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

DATAZUL: UN PRIMER CASO DE ANALYTICS APLICADO AL FÚTBOL PROFESIONAL EN CHILE

TESIS PARA OPTAR AL GRADO DE MAGÍSTER EN GESTIÓN DE OPERACIONES

MEMORIA PARA OPTAR AL TÍTULO DE INGENIERO CIVIL INDUSTRIAL

PABLO ANDRÉS GALAZ CARES

PROFESOR GUÍA:
DENIS SAURÉ VALENZUELA

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

Este trabajo ha sido parcialmente financiado por:
CONICYT/Proyecto FONDECYT/1181513

SANTIAGO, CHILE
2020

RESUMEN DE LA MEMORIA PARA OPTAR
AL TÍTULO DE MAGÍSTER EN
GESTIÓN DE OPERACIONES
POR: **PABLO ANDRÉS GALAZ CARES**
FECHA: 2020
PROF. GUÍA: DENIS SAURÉ VALENZUELA

DATAZUL: UN PRIMER CASO DE ANALYTICS APLICADO AL FÚTBOL PROFESIONAL EN CHILE

Durante los últimos años, los clubes de fútbol de Europa han desarrollado sistemas basados en análisis de datos para apoyar la gestión de contratación de jugadores tal como muestra el New York Times en dos de sus artículos: "El arma secreta del Liverpool" [4] y "La ciencia del mercado de transferencia" (Sevilla FC) [15]. Es por lo anterior, que el club de fútbol Universidad de Chile decidió investigar e implementar este tipo de sistemas en su gestión.

El objetivo principal de este estudio es crear un sistema de recomendación de jugadores para el club de fútbol Universidad de Chile, en base al rendimiento deportivo de los jugadores en la temporada 2019 del torneo chileno con datos proporcionados por Opta Sports.

La metodología consta de 3 etapas. En primera instancia, se realiza la definición de variables a considerar para medir el rendimiento de los jugadores en cada una de las posiciones del campo de juego y la importancia relativa de cada una de ellas, en conjunto con directores y analistas de la secretaría técnica del club. En una segunda etapa, se realiza un modelo de recomendación basado en estadística descriptiva en el cual se evalúa cada jugador de acuerdo al *Score* calculado en base a las variables definidas y el rendimiento propio del jugador en cada una de las posiciones. Finalmente, la tercera etapa consta del modelamiento de un sistema de recomendación basado en simulaciones. Para esto se crea un modelo de Cadenas de Markov, basado en el *estilo de juego* de cada uno de los jugadores en distintas zonas del terreno de juego y se compara el rendimiento predictivo de resultados con el modelo de Poisson [12].

El resultado del modelo basado en estadística descriptiva es una lista jugadores en base a su *Score* para cada una de las posiciones, donde destacan Jorge Valdivia (Colo Colo) como el mejor volante ofensivo del fútbol chileno o José Pedro Fuenzalida (U. Católica) como el mejor delantero del torneo. En el caso del modelo basado en Cadenas de Markov, tiene un poder predictivo un poco menor que el de Poisson (33.7 % vs 35.8 %) y un error un poco mayor (43,894 vs 39,549). Además, los jugadores que más benefician al club son Lucas Passerini (Palestino), Jorge Valdivia (Colo Colo) y José Pedro Fuenzalida (U. Católica) con un aumento en la probabilidad de ser campeón de un 7.61 %, 3.25 % y 2.66 %, respectivamente.

Se concluye que el modelo propuesto basado en Cadenas de Markov, a pesar de tener un rendimiento predictivo levemente peor que el modelo estándar de Poisson, permite identificar individualmente qué jugadores aportan a obtener un mejor rendimiento colectivo y, por otra parte, que la implementación de este tipo de sistemas en el club permite tomar decisiones basadas en métricas objetivas, reduciendo el error y los sesgos del juicio de experto y, además, puede traer beneficios tanto en lo deportivo como en lo financiero.

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

Actualmente está muy desarrollado el campo del análisis de datos para tomar decisiones. En términos formales, John Tukey en 1961 definió el análisis de datos como "*... los procedimientos para analizar datos, técnicas para interpretar los resultados de dichos procedimientos, las formas de planear la recolección de datos para hacer un análisis más fácil, más preciso o exacto*" [1].

Haciendo el simple ejercicio de utilizar el buscador del portal web del New York Times con el término *Data Analysis*, podemos observar la gran variedad de noticias relacionadas a toma de decisiones de organizaciones privadas dedicadas al retail, transporte, tecnología, entre otras y, otras noticias relacionadas a análisis de expertos en temas económicos, políticos y sociales de los distintos países de mundo [2]. Adicionalmente, aunque en menor medida, existen aplicaciones de análisis de datos en organizaciones ligadas al deporte para mejorar su rendimiento financiero y deportivo.

En el año 2003 se publica el libro *Moneyball: the art of winning an unfair game* [3], se dio a conocer la historia de un equipo con bajo presupuesto y poca historia, como los Oakland Athletics de la Major League Baseball (MLB) de los Estados Unidos, logra competir de igual a igual con las grandes franquicias del béisbol estadounidense a través un nuevo e innovador método para evaluar a los jugadores, en desmedro de la intuición y expertiz de los antiguos analistas. Así es como Billy Bean, entrenador del equipo le da la oportunidad a un economista de la Universidad de Yale, llamado Peter Brand para reinventar la forma de contratar jugadores. Esta nueva forma le entrega tremendos resultados deportivos lo que generó que las grandes franquicias de los distintos deportes, hayan ido transformando y profesionalizando el proceso de contratación de jugadores (scouting) para reducir el error al contratar nuevas figuras para sus equipos.

El día 29 de mayo de 2019, el New York Times publicó un artículo titutlado "El arma secreta del Liverpool: el análisis de datos" [4], el cual relata la anécdota vivida entre el director de investigación del Liverpool, Ian Graham, y el entrenador del primer equipo del Liverpool, Jurgen Klopp. En esta anécdota, Graham le revela que no necesitó ver partidos del Borussia Dortmund (ex-equipo de Jurgen Klopp), para darse cuenta de la calidad de entrenador que él era, a pesar de los malos resultados obtenidos en la temporada 2013-2014 y, que era una gran oportunidad de ficharlo para el club. El Liverpool actualmente cuenta con un equipo de 5 analistas entre los que tiene a un PhD en Física y un ex-campeón juvenil de ajedrez

para analizar a rivales y futuras contrataciones con sofisticados modelos matemáticos, lo que le ha traído grandes resultados desde lo deportivo, ya que son los actuales campeones de la Champions League (torneo de clubes más importante del mundo) y corre con gran ventaja en la Premier League de Inglaterra frente a grandes clubes como Manchester City FC de Pep Guardiola, Totenham Hotspur FC de José Mourinho o Manchester United FC de Ole Gunnar Solskjær.

En la literatura existen 2 tipos de estudios: relacionado al rendimiento individual de jugadores y relacionado al rendimiento colectivo de equipos. Hughes et al [5], realizaron una serie de focus group con expertos del fútbol, para definir los indicadores clave (KPI) en 5 diferentes categorías (fisiológicas, tácticas, técnicas defensivas, técnicas ofensivas y sicológicas) para medir en rendimiento de los futbolitstas en 7 posiciones del campo de juego (arquero, defensa central, defensa lateral, mediocampista central, mediocampista ofensivo, mediocampista exterior y delantero). En general, en las categorías técnicas se mencionaban características similares pero en distinto orden de importancia, por ejemplo, precisión en los pases en los mediocampistas era fundamental pero en los defensas centrales no era primordial.

Pappalardo et al [6] especifica 2 medidas para analizar el rendimiento de jugadores en forma individual. Estas medidas buscan representar la importacia relativa de un jugador en su equipo y el rendimiento de un jugador en un partido. Las medidas son la Centralidad de Flujo (*Flow Centrality*) y el Puntaje de Rankeo de Jugadores (*Player Rank Score).* La primera medida refleja, en forma agregada por temporada, la importancia relativa de un jugador en su equipo considerando medidas de redes y la segunda medida es un indicador que muestra el puntaje (*score*) que representa el rendimiento de un jugador en cada uno de sus partidos. En temas de análisis de equipos, se especifican otros 2 indicadores: el Índice de Invasión (*Invasion Index*) y el Índice de Aceleración (*Aceleration Index*). El primero, mide qué tan cerca del arco rival juega un equipo durante un partido y el segundo, es la velocidad con la que un equipo llega a ciertas posiciones del campo rival.

Continuando con análisis del estilo de juego de los equipos, Decroos et al [7] caracterizaron secuencias de pases de los equipos de fútbol, para luego clusterizarlas de acuerdo a variables espaciales (zonas del campo de juego) y temporales (momento del partido en que se realiza). Posterior a la clusterización, se ranquean las secuencias clusterizadas en base a cantidad de pases y remates al arco, para determinar cuáles son las más utilizadas por los equipos. Este análisis se realizó para comparar los estilos de ataque de Arsenal FC, Leicester City FC y Manchester City FC, clubes de la primera división de Inglaterra.

Wang & Yao [8] extienden el análisis de patrones de juego a través de un modelo de inferencia bayesiana para identificar distintos estilos de ataque de los equipos. Además, se genera un Índice de Goles (*Goal Rate*) para cuantificar la efectividad de conversión de goles a través de cierto patrón de juego. Además, se calculó el rol de cada jugador (*Player Role*) en cada patrón de juego, es decir, la importancia que tiene cierto jugador para cada uno de los patrones identificados. Este análisis fue realizado para todos los partidos del FC Barcelona, obteniendo los patrones que más goles le entregaron al equipo durante la temporada 2013-2014.

En este trabajo, se desea resolver el problema de qué jugadores contratar de acuerdo a

las restricciones de periodos de contratación, presupuesto, cantidad de jugadores a contratar y aminorar los sesgos que pueden existir al tomar decisiones basadas en juicio de experto. Actualmente, en ciertos periodos del año, se abre el periodo de traspasos de jugadores, en el cual los clubes pueden incorporar nuevos jugadores provenientes de otros clubes (de la misma liga nacional o de otros países) a su plantilla. Al contratar un jugador, se pueden presentar dos situaciones: 1) el jugador está en condición de "Jugador Libre", es decir, no tiene contrato actual con ningún otro club o, 2) el jugador tiene contrato vigente con algún otro club. En este segundo caso, el club interesado en contar con los servicios del jugador, tendrá que pagar un precio al club propietario del jugador para tenerlo en su plantilla y, en la mayoría de los casos, los jugadores aspiran a un aumento en su salario al cambiarse de club.

Las personas que eligen a los potenciales jugadores a ser contratados son analistas (también llamados scouts) de un área llamada Secretaría Técnica. El proceso de elección es, principalmente, a través de visualización de videos de partidos de las ligas de otros países para, a través de juicio de experto, generar una lista de candidatos a ser contratados por el club. Esta metodología genera ciertos sesgos a la hora de escoger al refuerzo ideal para el club, dado que los partidos son escogidos al azar y ven una baja cantidad de partidos del mismo club o del mismo jugador, por lo que el rendimiento observado puede tener alta varianza. Además, a los analistas o scouts, les solicitan revisar ciertas posiciones del campo de juego a las cuales destinar sus esfuerzos. Sin embargo, no existen metodologías para determinar qué posición es prioritaria a la hora de buscar un refuerzo.

Es por lo anterior, que no es una decisión trivial qué jugador contratar en cada periodo de traspaso, pues tiene altos costos asociados y los clubes cuentan con presupuestos finitos para el ítem de contratación. Además del costo monetario, existe incertidumbre sobre el rendimiento esperado del jugador: un jugador puede mejorar su rendimiento y ser gran aporte al equipo, incluso puede volver a ser vendido a otro club que esté dispuesto a pagar un mayor valor por su traspaso o, el jugador puede empeorar su rendimiento y no cumplir las expectativas del club lo que se traduce en una pérdida de dinero (pues el jugador disminuye su valor y no se podrá recuperar lo invertido en él) y en el costo de oportunidad de haber contratado un jugador que si cumpliese el rendimiento esperado.

El objetivo general de este trabajo, es generar un modelo de recomendación de jugadores de fútbol basado en características cuantitativas de los jugadores y en indicadores claves (KPI) del rendimiento agregado del equipo. Con el modelo propuesto en este trabajo, se espera determinar cuáles son las posiciones en las que el club está más débil, por lo que se vuelve prioritario reforzarse en esa zona y cuáles son los jugadores que mejor rinden en cada una de las posiciones requeridas. Así se espera quitar ciertos sesgos de la metodología actual para reducir la incertidumbre sobre el rendimiento deportivo al momento de contratar nuevos jugadores.

Para cumplir el objetivo general, se debe acceder a la API de Opta Sports, a través de consultas, para obtener los datos correctamente. Luego, se debe realizar un análisis exploratorio de datos para de los distintos *feeds* que proporciona la API de Opta Sports. Por otra parte, se deben definir variables que permitan medir el rendimiento deportivo de los jugadores en cada una de las posiciones del campo de juego, en conjunto con los colaboradores de Azul Azul S.A. Más adelante, se debe generar un modelo que permita, a través de estadística

descriptiva, obtener una lista de jugadores en base a su rendimiento deportivo. Finalmente, se debe modelo un proceso estocástico que permita representar un partido de fútbol para generar, de una nueva forma, una lista de recomendación de jugadores que permitan mejorar el rendimiento deportivo del club. Para esto, se creará una medida de rendimiento, basado en la probabilidad del club de salir campeón a final del torneo.

