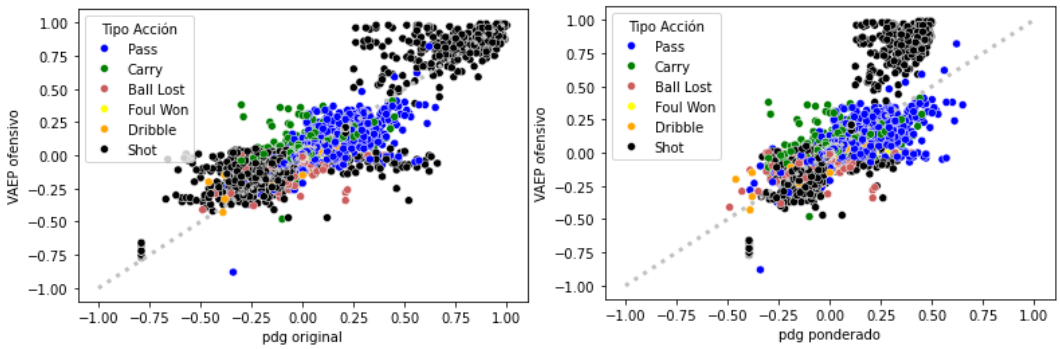
Como se mencionó en el capítulo anterior, el modelo VAEP valoriza tanto acciones ofensivas como defensivas y, para comparar contra *Peligro de Gol*, se considera

[4]Disponible en GitHub: https://github.com/ML-KULeuven/socceraction

solamente la componente ofensiva. De todas maneras, a partir de un análisis de correlación, se observa que el valor del VAEP está dominado por la parte ofensiva:

Correlación VAEP vs. VAEP Ofensivo	0.93
Correlación VAEP vs. VAEP Defensivo	0.40

Al comparar *pdg* (original y ponderado) contra *VAEP Ofensivo*, se puede notar visualmente una fuerte correlación entre ambos:



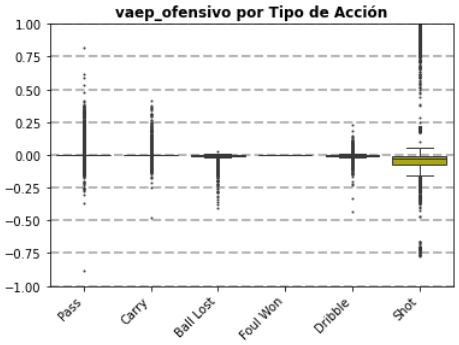
(A) Scatterplot VAEP Ofensivo vs. pdg original

(B) Scatterplot VAEP Ofensivo vs. pdg ponderado

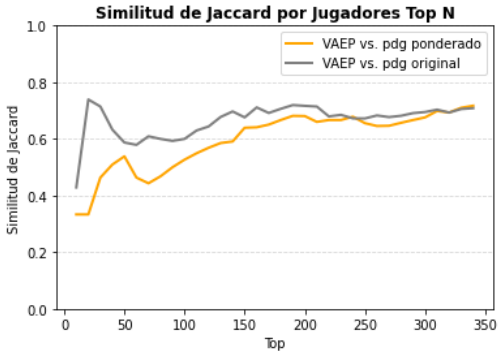
Correlación pdg (Original) vs. VAEP Ofensivo	0.89
Correlación pdg (Ponderado) vs. VAEP Ofensivo	0.80

Es posible apreciar que la ponderación diferente de los remates al arco produce una mayor diferenciación entre los modelos, bajando 0.09 la correlación entre ambos.

Al analizar la distribución por tipo de acción, se observa que VAEP Ofensivo tiene un comportamiento similar a *pdg*, mostrando una mayor dispersión en las acciones de remate. A su vez, las faltas recibidas no son consideradas por este modelo, lo que se traduce en una concentración de valores en el 0.



Posteriormente, enfocando el análisis al Mundial 2022, se obtiene el listado de jugadores top por métrica y se calcula la *Similitud de Jaccard*, la cual toma valores cercanos a 1 para conjuntos similares y ronda el 0 para disímiles.



Se observa mayor similitud entre *pdg* original y VAEP Ofensivo, estabilizándose cerca del 0.7 a partir del top 200.

Finalmente, se analiza la correlación entre las métricas y los goles anotados, apareciendo el VAEP Ofensivo como la medida con correlación más fuerte con los goles.

Correlación pdg (Original) vs. Goles Anotados	0.68
Correlación pdg (Ponderado) vs. Goles Anotados	0.57
Correlación VAEP Ofensivo vs. Goles Anotados	**0.78**

5.8. Resumen

Como resumen, se ha logrado desarrollar un modelo de *Possession Value*, el cual es capaz de capturar el concepto de peligrosidad, con resultados interpretables y consistentes con el ojo experto. A su vez, *Peligro de Gol* resulta comparable contra uno de los modelos desarrollados en la actualidad, tanto a nivel métricas como en distribución de sus valoraciones. Además, consigue una menor correlación con los goles anotados, otorgando un mayor balance a sus valoraciones y resaltando características que hoy no emergen de las estadísticas tradicionales. La utilización exclusiva de las asistencias y los goles puede llegar a destacar únicamente a aquellos que cerraron la jugada, ocultando a otros participantes que tuvieron un rol clave durante el desarrollo de la acción. Al mismo tiempo, permite separar los conceptos de cantidad y calidad, ya que no sólo se cuenta el número de pases o conducciones si no que se pondera a cada acción por su valor ofensivo real. De esta forma, jugadores que generen peligro ofensivo de manera consistente, independientemente del equipo donde se desempeñen, podrán ser detectados. En las secciones siguientes se aplica detalladamente este modelo, en particular la versión ponderada.

Capítulo 6

Caso de Estudio: Mundial Qatar 2022

6.1. Introducción

A lo largo de este capítulo se lleva a cabo un análisis **no tradicional** acerca de lo sucedido en el **Mundial de Qatar 2022**. Si bien la Copa del Mundo ya finalizó (todos conocemos su feliz desenlace) y sus datos se encuentran totalmente disponibles, se propone un enfoque distinto al habitual. En lugar de estudiar los desempeños de jugadores o equipos usando las estadísticas y métodos tradicionales, la propuesta consiste en simular la participación de un *científico de datos* dentro del cuerpo técnico de la *Selección Argentina*, sitúandose en los días previos de cada partido y analizando únicamente los datos disponibles hasta ese entonces. Todas las hipótesis y conclusiones son respaldadas por datos, sin necesidad de ver ningún video de los partidos. Sólo se considera la información del conjunto de datos disponible, una limitante con la cual es necesario lidiar. Se emplean las técnicas descriptas en los capítulos de *Estado del Arte* y se proponen nuevas visualizaciones basadas en el modelo de **Peligro de Gol**. Para cada encuentro se analiza el pre-partido, buscando hacer una *radiografía* del equipo rival, y también el post-partido, tratando de entender y explicar lo sucedido en el encuentro.

6.2. Experiencia Mundialista

Antes de comenzar a analizar a los rivales, se resume la última actuación argentina en Mundiales (Rusia 2018).

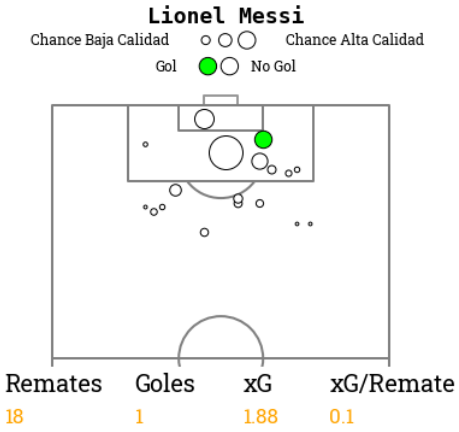
Del plantel actual solamente 7 jugadores fueron parte del Mundial anterior:

Jugador	Minutos Jugados	Partidos Jugados
Nicolás Otamendi	431	4
Lionel Messi	431	4
Nicolás Tagliafico	416	4
Ángel Di María	291	3
Franco Armani	239	2
Marcos Acuña	96	1
Paulo Dybala	27	1

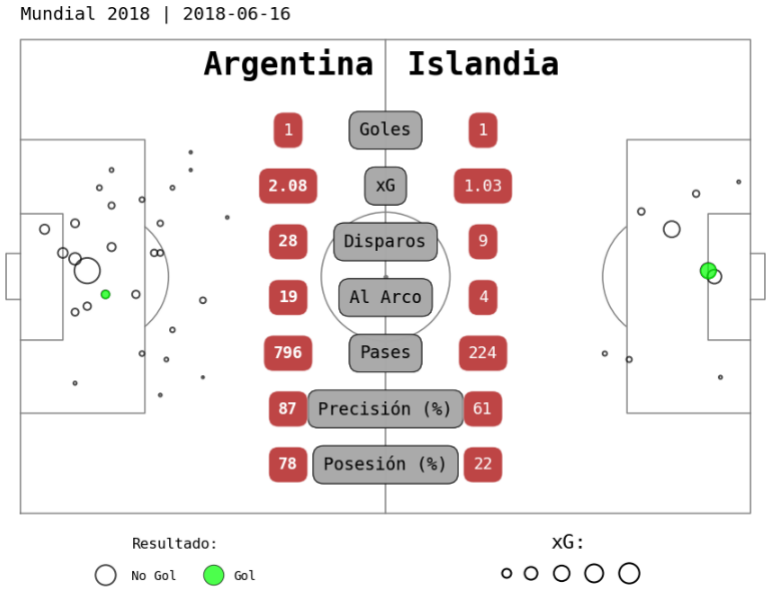
El recorrido del seleccionado consistió en:

Ronda	Partido	Goles ARG
Fase de Grupos	Argentina 1 - Islandia 1	Sergio Agüero
Fase de Grupos	Argentina 0 - Croacia 3	-
Fase de Grupos	Argentina 2 - Nigeria 1	Lionel Messi - Marcos Rojo
8vos de Final	Argentina 3 - Francia 4	Ángel Di María - Gabriel Mercado - Sergio Agüero

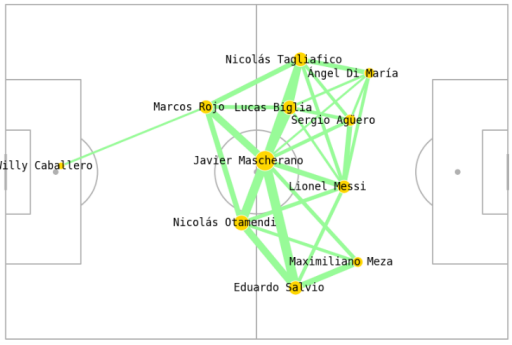
En cuanto a los disparos, quien más remató en el equipo argentino fue por lejos *Messi* (18 veces, 1 gol), concentrando casi un tercio de los tiros del equipo. Como puede observarse en el mapa de tiros de más abajo, la mitad de sus remates se realizaron desde fuera del área.



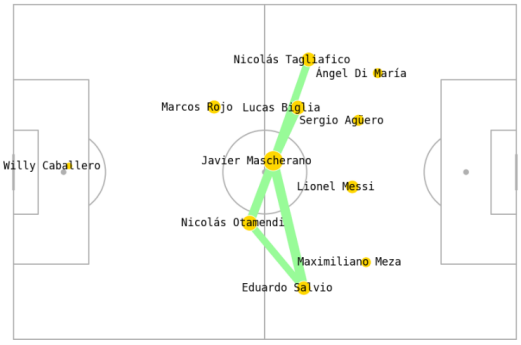
Para esta sección no se analiza toda la participación mundialista si no que se hace foco en el primer y último encuentro del Mundial.



El debut con Islandia estuvo lejos de las expectativas, donde el equipo tuvo amplio dominio de la pelota, cuadriplicando la cantidad de pases del rival, y remató más que el oponente pero no pudo llevarse la victoria.



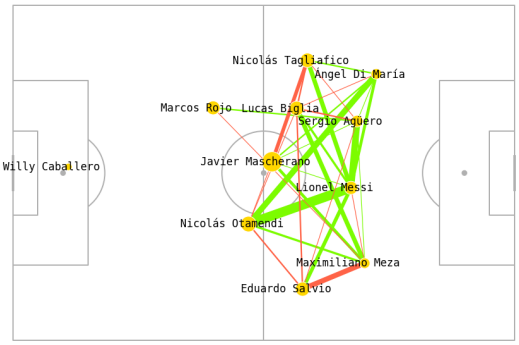
(A) Red de pases Argentina vs. Islandia (completo)



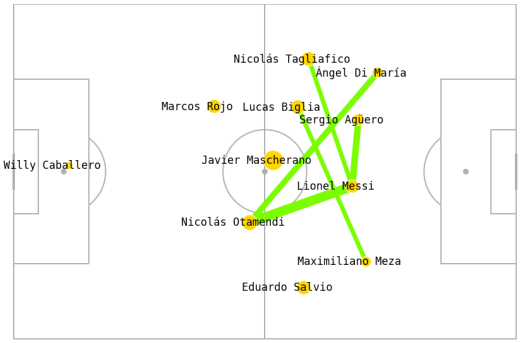
(B) Red de pases Argentina vs. Islandia (top 5 conexiones)

El mapa de pases de este partido muestra una participación muy activa de *Mascherano*, jugando un papel central en la red de pases (realizó 134 y recién Rojo lo sigue con 70). El *índice de centralidad* confirma esta percepción, arrojando un valor alto de **0.14**.

A continuación se presenta una variante de red de pases, donde se pondera cada conexión por su suma de **Peligro de Gol** en lugar de la cantidad de pases entre jugadores. De esta manera, se resaltan aquellos vínculos que generaron mayor peligrosidad en el partido. El color verde indica peligrosidad positiva (aumento en la probabilidad de gol), mientras que el rojo denota peligrosidad negativa (reducción en la probabilidad de gol).

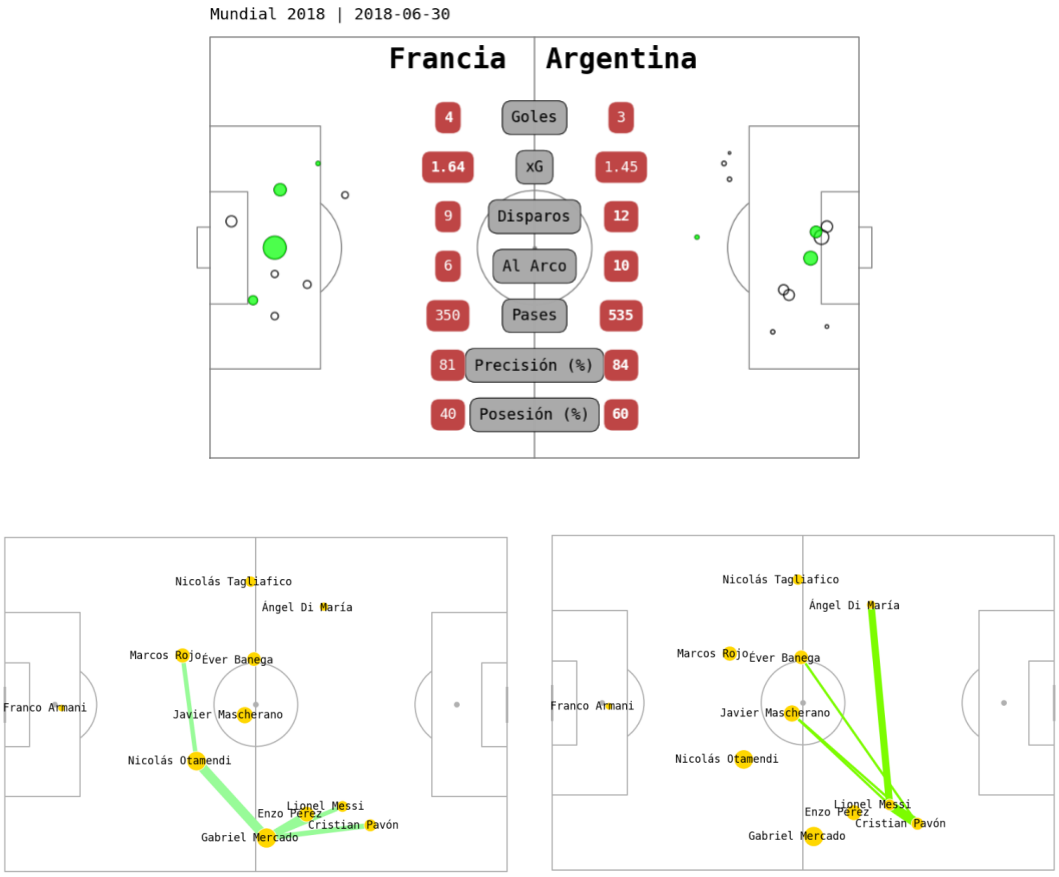


(A) Red de pases por pdg Argentina vs. Islandia (completo)



(B) Red de pases por pdg Argentina vs. Islandia (top 5 conexiones)

A partir de esta nueva visualización se puede observar un claro contraste, donde *Mascherano* ya deja de ser el centro de la red y emerge *Messi* como el jugador con más peligrosidad generada. De esta manera, se puede diferenciar entre cantidad y calidad de pases, trazando una línea divisoria entre acciones comunes de juego y aquellas con mayor incidencia ofensiva. Como consecuencia, es posible detectar a través de qué conexiones un equipo habitualmente provoca daño a sus rivales y así diseñar estrategias que busquen anular o dificultar esas sociedades.



