UNIVERSIDAD DE CHILE
FACULTAD DE CIENCIAS FÍSICAS Y MATEMÁTICAS
DEPARTAMENTO DE INGENIERÍA INDUSTRIAL

IMPACTO DE LOS FUTBOLISTAS DE LA "PREMIER LEAGUE 2017-2018" EN LA PROBABILIDAD DE SALIR CAMPEÓN A TRAVÉS DE UN MÉTODO DE SIMULACIÓN CON INFERENCIA BAYESIANA

TESIS PARA OPTAR AL GRADO DE MAGÍSTER EN GESTIÓN DE OPERACIONES

MEMORIA PARA OPTAR AL TÍTULO DE INGENIERO CIVIL INDUSTRIAL

SEBASTIÁN ANDRÉS MENA DUCE

PROFESOR GUÍA:
DENIS SAURÉ VALENZUELA

MIEMBROS DE LA COMISIÓN:
ANDRÉS WEINTRAUB POHORILLE
MARCELO OLIVARES ACUÑA
GUILLERMO DURÁN

Este trabajo ha sido parcialmente financiado por:
ANID - CONICYT

SANTIAGO DE CHILE
2021

RESUMEN DE LA TESIS PARA OPTAR
AL GRADO DE MAGÍSTER EN GESTIÓN
DE OPERACIONES
POR: **SEBASTIÁN ANDRÉS MENA DUCE**
FECHA: 2021
PROF. GUÍA: DENIS SAURÉ VALENZUELA

IMPACTO DE LOS FUTBOLISTAS DE LA "PREMIER LEAGUE 2017-2018" EN LA PROBABILIDAD DE SALIR CAMPEÓN A TRAVÉS DE UN MÉTODO DE SIMULACIÓN CON INFERENCIA BAYESIANA

El proceso de *scouting* que tiene por objetivo la búsqueda de nuevos jugadores en los clubes de fútbol profesional, actualmente es deficiente por la inexistencia de una herramienta que ayude a analizar la cantidad de datos e información que se tienen disponibles. En los últimos años, clubes de Europa y algunos clubes de América han desarrollado sistemas basados en datos de rendimiento para apoyar la gestión de contratación de jugadores, sin embargo estos aún no se han desarrollado desde una mirada más global.

El **objetivo principal** de esta tesis es construir un sistema para cuantificar el efecto de un jugador en las probabilidades de campeonar de un equipo de fútbol, en base al rendimiento deportivo de los jugadores en la temporada 2017-2018 de la Premier League con datos proporcionados por Wyscout.

La metodología consta de tres etapas. En primer lugar, se calculan las probabilidades de elegir una acción por un jugador en cierta zona del campo, bajo las condiciones del partido, utilizando un modelo *Multi-logit* e Inferencia Bayesiana para estimar los parámetros. Luego, se calculan las probabilidades de ejecutar de forma correcta esa acción con un modelo *Logit*, basándose tanto en las condiciones del partido como en los jugadores involucrados en dicho evento, estimando los parámetros que definen esta probabilidad con Inferencia Bayesiana. Finalmente, la última etapa consta del modelamiento de un sistema de recomendación basado en simulaciones. Para esto, se construye un modelo de simulación de Cadenas de Markov en donde cada uno de los eventos es un nodo de la cadena y las probabilidades de transición entre un nodo y otro se construyen a partir de la combinación de las probabilidades calculadas en la etapa uno y dos.

De esta forma, se simula el torneo completo en una gran cantidad de iteraciones, permitiendo en primer lugar, comparar el rendimiento predictivo de los resultados con el modelo de Poisson, que supera levemente al modelo de Cadenas de Markov con un acierto predictivo de 39,37 % vs. 44,45 % y en segundo lugar, realizar experimentos para evaluar distintos jugadores que podrían llegar a un club y cuantificar el efecto de ese jugador en la probabilidad de campeonar del equipo. La flexibilidad para ensayar y probar en la simulación permite hacer más eficiente la búsqueda de nuevos jugadores para cada club, reducir el campo de opciones a explorar, el tiempo que le dedican los *scouts* a este proceso y eliminar el sesgo que se genera al validar este proceso solo con el juicio de expertos. De esta forma, ayuda a concentrarse en aquellos deportistas que podrían ser una buena opción a contratar, haciendo el proceso más eficiente y ahorrando recursos para los clubes.

Para las y los que día a día,
buscan una mejor versión de sí mismos.

Saludos

Tabla de Contenido

Índice de Tablas

Índice de Ilustraciones

Capítulo 1

Introducción

La relevancia que han tomado en la última década los datos en las organizaciones ha aumentado exponencialmente. Muestra de ello es que cada vez escuchamos con mayor frecuencia términos como: "*Big Data*", "*Data Mining*" o "Inteligencia Artificial". Si miramos hacia atrás, el origen de la palabra "datos" viene del latín "*datum*" y que su etimología refiere: *"su significado en un comienzo era "lo que se da", es decir, cualquier cosa dada pero que luego se especializó para hechos o información"* [1].

La cantidad de datos que se pueden recopilar y, en algunos casos, la facilidad para hacerse con estos ha dado pie para que el manejo masivo de datos sea más común de lo que creemos. El *Big Data* es un término que define la recopilación de datos masivos y completos que, por su tamaño, complejidad, variabilidad y velocidad de crecimiento, no podrían ser procesados mediante herramientas convencionales.

Uno de los campos que más se ha explorado es el análisis de datos, proceso que se dedica a examinar un conjunto de datos con el objetivo de concluir sobre la información para poder tomar decisiones, o simplemente *"...ampliar los conocimientos sobre diversos temas, a través de la inspección, limpieza, transformación y modelamiento de estos datos que permiten actualmente a las empresas y organizaciones operar de manera más eficaz"* [2]. De forma más precisa, Charles Judd y Gary McCleland en 1989 describieron una de las funciones del análisis de datos: *"...los datos se coleccionan y analizan para indagar en cuestiones, probar conjeturas o refutar teorías"* [3].

Con una fortuna valuada en más de 187,5 mil millones de dólares Jeff Bezos es el hombre más rico del mundo según señala la revista Forbes [4] al día de hoy. Su patrimonio se debe mayoritariamente a la empresa que él mismo fundó en 1994, Amazon. Son muchos los factores que hicieron de esta empresa una de las más grandes en la actualidad, no obstante la recolección de datos y el posterior análisis de estos jugaron un rol decisivo en este crecimiento. El fundador de Amazon, en relación a su "obsesión por los clientes", dice que la primera prioridad de la firma es *"descubrir lo que (los clientes) quieren, lo que es importante para ellos"*. Una obsesión que se entiende, luego de leer lo que James Thomson, uno de los ex ejecutivos de Amazon comenta: *"Cada oportunidad para interactuar con un cliente es otra oportunidad para recolectar datos"* [5].

El trabajo con datos, en los últimos años, se ha introducido de manera creciente en distintos campos de trabajo. Uno de ellos ha sido la industria del deporte, que se ha percatado de la posibilidad de entendimiento que se puede llegar a tener usando y analizando los datos que se generan en cualquier instancia deportiva. Como lo indica la escuela de negocios digitales *Three Points* en su página: *"durante un partido de fútbol se pueden llegar a capturar alrededor de ocho millones de datos. Sin embargo, el ojo humano solo es capaz de retener el 30 % de esa información."*[6]

En el año 1970 Bill James escritor, historiador y estadístico del béisbol estadounidense empieza a analizar los registros históricos de los jugadores de las principales ligas americanas y a través de la evidencia objetiva poder crear un registro eficiente de los eventos que suceden dentro del campo de juego. Desde entonces, muchos son los deportes que han tomado medidas con respecto al uso de los datos y han podido sacar provecho de las posibilidades del Big Data para mejorar los análisis predictivos que ayudan a diferentes áreas del deporte, desde prevenir lesiones, cambiar las tácticas de juego, mejorar el rendimiento o encontrar errores en el funcionamiento [6].

Desde que Bill James comenzó con la recopilación de datos y posterior análisis de estos en el deporte, han sido muchos los casos, algunos muy famosos, donde han utilizado esta estrategia para mejorar el rendimiento tanto financiero como deportivo de los clubes o atletas. Fue el caso de los Oakland Athletics de la Major League Baseball (MLB) de los Estados Unidos, historia que Michael Lewis describe en su libro el año 2003 [7], quienes en un momento de precaria situación tanto deportiva como económica deciden invertir en un economista de la Universidad de Yale llamado Peter Brand para que los ayudase con la contratación de jugadores. Peter, fiel a su esencia basada en los números, modifica el proceso de contratación de jugadores (*scouting*) y crea una metodología innovadora de evaluación netamente estadística que se antepone a la intuición y expertiz de los antiguos encargados de este proceso. Esta nueva forma resulta en un éxito para el equipo, y exhibe una estrategia vanguardista en esa época para grandes franquicias en distintos deportes, quienes han ido transformando y profesionalizando este tipo de procesos.

Es mucha la literatura que se puede encontrar con respecto al uso de los datos en el deporte y específicamente en el fútbol. Hughes et al. [8] en la primavera del 2011 aprovecharon la oportunidad única de la concentración de un gran número de analistas de rendimiento reunidos para discutir el problema de la variación de indicadores de rendimiento clave (KPI) en el fútbol entre un entrenador y otro. Con ello, lograron definir 5 diferentes categorías relacionadas a ámbitos: fisiológicos, tácticos, técnicas defensivas, técnicas ofensivas y psicológicos, para medir el rendimiento de los futbolistas en 7 posiciones distintas del campo: arquero, defensa central, defensa lateral, mediocampista central, mediocampista ofensivo, mediocampista exterior y delantero. En general, estos KPI eran diferentes de una posición a otra dentro del equipo. Los KPI para los jugadores de campo eran muy similares, difiriendo solo en su orden de importancia según la posición en la que se encuentran.

El 8 de Octubre del 2015 el actual entrenador de fútbol alemán Jürgen Klopp fue presentado en uno de los clubes ingleses con más trayectoria de Europa, el Liverpool F.C.. Como lo describe el New York Times en su artículo publicado el 2019 y titulado: *"El arma secreta del Liverpool: el análisis de datos"* [9], fue el propio director de investigación del Liverpool,

Ian Graham doctorado en Física Teórica por la Universidad de Cambridge, quien le revela que: *"no fue necesario ver ningún partido del ex-equipo de Klopp, el Borussia Dortmund, para determinar y validar su contratación como nuevo director técnico del club, a pesar de los malos resultados que había obtenido Klopp en la temporada anterior con el B. Dortmund"* [9]. La decisión de Graham no fue para nada errada, de hecho el Liverpool luego de muchos años de ausencia en la élite del fútbol, ha logrado renacer futbolísticamente hablando y ha vuelto a ganar el torneo inglés (Premier League 2019-2020) que se les hacía esquivo desde principios de los '90 y volvió a ganar el máximo torneo de clubes internacional (Champions League 2018-2019). Actualmente, el equipo de analistas del Liverpool no solo ayudan al club en la contratación de jugadores, sino también son un apoyo a la dirección técnica del equipo, entregan informes deportivos tanto del equipo como de los rivales y apoyan a la creación de la estrategia de los partidos.

Ramón Rodríguez Verdejo, conocido deportivamente como *"Monchi"* es un ex-arquero del club de fútbol español Sevilla F.C.. Actualmente es el Director General Deportivo del club y es el encargado principal de la venta y contratación de nuevos jugadores. Rodríguez, lleva 20 años realizando esta actividad y como lo muestra el New York Times en su artículo publicado en 2019 y titulado: *"La ciencia del mercado de transferencias del fútbol"* [10]: *"...la fama en este puesto no ha sido gracias al azar, ha logrado grandes fichajes y ventas que han mantenido al club en una posición privilegiada en Europa y con números azules tanto desde una perspectiva económica como deportiva. (...) este trabajo lo ha llevado a cabo, en conjunto con su equipo, a través de un seguimiento de los potenciales jugadores que el club podría contratar"*. Son muchos los parámetros e indicadores que *"Monchi"* le entrega al club para discutir y decidir sobre los nuevos fichajes. Sin embargo, en muchas ocasiones se determina fichar a un jugador sólo considerando sus datos estadísticos de rendimiento.

Este tipo de decisiones de algunos equipos logra apoyar el proceso de contratación de jugadores en base a lo que el club necesita deportivamente hablando. Hay muchas formas de ir desarrollando y especializando esta actividad. El intentar cuantificar de una forma estadística y basado en los datos generados el efecto de la inclusión de un nuevo jugador en el esquema táctico del equipo, o incluso en el reemplazo de un jugador nuevo por otro perteneciente a la plantilla del club, puede ser una herramienta importante a la hora de tomar decisiones en relación a quién contratar. Considerando que los datos están y acompañado de una inversión por parte de los clubes, no sería difícil el acceso y análisis propio de éstos.

Actualmente, al término e inicio de nuevas temporadas deportivas comienza el periodo de traspaso de jugadores, en el que cada equipo puede incorporar nuevos jugadores provenientes de otros clubes (de la misma liga nacional o de otras ligas extranjeras) a su plantilla. Las personas encargadas de elegir a los potenciales futbolistas a ser contratados son analistas (también llamados *scouts*). La metodología que se ocupa en la mayoría de los equipos se basa principalmente en la visualización de vídeos de partidos completos de distintos equipos y ligas accesibles para los *scouts*, en dónde a través de su propio juicio van determinando jugadores potenciales a los que se les realiza un seguimiento durante la temporada previa al período de fichaje en cuestión. Esto genera un sesgo de selección al determinar qué jugador potencial podría ser ideal para el club, ya que se visualizan pocos partidos y al determinar qué jugadores se comenzarán a observar, se deja de seguir a jugadores que podrían ser buenas opciones. Además, en la elección de los potenciales jugadores a incluir se suele caer en

un sesgo sólo por el hecho de que el partido que se observó del jugador era anormal a su desempeño regular, generando una pérdida de tiempo en su análisis posterior.

Es por lo anterior que la decisión de a quién contratar no se hace nada fácil para quien tiene que decidir, considerando que las sumas de dinero por la transacción cada vez son más altas; de hecho, de los 30 fichajes más caros en la historia del fútbol 29 de ellos se realizaron en los últimos 10 años, por ende, la inversión que hace el club con la contratación de un jugador puede dar pie para estar bien preparado tanto deportiva como económicamente durante la temporada. Esto se revela con el desempeño del jugador, si se apostó por contar con los servicios de un jugador y éste obtiene un alto rendimiento durante la(s) temporada(s), otros equipos estarían interesados en contratarlo, dándole la opción al club de venderlo a un precio mayor al que lo compró y generando un diferencial de ganancia. Por otro lado, si el jugador no rinde como se esperaba que lo hiciese, esto se traduce en una pérdida de dinero para el equipo, pues el jugador disminuye su valor y no podrá recuperar lo invertido en él, en desmedro de lo que podría haber generado con la contratación de un jugador que si cumpliese con el rendimiento esperado.

El objetivo general de esta tesis es cuantificar el efecto de un jugador en un equipo de fútbol, medido a través de su aporte en la probabilidad de salir campeón del club en el torneo y que se pueda utilizar como una herramienta de apoyo al proceso de contratación de jugadores (*scouting*) basado en determinar las posiciones en las que el equipo está más débil y proponer nuevos jugadores que puedan satisfacer esas carencias deportivas. Con el modelo de simulación propuesto más adelante y basado en características cuantitativas del desempeño de los jugadores se puede determinar qué jugadores podrían aumentar la probabilidad de campeonar del equipo y así evidenciar de forma estadística si son potenciales jugadores a contratar. Así, se espera quitar los sesgos de la actual metodología y reducir el espectro de jugadores a observar para los *scouts*.

En el siguiente capítulo de esta tesis se detallarán los datos usados en esta investigación, la fuente de donde se extraen y cómo se obtienen. También, cómo éstos fueron reestructurados para facilitar el trabajo que se desarrolla y su respectivo análisis exploratorio de datos (EDA). En el capítulo posterior se detallan las técnicas y conceptos teóricos que se utilizan para el modelo de simulación y por qué estas son, en principio, una buena opción para lo que se busca desarrollar. Para luego, dar paso al capítulo donde se explica el modelo de simulación realizado y cómo es que se construye la simulación desde un evento en un partido hasta el campeonato completo. Por último, se muestra un capítulo con los resultados de esta tesis, las dificultades y mejoras que se pueden hacer y las conclusiones que se obtuvieron de esta investigación. Además, se pueden observar distintos gráficos y tablas en la sección de apéndice que complementan la información que se expone en los distintos capítulos.

Capítulo 2

Datos

En este capítulo se detallan y describen los datos usados en la investigación y desde dónde y cómo se extraen. También, cómo es que se reestructuran en función del trabajo que se realiza y por último, se muestran algunos hallazgos interesantes que se obtienen directamente al revisar el conjunto de datos previamente estructurados.

2.1. Datos

2.1.1. Extracción de los datos

Los datos utilizados en este trabajo fueron recolectados por la compañía *Wyscout* [11] y contiene la estructura de datos completa del torneo de fútbol inglés Premier League 2017-2018. Estos datos fueron publicados por Luca Pappalardo [12] y Emanuele Massucco a través de la página web "Figshare", con el objetivo que pudieran estar disponibles para toda la comunidad científica. Esta es la colección de datos abierta publicada más grande de registros de fútbol y utilizada el 2019 en el *Soccer Data Challenge*[13]. *Wyscout* es una empresa italiana que apoya la exploración del fútbol, el análisis de partidos y la dinámica de transferencias. La empresa se fundó en Génova, Italia, en 2004, y está ubicada en Chiavari desde enero del 2008 [14].

El acceso a los datos, entrega información general asociada a las plantillas de los equipos, eventos de cada uno de los partidos e información del rendimiento deportivo de los jugadores y técnicos de los 20 clubes pertenecientes a la Premier League 2017-2018. Se puede acceder fácilmente a la estructura de datos que *Wyscout* ofrece desde su página, en su versión gratuita o pagada, según los permisos que ellos mismos asignen a sus usuarios. Por su parte, la profundidad de acceso a la información también se limita según los permisos que se asignen al cliente, ofreciendo así datos agrupados sobre jugadores, partidos o equipos, de fácil acceso para todo publico o el detalle de cada uno de los eventos de un jugador, partido o equipo si es que se obtienen los permisos de acceso para esos datos desagrupados.

Estos permisos permiten al acceso a la API (*Application Programming Interface*) de *Wyscout* desde donde se pueden acceder a los datos a través de consultas tal y como se obtienen de la publicación desde donde se extrajeron los datos de esta tesis.

2.1.2. Contexto de los datos

La Premier League 2017-2018 jugada entre el 11 de agosto del 2017 y el 13 de mayo del 2018 fue la vigesimosexta temporada de la máxima división de fútbol inglesa, desde su creación en 1992. Un total de 20 equipos participan en la competición, incluyendo 17 equipos de la temporada anterior y 3 ascendidos de la English Football League Championship 2016-17 (segunda división de fútbol inglés). 1.018 goles se convirtieron durante los 380 partidos del torneo, con un promedio de 2,68 goles por encuentro. El campeón fue el Manchester City, estableciendo un récord de obtención de 100 puntos, siendo el primer equipo en lograr esa marca en la historia de la liga inglesa [15]. El premio al mejor jugador y al goleador del torneo se concedió al jugador del Liverpool F.C. Mohamed Salah, quien anotó 32 goles convirtiéndose además, en el máximo goleador del torneo en toda su historia. Ian Graham, estadístico del Liverpool, argumentaba que uno de los motivos por los que Salah llegó a esa cifra se debió a que Roberto Firmino, compañero de equipo de Salah, creó con sus pases más goles esperados (xG, *Expected Goals*) que cualquier otro delantero del mundo [16].

Tabla 2.1: Tabla de posiciones Premier League 2017-2018

Pos.	Equipo	Pts.	PJ	G	E	P	Dif.
1	Manchester City	**100**	38	32	4	2	+79
2	Manchester United	**81**	38	25	6	7	+40
3	Tottenham Hotspur	**77**	38	23	8	7	+38
4	Liverpool	**75**	38	21	12	5	+46
5	Chelsea	**70**	38	21	7	10	+24
6	Arsenal	**63**	38	19	6	13	+23
7	Burnley	**54**	38	14	12	12	-3
8	Everton	**49**	38	13	10	15	-14
9	Leicester City	**47**	38	12	11	15	-4
10	Newcastle United	**44**	38	12	8	18	-8
11	Crystal Palace	**44**	38	11	11	16	-10
12	Bournemouth	**44**	38	11	11	16	-16
13	West Ham United	**42**	38	10	12	16	-20
14	Watford	**41**	38	11	8	19	-20
15	Brighton & Hove Albion	**40**	38	9	13	16	-20
16	Huddersfield Town	**37**	38	9	10	19	-30
17	Southampton	**36**	38	7	15	16	-19
18	Swensea City	**33**	38	8	9	21	-28
19	Stoke City	**33**	38	7	12	19	-33
20	West Bromwich Albion	**31**	38	6	13	19	-25

Con respecto a la contratación de jugadores, las transacciones más caras en ambos periodos de transferencias fueron: Romelu Lukaku, quien fue comprado durante el mercado de verano por el Manchester United F.C. al Everton F.C. por la suma de €84.700.000 y Virgil van Dijk, comprado por el Liverpool F.C. durante el mercado de invierno al Southampton F.C. de inglaterra por €78.800.000.

2.1.3. Estructura inicial

Para poder acceder a los datos desagregados que ofrece *Wyscout* es necesario realizar una consulta directamente en la API que ofrecen para sus clientes. En ella se encuentra la información completa del campeonato y de cada uno de los eventos de los 380 partidos, además de información detallada de los clubes, jugadores, técnicos y árbitros y el rendimiento correspondiente de cada uno de ellos medido a través de múltiples parámetros.

Como resultado a las consultas en la API se obtienen paquetes estructurados en forma de "árbol" similar a un archivo tipo ".JSON". En donde se va ramificando la información desde lo más general a lo más específico. Cada paquete de datos corresponde a distinta información recopilada por *Wyscout*:

1. **Teams**: contiene información respectiva a todos los equipos del campeonato, incluyendo el nombre oficial de cada club y un número único "id" asignado para identificarlo.

2. **Players**: contiene información personal de cada uno de los jugadores del torneo incluyendo su nombre completo, fecha de nacimiento, nacionalidad, peso y altura, además de datos relacionados al deporte como la posición en la que se desempeña habitualmente, el pie con el que juega, "id" del equipo al que pertenece y un "id" único de identificación personal.

3. **Coaches**: contiene información personal de cada uno de los técnicos de los clubes, como su nombre, nacionalidad y fecha de nacimiento, además del "id" del equipo que dirige y un "id" único de identificación personal.

4. **Matches**: este paquete es uno de los que contiene mayor información relevante para esta tesis, incluye información agrupada relativa al partido: fecha del partido, "id" único asociado al partido, equipos participantes, jornada del torneo, resultado y localía. Además, contiene detalles de ambos equipos, como: formación inicial, banca, sustituciones, tarjetas amarillas, tarjetas rojas y goles.

5. **Events**: contiene el detalle de cada uno de los eventos del partido, ordenados en forma cronológica: pases, duelos, tiros al arco, faltas, tarjetas amarillas, tarjetas rojas, penales, balones fuera del campo, entre otros. Es el paquete con mayor información y detalle, donde para cada uno de los eventos se especifica el jugador y equipo involucrado, el tiempo relativo al incio del partido cuando ocurre el evento, la zona del campo en la que ocurre y si el evento fue considerado como exitoso o no para el jugador que realiza la acción.

Es con la información contenida en estos paquetes donde se basa mayoritariamente el modelo de simulación que se detalla más adelante. Principalmente los paquetes de "*matches*" y "*events*" contienen la mayor cantidad de datos, los que servirán posteriormente como *inputs* para generar la simulación de cada uno de los eventos y finalmente del torneo completo, además servirán para evaluar y medir errores del modelo. Cabe destacar también, que según lo que *Wyscout* declara oficialmente, existe un error menor al 1 % en la recolección de estos datos. La empresa afirma que los datos son almacenados a tiempo real por sus trabajadores, sin embargo son revisados e inspeccionados nuevamente para lograr minimizar el error al momento de su recopilación.

2.2. Reestructuración de los datos

Para facilitar el trabajo posterior y dado que la estructura inicial contenida en los paquetes es compleja para el tratamiento y análisis de los datos, se decide reestructurarlos a través de tablas ordenadas. Así, a partir de los paquetes anteriormente detallados se generan las siguientes tablas con las que se trabaja:

1. **Tabla de equipos**: esta tabla de datos contiene la información relativa a los 20 equipos, las columnas son: "id" del equipo y nombre del equipo.

2. **Tabla de jugadores**: esta tabla contiene la información relacionada a cada uno de los 515 jugadores participantes del torneo, las columnas son: "id" del jugador, nombre y apellido, posición donde juega, "id" del equipo donde juega y nombre del equipo que pertenece.

3. **Tabla de partidos**: esta tabla de datos contiene la información relativa a los 380 partidos, las columnas son: fecha y hora del partido, "id" del equipo local, "id" del equipo visitante, nombre del equipo local, nombre del equipo visita, goles del equipo local, goles del equipo visitante.