Es preciso mencionar que este trabajo es parte del proyecto DatAzul en conjunto con Azul Azul S.A. que abarca otras aristas como por ejemplo, análisis de rivales. Además, este trabajo es una versión simplificada de la realidad de un partido de fútbol y tiene amplios espacios de mejora. El documento muestra una sección en la cual se explica los datos con los que se desarrolló el trabajo, para dar paso a la sección en la que se especifican los modelos de recomendación, los cuales constan de una etapa descriptiva y otra de simulación, para finalmente mostrar la sección de los resultados obtenidos junto a las conclusiones finales. Se podrán encontrar tablas y gráficos adicionales en la sección de apéndice.

Capítulo 2

Datos

En el capítulo presente se explica el origen de los datos, los paquetes de datos disponibles para el trabajo y análisis exploratorio de datos para tener mayor claridad de la estructura y la calidad de los datos con los que se cuenta.

2.1. Datos

2.1.1. Origen de los datos

Los datos utilizados son proporcionados al club de fútbol Universidad de Chile por la compañía británica Opta Sports, a través de un acuerdo comercial entre ambas partes. Los datos a los cuales se tiene acceso, corresponden a partidos y jugadores que participan en el campeonato de primera división del fútbol chileno edición 2019.

El acuerdo entrega acceso a la API (Application Programming Interface) de Opta Sports desde la cual se pueden acceder a los datos a través de consultas de acuerdo a los accesos acordados.

2.1.2. Feeds disponibles

El resultado de cada consulta tiene una estructura de árbol, similar a un archivo tipo JSON. Cada paquete de datos recibe el nombre de *Feed* y se tiene acceso a los siguientes *Feeds*:

1. **MA0 - Tournament Schedule**: entrega información básica de partidos divididos por día del calendario.

2. **MA1 - Matches**: entrega información básica de todos los partidos jugados en el campeonato como resultado, público asistente o goleadores, .

3. **MA3 - Match Events**: entrega todos los eventos e interacciones de un jugador en un partido con el balón con su respectiva coordenada (x, y), el tipo de evento, minuto de juego asociado entre otros. Entrega también la formación de cada equipo en cada partido. La coordenada (x, y) representa la posición en el largo y ancho del terreno de juego respectivamente.

4. **MA4 - Pass Matrix**: entrega el detalle de la frecuencia de los pases entregados por cada jugador al resto de sus compañeros en un partido.

5. **MA5 - Possession**: entrega el detalle de la posesión del balón de cada equipo, dividido en diferentes intervalos de tiempo. También entrega información del sector del campo donde estuvo el balón, dividido en diferentes intervalos de tiempo.

6. **MA12 - Match Expected Goals**: entrega información de remates al arco y el Expected Goal[1] asociado, por partido de manera acumulada por equipo y jugador.

7. **MA13 - Possession Events**: entrega el detalle de la posesión de los equipos en distintos intervalos de tiempo y dividido en distintos sectores del terreno de juego.

8. **MA19 - Penalties**: muestra el detalle de a qué lugar fueron lanzados los penales de los jugadores, con sus respectivas coordenadas (y, z), donde y representa el ancho del terreno y z representa la altura del arco.

9. **PE4 - Soccer Rankings**: entrega el detalle estadístico de los jugadores para cada uno de los partidos jugados. Se obtienen alrededor de 130 variables para cada jugador.

10. **TM3 - Team Squads**: entrega el detalle de los jugadores que conforman la plantilla de cada uno de los equipos participantes en el campeonato.

2.1.3. Bases de datos creadas

El objetivo es, que a partir de los distintos *Feeds*, se puedan construir tablas con los datos que se utilizarán más adelante para los distintos análisis. Las tablas creadas a partir de los *Feeds* se detallan a continuación:

1. **Equipos en competencia**: Id y nombre de todos los equipos.

2. **Jugadores en competencia**: Id, nombre, apellido, posición y equipo al que pertenece cada jugador.

3. **Entrenadores en competencia**: Id, nombre, apellido y equipo que dirige cada entrenador.

4. **Partidos de la competencia**: Id partido, id y nombre de equipo local, id y nombre de equipo visitante, fecha del partido y una variable binaria que indica si el partido ya se jugó.

5. **Eventos por partido**: Id del evento, Id y nombre del tipo de evento, periodo, minuto y segundo del partido en que ocurrió el evento, id del equipo, id y nombre del jugador relacionado al evento, coordenada (x, y) del campo de juego donde ocurrió el evento y *qualifiers*[2] que agregan detalles al evento asociado.

[1] El Expected Goal (xG) mide la probabilidad de que un remate desde cierta posición, sea gol [9]

[2] Los *qualifiers* son una lista de atributos que entregan mayor detalle de cada uno de los eventos, por ejemplo, si un evento corresponde a un *Remate al arco*, los *qualifiers* agregan información como coordenada (y, z) donde **y** es la posición de ancho del terreno de juego y **z** la posición de la altura a la que se dirigió el remate.

6. **Resultados del torneo**: Equipo local, equipo visita, goles convertidos por el equipo local y goles convertidos por el equipo visitante.

2.2. Análisis exploratorio de datos

Las tablas de *Equipos en competencia*, *Jugadores en competencia* y *Entrenadores en competencia*, se utilizan para relacionar el *Id* con los nombres de jugadores, equipos y entrenadores. La tabla de *Partidos de la competencia*, se utiliza para identificar los distintos partidos a través de su *Id*. Es por esto que se realiza el análisis descriptivo de la tabla de eventos por partido y de resultados del torneo.

2.2.1. Eventos por partido

Se agregaron las tablas de eventos de todos los partidos, para realizar una visualización de los eventos de todo el torneo. En la figura 2.1, se puede observar el porcentaje de realización de cada uno de los eventos con un porcentaje mayor a 1 %. Se destaca la gran cantidad de pases que se realizan (53.6 %) y que no aparece, entre los eventos filtrados, los goles ni ningún tipo de remate al arco. En rojo está el evento con mayor frecuencia (Pases), en verde se destacan eventos que tienen relación a cambios de posesión del balón por disputas entre jugadores por el balón. Los eventos en azul son las faltas realizadas (3.1 %) y los tiros de esquinas concedidos (1.1 %).

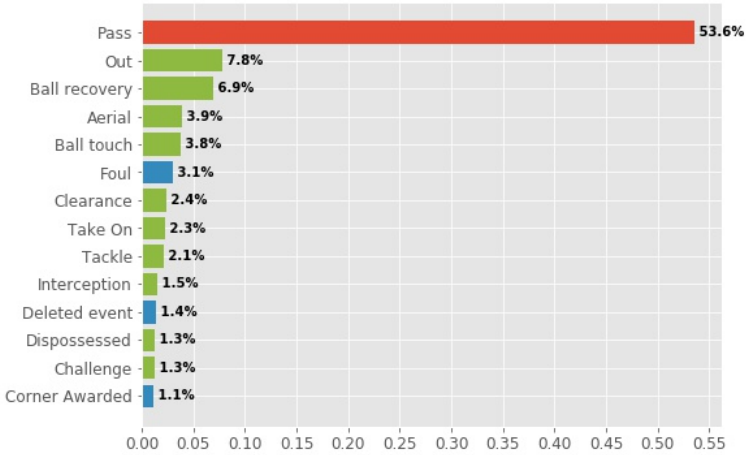


Figura 2.1: Distribución eventos con porcentaje $> 1\,\%$

Para simplificar el análisis se escogerán los eventos más relevantes tanto en frecuencia de realización como en importancia en el juego. Los eventos escogidos son **Pases**, por la alta frecuencia con la que se realizan, **Regates**, pues permite concentrar los eventos de disputas de balón (eventos en verde en figura 2.1) y **Remates** pues permite representar el evento más importante en el fútbol, que son los goles. En la figura 2.2 se observa la nueva distribución de realización de estos eventos.

Evento	%	Color
Pases	93.2	Rojo
Regates	4.0	Verde
Remates	2.8	Azul

Tabla 2.1: Resumen porcentaje de ocurrencia de eventos simplificado

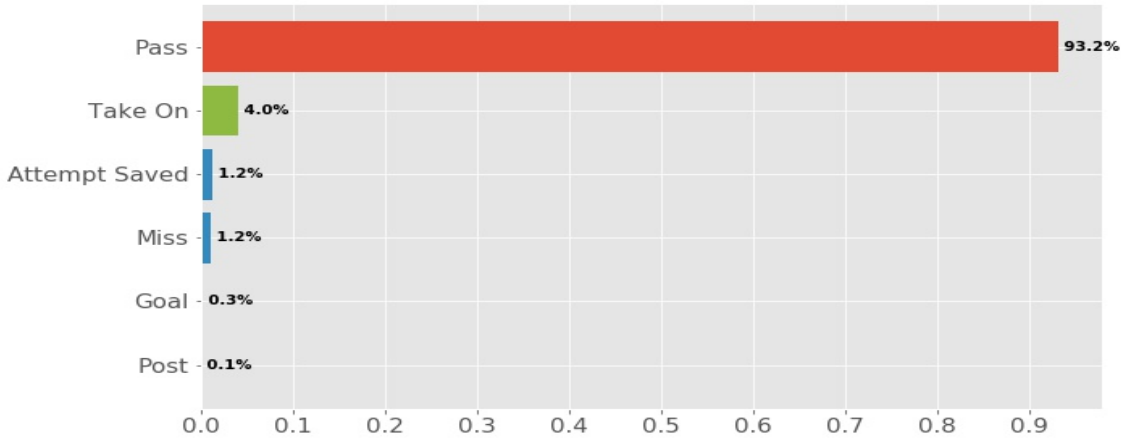


Figura 2.2: Distribución eventos simplificado

Uno de los eventos más relevantes en el fútbol, es sin duda, la conversión de goles, ya que un partido es ganado por el equipo que convierte más goles en el arco del equipo rival. En la figura 2.3, se observa la distribución de goles realizada por equipo. Los equipos están ordenados de izquierda a derecha, según la posición en que terminaron en el torneo donde Universidad Católica fue 1° y Universidad Concepción finalizó 16°.

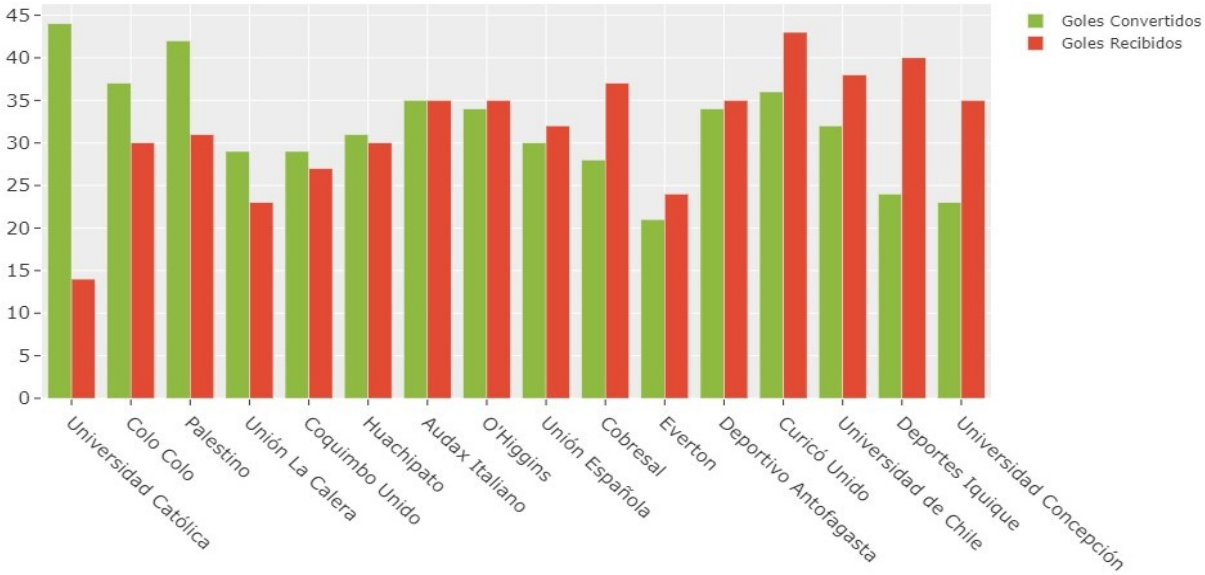


Figura 2.3: Goles convertidos y recibidos por equipo

Definido el subconjunto de eventos a estudiar, es interesante observar en qué zonas del terreno de juego suceden estos eventos. En las figuras 2.4, 2.5, 2.6, 2.7, 2.8 se observan mapas de calor de realización de pases, regates, recuperaciones de balón, remates al arco y de los goles. Es claro que cada evento tiene distribuciones diferentes en las distintas zonas de terreno de juego. El sentido de juego es de izquierda a derecha, donde el arco rival se encuentra en el extremo derecho.