(A) Red de pases Argentina vs. Francia (top 5 conexiones)

(B) Red de pases por pdg Argentina vs. Francia (top 5 conexiones)

En el partido de la eliminación con Francia se observa una concentración de juego sobre la banda derecha, con *Messi* como mayor generador de peligro.

Con la intención de resumir la actuación de los principales jugadores que repiten presencia en el Mundial 2022, se propone combinar el clásico mapa de calor de cada jugador con sus jugadas más peligrosas (usando la métrica de *pdg*).

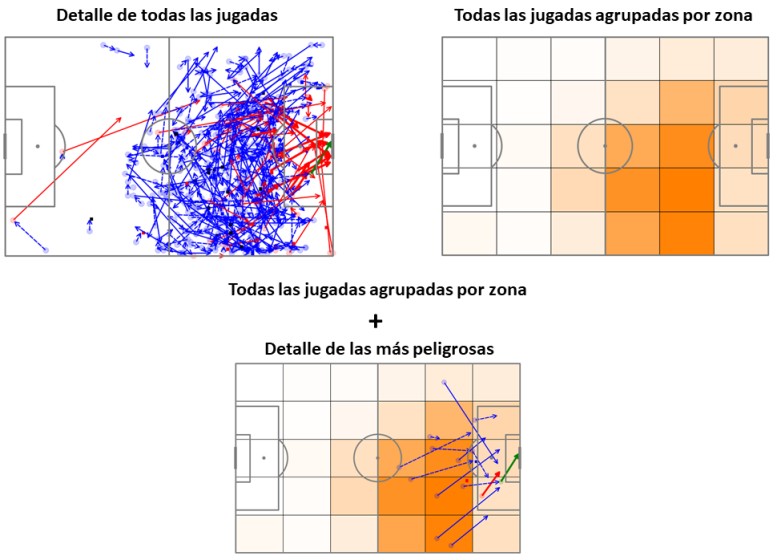
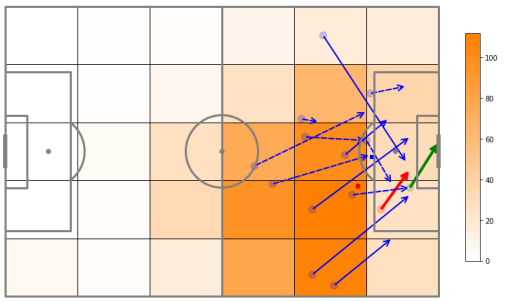
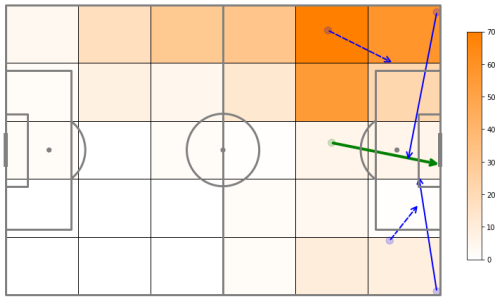


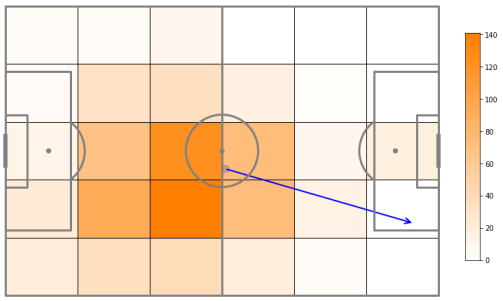
FIGURA 6.4: Nueva visualización propuesta que resume jugadas y aquellas de mayor peligrosidad[1]



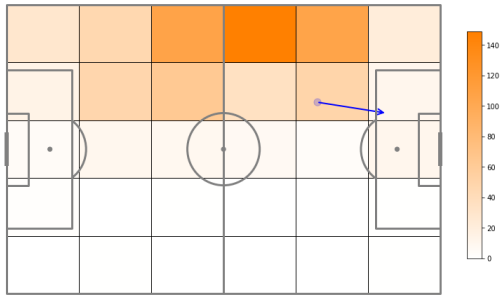
(A) Mapa de calor de **Lionel Messi** y jugadas con más peligro generado (umbral de 0.05)



(B) Mapa de calor de **Ángel Di María** y jugadas con más peligro generado (umbral de 0.05)



(A) Mapa de calor de **Nicolás Otamendi** y jugadas con más peligro generado (umbral de 0.05)



(B) Mapa de calor de **Nicolás Tagliafico** y jugadas con más peligro generado (umbral de 0.05)

A partir de las visualizaciones anteriores, se puede mencionar lo siguiente de cada uno de estos jugadores:

- *Lionel Messi* se movió principalmente por el sector derecho del ataque y fue quien más peligro generó
- *Ángel Di María* volcó su juego sobre la franja izquierda, bien adelantado
- *Nicolás Otamendi* se desempeñó como primer marcador central, involucrado activamente en el circuito de pases del equipo argentino
- *Nicolás Tagliafico* jugó principalmente de lateral izquierdo (en un partido lo hizo de segundo marcador central)

Adicionalmente, se presenta una tabla que llamaremos **Resumen pdg**, la cual muestra características generales sobre la peligrosidad del equipo bajo análisis:

Resumen pdg - Argentina - Mundial 2018

Métrica	Detalle
Conexiones de pases más peligrosas	Messi → Agüero (pdg 0.39)
Pasadores más peligrosos	Messi (pdg 1.06), Banega (pdg 0.68)
Receptores más peligrosos	Agüero (pdg 0.73), Messi (pdg 0.73)
Altura de pase más peligrosa	Ras del piso (54 % del peligro por esta vía)
Gambeteadores más peligrosos	Messi (pdg 0.55), Meza (pdg 0.16)
Conductores más peligrosos	Messi (pdg 0.58), Meza (pdg 0.46), Di María (pdg 0.29)

[1]Referencia de la visualización: línea punteada=*conducción*, línea sólida delgada=*pase*, línea sólida gruesa=*disparo*, rectángulo=*gambeta*, color azul=*resultado OK*, color rojo=*resultado fallido*, color verde=*gol*

Por último, volviendo a tomar como ejemplo el primer partido de Argentina-Islandia, la métrica de *Peligro de Gol* podría ser de utilidad para enriquecer transmisiones deportivas, dejando de lado la clásica medición de cantidad de pases y precisión, reemplazándola por la ponderación de las jugadas realmente influyentes.

FIGURA 6.7: Placa clásica de transmisiones deportivas vs. posibles variantes usando *Peligro de Gol* (Fuente fotos: TyC Sports y Clarín)

6.3. 22/11/2022 (Fase de Grupos): Arabia Saudita

6.3.1. Pre-partido

Con la mente puesta en el debut mundialista frente al equipo saudí, se procede a analizar la única información disponible que poseemos en el conjunto de datos: su participación en el Mundial 2018.

Con respecto al Mundial anterior, 9 jugadores repiten asistencia:

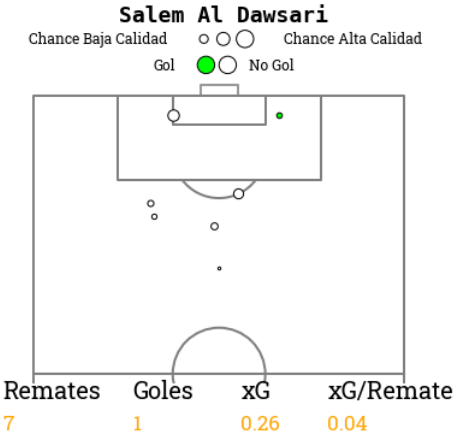
Jugador	Minutos Jugados	Partidos Jugados
Salem Al Dawsari	293	3
Salman Al Faraj	293	3
Yasir Al Shahrani	293	3
Mohammed Al Burayk	293	3
Abdullah Otayf	262	3
Hattan Babhir	169	3
Mohammed Al Owais	96	1
Ali Albulayhi	96	1
Mohammed Kanoo	20	1

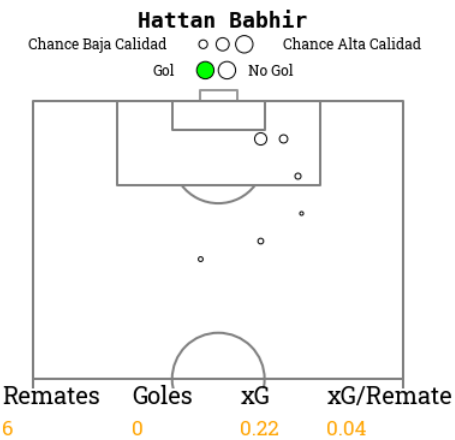
Su recorrido previo fue el siguiente:

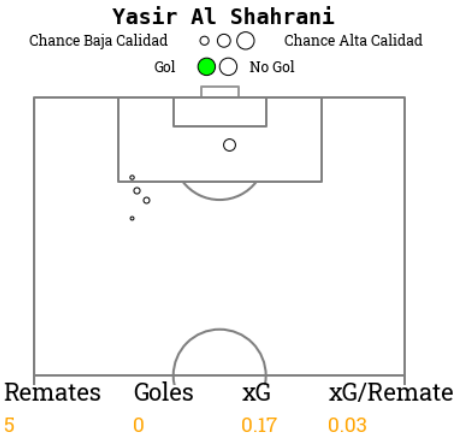
Ronda	Partido	Goles Arabia Saudita
Fase de Grupos	Arabia Saudita 0 - Rusia 5	-
Fase de Grupos	Arabia Saudita 0 - Uruguay 1	-
Fase de Grupos	Arabia Saudita 2 - Egipto 1	Salman Al Faraj (penal) - Salem Al Dawsari

A nivel formación, las predominantes fueron 4-1-4-1 y 4-2-3-1.

En cuanto a los remates, se grafican los mapas de disparos de los 3 jugadores que más remataron.



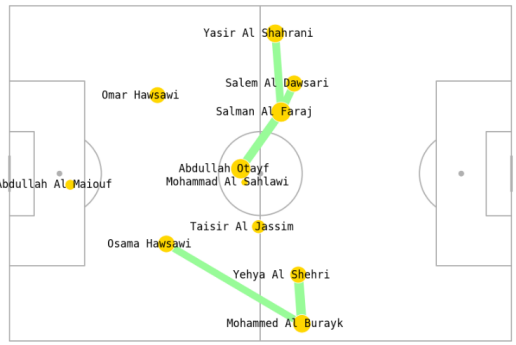




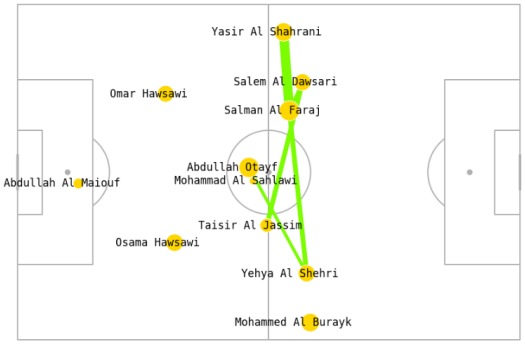
Se puede observar que:

- *Salem Al Dawsari*, volante por izquierda, remata principalmente de afuera y por izquierda o el centro (aunque su gol lo hace por derecha, su pierna hábil es la derecha)
- *Hattan Babhir*, wing derecho, remata principalmente por derecha
- *Yasir Al Shahrani*: lateral izquierdo, llega a posiciones de remate.

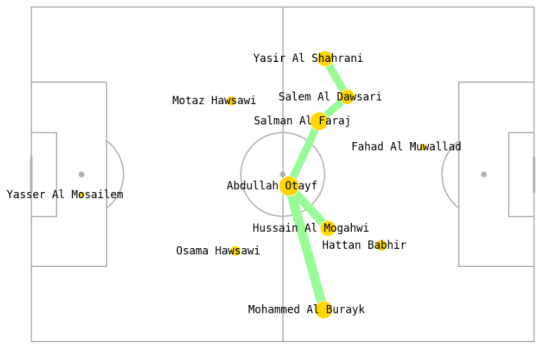
Se analizan las redes de pases de su primer y último encuentro:



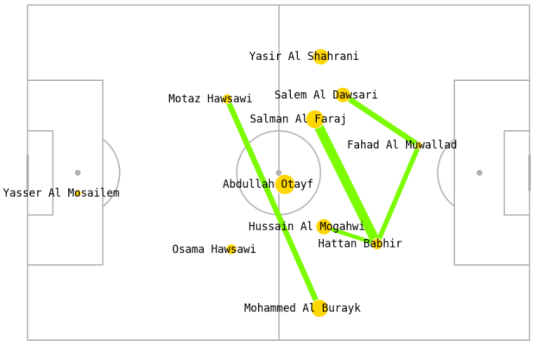
(A) Red de pases Arabia Saudita - Rusia (top 5 conexiones)



(B) Red de pases por pdg Arabia Saudita - Rusia (top 5 conexiones)



(A) Red de pases Arabia Saudita - Egipto (top 5 conexiones)



(B) Red de pases por pdg Arabia Saudita - Egipto (top 5 conexiones)

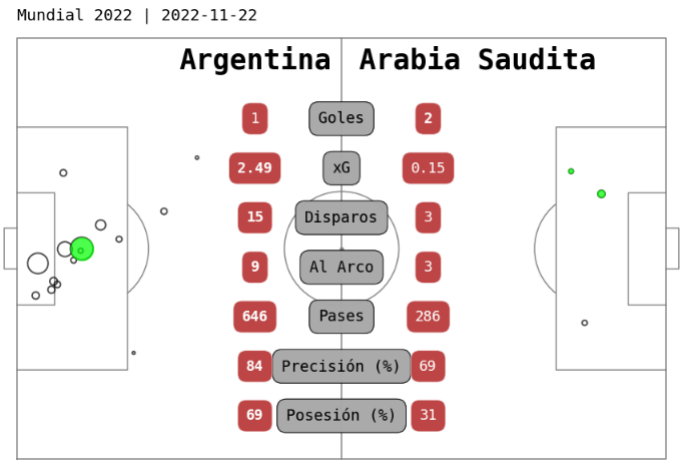
Se puede apreciar un equipo bastante retrasado en el primer partido, con laterales bien abiertos y participación activa en el juego (con Uruguay repite situación). Además, fuerte involucramiento en el juego de su eje central formado por *Salman Al Faraj* (tendencia a volcarse sobre parte central izquierda) y *Abdullah Otayf* (bien centralizado).

Resumen pdg - Arabia Saudita - Mundial 2018

Métrica	Detalle
Conexiones de pases más peligrosas	Salman Al Faraj → Salem Al Dawsari (pdg 0.13)
Pasadores más peligrosos	Salman Al Faraj (pdg 0.46, principalmente pases altos)
Receptores más peligrosos	Salem Al Dawsari (pdg 0.41)
Altura de pase más peligrosa	Ras del piso (54 % del peligro por esta vía)
Gambeteadores más peligrosos	Sin destacados
Conductores más peligrosos	Salem Al Dawsari (pdg 0.23), Salman Al Faraj (pdg 0.17)

6.3.2. Post-partido

El resumen del partido fue el siguiente:

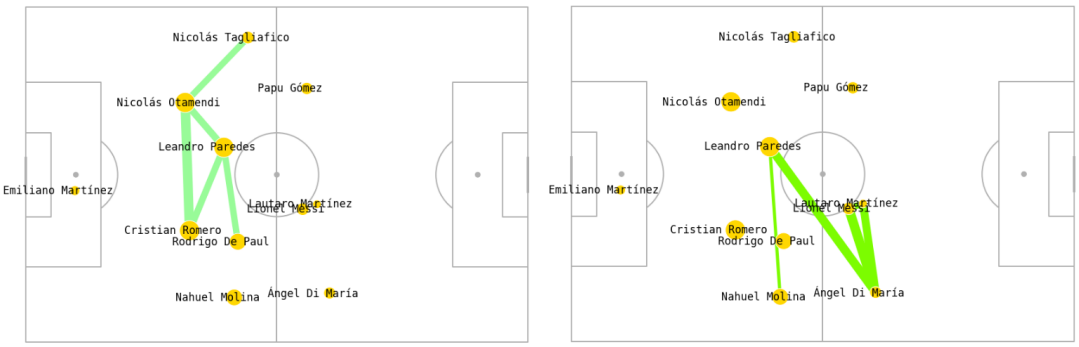


Goles

Equipo	Tiempo	Minuto	Detalle
Argentina	1	9	Messi (penal)
Arabia Saudita	2	47	Saleh Al Shehri (asistencia Firas Albirakan)
Arabia Saudita	2	52	Salem Al Dawsari

Un debut inesperado para un equipo que llevaba un invicto de 36 partidos. Sin hablar de merecimientos, los números reflejan un dominio argentino: 15 vs. 3 disparos, 69 % de posesión de la pelota y 2.49 vs. 0.15 en xG, lo que denota una mayor calidad de remates del conjunto argentino. En particular, el 0.15 xG de Arabia Saudita indica que sus goles llegaron desde situaciones de baja probabilidad de gol. El segundo gol llegó a través de *Salem Al Dawsari*, quien había sido identificado como uno de los hombres más peligrosos del equipo saudí.

A nivel formaciones, Arabia Saudita comenzó el partido con un 4-1-4-1 para luego pasar a una 4-4-1-1 y Argentina utilizó un 4-4-2.