4. **Tabla de eventos**: en esta tabla de datos se encuentra toda la información relativa a cada uno de los eventos de cada partido del torneo. Cada fila de esta tabla representa un evento en particular, ya sea un pase, duelo, tiro, tarjeta, gol, entre muchos otros. Cada una de las columnas representan: "id" único de identificación del evento, el nombre del evento, el nombre del sub-evento que detalla información adicional del tipo de evento, el minuto relativo al comienzo del partido cuando ocurre, el "id" asociado al partido, el "id" asociado al jugador involucrado en el evento y su nombre, la posición en el "eje x" y "eje y" donde comienza el evento y la posición en el "eje x" y "eje y" donde termina el evento, una columna de valor binario que asigna el valor 1 si el equipo asociado al evento es local o 0 si no lo es, una columna de valor binario que asigna el valor 1 si es que el evento es realizado con éxito por el jugador que lo realiza o 0 si no lo es y por último, el resultado actual del partido al momento que ocurre aquel evento.

Además, se construyen diccionarios que posteriormente facilitan el trabajo y la búsqueda de información más eficiente para la simulación. Los diccionarios se detallan a continuación:

1. **Partidos/Equipos**: se le asigna una llave única asociada a un partido y el diccionario muestra el "id" del equipo local y visita.

2. **Partidos/Alineaciones**: se le asigna una llave única asociada a un partido y un equipo, y el diccionario muestra la alineación inicial, banca y sustituciones.

3. **Partidos/Nombre**: se le asigna una llave única asociada a un partido y el diccionario muestra el nombre de los equipos local y visita.

2.3. Análisis exploratorio de datos

Con la reestructuración de los datos se hace más fácil el manejo y la relación entre ellos. Si bien la estructura en forma de "árbol" es muy ordenada, cuando se trata de muchos niveles de información se torna mas complejo localizar el dato que se busca. Por otra parte, la estructura de tablas interrelacionadas permite por un lado, que esa búsqueda sea más eficiente, y también que la visualización sea más limpia.

De esta forma, las tablas de equipos, jugadores, partidos y eventos conversan entre ellas a través de los indicadores únicos "id" que asocian un número natural exclusivo para cada uno de los equipos, jugadores, partidos o eventos.

2.3.1. Distribución de los eventos

Como se detalla en la sección anterior, la tabla de eventos detalla cada uno de los sucesos que tienen lugar en un partido de fútbol. Los eventos vienen agrupados y organizados en 10 diferentes categorías, éstas a su vez se fraccionan en diferentes sub-categorías que en conjunto detallan las 36 sub-categorias de eventos que existen.

En la siguiente tabla se detallan las 10 categorías de eventos con la cantidad de apariciones en la tabla de eventos y el porcentaje en relación al total de eventos:

Tabla 2.2: Distribución del total de los eventos.

Nombre evento	Cantidad	% del total
Pass	328.657	51,1 %
Duel	176.688	27,5 %
Others on the ball	51.085	7,9 %
Free Kick	36.423	5,7 %
Interruptions	27.535	4,3 %
Foul	8.318	1,3 %
Shot	8.451	1,3 %
Save attempt	3.349	0,5 %
Goalkeeper leaving line	1.266	0,2 %
Offside	1.558	0,2 %
Total	**643.330**	**100 %**

Se observa que el evento con mayor ocurrencia durante el transcurso de la totalidad del torneo son los pases, con más del 50 % del total de eventos. Luego, el evento relacionado a los duelos entre un jugador de un equipo y otro son más del 25 % de los eventos. Por último, se tiene que la categoría de faltas y tiros tienen cada una solo un 1,3 % del total de eventos.

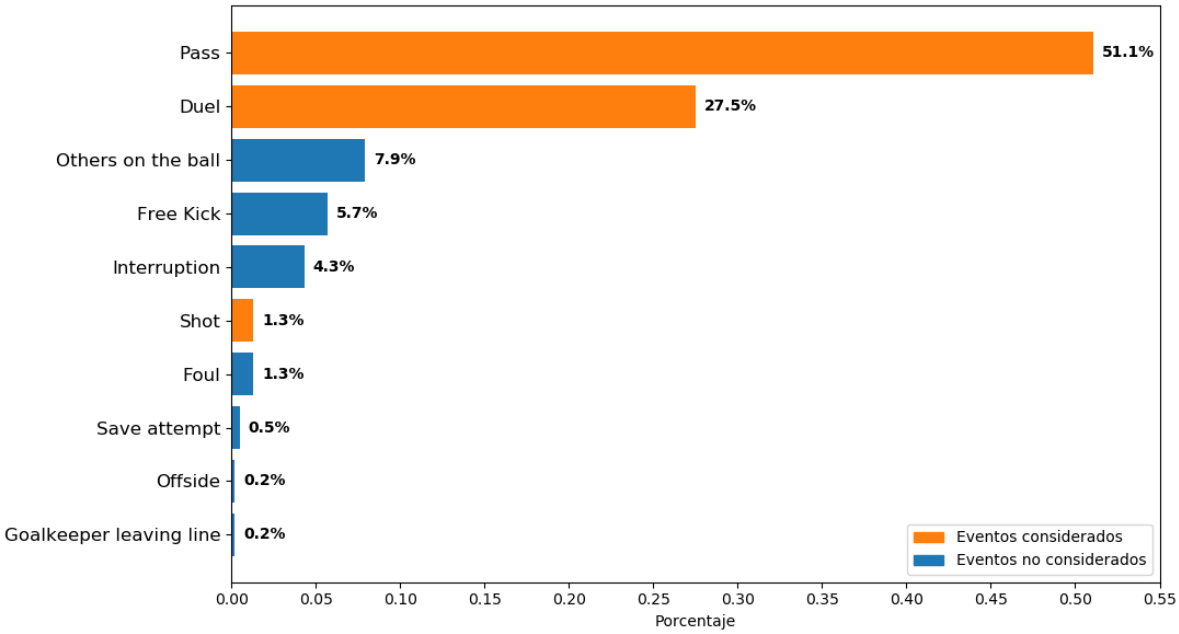

Figura 2.1: Porcentaje del total de eventos por categoría. Separados por eventos a considerar en la simulación.

Con el objetivo de simplificar la simulación, y como se muestra en la Figura 2.1 se consideran los eventos más relevantes del juego. En primer lugar, aquellos que ocurren con mayor frecuencia, como son los eventos de **Pases** (*Pass*) y **Duelos** (*Duels*) que tienen que ver con el traslado y la disputa del balón, y que entre ambos abarcan más del 75 % de los eventos que ocurren en un partido de fútbol. Y en segundo lugar, el evento de **Tiros** (*Shot*) que da pie a representar los goles, el evento con mayor importancia en el fútbol.

Si bien, se dejan de considerar eventos importantes en el desarrollo de un partido para la simulación como son las **Faltas** (*Fouls*), **Tiros libres** (*Free Kicks)* o **Fueras de juego** (*Offside*) estos no son tan relevantes en las incidencias de un partido, y pueden ser considerados como parte de otro grupos de eventos. Por ejemplo, las faltas pueden ser omitidas y consideradas como una detención del juego y un posterior pase del equipo contrario (equipo que recibe la infracción) o los tiros libres y tiros de esquina, pueden ser asociados al grupo de eventos de pases, ya que no son más que un traslado del balón desde una zona de la cancha a otra en el caso del tiro de esquina o añadidos al grupo de eventos de tiros el caso de un tiro libre que tiene como dirección el arco contrario.

2.3.2. Eventos por equipo

Como cada uno de los eventos está asociado a un equipo, es interesante un análisis simple sobre aquellos equipos que son partícipes de una gran cantidad de eventos, en contraste con aquellos que tienen menos. Más aún, como vimos en la tabla anterior, más del 75 % de los eventos son pases o duelos, es decir, se puede inferir el estilo de juego de un equipo o simplemente la posesión, que es el tiempo de acciones donde está involucrado un equipo que tiene el balón con respecto al total, sólo analizando un poco los datos.

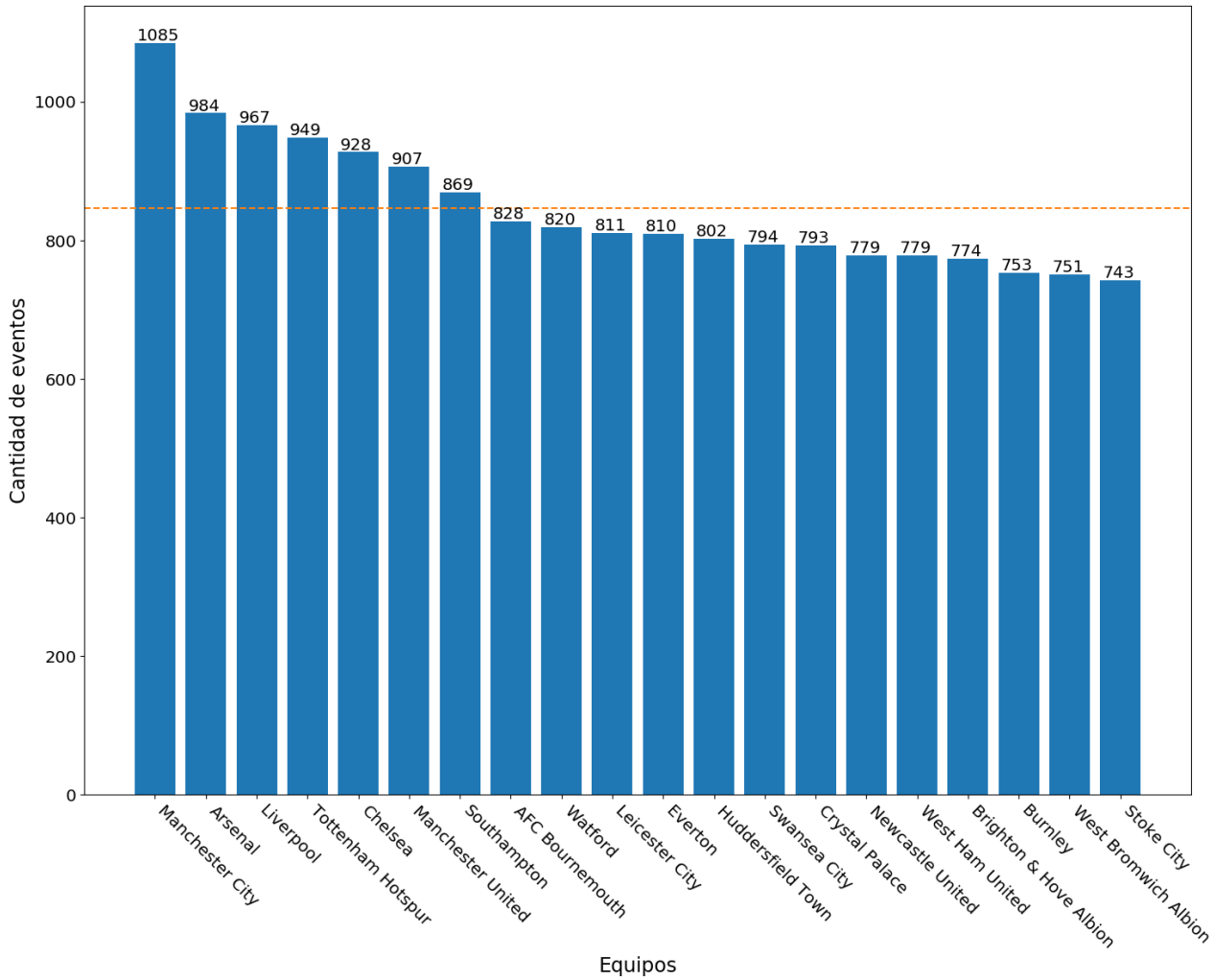

Figura 2.2: Promedio de eventos totales de los 20 equipos en los 38 partidos que juega cada uno en el torneo.

Como se muestra en la Figura 2.2 existe una variación entre la cantidad de eventos en que participa un equipo en promedio durante un partido. Esto puede hablar mucho del juego de cada uno de ellos, de hecho seis de los siete equipos que se encuentran sobre el promedio (linea punteada) terminaron el torneo en las primeras seis posiciones: Manchester City, Manchester United, Tottenham Hotspur, Liverpool, Chelsea y Arsenal (ver Tabla 2.1).

Puede ser interesante también, observar la diferencia en la cantidad de eventos que un equipo participa en situación de local, en contraposición a cuando éste se encuentra de visita. Muchas veces se dice que un equipo suele tener más el balón o tener un rol más participativo en las acciones del juego cuando juega en su propio estadio, con el público a su favor y las condiciones que está acostumbrado. A diferencia de lo que es jugar en situación de visita, con el público en contra y situaciones poco comunes.

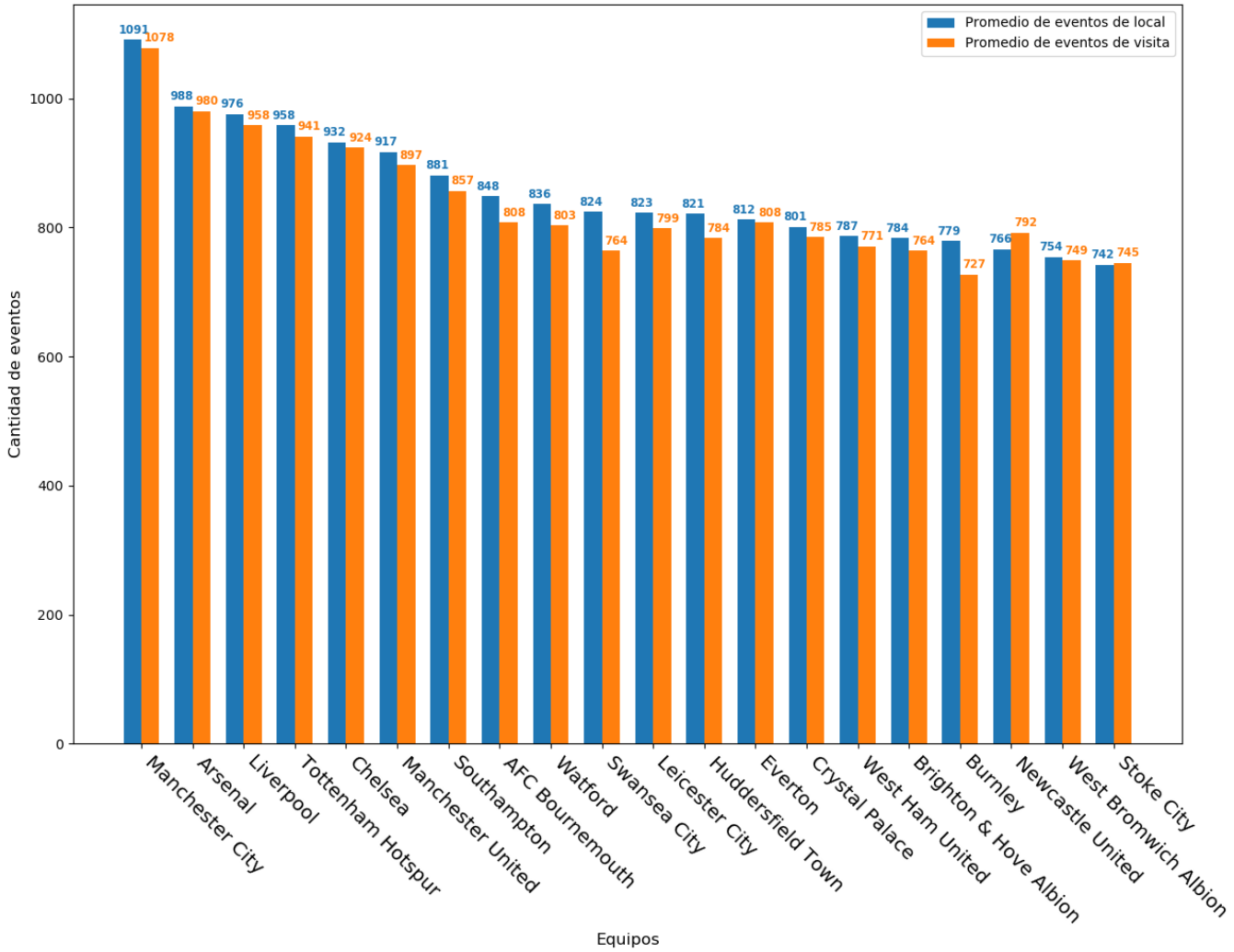

Figura 2.3: Promedio de eventos de local y de visita de los 20 equipos del torneo.

Como se observa en la Figura 2.3 y en concordancia con lo expuesto anteriormente, en su mayoría los equipo participan de más eventos en promedio cuando se encuentran de local en comparación a cuando son visita. Sin embargo, esta diferencia no es estadísticamente significativa en la mayoría de los equipos, en general todos los equipos mantienen un promedio parecido de participación en los eventos de sus partidos tanto de local como de visita. Son solo dos los equipos que tienen más eventos en promedio de visita que de local: el Newcastle United y el Stoke City.

Se puede desprender entonces de estas dos últimas figuras que la participación en los eventos sí tiene mucha relación con el juego que tienen los equipos y como ellos terminan posicionados al final del torneo, sin embargo no existe una diferencia significativa entre estos promedio de participación en los eventos entre los partidos de local y los partidos de visita.

2.3.3. Goles por equipo

Sin duda alguna, el evento más importante en un partido de fútbol son los **Goles**. La realización de ellos y el evitar que el equipo contrario haga un gol en el arco propio es el sentido más global de este juego. Por ende, aquél equipo que convierte más goles que su rival gana el partido y obtiene +3 puntos, en el caso que ambos equipos marquen la misma cantidad de goles se considera empate y cada uno obtiene +1 punto, por último el equipo perdedor recibe +0 puntos. Así, es trivial pensar que aquél equipo que convierte más goles de los que recibe durante el torneo obtiene una mayor cantidad de puntaje, por lo tanto debería estar ubicado en una mejor posición en la tabla al finalizar el torneo.

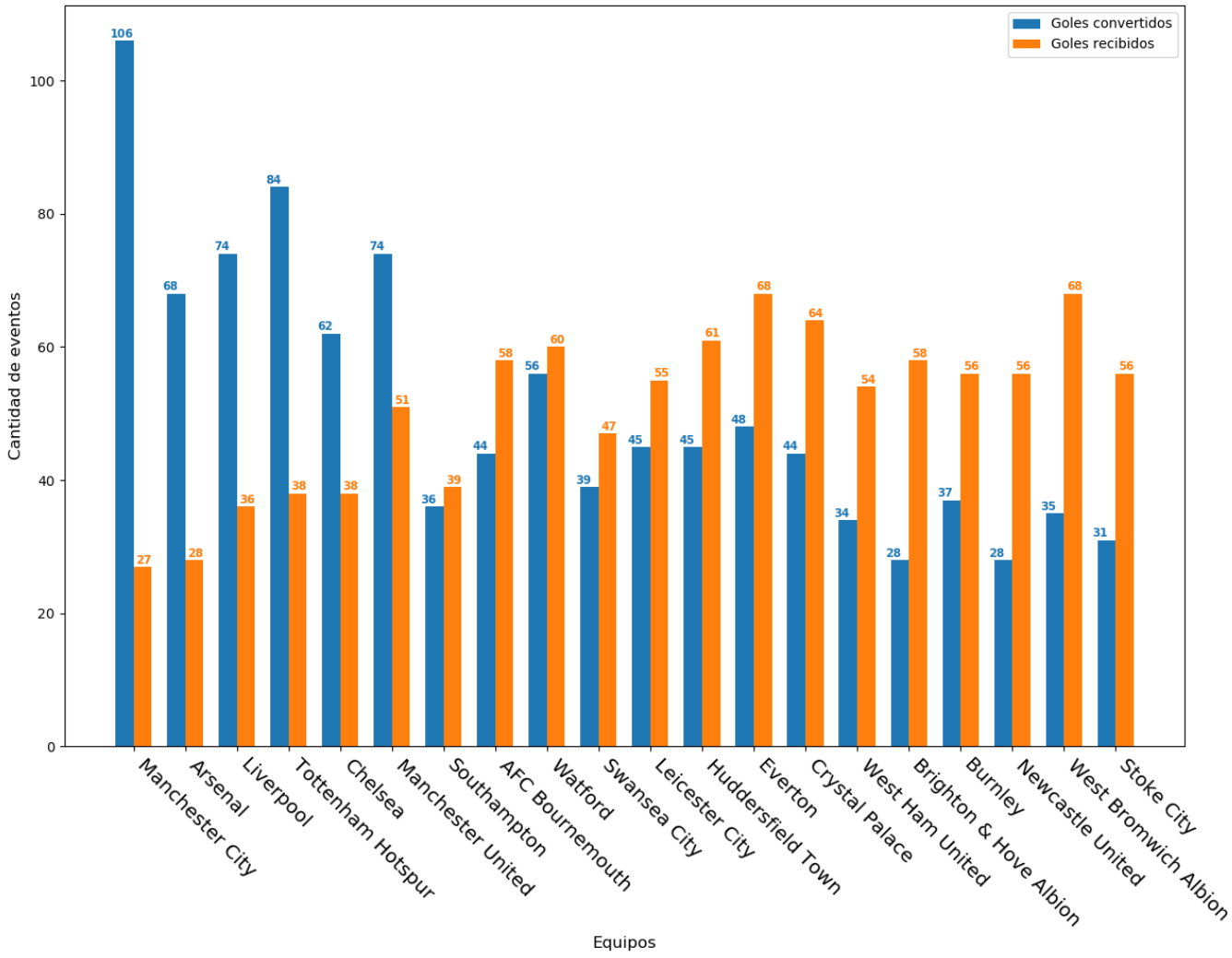

Figura 2.4: Goles anotados y recibidos por cada uno de los 20 equipos del torneo.

En la Figura 2.4 se puede observar claramente que aquellos equipos que tienen una mayor diferencia positiva entre goles convertidos y goles recibidos son justamente aquellos que terminan en los primero puestos al finalizar el torneo: Manchester City, Arsenal, Liverpool, Tottenham Hotspur, Chelsea y Manchester United. En cambio, aquellos que tienen una diferencia negativa entre goles convertidos y recibidos son justamente los equipos que ocupan las últimas posiciones de la liga (ver Tabla 2.1).

2.3.4. Minutos y eventos por jugador

Antes de analizar la participación de los jugadores en los eventos, es bueno tener en consideración que esta participación está directamente relacionada con los minutos jugados por cada uno de los deportistas.

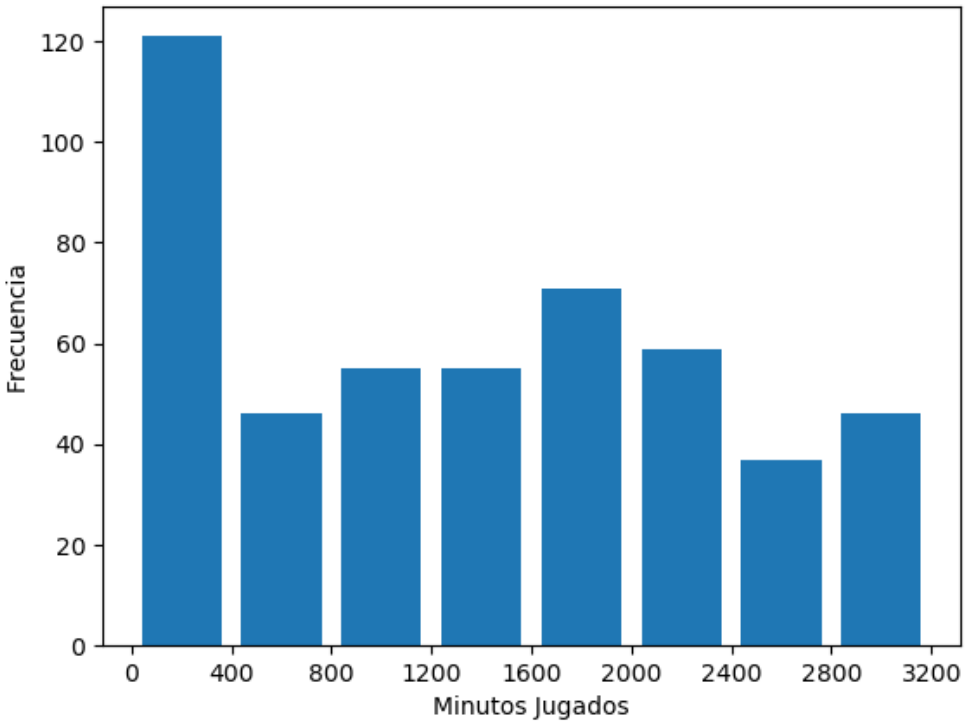

Figura 2.5: Histograma de los minutos jugados durante todo el torneo por los 515 jugadores.

En la Figura 2.5 se observa la frecuencia de participación en minutos jugados durante todo el torneo por los 515 jugadores. Casi un cuarto de los jugadores no jugó más de 400 minutos (cerca de cuatro partidos), esto tiene relación con la repetición de las alineaciones que tienen los equipos durante el torneo, dejando así gran parte de los encuentros a jugadores en la banca. En el Anexo A.1 se puede encontrar una tabla con los jugadores con mayor y menor cantidad de minutos jugados en el torneo. En el Anexo A.2 se puede observar una tabla con los jugadores con mayor y menor cantidad de participación en algún evento en el torneo.