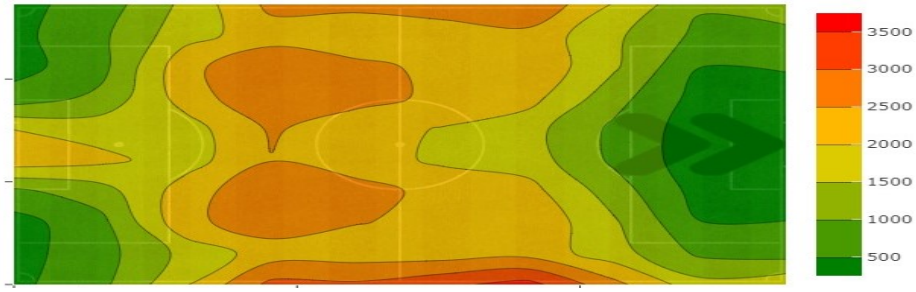


Figura 2.4: Mapa de calor de pases

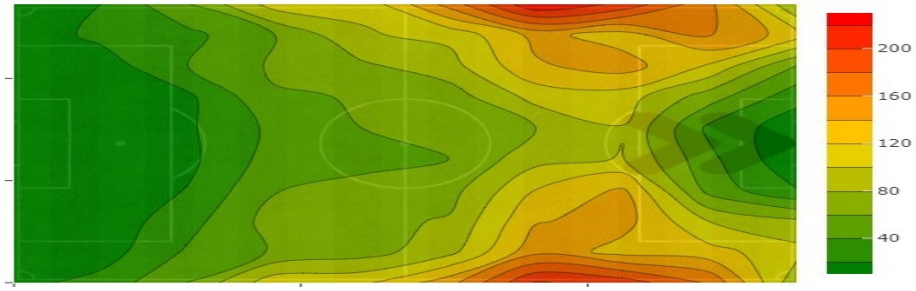


Figura 2.5: Mapa de calor de regates intentados

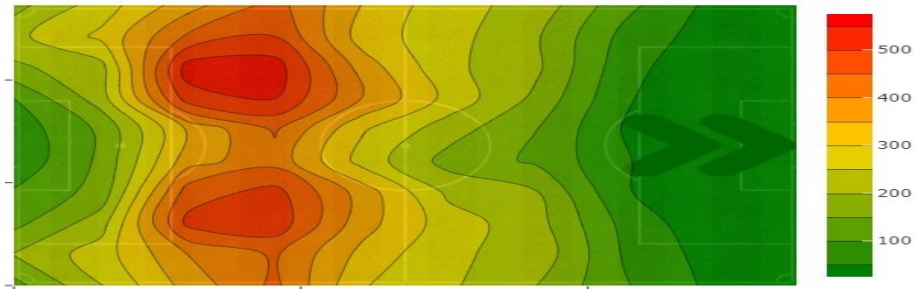


Figura 2.6: Mapa de calor de recuperaciones de balón

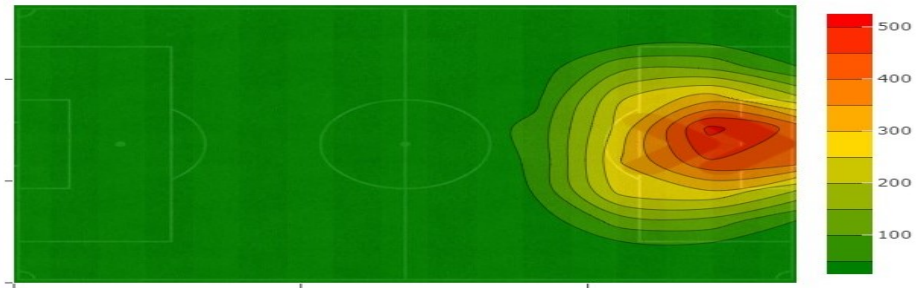


Figura 2.7: Mapa de calor de remates

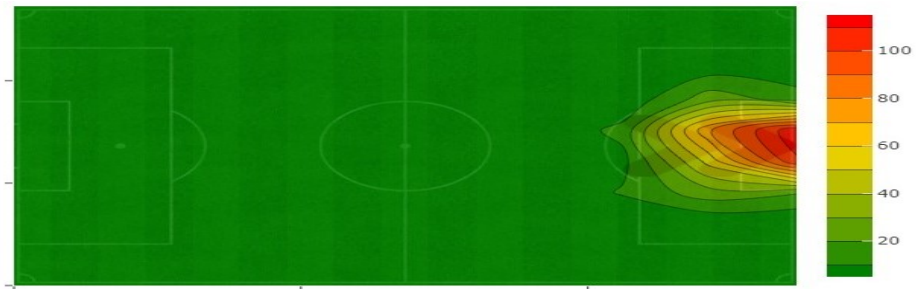


Figura 2.8: Mapa de calor de goles

Para realizar el análisis de distribución de eventos, se divide el terreno de juego en 3 tercios, tal como muestra la figura 2.9, con el objetivo de caracterizar el comportamiento o estilo de juego de los futbolistas en las distintas zonas del terreno de juego.

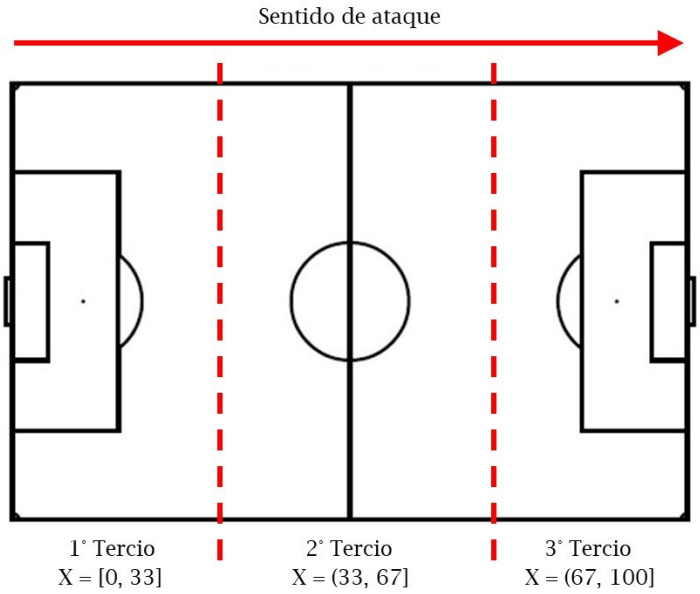


Figura 2.9: División terreno de juego

Dado que se considerarán solo 3 eventos y existe diferencia en las distribuciones de éstos, se revisará la distribución de realización de eventos para un jugador en 6 posiciones distintas. En la tabla 2.2, se detallan los jugadores escogidos. El criterio de elección fue los 4 con más acciones realizadas, el arquero menos batido y al goleador del torneo respectivamente. En las tablas 2.3, se observa la frecuencia relativa de la realización de cada acción, en cada una de las zonas del terreno de juego.

En primera instancia, se ve la similitud entre Ramón Fernández (volante ofensivo) y Luciano Aued (volante central) en cada una de las zonas del terreno de juego, esto debido a que utilizan posiciones similares en el terreno de juego. Sin embargo, cuando se comparan a los defensas Yonathan Opazo (lateral derecho) y Sebastián Silva (defensa central) se observa una gran diferencia en el tercer tercio. Esto se debe a que Y. Opazo tiene alta participación en la elaboración del juego en esa zona, mientras que S. Silva tiene participación en jugadas de finalización, tales como balones detenidos (tiros libres y tiros de esquina). Finalmente, se observa la diferencia entre Matías Dituro (arquero) y Lucas Passerini (delantero centro) ya que M. Dituro no participa en el último tercio (debido a su posición), mientras que L. Passerini es el encargado de finalizar las jugadas ofensivas de su equipo, esto último se evidencia en el alto porcentaje de remates en el tercer tercio.

Jugador	Posición	# Acciones
Ramón Fernández	Volante ofensivo	1630
Luciano Aued	Volante central	1537
Yonathan Opazo	Lateral derecho	1366
Sebastián Silva	Defensa central	1365
Matías Dituro	Arquero	745
Lucas Passerini	Delantero centro	505

Tabla 2.2: Nombre de jugadores de distintas posiciones y número de acciones realizadas en el torneo

Ramón Fernández			
Evento	**1° T**	**2° T**	**3° T**
Pases	97 %	96 %	89 %
Remates	0 %	0 %	7 %
Regates	3 %	4 %	4 %

Luciano Aued			
Evento	**1° T**	**2° T**	**3° T**
Pases	99 %	97 %	89 %
Remates	0 %	0 %	6 %
Regates	1 %	3 %	5 %

Yonathan Opazo			
Evento	**1° T**	**2° T**	**3° T**
Pases	99 %	99 %	97 %
Remates	0 %	0 %	1 %
Regates	1 %	1 %	2 %

Sebastián Silva			
Evento	**1° T**	**2° T**	**3° T**
Pases	100 %	99 %	60 %
Remates	0 %	0 %	38 %
Regates	0 %	1 %	2 %

Matías Dituro			
Evento	**1° T**	**2° T**	**3° T**
Pases	100 %	100 %	- [3]
Remates	0 %	0 %	- [3]
Regates	0 %	0 %	- [3]

Lucas Passerini			
Evento	**1° T**	**2° T**	**3° T**
Pases	86 %	90 %	71 %
Remates	0 %	1 %	19 %
Regates	14 %	9 %	10 %

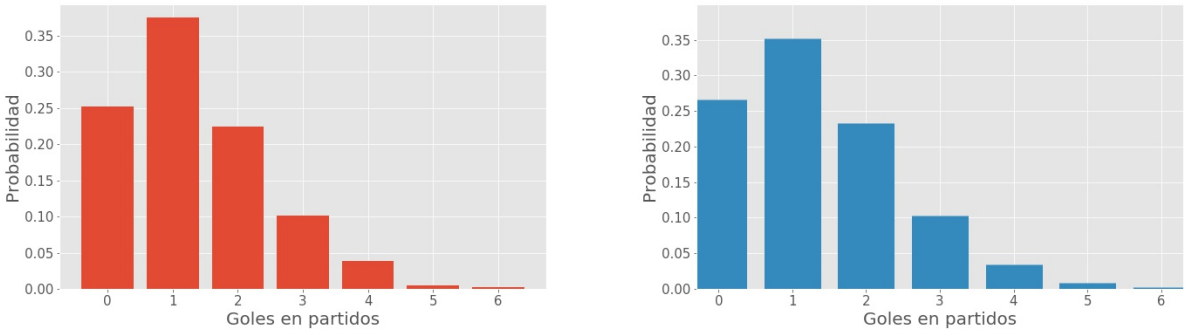
Tabla 2.3: Distribución de acciones por zona del terreno de juego

2.2.2. Resultados del torneo

Resulta interesante observar la cantidad de goles convertidos durante el torneo y cómo estos distribuyen. En la figura 2.10a, se observa la distribución real de los goles convertidos durante el torneo y en la figura 2.10b, la distribución teórica de una distribución Poisson de parámetro $\lambda = 1.326$, donde el valor de λ corresponde al promedio de goles por partido. Se realiza el test Kolmogorov-Smirnov (*KS Test*) [10] para comprobar si la distribución de los goles realizados corresponde a una distribución de Poisson de parámetro λ. La hipótesis nula, es que ambas distribuciones son idénticas en distribución, con cierto nivel α de significancia. En este caso, la hipótesis nula, es que los goles convertidos en los partidos del torneo chileno distribuyen Poisson de parámetro $\lambda = 1.326$ y se considera una significancia de $\alpha = 1\,\%$. El test entregó dos resultados, el estadística D = 0.368 y el p-valor < 0.001. Como el p-valor

[3] No registra acciones en el 3° Tercio del terreno de juego

es menor a $\alpha = 0.01$, no se puede rechazar la hipótesis nula y se concluye que con un nivel de significancia del 1 %, los goles convertidos en los partidos del torneo chileno distribuyen Poisson.



(a) Distribución real

(b) Distribución teórica

Figura 2.10: Distribuciones de goles real y teórica

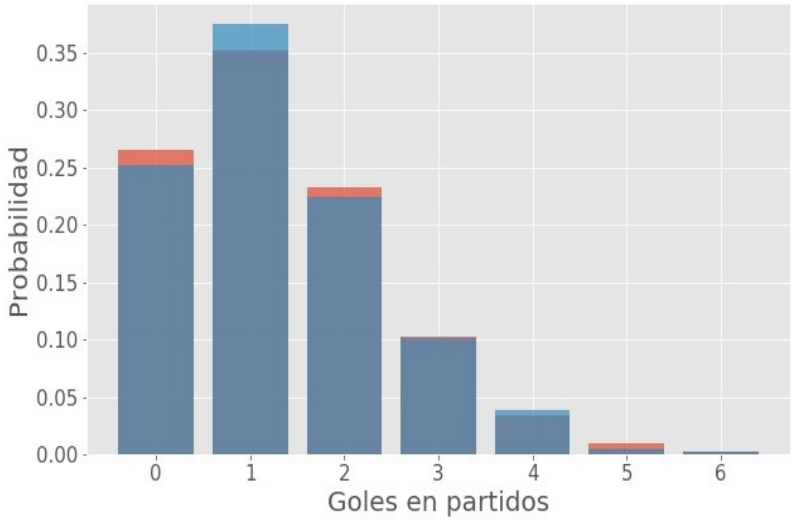


Figura 2.11: Comparación distribución de goles

Por otra parte, se desea observar la frecuencia de resultados en el torneo, es decir, cuántas veces se repitió el resultado compuesto por goles del local versus goles de la visita. En la figura 2.12 se observa que, a mayor tamaño del círculo, mayor es la frecuencia de ese resultado. El resultado que más se repite es el empate 1-1 con un 15.1 % seguido del 1-0 a favor del local con un 12.5 %.