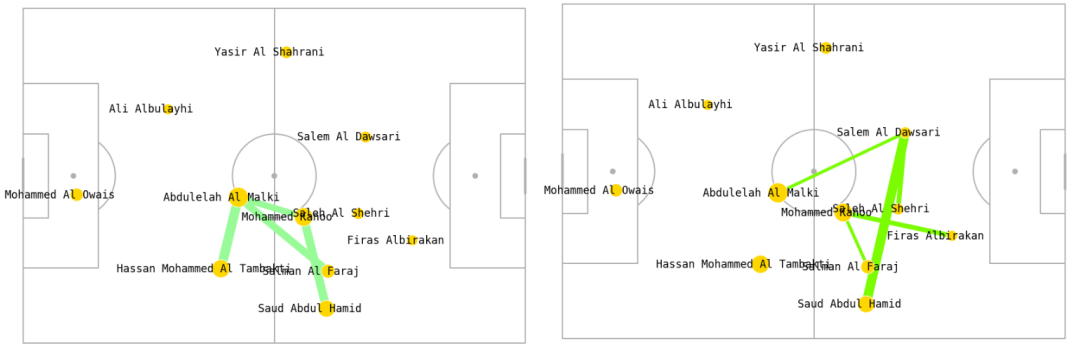
(A) Red de pases Argentina (top 5 conexiones)

(B) Red de pases por pdg Argentina (top 5 conexiones)

En cuanto a las redes de pases de Argentina, se observa un equipo bastante retrasado con una concentración de pases en zona defensiva y sobre sector izquierdo. Además, las interacciones de *Messi* y *Lautaro Martínez* parecen solaparse. Si analizamos la peligrosidad, se destaca *Di María* sobre el sector derecho de la cancha.

Resumen pdg - Argentina

Métrica	Detalle
Conexiones de pases más peligrosas	Messi → Lautaro Martínez (pdg 0.37), Di María → Messi (pdg 0.25)
Pasadores más peligrosos	Messi (pdg 0.55), Di María (pdg 0.40)
Receptores más peligrosos	Lautaro Martínez (pdg 0.47), Messi (pdg 0.42, sobre todo pases altos)
Altura de pase más peligrosa	Pases altos (47 % del peligro por esta vía)
Gambeteadores más peligrosos	Sin destacados
Conductores más peligrosos	Di María (pdg 0.31)



(A) Red de pases Arabia Saudita (top 5 conexiones)

(B) Red de pases por pdg Arabia Saudita (top 5 conexiones)

Por el lado de Arabia, se aprecia un equipo abierto, con concentración de juego en la zona derecha de la cancha. Si bien en ese sector se detecta peligrosidad, se destaca como la conexión más peligrosa la que va desde el lateral derecho hacia el puntero izquierdo (*Saud Abdul Hamid* → *Salem Al Dawsari*).

Resumen pdg - Arabia Saudita

Métrica	Detalle
Conexiones de pases más peligrosas	Saud Abdul Hamid → Salem Al Dawsari (pdg 0.06)
Pasadores más peligrosos	Saud Abdul Hamid (pdg 0.07, sobre todo pases altos)
Receptores más peligrosos	Salem Al Dawsari (pdg 0.12, sobre todo pases altos)
Altura de pase más peligrosa	Pases altos (69 % del peligro por esta vía)
Gambeteadores más peligrosos	Salem Al Dawsari (pdg 0.22)
Conductores más peligrosos	Saleh Al Shehri (pdg 0.13)

6.4. 26/11/2022 (Fase de Grupos): México

6.4.1. Pre-partido

En este caso, ya se dispone de un partido del torneo actual y también se analiza el comportamiento en la anterior edición de la Copa Mundial.

Son 8 los jugadores que repiten cita mundialista:

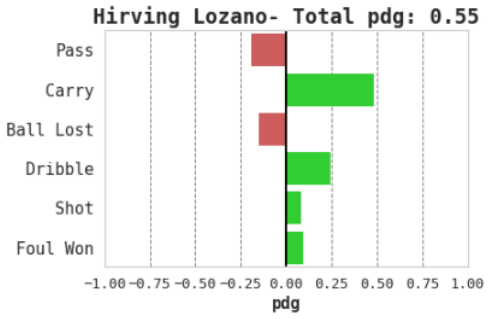
Jugador	Minutos Jugados	Partidos Jugados
Guillermo Ochoa	382	4
Héctor Herrera	382	4
Jesús Gallardo	352	4
Hirving Lozano	329	4
Andrés Guardado	313	4
Héctor Moreno	286	3
Edson Álvarez	282	4
Raúl Jiménez	65	2

Su recorrido previo fue:

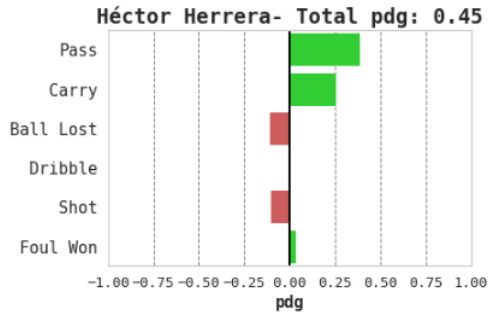
Ronda	Partido	Goles México
Fase de Grupos	México 1 - Alemania 0	Hirving Lozano
Fase de Grupos	México 2 - Corea del Sur 1	Carlos Vela (penal) - Javier Hernández (asistencia Hirving Lozano)
Fase de Grupos	México 0 - Suecia 3	-
8vos de Final	México 0 - Brasil 2	-

Su formación predominante fue 4-2-3-1.

Sus dos jugadores más destacados según *pdg* fueron:



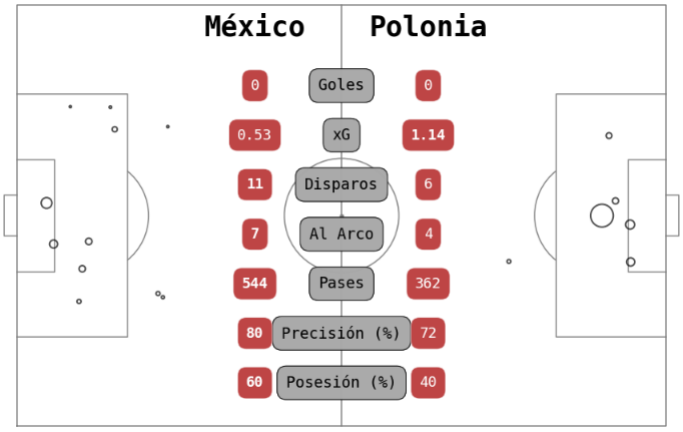
(A) pdg por tipo de acción para Hirving Lozano en Mundial 2018

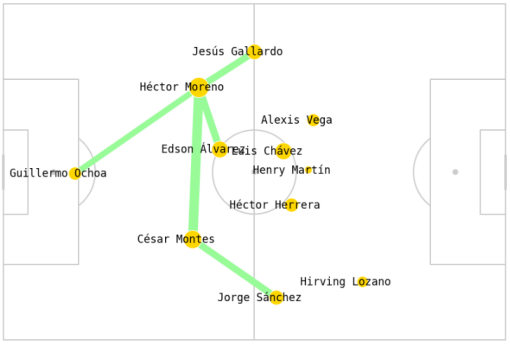


(B) pdg por tipo de acción para Héctor Herrera en Mundial 2018)

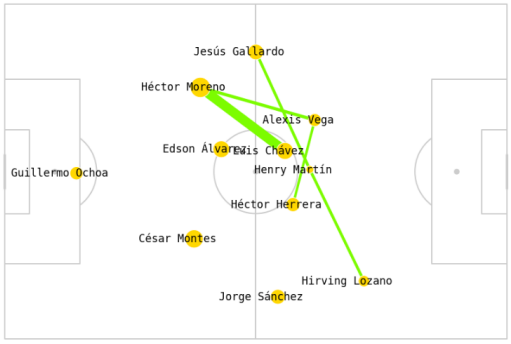
Es posible notar que *Lozano* se caracteriza por conducir y gambetear mientras que *Herrera* genera peligrosidad principalmente con sus pases y conducciones.

En cuanto al Mundial 2022, su debut fue un 0-0 contra Polonia donde utilizó una formación 4-3-3.





(A) Red de pases México - Polonia (top 5 conexiones)



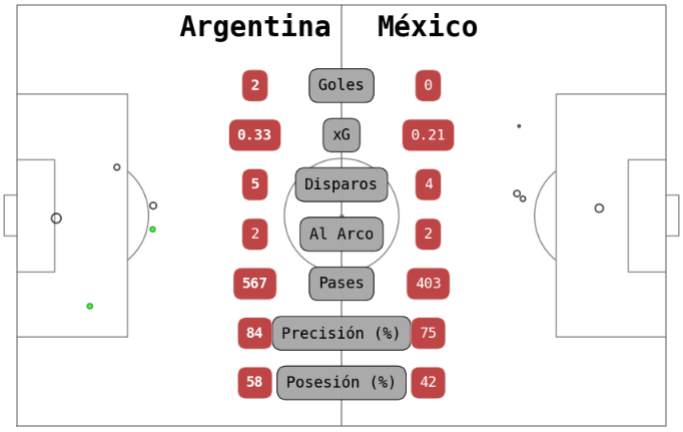
(B) Red de pases por pdg México - Polonia (top 5 conexiones)

Las redes muestran mucha circulación de pelota entre la última línea y el arquero, al central *Héctor Moreno* interactuando peligrosamente con *Chávez* y *Vega* y lo mismo entre *Henry Martín* con *Hirving Lozano* y *Jesús Gallardo*. Como curiosidad, las acciones de peligro entre *Moreno* y los otros dos jugadores lo tienen a *Moreno* como receptor de centros. Adicionalmente, quien más remató al arco durante el partido fue *Alexis Vega* con 5 remates (sobre 11 del equipo).

Resumen pdg - México - Mundial 2022

Métrica	Detalle
Conexiones de pases más peligrosas	Luis Chávez → Héctor Moreno (pdg 0.27)
Pasadores más peligrosos	Luis Chávez (pdg 0.30, pases altos), Hirving Lozano (pdg 0.15, a ras del piso)
Receptores más peligrosos	Héctor Moreno (pdg 0.32, pases alto), Henry Martín (pdg 0.19, a ras del piso)
Altura de pase más peligrosa	Pases altos (57 % del peligro por esta vía)
Gambeteadores más peligrosos	Luis Chávez (pdg 0.04)
Conductores más peligrosos	Hirving Lozano (0.10 pdg)

6.4.2. Post-partido

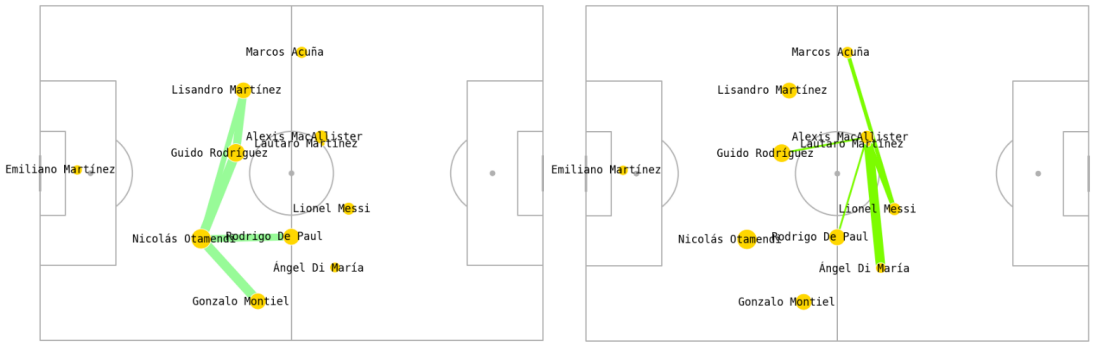


Goles

Equipo	Tiempo	Minuto	Detalle
Argentina	2	63	Lionel Messi (asistencia Di María)
Argentina	2	86	Enzo Fernández (asistencia Messi)

Los xG de cada equipo reflejan que las chances generadas fueron de baja calidad (ambos goles argentinos llegaron desde posiciones de muy difícil ejecución) y quien más remató del equipo visitante fue *Alexis Vega* (2 de 4 remates), repitiendo situación vs. Polonia. Sin dudas, un partido muy parejo desde los números, cargado de tensión hasta que *Messi* logró abrir el marcador.

México usó la formación 3-5-2, mientras que Argentina, 4-4-2.



(A) Red de pases Argentina (top 5 conexiones) (B) Red de pases por pdg Argentina (top 5 conexiones)

En las redes de pases se puede observar un planteo similar al de Arabia (con algunos cambios de nombres), con *Otamendi* de primer marcador central y activo en el circuito de pases (como ya se lo observó en partidos del Mundial pasado). Vuelven a aparecer *Di María* (por derecha) y *Messi* con peligrosidad y emerge *Alexis MacAllister*, quien había reemplazado a *Papu Gómez* con respecto al debut. Un dato interesante de *MacAllister* es que se vuelve una frecuente descarga de pases para *Messi* (junto con *De Paul* reciben casi el 50 % de los pases de la estrella argentina).

Resumen pdg - Argentina

Métrica	Detalle
Conexiones de pases más peligrosas	Sin destacados
Pasadores más peligrosos	Di María (pdg 0.11)
Receptores más peligrosos	Sin destacados
Altura de pase más peligrosa	Ras del piso (69 % del peligro por esta vía)
Gambeteadores más peligrosos	Sin destacados
Conductores más peligrosos	Di María (pdg 0.11)

No hay mucho para destacar de México, a excepción de que la única peligrosidad que generaron fue a través de pelotas altas (pdg 0.14).

6.5. 30/11/2022 (Fase de Grupos): Polonia

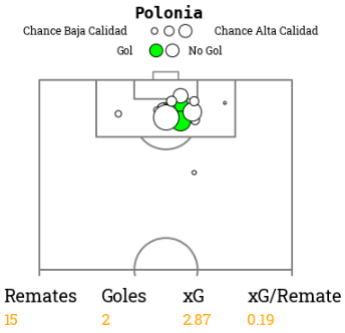
6.5.1. Pre-partido

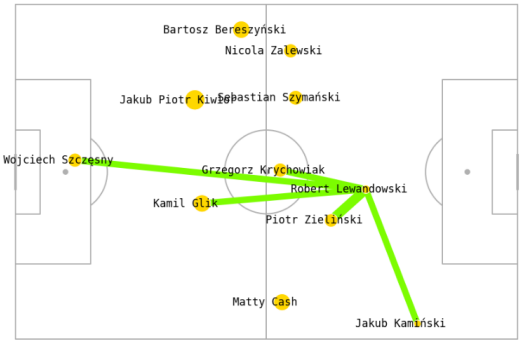
Su recorrido en este Mundial fue el siguiente:

Ronda	Partido	Goles Polonia
Fase de Grupos	Polonia 0 - México 0	-
Fase de Grupos	Polonia 2 - Arabia Saudita 0	Piotr Zieliński (asistencia Lewandowski) - Robert Lewandowski

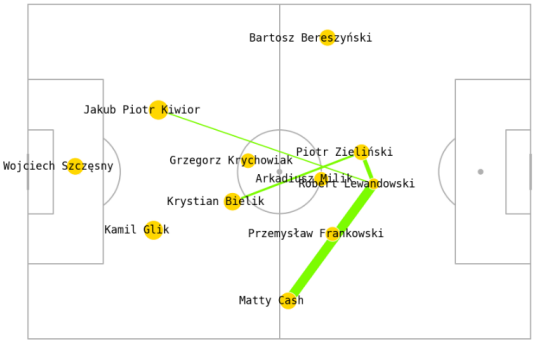
Como formaciones, utilizó 4-1-4-1, 4-4-1-1 y 4-4-2.

Polonia concentra sus tiros en posiciones cercanas al arco (xG alto, cercano a goles convertidos) y se destaca *Lewandowski* con casi el 50 % de los disparos del equipo.





(A) Red de pases por pdg Polonia - México (top 5 conexiones)



(B) Red de pases por pdg Polonia - Arabia Saudita (top 5 conexiones)

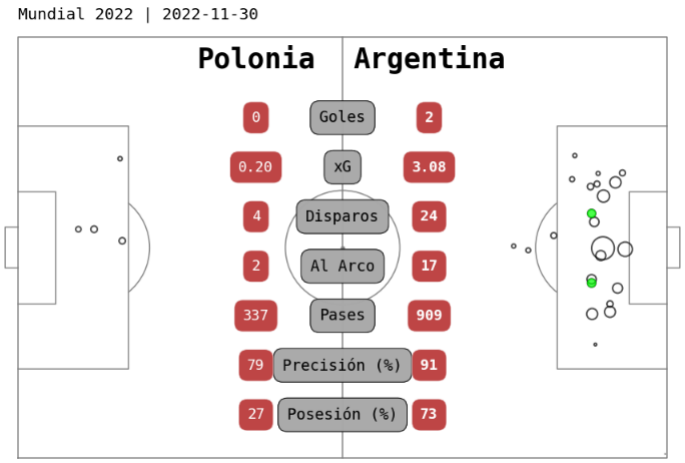
En cuanto a sus redes de pases por *pdg*, en el primer partido se puede observar a *Lewandowski* siendo el centro de la peligrosidad (se aprecian incluso pelotazos largos

desde el arquero hacia él). En el segundo partido, también se destaca el centrodelantero pero se observa peligrosidad por la zona derecha, sobre todo del lateral *Matty Cash*.