Capítulo 3

Marco Teórico

En el siguiente capítulo se detallan las técnicas y conceptos teóricos que se utilizan para el modelo de simulación y por qué éstas son, en principio, una buena opción para el modelo que se detalla en el capítulo posterior.

3.1. Cadenas de Markov

En la teoría de probabilidad, se conoce como Cadena de Markov o modelo de Markov a un tipo especial de proceso estocástico discreto en el que la probabilidad de que ocurra un evento depende solamente del evento inmediatamente anterior. Esta característica de "falta de memoria" recibe el nombre de propiedad de Markov. Las Cadenas de Markov adoptan su nombre del matemático ruso Andréi Markov (1856-1922), que las introdujo en 1906 [17].

3.1.1. Procesos estocásticos

Un proceso estocástico es un fenómeno aleatorio que surge en una sucesión de eventos que se desarrollan en el tiempo de una manera controlada por medio de leyes probabilísticas. Es decir, está regulado bajo la incertidumbre que generan las probabilidades de ocurrencia de cada uno de los sucesos. Así, un proceso estocástico es una familia de variables aleatorias que proporcionan una descripción de la evolución de un determinado fenómeno físico a través del tiempo. A los posibles valores que puede tomar la variable aleatoria se le llaman estados por lo que el estado puede ser de un espacio discreto o continuo. Donde $X(t) \in E$ es el estado del proceso en el instante t dentro del conjunto de estado E y T es el conjunto de índices del proceso.

$$\{X(t) \in E, t \in T\} \tag{3.1}$$

3.1.2. Propiedad de Markov

Una Cadena de Markov es un proceso estocástico, pero se diferencia de un proceso estocástico general en que una Cadena de Markov debe ser "sin memoria". Es decir, (la probabilidad de) acciones futuras no depende de los pasos que condujeron al estado presente. Esto se

llama **propiedad de Markov**. Si bien la teoría de las Cadenas de Markov es importante precisamente porque muchos procesos "cotidianos" satisfacen la propiedad de Markov, existen muchos ejemplos comunes de propiedades estocásticas que no satisfacen la propiedad de Markov. Así, esta propiedad queda definida como se muestra a continuación:

Para cualquier entero positivo n y posibles estados $i_0, i_1, ..., i_n \in E$ de las variables aleatorias, se tiene que, la distribución de probabilidad P del estado $X_{n-1} = i_{n-1}$ al estado $X_n = i_n$ es:

$$P(X_n = i_n | X_0 = i_0, X_1 = i_1, ..., X_{n-1} = i_{n-1}) = P(X_n = i_n | X_{n-1} = i_{n-1}) \tag{3.2}$$

En otras palabras, el conocimiento del estado anterior es todo lo que se necesita para determinar la distribución de probabilidad del estado actual.

3.1.3. Cadenas de Markov en tiempo discreto

Una Cadena de Markov es un proceso estocástico que experimenta transiciones de una estado a otro según ciertas reglas probabilísticas. La característica principal de una Cadena de Markov es que no importa cómo llegó el proceso al estado actual, los posibles estados futuros dependen sólo del estado actual y tienen probabilidades fijas. En otras palabras, la probabilidad de pasar a cualquier estado en particular depende únicamente del estado actual y del tiempo transcurrido. El espacio de estados, o conjunto de todos los estados posibles puede ser cualquier cosa: letras, números, condiciones climáticas, puntajes de un torneo, eventos de un partido de fútbol o valores de acciones financieras.

Las Cadenas de Markov en tiempo discreto pueden modelarse mediante máquinas de estado finitos, que van pasando de un estado a otro a través de ciertas probabilidades definidas. Son utilizadas ampliamente en ámbitos económicos, teoría de juegos, teoría de colas, genética, finanzas, entre muchas otras aplicaciones. Si bien es posible describir las Cadenas de Markov con cualquier tamaño de espacio de estados, la teoría inicial y la mayoría de las aplicaciones se centran en casos con un número finito (o numerablemente infinito) de estados.

Para entenderlo de mejor manera, puede ser útil graficarlo como redes de nodos. De esta forma, cada uno de los nodos representa un posible estado de la cadena y cada arco que une dos nodos caracteriza la probabilidad de transición de un estado (nodo) a otro, como lo muestra la Figura 3.1

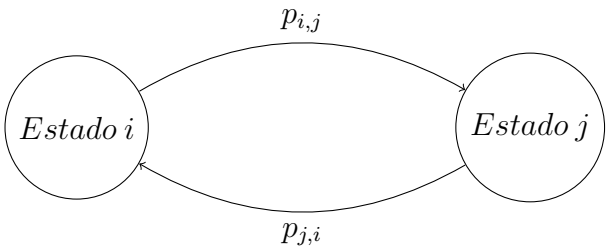

Figura 3.1: Ejemplo de grafo de una cadena de Markov.

3.1.4. Matrices de transición

Una matriz de transición $P_t \in NxN$ para la cadena de Markov $\{X\}$ en el tiempo t es una matriz que contiene información sobre la probabilidad de transición entre estados. En particular, dado un orden de las filas y columnas de una matriz por el espacio de estados E con $E = \{1, 2, ..., N\}$ el conjunto de estados, el elemento de la matriz P_t en la posición (i, j), está dado por:

$$(P_t)_{i,j} = \mathbb{P}(X_{t+1} = j | X_t = i) \tag{3.3}$$

esto significa, que cada fila de la matriz P_t es un vector de probabilidad y la suma de sus entradas es 1. El **vector de estado inicial**, de dimensiones $Nx1$ representa la distribución de probabilidad de comenzar en cada uno de los N estados posibles. Cada elemento del vector representa la probabilidad de comenzar en ese estado.

3.2. Inferencia Bayesiana

La inferencia es un método que se usa para poder estimar algún elemento probabilístico que se desconozca en exactitud, observando la evidencia. Las inferencias pueden tomar la forma de un estimador puntual, un intervalo de confianza, una prueba de hipótesis, o simplemente un pronóstico. Existen tipos de inferencia:

- **Inferencia paramétrica:** donde en ocasiones resulta conveniente suponer que

  $$\mathbb{P}[X = x] = p(x|\theta), \qquad (si \quad X \quad es \quad discreta)$$

  donde $p(\cdot|\theta)$ tiene forma conocida pero el valor de θ es desconocido. De esta forma se describe el fenómeno, si sólo sí, se caracteriza el valor del parámetro θ

- **No paramétrica:** en estos casos, la propia forma funcional de $\mathbb{P}[X = x]$ se supone conocida

Para el caso de las inferencias paramétricas se realizan especificando un modelo probabilístico, $p(x|\theta)$, que determina las probabilidades de los posibles valores de X para un valor dado de θ, es decir:

$$X \sim Bin(\theta, n)$$

de manera que el problema de inferencia estadística se reduce a hacer inferencia sobre θ con base en el valor observado $X = x$

Las diferentes metodologías de inferencia se pueden ver como un conjunto de fórmulas que resultan aplicables en determinados casos y bajo ciertas condiciones. La metodología bayesiana está basada en la interpretación subjetiva de la probabilidad y tiene como punto central el **Teorema de Bayes**.

3.2.1. Teorema de Bayes

Thomas Bayes (1702 - 1761) fue un matemático y ministro de la iglesia presbiteriana [18], que tres años después de su muerte en 1764 se publica una memoria en la que aparece, por vez primera, la determinación de la probabilidad de las causas a partir de los efectos que han podido ser observados. El cálculo de esas probabilidades recibe el nombre de Teorema de Bayes.

Dentro de las aplicaciones de la teoría de la probabilidad, según Mesa et. al (2011) [19] "*(...) es válido enunciar el Teorema de Bayes como una expresión de probabilidad condicional que demuestra los beneficios obtenidos en las estimaciones basadas en c**onocimientos propios**. La metodología bayesiana específica un modelo de probabilidad que contiene algún tipo de conocimiento previo acerca de un parámetro investigativo, de este modo se acondiciona al modelo de probabilidad para realizar el ajuste de los supuestos*". La fórmula básica para la probabilidad condicional en circunstancias de dependencia se conoce como Teorema de Bayes:

$$\mathbb{P}(A|B) = \frac{\mathbb{P}(B \cap A)}{\mathbb{P}(A)} = \frac{\mathbb{P}(B|A)\mathbb{P}(A)}{\mathbb{P}(B)} \tag{3.4}$$

Si $\{A_i : i = 1, 2, ..., M\}$ es un conjunto exhaustivo de eventos mutuamente excluyentes, entonces:

$$\mathbb{P}(A_i|B) = \frac{\mathbb{P}(B|A_i)\mathbb{P}(A_i)}{\sum_{j=1}^{M} \mathbb{P}(B|A_j\mathbb{P}(A_j)} \tag{3.5}$$

En términos menos matemáticos, la importancia de este teorema recae en que vincula la probabilidad de A dado que se conoce B con la probabilidad de B dado que se conoce A.

3.2.2. Método de Inferencia Bayesiana

La incertidumbre es algo común dentro de los proceso de razonamiento en los que se pueden establecer condiciones para inferir de manera deductiva una proposición determinada. Uno de estos métodos de razonamiento es el **Modelo Bayesiano** que tiene como característica la asignación de la probabilidad como medida de creencia de una hipótesis, a través de un reajuste de medidas de creencias de nuevos axiomas. De esta forma, cuando se determina que una suposición es cierta basándose en evidencias y observaciones, entonces se habla de una **Inferencia Bayesiana**

Para esto se debe en primer lugar, definir una probabilidad a priori, que sea descriptiva con respecto a la incertidumbre y asocia la estimación de probabilidad con la ocurrencia de una evento al azar. Por el contrario, la distribución de probabilidad, más bien, es la estimación de un evento dado que se basa en un juicio informado. Ahora, siempre es bueno recordar que al momento de toma de decisiones se debe estar consciente de una infinidad de características asociadas a la incertidumbre que pueden influir en el valor de los resultados.

3.2.2.1. Interpretación Frecuentista y Bayesiana

En particular, la probabilidad tiene dos grandes formas de representación e interpretación, inmortalizadas como dos corrientes de pensamiento estadístico, una forma **frecuentista** y otra **bayesiana**.

La primera, considera la probabilidad como la frecuencia relativa de un experimento aleatorio. Es decir, la probabilidad se interpreta como la razón entre los casos favorables del experimento en relación al total de eventos aleatorios realizados. La segunda, el enfoque bayesiano, interpreta la probabilidad de manera subjetiva, y la utiliza para expresar su creencia respecto a una afirmación, dada cierta evidencia, es decir está relacionada con el concepto de probabilidad condicionada visto anteriormente en el Teorema de Bayes.

Durante mucho tiempo la probabilidad se vio como algo netamente frecuentista, es decir si se considera el experimento de lanzar una moneda al aire (no cargada), la experiencia muestra que en promedio tras un número suficientemente grande de lanzamientos, la cantidad de caras y sellos obtenidas es la misma, de esta forma se determina que la probabilidad de obtener una cara o sello de un lanzamiento al aire de la moneda es 0,5; sin embargo, bajo el mismo experimento, un individuo podría pensar que la probabilidad de obtener una cara o sello es 0,5, pero no por el hecho de tener esa experiencia previa, más bien por lo que el individuo espera que suceda, es decir, una probabilidad netamente subjetiva.

Estos dos enfoques no se encuentran tan distanciados. De hecho, una de las similitudes es que en ambos se utilizan modelos de estimación para inferir parámetros que se desconocen, en ambos también, se hace uso de la recolección de la información como soporte para la estimación de los parámetros desconocidos.

En particular, la estadística bayesiana da opción para incorporar información relevante externa de manera subjetiva, que reitera la posibilidad de modificar la probabilidad ante el acontecimiento de otro evento, mediante la utilización del Teorema de Bayes. Es decir, es un proceso que mediante una distribución de probabilidad, permite ajustar un modelo probabilístico y así obtener información de los parámetros sobre los que se desea realizar alguna estimación. Lo que conduce al bien llamado "mantra bayesiano":

$$\mathbb{P}(\theta|datos) \propto \mathbb{P}(\theta) \cdot \mathbb{P}(datos|\theta)$$

Que en simples palabras es que la relación entre la densidad a posteriori, es directamente proporcional a la densidad a priori por la verosimilitud.

3.2.3. Predicción de valores

Para poder estimar la probabilidad de los datos futuros, se debe usar el conocimiento que se tiene actualmente. De esta forma, la probabilidad predictiva de un dato y (no observado) se determina promediando las probabilidades predictivas de los datos a lo largo de todos los posibles valores de los parámetros y ponderados por la creencia en los valores de los parámetros (verosimilitud). Cuando solo se cuenta con la experiencia previa, se tiene:

$$p(y) = \int p(y|\theta)\, p(\theta)\, d\theta \tag{3.6}$$

Así, luego que se observan datos y se conoce una evidencia, la distribución predictiva posterior queda como:

$$p(y|x) = \int p(y|\theta)\, p(\theta|x)\, d\theta \tag{3.7}$$

Vale la pena destacar, que las predicciones son probabilidades de cada posible valor condicional al modelo de creencias actuales. Si interesa predecir un valor en particular en lugar de una distribución a lo largo de todos los posibles valores se puede usar la media de la distribución predictiva. Por lo tanto, el valor a predecir es:

$$p(y) = \int y\, p(y)\, dy \tag{3.8}$$

Donde la integral anterior únicamente tiene sentido si y es una variable continua. Si y es nominal entonces se puede usar el valor más probable.

3.2.3.1. Distribución a priori

Como se detalla en la sección anterior (Sección 3.2.2.1), aplicando el Teorema de Bayes, la probabilidad a priori se multiplica por la verosimilitud; al normalizar se obtiene la distribución de probabilidad a posteriori, la cual es la probabilidad de la distribución condicional dados los datos. Sin embargo, la elección de una buena distribución a priori puede ser fundamental para encontrar de forma precisa los valores de los parámetros buscados.

En inferencia estadística Bayesiana, una distribución de probabilidad a priori de una cantidad θ desconocida, es la distribución de probabilidad que expresa alguna incertidumbre acerca de θ antes de tomar en cuenta los datos. Los parámetros de las distribuciones a priori son llamados **hiperparámetros**, para distinguirlos de los parámetros del modelo.

En esta tesis se utilizan dos tipos de distribuciones a priori, la distribución Normal y la distribución Beta con sus respectivos parámetros que se detallarán en el próximo capítulo.

3.2.3.2. Distribución a posteriori

Como se nombra anteriormente, se tiene en primer lugar los datos, y cantidades o parámetros desconocidos θ cuyo valor interesa calcular. Se postula en primer lugar un modelo de probabilidad $p(y|\theta)$ donde, visto desde el punto de vista bayesiano, θ debe tener una distribución de probabilidad $p(\theta)$ que refleja la incertidumbre inicial acerca de su valor, además Y es conocido por lo que se debe condicionar a su valor observado y. Por lo tanto, el conocimiento acerca del valor de θ queda descrito a través de su distribución final (o distribución posteriori):

$$p(\theta|y)$$

que usando el Teorema de Bayes se puede encontrar:

$$p(\theta|y) = \frac{p(y|\theta)\, p(\theta)}{\int p(y|\theta)\, p(\theta)\, d\theta} \tag{3.9}$$

El cálculo de la integral en el denominador de la distribución posterior no es trivial, sin embargo hay manera de evitar calcularlo directamente:

- Un camino tradicional consiste en usar funciones de verosimilitud con distribuciones iniciales conjugadas. Cuando una distribución inicial es conjugada de la verosimilitud resulta en una distribución posterior con la misma forma funcional que la distribución inicial.
- Otra alternativa es aproximar la integral numéricamente. Eso sólo es posible cuando el espacio de parámetros es de una dimensión pequeña.
- Por último, en los últimos años se ha desarrollado una clase de métodos de simulación para poder calcular la distribución posterior, estos se conocen como cadenas de Markov vía Monte Carlo (MCMC por sus siglas en inglés). El desarrollo de los métodos MCMC es lo que ha propiciado el desarrollo de la estadística bayesiana en años recientes.

Para la resolución de esa integral compleja en la Ecuación (3.9), en esta tesis se utiliza un método de simulación a través de cadenas de Markov vía Monte Carlo, utilizando un programa de simulación que se detalla en el siguiente capítulo.

Capítulo 4

Modelo

En este capítulo se explica el paso a paso del desarrollo de un modelo de simulación del torneo "Premier League 2017-2018" que puede ser replicado, considerando los datos iniciales necesarios, para cualquier tipo de campeonato de fútbol.

Como se nombra anteriormente, el objetivo general de esta tesis es cuantificar el efecto de un jugador en un equipo de fútbol, medido a través de su aporte en la probabilidad de salir campeón del club en el torneo. Esto, para luego ser utilizado en el proceso de *scouting* (búsqueda de jugadores) y reducir el espectro de posibilidades que se tienen. Para poder lograr este objetivo se realiza una simulación completa del torneo. Es decir, se deben simular cada uno de los partidos que juegan los 20 equipos, como cada equipo juega 19 partidos tanto de local como de visita (no puede jugar contra sí mismo) todos los equipos juegan un total de 38 partidos durante el torneo, vale decir 380 partidos en total.

Para la simulación de todos los partidos, se intenta reproducir el accionar de los 22 futbolistas, cuando alguno de estos tiene el balón. O sea, se intenta simular cada uno de los llamados "eventos" en los partidos. De manera más simplificada, se consideran solo los eventos importantes: pases, tiros y duelos. Y así, se espera que se pueda representar de manera genuina el desarrollo de cada partido y por ende, la actuación de cada uno de los jugadores. Para finalmente, poder medir el error del modelo en comparación a la realidad.

4.1. Cadena de Markov

Esta simulación está descrita a través de Cadenas de Markov en tiempo discreto, que se pueden explicar gráficamente como una red de nodos y aristas que unen a estos nodos, en donde cada uno de éstos representa un estado particular del juego (evento) y cada arista una probabilidad de transición de un estado a otro. Así, **cada estado representa un evento particular del juego**: pase, tiro o duelo. Además, cada estado contiene la información respecto a qué **futbolista** se asocia el evento, en qué **tiempo** (minuto del partido) está ocurriendo el evento, cuál es el **resultado actual** del partido, en qué **zona** del campo ocurre el evento y si el jugador se encuentra de **local o visita**. De esta forma, para transcurrir de un estado (nodo) a otro, se necesita conocer la probabilidad de transición (arista) entre estos dos estados, tal como lo muestra la Figura 4.1.

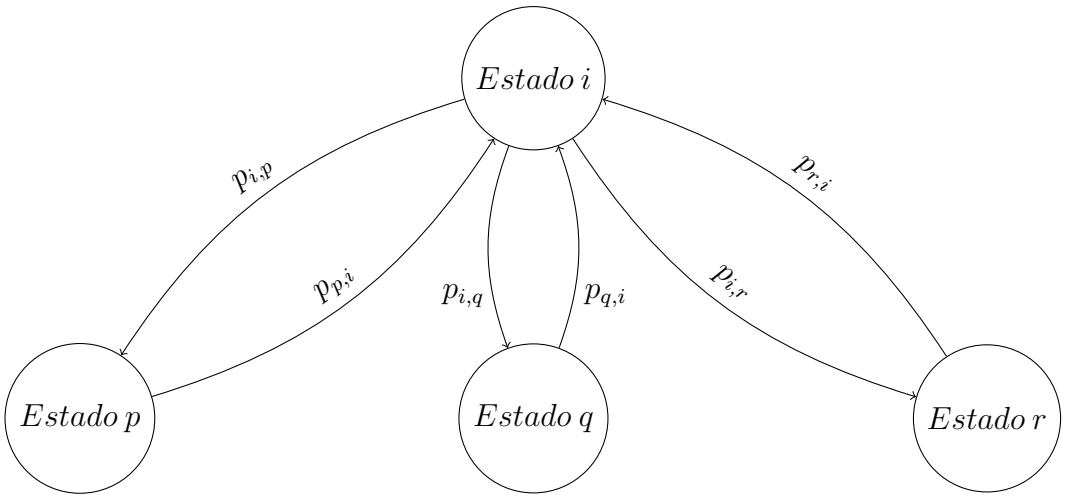

Figura 4.1: Grafo tipo "árbol de nodos": cada nodo representa un estado y cada arista una probabilidad de transición entre los estados.

Luego, para conocer el valor de la probabilidad de transición $\mathbb{P}_{transición}$ entre cada estado, se realiza un cálculo de probabilidad de dos instancias: la primera es la probabilidad de elección $\mathbb{P}_{elección}$ para determinar qué tipo de evento (acción) va a elegir el jugador, que tiene tres opciones posibles: realizar un pase, un regate a un rival o un tiro al arco. Luego que se determina la acción que el jugador decide realizar, se debe calcular la probabilidad de ejecución para determinar si la acción se realiza de forma correcta o incorrecta $\mathbb{P}_{ejecución}$. Así, después de estas dos instancias se pueden agrupar y combinar ambas probabilidades en una probabilidad única de transición, como lo muestra la Ecuación (4.1) y simular aleatoriamente la cadena de estados bajo la incertidumbre que estas mismas proponen; que gráficamente podremos describir como un nuevo nodo donde se ve involucrado un jugador y el balón en una nueva situación.

$$\mathbb{P}_{transición} = \mathbb{P}_{acción} \cdot \mathbb{P}_{ejecución} \tag{4.1}$$

En resumen, estableciendo un estado inicial, que de forma trivial en el fútbol se da con el evento del primer pase de un jugador a un compañero desde la mitad de la cancha en el inicio del partido, se comienza a generar un espectro de posibles próximos estados que involucran al segundo jugador en tocar el balón, pero difieren en primera instancia, en la decisión que toma ese jugador en el qué hacer: dar otro pase, intentar un duelo con un rival o tirar al arco contrario y en segunda instancia en la realización exitosa o fallida de tal acción. De esta forma, esta estructura de estados consecutivos tiene forma de árbol, tal como se muestra en la Figura 4.1 ya que para cada uno de los posibles estados a los que se llega, existe un nuevo espectro finito de estados próximos a ese último, y que van avanzando de forma consecutiva en el tiempo, determinado por las probabilidades de transición de cada uno.

4.2. Probabilidad de la elección

Como se describe en la sección anterior, para conocer la probabilidad de transición de un estado (nodo) a otro, en primer lugar se debe conocer que acción toma el jugador que tiene el balón. Para esto, en esta sección se describe el cálculo de esta probabilidad que luego se combina con la probabilidad de ejecución en la siguiente sección, para conocer la probabilidad final de transición entre los estados del modelo, como se muestra en la Ecuación (4.1).

Para el cálculo de esta probabilidad se utiliza la tabla de eventos detallada en la Sección 2.2, donde en la columna "*nameEvent*" se conoce el nombre (Tiro, Pase o Duelo) de cada una de las acciones tomadas por todos los jugadores durante el torneo. Además, en las otras columnas se tiene información respecto a cada uno de esos eventos: qué jugador está involucrado, en qué zona del campo de juego ocurre, cuál es el resultado actual del encuentro, el minuto de juego, entre otras.

Existe una dicotomía interesante en el mundo de la ciencia de datos entre aquellos que prefieren el *Machine Learning* (cada vez más relacionado al *Deep Learning*) y los estadísticos clásicos (tanto frecuentistas como bayesianos, ver Sección 3.2.2.1). Sin embargo, existen algunas herramientas interesantes que intentan acercar la brecha entre los dos campos, especialmente utilizando técnicas de Inferencia Bayesiana para estimar la incertidumbre de los modelos de *Deep Learning*. El mayor beneficio de aplicar el conocimiento bayesiano es que obliga a diseñar explícitamente todos los supuestos que entran en el modelo. Es difícil realizar Inferencia Bayesiana sin ser plenamente consciente de todas las opciones de modelado. La mayor desventaja de la Inferencia Bayesiana es el tiempo necesario para ejecutar incluso modelos de tamaño moderado como los que se presentan en esta tesis [25].

Por lo anterior, y previo a la descripción del modelo de inferencia bayesiana, es razonable pensar en algunos factores que podrían influir en la elección de la acción a tomar por un jugador y así incluirlos en el cálculo de esta probabilidad, sin olvidar considerar una variable relacionada al jugador que decide. A continuación se presentan las variables que se consideran pertinentes para describirla:

- **Zona:** La zona donde se está tomando la decisión de qué acción tomar influye en que decisión tomar, mientras más cerca del arco contrario, por ejemplo, aumenta la probabilidad de elegir tirar al arco, a diferencia de los duelos y pases que son más esperables en zonas del centro de la cancha.
- **Tiempo:** El minuto de juego puede ser un factor influyente a la hora de tomar la decisión de que acción tomar. Un equipo que está ganando por holgura en los últimos minutos, tiende a dar una mayor cantidad de pases y evitar los duelos o tiros al arco.
- **Resultado:** El resultado actual del partido podría influir a la hora de tomar la decisión de qué acción tomar. Un jugador que va perdiendo podría estar más presionado y tomar una decisión apurada con respecto a qué hacer.
- **Localía:** Es evidente que un equipo que se encuentra en condición de local, tiene mayor confianza en sus jugadores que estando en condición de visita, por lo que la localía podría ser un factor al momento de la toma de decisiones.