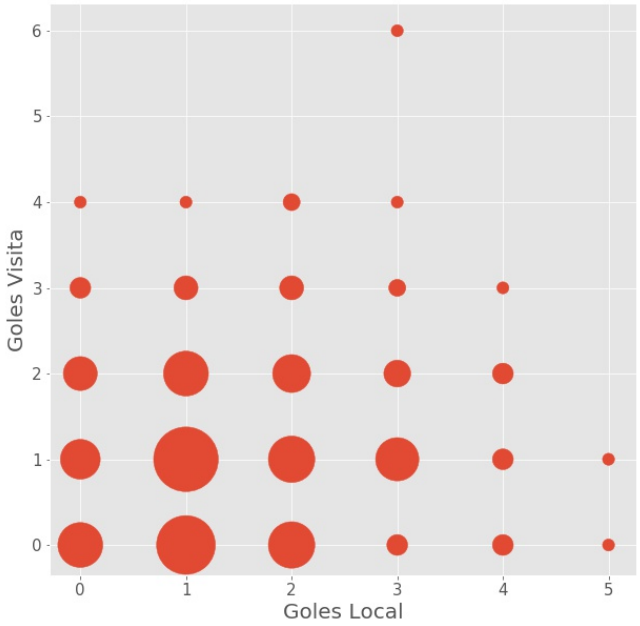


Figura 2.12: Distribución resultados

En resumen, se puede mencionar que se considerarán 3 tipos de eventos que son **pases**, pues representan en gran porcentaje los eventos ocurridos durante el torneo, **regates**, ya que captura el efecto de cuando la posesión del balón cambia de equipo y **remates**, dado que permite representar el evento más importante para el resultado final del partido, que son los goles. Además, se realizó un análisis de la distribución de ocurrencia de los eventos y se observa que existen heterogeneidad en sus distribuciones en el terreno de juego, es decir, los eventos seleccionados ocurren, evidentemente, en distintas zonas del campo. Observando los goles convertidos, se ve que los equipos que tienden a convertir más goles, obtienen mejores posiciones al final del campeonato y que la distribución de conversión de resultados se puede aproximar a través de un proceso de Poisson. Finalmente, se ve que los equipos locales tienden a convertir más goles que los equipos visitantes.

Capítulo 3

Modelo de Recomendación

3.1. Modelo basado en estadística descriptiva

El objetivo del modelo, es entregar una lista de jugadores que cumplan con cierto criterios definidos por el club, tales como superar un umbral de minutos jugados y tengan un rendimiento sobresaliente. Se obtiene una lista distinta para cada una de las posiciones y las medidas de rendimiento son calculadas en base al rendimiento de cada jugador en cada posición distinta en la que jugó, de esta forma, podemos comparar a un jugador consigo mismo en distintas posiciones, en base a los minutos jugados en cada una de ellas.

El primer paso es definir las posiciones del campo de juego que se considerarán. Además, para cada posición se debe definir un conjunto de variables que sean representativas del rendimiento de un jugador y, finalmente, definir la importancia de cada una de estas variables en base a ponderaciones. La definición de estos ítems, fue en conjunto con la secretaría técnica del club de fútbol Universidad de Chile compuesta por Rodrigo Carrasco, Manuel Mayo y Fabián Pacheco y, los directores deportivos Rodrigo Goldberg y Sergio Vargas. En la tabla 3.1 se especifican las posiciones, variables y sus respectivas ponderaciones.

Arquero	**Defensa Central**	**Laterales**	**Volante Central**	**Volante Interior**	**Volante Ofensivo**	**Extremo**	**Delantero Centro**
Expected Goal en Contra (xGC) (40 %)	% Duelos defensivos ganados (40 %)	% Centros correctos (40 %)	% Pases hacia adelante (35 %)	% Pases hacia adelante (30 %)	Asistencias cada 90 mins (30 %)	% Duelos ofensivos ganados (30 %)	Diferencia goles vs xG (25 %)
% Duelos aéreos ganados (30 %)	% Duelos aéreos defensivos ganados (30 %)	% Duelos defensivos ganados (30 %)	% Duelos defensivos ganados (25 %)	Intercepciones cada 90 mins (20 %)	% Remates fuera del área al arco (30 %)	% Centros correctos (30 %)	% Remates al arco (25 %)
% Pases cortos correctos (20 %)	% Pases cortos correctos (10 %)	% Duelos ofensivos ganados (20 %)	% Balones largos correctos (20 %)	Asistencias cada 90 mins (20 %)	% Remates dentro del área al arco (20 %)	% Remates al arco (15 %)	% Duelos aéreos ganados (25 %)
% Pases largos correctos (10 %)	% Pases largos correctos (10 %)	% Duelos aéreos defensivos ganados (10 %)	Intercepciones cada 90 min (10 %)	% Duelos defensivos ganados (15 %)	% Duelos ofensivos ganados (20 %)	Asistencias cada 90 mins (15 %)	Asistencias cada 90 mins (15 %)
	Faltas cada 90 min (10 %)	-	% Remates al arco (10 %)	% Remates al arco (15 %)	-	% Duelos defensivos ganados (5 %)	% Duelos defensivos ganados (5 %)
	-	-	-	-	-	Intercepciones cada 90 mins (5 %)	Intercepciones cada 90 mins (5 %)

Tabla 3.1: Variables por posición con respectivas ponderaciones

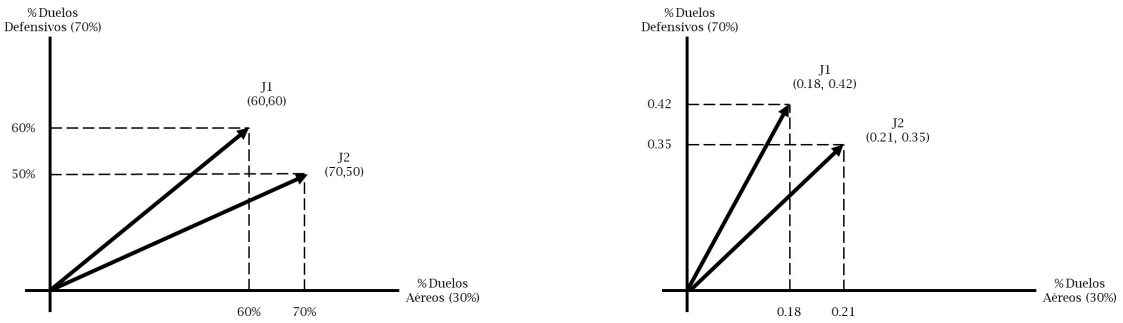
Ya definida las variables, se calcula el rendimiento para cada una de las posiciones para los jugadores que hayan jugado más de una cantidad fija de minutos. El umbral fijado fue de 400 minutos disputados en el torneo nacional. Existen jugadores que se desempeñan en más de una posición a lo largo del torneo y es por esto que puede aparecer el mismo jugador en más de una posición y con rendimientos distintos, tal como se muestra en la tabla 3.2.

Jugador	**Mins Jugados**	**Posición**	**% Duelos defensivos ganados**	**% Duelos aéreos defensivos ganados**	**% Pases cortos correctos**	**% Pases largos correctos**	**Faltas cada 90 mins**
Benjamín Kuscevic	564	Defensa Central Derecho	47 %	51 %	87 %	45 %	1.11
Benjamín Kuscevic	450	Defensa Central Izquierdo	62 %	50 %	90 %	69 %	1.8

Tabla 3.2: Rendimiento de Benjamín Kuscevic en 2 posiciones distintas

Una vez calculados los indicadores de rendimiento para todos los jugadores y cada una de las posiciones, se obtiene una lista desordenada de los rendimientos de los jugadores y lo que continúa es generar un ranking de cuáles son los mejores jugadores en cada una de las posiciones. Para ordenar a los jugadores, se genera un *Score* para cada posición. El *Score* es igual al módulo del vector generado por las variables de rendimiento ponderadas respectivamente. Cada posición tiene su propio puntaje asociado, dado que se consideran distintas variables para cada una de ellas y este puntaje no es comparable entre posiciones. El cálculo del *Score* se ejemplifica con 2 variables en los gráficos de la figura 3.1

Supongamos a dos jugadores distintos llamados J1 y J2, donde cada uno tiene cierto rendimiento en las variables *% Duelos aéreos ganados* (eje X) y *% Duelos defensivos ganados* (eje Y), las cuales tienen una ponderación de 30 % y 70 % respectivamente.



(a) Variables normales

(b) Variables ponderadas

Figura 3.1: Transformación de 2 variables a través de ponderaciones

Las ponderaciones, permiten que el *Score* represente realmente a los jugadores que tienen un buen rendimiento deportivo. En caso contrario, sucedería lo siguiente:

$$Score_1 = \sqrt{0.6^2 + 0.6^2} = \sqrt{0.36 + 0.36} = \sqrt{0.72} = 0.85 \tag{3.1}$$

$$Score_2 = \sqrt{0.7^2 + 0.5^2} = \sqrt{0.25 + 0.49} = \sqrt{0.74} = 0.86 \tag{3.2}$$

Con esto, concluiríamos que el Jugador 2 está por sobre el Jugador 1. Sin embargo, si consideramos las ponderaciones:

$$x_1 = 0.6 \cdot 0.3 = 0.18 \quad y_1 = 0.6 \cdot 0.7 = 0.42 \tag{3.3}$$

$$x_2 = 0.7 \cdot 0.3 = 0.21 \quad y_2 = 0.5 \cdot 0.7 = 0.35 \tag{3.4}$$

$$Score_1 = \sqrt{0.18^2 + 0.42^2} = \sqrt{0.0324 + 0.1764} = \sqrt{0.2088} = 0.47 \tag{3.5}$$

$$Score_2 = \sqrt{0.21^2 + 0.35^2} = \sqrt{0.0441 + 0.1225} = \sqrt{0.1666} = 0.41 \tag{3.6}$$

Como *Score 1* es mayor que *Score 2*, concluimos correctamente que el Jugador 1 tiene mejor rendimiento que el Jugador 2.

Es preciso mencionar, que se debe hacer 2 transformaciones a las variables antes de ponderarlas. La primera transformación es necesaria porque algunas variables son de frecuencia de ocurrencia de eventos cada 90 minutos de juego. Estas variables pueden tener valores mayores a 1 y es necesario escalarlas entre [0, 1] para que sean comparables con las variables de porcentaje que ya están entre el rango [0, 1]. La segunda transformación es necesaria porque, la gran mayoría de las variables, entre más grande es el valor, se considera un mejor rendimiento, sin embargo, existen algunas variables (como *Faltas cada 90 min*) que un mayor valor representa un menor rendimiento deportivo. Para solucionar este problema y que en todas las variables se considere que un mayor valor es mejor, se realiza la transformación $1 - variable$. Se puede realizar esto dado que todas las variables están escalados para que pertenezcan al rango [0, 1]. Luego de realizar esto para todos las posiciones, se obtiene la lista de jugadores ordenada de mejor a peor rendimiento en base a su *Score* para cada una de las posiciones.

3.2. Modelo basado en simulación: Cadenas de Markov

3.2.1. Descripción del modelo

El objetivo del modelo, basado en simulación, es encontrar jugadores de otros clubes que ayuden al equipo a tener un mejor rendimiento. En primera instancia, se identifica la posición que se desea fortalecer para luego reemplazar al actual jugador del club por otros potenciales jugadores que se desempeñan en esa misma posición. Finalmente, se define una métrica de desempeño del equipo para cuantificar y comparar si la inclusión de un jugador mejora el rendimiento colectivo del equipo.

Para llevar a cabo esta simulación, se considera un partido como un proceso estocástico, donde los jugadores van decidiendo qué acción realizar con el balón durante el desarrollo del encuentro. De esta forma, se define el *Estilo de Juego* de un futbolista como la *distribución de las acciones que realiza durante una temporada*. Para este trabajo, se consideran 3 acciones posibles bajo las cuales un jugador puede decidir: **dar un pase, regatear a un rival y rematar al arco**. No se consideran eventos cuando el balón sale del terreno de juego (tiros de esquina o saques de banda) o existen tiempos muertos (faltas, tiros libres, posiciones de adelanto y cambios de jugadores) debido a que se desea modelar en tiempo discreto.

De acuerdo a las acciones durante la temporada, cada jugador tiene una distribución distinta de las 3 acciones antes mencionadas cuando está en condición de local o de visita y, sumado a lo anterior, esta distribución cambia de acuerdo a la zona del campo de juego donde está ubicado. Para este trabajo se divide el terreno de juego en 3 zonas (1° Tercio, 2° Tercio y 3° Tercio) como muestra la figura 2.9

Se desea modelar un partido como un proceso estocástico. Se utilizará un modelo de **Cadenas de Markov a tiempo discreto** para representar la información necesaria con el objetivo de simular un partido. Una Cadena de Markov a tiempo discreto, se define a través de *Estados* y una *Matriz de Transición* y, puede ser representada a través de grafos. La información que la cadena representa en cada uno de sus estados es: **el minuto del partido (M), el jugador que posee el balón (j) y la zona del campo en la que se encuentra (z)**.

$$Estado_i \ = \ (M, \ j, \ z)$$

Los conjuntos sobre los que se trabaja se detallan a continuación. Es preciso mencionar que, cada probabilidad detallada más adelante está indexada por el partido t que se está jugando, pero para simplificar la notación, se decide omitir el subíndice.