Resumen pdg - Polonia - Mundial 2022

Métrica	Detalle
Conexiones de pases más peligrosas	Matty Cash → Lewandowski (pdg 0.37), Kamiński → Lewandowski (pdg 0.30)
Pasadores más peligrosos	Cash (pdg 0.41, sobre todo a ras del piso), Kamiński (pdg 0.38, mixto)
Receptores más peligrosos	Lewandowski (pdg 0.96, por lejos destacado)
Altura de pase más peligrosa	Mixto
Gambeteadores más peligrosos	Lewandowski (pdg 0.19)
Conductores más peligrosos	Sin destacados

6.5.2. Post-partido

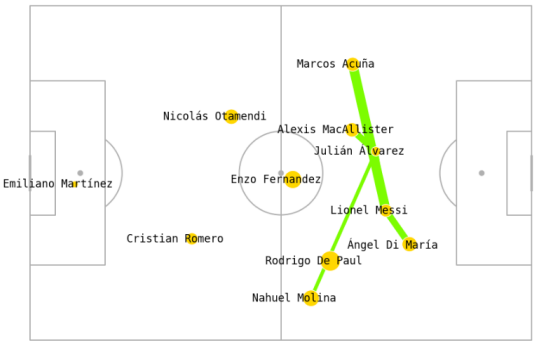


Goles

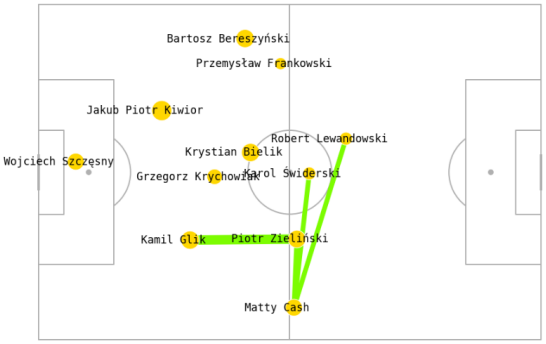
Equipo	Tiempo	Minuto	Detalle
Argentina	2	45	Alexis MacAllister (asistencia Nahuel Molina)
Argentina	2	66	Julián Álvarez (asistencia Enzo Fernández)

Argentina fue un claro dominador del encuentro, con un 73 % de posesión y superando ampliamente en cantidad y calidad de tiros al rival. La peligrosidad de *Lewandowski* fue neutralizada: el centrodelantero polaco ni siquiera pudo rematar una vez al arco contrario. A su vez, el índice de centralidad de Argentina tocó un valor altísimo (**0.17**), con *De Paul* realizando 140 pases, casi la mitad de todo el equipo polaco y seguido recién por *Otamendi* con 89.

Para este encuentro, Polonia utilizó la formación 4-4-2 y Argentina, 4-3-3.



(A) Red de pases por pdg Argentina (top 5 conexiones)



(B) Red de pases por pdg Polonia (top 5 conexiones)

La red de pases de Argentina muestra un equipo parado muchos metros más adelante que los partidos anteriores. *Enzo Fernández* ingresa y se ubica casi en el círculo central, lejos de meterse entre ambos defensores centrales argentinos. También se aprecia una mayor peligrosidad de ambos laterales, con acciones en posiciones altas y el aporte ofensivo de *Julián Álvarez*, asociándose principalmente con *Messi* y *MacAllister*. Nuevamente, *Di María* aparece por el sector derecho de la cancha. En el caso de Polonia, se observa un equipo muy retrasado y repite su mayor peligrosidad por la zona derecha con *Cash* como protagonista (sin mucho más por resaltar).

Resumen pdg - Argentina

Métrica	Detalle
Conexiones de pases más peligrosas	Messi → Julián Álvarez (pdg 0.26)
Pasadores más peligrosos	Messi (pdg 0.54)
Receptores más peligrosos	Julián Álvarez (pdg 0.50)
Altura de pase más peligrosa	Ras del piso (81 % del peligro por esta vía)
Gambeteadores más peligrosos	Di María (pdg 0.17), Messi (pdg 0.14)
Conductores más peligrosos	Messi (pdg 0.25), MacAllister (pdg 0.14), Enzo Fernández (pdg 0.11)

6.6. 03/12/2022 (8vos de Final): Australia

6.6.1. Pre-partido

Su recorrido en este torneo fue el siguiente:

Ronda	Partido	Goles Australia
Fase de Grupos	Australia 1 - Francia 4	Craig Goodwin (asistencia Mathew Leckie)
Fase de Grupos	Australia 1 - Túnez 0	Mitchell Duke (de cabeza)
Fase de Grupos	Australia 1 - Dinamarca 0	Mathew Leckie (asistencia Riley McGree)

Sus formaciones más frecuentes son 4-4-2 y 4-1-4-1.

En cuanto a remates, un tercio del total del equipo (7 de 21) fueron ejecutados por el centrodelantero *Mitchell Duke* (5 de ellos de cabeza, mide 1.86m).

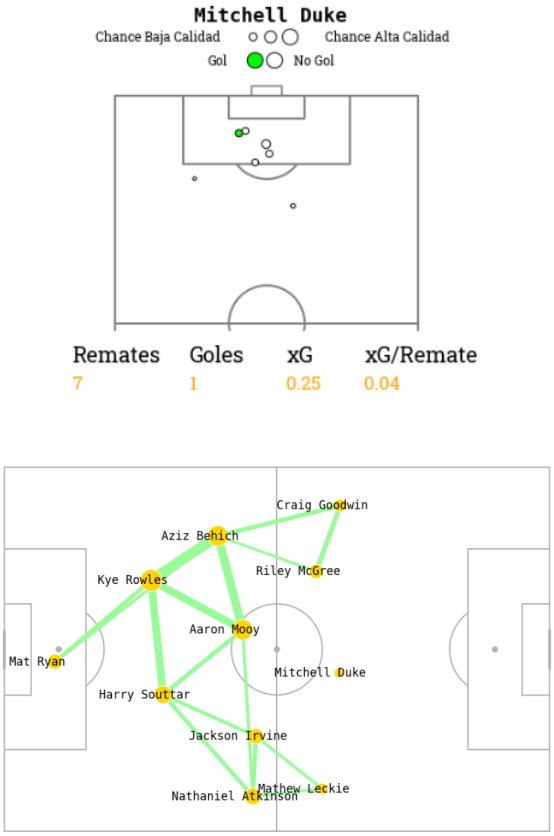


FIGURA 6.18: Red de pases de Australia - Francia (completa)

Tomando como ejemplo el partido con Francia (en los otros se ven comportamientos similares), la red de pases muestra un equipo muy largo y ancho, con centrales bien retrasados, laterales bien abiertos, dos volantes centrales a la altura de los laterales, dos volantes por fuera y dos delanteros por el centro. En sus tres partidos, su xG es bajo (0.36 vs. Francia, 0.36 vs Túnez y 0.47 vs. Dinamarca), demostrando que pudo convertir a pesar de tener chances de baja calidad.

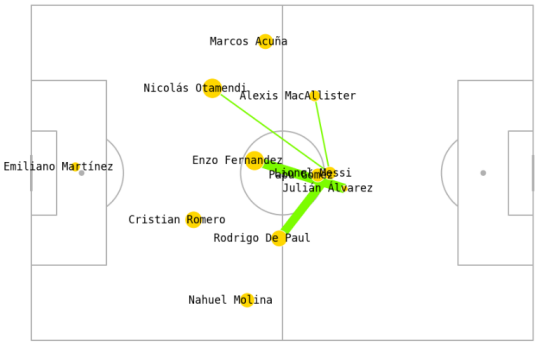
Resumen pdg - Australia - Mundial 2022

Métrica	**Detalle**
Conexiones de pases más peligrosas	Karačić → Duke (pdg 0.13), Leckie → Goodwin (pdg 0.12)
Pasadores más peligrosos	Riley McGree (pdg 0.17), Mathew Leckie (pdg 0.16)
Receptores más peligrosos	Duke (pdg 0.40, mayoría pases altos), Goodwin (pdg 0.18, mayoría pases bajos)
Altura de pase más peligrosa	Pases altos (69 % del peligro por esta vía)
Gambeteadores más peligrosos	Aziz Behich (pdg 0.20)
Conductores más peligrosos	Aziz Behich (pdg 0.25), Mathew Leckie (pdg 0.25)

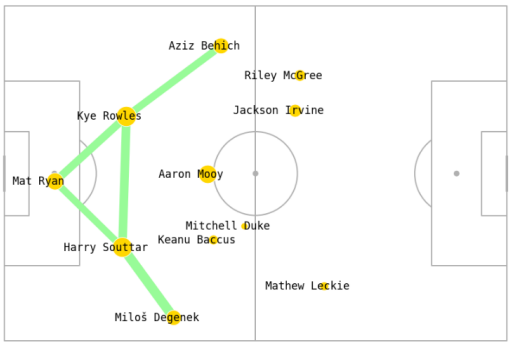
6.6.2. Post-partido

Goles

Equipo	Tiempo	Minuto	Detalle
Argentina	1	34	Lionel Messi (asistencia Nicolás Otamendi)
Argentina	2	56	Julián Álvarez
Australia	2	77	Enzo Fernández (EC)



(A) Red de pases por pdg Argentina (top 5 conexiones)



(B) Red de pases Australia (top 5 conexiones)

En la red de pases por *pdg* de Argentina se puede destacar el solapamiento en interacciones de *Papu Gómez* y *Messi* (cerca también *Julián Álvarez*), además de las conexiones peligrosas entre *Enzo Fernández-Julián Álvarez* y *De Paul-Messi*. Por otro lado, se verifica que Australia nuevamente mostró un equipo abierto con mucha interacción entre su última línea y el arquero, lo que probablemente haya favorecido la presión por parte de Argentina que derivó en el segundo gol. Adicionalmente, el equipo argentino volvió a contener al centrodelantero rival (*Duke* en este caso), quien tampoco logró rematar al arco en todo el encuentro.

Australia eligió la formación 4-4-2 y Argentina se inclinó por la 4-3-3.

Resumen pdg - Argentina

Métrica	Detalle
Conexiones de pases más peligrosas	Messi → Lautaro Martínez (pdg 0.29), De Paul → Messi (pdg 0.16)
Pasadores más peligrosos	Rodrigo De Paul (pdg 0.37)
Receptores más peligrosos	Lautaro Martínez (pdg 0.50), Lionel Messi (pdg 0.34)
Altura de pase más peligrosa	Ras del piso (85 % del peligro por esta vía)
Gambeteadores más peligrosos	Lionel Messi (pdg 0.06)
Conductores más peligrosos	Lionel Messi (pdg 0.26), Lautaro Martínez (pdg 0.24)

Resumen pdg - Australia

Métrica	Detalle
Conexiones de pases más peligrosas	Sin destacados
Pasadores más peligrosos	Goodwin (pdg 0.13, pases altos), Karačić (pdg 0.09, mayoría pases altos)
Receptores más peligrosos	Mathew Leckie (pdg 0.11)
Altura de pase más peligrosa	Pases altos (88 % del peligro por esta vía)
Gambeteadores más peligrosos	Aziz Behich (pdg 0.09)
Conductores más peligrosos	Aziz Behich (pdg 0.26)

6.7. 09/12/2022 (4tos de Final): Países Bajos

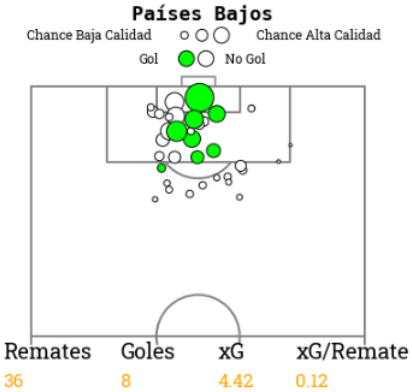
6.7.1. Pre-partido

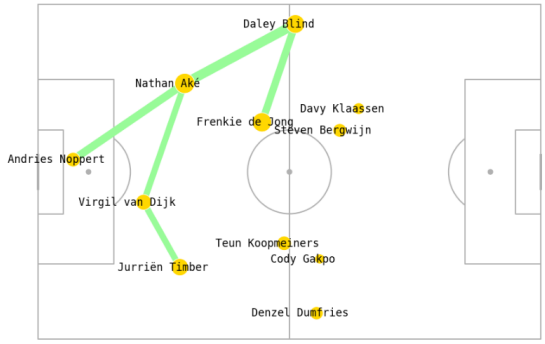
Su recorrido en este Mundial fue:

Ronda	Partido	Goles Países Bajos
Fase de Grupos	Países Bajos 2 - Senegal 0	Cody Gakpo (cabeza, asistencia Frenkie de Jong) - Davy Klaassen
Fase de Grupos	Países Bajos 1 - Ecuador 1	Cody Gakpo (asistencia Davy Klaassen)
Fase de Grupos	Países Bajos 2 - Qatar 0	Cody Gakpo (asistencia Davy Klaassen) - Frenkie de Jong
8vos de Final	Países Bajos 3 - EEUU 1	Depay (asist. Dumfries) - Blind (asist. Dumfries) - Dumfries (asist. Blind)

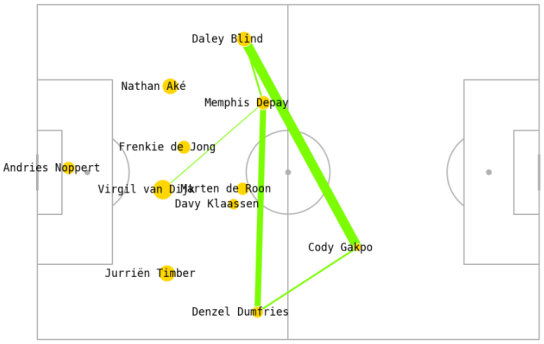
Su formación siempre fue 3-4-1-2.

En cuanto a sus remates, se puede contemplar que tienen predilección por remates centrales.



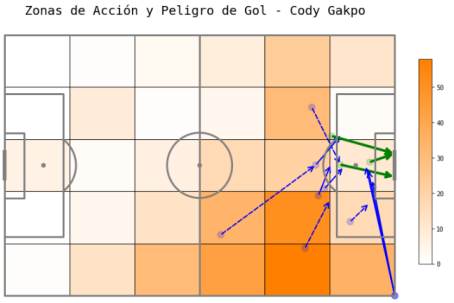


(A) Red de pases Países Bajos - Ecuador (top 5 conexiones)

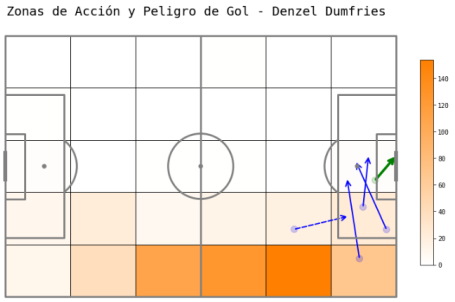


(B) Red de pases por pdg Países Bajos - EEUU (top 5 conexiones)

Las redes de pases muestran que, a excepción del partido con el débil Qatar, Países Bajos es un equipo largo que se para bastante retrasado. Tiene mucha circulación de pelota en la última línea (sobre todo por zona izquierda con *Nathan Aké*) y juega con *Blind* y *Dunfries* bien abiertos, quienes generaron mucho peligro en el partido de 8vos de Final. *Gakpo* es quien aparece más adelantado por derecha.



(A) Mapa de calor de **Cody Gakpo** y jugadas con más peligro generado (umbral de 0.05)



(B) Mapa de calor de **Denzel Dumfries** y jugadas con más peligro generado (umbral de 0.05)

Si analizamos las zonas y peligrosidad de *Gakpo* (su goleador), se puede apreciar que se mueve principalmente por derecha pero tiende a tirarse hacia el centro, desde donde convirtió sus tres goles. Por otro lado, uno de los destacados del partido de 8vos fue *Dumfries*, quien se mueve bien abierto por derecha y genera peligro desde ese sector.

Resumen pdg - Países Bajos - Mundial 2022

Métrica	Detalle
Conexiones de pases más peligrosas	de Jong → Gakpo (pdg 0.38), Depay → Bergwijn (pdg 0.27)
Pasadores más peligrosos	Dumfries (pdg 0.50), Blind (pdg 0.48), de Jong (pdg 0.44), Gakpo (pdg 0.43)
Receptores más peligrosos	Gakpo (pdg 0.70), Depay (pdg 0.60), Bergwijn (pdg 0.58)
Altura de pase más peligrosa	Ras del piso (60 % del peligro por esta vía)
Gambeteadores más peligrosos	Sin destacados
Conductores más peligrosos	Depay (pdg 0.29), Gakpo (pdg 0.24)

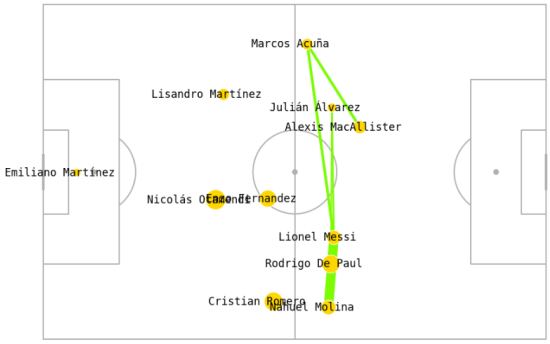
6.7.2. Post-partido

Goles

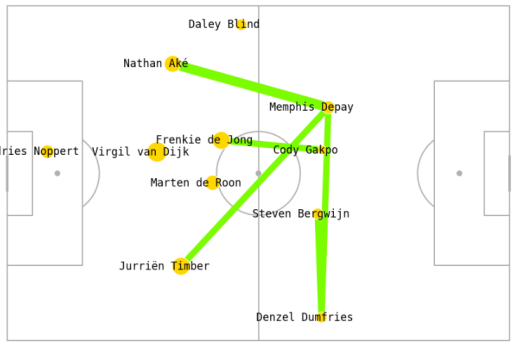
Equipo	Tiempo	Minuto	Detalle
Argentina	1	34	Nahuel Molina (asistencia Lionel Messi)
Argentina	2	72	Lionel Messi (penal)
Países Bajos	2	82	Wout Weghorst (asistencia Steven Berghuis)
Países Bajos	2	100	Wout Weghorst (asistencia Teun Koopmeiners)

Luego del empate en tiempo reglamentario, se definió por penales (4-3 en favor de Argentina).