Existen otras variables que se podrían agregar a este modelo, para poder hacerlo más robusto. Sin embargo, al considerar las variables nombradas anteriormente se puede hacer un buen primer acercamiento a la representación de esta probabilidad. El modelo entonces, queda definido como lo muestra la forma funcional de la Ecuación (4.2).

$$\begin{aligned}\theta_{elección} = \beta_0 + \beta_{jugador} \cdot jugador + \beta_{zona} \cdot zona + \beta_{tiempo} \cdot tiempo \\ + \beta_{resultado} \cdot resultado + \beta_{localía} \cdot localía\end{aligned} \tag{4.2}$$

Para el cálculo de la probabilidad de ejecución, explicado en la siguiente sección, se utiliza un modelo de regresión logística simple, porque en ese caso la variable dependiente es una variable binaria (toma los valores 0 o 1), que indica si es gol o no, por ejemplo, para el caso de los tiros. En cambio, para modelar la probabilidad de elección, se tiene que considerar que son 3 categorías (pase, tiro o duelo), por ende la suma de las probabilidades de cada acción debe ser igual a 1. De esta forma se propone realizar un modelo de **regresión logística multinomial**.

4.2.1. Inferencia Bayesiana utilizando "Stan"

Como se desea conocer el valor del parámetro $\theta_{elección}$ basándose en las expectativas previas que se obtienen de los propios datos (β_i, en la Fórmula 4.2), se realiza Inferencia Bayesiana sobre este parámetro. Para esto, se usan los datos ordenados en la tabla de eventos. Como estos datos relacionan la probabilidad de elección con las variables que se consideran influyentes, se pueden realizar los pasos descritos en la Sección 3.2.3 para la predicción.

Sin embargo, como el cálculo de la integral en el denominador de la distribución posterior en la Fórmula 3.9 no es trivial, se nombran algunas maneras de evitar su calculo directamente y acercarse a su valor, donde la opción elegida en esta tesis es: métodos de simulación a través de cadenas de Markov vía Monte Carlo, donde para hacer esto se utiliza el programa "Stan".

Stan es un programa que se usa para generar muestras de una distribución posterior de los parámetros de un modelo, justamente lo que se necesita para encontrar la distribución de los parámetros β_i. El programa tiene ese nombre en referencia a Stanislaw Ulam (1904-1984) [22] quien fue pionero en utilizar los métodos de Monte Carlo. Para generar estas distribuciones, Stan realiza pasos de una Cadena de Markov con un método llamado Monte Carlo Hamiltoniano (HMC). Que suele ser más eficiente que otros muestreadores en JAGS y BUGS, en particular cuando se ajustan a modelos complejos con variables con correlación alta.

En física computacional y estadística, el algoritmo Monte Carlo Hamiltoniano (también conocido como Monte Carlo híbrido), es un método Monte Carlo de cadena de Markov para obtener una secuencia de muestras aleatorias que convergen para distribuirse de acuerdo con una distribución de probabilidad objetivo para la que se realiza un muestreo directo. Esta secuencia se puede utilizar para estimar integrales con respecto a la distribución objetivo (valores esperados) [23]. El lenguaje de programación que se utiliza en Stan es C++, y se puede acceder al lenguaje Stan en sí a través de varias interfaces, en esta tesis se realiza desde la interfaz *PyStan* que está integrado con el lenguaje de programación *Python*.

4.2.2. Modelo *Multi-Logit* en "Stan"

En estadística, la Regresión Logística Multinomial generaliza el método de regresión logística para problemas multiclase, es decir, con más de dos posibles resultados discretos. Vale decir, se trata de un modelo que se utiliza para predecir las probabilidades de los diferentes resultados posibles de una distribución categórica como variable dependiente, dado un conjunto de variables independientes (que pueden ser de valor real, valor binario, categórico-valorado, entre otras) que ayudan a explicar el modelo, en este caso: tiros, pases y duelos.

Para el cálculo de la probabilidad se propone un modelo *Multi-Logit*, que en primera instancia se le asignan distribuciones normales estándar como distribuciones a priori de los parámetros β_i. En el caso de este conjunto de datos, se tienen un total de 4 características diferentes; sólo la variable relacionada a la zona es una característica categórica y el resto son características codificadas de forma única con un **id** correlativo para cada valor. Por lo tanto, el número de parámetros a estimar es de cuatro por categoría. Dado que hay tres categorías (acciones), son un total de 12 parámetros para estimar. Para cada categoría se calcula la suma de los coeficientes y los valores de las características como se muestra en la Ecuación (4.2). Y que se resume en la ecuación que sigue con $N = 4$:

$$y_{ik} = \sum_{i=1}^{N} \beta_{ik} \cdot x_i \tag{4.3}$$

El código de programación de este modelo Multi-Logit (Ver Anexo B.2) se realiza en lenguaje C++ en la aplicación de Stan, utilizando la interfaz PyStan desde el programa Jupyter Notebook, donde se realizan todos los códigos de esta tesis. Una vez que se asignan el número de cadenas para que Stan realice las iteraciones, y se determine tanto el número de iteraciones como de las iteraciones de "calentamiento", se pueden comenzar a analizar los resultados del muestreo y la precisión de las distribuciones a posteriori del modelo.

Una vez que Stan realiza el muestreo de las distribuciones a posteriori de cada uno de los parámetros β_i existen dos opciones para proceder. En primer lugar, se puede utilizar las distribuciones a priori que describen el parámetro, para *samplear* (realizar una muestra) y obtener un valor aleatorio para el parámetro β_i de la distribución a posteriori. Por el contrario, y de manera de simplificar el cálculo se puede optar por utilizar el promedio de las muestras que realiza Stan para cada uno de los parámetros. En esta tesis, se decide utilizar los valores promedios de las distribuciones y utilizarlos como valor de los parámetros β_i.

Así, una vez que se tiene un valor para los parámetros β_i se procede al cálculo de la probabilidad de elección $\mathbb{P}_{elección} = \theta_{elección}$ que asigna un vector unidimensional de largo tres con las probabilidades de realizar: un pase, tiro o duelo. Teniendo en cuenta que estas probabilidades varían dependiendo del jugador que decide, la zona del campo donde se encuentra, el minuto de tiempo del partido, el resultado actual y la localía.

4.3. Probabilidad de ejecución

Tal como se describe en el inicio de esta sección, y luego de determinar la acción a realizar por parte del jugador asociado al estado actual de la Cadena de Markov, se debe determinar la probabilidad de realizar dicha acción de forma exitosa o fallida. Para esto, se detalla a continuación el proceso para encontrar la probabilidad de realizar un tiro exitosamente, pero se replica la metodología para las demás acciones.

4.3.1. Tiros

Una vez que se determina que el jugador relacionado con el estado actual decide realizar un "tiro", se debe definir una probabilidad de que realice esta acción de forma exitosa o fallida. Para esto, puede ser útil pensar qué factores directos y relacionados al juego en cancha pueden influir en esta probabilidad. En primer lugar, parece trivial que la probabilidad esté caracterizada por quién patea, es decir el **jugador asociado al evento** y por el **arquero rival**, quién intenta detener el tiro y evitar el gol. Además, tiene sentido considerar otros factores como por ejemplo:

- **Zona:** la zona de la cancha desde donde se realiza el tiro. Suena prudente considerar la zona, ya que un tiro en una zona más alejada del arco adversario debería tener una menor probabilidad de convertirse en gol en relación a una zona cercana.
- **Tiempo:** el minuto de juego en el que se encuentra el partido podría ser un factor importante a la hora de determinar si un tiro puede ser gol o no.
- **Resultado:** el resultado actual de un partido puede ser un factor influyente en determinar si un tiro se convierte en gol o no, esto ya que si un equipo va perdiendo o empatando, se puede pensar que un tiro podría asegurarse más.
- **Localía:** tiene sentido pensar que la localía puede ser un factor a considerar en la probabilidad de que un tiro sea gol o no.

Siempre un modelo se considera más robusto mientras se agregan más variables, sin embargo el exceso de estas puede ocasionar el efecto contrario. Por otro lado, hay variables externas a lo que sucede en el campo de juego que pueden ser muy influyentes, como las horas de sueño o problemas personales del jugador pero que el acceso a esos datos no es fácil. De esta manera, este conjunto de variables a considerar en el efecto de si un tiro se convierte en gol o no, pueden considerarse como una buena primera aproximación.

Para encontrar la probabilidad de ejecución se realiza Inferencia Bayesiana sobre el parámetro $\theta_{ejecución}$, por lo que se tiene que encontrar una distribución a priori (como se explica en la Sección 3.2.3.1) para este parámetro. Ahora, como se detalla más arriba, este parámetro se puede caracterizar y dejar en función de todos los factores: jugador involucrado en la acción, arquero rival, zona de la cancha, tiempo, resultado actual y localía. De esta forma se puede escribir el parámetro θ como una regresión simple con parámetros β_i asociados a inferir.

$$\begin{aligned}\theta_{ejecución} = \beta_0 + \beta_{jugador} \cdot jugador + \beta_{arquero} \cdot arquero + \beta_{zona} \cdot zona \\ + \beta_{tiempo} \cdot tiempo + \beta_{resultado} \cdot resultado + \beta_{localía} \cdot localía\end{aligned} \tag{4.4}$$

Para este caso se debe encontrar la distribución a priori para cada una de estos nuevos parámetros β_i con $i = \{jugador, arquero, zona, tiempo, resultado, localía\}$. A modo de simplificar el informe de esta tesis se analiza sólo el parámetro asociado al jugador, pero se utiliza la misma metodología para los parámetros asociados al arquero rival, zona, tiempo, resultado actual y localía. Algunas distribuciones como la Normal o la Beta podrían ser una buena aproximación que describan estos parámetros a priori.

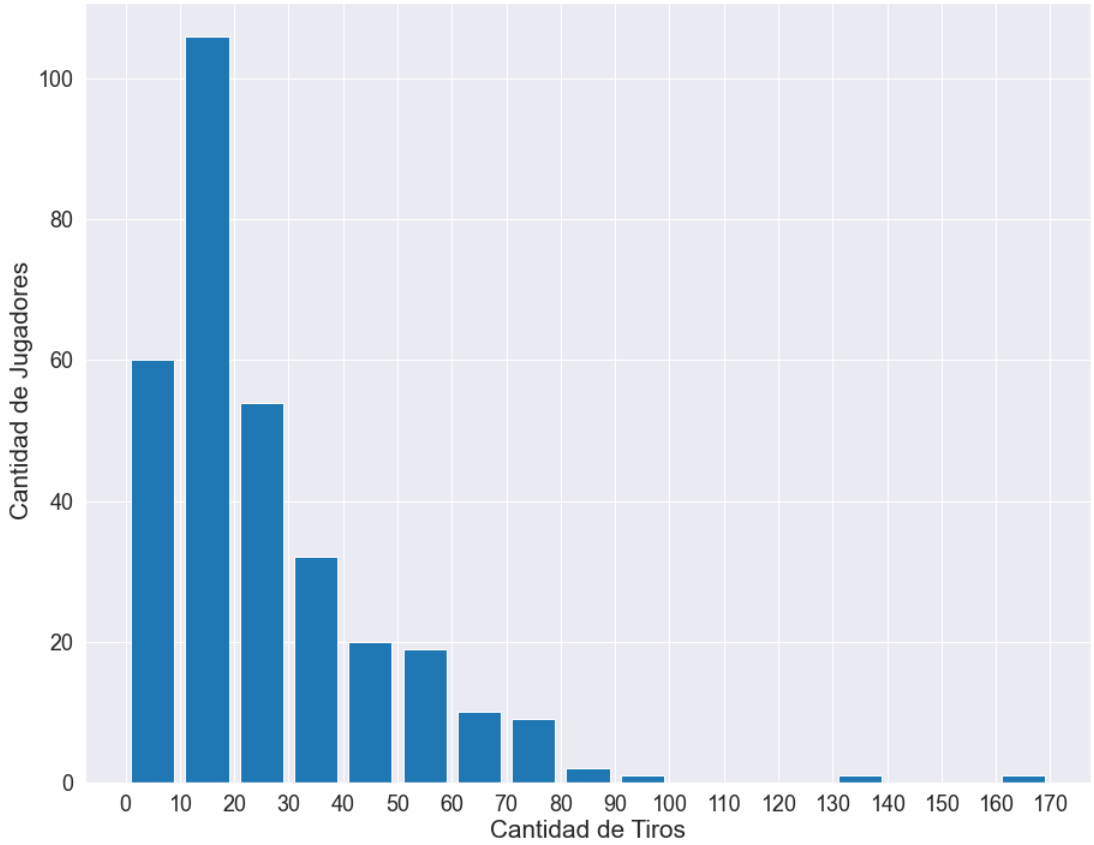

Figura 4.2: Histograma de la cantidad de tiros de los jugadores con más de cinco tiros durante todo el campeonato.

En la Figura 4.2 se observa como se distribuye la cantidad de tiros totales de los jugadores con más de cinco tiros. Se advierte que si bien los jugadores entre cero y diez tiros son aproximadamente 60 (19 % del total), al realizar el corte se ha dejado de lado 111 jugadores que tienen menos de 5 tiros. Esto se realiza con el fin de no "ensuciar" la predicción del parámetro considerando jugadores de los que se tiene poca información, para tener una buena expectativa previa. Esto se explica a continuación.

Cualquiera que siga el fútbol estaría interesado en conocer los promedios de conversión de goles de los jugadores, que, desde una mirada frecuentista, simplemente es el número de veces que un jugador realiza un gol divido por el número de veces que realiza un tiro (por lo que es solo un porcentaje entre 0 y 1). En este torneo el promedio de goles por tiro es de **0.319**, mientras que 0.450 se considera excelente.

Si se tiene un jugador de fútbol y se quiere predecir su promedio de goles por tiro durante toda la temporada, desde una perspectiva frecuentista se podría decir que se puede usar su promedio de gol hasta ahora, pero esta sería una medida muy pobre al comienzo de la temporada. Si un jugador tira una vez y hace un gol, su promedio de gol simplemente sería 1, mientras que si lo falla, su promedio es 0. No mejora mucho si patea tres o cuatro veces; podría tener una racha de suerte y obtener un promedio de 1, o una racha de mala suerte y obtener un promedio de 0, ninguna de las cuales es un predictor ni remotamente bueno de como patearía ese jugador en la temporada.

Lo anterior no es un buen predictor de como será su temporada. Ahora, en el caso que el primer tiro de un jugador en la temporada es un fallo, no se podría predecir que no hará un gol durante todo el transcurso del torneo, porque se tienen que considerar las **expectativas previas**, es por eso que se analiza la distribución de densidad de aquellos futbolistas que tienen más de cinco tiros desde una perspectiva bayesiana.

4.3.1.1. Distribución Beta

Dado el problema de promedio de gol del total de tiros, que se puede representar con una distribución binomial (una serie de éxitos y fracasos), la mejor manera de representar estas expectativas previas (lo que en estadísticas se llama a priori) es con la distribución Beta, es decir, antes de que hayamos visto al jugador hacer su primer tiro, lo que aproximadamente esperamos que sea su promedio de goles por tiros es 0,319, como se muestra en el gráfico de la derecha de la Figura 4.3. El dominio de la distribución beta es $(0, 1)$, como una probabilidad, por lo que ya se sabe que se está en el camino correcto, pero la idoneidad de la beta para esta tarea va mucho más allá.

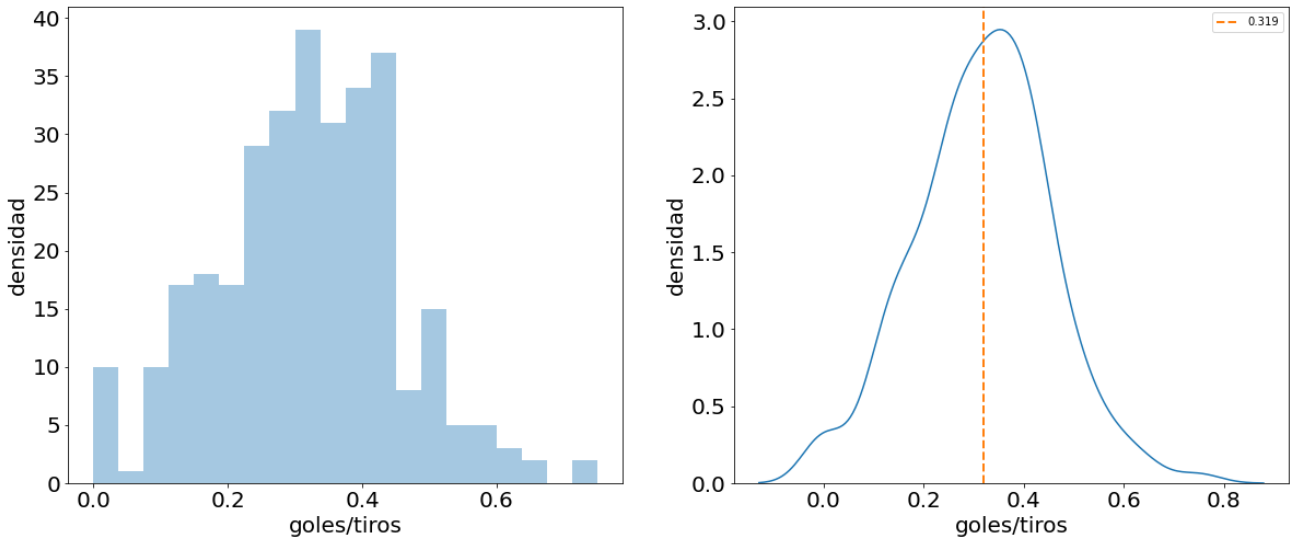

Figura 4.3: Histograma y densidad de la proporción de goles en el total de tiros de los jugadores con más de cinco tiros durante el campeonato.

Se espera que el promedio de goles del jugador durante toda la temporada sea más probable alrededor de 0.319, pero que razonablemente podría oscilar entre 0.15 y 0.50 donde se encuentra el 80 % de los tiradores de este campeonato. Esto se puede representar con una distribución beta con parámetros $\alpha = 3,44$ y $\beta = 7,34$:

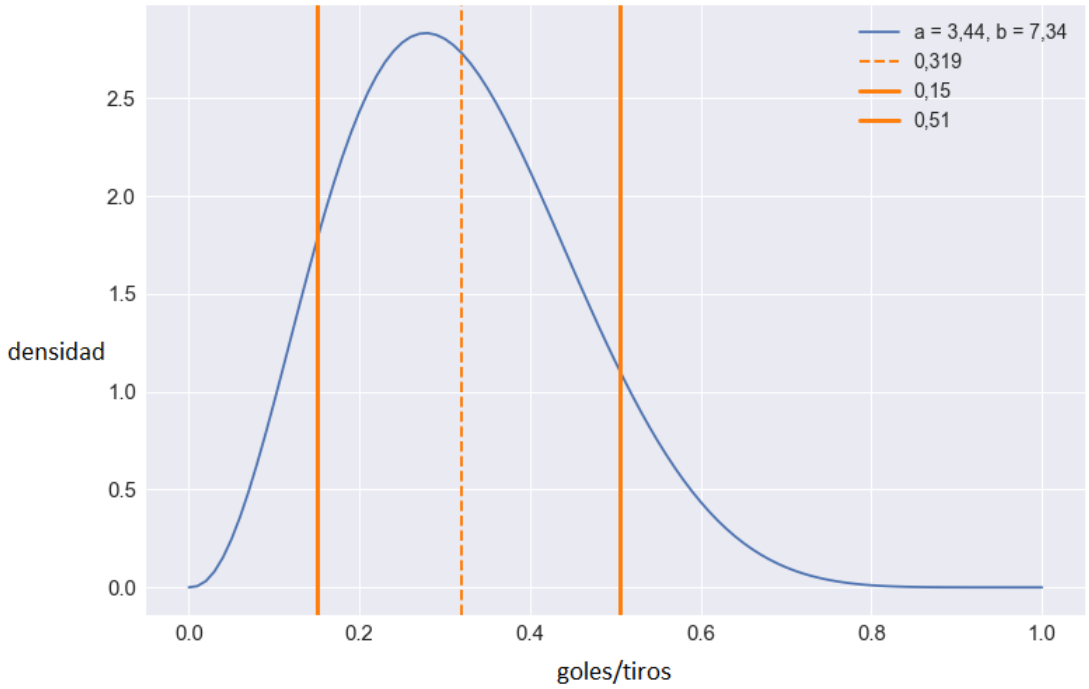

Figura 4.4: Distribución Beta de parámetros $Beta(\alpha, \beta)$ con $\alpha = 3,44$ y $\beta = 7,34$. Las líneas naranjas continuas demarcan el intervalo que contiene al 80 % de la población, entre $x_1 = 0,15$ y $x_2 = 0,50$.

Estos parámetros de la distribución $Beta(\alpha, \beta)$ corresponden a:

- La esperanza es $\frac{\alpha}{\alpha+\beta} = \frac{3,44}{3,44+7,34} = \mathbf{0,319}$
- Como se puede ver en la Figura 4.4, esta distribución se encuentra casi por completo (un 80 % de la población) dentro de (0,15, 0,51), el rango razonable para un promedio de goles por tiros.

Una buena intuición es interpretar qué representa el eje x en un gráfico de densidad de distribución beta; aquí representa el promedio de goles. Por lo tanto, se observa que en este caso, no solo el eje y es una probabilidad (o más precisamente una densidad de probabilidad), sino que el eje x también lo es, ya que el promedio de aciertos es solo una probabilidad de un acierto, después de todo. **La distribución Beta representa una distribución de probabilidades de probabilidades.**

Pero he aquí por qué la distribución Beta es tan apropiada. En el caso, que un jugador realiza un gol en un tiro. Sus estadísticas de la temporada serían "un tiro; un gol". Luego, tenemos que actualizar la probabilidad que en un comienzo definimos como el promedio: 0,319;

se quiere cambiar esta curva completa solo un poco para reflejar esta nueva información. Si bien las matemáticas para demostrar eso son complejas, se intenta resumir en lo siguiente:

La forma del conjugado a priori generalmente se puede determinar mediante la inspección de la densidad de probabilidad o la función de masa de probabilidad de una distribución. Por ejemplo, considerando una variable aleatoria que representa el número de sucesos s en n ensayos Bernoulli con probabilidad de suceso desconocida $q \in [0, 1]$. Esta variable aleatoria seguirá una distribución Binomial, con una función de probabilidad de masa de la forma:

$$p(s) = \binom{n}{s} q^s (1-q)^{(n-s)} \tag{4.5}$$

El prior conjugado usual es la distribución Beta con parámetros (α, β):

$$p(q) = \frac{q^{\alpha-1}(1-q)^{n-s}}{\mathbb{B}(\alpha, \beta)} \tag{4.6}$$

Donde α y β son elegidos para reflejar cualquier creencia o información existente ($\alpha = 1$ y $\beta = 1$ es una función Beta que actúa como una distribución contante). En este caso, α y β son llamados *hiperparámetros* (parámetros del prior), para distinguirlos de los parámetro subyacente (en la Ecuación (4.6) el parámetro q). Si luego, se toma una muestra de esta variable aleatoria y se obtiene s éxitos y f fracasos, se tiene[20]:

$$\begin{aligned} P(s, f|q = x) =& \binom{s+f}{s} x^s (1-x)^f \\ P(q = x) =& \frac{x^{\alpha-1}(1-x)^{\beta-1}}{B(\alpha, \beta)} \end{aligned} \tag{4.7}$$

$$\begin{aligned} P(q = x|s, f) =& \frac{P(s, f|x)P(x)}{\int P(s, f|y)P(y)dy} \\ =& \frac{\binom{s+f}{s} x^{s+\alpha-1}(1-x)^{f+\beta-1}/B(\alpha, \beta)}{\int_{y-0}^{1}(\binom{s+f}{n} y^{s+\alpha-1}(1-y)^{f+\beta-1}/B(\alpha, \beta))dy} \\ =& \frac{x^{s+\alpha-1}(1-x)^{f+\beta-1}}{B(s+\alpha, f+\beta)} \end{aligned} \tag{4.8}$$

que es otra distribución Beta con parámetros $(\alpha + s, \beta + f)$. Que según el ejemplo queda como: $(\alpha_0 + tiros, \beta_0 + fracasos)$. Donde α_0 y β_0 son los parámetros con los que se empieza, esto es: 3,44 y 7,34 respectivamente. Por lo tanto, y siguiendo el ejemplo, en este caso α ha aumentado en 1 (su único tiro), mientras que β no ha aumentado en absoluto (todavía no ha fallado, ya que fue gol). Esto significa que la nueva distribución es $Beta(3.44 + 1, 7.34 + 0) = Beta(4.44, 7.34)$. Esta nueva distribución queda descrita por la siguiente curva, comparada con la original:

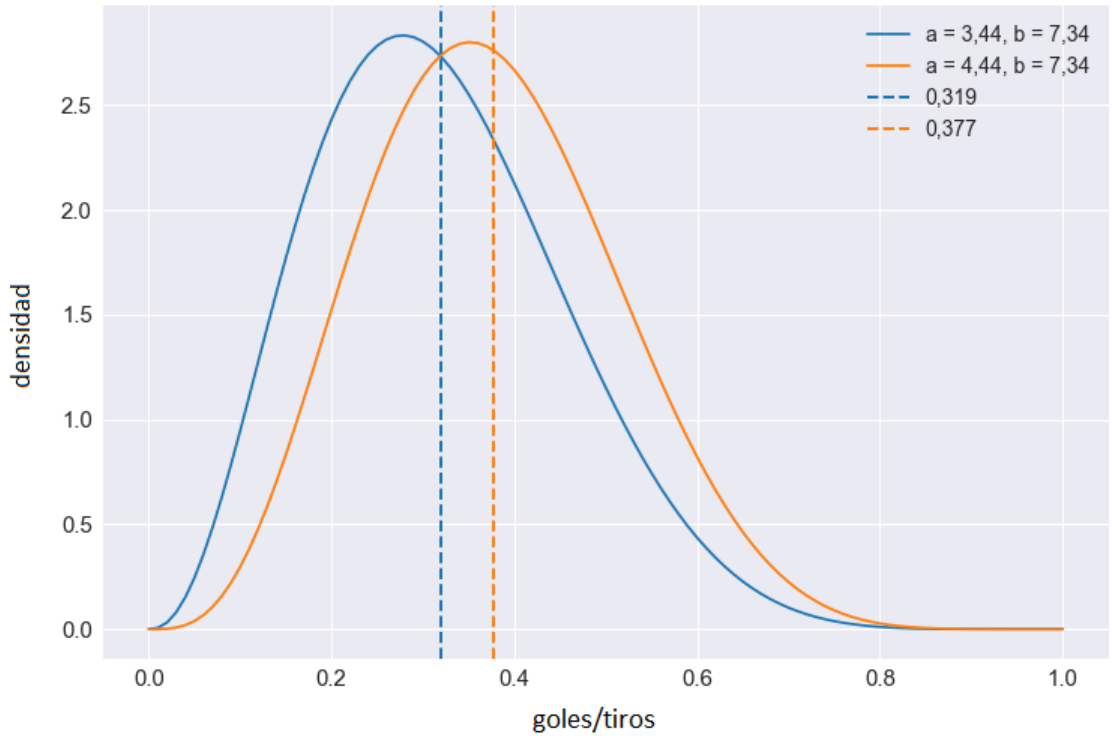

Figura 4.5: Distribución Beta de parámetros $Beta(\alpha, \beta)$ con $\alpha = 3,44$ y $\beta = 7,34$ (azul) y Distribución Beta de parámetros $B(\alpha + 1, \beta + 0)$ (naranjo). La primera con media=0,319 y la segunda con media=0,377.