- $j \in J$ conjunto de jugadores
- $z \in Z$ conjunto de zonas del terreno de juego
- $j' \in J/\{j\}$ conjunto de compañeros del jugador j
- $r \in R_j$ conjunto de rivales del jugador j

Se comienzan desde un estado inicial, donde el delantero del equipo que comienza el saque de inicio de partido entrega un pase correcto a un compañero. Así, el nuevo jugador debe decidir qué acción realizar dentro de las opciones posibles: **dar un pase, regatear a un rival o rematar al arco**. Cada jugador tiene una distribución distinta de la probabilidad de realizar una acción en cada zona del terreno de juego y dependiendo también de la condición en que se realiza (local o visita). La suma de la distribución de acciones en cada zona suma 1. La notación de la distribución es de la siguiente forma:

$$Estilo\ de\ juego_{j,z} = (p_{j,z}, d_{j,z}, r_{j,z}) \qquad \forall j, z \tag{3.7}$$

con:

$$p_{j,z} + d_{j,z} + r_{j,z} = 1 \qquad \forall j, z \tag{3.8}$$

donde:

- $p_{j,z}$ es la probabilidad de que el jugador j intente un pase en la zona z
- $d_{j,z}$ es la probabilidad de que el jugador j intente un regate en la zona z
- $r_{j,z}$ es la probabilidad de que el jugador j realice un remate en la zona z

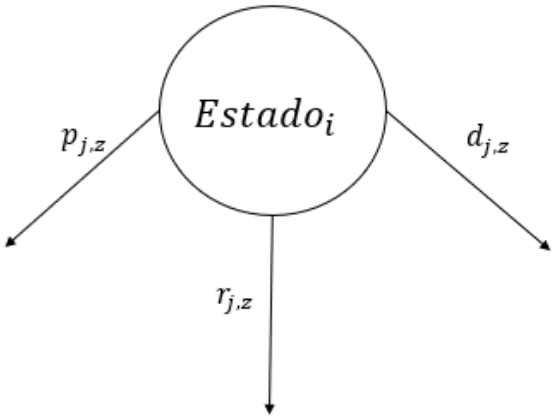


Figura 3.2: Posibilidades de acciones desde el estado i - Estilo de juego

Definida las posibles acciones que un futbolista puede realizar, se definen dos distribuciones: distribución de compañeros y distribución de rivales.

1. **Distribución de compañeros:** la distribución compañeros cuantifica la probabilidad de que un $compañero_{j'}$ sea el posible receptor de un pase en una zona z'. Se denota la probabilidad como $m_{j',z'}$ considerando $\sum_{j' \neq j} m_{j',z'} = 1 \quad \forall z'$. Esta probabilidad se calcula como la cantidad de pases recibidos por el jugador j' en la zona z' dividido por la cantidad de pases totales recibidos por su equipo en la zona z'

2. **Distribución de rivales:** la distribución de rivales cuantifica la probabilidad de que el balón quede en posesión de un rival r luego de realizar pase o regate en forma incorrecta.

   - **Intercepción de pases**: se denota como $n_{r,z'}$ la probabilidad de que un rival r sea quien intercepta un pase en la zona z' considerando $\sum_r n_{r,z'} = 1 \quad \forall z'$. Esta probabilidad se calcula como la cantidad de pases interceptados por el rival r en la zona z' dividido por la cantidad total de pases interceptados por su equipo en la zona z'
   - **Recuperación de balón**: se denota como $q_{r,z}$ la probabilidad de que un rival r sea quien intenta recuperar el balón ante un regate de un jugador en la zona z considerando $\sum_r q_{r,z} = 1 \quad \forall z$. Esta probabilidad se calcula como la cantidad de recuperaciones intentadas por el jugador r en la zona z dividido por la cantidad de recuperaciones intentadas por su equipo en la zona z.

Una vez definidas las distribuciones, se define qué es una acción realizada con éxito, cuáles son las probabilidades asociadas y cómo se calculan.

1. **Dar un pase**: un jugador j ubicado en la zona z intenta dar un pase a un compañero j' a la zona z' de acuerdo a la distribución $m_{j',z'}$. Sin embargo, el equipo rival tiene cierta capacidad de interceptar pases en la zona z'. Entonces, la probabilidad de que el pase llegue al jugador j' en la zona z' depende de la capacidad del jugador j de no errar su pase y la capacidad del equipo rival de interceptar pases.

   - **Capacidad de no errar un pase**: cantidad de pases entregados correctamente desde la zona z a la zona z' dividido por la cantidad total de pases intentados desde la zona z a la zona z'. Se denota $u_{j,z,z'} \quad \forall j, z, z'$

- **Capacidad del equipo rival de interceptar un pase**: cantidad de pases interceptados en la zona z' dividido por la cantidad total de pases que les realizaron en la zona z'. Se denota $v_{t',z'}$ donde t' representa al equipo rival. Esta probabilidad actúa como un factor de descuento para incorporar la habilidad del equipo rival para interceptar.

Finalmente, la probabilidad de que el pase del jugador j desde la zona z al jugador j' en la zona z' sea correcto se define como:

$$c_{j,z,j',z'} = u_{j,z,z'} \cdot (1 - v_{t',z'}) \quad \forall j, z, j', z' \tag{3.9}$$

$$\sum_{j'} c_{j,z,j',z'} = Pase\ es\ correcto \quad \forall j, z, z'$$

$$1 - \sum_{j'} c_{j,z,j',z'} = Pase\ es\ interceptado \quad \forall j, z, z'$$

Si el pase es interceptado, se utiliza la distribución $n_{r,z'}$ para identificar qué jugador rival r queda en posesión del balón.

2. **Regatear a un rival**: un jugador j ubicado en la zona z intenta regatear a un rival r que intenta quitarle el balón de acuerdo a la distribución $q_{r,z}$. El rival r tiene cierta capacidad de recuperar el balón ante un regate del jugador j. Entonces, la probabilidad de que el regate del jugador j sea correcto es:

   - **Capacidad de encarar bien**: cantidad de veces que regateó correctamente a un rival en la zona z dividido por la cantidad de regates intentados en la zona z. Se denota $s_{j,z} \quad \forall j, z$
   - **Capacidad del rival de recuperar el balón**: el jugador r que intenta quitar el balón al jugador j tiene cierta habilidad de recuperar balones. Esta probabilidad se calcula como la cantidad de veces que recuperó el balón cuando intentó una recuperación en la zona z dividido por la cantidad total de recuperaciones intentadas en la zona z. Se denota $t_{r,z} \quad \forall r, z$

   Finalmente, la probabilidad de que el jugador j mantenga la posesión del balón luego del regate ante un rival r en la zona z se define como:

$$e_{j,z,r} = \frac{\alpha \cdot s_{j,z}}{\alpha \cdot s_{j,z} + \beta \cdot t_{r,z}} \quad \forall j, z, r \tag{3.10}$$

   donde α representa la importancia de la capacidad de regatear bien del jugador j y β representa la importancia de la capacidad del rival r de recuperar bien el valor considerando $\alpha + \beta = 1$.

   Si el regate es efectivo, el jugador j mantiene la posesión del balón en la zona z, en caso contrario, si no es efectivo, el rival r queda en posesión del balón en la zona z.

3. **Rematar al arco**: un jugador j ubicado en la zona z intenta un remate al arco. La probabilidad de dirigir el remate al arco depende exclusivamente de la capacidad del rematador. Por otra parte, si el remate va al arco, este puede ser gol o no. Esta probabilidad depende de la habilidad del rematador y del arquero rival.

   - **Remate al arco**: la probabilidad de que un remate vaya al arco se calcula como la razón entre la suma del xG (Expected Goal) de los remates que van efectivamente al arco (los que son atajados y los que terminan en gol) dividido por la suma del xG de todos los remates intentados en cada zona z. Se denota $f_{j,z}$ la probabilidad de que el remate vaya al arco (1 - $f_{j,z}$ el remate se va afuera).
   - **Capacidad de atajar del arquero**: la capacidad de un arquero de atajar un remate se calcula como la razón entre la suma del xG de los remates que atajó dividido por la suma del xG de todos los remates que le lanzaron. Se denota $cap_r^{gk} \quad \forall r$ donde r representa al arquero del equipo rival.

   Finalmente, la probabilidad de que un remate del jugador j desde la zona z contra el arquero r sea gol (condicionado a que efectivamente va al arco) se define como:

$$g_{j,z,r} = f_{j,z} \cdot (1 - cap_r^{gk}) \quad \forall j, z, r \tag{3.11}$$

   Si el remate es efectivo, es decir, se convirtió en gol, el equipo rival reanuda el juego en el 2° Tercio del campo de juego simulando el saque de mitad de campo. Si el remate no es efectivo, es decir, es atajado por el arquero rival o va fuera del arco, el arquero reanuda el juego desde el 1° Tercio simulando que el arquero continúa con el balón e intenta dar un pase a un compañero o un saque de arco, respectivamente.

Una representación gráfica de la cadena se muestra en la figura 3.3

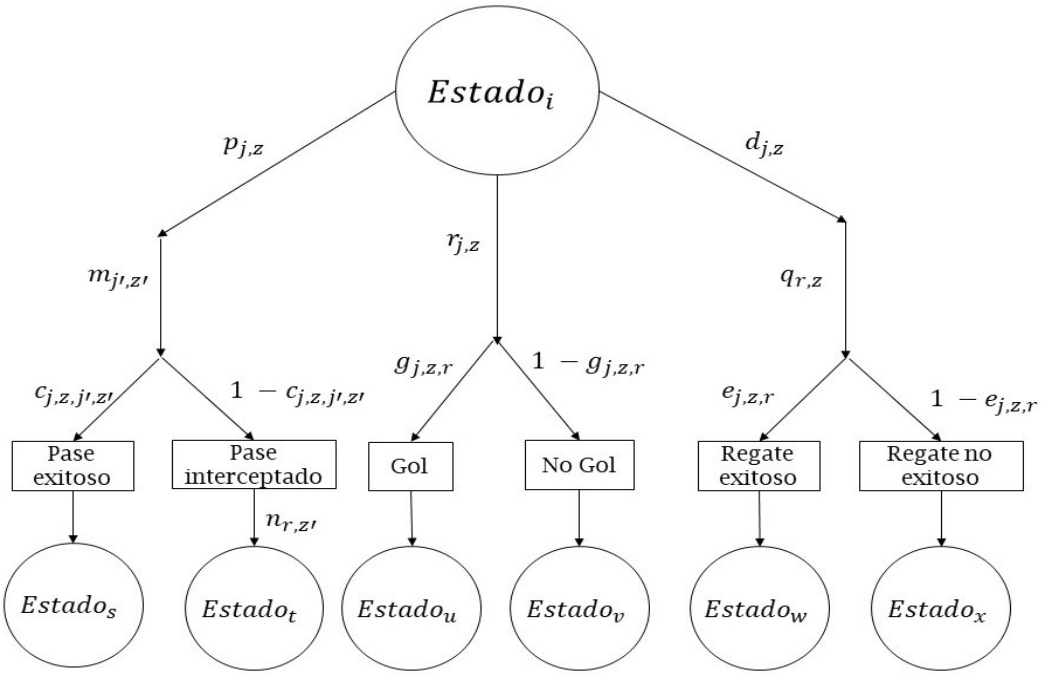


Figura 3.3: Modelo estocástico de un partido de fútbol

donde los estados representan:

- $Estado_s$ = la nueva zona z' dónde está el balón y el compañero j' que lo tiene
- $Estado_t$ = la nueva zona z' dónde está el balón y el rival r que lo tiene
- $Estado_u$ = se marcó un gol y ahora el balón lo tiene el delantero del equipo que recibió el gol para reanudar el juego desde mitad de cancha (2° Tercio)
- $Estado_v$ = el balón se fue por la línea de fondo o el arquero rival atajó el remate, en ambos casos, el arquero rival tiene el balón y reanuda el juego con un pase desde el 1° Tercio
- $Estado_w$ = el jugador j que realizó correctamente el regate tiene el balón en la misma zona z
- $Estado_x$ = el jugador j perdió el balón en la zona z por la recuperación del rival r. Ahora el balón lo posee el rival r en la zona z

3.2.2. Criterio de recomendación de jugadores

Como este modelo está representado por jugadores y sus estilos de juego, se puede utilizar para recomendar jugadores si, en la alineación inicial de cada partido, se reemplaza el jugador que se desea contratar, por el actual jugador del club Universidad de Chile que se desempeña en esa posición. Para definir si un jugador es o no un buen refuerzo para el club, se utilizará como medida del rendimiento del equipo la probabilidad de salir campeón. La probabilidad de salir campeón se calcula como la cantidad de veces que el equipo terminó el torneo en la primera posición dividido en el número total de simulaciones. Para esto se considerarán 24 fechas y los criterios de desempate del torneo chileno [11].