Países Bajos usó su formación habitual de 3-4-1-2 y Argentina utilizó un inusual 3-5-2, probablemente para emparejar en formacion al rival.



(A) Red de pases por pdg Argentina (top 5 conexiones)

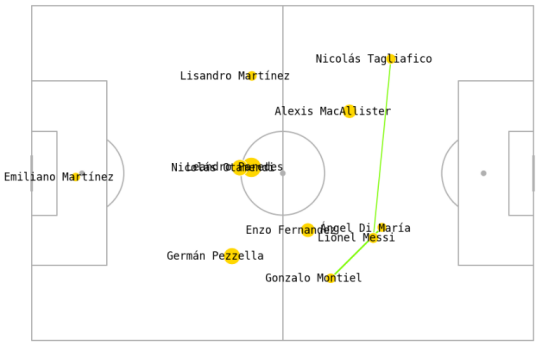


(B) Red de pases por pdg Países Bajos (top 5 conexiones)

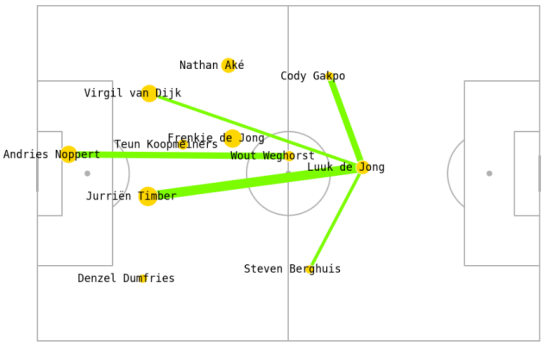
Las redes de pase muestran a Argentina con los laterales bien abiertos y generando peligro, sobre todo por la derecha con el tándem *Messi-Molina-De Paul*. Por su parte, Países Bajos exhibe nuevamente un equipo ancho, con peligrosidad por derecha con *Dumfries* y con *Aké* desde el fondo. En esta ocasión se ve poca participación del lateral izquierdo neerlandés, probablemente ocupado (y preocupado) por la ofensiva argentina.

El empate agónico de Países Bajos forzó el suplementario (nuevamente sus goles llegaron por zonas centrales como se había destacado previamente) y allí se observa un comportamiento *particular*. Del lado del conjunto argentino, no parece haberse sentido el impacto anímico ya que las redes de pases (limitadas al tiempo suplementario) exhiben un equipo bien adelantado buscando definir el partido, sobre todo

del lado de *Tagliafico*. A su vez, Países Bajos se muestra retrasado jugando a pelotas largas a sus dos delanteros centros (*Weghorst* y *Luuk de Jong*).



(A) Red de pases por pdg Argentina (tiempo suplementario - top 5 conexiones)



(B) Red de pases por pdg Países Bajos (tiempo suplementario - top 5 conexiones)

Similar comportamiento se puede ver a nivel pases y remates, superando *ampliamente* en números a su rival.

Período	PBAJ-Pases	ARG-Pases	PBAJ-Remates	ARG-Remates	PBAJ-xG	ARG-xG
1	313	238	1	5	0.09	0.58
2	299	181	5	2	0.43	0.86
3	57	**107**	0	0	0	0
4	22	**114**	1	**7**	0.05	**0.50**

Sin dudas, una gran exhibición de personalidad y caracter del equipo para sobreponerse a situaciones adversas como la sucedida al final del partido.

Resumen pdg - Argentina

Métrica	Detalle
Conexiones de pases más peligrosas	Messi → Molina (pdg 0.20), Enzo Fernández → Lautaro Martínez (pdg 0.18)
Pasadores más peligrosos	Messi (pdg 0.30), Enzo Fernández (pdg 0.18)
Receptores más peligrosos	Lautaro Martínez (pdg 0.24), Molina (pdg 0.21)
Altura de pase más peligrosa	Ras del piso (84 % del peligro por esta vía)
Gambeteadores más peligrosos	Enzo Fernández (pdg 0.03), Messi (pdg 0.02)
Conductores más peligrosos	Molina (pdg 0.43), MacAllister (pdg 0.12)

Resumen pdg - Países Bajos

Métrica	Detalle
Conexiones de pases más peligrosas	Sin destacadas
Pasadores más peligrosos	Steven Berghuis (pdg 0.14, pases altos), Cody Gakpo (pdg 0.12, pases altos)
Receptores más peligrosos	Weghorst (pdg 0.21, mayoría pases altos), Luuk de Jong (pdg 0.20, mayoría pases altos)
Altura de pase más peligrosa	Pases altos (60 % del peligro por esta vía)
Gambeteadores más peligrosos	Sin destacados
Conductores más peligrosos	Wout Weghorst (pdg 0.12), Frenkie de Jong (pdg 0.09)

6.8. 13/12/2022 (Semifinal): Croacia

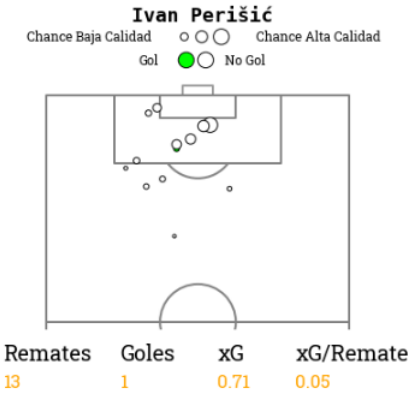
6.8.1. Pre-partido

Su recorrido previo en este Mundial fue el siguiente:

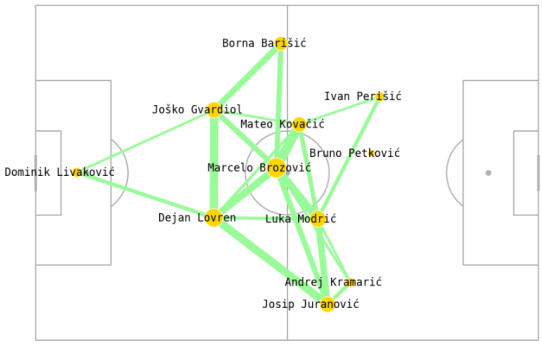
Ronda	Partido	Goles Croacia
Fase de Grupos	Croacia 0 - Marruecos 0	-
Fase de Grupos	Croacia 4 - Canadá 1	Kramarić (asist. Perišić) x 2 - Livaja (asist. Juranović) - Majer (asist. Oršić)
Fase de Grupos	Croacia 0 - Bélgica 0	-
8vos de Final	Croacia 1 - Japón 1	Ivan Perišić (de cabeza, asistencia Dejan Lovren)
4tos de Final	Croacia 1 - Brasil 1	Bruno Petković (asistencia Mislav Oršić)

Su formación más frecuente fue 4-3-3.

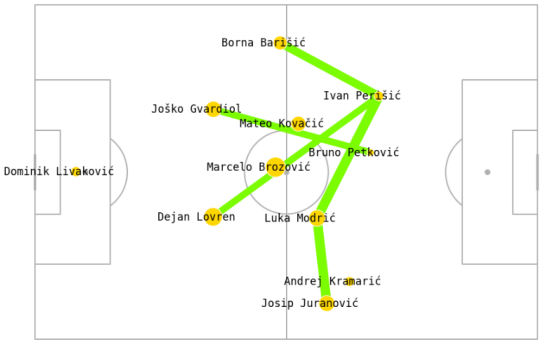
En cuanto a remates, quien aparece en primer lugar es el punta izquierdo *Ivan Perišić* con 13 disparos, desde zona izquierda y central, de afuera y dentro del área.



Muestra partidos con índice de centralidad alto: 0,16 vs. Marruecos (con *Gvardiol* realizando 102 pases), 0,21 vs. Japón (con *Brozović* ejecutando 116 pases) y 0,15 vs. Brasil (con 105 pases de *Modrić*).



(A) Red de pases Croacia - Japón (completa)



(B) Red de pases por pdg Croacia - Japón (top 5 conexiones)

El equipo croata exhibe un mediocampo de mucha circulación con *Brozović* posicionado en el círculo central, *Modrić* a derecha y *Kovačić* a izquierda. Despliega a sus laterales bien abiertos y cuenta con centrales activos en el circuito de pases y generando peligro. Como más peligroso, asoma *Perišić* posicionado en la franja izquierda.

Como dato interesante, un jugador que aparece dando asistencias pero no figura en las redes de pases (dado que es habitualmente suplente) es *Mislav Oršić*. Jugó un total de 44 minutos en el Mundial, brindó 2 asistencias, juega por franja izquierda y acumula un *pdg* de 0.43 (siendo el tercero del equipo a pesar de su escasa participación).

Resumen pdg - Croacia - Mundial 2022

Métrica	Detalle
Conexiones de pases más peligrosas	Kovačić → Livaja (pdg 0.40), Perišić → Kramarić (pdg 0.34)
Pasadores más peligrosos	Modrić (pdg 0.87), Kovačić (pdg 0.71)
Receptores más peligrosos	Perišić (pdg 0.95), Kramarić (pdg 0.77)
Altura de pase más peligrosa	Ras del piso (64 % del peligro por esta vía)
Gambeteadores más peligrosos	Brozović (pdg 0.05)
Conductores más peligrosos	Kovačić (pdg 0.48), Perišić (pdg 0.29)

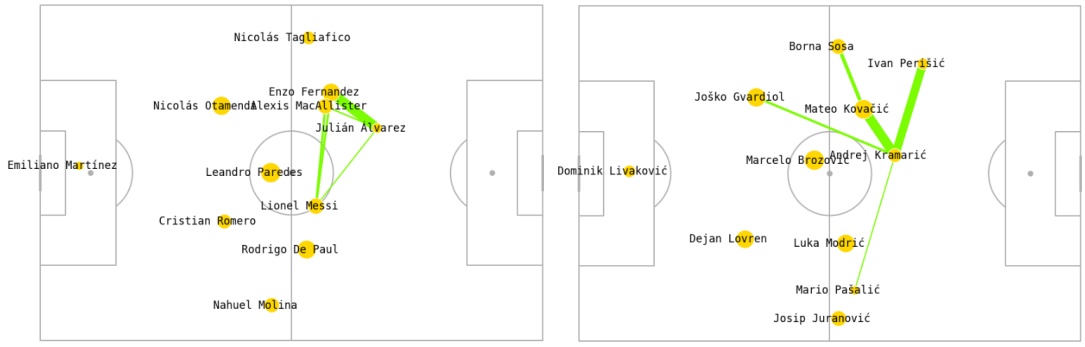
6.8.2. Post-partido

Goles

Equipo	Tiempo	Minuto	Detalle
Argentina	1	33	Lionel Messi (penal)
Argentina	1	38	Julián Álvarez
Australia	2	68	Julián Álvarez (asistencia Lionel Messi)

Croacia formó con su habitual 4-3-3, mientras que la Argentina optó por 4-4-2, con un medio menos ofensivo que en los últimos partidos (*Paredes* en el centro, *De Paul* por derecha y *Enzo-Mac Allister* por izquierda), probablemente con el objetivo de contrarrestar al fuerte mediocampo croata.

El partido se desarrolló favorablemente para Argentina, quien concentró las llegadas de mayor calidad. A pesar de esto, fue la primera vez en el torneo que Argentina se vio ampliamente superado en posesión (40 % vs. 60 %, siempre había estado arriba a excepción de Países Bajos con 48 % vs 52 %) y en cantidad de pases (432 vs. 648, la marca más baja del torneo para Argentina).



(A) Red de pases por pdg Argentina (top 5 conexiones) (B) Red de pases por pdg Croacia (top 5 conexiones)

En las redes de pase se puede contemplar la peligrosidad argentina concentrada en la zona izquierda del ataque (*Enzo-MacAllister-Julián Álvarez*) sumado a *Messi*, sector donde defiende *Modrić*. Por el lado croata, vuelve a utilizar laterales bien abiertos y peligrosidad por la zona de *Perišić*.

Resumen pdg - Argentina

Métrica	Detalle
Conexiones de pases más peligrosas	Messi → Julián Álvarez (pdg 0.29), Enzo Fernández → Julián Álvarez (pdg 0.25)
Pasadores más peligrosos	Enzo Fernández (pdg 0.35), Messi (pdg 0.34)
Receptores más peligrosos	Julián Álvarez (pdg 0.59)
Altura de pase más peligrosa	Ras del piso (57 % del peligro por esta vía)
Gambeteadores más peligrosos	Messi (pdg 0.06)
Conductores más peligrosos	Julián Álvarez (pdg 0.42)

Resumen pdg - Croacia

Métrica	Detalle
Conexiones de pases más peligrosas	Modrić → Lovren (pdg 0.17, pases altos), Kovačić → Kramarić (pdg 0.06)
Pasadores más peligrosos	Kovačić (pdg 0.17), Modrić (pdg 0.16)
Receptores más peligrosos	Kramarić (pdg 0.17)
Altura de pase más peligrosa	Ras del piso (67 % del peligro por esta vía)
Gambeteadores más peligrosos	Sin destacados
Conductores más peligrosos	Kovačić (pdg 0.09)

6.9. 18/12/2022 (Final): Francia

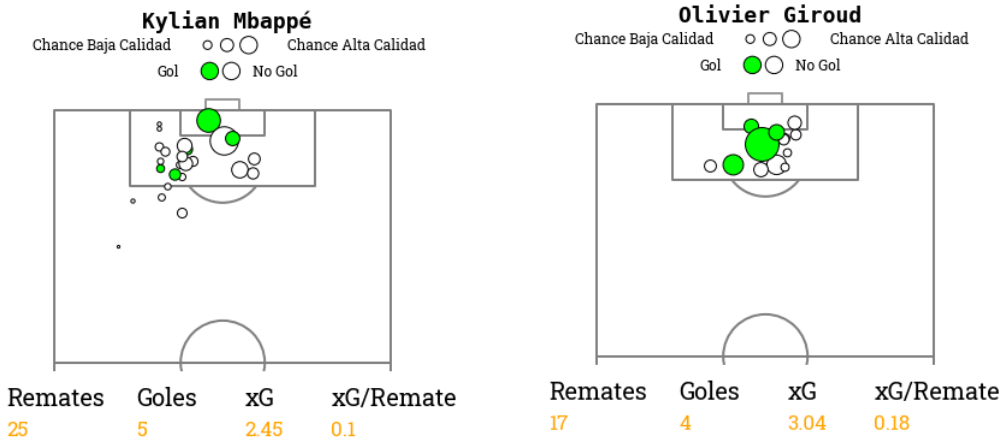
6.9.1. Pre-partido

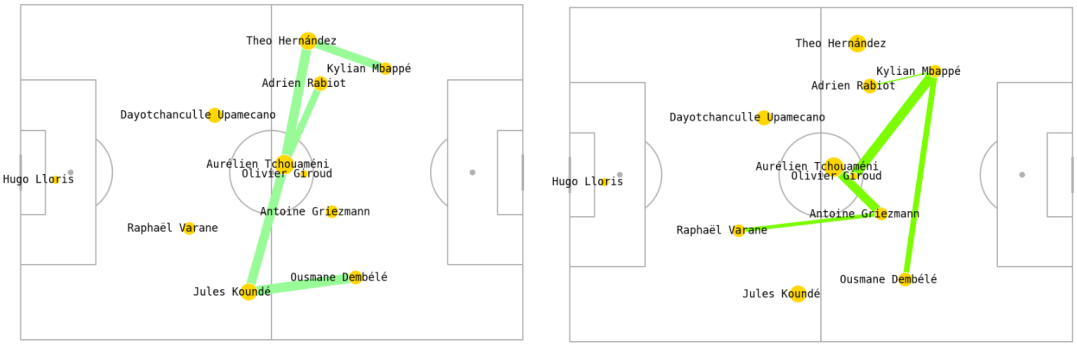
Su recorrido previo en este Mundial fue el siguiente:

Ronda	Partido	Goles Francia
Fase de Grupos	Francia 4 - Australia 1	Rabiot(asist. T.Hernández)-Giroud x 2(asist. Rabiot y Mbappé)-Mbappé(asist. Dembelé)
Fase de Grupos	Francia 2 - Dinamarca 1	Mbappé (asistencia Theo Hernández) - Mbappé (asistencia Griezmann)
Fase de Grupos	Francia 0 - Túnez 1	-
8vos de Final	Francia 3 - Polonia 1	Giroud (asist. Mbappé) - Mbappé (asist. Dembelé) - Mbappé (asist. Marcus Thuram)
4tos de Final	Francia 2 - Inglaterra 1	Tchouaméni (asistencia Griezmann) - Giroud (asistencia Griezmann)
Semifinal	Francia 2 - Marruecos 0	Theo Hernández - Randal Kolo Muani

Su formación principal es 4-2-3-1 y, eventualmente, utiliza 4-3-3.