En la Figura 4.5 se observa que, si bien existe una diferencia significativa estadísticamente entre ambas distribuciones el cambio es mínimo y se ve reflejado en el promedio, donde éste sólo varía 0,058 unidades. Esto tiene sentido tomando en cuenta que un tiro realmente significa poco en comparación con el total de tiros en una temporada. En la Figura 4.2 se observa la distribución de tiros a los largo del campeonato, donde el promedio es $\overline{x}_{tiros} = 25.89$ tiros.

Por otro lado, mientras más tiros realice el jugador durante el transcurso de una temporada, más se irá desplazando la curva de densidad de distribución para adaptarse a la nueva evidencia y, además, se reducirá más en función del hecho de que se tienen más pruebas. En el caso que un jugador tire 30 veces y marque 10 goles, la nueva distribución beta sería $Beta(3.44 + 10, 7.34 + 20) = Beta(13.44, 27.20)$, que se muestra en color rojo en la Figura 4.6 en comparación con las otras dos distribuciones beta:

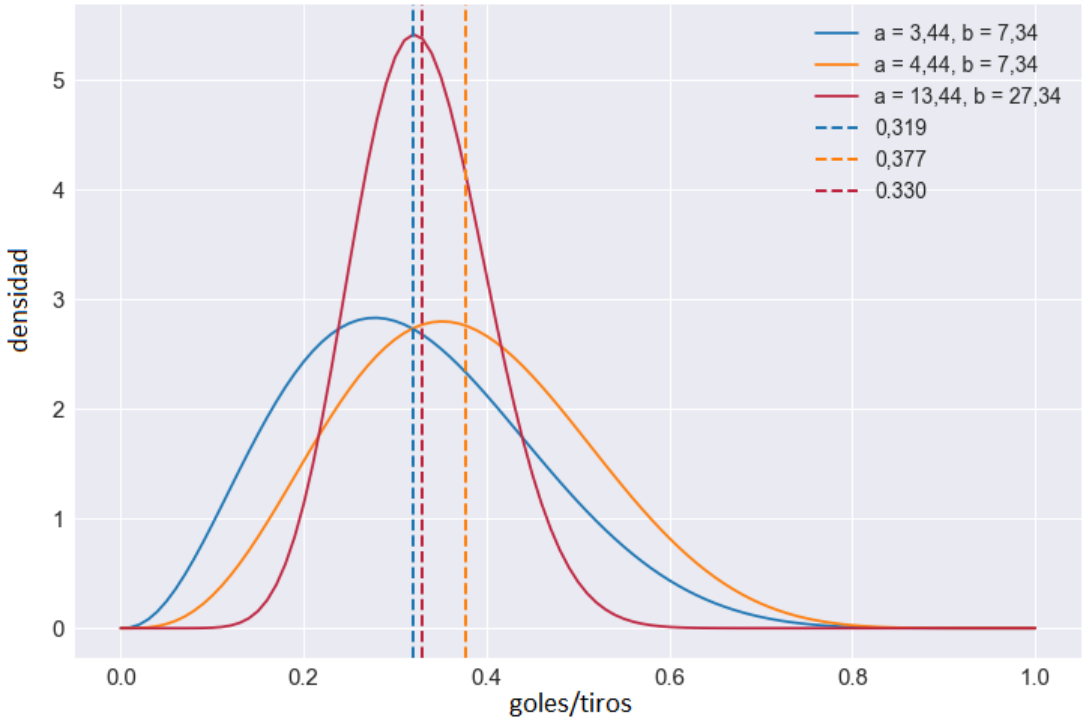

Figura 4.6: Distribuciones Beta de parámetros $Beta(3.44, 7.34)$ con media 0,319 (azul), $Beta(4.44, 7.34)$ con media 0,377 (naranjo), $Beta(13.44, 27.34)$ con media 0,330 (rojo).

Se observa en esta nueva curva (rojo) que es una distribución beta más delgada y levemente desplazada hacia la derecha (dado que tiene un promedio de gol más alto) de lo que era antes; de esta forma se tiene una mejor idea de cuál es el promedio de goles del jugador.

Uno de los resultados más interesantes de esta fórmula es el valor esperado (esperanza) de la distribución beta modificada, que es básicamente su nueva estimación. Recordando que el valor esperado de la distribución beta es: $\frac{a}{a+b}$. Por lo tanto, después de 10 goles en 30 tiros, el valor esperado de la nueva distribución beta es $\frac{3,44+10}{3,44+10+7,34+20} = 0,330$ que es menor que la estimación directa (sin considerar el conocimiento previo) $\frac{10}{10+20} = 0,333$ pero superior a la estimación con la que se comienza la temporada $\frac{3,44}{3,44+7,34} = 0,319$.

Por lo tanto, la distribución beta es mejor para representar una **distribución probabilística de probabilidades**; es decir, el caso en el que no se sabe que es una probabilidad de antemano, pero se tienen algunas conjeturas razonables.

En resumen, analizando el promedio de goles (goles en tiros realizados) podemos usar la distribución Beta como distribución a priori del parámetro a inferir $\beta_{jugador}$. De la misma forma, se realiza para los demás parámetros β_i de la Ecuación (4.4) para encontrar las demás distribuciones a priori. Ahora que se tienen la distribuciones a priori, el siguiente paso es encontrar la distribución a posteriori a través del método de Inferencia Bayesiana, obteniendo para cada parámetro β_i una distribución a posteriori con la ecuación nombrada

anteriormente y así inferir el parámetro $\theta_{ejecución}$, que representa la probabilidad de hacer un gol en un tiro condicionada al jugador involucrado, al arquero rival, a la zona desde donde se patea, al tiempo, al resultado actual del partido y a la localía del equipo de quién patea.

Ahora que se muestra el efecto que puede hacer el agregar un tiro más a la hora de calcular la distribución a priori, tiene sentido fijar un mínimo de tiros al arco a considerar para que la distribución a posteriori no se vea afectada por jugadores que tengan pocos tiros (menor a cinco tiros). La solución posterior que se realiza con estos jugadores que no tienen asociado un $\beta_{jugador}$ se hace utilizando un método de emparejamiento según la **"mínima distancia euclidiana"**.

4.3.1.2. Distancia Euclidiana

Para asignar el parámetro beta $\beta_{jugador}$ a un jugador que previamente se le excluye por la poca información que se tiene para realizar una expectativa previa, se utiliza la "Distancia Euclidiana". Ésta, consiste en encontrar al jugador que tiene un beta asociado (que tiró más de cinco tiros) y que se parezca más a algún jugador que no tiene (que tiene menos de cinco tiros), según ciertas características, y asignarle a ese jugador el parámetro $beta_{jugador}$ correspondiente.

Para realizar esto, se considera que aquellos jugadores que tienen menos de cinco tiros son jugadores que no juegan regularmente como titulares, por ende no se les puede comparar con aquellos que tienen más eventos asociados, de lo contrario se estaría incurriendo en un sesgo de selección, de hecho el promedio de minutos jugados por parte de los jugadores que tienen menos de cinco tiros es de **35 minutos** en promedio por partido, en contraposición con los **65 minutos** en promedio que juegan los jugadores con más de 5 tiros por partido. Por otra parte, es necesario encontrar a los jugadores que más se parecen teniendo en cuenta que se deben comparar jugadores de la misma posición: defensas, mediocampistas o delanteros. Las características elegidas se muestran en la Tabla 4.1:

Tabla 4.1: Variables a considerar para cálculo de la Distancia Euclidiana

Posición	**Características**			
Defensas	% Duelos defensivos ganados	% Duelos aéreos ganados	% Pases correctos	Faltas cada 90 minutos
Mediocampistas	% Pases correctos	Pases clave cada 90 minutos	% Duelos defensivos ganados	% Duelos ofensivos ganados
Delanteros	% Duelos ofensivos ganados	Pases clave cada 90 minutos	% Remates al arco	% Conversión de goles

Considerando estas características se debe calcular la distancia entre todos los jugadores con la fórmula mostrada en la Ecuación (4.9), donde cada característica puede ser un elemento de un vector unidimensional $p_i = p_1, p_2, ..., p_n$ de largo n, que para el ejemplo de tiros se utiliza $n = 4$ correspondiente a las 4 características de cada una de las posiciones, mostradas en la Tabla 4.10. De esta forma, la menor distancia espacial (en un espacio N^4) empareja a un jugador sin parámetro β_i asociado con otro que sí tiene. Pese a que no es un valor preciso, sí se espera que se acerque al valor real, dado la similitud entre un jugador y otro bajo esas características.

La Distancia Euclidiana entre los puntos $P = (p_1, p_2, ..., p_n)$ y $Q = (q_1, q_2, ..., q_n)$ del espacio euclídeo n-dimensional, se define como[21]:

$$D_E(P,Q) = \sqrt{(p_1 - q_1)^2 + (p_2 - q_2)^2 + ... + (p_n - q_n)^2} = \sqrt{\sum_{i=1}^{n}(p_i - q_i)^2} \tag{4.9}$$

4.3.1.3. Modelo *Logit* en "Stan"

Una vez resuelto el problema con aquellos jugadores que tienen una cantidad baja de tiros al arco (menos de 5 tiros), se puede calcular la distribución a posteriori de los parámetros β_i tal como se describe en la Sección 3.2.3. Para esto, nuevamente se utiliza el programa "Stan", que a modo de recordatorio se usa para generar muestras de una distribución posterior de los parámetros de un modelo, justamente lo que se necesita para encontrar la distribución de los parámetros β_i.

El modelo que se propone para realizar la simulación en Stan, describe un modelo lineal generalizado con verosimilitud de Bernoulli y función de enlace *logit*. Justamente Stan proporciona una única primitiva para este modelo, es decir, una primitiva para una **regresión logística**. En Stan, esto proporciona una implementación más eficiente de la regresión logística que una regresión escrita manualmente en términos de probabilidad de Bernoulli y multiplicación de matrices.

La regresión logística analiza datos distribuidos binomialmente de la forma $Y_i\ B(p_i, n_i)$, para $i = \{1, ..., m\}$, donde los números de ensayos Bernoulli n_i son conocidos y las probabilidades de éxito p_i son desconocidas. El modelo es entonces obtenido a base de lo que cada ensayo (valor de i) y el conjunto de variables explicativas puedan informar acerca de la probabilidad final. Estas variables explicativas pueden pensarse como un vector X_i que es $k-$dimensional y el modelo toma la siguiente forma:

$$p_i = E\left(\frac{Y_i}{n_i}\middle| X_i\right) \tag{4.10}$$

Los *logit* de las probabilidades binomiales desconocidas son modeladas como una función lineal de los X_i.

$$logit(p_i) = ln\frac{p_i}{1 - p_i} = \beta_0 + \beta_1 x_{1,i} + ... + \beta_k x_{k,i} \tag{4.11}$$

En estas condiciones, la probabilidad aproximada de hacer un gol se aproximará mediante una función logística del tipo:

$$\pi(x) = \frac{e^{(\beta_0+\beta_1 x)}}{e^{(\beta_0+\beta_1 x)}} = \frac{1}{e^{-(\beta_0+\beta_1 x)+1}} \tag{4.12}$$

que puede reducirse al cálculo de una regresión lineal para la función *logit* de la probabilidad:

$$g(x) = ln\frac{\pi(x)}{1-\pi(x)} = \beta_0 + \beta_1 x \tag{4.13}$$

De esta forma el modelo de regresión logística programado en Stan se muestra al detalle en el código del modelo que se puede observar en el Anexo B.1.

4.3.2. Cálculo probabilidad final

Por último, conociendo las distribuciones a posteriori de los parámetros del modelo para definir si un tiro es gol o no, caracterizado directamente por las variables asociadas al jugador, al arquero rival, a la zona, al tiempo, al resultado actual y a la localía, queda definido por:

$$u = \beta_0 + \beta_{jugador} + \beta_{arquero} + \beta_{zona} + \beta_{tiempo} + \beta_{resultado} + \beta_{localía} \tag{4.14}$$

Remplazando,

$$\begin{aligned} \mathbb{P}_{ejecución} = \mathbb{P}_{gol} &= \frac{1}{e^{-u}+1} \\ &= \frac{e^{u}}{e^{u}+1} \\ &= \frac{e^{(\beta_0+\beta_{jugador}+\beta_{arquero}+\beta_{zona}+\beta_{tiempo}+\beta_{resultado}+\beta_{localía})}}{e^{(\beta_0+\beta_{jugador}+\beta_{arquero}+\beta_{zona}+\beta_{tiempo}+\beta_{resultado}+\beta_{localía})}+1} \end{aligned} \tag{4.15}$$

Así, la probabilidad de ejecución para los tiros, se puede calcular como se muestra en la Ecuación (4.15). Cabe destacar, que una vez que se obtiene la distribución a posteriori de los valores de cada uno de los parámetros β_i muestreado a través de Stan, existen distintas formas para determinar el valor a utilizar del parámetro en cuestión. La primera forma se realiza tomando un valor aleatorio de la función de densidad de distribución del parámetro. Otra opción, y la que se decide realizar en esta tesis, es simplemente tomar el valor esperado de la distribución, es decir, el valor promedio de cada uno de los β_i.

En la siguiente sección, se detalla el proceso de simulación del torneo considerando todos estos cálculos previos. Para facilitar la simulación, y considerando que se realizan muchas iteraciones del campeonato, se intenta reducir la cantidad de cálculos que se deban hacer dentro de ésta. Por esto, se construye un diccionario con todas las combinaciones posibles entre las características, para definir la probabilidad de si un tiro es gol o no. De esta forma, al realizar la simulación y en cualquier "estado" en un tiro, es decir, fijando un jugador, zona, tiempo, resultado actual y localía, se puede acceder directamente a la probabilidad de gol de ese tiro, sin incurrir en el calculo de esa probabilidad y haciendo más eficiente la simulación. Esto se replica para las demás acciones (pases y duelos) y para la probabilidad de elección.

4.4. Simulación del campeonato

Una vez que se calculan las probabilidades tanto de elección como de ejecución y ya se encuentran calculadas para cualquier posible estado (nodo) de la cadena. Se almacenan en diccionarios de fácil acceso, que contienen todas las probabilidades calculadas para cualquier combinación de parámetros β_i. Haciendo las iteraciones de la simulación más eficientes computacionalmente hablando.

4.4.1. Simulación de un partido

Para comenzar a realizar la simulación de cada uno de los eventos de un partido es necesario definir condiciones iniciales de la cadena de Markov. De forma trivial, se fija el primer evento con un "pase exitoso", pero se debe encontrar qué compañero recibe el balón y en qué zona. De la misma forma en el caso que un pase fuera incorrecto, se debe encontrar al rival que se queda con el balón y en qué zona.

Para lo anterior, se busca la probabilidad que un jugador se encuentre en cierta zona del campo cuando un jugador de su mismo equipo realiza un pase. Esto se hace desde una mirada frecuentista, donde se calcula la participación en los eventos de cada jugador en cada zona. Así, para un pase exitoso se tiene un probabilidad de participación de un compañero en relación al total de eventos en donde participan todos (sin considerar a quien realiza el pase). Lo mismo ocurre cuando el pase es fallido, se busca la participación de algún jugador rival en una zona con respecto a la participación de los 11 jugadores de ese equipo.

Así, una vez que se realiza el primer pase exitoso, la cadena se encuentra en su segundo estado (nodo), asociado a un jugador que tiene el balón en una zona en particular de la cancha. Luego, como se está en un segundo estado, se busca en el diccionario de elección el vector de probabilidades que indica que ese jugador realiza un tiro, pase o duelo, generando una muestra aleatoria considerando esos "pesos" para los valores del vector.

Posteriormente, una vez que se conoce la acción a realizar por parte del jugador, se debe volver a generar un valor aleatorio dependiendo de las probabilidades de ejecución para saber si la acción se realiza de forma exitosa o fallida. Así, de forma repetitiva se puede ir simulando cada uno de los eventos consecutivamente.

De esta forma, se va simulando cada uno de los eventos en un partido de fútbol. Sin embargo, aún no se ha nombrado nada con respecto al tiempo. Para cada uno de los equipos se calcula el promedio de realizar un pase, tiro o duelo, considerando los casos especiales de un tiro que es gol, que demora más que un tiro que no lo es. Así, cada vez que ocurre un evento se le va adicionando el tiempo promedio que ese equipo demora en ese evento, hasta que el partido llega a los 95 minutos (tiempo máximo considerando el agregado).

Capítulo 5

Resultados

En este capítulo se detallan los resultados del modelo de simulación del torneo "Premier League 2017-2018". Para esto se realiza una comparación con la realidad. Dado que los datos utilizados como *inputs* del modelo son los eventos y resultados del torneo completo (las 38 fechas), para poder compararlo con la realidad, sin caer en sesgos, se deben separar los datos usados como *inputs* de algunos partidos, y predecir un grupo de partidos para el que no se han utilizado sus datos originales en la construcción del modelo, en estadística a esto se le llama "*Cross Validation*". De esta forma se evita evaluar un modelo calibrado con los mismos datos que se desean predecir.

Para esto, en primer lugar se detallarán los resultados particulares de la simulación de un partido para luego analizar los resultados generales del torneo simulado por completo. Así, en este capítulo se expone una comparación de la tabla de posiciones final del torneo simulado en relación a la tabla de posiciones real. Luego, se cuantifica el acierto del modelo de simulación, también en contraste con la realidad y en comparación con un modelo de *Poisson* que simula cada uno de los partidos a través de una competencia de tasas basadas en los goles a favor y en contra. Por último, se exhiben algunos ejemplos de intercambio de jugadores y la respectiva variación en la probabilidad de campeonar de los equipos involucrados al integrar un jugador en un equipo.

5.1. Simulación de un partido

Como se detalla en el capítulo anterior, la secuencia en la Cadena de Markov genera al mismo tiempo una sucesión de eventos que ocurren en el campo de juego, determinados siempre por las probabilidades de ocurrencia de cada evento según las condiciones del estado anterior. Así, al simular un partido se puede conocer el detalle de cada uno de los eventos que ocurren incluyendo pases, tiros y duelos, los jugadores involucrados en cada uno de ellos además de todas las variables anexas que involucran una acción en el modelo. A modo de ejemplificar como es el resultado de una simulación de un partido en específico, se muestra el resultado del primer partido de este torneo entre los equipos Arsenal y Leicester City, que en un entretenido encuentro terminó con un resultado real de 4-3 a favor del local.

Luego de simular 10.000 veces este partido, se obtienen los siguientes resultados mostrado en la Tabla 5.1 con respecto a la cantidad de eventos y goles por equipo en promedio.

Tabla 5.1: Promedio de eventos y goles: Arsenal vs Leicester City.

Equipo	**Promedio eventos**	**Promedio goles**	**Desv. estándar goles**
Arsenal	794,78	3,14	1,23
Leicester City	657,02	2,66	1,10

Dado que es una simulación que depende de las probabilidades que van ocurriendo en cada uno de los estados de la Cadena de Markov, no se puede analizar una sola simulación de forma individual, ya que ésta por sí sola no es representativa de lo que se intenta reproducir. Sin embargo, por la ley de los grandes números, si se simula el mismo partido, en este caso 10.000 veces, se puede llegar a aproximaciones reales de lo que el modelo de simulación desea representar. En este caso, como se observa en la Tabla 5.1 el promedio de eventos totales de los equipos se encuentra cercano al 79,9 % del total de eventos de la muestra, ver Tabla 2.2 tanto para el equipo Arsenal, que tiene un promedio de eventos total de local de 988 (794,78 promedio simulación) y el equipo Leicester City, que tiene un promedio de eventos total de visita de 799 (657,02 promedio simulación), como lo muestra la Figura 2.3.

Por otra parte, si bien el resultado real del encuentro fue 4-3 a favor del local, el promedio de goles en la simulación es bastante cercano a la cantidad de goles del partido real, lo que, en primera instancia, refleja un buen acercamiento del modelo a la realidad del partido. Notemos que al realizar muchas simulaciones, estás varían en la cantidad de goles, por lo que es interesante analizar la desviación estándar de los goles para las 10.000 simulaciones. Como se muestra en la Tabla 5.1 el promedio de goles del equipo Arsenal es de $3,14$ con una desviación estándar de $1,23$, por lo que es de esperarse que en una simulación cualquiera en promedio, esta se desvié $1,23$ goles de la media. Para el caso del equipo Leicester, este promedio de desviación de los datos con respecto a la media es de $1,10$, es decir, levemente más preciso (cercano a la media) que el local.

Ahora, lo que interesa analizar es como se comporta el modelo de simulación en un nivel más general. Para esto en la siguiente sección se estudia su comportamiento y efectividad a partir de la simulación de todo el torneo.

5.2. Simulación del campeonato

Antes de realizar una comparación con otros modelos y verificar la efectividad del modelo que se detalla en esta tesis, se muestra a continuación el resultado de la simulación del torneo “Premier League 2017-2018” observando la probabilidad de campeonar de los 20 equipos participantes simulando los 380 partidos del campeonato.

Recordando las posiciones reales al término del campeonato como lo muestra la Tabla 2.1, se procede a una comparación respecto a las probabilidades de campeonar de los equipos. Para esto, cada simulación del torneo tiene a un equipo "virtualmente" campeón, si se realizan 100 simulaciones del torneo, de las cuales, por ejemplo, un equipo es campeón en 34 ocasiones, entonces se puede decir que el equipo tiene un 34 % de probabilidades de campeonar. De esta forma, si se aumenta el número de simulaciones, las probabilidades de campeonar se hacen cada vez más precisas.

La Tabla 5.2 muestra las probabilidades de campeonar de todos los equipos para distintas cantidades de iteraciones a simular del torneo, comparándola con la tabla de posiciones real, donde el campeón fue el equipo "Manchester City" seguido de su clásico rival "Manchester United".

Tabla 5.2: Probabilidad de salir campeón de los equipos, luego de 100, 1.000 y 10.000 iteraciones del torneo.