La comparación se realizará entre el torneo simulado con las plantillas actuales y se calculará la probabilidad de salir campeón a todos los clubes participantes. Esta será la línea base del rendimiento deportivo. Luego, se reemplazarán, uno a uno, los jugadores con mayor rendimiento deportivo, de acuerdo al modelo de recomendación de estadística descriptiva para simular nuevamente el torneo completo. Finalmente, se calcula la probabilidad de salir campeón en estas nuevas simulaciones y se compara con la línea base de acuerdo a la expresión 3.12. La lista de recomendación será conformada por los jugadores que el aumento de la probabilidad de campeonar sea mayor.

$$\triangle\mathbb{P}(Campeón\ refuerzo\ i) = \mathbb{P}(Campeón\ \ simulación\ refuerzo\ i) - \mathbb{P}(Campeón\ línea\ base) \tag{3.12}$$

3.3. Modelo de benchmarking basado en simulación: Poisson

Para tener una referencia de la precisión del modelo, es que se desea comparar el modelo de Cadenas de Markov con otros modelos que existen en la actualidad para predecir resultados de partidos de fútbol. El modelo que se utilizará para comparar es el modelo propuesto

por Dixon & Coles [12], basado en que los goles convertidos por un equipo siguen un proceso de Poisson. Para esto, utilizaremos datos de las primeras 18 fechas del torneo chileno 2019 para predecir los resultados de las últimas 6 fechas jugadas (fecha 19 - fecha 24).

Para simular con esta metodología, se deben estimar las tasas de conversión de goles de cada uno de los 16 equipos del torneo a través de una regresión de Poisson [13]. En este modelo, se busca una relación lineal entre los goles convertidos por un equipo a través de 3 parámetros: localía, equipo, rival. Con los parámetros estimados, se puede simular un partido a través de la representación de los goles como la realización de variables aleatorias que distribuyen Poisson con λ_1 y λ_2 para el equipo local y equipo visitante. Para el caso del equipo local, la tasa de conversión de goles incluye el parámetro de localía. Si se realiza este ejercicio de simulación una cantidad lo suficientemente grande, permite calcular la probabilidad de que un equipo derrote al otro o empaten. Para el caso de simular un torneo, se deben simular todos los partidos y calcular el puntaje obtenido según las reglas FIFA, en que un partido ganado otorga 3 puntos, partido empatado otorga 1 punto y uno perdido 0 puntos [14].

3.3.1. Comparación

Para comparar el rendimiento y la precisión de los modelos basados en simulación con Cadenas de Markov y Poisson se revisarán que tan bien predicen estos modelos en las fechas 19 a la fecha 24. Para esto, se define la métrica de error como :

$$Error_m = \frac{1}{N}\sum_s\sum_i\sum_p Rr_{ip}(\sqrt{\alpha(Rr_{ip} - Rs_{msip})^2 + \beta(GLr_{sp} - GLs_{msp})^2 + \beta(GVr_{sp} - GVs_{msp})^2}) \tag{3.13}$$

donde:

- $Modelos\ m\ \in\ \{CM, Poisson\}$
- $Simulación\ s\ \in\ \{1, \_\_\_\_, 10000\}$
- $Resultado\ i \in\ \{Local, Empate, Visita\}$
- $Partido\ p \in\ \{1, \_\_\_\_, 192\}$
- $Rr_{ip} \in\ \{0, 1\}$ Resultado real i del partido p
- $Rs_{msip} \in \{0, 1\}$ Resultado i de la simulación s del modelo m en el partido p
- $GLr_p \in \mathbb{R}_0^+$ Goles reales del local del partido p
- $GLs_{msp} \in \mathbb{R}_0^+$ Goles del local de la simulación s del modelo m en el partido p
- $GVr_p \in \mathbb{R}_0^+$ Goles reales de la visita del partido p
- $GVs_{msp} \in \mathbb{R}_0^+$ Goles de visita de la simulación s del modelo m en el partido p
- $\alpha \in \mathbb{R}_0^+$ ponderación de la diferencia del resultado
- $\beta \in \mathbb{R}_0^+$ ponderación de la diferencia de los goles del local y la visita

Capítulo 4

Resultados

En esta sección se encuentran tablas y gráficos para visualizar y comparar los resultados del modelo de rendimiento descriptivo en algunas posiciones. Aquí se mostrará el ranking de los mejores jugadores y figuras que permiten una comparación visual de este ítem, además de una reseña en modo análisis y conclusión de cada una de estas posiciones. Luego, se encuentra una comparación entre los modelos de simulación de Poisson y el modelo de simulación de Cadenas de Markov para ver su capacidad predictiva con algunos partidos del torneo chileno de primera división 2019, para dar paso a el cálculo de la métrica de error de ambos modelos. Finalmente, se encuentra la utilización del modelo de simulación de Cadenas de Markov como recomendador de jugadores. Se muestra qué jugadores se utilizaron para testear y cuáles son los que aportan mayor rendimiento al equipo de la Universidad de Chile en base a la métrica de rendimiento de cuánto aumenta la probabilidad de salir campeón.

4.1. Visualización del modelo de estadística descriptiva

Para mostrar los resultados de este modelo, se seleccionaron 4 posiciones del campo de juego: Defensa central derecho, mediocampista central, mediocampista ofensivo y delantero. Para cada posición se muestra el top 3 de los jugadores con sus estadísticas asociadas y un gráfico polar, donde se puede comparar al jugador perteneciente de plantel de la Universidad de Chile de mejor rendimiento con el mejor jugador de esa categoría. En el gráfico se puede observar el valor de la ponderación de la variable al lado de su nombre. Las variables que no son porcentaje, fueron escaladas para representar el percentil al que pertenece el jugador en esa variable.

1. Defensa central derecho

Nombre	Club	Mins Jug	% Duelos Defensivos	% Duelos Aéreos	% Pases Cortos	% Pases Largos	Faltas c/90 mins	Score
Pond.	-	-	**40 %**	**30 %**	**10 %**	**10 %**	**10 %**	-
Julio Barroso	Colo Colo	1080	71	51	87	44	0.2	1.066
Franco Bechtholdt	Curicó Unido	1687	66	65	86	37	0.9	1.061
Germán Lanaro	U. Católica	1584	63	58	88	63	0.6	1.056

Tabla 4.1: Estadísticas top 3 defensas centrales derechos

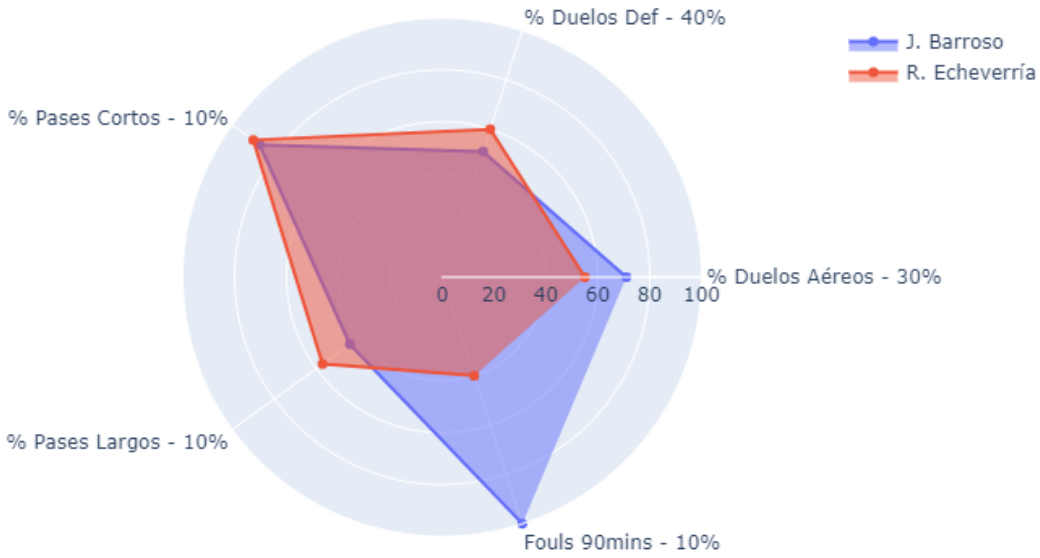


Figura 4.1: Julio Barroso (CC) vs Rodrigo Echeverría (UCH)

En la tabla 4.1 se observa el top 3 de jugadores, ordenados de mejor rendimiento a peor rendimiento según el *Score* calculado. En estos jugadores, la diferencia del rendimiento no la genera la variable *% Pases Cortos* ya que todos tienen un rendimiento muy similar. Sin embargo, Julio Barroso destaca por el *% Duelos Defensivos*, en que saca una leve ventaja al resto de los jugadores, pero que entrega mucho puntaje para el *Score* final dado que es la variable que mayor ponderación tiene (40 %). Por otra parte, en la figura 4.1, cuando comparamos a Julio Barroso (CC) con Rodrigo Echeverría (UCH), se observa un rendimiento similar en las variables de *% Pases Cortos, % Duelos Defensivos y % Pases Largos*, pero Barroso tiene un rendimiento sobresaliente en la capacidad de cometer pocas faltas cada 90 minutos de juego y una diferencia significativa en *% Duelos Aéreos*, que es la segunda variable con mayor ponderación (30 %).

2. Mediocampista central

Nombre	Club	Mins Jug	% Duelos Defensivos	Inter 90mins	% Pases Adelante	% Pases Largos	% Remates al arco	Score
Pond.	-	-	**25 %**	**10 %**	**35 %**	**20 %**	**10 %**	-
Eduardo Farías	Cobresal	1305	47	1.9	81	54	3	0.411
Esteban Pavez	Colo Colo	630	44	2.9	75	84	40	0.402
Nery Leyes	Antofagasta	450	45	1.4	74	80	30	0.376

Tabla 4.2: Estadísticas top 3 mediocampistas centrales

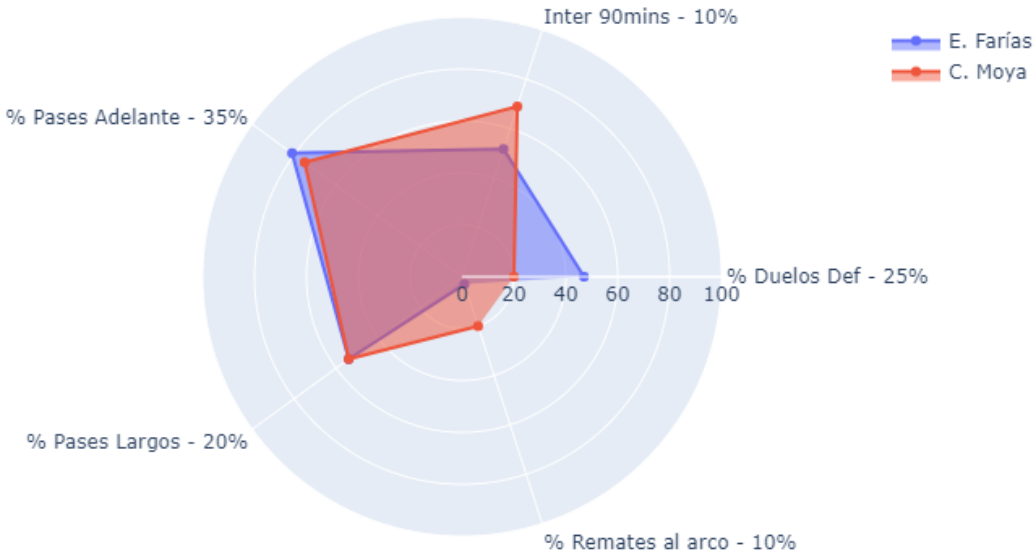


Figura 4.2: Eduardo Farías (COB) vs Camilo Moya (UCH)

En la tabla 4.2, se observa que Eduardo Farías (COB), es el jugador con mejor rendimiento en esta posición y con una gran cantidad de minutos jugados con respecto a los jugadores que completan el podio. Se observa que Eduardo Farías tiene bajó *% Remates al arco* (3 %), pero esa variable tiene una baja ponderación (10 %) con respecto a las variables donde Farías tiene un rendimiento sobresaliente, como lo es *% Pases Adelante* que pondera un 35 % y una pequeña ventaja en *% Duelos Defensivos* con una ponderación del 25 %. En la figura 4.2, se observar el rendimiento comparado con Camilo Moya, el jugador de la Universidad de Chile con mejor rendimiento en esta posición, y la principal diferencia se genera en que Camilo Moya tiende a interceptar mayor cantidad de balones cada 90 minutos, pero esa variable tiene baja ponderación (10 %) y, en cambio, en la variable de *% Duelos Defensivos*, Eduardo Farías obtiene un rendimiento de más del doble de conversión en ese ítem y esa variable es la segunda en ponderación con un 25 %.