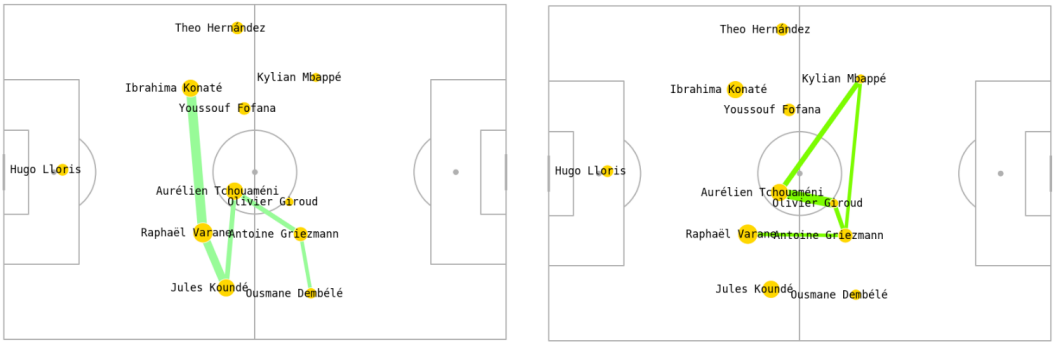
Quienes más rematan al arco son sus dos delanteros: *Mbappé* primero con 25 disparos (desde zona central o izquierda, 3 goles originados desde esta última), lo sigue *Giroud* con 17 (8 de ellos de cabeza).





(A) Red de pases por Francia - Polonia (top 5 conexiones)

(B) Red de pases por pdg Francia - Polonia (top 5 conexiones)



(A) Red de pases por Francia - Marruecos (top 5 conexiones)

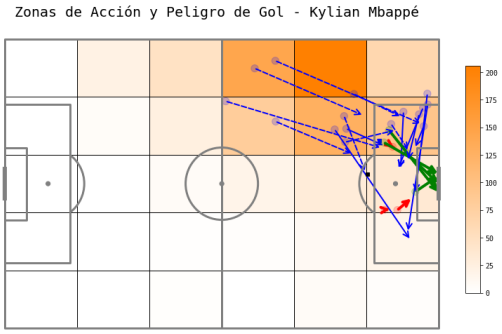
(B) Red de pases por pdg Francia - Marruecos (top 5 conexiones)

Sus redes de pases muestran un equipo corto, con su lateral izquierdo *Theo Hernández* moviéndose en posiciones mucho más altas y abiertas que *Koundé*, el lateral derecho. El mediocampista *Tchouaméni* se para en el círculo central, con *Rabiot* volcado a izquierda y *Griezmann* a su derecha, un poco más cerrado. Más adelante, *Mbappé* y *Dembelé* bien abiertos. *Giroud*, su centroatacante, llega a posiciones de gol pero también baja mucha veces para entrar en juego con la pelota.

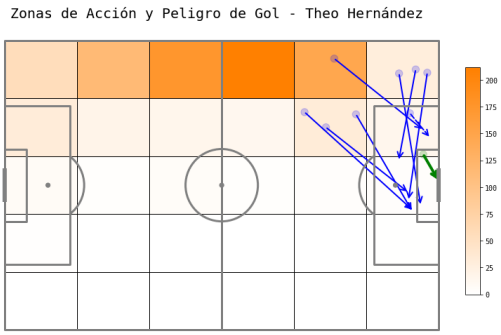
Resumen pdg - Francia - Mundial 2022

Métrica	Detalle
Conexiones de pases más peligrosas	Griezmann → Mbappé (pdg 1.25), Dembelé/Mbappé→ Giroud (pdg 0.5/0.5)
Pasadores más peligrosos	Griezmann (pdg 2.31, pases altos), Mbappé (pdg 0.97), T.Hernández (pdg 0.94)
Receptores más peligrosos	Mbappé (pdg 2.90), Giroud (pdg 2.15)
Altura de pase más peligrosa	Pases altos (54 % del peligro por esta vía)
Gambeteadores más peligrosos	Kylian Mbappé (pdg 0.13), Kingsley Coman (pdg 0.12)
Conductores más peligrosos	Mbappé (pdg 0.73), Griezmann (pdg 0.53), Dembelé (pdg 0.43)

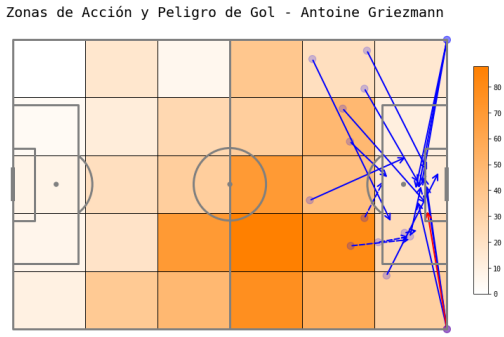
Al analizar las zonas y peligrosidad de sus principales jugadores, se puede contemplar una clara tendencia ofensiva por la banda izquierda.



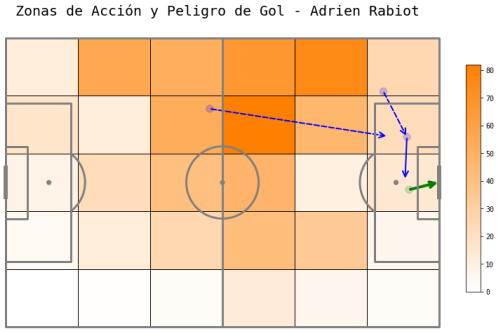
(A) Mapa de calor de **Kylian Mbappé** y jugadas con más peligro generado (umbral de 0.05)



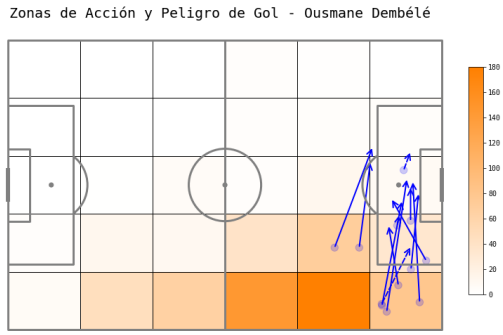
(B) Mapa de calor de **Theo Hernández** y jugadas con más peligro generado (umbral de 0.05)



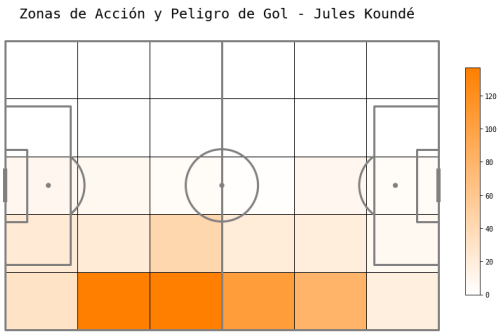
(A) Mapa de calor de **Antoine Griezmann** y jugadas con más peligro generado (umbral de 0.05)



(B) Mapa de calor de **Adrien Rabiot** y jugadas con más peligro generado (umbral de 0.05)



(A) Mapa de calor de **Ousmane Dembélé** y jugadas con más peligro generado (umbral de 0.05)

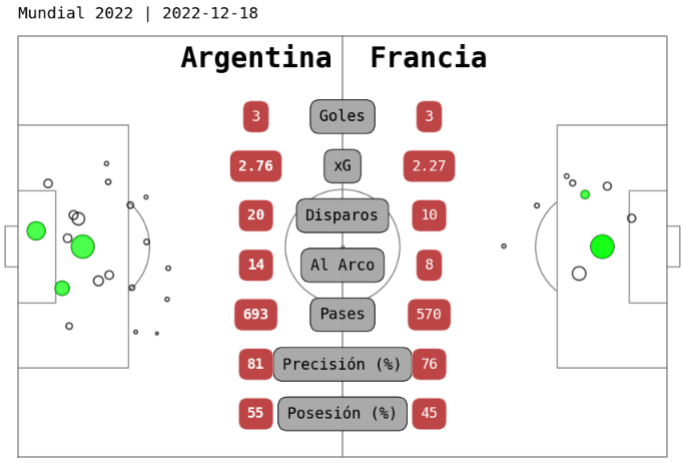


(B) Mapa de calor de **Jules Koundé** y jugadas con más peligro generado (umbral de 0.05)

A partir de estos gráficos, es posible destacar:

- *Mbappé* es quien más peligro genera desde la izquierda, muchas veces conduciendo desde posiciones retrasadas y también mediante centros atrás
- *Theo Hernández* produce daño mediante centros desde sector izquierdo
- El caso de *Griezmann* es interesante, ya que si bien se mueve por la zona derecha/central, su mayor peligro lo genera desde zona izquierda y córners, a través de centros principalmente
- *Rabiot* también se mueve por el sector izquierdo y *lastima* ofensivamente desde allí
- *Dembelé* es el único con incidencia ofensiva en el sector derecho, muchas veces llegando a línea final
- Por último, *Koundé*, lateral derecho, no produce peligro y se lo observa bastante retrasado posicionalmente

6.9.2. Post-partido

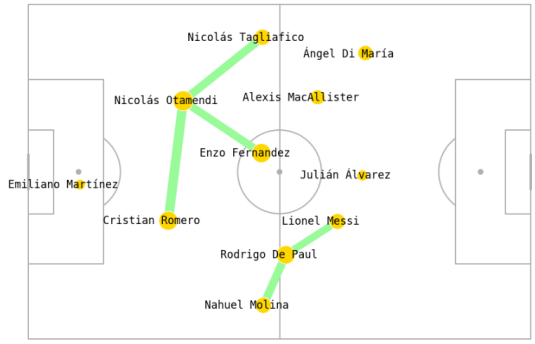


Goles

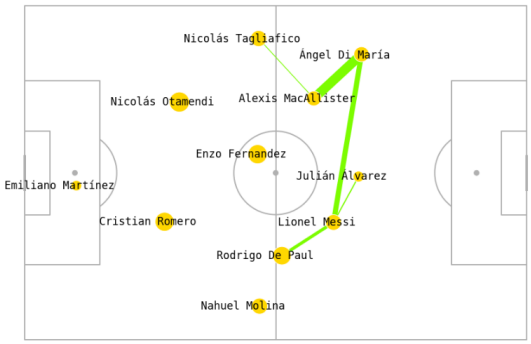
Equipo	Tiempo	Minuto	Detalle
Argentina	1	22	Lionel Messi (penal)
Argentina	1	35	Ángel Di María (asistencia Alexis MacAllister)
Francia	2	79	Kylian Mbappé (penal)
Francia	2	80	Kylian Mbappé (asistencia Marcus Thuram)
Argentina	4	107	Lionel Messi
Francia	4	117	Kylian Mbappé (penal)

Una final emocionante, cargada de tensión y dramatismo, que se terminó definiendo por penales (4-2 en favor de Argentina).

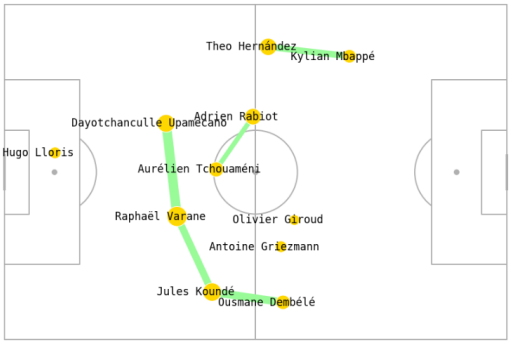
Argentina utilizó la formación 4-3-3 y Francia la 4-2-3-1. Otra vez el técnico argentino hizo ajustes en el equipo adaptándose al rival, sorprendiendo con *Di María* por la punta izquierda. Si bien no era una posición desconocida por el atacante argentino (por ejemplo, todo el Mundial pasado lo había hecho por ese sector), *Scaloni* venía utilizándolo por la banda derecha.



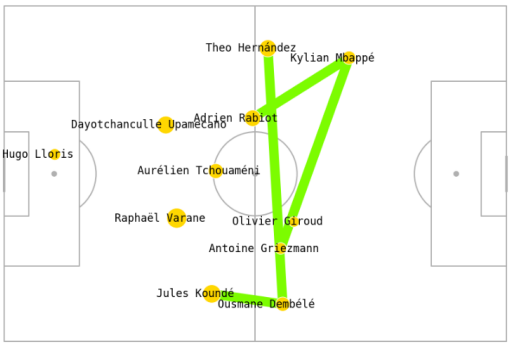
(A) Red de pases por Argentina (top 5 conexiones)



(B) Red de pases por pdg Argentina (top 5 conexiones)



(A) Red de pases por Francia (top 5 conexiones)



(B) Red de pases por pdg Francia (top 5 conexiones)

Las redes de pase muestran que el planteo de *Scaloni* provocó que Argentina hiciera daño por el sector izquierdo, con *Di María* y *MacAllister* asociándose de manera peligrosa. Esto terminó obligando a *Dembélé* a retrasarse (termina cometiendo el penal, no acostumbrado a tareas defensivas).

El equipo francés recién realizó su primer remate en el minuto 67 y una ráfaga le permitió emparejar el partido (con *Mbappé* y *Thuram* dañando por su zona más peligrosa, la izquierda de la cancha).

Al igual que en 4tos de final, el conjunto argentino volvió a mostrar su personalidad, manteniendo intensidad y vocación ofensiva a pesar del golpe anímico.

Período	ARG-Pases	FRA-Pases	ARG-Remates	FRA-Remates	ARG-xG	FRA-xG
1	312	210	6	0	1.29	0
2 - hasta minuto 80 (2-2)	155	204	4	4	0.26	1.03
2 - desde minuto 80 (2-2)	**65**	62	2	2	0.07	**0.13**
3	**90**	57	**4**	1	**0.41**	0.02
4	**71**	37	**4**	3	0.72	**1.09**

Resumen pdg - Argentina

Métrica	Detalle
Conexiones de pases más peligrosas	Messi → Lautaro Martínez (pdg 0.24), MacAllister → Di María (pdg 0.22)
Pasadores más peligrosos	Messi (pdg 0.42), Alexis MacAllister (pdg 0.25)
Receptores más peligrosos	Lautaro Martínez (pdg 0.63), Di María (pdg 0.24)
Altura de pase más peligrosa	Ras del piso (70 % del peligro por esta vía)
Gambeteadores más peligrosos	Sin destacados
Conductores más peligrosos	Messi (pdg 0.30), Lautaro Martínez (pdg 0.25)

Resumen pdg - Francia

Métrica	Detalle
Conexiones de pases más peligrosas	Griezmann → Kolo Muani (pdg 0.07), Thuram → Mbappé (pdg 0.06)
Pasadores más peligrosos	Griezmann (pdg 0.10), Thuram (pdg 0.09)
Receptores más peligrosos	Kolo Muani (pdg 0.17), Mbappé (pdg 0.14)
Altura de pase más peligrosa	Ras del piso (46 % del peligro por esta vía)
Gambeteadores más peligrosos	Mbappé (pdg 0.06)
Conductores más peligrosos	Mbappé (pdg 0.09)

6.10. Jugadores Más Peligrosos del Mundial

Finalmente, presentamos el top 10 de jugadores más peligrosos según *pdg* (absoluto y rating, es decir, su valor cada 90 minutos). Para el rating, se consideran jugadores que hayan tenido más de 270 minutos (al menos 3 partidos completos).