Tabla de posiciones real		**Probabilidad de campeonar**		
Equipo	**Ptos.**	**100 its.**	**1.000 its.**	**10.000 its.**
		%	%	%
Manchester City	**100**	25,00	20,34	20,50
Manchester United	**81**	**27,00**	**31,11**	**31,20**
Tottenham Hotspur	**77**	11,00	18,58	18,52
Liverpool	**75**	11,00	9,32	9,29
Chelsea	**70**	8,00	4,81	5,35
Arsenal	**63**	2,00	1,23	1,24
Burnley	**54**	6,00	4,16	4,12
Everton	**49**	0,00	0,93	0,81
Leicester City	**47**	10,00	7,92	7,45
Newcastle United	**44**	0,00	0,21	0,19
Crystal Palace	**44**	0,00	0,00	0,01
Bournemouth	**44**	0,00	0,17	0,11
West Ham United	**42**	0,00	0,97	0,67
Watford	**41**	0,00	0,00	0,00
Brighton & H. A.	**40**	0,00	0,00	0,04
Huddersfield Town	**37**	0,00	0,00	0,00
Southampton	**36**	0,00	0,00	0,02
Swensea City	**33**	0,00	0,00	0,00
Stoke City	**33**	0,00	0,25	0,26
West Bromwich A.	**31**	0,00	0,00	0,04

Se observa que a medida que aumenta el número de simulaciones (iteraciones) del torneo, las distintas probabilidades se tornan más precisas. Por otro lado, si bien el equipo Manchester City fue campeón por amplia ventaja de sus seguidores, el modelo predice que el equipo con mayor probabilidad de salir campeón es el Manchester United, precisamente aquél que

obtuvo el segundo lugar con más de 80 puntos. Estas diferencias se pueden deber a varios factores: uno de ellos puede ser que efectivamente al mirar el plantel del Manchester United, este tuviera (por nombres y calidad de futbolistas) un mejor equipo que sus adversarios y, por ende, un mejor rendimiento "teórico", sin embargo en la práctica esto no ocurre siempre y se dan resultados inesperados. También, el modelo no considera la posición actual de los equipos al momento de jugar un partido, así, por ejemplo, cuando se está peleando por las primeras posiciones y se aproxima el final del torneo, los equipos tienden a intentar mantener o sacar aún más ventaja de sus seguidores, cosa que no se ve reflejada en el modelo y que podría explicar esa diferencia.

Si bien, estos resultados muestran como se comporta el modelo a nivel agregado y al simular muchas veces un torneo, es importante tener en cuenta que los datos utilizados como *inputs* para el cálculo de las probabilidades y la ejecución de las simulaciones son todos los datos de cada uno de los partidos del torneo real. Para no caer en estos sesgos, sería interesante analizar el comportamiento del modelo separando los partidos de donde se extraen los datos utilizados como "inputs" de los partidos a simular, a esto se le denomina en estadística "*Cross Validation*".

5.2.1. Efectividad del Modelo

5.2.1.1. Tabla de posiciones relativa

Para realizar lo anterior, se divide la muestra en dos conjuntos de partidos, la primera rueda (partidos de ida) que contiene las primeras 19 fechas del torneo, con un total de 190 partidos, y la segunda rueda (partidos de vuelta) que abarca la misma cantidad de juegos. De esta forma, se pueden utilizar los datos contenidos en una muestra para poder predecir y simular el otro conjunto de partidos y luego realizarlo de forma inversa. Esto permite no caer en ningún sesgo al utilizar datos de partidos distintos a los que se desea predecir.

Para esto, en primer lugar se consideran los datos de los partidos de la primera rueda en donde se calculan todas las probabilidades de transición desde cualquier estado a otro en la cadena de Markov. Se construyen nuevamente los diccionarios que contienen esta probabilidades y con esto se simulan cada uno de los partidos de a segunda rueda. La tabla de posiciones se empieza a construir desde esa fecha, sin considerar las 19 anteriores, como si fuese un nuevo campeonato y sin considerar los puntos obtenidos de los partidos de la primera rueda, formando una **tabla de posiciones relativa** desde la fecha 20 en adelante, simulando los 190 partidos de la segunda rueda.

La Tabla 5.3 muestra los resultados de la tabla de posiciones relativa en comparación con la tabla de posiciones real luego de 1.000 y 10.000 iteraciones del torneo, en donde además para cada uno de los equipos se calcula la diferencia de su posición relativa promedio en las simulaciones con respecto a la posición real obtenida en el torneo. Además, se muestra la diferencia entre estas dos posiciones para cuantificar la efectividad del modelo en términos de posiciones relativas finales de los equipos. Por ultimo, también se muestra el promedio general de estas diferencias de posición para ambas simulaciones.

Tabla 5.3: Tabla de posiciones relativa de la segunda mitad del campeonato (fechas 20-38), simulando el torneo en 1.000 y 10.000 iteraciones.

Tabla de posiciones real		**Tabla de posiciones relativa**			
Pos.	**Equipo**	**1.000 its.**		**10.000 its.**	
		Pos. prom.	**Dif. Pos.**	**Pos. prom.**	**Dif. Pos.**
1	**Manchester C.**	5,35	4,35	5,18	4,18
2	**Manchester U.**	4,29	2,29	4,32	2,32
3	**Tottenham H.**	4,68	1,68	4,65	1,65
4	**Liverpool**	5,30	1,30	5,26	1,26
5	**Chelsea**	7,69	2,69	7,61	2,61
6	**Arsenal**	8,46	2,46	8,59	2,59
7	**Burnley**	7,43	0,43	7,23	0,23
8	**Everton**	9,18	1,18	9,26	1,26
9	**Leicester City**	7,71	1,29	7,66	1,34
10	**Newcastle U.**	11,00	1,00	11,14	1,14
11	**Crystal Palace**	13,33	2,33	13,42	2,42
12	**Bournemouth**	11,82	0,18	12,11	0,11
13	**West Ham U.**	13,50	0,50	9,95	3,05
14	**Watford**	17,51	3,51	17,56	3,56
15	**Brighton & H.A.**	14,33	0,67	14,14	0,86
16	**Huddersfield**	16,36	0,36	16,27	0,27
17	**Southampton**	17,27	0,27	17,23	0,23
18	**Swensea City**	14,51	3,49	14,50	3,50
19	**Stoke City**	10,25	8,75	10,52	8,48
20	**West Brom. A.**	13,50	6,50	13,39	6,61
		Dif. prom.	**2,26**	**Dif. prom.**	**2,38**

Como se observa en la Tabla 5.3 las posiciones promedio luego de 1.000 iteraciones de la simulación de la segunda rueda (fecha 20 a 38), en general se ajustan bien a la tabla de posiciones real, salvo algunas excepciones como los equipos: Stoke City, West Bromwich A. y Manchester City, quienes tienen las mayores diferencias entre su posición promedio relativa y real: 8,75, 6,50 y 4,35 puestos respectivamente. Por otro lado, el Bournemouth con 0,18 es el equipo con mejor predicción en términos de su posición, seguido por el Southampton (0,27) y el Huddersfield (0,36). La diferencia promedio entre las posiciones reales y relativas del torneo es de **2,26 puestos**, esto es, en promedio las posiciones relativas se desviaron de las reales en 2,26 puestos. Para el caso de las 10.000 iteraciones del torneo los resultados no varían significativamente y la diferencia de posiciones promedio es de **2,38 puestos**. Esa mínima diferencia en la décima, se puede otorgar a la estocasticidad del modelo.

Para completar el análisis, se realiza el procedimiento inverso, ahora se toman los datos de los partidos de la segunda rueda como *inputs* del modelo, y se simulan las 19 primeras fechas, como se observa en la siguiente tabla.

Tabla 5.4: Tabla de posiciones relativa de la primera mitad del campeonato (fechas 1 a 19), simulando el torneo en 1.000 y 10.000 iteraciones.

Tabla de posiciones real		**Tabla de posiciones relativa**			
Pos.	**Equipo**	**1.000 iteraciones**		**10.000 iteraciones**	
		Pos. prom.	**Dif. Pos.**	**Pos. prom.**	**Dif. Pos.**
1	**Manchester C.**	4,41	3,41	4,51	3,51
2	**Manchester U.**	3,96	1,96	3,95	1,95
3	**Tottenham H.**	5,28	2,28	5,32	2,32
4	**Liverpool**	6,87	2,87	6,97	2,97
5	**Chelsea**	6,48	1,48	6,41	1,41
6	**Arsenal**	10,36	4,36	10,47	4,47
7	**Burnley**	7,15	0,15	7,05	0,05
8	**Everton**	10,63	2,63	10,80	2,80
9	**Leicester City**	5,36	3,64	5,33	3,67
10	**Newcastle U.**	11,76	1,76	11,70	1,70
11	**Crystal Palace**	15,48	4,48	15,45	4,45
12	**Bournemouth**	12,56	0,56	12,47	0,47
13	**West Ham U.**	10,98	2,02	11,18	1,82
14	**Watford**	16,16	2,16	16,18	2,18
15	**Brighton & H.A.**	12,64	2,36	12,62	2,38
16	**Huddersfield**	16,53	0,53	16,44	0,44
17	**Southampton**	14,04	2,96	13,93	3,07
18	**Swensea City**	13,30	4,70	13,41	4,59
19	**Stoke City**	12,85	6,15	12,54	6,46
20	**West Brom. A.**	13,19	6,81	13,26	6,74
		Dif. prom.	**2,86**	**Dif. prom.**	**2,87**

Como se observa en la Tabla 5.4 las posiciones promedio luego de 1.000 iteraciones de la simulación de la primera rueda (fechas 1 a 19), nuevamente se ajustan bien a lo que fue la tabla de posiciones real, pese a que la diferencia promedio entre las posiciones reales de los equipos y las relativas a la primera rueda simulada aumentó a **2,86 puestos** en el caso de las 1.000 iteraciones y a **2,87 puestos** para las 10.000 iteraciones, en comparación a la simulación de la segunda rueda vista anteriormente. Los equipos con menor diferencia de posición son: Burnley, Huddersfield y Bournemouth, quienes tuvieron 0,15, 0,53 y 0,56 puestos de diferencia respectivamente luego de 1.000 iteraciones.

Por otro lado, cuando se realizan 10.000 iteraciones los resultados no varían significativamente, se repiten los equipos con menor diferencia de posiciones real y relativa y los equipos con mayor diferencia, al igual que para las 1.000 iteraciones son: West Bromwich Albion, Stoke City y Swensea City, con 6,74, 6,46 y 4,59 respectivamente, justamente aquellos equipos que ocuparon los últimos puestos de la tabla.

5.2.1.2. Comparación con modelo Poisson

Para medir el rendimiento del modelo de simulación utilizando Cadenas de Markov e Inferencia Bayesiana, se realiza un contraste con un modelo de Poisson que tiene como cualidad principal que basa su predicción utilizando solo tres variables como *inputs* del modelo: goles del local, goles del visita y localía. Con esto, utilizando los datos de los partidos jugados genera una **tasa de ataque** construida desde los goles realizados por los equipos, una **tasa de defensa** construida desde los goles recibidos por cada uno de los equipos y por ultimo un factor de localía que favorece al equipo local. De esta forma, el modelo de Poisson simula los partidos del torneo a través de una competencia de tasas entre el ataque y defensa de cada uno de los equipos considerando adicionalmente el factor de localía.

Para esta comparación se consideran cinco parámetros que entregan información respecto a la efectividad y rendimiento de los modelos basados en el resultado general de los partidos, como también en el acierto de los goles y diferencia de goles del local y visita. Los parámetros a considerar son:

- **Acierto al resultado:** este parámetro indica el rendimiento del modelo al coincidir con el resultado real del partido (local, empate o visita). Si por ejemplo, el resultado real fue un empate, luego de realizar 1.000 simulaciones este parámetro indica cual es el porcentaje de empates respecto al total de simulaciones.
- **Acierto diferencia de goles absoluta:** este parámetro refleja el porcentaje de acierto del modelo al valor absoluto de la diferencia de goles real que tuvo el partido entre el local y la visita.
- **Acierto diferencia de goles no absoluta:** este parámetro refleja el porcentaje de acierto del modelo a la diferencia de goles real que tuvo el partido entre el local y la visita. Por ejemplo, si el resultado real fue 3-2, este parámetro indica cuantas veces el partido simulado tuvo una diferencia de gol de -1 (tres menos dos) en relación al total de simulaciones.
- **Diferencia de gol absoluta promedio:** este parámetro indica el promedio de la diferencia de goles absoluta promedio entre el resultado real y el simulado de todos los partidos. Mide que tan alejado está la simulación con la realidad en términos del valor absoluto de la diferencia de goles.
- **Diferencia de gol no absoluta promedio:** este parámetro indica el promedio de la diferencia de goles no absoluta promedio entre el resultado real y el simulado de todos los partidos. Mide que tan alejado está la simulación con el resultado real en términos de la diferencia de goles no absoluta.

En la siguiente tabla se muestra el resumen de la comparación de estos parámetros entre ambos modelos para 1.000 y 10.000 iteraciones del torneo, dividiendo el torneo en dos conjuntos de partidos, la primera rueda (partidos de ida) que considera desde la primera fecha hasta la fecha 19 y la segunda rueda (partidos de vuelta) que considera desde las fecha 20 a la fecha 38 del torneo. Así, se utilizan los datos de una rueda para simular los partidos de la otra rueda.

Tabla 5.5: Rendimiento de los modelos de cadena de Markov y Poisson para 1.000 y 10.000 iteraciones según diferentes parámetros.

Parámetros	**Modelo C. de Markov**				**Modelo de Poisson**			
	1.000 its.		**10.000 its.**		**1.000 its.**		**10.000 its.**	
	1-19	**20-38**	**1-19**	**20-38**	**1-19**	**20-38**	**1-19**	**20-38**
% Acierto al resultado	39,01	39,37	39,09	39,12	44,43	**44,45**	44,28	44,28
% Acierto dif. goles absoluta	26,76	**27,61**	26,80	27,57	25,51	25,97	25,40	26,03
% Acierto dif. goles no abs.	18,33	**19,41**	18,38	19,26	17,77	18,59	17,64	18,54
Dif. de gol abs. prom.	1,27	1,21	1,27	1,21	0,87	**0,84**	0,87	0,84
Dif. de gol no abs. prom.	1,80	1,72	1,80	1,72	1,38	1,35	1,37	**1,34**

En la Tabla 5.5 se puede observar la capacidad de predicción o rendimiento del modelos de Cadenas de Markov y el modelo de Poisson. Utilizando como *inputs* la segunda rueda y simulando las primeras 19 fechas del torneo, se observa que luego de 10.000 iteraciones el modelo de Cadenas de Markov tiene un rendimiento aproximado del 39,09 % en contraste con el 44,28 % del modelo de Poisson. Es decir, este último es capaz de predecir en un 44,28 % el resultado (local, empate o visita) de los partidos, calibrando el modelo sólo con la mitad de los datos (fecha 20 a 38). De la misma forma, cuando se utiliza la primera rueda de partidos para calibrar los modelos y se simulan los partidos de la segunda rueda (fecha 20 a 38) se mantienen los porcentajes, con 39,12 % de efectividad para el modelo de Cadenas de Markov y un 44,28 % para el modelo de Poisson, cuando se realizan 10.000 iteraciones.

Por otro lado, el modelo de cadenas de Markov se comporta de mejor manera al predecir la diferencia de goles y diferencia de goles absoluta de los partidos en comparación con el modelo de Poisson. El porcentaje de acierto del primer modelo luego de 10.000 iteraciones es de 26,80 % y 27,57 % cuando se simula la primera y segunda rueda respectivamente y el modelo de Poisson sólo acierta en un 25,40 % y 26,03 % al valor absoluto de la diferencia de goles.

Además, si se analiza el comportamiento de ambos modelos en relación al acierto en la diferencia de goles directa (sin valor absoluto), la diferencia entre ambos modelos se mantiene, acertando en un 18,38 % y 19,26 % cuando se realiza la predicción de ambas rondas del torneo con el modelo descrito en esta tesis, en contraste con el 17,64 % y 18,54 % que predice el modelo de Poisson luego de realizar 10.000 iteraciones del torneo.

Por último, para cuantificar estas diferencias en los resultados de los partidos simulados por ambos modelos en relación a la realidad, es interesante observar qué tan desviados se encuentran en relación a los goles convertidos durante el partido y la diferencia de goles. El valor absoluto de la diferencia de goles entre el local y el visita en promedio, refleja que tan acertado es el modelo para predecir la diferencia entre los equipos que se ve reflejada

en goles. El modelo de Cadenas de Markov se aleja 1,2 goles promedio con respecto a la diferencia real que hay entre los equipos, en el modelo de Poisson esta diferencia disminuye a 0,8 goles aproximadamente. Al ser un valor absoluto, esta diferencia solo compara los goles en el marcador y pierde la conexión con el acierto del resultado: local, empate o visita.

Por lo anterior, uno de los parámetros más importantes a observar es el valor absoluto de la diferencia de goles promedio, ya que esta sí mantiene ligado el resultado final del partido con la diferencia de goles entre el local y visita. Para el caso del modelo de simulación con Cadenas de Markov luego de 1.000 y 10.000 iteraciones esta diferencia es de 1,8 goles cuando se simula la primera rueda de partidos y de 1,72 goles cuando se simula la segunda rueda. El modelo de Poisson tiene un mejor rendimiento ya que su diferencia de goles promedio con respecto a la realidad es 1,37 y 1,34 luego de 10.000 iteraciones, simulando la primera y segunda rueda respectivamente.

Es importante señalar que aquellos parámetros que muestran un porcentaje de acierto, como los tres primeros de la Tabla 5.5 mientras mayor sea el porcentaje, mejor rendimiento tiene el modelo, un porcentaje del 100 % indica que el modelo logra predecir por completo lo sucedido en la realidad en relación a ese parámetro. Por otro lado, los últimos dos parámetros de la tabla muestran la distancia en goles de las diferencias de gol promedio de los partidos. De esta manera, una diferencia de gol de 0 goles indica que el modelo logra predecir por completo la diferencia de goles de los partidos simulados en comparación con la realidad.

5.3. Recomendación de jugadores

En esta sección se exponen ocho ejemplos de la inclusión de nuevos jugadores de distintas posiciones y en distintos equipos para analizar la variación en la probabilidad de campeonar de los equipos, tomando como base las probabilidades que se muestran en la Tabla 5.2.

Como el objetivo de esta tesis es exactamente medir el efecto de un jugador en la probabilidad de campeonar de un equipo, previamente se debe tener en consideración lo siguiente. El efecto de la inclusión de un nuevo jugador en un equipo genera dos modificaciones en la situación preliminar del torneo, en primer lugar, el jugador A que es incluido en el equipo B ocasiona un cambio en la estructura y formación inicial del equipo B, pero además genera el mismo cambio en el equipo A, que ya no cuenta con él para su formación inicial. Por ende, en ese caso no solo se está midiendo a inclusión del nuevo jugador sino también el efecto que genera que el deportista deje su equipo original. Para esto, a continuación se describen tres posibles mitigaciones a ese efecto:

1. **Reemplazo del jugador por un suplente:** Para solucionar la inclusión de un jugador del equipo A en el equipo B, se puede reemplazar el lugar que deja en el equipo A por un jugador suplente del mismo equipo. Sin embargo, esta medida no es efectiva, ya que el jugador suplente, por lo general, tiene pocos minutos jugados, y por consiguiente, pocos datos que describan de forma robusta su comportamiento en el campo, por lo que difícilmente el modelo logra describir de forma integra su desempeño.

2. **Mantener al jugador en ambos equipos:** Otra forma de solucionar el lugar que deja el jugador A en el equipo A al incorporarse al equipo B, puede ser reemplazándolo por él mismo, es decir, al momento de simular el torneo con el modelo descrito en la tesis, tanto el equipo A como el equipo B tendrían a disposición el jugador A. Si bien, esto puede ser una opción factible para medir el desempeño del jugador en el equipo B, sin alterar en demasía el desarrollo del torneo, en la aplicación se vuelve poco realista, ya que esta situación es imposible de realizar en la vida real.

3. **Intercambio de jugadores entre los equipos:** La inclusión del jugador A en el equipo B, además de dejar un lugar vacío en la alineación inicial del equipo A, ocasiona que el jugador B del equipo B quede fuera del planteamiento del equipo B. Por lo mismo, el intercambio de jugadores entre el equipo A y B puede ser una solución más efectiva y realista que las anteriores. Así, el jugador A ahora pertenece al equipo B y el jugador B ahora es parte del equipo A.

Si bien, ninguna de estas opciones mide únicamente la inclusión de un jugador en un nuevo equipo, la alternativa de intercambiar jugadores parece ser la más razonable e interesante dado el contexto de esta tesis. De hecho, si se tuvieran datos de otras ligas o campeonatos para alimentar este modelo, podría ser factible la inclusión de un jugador A de la liga A en algún equipo de la liga B y de esta forma se podría medir el efecto auténtico de la incorporación de ese único jugador, sin afectar el desarrollo del resto del torneo.

5.3.1. Defensas

5.3.1.1. Harry Maguire por Rob Holding

El primer intercambio de jugadores que se realiza es el de Harry Maguire, jugador perteneciente al Leicester City por Rob Holding, perteneciente al equipo Arsenal. Ambos son jugadores que juegan en la posición de defensas centrales y participaron en 3.420 minutos y 820 minutos respectivamente. La Figura 5.1 muestra el desempeño de ambos considerando los parámetros: porcentaje de duelos defensivos ganados, porcentaje de duelos aéreos ganados, porcentaje de pases cortos correctos, porcentaje de pases largos correctos y porcentaje de minutos jugados en relación a todo el torneo.

Si bien, ambos son jugadores relativamente parejos en relación al desempeño que tuvieron durante el campeonato, el jugador Harry Maguire le saca ventajas a Holding en el porcentaje de duelos aéreos ganados, porcentaje de duelos defensivos ganados y estuvo en cancha más de cuatro veces los minutos en los que participó Rob Holding. Por otro lado, Holding, es levemente superior en los parámetros de pases cortos y pases largos, con un acierto cercano al 80 % y 60 % respectivamente.

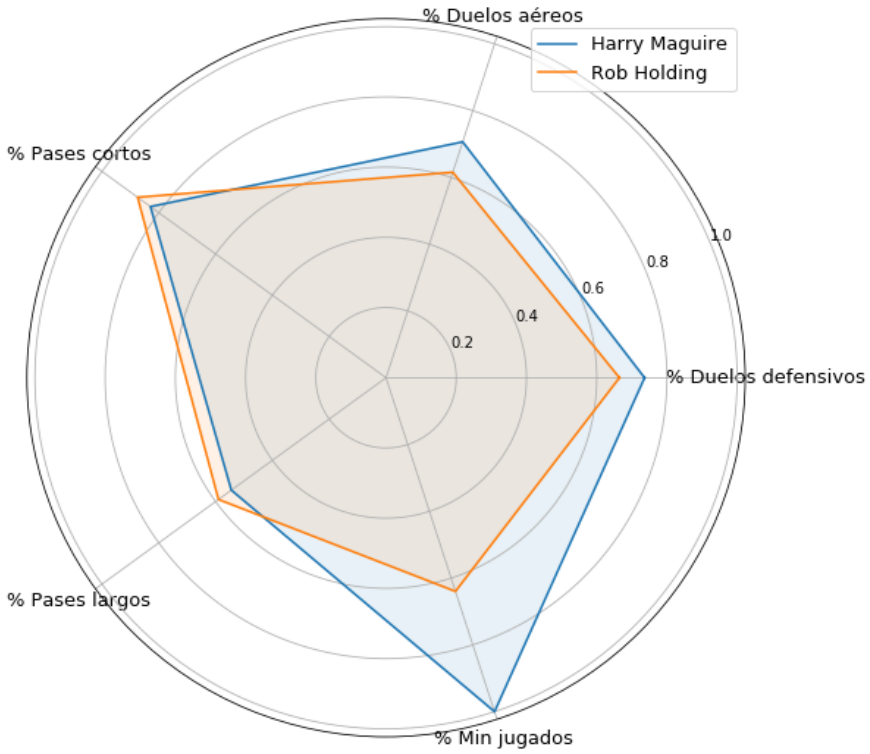

Figura 5.1: Comparación de rendimiento defensivo del jugador Harry Maguire (Leicester City) y Rob Holding (Arsenal).

Al realizar el intercambio de estos jugadores y volver a simular el torneo iterando cada uno de sus partidos 10.000 veces, los resultados de las nuevas probabilidades de campeonar de los equipos se muestran de forma resumida en la Tabla 5.6, para ver el detalle completo de las nuevas probabilidades de campeonar de todos los equipos, revisar Anexo C.1.

Tabla 5.6: Probabilidades de campeonar intercambiando a Maguire por Holding.

Equipo	**Probs. sin cambio**	**Probs. con cambio**	**Delta**
Arsenal	1,24	2,40	+ 1,16
Leicester City	7,45	6,70	- 0,75

Se observa que con la salida de Harry Maguire y la inclusión de Holding en el Leicester City, las probabilidades de campeonar de este equipo bajan en 0,75 puntos porcentuales, de forma inversa la inclusión de Maguire en el Arsenal hace que este equipo se refuerce de mejor manera y sus probabilidades aumentan en 1,16 puntos porcentuales.

Como se detalla al comienzo de esta tesis, este tipo de investigaciones permite facilitar el proceso de *scouting* de los equipos, analizando así posible refuerzos que según su desempeño podrían tener buenos resultados en sus nuevos equipos, centrando el análisis sólo en aquellos que potencialmente podrían ser un buen refuerzo para el equipo.