3. Mediocampista ofensivo

Nombre	Club	Mins Jug	Asist. c/90mins	% Remates fuera área	% Remates dentro área	% Duelos Ofensivos	Score
Pond.	-	-	**30 %**	**30 %**	**20 %**	**20 %**	-
J. Valdivia	Colo Colo	575	0.5	47	63	88	0.402
M. Salas	O'Higgins	586	0.3	62	67	62	0.400
H. Droguett	U. de Concepción	743	0.8	34	40	6	0.362

Tabla 4.3: Estadísticas top 3 mediocampistas ofensivos

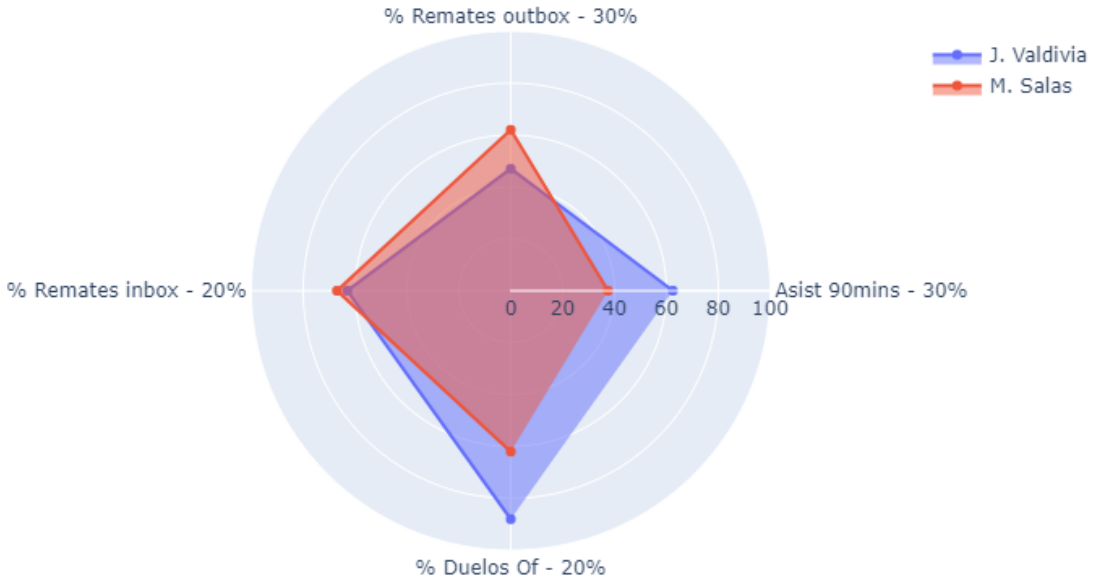


Figura 4.3: Jorge Valdivia (CC) vs Maximiliano Salas (OHI)

En la tabla 4.3 se observa que el mejor mediocampista ofensivo es Jorge Valdivia (CC), seguido de dos jugadores que solo jugaron la primera mitad del torneo en sus clubes. Se destaca la poca cantidad de minutos jugados por Jorge Valdivia con respecto a sus compañeros de posición y de otras posiciones debido a las constantes lesiones que sufrió el jugador. En esta posición, las variables tienen ponderaciones similares y no se ve una que destaque sobre las otras. Dado que no existe un jugador de la Universidad de Chile que haya cumplido el umbral de minutos jugados en esta posición, se centra el análisis en la comparación entre Jorge Valdivia y Maximiliano Salas (OHI). En la figura 4.3 se observa que Valdivia tiene un rendimiento sobresaliente en las variables *% Duelos Ofensivos* con un alto porcentaje de conversión (88 %) y en *Asistencias cada 90 minutos* donde pertenece al percentil 61 % y Salas solo al percentil 39 %. Sin embargo, Salas tiene un mejor rendimiento en *% Remates fuera del área* lo que explica la poca diferencia del *Score* final entre ambos jugadores (Valdivia: 0.402, Salas: 0.400).

4. **Delantero**

Nombre	Club	Mins Jug	% Duelos Defensivos	Inter. c/90mins	Asist. c/90mins	% Remates al arco	Dif xG	% Duelos Aéreos	Score
Pond.	-	-	**5 %**	**5 %**	**15 %**	**25 %**	**25 %**	**20 %**	-
J. Fuenzalida	U. Católica	456	0	0.4	0.4	84	3.8	0	0.362
N. Orellana	U. de Concepción	621	36	0.4	0.1	81	0.1	64	0.336
J. Sánchez	Huachipato	807	20	0	0.1	72	3.3	33	0.325

Tabla 4.4: Estadísticas top 3 delanteros

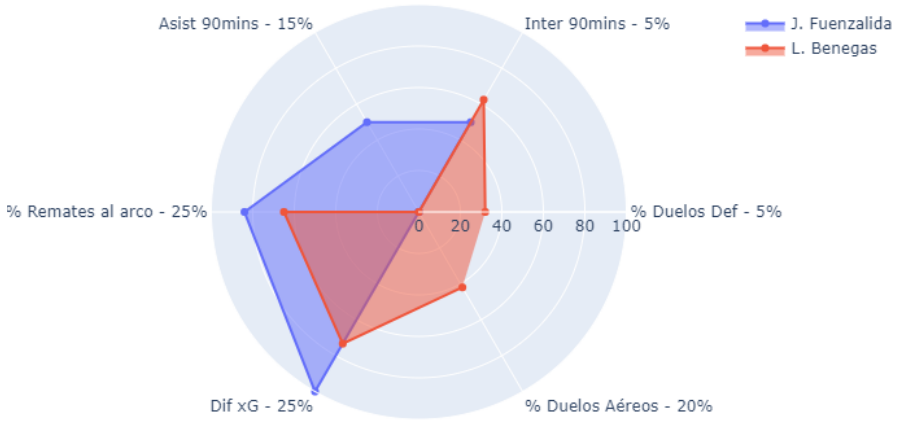


Figura 4.4: José Pedro Fuenzalida (UC) vs Leandro Benegas (UCH)

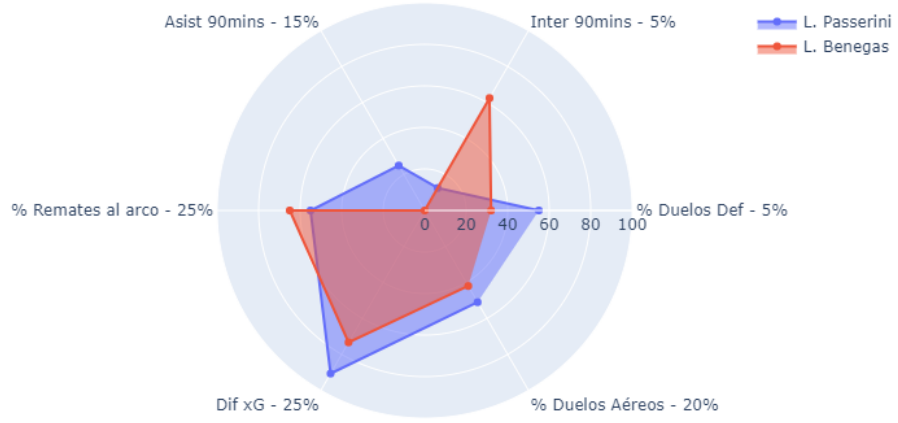


Figura 4.5: Lucas Passerini (PAL) vs Leandro Benegas (UCH)

En la tabla 4.4 se puede observar el rendimiento de los tres mejores delanteros del torneo. José Pedro Fuenzalida (UC) es el que tiene un mayor *Score*, principalmente, porque tiene una alta diferencia a favor en Expected Goals, es decir, de acuerdo a la calidad de sus oportunidades generadas, realizó 3.8 goles más que el jugador promedio. Esta variable tiene una alta ponderación del 25 %. Además, sus remates son de alta calidad, dado el alto rendimiento en la variable *% Remates al arco* que también tiene una ponderación del 25 %. Todo lo anterior es suficiente para sopesar el bajo rendimiento en las variables *% Duelos Aéreos* y *% Duelos Defensivos* que la ponderación es un 20 % y 5 %, respectivamente.

En las figuras 4.4 y 4.5 se puede observar la comparación del mejor delantero de la Universidad de Chile, Leandro Benegas con José Pedro Fuenzalida y con Lucas Passerini, goleador del torneo con 14 goles. A primera vista, se observa una diferencia en los estilos de juego entre Fuenzalida y Benegas dado que presentan una alternancia de rendimiento en las variables en comparación. Benegas tiene una alta componente defensiva en temas de intercepciones y duelos defensivos ganados, pero que tienen bajas ponderaciones (5 % cada una), en cambio, Fuenzalida es percentil 100 % en la variable *Dif xG*, dado que fue el mejor en ese ítem y tiene buen rendimiento en la entrega de asistencias a sus compañeros. En cambio, cuando comparamos a Benegas y Passerini, se observa que sus estilos de juego son más similares entre sí, sin embargo, Passerini tiene mejor rendimiento en conversión de goles a través de la variable *Dif xG* y en la entrega de asistencias a sus compañeros.

4.2. Modelo Cadenas de Markov vs modelo Poisson

Para simular ambos modelos se utilizó un notebook Asus X456UB, con procesador Intel(R) Core(TM) i5-6200U CPU 2.40Ghz, 8.00 GB de RAM y un sistema operativo de 64bits. El tiempo de ejecución de N = 10000 torneos del modelo de Poisson con 6 fechas es, en promedio, de 6 minutos y para el modelo de Cadenas de Markov para 6 fechas es de, en promedio, 1 hora aproximadamente. Para simular el torneo completo con el modelo de Cadenas de Markov (192 fechas), el tiempo de ejecución bordea las 5 horas y 30 minutos.

Para validar el modelo de simulación de Cadenas de Markov, se utiliza una medida de error para ver cuánto se desvían los resultados de las simulaciones con respecto a la realidad. El rendimiento base será calculado con el modelo de simulación de Poisson. Para esto se calibran ambos modelos con los datos obtenidos de las primeras 18 fechas de 24 equivalente a 144 partidos de 192 (75 %) y se testeará con las últimas 6 fechas jugadas, es decir, de la fecha 19 a la 24 equivalente a 48 partidos de 192 (25 %). La medida de error a utilizar es la mencionada en 3.13. En la tabla 4.5 se muestra los resultados de la fecha 24 y el formato en el que están los resultados de las últimas fechas. Para ver los resultados de las últimas fechas, revisar Anexos 5.1.

Local	Visita	GL	GV
O'Higgins	Universidad Concepción	1	2
Cobresal	Palestino	3	2
Unión Española	Coquimbo Unido	3	1
Deportivo Antofagasta	Audax Italiano	1	0
Unión La Calera	Universidad Católica	1	0
Colo Colo	Huachipato	2	2
Everton	Curicó Unido	2	1
Universidad de Chile	Deportes Iquique	2	1

Tabla 4.5: Resultados Fecha 24

Una medida de error preliminar, es calcular el porcentaje de acierto al resultado del partido: ganó el local, hubo empate o ganó la visita. Para esto, se simularon las últimas 6 fechas (48 partidos) 10,000 veces cada partido. El porcentaje de acierto es del modelo de Poisson es de un 35.8 % y el del modelo de Cadenas de Markov un 33.8 % . Además, se computa el error de ambos modelos de acuerdo a la métrica 3.13 para distintas ponderaciones de α y β, donde α representa la importancia de acertar a la condición ganador (local, empate, visita) y β la importancia de la diferencia de goles real en comparación con la simulada.

Alpha	Beta	Modelo Poisson	Modelo CM
0.50	0.250	55.18	66.41
0.75	0.125	46.51	53.89
0.90	0.050	39.55	43.79
1.0	0.0	30.81	31.88

Tabla 4.6: Error del modelo para distintos valores de α y β

En resumen, el modelo de Cadenas de Markov tiene un rendimiento de aciertos menor al de Poisson y con un error mayor. Esta diferencia en el error, se puede explicar por que en el modelo de Cadenas de Markov se tienden a convertir más goles que los realmente realizados. Sin embargo, la gran ventaja del modelo de Cadenas de Markov, es que podemos analizar el rendimiento del equipo, no sólo en base a la historia de goles convertidos, sino que en base a los rendimientos individuales de los jugadores y eso permite individualizar qué sectores del terreno de juego se deberían mejorar y con qué jugadores se podría reforzar.

4.3. Recomendación de jugadores

Para utilizar el modelo de simulación de Cadenas de Markov como recomendador de jugadores, es necesario calcular la probabilidad de ser campeón de cada uno de los equipos sin realizar modificaciones a las alineaciones utilizadas. Luego, se realizarán reemplazos de jugadores y se calculará la nueva probabilidad de ser campeón del equipo, con el objetivo de ver si la diferencia de la inclusión del nuevo jugador es significativa (recordar la expresión 3.12). En la tabla 4.7 se observa la línea base de probabilidades de salir campeón de cada equipo.

Equipo	Probabilidad Campeón
Universidad Católica	46.36 %
Audax Italiano	16.92 %
Curicó Unido	8.40 %
Universidad de Chile	7.98 %
Everton	6.80 %
Deportes Antofagasta	3.20 %
Unión Española	3.07 %
Colo Colo	2.68 %
O'Higgins	1.85 %
Unión La Calera	1.24 %
Cobresal	0.49 %
Coquimbo Unido	0.43 %
Palestino	0.33 %
Universidad Concepción	0.22 %
Huachipato	0.03 %
Deportes Iquique	0.00 %

Tabla 4.7: Probabilidad de salir campeón según Cadena de Markov

Los jugadores utilizados para reemplazar son los siguientes: Matías Dituro (por Fernando De Paul y Johnny Herrera), Julio Barroso (por Rodrigo Echeverría), Eduardo Farías (por Camilo Moya), Jorge Valdivia (por Gonzalo Espinoza), José Pedro Fuenzalida (por Ángelo Henríquez) y Lucas Passerini (por Leandro Benegas).

Caso	Probabilidad Campeón	Delta
Nivel Base	7.98 %	0 %
Matías Dituro	10.17 %	+2.19 %
Julio Barroso	8.68 %	+0.70 %
Eduardo Farías	7.07 %	-0.91 %
Jorge Valdivia	11.23 %	+3.25 %
José Pedro Fuenzalida	10.64 %	+2.66 %
Lucas Passerini	15.59 %	+7.61 %

Tabla 4.8: Cambio en la probabilidad de ser campeón de U. de Chile

Los cambios en las probabilidades de ser campeón del club Universidad de Chile, se muestran en la tabla 4.8. Se observa que las posiciones que tienen directa relación con la realización de goles, son las que el delta de probabilidad es mayor (delanteros, mediocampista ofensivo y arquero) en desmedro de las posiciones que no tienen relación directa con al conversión de goles, tales como defensas o mediocampista centrales. Para observar las probabilidades de todos los equipos, dirigirse a 5.2 en el apéndice.