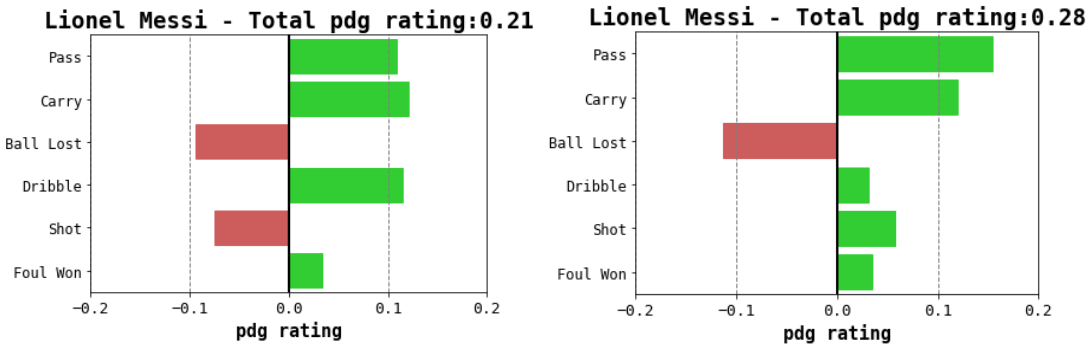
Top 10 - pdg (absoluto)

#	Jugador	Equipo	pdg total	# acciones	Pass	Shot	Carry	Dribble	Foul Won	Ball Lost
1	Kylian Mbappé	Francia	2.94	327	0.27	2.23	0.82	0.19	0.15	-0.72
2	Lionel Messi	Argentina	2.47	426	1.35	0.50	1.04	0.27	0.30	-0.99
3	Julián Álvarez	Argentina	1.68	106	-0.20	0.83	0.55	-0.04	0.33	0.20
4	Bukayo Saka	Inglaterra	1.54	102	0.14	1.01	0.45	-0.05	0.11	-0.12
5	Antoine Griezmann	Francia	1.48	200	1.60	-0.20	0.54	-0.07	0.03	-0.43
6	Mateo Kovačić	Croacia	1.43	190	0.71	-0.14	0.58	0.11	0.06	0.11
7	Theo Hernández	Francia	1.38	92	0.79	0.28	0.31	0.00	0.03	-0.04
8	Vinícius Júnior	Brasil	1.36	115	0.76	0.44	0.41	-0.11	0.07	-0.21
9	Kevin De Bruyne	Bélgica	1.35	118	1.15	-0.14	0.38	-0.01	0.02	-0.05
10	Mohammed Kudus	Ghana	1.29	94	0.35	0.65	0.14	0.02	0.18	-0.04

Top 10 - pdg (rating)

#	Jugador	Equipo	pdg rating	Minutos Jugados	Total	Pass	Shot	Carry	Dribble	Foul Won	Ball Lost
1	Bukayo Saka	Inglaterra	0.45	311	1.54	0.14	1.01	0.45	-0.12	0.11	-0.05
2	Kevin De Bruyne	Bélgica	0.41	294	1.35	1.15	-0.14	0.38	-0.05	0.02	-0.01
3	Vinícius Júnior	Brasil	0.39	311	1.36	0.76	0.44	0.41	-0.21	0.07	-0.11
4	Kylian Mbappé	Francia	0.39	686	2.94	0.27	2.23	0.82	-0.72	0.15	0.19
5	Raphaël Guerreiro	Portugal	0.32	311	1.11	0.44	0.47	0.26	-0.05	0.00	0.00
6	Julián Álvarez	Argentina	0.29	512	1.68	-0.20	0.83	0.55	0.20	0.33	-0.04
7	Dušan Tadić	Serbia	0.29	280	0.91	0.99	-0.03	-0.11	-0.06	0.03	0.09
8	Silvan Widmer	Suiza	0.28	286	0.91	0.88	-0.03	0.10	-0.05	0.00	0.00
9	Jordan Henderson	Inglaterra	0.28	291	0.91	0.48	0.35	0.06	0.00	0.03	-0.01
10	Lionel Messi	Argentina	0.28	786	2.47	1.35	0.50	1.04	-0.99	0.30	0.27

Además de poder encontrar los jugadores destacados, esta métrica nos permite también ver cómo contribuye cada jugador y cómo cambia esa contribución en el tiempo. Por ejemplo, si comparamos al *Messi* de 2018 con el de 2022, obtenemos:



(A) Distribución de pdg rating de Messi en Mundial 2018 (B) Distribución de pdg rating de Messi en Mundial 2022

A partir de estos gráficos, se puede observar que *Messi* tuvo mayor injerencia ofensiva durante el último Mundial y que, más allá de mejorar su efectividad en remates al arco, aumentó su contribución en pases y disminuyó en gambetas.

6.11. Extra: Copa África 2023

Por último, otro de los torneos que había formado parte del conjunto de validación a la hora de entrenar el modelo de *Peligro de Gol* fue la última Copa África. Aprovechando que no suele ser un torneo tan popular como los previamente analizados, se corrió el modelo para ver cuáles eran sus jugadores destacados.

Top 10 - pdg (absoluto)

#	Jugador	Equipo	pdg total	# acciones	Pass	Shot	Carry	Dribble	Foul Won	Ball Lost
1	Simon Adingra	Costa de Marfil	2.02	92	1.32	0.24	0.30	0.18	0.06	-0.08
2	Gelson	Angola	1.95	76	0.56	1.00	0.36	0.08	0.05	-0.11
3	Lamine Camara	Senegal	1.71	80	0.41	0.92	0.43	0.00	0.04	-0.09
4	Ademola Lookman	Nigeria	1.71	161	1.17	0.70	0.37	-0.50	0.06	-0.09
5	Hakim Ziyech	Morocco	1.53	82	1.04	0.48	0.26	-0.01	0.02	-0.26
6	Themba Zwane	Sudáfrica	1.37	165	0.53	0.86	0.06	0.10	0.03	-0.21
7	Sidi Bouna Sidi Amar	Mauritania	1.17	53	0.46	0.17	0.55	0.01	0.02	-0.04
8	Teboho Mokoena	Sudáfrica	1.13	173	1.18	-0.26	0.18	-0.01	0.08	-0.04
9	Seko Fofana	Costa de Marfil	1.12	137	0.03	0.27	0.54	0.06	0.01	0.21
10	Jean-Charles Castelletto	Camerún	1.10	51	0.72	0.39	0.01	0.01	0.00	-0.02

Tomando en cuenta los 4 primeros nombres:

- **Simon Adingra**: revelación de la Copa África, MVP de la final, extremo izquierdo del Brighton de Inglaterra (Cordovilla, 2024)
- **Gelson**: figura de Angola, medios mencionan su posible regreso al fútbol europeo, juega en Qatar y tuvo un paso por Portugal (Rubio, 2024)
- **Lamine Camara**: joven mediocampista senegalés que se encuentra en el radar de un grande como el Barcelona (Juanmartí, 2024)
- **Ademola Lookman**: en la última final de Europa League jugada en mayo 2024 le dio el triunfo al Atalanta con un *hat-trick* (Ordoñez, 2024)

De manera automática, en pocos segundos y sin ver ningún video, fue posible detectar jugadores destacados de un torneo completo.

Capítulo 7

Oficina de Datos

7.1. Introducción

En este capítulo, a modo de cierre, se presenta una propuesta describiendo los pasos iniciales para dar forma a una **Oficina de Datos** en un club del **Fútbol Argentino**.

En principio, deben darse tres condiciones para comenzar con una iniciativa de este tipo:

1. Contar con el interés del cuerpo técnico y/o de la dirigencia (idealmente de ambos)
2. Disponer de presupuesto para asignar a este proyecto
3. Asignar a los profesionales adecuados para llevarlo a cabo

Si bien contar con una *Oficina de Datos* es una decisión a largo plazo (asociado a una política del club), transitando las primeras fases de esta propuesta ya será posible ver resultados. Será clave avanzar en etapas, de menor a mayor, pudiendo generar la *confianza* necesaria que ayude a crecer en objetivos y que la inversión se justifique en base a los beneficios generados.

Las secciones siguientes describen una hoja de ruta realista y concreta, basada en las mejores prácticas del mundo empresarial para dar lugar a la aplicación de *Football Analytics* en una entidad deportiva.

Como aclaración, lamentablemente muchos de los productos o proveedores mencionados no publican los precios, motivo por el cual se brindan los datos que están disponibles para algunas alternativas, de manera que se pueda tener un **orden de magnitud** de los costos asociados.

7.2. Adquisición/Generación de Datos

Para comenzar a trabajar, el insumo indispensable son los *datos*. De mínima, es necesario contar con *ball event data* y sería deseable también disponer de información de *tracking* y datos biométricos.

En cuanto a *ball event data* y *tracking*, es necesario adquirir la información que genera un proveedor como los descriptos en capítulos anteriores (por ejemplo, *Statsbomb* o *Stats Perform*). La elección dependerá de diversos factores como:

- Existencia de datos para la liga donde participa el club
- Costo
- Detalle de la información ofrecida
- Disponibilidad de los datos (si están en tiempo real durante el partido, cuánto demoran en completarse después de finalizado el encuentro)

- Formato abierto de los datos

Además de la liga donde participa el club actualmente, es importante contar con:

- Datos históricos de la liga local (preferentemente 3 años de historia, muchos de los modelos requieren este tipo de información para ser entrenados)
- Información sobre otras ligas sudamericanas (posibles rivales de torneos internacionales y también potenciales mercados de jugadores)
- De participar en torneos internacionales (por ejemplo, *Mundial de Clubes*), información de las ligas de eventuales rivales
- Datos de otras categorías (*Ascenso*), posibles rivales de *Copa Argentina* y mercado de pases

Si bien todos los puntos anteriores son importantes, hay dos que son centrales. En primer lugar, es vital conocer **cuándo** y **cómo** el proveedor brindará la información. De esto dependerá si se pueden hacer análisis *real-time* o cuánto se deberá esperar para poder procesar la información generada. Por ejemplo, algunos proveedores pueden demorar alrededor de 84 hs (3 días y medio) en brindar los datos del partido de forma completa. El segundo punto está relacionado al **formato abierto de la información**. Muchas empresas ofrecen una suite de productos desde la cual se puede analizar el rendimiento pero no permiten extraer la información para generar análisis propios. Se debe estar preparado para la convivencia entre múltiples proveedores y también al cambio de empresa a lo largo del tiempo. Por este motivo, es *fundamental* que los datos puedan ser extraídos de cada plataforma para así generar una **base propia de conocimiento**.

Por último, para datos biométricos, la mayoría de los clubes ya cuentan con su propio proveedor (*Catapult* o similar), con lo cual habría que analizar qué tan abierto es su formato y cómo se pueden cruzar estos datos con los eventos con pelota o de tracking para poder realizar análisis integrados.

Para tener una referencia de costos para la provisión de datos, un proveedor de primera línea cuenta con estos precios (a julio 2023, no real-time, impuestos incluidos):

Descripción	Ball Event Data	Tracking Data
1 liga	1.800 USD mensuales	800 USD mensuales
5 ligas	3.200 USD mensuales	1.600 USD mensuales
10 ligas	4.050 USD mensuales	2.750 USD mensuales

7.3. Construcción de Base de Conocimiento

El siguiente paso es uno de los puntos centrales de la *Oficina de Datos*: generar una base de conocimiento propia con la información producida por distintos proveedores, diferentes tipos de datos y que permita conocer y detectar tendencias históricas. Esta estructura es usualmente conocida como *data-warehouse* y es desde donde se generan todos los análisis, informes y se pueden entrenar modelos predictivos.

En cuanto a infraestructura, se propone pensar en una solución *cloud*, caracterizada por ser escalable y elástica. Muchos de los productos del stack tecnológico relacionado a datos son auto-administrados, lo cual simplifica la gestión y mantenimiento de los mismos. Además, el esquema de *pay-as-you-go* ofrece flexibilidad para que los costos vayan acompañando el crecimiento mismo de la *Oficina de Datos* y no sea necesario realizar una alta inversión inicial.

Dentro de las nubes más populares, productos como *BigQuery* (nube *Google*), *Amazon RedShift* (nube *AWS*) o *Azure Synapse Analytics* (nube *Azure*) podrían ser de utilidad para esta etapa.

A modo de referencia, se brindan costos asociados a los productos de la nube de *Google (GCP*[1]):

Producto	Costo	Detalle
BigQuery	1.200 USD mensuales	Región us-central1, 20 TB active storage, 20 TB análisis (on-demand)
Máquinas Virtuales	1.500 USD mensuales	Región us-central1, 3 máquinas estándar (no spot), para entrenamiento de modelos y generalidades

7.4. Visualizaciones

Muchos de los proveedores de datos ofrecen sus propias herramientas de visualización, sin embargo es fundamental contar con una propia que habilite cruzar distintos orígenes de datos y muestre informes integrados.

Dentro de las herramientas de BI más populares de la actualidad, se destacan *Power BI*, *Tableau* y *Looker*, las cuales brindan facilidades para construir informes de manera rápida y que pueden compartirse con múltiples usuarios.

Para tomar de referencia, *Power BI* ofrece licencias por usuario a 10 USD mensual[2].

7.5. Posibles Aplicaciones

Dentro de los escenarios donde podrían aplicarse estas técnicas, se destacan:

Aplicación	Descripción
Análisis de Rivales	Generar informes en la previa de cada partido que ayuden a anticipar situaciones y a comprender el juego del rival
Análisis Real-Time	Brindar asistencia e información mientras está transcurriendo el partido
Análisis Post-Partido	Proveer informes que resuman el rendimiento del equipo y ayuden a explicar el resultado del encuentro
Análisis Rendimiento Jugador	Analizar el rendimiento individual de un jugador, comprender situaciones, detectar fortalezas y puntos a mejorar
Definición de Métricas del Equipo	Definición y seguimiento de indicadores de rendimiento del equipo (tiempo promedio de posesión, ancho y largo en metros, entre otras posibles métricas)
Reclutamiento	Buscar, analizar y realizar seguimiento de jugadores de interés para el club
Seguimiento de Jugadores a Préstamo	Monitorear rendimiento de aquellos jugadores que se encuentran en calidad de cedidos
Bitácora de Lesiones	Mantener un histórico de las lesiones de cada jugador (servirá de base para modelos que puedan anticipar lesiones)
Detección de Talento	Analizar rendimiento de jugadores de las divisiones inferiores para colaborar en el desarrollo y detección de jóvenes promesas

[1] Calculadora GCP
[2] Licencia Power BI Pro

7.6. Perfiles Necesarios y Roles

Para poder llevar a cabo este proyecto de manera exitosa se precisa contar con perfiles *senior* con mentalidad científica pero, al mismo tiempo, una fuerte orientación práctica. En la actualidad, la posición en la que se encuadran estos roles es la de **científicos de datos**. Dentro de los conocimientos necesarios, se especifican:

- Formación en Informática, Matemáticas, Estadística o disciplinas relacionadas
- Conocimientos de fútbol y táctica
- Fuertes conocimientos en Programación, *Machine Learning* y Visualización de Datos

En cuanto a estructura de equipo, es natural que se comience con un grupo reducido de profesionales (asignación parcial o completa según el caso), creciendo en staff y carga a medida que los proyectos y resultados lo demanden. Según la encuesta realizada en 2023 por la consultora especializada *Left Field* (Left Field, 2023), el staff de un área de este estilo está compuesto por entre 2 y 3 personas. En este punto será clave buscar perfiles versátiles (en su artículo, Sormaz, 2023 habla de personas que deberán utilizar muchos sombreros al mismo tiempo) y que puedan cubrir el punta a punta desde la captura de datos, la generación de informes, el entrenamiento de modelos y la administración de la infraestructura asociada. Además, deben estar preparados para trabajar bajo presión y en horarios y días no habituales (el cronograma lo irá marcando el fixture y el desempeño del equipo).

Un punto clave es la definición de roles que faciliten la interacción entre la Oficina de Datos-Cuerpo Técnico-Dirigencia, seleccionando las personas adecuadas que actúen como nexo entre estas áreas.

Como valor de referencia, según estimaciones de mercado, el valor hora de un *científico de datos* puede oscilar entre los 60 USD y 75 USD.

7.7. Presupuesto a Asignar

Los elementos que componen el presupuesto necesario para llevar adelante una iniciativa de este estilo entonces son:

- Adquisición de Datos (liga propia y externas)
- Infraestructura Cloud
- Salarios de profesionales de la *Oficina de Datos*

A continuación se describen tres posibles configuraciones, cada una con diferente alcance y presupuesto requerido. Vale la pena aclarar que lo expresado debe tomarse como *valores referenciales*, calculados en base a información que pudo conseguirse de manera online o a través de interacciones con personal del área de Ventas de cada solución. Las alternativas presentadas son: una *básica*, con información de una sola liga y poca asignación de recursos humanos y tecnológicos; una *estándar* que incluye más ligas y más horas-hombre; y una *avanzada* que contempla más ligas, recursos e infraestructura. Los costos están calculados de manera aproximada (sobre precios de lista) y se realizaron para un proveedor de *primera línea*, nube *Google* y herramienta *Power BI*. Los valores están expresados en **USD** y son **mensuales**. Para todos los ítems indicados no están contemplados descuentos que puedan llegar a negociarse por cantidad o duración de los contratos.

Ítem	Básica	Estándar	Avanzada
Ball Event Data	1.800 USD (1 liga)	3.200 USD (5 ligas)	4.050 USD (10 ligas)
Tracking Data	-	800 USD (1 liga)	1.600 USD (5 ligas)
Infraestructura	1.000 USD (consumo bajo)	2.000 USD (consumo medio)	3.000 USD (consumo medio/alto)
Visualización	100 USD (10 licencias)	300 USD (30 licencias)	500 USD (50 licencias)
Personal	1.875 USD (25 horas)	3.750 USD (50 horas)	12.000 USD (160 horas)
Total	**4.775 USD**	**10.050 USD**	**21.150 USD**

7.8. Etapas de Implementación

La velocidad y los tiempos con los que este tipo de iniciativas pueden avanzar dependenderán fuertemente de las horas asignadas tanto por los científicos de datos como también por las partes interesadas en los informes y datos que se generen.

En líneas generales, luego de un primer mes de *set-up* inicial y diseño básico de la base de conocimiento, estarán dadas las condiciones para que cada encuentro del equipo cuente con un análisis pre y post-partido. Luego del primer trimestre, se podrían incorporar análisis relacionados al mercado de pases para colaborar en las futuras incorporaciones del equipo. Transcurridos los primeros 6 meses, sería posible incursionar en las divisiones formativas, realizando estudios sobre futuros talentos.

Las primeras etapas serán de conocimiento mutuo, de idas y vueltas constantes que ayuden a encontrar las oportunidades donde estas técnicas pueden aportar la información que el cuerpo técnico y club necesitan. Transcurrido el primer año, se espera que la *Oficina de Datos* haya alcanzado la madurez necesaria para ser una herramienta clave a disposición del club.

Capítulo 8

Conclusión

8.1. Trabajos Futuros

Si bien la solución propuesta está orientada a la realidad del fútbol argentino, su aplicación no necesariamente está limitada a él. Estas ideas podrían llevarse a otras ligas o selecciones que busquen aplicar datos para enriquecer sus análisis y generar una ventaja competitiva.

A su vez, otros actores del ecosistema del fútbol como federaciones, representantes/agentes, medios y periodistas podrían agregar una perspectiva más a su mirada actual del fútbol, incorporando informes o análisis generados a partir de datos.