5.3.2. Mediocampistas

5.3.2.1. David Silva por Dele Alli

Los resultados del intercambio de David Silva (2.438 min. jugados) jugador del equipo campeón Manchester City por el jugador del Tottenham Hotspur Dele Alli (2.971 min. jugados), se muestran a continuación. En primer lugar, se analiza el desempeño de ambos jugadores en el torneo, tal y como se muestra en la Figura 5.2.

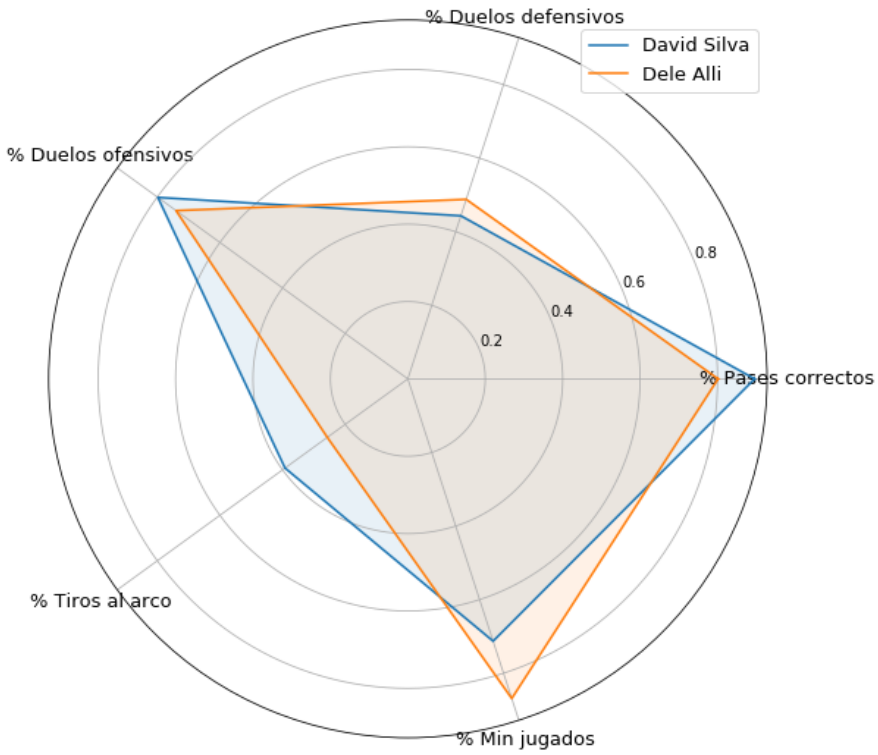

Figura 5.2: Comparación de rendimiento en el mediocampo del jugador David Silva (Manchester City) y Dele Alli (Tottenham Hotspur).

Con respecto a la variación en la probabilidad de campeonar, como se muestra en la Tabla 5.7 el Manchester C. disminuye en 9,90 puntos porcentuales su probabilidad de campeonar, mientras que el Tottenham con la inclusión de D. Silva aumenta sus probabilidades en 20,58 puntos porcentuales. Este es un buen ejemplo, para determinar que Silva podría ser un gran refuerzo en un equipo como el Tottenham dado que con su inclusión las probabilidades de campeonar de este último equipo aumentan considerablemente. Para ver el detalle completo de las probabilidades de todos los equipos, ver Anexo C.2.

Tabla 5.7: Probabilidades de campeonar intercambiando a Silva por Alli.

Equipo	**Probs. sin cambio**	**Probs. con cambio**	**Delta**
Manchester City	20,50	10,60	- 9,90
Tottenham Hotspur	18,52	39,10	+ 20,58

5.3.2.2. Kevin de Bruyne por Mark Noble

Los resultados del intercambio entre Kevin De Bruyne (3.084 min. jugados) jugador del equipo Manchester City y Mark Noble (2.404 min. jugados), histórico medicampista del West Ham United desde el año 2004, se detallan a continuación. En la Figura 5.3 se realiza una comparación del desempeño durante el torneo de estos dos jugadores.

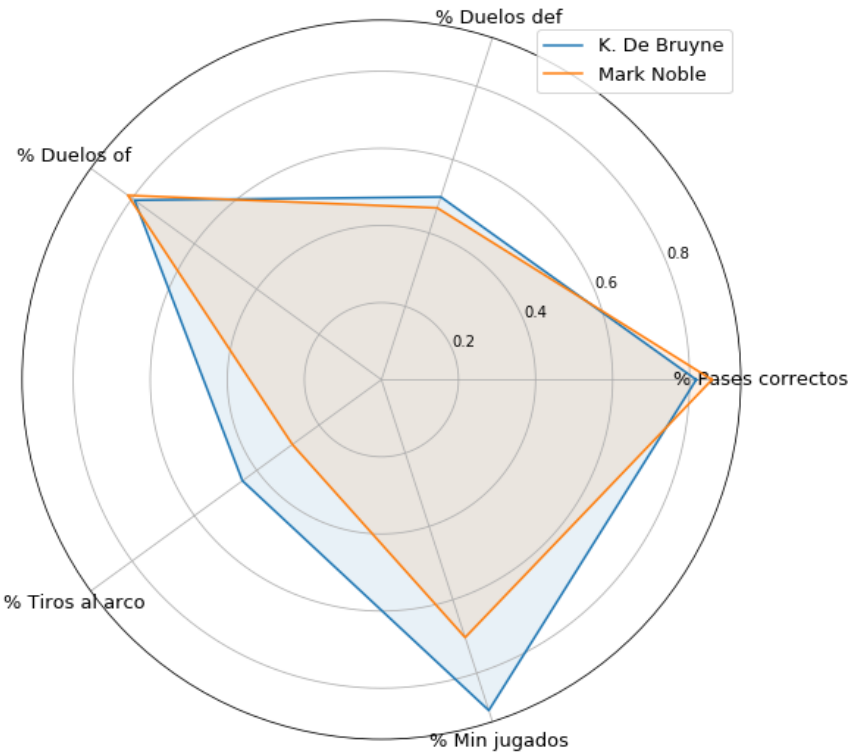

Figura 5.3: Comparación de rendimiento en el mediocampo del jugador Kevin De Bruyne (Manchester City) y Mark Noble (West Ham United).

Con respecto a la variación en la probabilidad de campeonar, como se muestra en la Tabla 5.8 el Manchester C. disminuye en 5,50 puntos porcentuales su probabilidad de campeonar, mientras que el West Ham U. con la incorporación de K. De Bruyne aumenta sus probabilidades en 1,83 puntos porcentuales. Aquí se observa lo importante del proceso de *scouting*, ya que la inclusión de un jugador puede incluso cuadriplicar las probabilidades de campeonar, tal como lo sería K. De Bruyne en el West Ham United. Para ver el detalle completo de las probabilidades de los demás equipos, ver Anexo C.2.

Tabla 5.8: Probabilidades de campeonar intercambiando a De Bruyne por Noble.

Equipo	**Probs. sin cambio**	**Probs. con cambio**	**Delta**
Manchester City	20,50	13,70	- 5,50
West Ham United	0,67	2,50	+ 1,83

5.3.3. Delanteros

5.3.3.1. Romelu Lukaku por Álvaro Morata

Los resultados del intercambio entre el traspaso más caro del torneo, Romelu Lukaku (2.869 min. jugados), jugador del equipo Manchester United y Álvaro Morata (2.068 min. jugados), jugador del Chelsea y de la selección española, se detallan a continuación. En la Figura 5.4 se muestra una comparación del desempeño durante el torneo entre estos dos jugadores. Lukaku es superior en todos los parámetros a analizar con excepción de los pases correctos, donde Morata tiene una mayor efectividad.

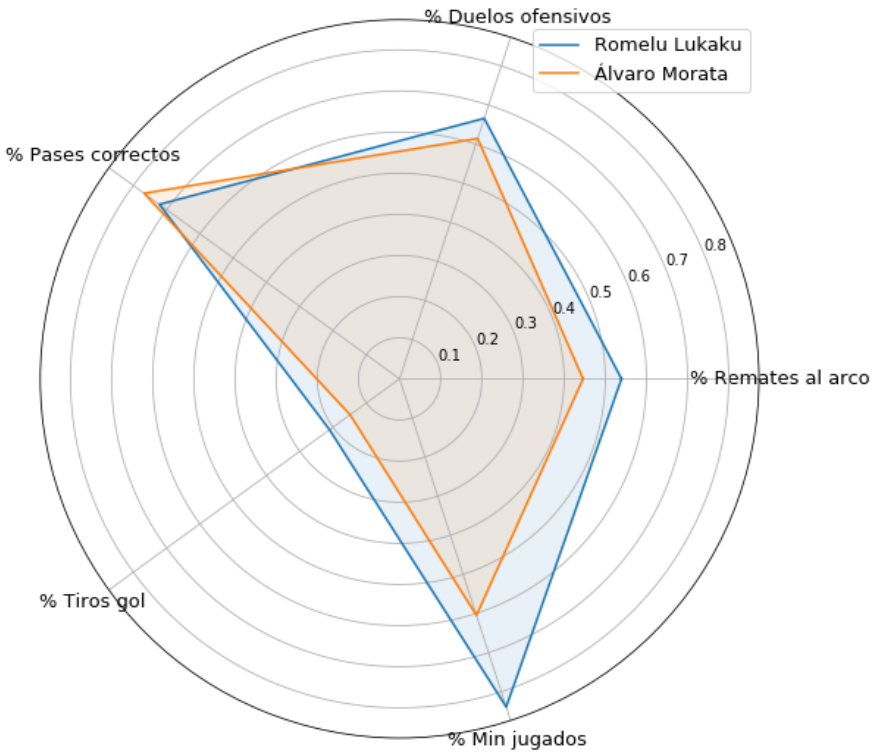

Figura 5.4: Comparación de rendimiento ofensivo del jugador Romelu Lukaku (Manchester United) y Álvaro Morata (Chelsea).

Tal como lo muestra la Figura 5.4 la salida de Lukaku y la integración de Morata en el Mnachester United beneficia al Chelsea, ya que Lukaku es un delantero con mejor rendimiento. En la Tabla 5.9 se ve que el intercambio entres estos dos ocasiona que el Manchester United se vea perjudicado, disminuyendo en 13,10 puntos porcentuales su probabilidad de campeonar, en contraposición al Chelsea que se vería beneficiado aumentándola en 5,87 puntos porcentuales. El detalle completo de las probabilidades se puede revisar en el Anexo C.3.

Tabla 5.9: Probabilidades de campeonar intercambiando a Lukaku por Morata.

Equipo	**Probs. sin cambio**	**Probs. con cambio**	**Delta**
Manchester United	31,20	18,10	- 13,10
Chelsea	5,53	11,40	+ 5,87

5.3.3.2. Sergio Agüero por Alexandre Lacazette

Los resultados del intercambio de Sergio Agüero (1.968 min. jugados), goleador histórico del equipo campeón Manchester City (21 goles en el torneo) y el jugador del Arsenal Alexandre Lacazette (2.211 min. jugados), se muestran a continuación. En primer lugar, se revisa el desempeño de ambos jugadores en el torneo, en el gráfico de la Figura 5.5.

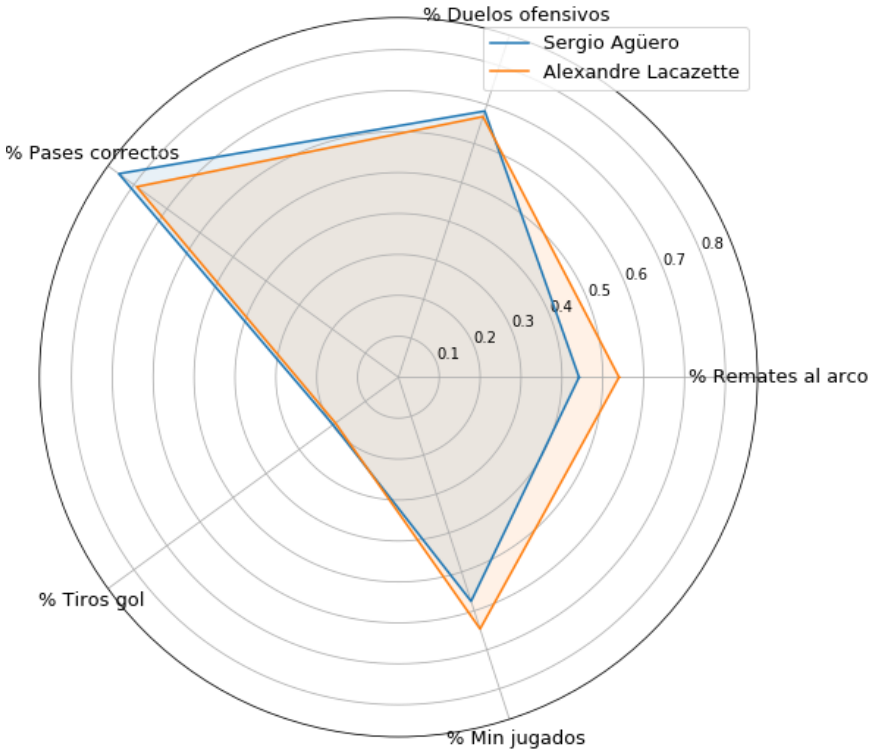

Figura 5.5: Comparación de rendimiento ofensivo del jugador Sergio Agüero (Manchester City) y Alexandre Lacazette (Arsenal).

Con respecto a la variación en la probabilidad de campeonar, como se muestra en la Tabla 5.10 el Manchester City disminuye en 6,80 puntos porcentuales su probabilidad de campeonar, mientras que el Arsenal con la incorporación de Sergio Agüero aumenta sus probabilidades en 2,46 puntos porcentuales, luego de simular 10.000 veces el torneo. Si bien, Lacazette tiene un mayor porcentaje de tiros al arco, Agüero hizo siete goles más en siete partidos menos jugados, lo que lo convierte en un atacante muy efectivo con 0,84 goles por partido, justo detrás del goleador del torneo Mohamed Salah, mientras que Alexandre Lacazette tiene 0,44 goles por partido, casi la mitad. Para ver el detalle completo de las probabilidades, ver Anexo C.3.

Tabla 5.10: Probabilidades de campeonar intercambiando a Agüero por Lacazette.

Equipo	**Probs. sin cambio**	**Probs. con cambio**	**Delta**
Manchester City	20,50	13,70	- 6,80
Arsenal	1,24	3,70	+ 2,46

5.3.3.3. Mohamed Salah por Mame Diouf

Los resultados del intercambio entre el goleador del torneo Mohamed Salah (2.922 min. jugados), jugador del Liverpool y Mame Diouf (2.601 min. jugados), histórico atacante del Stoke City desde el año 2014 al 2020, se detallan a continuación. En la Figura 5.6 se realiza una comparación del desempeño durante el torneo de estos dos jugadores. En donde se observa claramente que Salah es un jugador mucho más completo en relación a estos parámetros, ya que supera en todo nivel a M. Diouf.

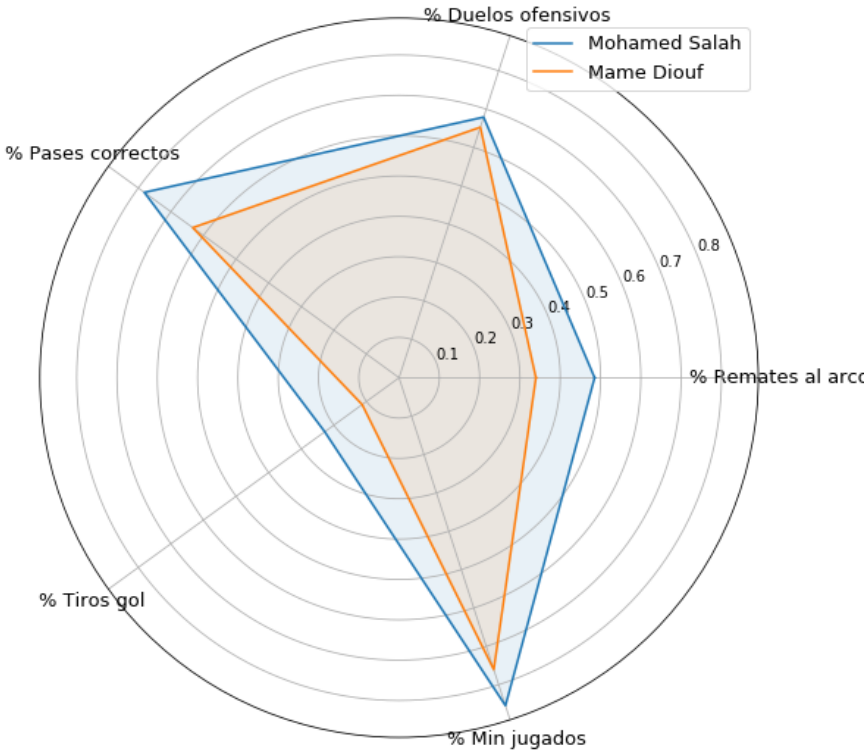

Figura 5.6: Comparación de rendimiento ofensivo del jugador Mohamed Salah (Liverpool) y Mame Diouf (Stoke City).

Con respecto a la variación en la probabilidad de campeonar, como se muestra en la Tabla 5.11 el Liverpool disminuye en 8,89 puntos porcentuales su probabilidad de campeonar, mientras que el Stoke City con la incorporación de Mohamed Salah aumenta sus probabilidades en 6,34 puntos porcentuales. La incorporación del goleador (32 goles) del campeonato en un equipo que terminó en el puesto 19° con tan solo 35 goles convertidos, casi los mismos goles convertidos por M. Salah en 36 partidos hace que las probabilidades de campeonar aumenten casi 25 veces. Para ver el detalle completo de las probabilidades, ver Anexo C.3.

Tabla 5.11: Probabilidades de campeonar intercambiando a Salah por Diouf.

Equipo	**Probs. sin cambio**	**Probs. con cambio**	**Delta**
Liverpool	9,29	0,40	- 8,89
Stoke City	0,26	6,60	+ 6,34

5.3.4. Arqueros

5.3.4.1. Ederson por Jack Butland

Al realizar el mismo ejercicio de intercambio con los arqueros, se obtienen resultados interesantes a comentar. A continuación se muestran los resultados al intercambiar a Ederson (3.191 min. jugados), arquero del equipo campeón del torneo por Jack Butland (3.150 min. jugados), arquero del Stoke City. En la Figura 5.7 se muestra la comparación del rendimiento de ambos arqueros en el torneo.

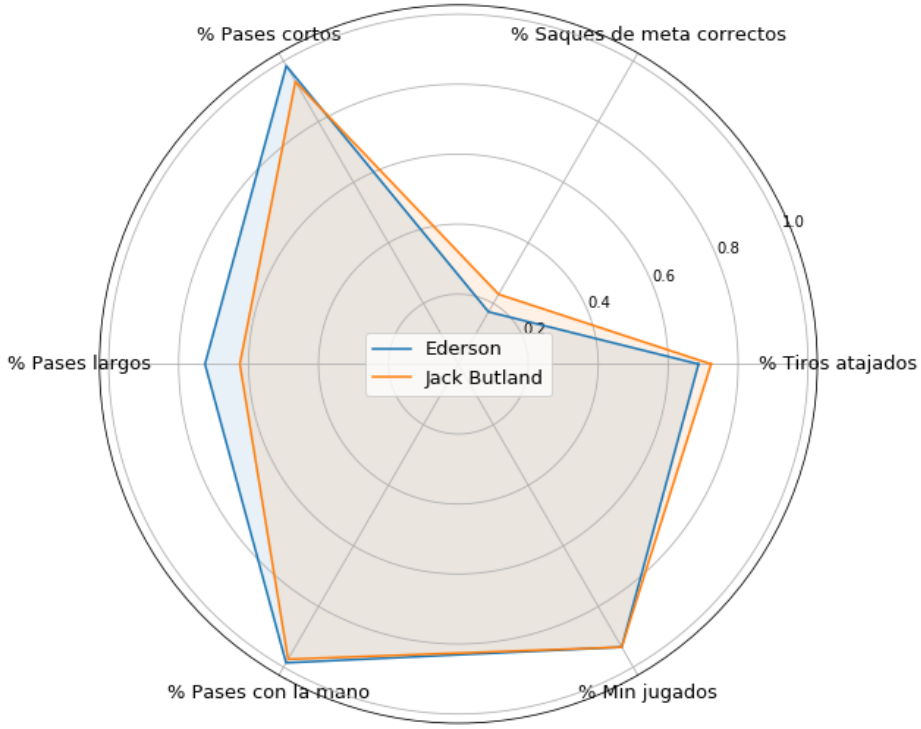

Figura 5.7: Comparación de rendimiento en el torneo de los arqueros Ederson (Manchester City) y Jack Butland (Stoke City).

Es interesante analizar que el rudimento de ambos es muy parejo, siendo que ambos pertenecen a equipos de rendimientos desiguales en el torneo. Por un lado, Ederson supera a Butland en caracteristicas de posesión de balón, como el porcentaje de pases cortos, pases largos y pases con la mano correctos, no obstante, Butland lo supera en el porcentaje de tiros atajados y saques de meta correctos. Este intercambio, como lo muestra la Tabla 5.12, al ser jugadores muy parecidos, no genera mayores diferencias en la probabilidad de campeonar para el Stoke City, en cambio, el Manchester City se ve beneficiado con un aumento de 29,20 puntos porcentuales. Esto se puede deber, a que Butland podría se un arquero que se ajuste mejor al esquema del equipo campeón. Sin desmerecer al gran rendimiento de Ederson a quién solo le convirtieron 27 goles, muy por debajo de los 68 que recibió el Stoke City. Para ver el detalle completo de las probabilidades, ver Anexo C.4.

Tabla 5.12: Probabilidades de campeonar intercambiando a Ederson por Butland.

Equipo	Probs. sin cambio	Probs. con cambio	Delta
Manchester City	20,50	49,70	+ 29,20
Stoke City	0,26	0,00	- 0,26

5.3.4.2. David De Gea por Nick Pope

El mismo fenómeno ocurre cuando se intercambian los porteros David De Gea, elegido como el mejor arquero de la liga, con tan solo 28 goles recibidos y el arquero de Burnley, Nick Pope a quien le embocaron 39 goles. Los resultados del intercambio se muestran a continuación y la Figura 5.8 hace una comparación del rendimiento deportivo entre ambos futbolistas durante el torneo.

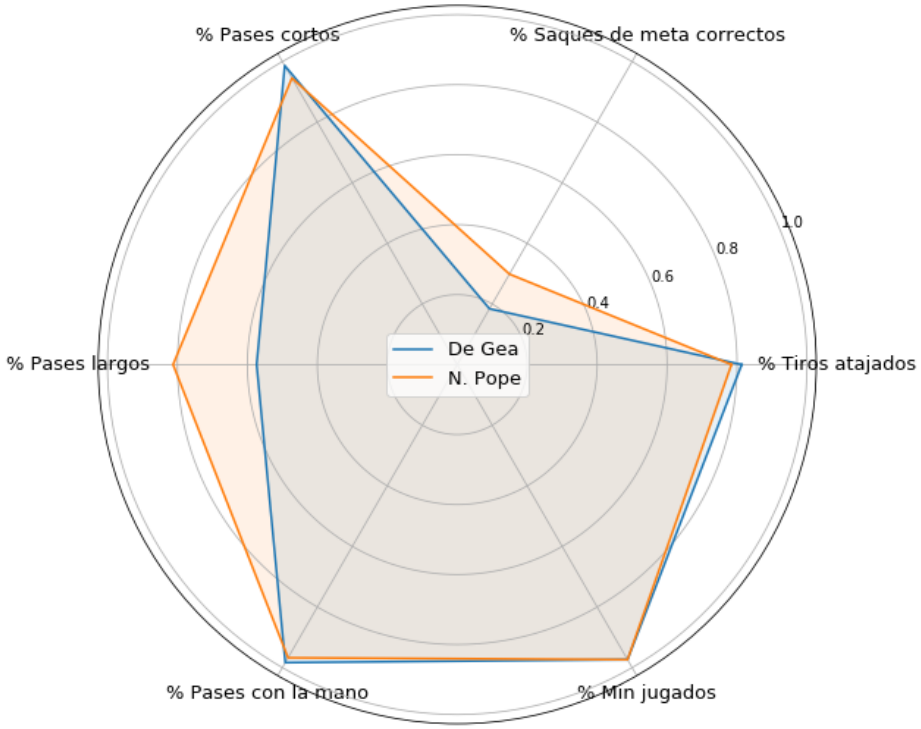

Figura 5.8: Comparación de rendimiento en el torneo de los arqueros David De Gea (Manchester United) y Nick Pope (Burnley).

Si bien, David De Gea fue elegido como el mejor portero del campeonato, al analizar su rendimiento en comparación con Nick Pope, queda en evidencia que este último lo supera en algunos parámetros. En el porcentaje de pases largos correctos y el porcentaje de saques de meta correctos Pope le saca una amplia ventaja a De Gea, a pesar de ello, este último lo supera levemente en el porcentaje de pases cortos, pases con la mano y tiros atajados. Estas diferencias entre ambos hacen que al realizar el intercambio se obtengan las siguientes probabilidades de campeonar, como lo muestra la Tabla 5.13. La salida de De Gea y el ingreso de Pope, hace que el Manchester United aumente en 23,60 puntos porcentuales sus chances

de campeonar, mientras que el Burnley disminuye sus probabilidades en 3,72 puntos porcentuales. Este extraño resultado nuevamente se puede adjudicar a que Pope podría resultar un mejor portero para el esquema de juego del Manchester United, sin embargo por la magnitudes de las diferencias de probabilidad esto se puede deber a otros aspectos no observados en esta tesis y que se dejan propuestos para trabajos posteriores. Para ver las probabilidades de campeonar de todos los equipos luego de este intercambio, ver Anexo C.4

Tabla 5.13: Probabilidades de campeonar intercambiando a De Gea por Pope.

Equipo	**Probs. sin cambio**	**Probs. con cambio**	**Delta**
Manchester United	31,20	54,80	+ 23,60
Burnley	4,12	0,40	- 3,72

Capítulo 6

Conclusiones

En esta tesis, se utilizan datos de partidos de la primera división del fútbol inglés del año 2017-2018 (Premier League 2017-2018) a través del proveedor de datos de fútbol *Wyscout*, que fueron puestos a disposición de la comunidad científica por Luca Pappalardo. Esta repositorio contenía resultados de los partidos además de una bitácora de eventos de cada uno de los partidos disputados durante el torneo. En el capítulo de Datos de esta tesis, se desarrolla un análisis exploratorio de datos, donde se pudo determinar que acciones del juego son las que más ocurren durante el torneo y aquellas que son las más influyentes en relación al resultado final del partido. También, se logró observar las distribuciones de acciones de cada uno de los jugadores condicionadas a la zona del campo de juego donde se realiza dicha acción, si se está jugando de local o visita, el tiempo de juego cuando se realiza la acción, entre otras.

Se logró construir un modelo de simulación que permite evaluar jugadores potencialmente con buenos rendimientos para reforzar el club en base a los datos de rendimiento de las temporadas anteriores de cada jugador. Se consiguió con este modelo un rendimiento predictivo del resultado final del partido un poco menor que el modelo de Poisson (39,37 % vs. 44,45 %), pero un porcentaje de acierto en la diferencia de goles de los partidos mayor que Poisson (27,61 % vs. 25,97 %) y que permite computar el aumento o disminución de la probablidad de ser campeón al incluir en el plantel a jugadores en distintos clubes: Harry Maguire (Leicester City) por Rob Holding (Arsenal) aumentó la probabilidad de campeonar del Arsenal en +1,16 puntos porcentuales, David Silva por Dele Alli aumentó en +20,58 % las probabilidades de campeonar del Tottenham, Kevin de Bruyne en el West Ham United en +1,83 %, Romelu Lukaku en el Chelsea aumentó en +5,87 %, jugador que actualmente se encuentra en ese club. Otros casos también, como el de Sergio Agüero en el Arsenal que aumentó las probabilidades de campeonar en +2,46 % o Mohamed Salah quien aumenta las probabilidades del Stoke City en +6,34 %.

Sin embargo, para el caso de los arqueros los resultados obtenidos tienen menos sentido ya que jugadores como Ederson o David De Gea, ambos arqueros de los equipos que terminaron en la primera y segunda posición respectivamente al probarlos en equipos de las últimas posiciones tuvieron resultados contrarios a los esperados: al intercambiar a Ederson (Manchester City) por Jack Butland (Stoke City) las probabilidades aumentaron para el campeón en +29,2 puntos porcentuales y la inclusión de Ederson en el Stoke City hicieron que este ultimo disminuyera sus probabilidades en -0,26 puntos porcentuales. Lo mismo ocurre al in-

tercambiar a De Gea por Pope, donde el Manchester United vio un aumento de +23,60 % y el Burnley disminuyó en -3,72 puntos porcentuales.

Se cumplió el objetivo general de cuantificar el efecto de un jugador en las probabilidades de campeonar de un equipo de fútbol, en base al rendimiento deportivo de los jugadores en la temporada 2017-2018 de la Premier League con datos proporcionados por *Wyscout*.

Para consumar el objetivo general, se tuvo que cumplir con cada uno de los objetivos específicos. Se logró consultar correctamente los datos proporcionados por *Wyscout*, a través de su API, para luego realizar un análisis exploratorio de los distintos *feeds*. Por otro lado, se logró reconocer el lenguaje de programación C++ con el que se desarrolló un modelo de Inferencia Bayesiana utilizando el programa *Stan* para estimar tanto las probabilidades de elección como de ejecución de los jugadores para el modelo de simulación. Además, se logró construir un modelo de Cadenas de Markov para simular todos los eventos de un partido y todos los partidos del torneo. Ese modelo fue programado correctamente permitiendo la realización de un método de *Cross Validation* para verificar su eficiencia y compararlo con el modelo de Poisson. Por último, se logró cuantificar el efecto de cualquier jugador de campo en un equipo de fútbol, en la probabilidad de campeonar de los equipos.

Desde el punto de vista técnico, se pueden realizar mejoras al modelo de simulación de Cadenas de Markov a través de, precisar las distribuciones a priori de los parámetros que determinan las distintas probabilidades, la aplicación de un ponderador que permita capturar el efecto de los partidos importantes o el nivel propio del equipo, agregar eventos a los tiempos muertos, como la incorporación de un evento asociado al "traslado del balón" o la realización de sustituciones durante el partido.

Visto desde un punto de vista comercial, este modelo puede hacerse aún más eficiente en relación a lo que el cliente busca, por ejemplo, se pueden incorporar filtros que permitan reducir el espectro de jugadores a observar para luego verificar con el modelo él o los jugadores que potencialmente tendrían un mejor rendimiento en el equipo. Por otra parte, la vinculación de este tipo de trabajos del área técnica con variables relacionadas al aspecto médico, fisiológico, psicológico, nutricional y otros ámbitos que tienen relación con la práctica deportiva harían de este, un modelo aun más robusto con resultados aún más precisos. Incluso, desde una mirada del negocio, se puede construir un modelo de optimización utilizando restricciones presupuestarias para la contratación de jugadores, o restricciones reglamentarias, como la restricción del número de jugadores extranjeros o número de minutos jugados por juveniles del club que permitirían estructurar las negociaciones y la toma de decisiones del equipo en cada periodo de traspasos.

Esto último queda propuesto al lector, ya que sería muy interesante estructurar en base a los datos relacionados al rendimiento deportivo de los jugadores, y según las restricciones presupuestarias y de reglamentación que rodea a los equipos, para así ahorrar recursos de los clubes y hacer más eficiente el proceso de *scouting* que hasta la fecha, al menos en América Latina es muy primitivo.

Bibliografía

[1] Etimologías de Chile. Etimología de Dato.
http://etimologias.dechile.net/?dato [Consulta: 13/03/2021]

[2] Definición de análisis de datos. Wikipedia.
https://es.wikipedia.org/wiki/An%C3%A1lisis_de_datos#.

[3] Definición análisis de datos de Judd, Charles; McCleland, Gary (1989).
https://archive.org/details/dataanalysismode0000judd/page/n3/mode/2up.

[4] Valor del patrimonio de Jeff Bezos según la revista Forbes. [Consulta: 21/05/2021]
https://www.forbes.com/profile/jeff-bezos/?sh=14a12a181b23

[5] *Amazon: ¿por qué debería preocuparnos todo lo que la compañía de Jeff Bezos sabe sobre nosotros?*, Leo Kelion. (2020).
https://www.bbc.com/mundo/noticias-51546041

[6] *ThreePoints "The School for digital business": Big Data aplicado a los deportes.*
https://www.threepoints.com/int/big-data-aplicado-a-los-deportes

[7] *Moneyball: the art of winning an unfair game*, Michael Lewis. (2004)
https://www.amazon.com/-/es/Michael-Lewis/dp/0393324818

[8] *Moneyball and soccer - an analysis of the key performance indicators of elite male soccer players by position*, Michel Hughes, Tim Caudrelier, Nic James, Athalie Redwood-Brown, Ian Donnelly, Anthony Kirkbride y Chrsitophe Duschesne. (2012)

[9] *How Data (and Some Breathtaking Soccer) Brought Liverpool to the Cusp of Glory*, The New York Times. (2019).
https://www.nytimes.com/es/2019/05/29/liverpool-champions/. [Consulta: 13/03/2021]

[10] Sevilla and the Science of Soccer's Summer Transfer Window. (2019)
https://www.nytimes.com/2019/08/02/sports/transfer-window-sevilla-monchi.html.

[11] Página web de Wyscout. Wyscout.
https://wyscout.com/es/

[12] Pappalardo, Luca; Massucco, Emanuele (2019): Soccer match event dataset. figshare. Collection. https://doi.org/10.6084/m9.figshare.c.4415000.v5

[13] Página web del Soccer Data Challenge. SDC.
https://sobigdata-soccerchallenge.it/

[14] Historia de Wyscout. Wikipedia.
https://en.wikipedia.org/wiki/Wyscout [Consulta: 22/04/2021]

[15] Premier League 2017-2018. Wikipedia.
https://es.wikipedia.org/wiki/Premier_League_2017-18 [Consulta: 04/04/2021]

[16] Biografía Mohamed Salah. Wikipedia.
https://es.wikipedia.org/wiki/Mohamed_Salah#Temporada_2017/18 [Consulta: 04/04/2021]

[17] Cadena de Markov. Wikipedia.
https://es.wikipedia.org/wiki/Cadena_de_M%C3%A1rkov [Consulta: 06/04/2021]

[18] Thomas Bayes y el Teorema de Bayes.
https://tereom.github.io/est-computacional-2018/regla-de-bayes-e-inferencia-bayesiana.html

[19] *Descripción general de la Inferencia Bayesiana y sus aplicaciones en los procesos de gestión*, Lesley Mesa, Miller Rivera, Jesús Romero. (2011)
https://www.urosario.edu.co/Administracion/documentos/investigacion/laboratorio/miller_2_2.pdf

[20] Prior Conjugado. Wikipedia.
https://en.wikipedia.org/wiki/Conjugate_prior#Example [Consulta: 27/05/2021]

[21] Distancia Euclidiana. Wikipedia.
https://es.wikipedia.org/wiki/Distancia_euclidiana [Consulta: 26/06/2021]

[22] Descripción mecanismo de trabajo de Stan, Tereza Ortiz. (2017)
https://tereom.github.io/est_computacional/13-Stan.html#:~:text=Stan%20genera%20muestras%20de%20la,para%20cualquier%20valor%20candidato%20%CE%B8. [Consulta: 22/06/2021]

[23] Monte Carlo Hamiltoniano. Wikipedia.
https://en.wikipedia.org/wiki/Hamiltonian_Monte_Carlo [Consulta: 22/05/2021]

[24] *Doing Bayesian Data Analysis*, John K. Kruschke. (2012)
https://sites.google.com/site/doingbayesiandataanalysis/1st-ed-stuff?authuser=0

[25] Multi-Logistic Regression With Probabilistic Programming, Krishna Yerramsetty. (2020)
https://medium.com/swlh/multi-logistic-regression-with-probabilistic-programming

Anexo A

Datos

A.1. Minutos jugados por jugador

En ambas tabla se muestran los 10 jugadores con más minutos (izquierda) y menos minutos (derecha) jugados durante el torneo.

Tabla A.1: Jugadores con mayor y menor cantidad de minutos jugados en el torneo.

Jugador	Minutos
Lewis Dunk	3.420
Jack Cork	3.420
Alfie Mawson	3.420
Asmir Begovic	3.420
Jordan Pickford	3.420
Mathias Jørgensen	3.420
Jonas Lössl	3.420
Harry Maguire	3.420
Lukasz Fabiański	3.420
Mathew Ryan	3.420

Jugador	Minutos
Dwight McNeil	0
Vincent Janssen	0
Axel Tuanzebe	0
Reece Oxford	4
Massadio Haïdara	5
Ben Woodburn	6
Harvey Barnes	7
Josh Cullen	7
Pape N'Diaye Souare	8
Michael Obafemi	8

A.2. Eventos por jugador

En ambas tablas se muestran los 10 jugadores con más eventos (izquierda) y menos eventos (derecha) en donde participa durante el torneo.

Tabla A.2: Jugadores con mayor y menor cantidad de eventos asociados.

Jugador	# Eventos
Granit Xhaka	4.060
Kevin De Bruyne	3.972
Nicolás Otamendi	3.938
Fernando Luiz Rosa	3.868
César Azpilicueta	3.710
Nemanja Matic	3.591
Christian Eriksen	3.408
Jan Vertonghen	3.393
Kyle Walker	3.388
Abdoulaye Doucouré	3.313

Jugador	# Eventos
Dwight McNeil	1
Vincent Janssen	1
Massadio Haïdara	2
Michael Obafemi	5
Reece Oxford	6
Divock Origi	7
Jeremy Boga	7
Lukas Nmecha	8
Josh Cullen	8
Rekeem Harper	8

Anexo B

Modelo

B.1. Código modelo *Logit* en Stan

El código asociado al modelo Logit que se ingresa en Stan en el lenguaje de programación C++ se muestra a continuación:

```
goals_model = """
data {
    int<lower=0> N; // number of observations (8238 solo jugadores con >5) (8451 todos los tiros)
    int players; // number of players (331 solo jugadores con >5) (426 todos los jugadores con tiros)
    int glk; // number of goalkeepers 38
    int zones; // number of field zones 8
    int time; // number of time frames 7
    int res; // types of results (winning, losing, tying)
    int loc; // localia

    int<lower=1,upper=players> player_id[N];
    int<lower=1,upper=glk> glk_id[N];
    int<lower=1,upper=zones> cat_zone[N];
    int<lower=1,upper=time> time_frame[N];
    int<lower=1,upper=res> cat_res[N];
    int<lower=1,upper=loc> localia[N];

    int goal[N]; // dependent variable
}
parameters {

    real alpha; // intercept

    vector[players] beta_player; // coefficient associated with each player
    vector[glk] beta_glk; // coefficient associated with each goalkeeper
    vector[zones] beta_zones; // coefficient associated with each zone
    vector[time] beta_time; // coefficient associated with each time frame
    vector[res] beta_res; // coefficient associated with each result
    vector[loc] beta_loc; // coefficient associated with each type of localia

    real epsilon; //Uncertainty / unexplained variance
}
model {
    // priors
    alpha ~ normal(0,1);
    beta_player ~ normal(0,1);
    beta_glk ~ normal(0,1);
    beta_zones ~ normal(0,1);
    beta_time ~ normal(0,1);
    beta_res ~ normal(0,1);
    beta_loc ~ normal(0,1);

    goal ~ bernoulli_logit(alpha + beta_player[player_id] + beta_glk[glk_id] +
        beta_zones[cat_zone] + beta_time[time_frame] + beta_res[cat_res] + beta_loc[localia]);

}
"""
```

Figura B.1: Fotografía del código Logit escrito en lenguaje C++ que ingresa a la plataforma integrada de Stan, PyStan.

B.2. Código modelo *Multi-Logit* en Stan

El código asociado al modelo Multi-Logit que se ingresa a la plataforma Stan en el lenguaje de programación C++ se muestra a continuación:

```
data {
    int<lower=0> Npairs;
    int Nplayers; // number of players 515
    int Nzones; // number of zones 12
    int Ntimes; // number of times 7
    int Nres; // number of results 3
    int Nloc; // number of localias 2

    int<lower=1,upper=Nplayers> player_id[Npairs];
    int<lower=1,upper=Nzones*Nloc*Nres*Ntimes> pred_index[Npairs]; //pred_index replaces (zone_id, loc_id, res_id, time_id)

    int<lower=0> actions[Npairs,3];
}
transformed data {
    int Nall = Nzones*Nloc*Nres*Ntimes; // 12*2*3*7 = 504
    int Kz = Nloc*Nres*Ntimes; // 2*3*7 = 42
    int Kl = Nres*Ntimes; // 3*7 = 21
    int Kr = Ntimes; // 7
}
parameters {
    matrix[2,Nplayers] beta_player;
    matrix[2,Nzones] beta_zone;
    matrix[2,Nloc] beta_loc;
    matrix[2,Nres] beta_res;
    matrix[2,Ntimes] beta_time;
}
model {
    matrix[2,Nall] beta_all;
    for (z in 1:Nzones) {
        vector[2] bz = beta_zone[:,z];
        for (l in 1:Nloc) {
            vector[2] bl = bz + beta_loc[:,l];
            for (r in 1:Nres) {
                vector[2] br = bz + beta_res[:,r];
                for (t in 1:Ntimes) {
                    beta_all[:,(z-1)*Kz+(l-1)*Kl+(r-1)*Kr+t] = br + beta_time[:,t];
                }
            }
        }
    }
    to_vector(beta_player) ~ std_normal();
    to_vector(beta_zone) ~ std_normal();
    to_vector(beta_loc) ~ std_normal();
    to_vector(beta_res) ~ std_normal();
    to_vector(beta_time) ~ std_normal();

    for (i in 1:Npairs) {
        vector[2] beta = beta_player[:,player_id[i]] + beta_all[:,pred_index[i]];
        actions[i] ~  multinomial(softmax(append_row(0.0, beta)));
    }
```

Figura B.2: Fotografía del código del modelo Multi-Logit escrito en lenguaje C++ que ingresa a la plataforma integrada de Stan, PyStan.

Anexo C

Resultados

C.1. Cambio de probabilidades: Defensas

En la siguiente tabla se muestra el detalle de todas las probabilidades de campeonar de todos los equipos, considerando el cambio realizado por Harry Maguire y Rob Holding.

Tabla Real		**Maguire por Holding**	
Equipo	**Probs.**	**New Prob.**	**Delta**
Manchester City	20,50	20,40	- 0,10
Manchester United	31,20	31,10	- 0,10
Tottenham Hotspur	18,52	18,50	- 0,02
Liverpool	9,29	9,00	- 0,29
Chelsea	5,53	6,00	+ 0,47
Arsenal	1,24	**2,40**	**+ 1,16**
Burnley	4,12	3,90	- 0,22
Everton	0,81	1,00	+ 0,19
Leicester City	7,45	**6,70**	**- 0,75**
Newcastle United	0,19	0,20	+ 0,01
Crystal Palace	0,01	0,00	- 0,01
Bournemouth	0,11	0,10	- 0,01
West Ham United	0,67	0,60	- 0,07
Watford	0,00	0,00	0,00
Brighton & H.A.	0,04	0,00	- 0,04
Huddersfield Town	0,00	0,00	0,00
Southampton	0,02	0,00	- 0,02
Swensea City	0,00	0,00	0,00
Stoke City	0,26	0,10	- 0,16
West Bromwich A.	0,04	0,00	- 0,04
	100,00	**100,00**	

C.2. Cambio de probabilidades: Mediocampistas

En la siguiente tabla se muestra el detalle de todas las probabilidades de campeonar de todos los equipos, considerando los cambios de: David Silva por Dele Alli y Kevin De Bruyne por Mark Noble.

Tabla Real		**Silva por Alli**		**De Bruyne por Noble**	
Equipo	**Probs.**	**New Prob.**	**Delta**	**New Prob.**	**Delta**
Manchester City	20,50	**10,60**	**- 9,90**	**15,00**	**- 5,50**
Manchester United	31,20	26,40	- 4,80	33,00	+ 1,80
Tottenham Hotspur	18,52	**39,10**	**+ 20,58**	17,40	- 1,12
Liverpool	9,29	7,00	- 2,29	10,70	+ 1,41
Chelsea	5,53	6,10	+ 0,57	5,20	- 0,33
Arsenal	1,24	0,80	- 0,44	1,20	- 0,04
Burnley	4,12	4,20	+ 0,08	4,90	+ 0,78
Everton	0,81	0,80	- 0,01	0,90	+ 0,09
Leicester City	7,45	5,80	- 1,65	8,80	+ 1,35
Newcastle United	0,19	0,20	+ 0,01	0,00	- 0,19
Crystal Palace	0,01	0,00	- 0,01	0,00	- 0,01
Bournemouth	0,11	0,20	+ 0,09	0,10	- 0,01
West Ham United	0,67	0,10	- 0,57	**2,50**	**+ 1,83**
Watford	0,00	0,00	0,00	0,00	0,00
Brighton & H.A.	0,04	0,00	- 0,04	0,00	- 0,04
Huddersfield Town	0,00	0,00	0,00	0,00	0,00
Southampton	0,02	0,00	- 0,02	0,00	- 0,02
Swensea City	0,00	0,00	0,00	0,00	0,00
Stoke City	0,26	0,10	- 0,16	0,30	+ 0,04
West Bromwich A.	0,04	0,00	- 0,04	0,00	-0,04
	100,00	**100,00**		**100,00**	

C.3. Cambio de probabilidades: Delanteros

En la siguiente tabla se muestra el detalle de todas las probabilidades de campeonar de todos los equipos, considerando los cambios de delanteros: Romelu Lukaku por Álvaro Morata, Sergio Agüero por Alexandre Lacazette y Mohamed Salah por Mame Diouf.

Tabla Real		**Lukaku - Morata**		**Agüero - Lacazette**		**Salah - Diouf**	
Equipo	**Prob**	**Prob**	**Delta**	**Prob**	**Delta**	**Prob.**	**Delta**
Manchester C.	20,50	23,60	+ 3,10	**13,70**	**- 6,80**	20,30	- 0,20
Manchester U.	31,20	**18,10**	**- 13,10**	31,90	+ 0,70	31,70	+ 0,50
Tottenham H.	18,52	20,70	+ 2,18	19,60	+ 1,08	19,50	+ 0,98
Liverpool	9,29	10,10	+ 0,81	9,50	+ 0,21	**0,40**	**- 8,89**
Chelsea	5,53	**11,40**	**+ 5,87**	6,10	+ 0,57	5,00	- 0,53
Arsenal	1,24	0,60	- 0,64	**3,70**	**+ 2,46**	1,20	- 0,04
Burnley	4,12	4,90	+ 0,78	4,80	+ 0,68	4,90	+ 0,78
Everton	0,81	1,00	+ 0,19	1,20	+ 0,39	1,20	+ 0,39
Leicester C.	7,45	8,40	+ 0,95	8,50	+ 1,05	8,00	+ 0,55
Newcastle U.	0,19	0,20	+ 0,01	0,10	- 0,09	0,20	+ 0,01
Crystal P.	0,01	0,00	- 0,01	0,00	- 0,01	0,00	- 0,01
Bournemouth	0,11	0,00	- 0,11	0,00	-0,11	0,30	+ 0,19
West Ham U.	0,67	0,50	- 0,17	0,70	+ 0,03	0,70	+ 0,03
Watford	0,00	0,00	0,00	0,00	0,00	0,00	0,00
Brighton H.A.	0,04	0,10	+ 0,06	0,00	- 0,04	0,00	- 0,04
Huddersfield	0,00	0,00	0,00	0,00	0,00	0,00	0,00
Southampton	0,02	0,00	- 0,02	0,00	- 0,02	0,00	- 0,02
Swensea City	0,00	0,00	0,00	0,00	0,00	0,00	0,00
Stoke City	0,26	0,30	+ 0,04	0,10	- 0,16	**6,60**	**+ 6,34**
West Brom. A.	0,04	0,10	+ 0,06	0,10	+ 0,06	0,00	-0,04
	100,00	**100,00**		**100,00**		**100,00**	

C.4. Cambio de probabilidades: Arqueros

En la siguiente tabla se muestra el detalle de todas las probabilidades de campeonar de todos los equipos, considerando los cambios de arqueross: REderson por Jack Butland y David De Gea por Nick Pope.

Tabla Real		**Ederson por Butland**		**De Gea por Pope**	
Equipo	**Prob**	**Prob**	**Delta**	**Prob**	**Delta**
Manchester C.	20,50	**49,70**	**+ 29,20**	15,30	- 5,20
Manchester U.	31,20	18,40	- 12,80	**54,80**	**+ 23,60**
Tottenham H.	18,52	13,30	- 5,22	13,40	- 5,12
Liverpool	9,29	5,40	- 3,89	5,70	- 3,59
Chelsea	5,53	3,00	- 2,53	3,60	- 1,93
Arsenal	1,24	0,90	- 0,34	1,00	- 0,24
Burnley	4,12	2,00	- 2,12	**0,40**	**- 3,72**
Everton	0,81	0,40	- 0,41	0,30	- 0,51
Leicester C.	7,45	6,40	- 1,05	5,20	- 2,25
Newcastle U.	0,19	0,20	+ 0,01	0,10	- 0,09
Crystal P.	0,01	0,00	- 0,01	0,00	- 0,01
Bournemouth	0,11	0,00	- 0,11	0,00	- 0,11
West Ham U.	0,67	0,20	- 0,47	0,10	- 0,57
Watford	0,00	0,00	0,00	0,00	0,00
Brighton H.A.	0,04	0,10	+ 0,06	0,00	- 0,04
Huddersfield	0,00	0,00	0,00	0,00	0,00
Southampton	0,02	**0,00**	**- 0,02**	0,00	- 0,02
Swensea City	0,00	0,00	0,00	0,00	0,00
Stoke City	0,26	0,00	- 0,26	0,10	- 0,16
West Brom. A.	0,04	0,00	- 0,04	0,00	- 0,04
	100,00	**100,00**		**100,00**	