Por otro lado, muchas de las propuestas e ideas planteadas podrían ser adaptadas a otros deportes como polo, rugby o hockey, donde la aplicación de estas técnicas analíticas podría ayudar a estudiar diferentes aspectos del juego. Para avanzar en este sentido, sería necesario disponer de datos históricos de partidos del deporte en cuestión y apoyo de expertos que colaboren para definir métricas y ayudar a encontrar oportunidades concretas para aplicar estas técnicas.

En cuanto al modelo de *Peligro de Gol*, un punto de interés sería el de incorporar datos de *tracking* para poder capturar más información sobre el contexto de juego. No existen límites técnicos para avanzar, solamente debe contarse con datos suficientes para realizar el entrenamiento y ajuste de modelos. Para llevar a cabo un trabajo similar al realizado en esta tesis, se precisaría contar con al menos 2000 partidos históricos que brinden la información suficiente para incorporar nuevas características relacionadas a la distribución de los jugadores dentro del campo. El proceso de *feature-engineering* sería de mayor complejidad, debiendo incorporar conceptos relacionados al análisis espacial. Para el entrenamiento y validación de los modelos sería posible reutilizar todo lo realizado en este trabajo.

8.2. Conclusión

A lo largo de este trabajo se realizó un profundo estudio sobre las diversas técnicas y aplicaciones relacionadas a *Football Analytics* y se logró evidenciar que su aplicación puede ser real y concreta, adaptándose a las distintas realidades de cada club. Tal como sucede en la mayoría de los campos, la *Inteligencia Artificial* no viene a reemplazar perfiles ni a brindar la fórmula mágica del éxito, pero su aplicación en el fútbol proporcionará una nueva herramienta, una *mirada distinta* de la que hoy ya existe en los clubes y que permitirá generar conocimiento que ayude a ganar partidos.

Adicionalmente, la velocidad y capacidad de procesamiento de estas técnicas y modelos permiten acelerar muchos de los procesos que hoy se realizan manualmente como el análisis de rivales o *scouting* mediante videos. Esto no significa que dejará

de ser necesario observar jugadas/partidos en vivo y en video, si no que se podrá ser mucho más selectivo (y efectivo) a la hora de realizarlo. Se podrán abarcar más rivales y más partidos, pudiendo decidir con mayor precisión sobre qué encuentros se hará un análisis manual y más detallado. Tomando como ejemplo el caso de *scouting* de jugadores, el modelo *Peligro de Gol* permitiría en minutos barrer todas las ligas sudamericanas y proveer una lista acotada de jugadores y de partidos destacados, que sirva para luego confirmar o descartar manualmente potenciales jugadores de interés.

Sin lugar a dudas, compatibilizar el mundo del fútbol con el de la *Inteligencia Artificial* no será sencillo, deberá ir de menor a mayor, generando poco a poco la confianza necesaria para que los análisis resulten aplicables a situaciones concretas de juego. Modelos como el propuesto (*Peligro de Gol*) pueden servir para establecer los primeros puentes, permitiendo generar valor a través de técnicas interpretables y consistentes con el ojo experto.

El hecho de que muchos clubes aún no hayan emprendido el camino de *Football Analytics* le proveerá a los primeros una ventaja competitiva por sobre el resto. En definitiva, *no hay excusas*, los datos y las técnicas están, pueden adaptarse a diferentes presupuestos, hoy la decisión final está en manos de los clubes y el destino del *Football Analytics* dependerá de su determinación y audacia.

Bibliografía

Aichner, Thomas (2019). «Football clubs' social media use and user engagement». En: *Marketing Intelligence and Planning* 37.3, págs. 242-257.

Altman, Dan (2020). *Who's who in a football data department*. URL: https://trainingground.guru/articles/dan-altman-whos-who-in-a-football-data-department.

Analyst, The (2023). *Brighton and Brentford: Two Smart Clubs Who Play the Game in Opposite Ways*. URL: https://theanalyst.com/eu/2023/03/brighton-and-brentford-two-smart-clubs-who-play-the-game-in-opposite-ways/.

Bellizzi, Germán (2020). *Big Data en el fútbol: qué es, su utilidad y cómo se aplica en la Argentina*. URL: https://www.tycsports.com/al-angulo/big-data-en-el-futbol-que-es-su-utilidad-y-como-se-aplica-en-la-argentina-20200726.html.

Bernath, Gonzalo (2021). «La utilización de BIG DATA para la gestión eficiente de clubes del fútbol argentino». Trabajo Final de Carrera. Tesis. Universidad de Belgrano, Buenos Aires, Argentina. URL: http://repositorio.ub.edu.ar/handle/123456789/10097.

Brier, Glenn W (1950). «Verification of forecasts expressed in terms of probability». En: *Monthly weather review* 78.1, págs. 1-3.

Brúgola, Luciano Ariel, Guillermo Durán y Andrés Farral (2021). «¿Cuándo se convierten los goles en el fútbol? Modelado Estadístico de la Ocurrencia de goles de las principales Ligas Profesionales Internacionales de Fútbol». Licenciatura en Matemáticas - Facultad de Ciencias Exactas y Naturales. Tesis. Universidad de Buenos Aires, Buenos Aires, Argentina. URL: https://web.dm.uba.ar/files/tesis_lic/2021/brugola.pdf.

Cordovilla, Iván (2024). *Adingra: de ser víctima de una estafa con 12 años a MVP de la final*. URL: https://as.com/futbol/internacional/adingra-de-ser-victima-de-una-estafa-con-12-anos-a-mvp-de-la-final-n/.

Cotton, Richie (2022). *How Chelsea FC Uses Analytics to Drive Matchday Success*. URL: https://www.datacamp.com/podcast/how-chelsea-fc-uses-analytics-to-drive-matchday-success.

Decroos, Tom et al. (jul. de 2019). «Actions Speak Louder than Goals: Valuing Player Actions in Soccer». En: págs. 1851-1861. ISBN: 978-1-4503-6201-6. DOI: 10.1145/3292500.3330758.

Diament, Yoel Hernán (2021). «Predicción del comportamiento de los jugadores en el fútbol argentino». Trabajo Final de Posgrado. Tesis. Universidad de Buenos Aires, Buenos Aires, Argentina. URL: http://bibliotecadigital.econ.uba.ar/download/tpos/1502-1913_DiamentYH.pdf.

Fawcett, Tom (jun. de 2006). «Introduction to ROC analysis». En: *Pattern Recognition Letters* 27, págs. 861-874. DOI: 10.1016/j.patrec.2005.10.010.

FC, Analytics (2021). *Kevin De Bruyne and football's data revolution*. URL: https://analyticsfc.co.uk/case-studies/kevin-de-bruyne-and-footballs-data-revolution/.

Ferreyra, Mario Ezequiel (2021). «Análisis de Sentimiento en Tweets de Fútbol Argentino». Trabajos Especiales de Licenciatura en Ciencias de la Computación.

Tesis. Universidad Nacional de Córdoba, Córdoba, Argentina. URL: https://rdu.unc.edu.ar/handle/11086/18384.

FIFA (2021). *El futuro es ahora: la FIFA lleva el análisis del rendimiento a una dimensión totalmente nueva*. URL: https://inside.fifa.com/es/tournaments/mens/arabcup/arabcup2021/media-releases/el-futuro-es-ahora-la-fifa-lleva-el-analisis-del-rendimiento-a-una-dimension.

Ford, C. (2000). *A Brief on Brier Scores*. URL: https://library.virginia.edu/data/articles/a-brief-on-brier-scores.

Gantman, Marcelo (2021). *El fútbol argentino incorpora la inteligencia artificial de Smart Gear*. URL: https://bigdatasports.media/2021/02/02/el-futbol-argentino-incorpora-el-tracking-optico-de-smart-gear/.

Gong, Hua y Su Chen (2023). «Estimating Positional Plus-Minus in the NBA». En: *MIT Sloan Sports Analytics Conference*.

Grund, Thomas (2012). «Network structure and team performance: The case of English Premier League soccer teams». En: *Social Networks* 34.4, págs. 682-690. ISSN: 0378-8733. DOI: https://doi.org/10.1016/j.socnet.2012.08.004. URL: https://www.sciencedirect.com/science/article/pii/S0378873312000500.

Juanmartí, Toni (2024). *Lamine Camara, otra opción 'low cost' para el Barça*. URL: https://www.sport.es/es/noticias/barca/lamine-camara-ofrece-barca-105110041.

Ke, Guolin et al. (2017). «LightGBM: A Highly Efficient Gradient Boosting Decision Tree». En: *Neural Information Processing Systems*. URL: https://api.semanticscholar.org/CorpusID:3815895.

Kloppy (2020). *Kloppy: standardizing soccer tracking and event data*. URL: https://kloppy.pysport.org/.

Kullowatz, Matthias (2020). *Goals Added: Deep dive methodology*. URL: https://www.americansocceranalysis.com/home/2020/5/4/goals-added-deep-dive-methodology.

Left Field (2023). *Analytics in Football: the state of play*. URL: https://leftfieldfc.com/analytics-in-football-the-state-of-play/.

Lundberg, Scott M y Su-In Lee (2017). «A Unified Approach to Interpreting Model Predictions». En: *Advances in Neural Information Processing Systems*. Ed. por I. Guyon et al. Vol. 30. Curran Associates, Inc. URL: https://proceedings.neurips.cc/paper_files/paper/2017/file/8a20a8621978632d76c43dfd28b67767-Paper.pdf.

Melville, William et al. (2024). «Optimizing Baseball Fielder Positioning with Consideration for Adaptable Hitters». En: *MIT Sloan Sports Analytics Conference*.

Merhej, Charbel et al. (2021). «What Happened Next? Using Deep Learning to Value Defensive Actions in Football Event-Data». En: *Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, págs. 3394-3403. DOI: 10.1145/3447548.3467090. eprint: 2106.01786.

Ntsoane, Tebogo (2023). *6 football clubs who are making great use of data analytics*. URL: https://www.sportskeeda.com/football/6-football-clubs-making-great-use-of-data-analytics.

O Tempo, Fernando Martins Y Miguel (2021). *Galo é pioneiro e lança setor analytics, que 'transforma' dados em conhecimento*. URL: https://www.otempo.com.br/sports/atletico/galo-e-pioneiro-e-lanca-setor-analytics-que-transforma-dados-em-conhecimento-1.2486872.

Ordoñez, Leandro (2024). *Quién es Lookman, el delantero que pulverizó al Bayer Leverkusen y le dio la Europa League al Atalanta*. URL: https://www.ole.com.ar/

futbol-internacional/europa-league/ademola-lookman-atalanta-bayer-leverkusen-europa-league-final-goles_0_cRPuDzxfwZ.html.

Palmeiras, Departamento de Comunicação (2024). *Palmeiras se torna o primeiro time sul-americano a usar os dados do Opta Vision*. URL: https://www.palmeiras.com.br/noticias/palmeiras-se-torna-o-primeiro-time-sul-americano-a-usar-os-dados-do-opta-vision/.

Pretto, Fabricio y Guido De Caso (2022). «Development of a Football Analytics Web Application for Player Scouting». Master in Management + Analytics. Tesis. Universidad Torcuato Di Tella, Buenos Aires, Argentina. URL: https://repositorio.utdt.edu/bitstream/handle/20.500.13098/11892/MiM_Pretto_2022.pdf.

Rubio, Alberto (2024). *Gelson Dala, la palanca de las 'Palancas': "De un momento a otro podría volver a Europa"*. URL: https://www.marca.com/futbol/copa-africa/2024/01/30/65b7f7e5e2704e06a58b45a3.html.

Rudd, Sarah (2011). «A Framework for Tactical Analysis and Individual Offensive Production Assessment in Soccer Using Markov Chains». En: *New England Symposium on Statistics in Sports*.

Schoenfeld, Bruce (2019). *El arma secreta del Liverpool: el análisis de datos*. URL: https://www.nytimes.com/2019/05/22/magazine/soccer-data-liverpool.html.

Singh, Karun (2018). *Introducing Expected Threat (xT)*. URL: https://karun.in/blog/expected-threat.html.

Song, Huan et al. (2023). «Explainable Defense Coverage Classification in NFL Games using Deep Neural Networks». En: *MIT Sloan Sports Analytics Conference*.

Sormaz, Mladen (2023). *Building a data analytics department*. URL: https://trainingground.guru/articles/mladen-sormaz-building-a-data-analytics-department.

StatsBomb (2018). *StatsBomb Release Free FIFA World Cup Data*. URL: https://statsbomb.com/news/statsbomb-release-free-fifa-world-cup-data/.

— (2021a). *StatsBomb Announce The Release Of Free StatsBomb 360 Data: Euro 2020 Available Now*. URL: https://statsbomb.com/news/statsbomb-announce-the-release-of-free-statsbomb-360-data-euro-2020-available-now/.

— (2021b). *StatsBomb Release Free Lionel Messi Data: All Seasons From 2004/05 – 2020/21 Now Available*. URL: https://statsbomb.com/news/statsbomb-release-free-messi-data-all-seasons-from-2004-05-2020-21-now-available/.

— (2022a). *StatsBomb Open Data Specification*. URL: https://github.com/statsbomb/open-data/tree/master/doc.

— (2022b). *StatsBomb Release Free FIFA World Cup Data*. URL: https://statsbomb.com/news/statsbomb-release-free-2022-world-cup-data/.

— (2022c). *What is Expected Threat (xT)? Possession Value models explained*. URL: https://statsbomb.com/soccer-metrics/possession-value-models-explained/.

— (2023a). *StatsBomb Release Free 2023 African Cup of Nations Data*. URL: https://statsbomb.com/news/statsbomb-release-free-2023-african-cup-of-nations-data/.

— (2023b). *The 2015/16 Big 5 Leagues Free Data Release: Bundesliga*. URL: https://statsbomb.com/news/the-2015-16-big-5-leagues-free-data-release-bundesliga/.

— (2023c). *The 2015/16 Big 5 Leagues Free Data Release: La Liga*. URL: https://statsbomb.com/news/the-2015-16-big-5-leagues-free-data-release-la-liga/.

— (2023d). *The 2015/16 Big 5 Leagues Free Data Release: Ligue 1*. URL: https://statsbomb.com/news/the-2015-16-big-5-leagues-free-data-release-ligue-1/.

StatsBomb (2023e). *The 2015/16 Big 5 Leagues Free Data Release: Premier League*. URL: https://statsbomb.com/news/the-2015-16-big-5-leagues-free-data-release-premier-league/.

— (2023f). *The 2015/16 Big 5 Leagues Free Data Release: Serie A*. URL: https://statsbomb.com/news/the-2015-16-big-5-leagues-free-data-release-serie-a/.

— (2024). *Como o Brasil está liderando a revolução dos dados no futebol latino-americano*. URL: https://statsbomb.com/pt-pt/noticias-pt/como-o-brasil-esta-liderando-a-revolucao-dos-dados-no-futebol-latino-americano/.

StatsPerform (2019). *Introducing a Possession Value Framework*. URL: https://www.statsperform.com/resource/introducing-a-possession-value-framework/.

— (2020). *Evolving Possession Value*. URL: https://www.statsperform.com/resource/evolving-our-possession-value-framework/.

Sumpter, David (2016). *Soccermatics: Mathematical Adventures in the Beautiful Game*. Bloomsbury sigma series. Bloomsbury Publishing Plc. ISBN: 9781472924131. URL: https://books.google.com.ar/books?id=IwW0jwEACAAJ.

— (2021). *Explaining Expected Threat*. URL: https://soccermatics.medium.com/explaining-expected-threat-cbc775d97935.

— (2022). *Soccermatics Course*. URL: https://soccermatics.readthedocs.io/en/latest/.

Tempone, Pablo Matías (2017). «Predicción de victorias en los equipos locales, del fútbol argentino, según las acciones de jugadores [Tesis de maestría no publicada]». Trabajo Final - Maestría en Explotación de Datos y Descubrimiento del Conocimiento. Tesis. Universidad de Buenos Aires, Buenos Aires, Argentina.

Tuyls, Karl et al. (mayo de 2021). «Game Plan: What AI can do for Football, and What Football can do for AI». En: *Journal of Artificial Intelligence Research* 71, págs. 41-88. DOI: 10.1613/jair.1.12505.

Van Roy, Maaike, Pieter Robberechts y Jesse Davis (2021). «Optimally Disrupting Opponent Build-ups». En: *Proceedings of the StatsBomb Innovation in Football Conference*. StatsBomb. London, United Kingdom.

Villarreal, Antonio (2019). *La reinvención de Monchi, el mago de los fichajes: Obviar el big data es anacrónico*. URL: https://www.elconfidencial.com/deportes/futbol/2019-10-17/entrevista-monchi-sevilla-big-data-443_2278023/.

Vovk, Vladimir (2015). *The fundamental nature of the log loss function*. eprint: 1502.06254.

Whitmore, Jonny (2023a). *Slice It Up: Introducing Opta Player Radars*. URL: https://theanalyst.com/eu/2023/06/introducing-opta-radars-compare-players/.

— (2023b). *What Is Expected Goals (xG)?* URL: https://theanalyst.com/eu/2023/08/what-is-expected-goals-xg/.

Zivkovic, Jason (2021). *worldfootballR*. URL: https://github.com/JaseZiv/worldfootballR.

Álvarez, David (2023). *Así es el gurú de los datos del Liverpool que descubrió a Firmino y Salah*. URL: https://elpais.com/deportes/futbol/2023-11-07/asi-es-el-guru-de-los-datos-del-liverpool-que-descubrio-a-firmino-y-salah.html.