Capítulo 5

Conclusiones

En este trabajo, se utilizan datos de partidos de la primera división del fútbol chileno año 2019 a través del proveedor Opta Sports, que logró un acuerdo comercial con el club de fútbol profesional Universidad de Chile. Esta información contenía resultados de los partidos y una bitácora de eventos de cada uno de los partidos disputados. En el análisis exploratorio de datos, se pudo determinar qué acciones del juego son las que más ocurren y las más relevantes para el resultado final del partido. Además, se logró observar las diferentes distribuciones de acciones de cada uno de los jugadores, condicionadas a la posición que usa dentro del terreno de juego, la zona del campo dónde se realiza la acción o si está jugando de local o de visita.

Se logró modelar dos tipos de modelos de recomendación, uno basado en estadística descriptiva y el otro basado en simulaciones. El de estadística descriptiva, muestra cuáles son los jugadores de cada posición, que tienen el mejor rendimiento en base a variables y ponderaciones de ellas, escogidas en conjunto con los directores deportivos y el área de secretaría técnica del club. Por otro lado, a través del modelo basado en simulaciones, se consiguió un rendimiento predictivo un poco menor que el modelo de Poisson (33.8 % vs 35.8 %) y que permite computar el aumento en la probabilidad de ser campeón del club Universidad de Chile a través de la inclusión de jugadores como Passerini (PAL), con un aumento del 7.61 % o Jorge Valdivia (CC), con un aumento del 3.25 % de ser campeón.

Se cumplió el objetivo general de crear un sistema de recomendación de jugadores para el club de fútbol Universidad de Chile, tomando como base el rendimiento deportivo de los jugadores en la temporada 2019 del torneo de fútbol chileno con los datos obtenidos gracias a Opta Sports.

Para cumplir el objetivo general, se tuvo que cumplir con cada uno de los objetivos específicos. Se logró consultar correctamente los datos proporcionados por Opta Sports, a través de su API, para realizar un correcto análisis exploratorio de los distintos *feeds*. Por otra parte, se realizó la definición de las variables con las que se evalúan a los jugadores, en cada una de las posiciones junto a sus respectivas ponderaciones. Además, se logró generar un ranking de jugadores, en base al modelo de estadística descriptiva, lo que permitió tener un input de qué jugadores podrían ser reemplazados en el modelo de simulación de Cadenas de Markov, el cual fue programado correctamente para obtener una nueva lista de recomendación de jugadores en base a la métrica del cambio en la probabilidad de salir campeón del club Universidad de Chile.

Desde el punto de vista técnico, se pueden realizar mejoras al modelo de simulación basado en Cadenas de Markov a través de, aumentar la cantidad de zonas en las que se divide el terreno de juego, utilizar un ponderador que permita capturar el efecto de los partidos importantes o el nivel propio del equipo, realizar una diferenciación de las distribuciones, incorporando estilos de juegos en el primer y segundo tiempo o, agregar los eventos de tiempos muertos. Otra mejora puede ser realizar esta simulación pero considerando el partido de fútbol como una Cadena de Markov en tiempo continuo donde existen tasas de transición entre los distintos eventos considerados para el modelamiento actual.

Desde el punto de vista del negocio, el modelo permite incorporar aún más requerimientos de parte del club en relación a los filtros que deseen para buscar jugadores y los resultados a través de gráficos y tablas son intuitivos para ellos. Sin embargo, este modelo es una representación de la realidad ,pero se puede mejorar incluyendo más eventos e información de los jugadores, no solo en el aspecto técnico, sino que también en los aspectos médicos, fisiológicos, sicológicos, nutricional y otros ámbitos que rodean la práctica deportiva. También se pueden agregar restricciones propias del negocio, como presupuestos finitos para la contratación de jugadores o restricciones reglamentarias, como la restricción del número de extranjeros en un plantel y el número de refuerzos que se pueden contratar en cada periodo de traspaso.

Apéndice

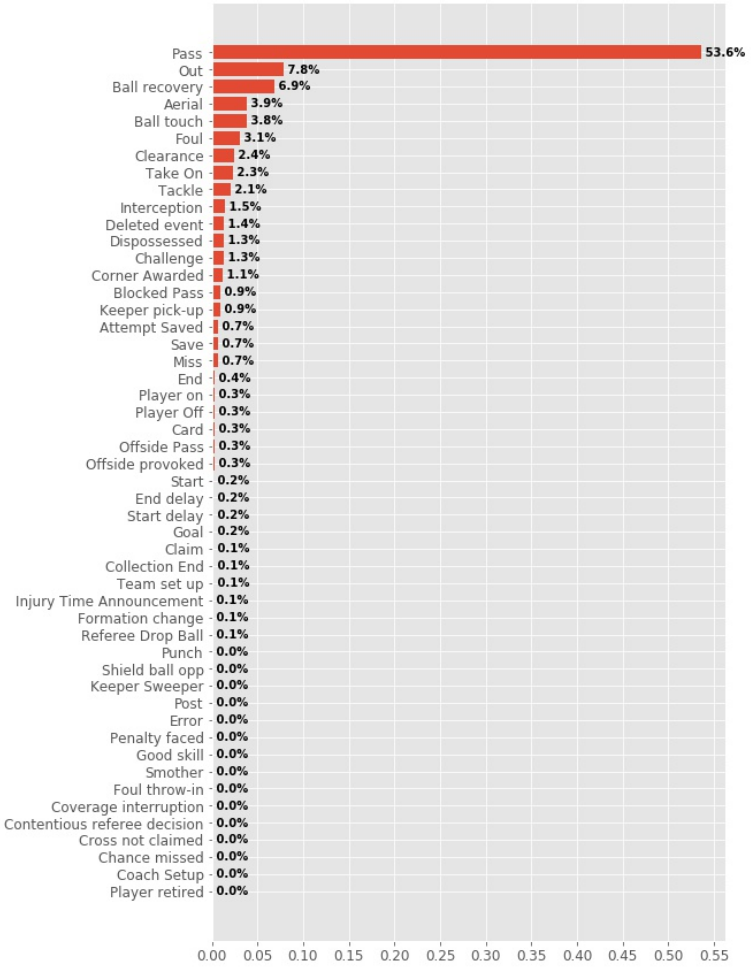


Figura 5.1: Distribución de eventos torneo primera división Chile 2019

Local	Visita	Goles Local	Goles Visita
Fecha 19			
Unión Española	O'Higgins	0	1
Cobresal	Unión La Calera	1	0
Palestino	Colo Colo	2	2
Everton	Coquimbo Unido	0	2
Curicó Unido	Audax Italiano	2	2
Deportes Iquique	Universidad Concepción	2	2
Universidad de Chile	Universidad Católica	1	1
Huachipato	Deportivo Antofagasta	4	3
Fecha 20			
Unión La Calera	Everton	0	0
Curicó Unido	Deportes Iquique	1	2
Coquimbo Unido	Universidad de Chile	2	2
Audax Italiano	Unión Española	3	2
Universidad Concepción	Palestino	0	0
Deportivo Antofagasta	Universidad Católica	1	2
O'Higgins	Huachipato	0	3
Colo Colo	Cobresal	0	2
Fecha 21			
Huachipato	Curicó Unido	2	0
Palestino	Coquimbo Unido	2	0
Deportes Iquique	Audax Italiano	1	3
Unión Española	Universidad de Chile	1	1
Universidad Concepción	Colo Colo	3	1
Cobresal	Everton	2	1
Universidad Católica	O'Higgins	0	0
Unión La Calera	Deportivo Antofagasta	2	3
Fecha 22			
Everton	Universidad Concepción	1	0
Coquimbo Unido	Deportivo Antofagasta	1	1
Universidad de Chile	Palestino	2	3
Curicó Unido	Universidad Católica	1	4
Unión Española	Deportes Iquique	4	0
Audax Italiano	Colo Colo	2	4
Cobresal	Huachipato	1	1
O'Higgins	Unión La Calera	3	1
Fecha 23			
Huachipato	Unión La Calera	1	2
Deportivo Antofagasta	O'Higgins	3	1
Colo Colo	Universidad de Chile	3	2
Universidad Concepción	Unión Española	1	1
Deportes Iquique	Everton	0	0
Palestino	Curicó Unido	4	2
Universidad Católica	Cobresal	5	0
Coquimbo Unido	Audax Italiano	1	0

Tabla 5.1: Resultados fecha 19 a la 23 torneo primera división Chile 2019

	M. Dituro	**J. Barroso**	**E. Farías**	**J. Valdivia**	**J. Fuenzalida**	**L. Passerini**
Universidad Católica	46.21 %	46.20 %	46.35 %	43.70 %	44.41 %	43.56 %
Audax Italiano	17.03 %	17.36 %	18.12 %	15.63 %	18.10 %	15.83 %
Universidad de Chile	10.17 %	8.68 %	7.07 %	11.23 %	10.64 %	15.59 %
Curicó Unido	7.61 %	7.91 %	8.48 %	8.74 %	7.84 %	7.69 %
Everton	6.22 %	6.22 %	6.50 %	7.13 %	6.49 %	5.43 %
Deportes Antofagasta	3.18 %	3.51 %	3.20 %	2.70 %	3.13 %	3.35 %
Unión Española	2.96 %	2.95 %	3.34 %	3.44 %	2.81 %	2.69 %
Colo Colo	2.40 %	2.59 %	2.60 %	2.43 %	2.41 %	2.27 %
O'Higgins	1.56 %	1.63 %	1.71 %	1.81 %	1.45 %	1.21 %
Unión La Calera	1.23 %	1.44 %	1.22 %	1.82 %	1.35 %	1.23 %
Cobresal	0.48 %	0.45 %	0.41 %	0.46 %	0.47 %	0.39 %
Coquimbo Unido	0.43 %	0.50 %	0.48 %	0.46 %	0.47 %	0.32 %
Palestino	0.24 %	0.33 %	0.34 %	0.27 %	0.25 %	0.27 %
Universidad Concepción	0.27 %	0.22 %	0.16 %	0.15 %	0.13 %	0.16 %
Huachipato	0.01 %	0.01 %	0.02 %	0.03 %	0.05 %	0.01 %
Deportes Iquique	0.00 %	0.00 %	0.00 %	0.00 %	0.00 %	0.00 %

Tabla 5.2: Probabilidades de ser campeón de cada club luego de reemplazar cada jugador en las alineaciones de la Universidad de Chile

Bibliografía

[1] Wikipedia. Definición análisis de datos de John Tukey *https://es.wikipedia.org/wiki/Análisis_de_datos*. [Consulta: 27/11/2019]

[2] New York Times. Resultados búsqueda *Data Analysis* en sitio web del New York Times *https://www.nytimes.com/search?dropmab=true&query=data %20analysis&sort=best*. [Consulta: 13/12/2019]

[3] Michel Lewis *Moneyball: the art of winning an unfair game*, 2003

[4] New York Times. Bruce Schoenfeld. How Data (and Some Breathtaking Soccer) Brought Liverpool to the Cusp of Glory *https://www.nytimes.com/es/2019/05/29/liverpool-champions/*. [Consulta: 10/08/2019]

[5] Michael Hughes, Tim Caudrelier, Nic James, Athalie Redwood-Brown, Ian Donnely, Anthony Kirkbride & Christophe Duschesne *Moneyball and soccer - an analysis of the key performance indicators of elite male soccer players by position*

[6] Luca Pappalardo, Paolo Cintia, Alessio Rossi, Emanuele Massucco, Paolo Ferragine, Dino Pedreschi & Fosca Giannotti *A public data set of spatio-temporal match events in soccer competitions*, 2019

[7] Tom Decroos, Jan Van Haaren & Jeese David *Automatic discovery of tactivs in spatio-temporal soccer match data*, 2018

[8] Qing Wang, Hengshu Zhu, Wei Hu, Zhiyong Shen & Yuan Yao *Discerning tactical patterns for professional soccer teams: an enhanced topic model with applications*, 2015

[9] Opta Sports. Advanced Metrics. *https://www.optasports.com/services/analytics/advanced-metrics/* [Consulta: 04/02/2020]

[10] Wikipedia, Test de Smirnov-Kolmogorov *https://en.wikipedia.org/wiki/Kolmogorov %E2 %80 %93Smirnov_test#Two-sample_Kolmogorov %E2 %80 %93Smirnov_test*. [Consulta: 05/01/2020]

[11] ANFP, Chile. Bases torneo primera división año 2019, artículo 82 *http://www.anfp.cl/documentos/1556311422-bases-primera-division-2019-26042019-publicar.pdf* [Consulta: 15/01/2020]

[12] Mark Dixon & Stuart Coles *Modelling association football scores and inefficiencies in the football betting market*, 1997

[13] Wikipedia. Regresión de Poisson *https://es.wikipedia.org/wiki/Regresi %C3 %B3n_de_Poisson* [Consulta: 20/01/2020]

[14] FIFA. Criterios de puntuación FIFA

https://es.fifa.com/mm/document/fifafacts/rawrank/ip-590_04s_wrlong_8784.pdf [Consulta: 02/02/2020]

[15] New York Times. Rory Smith. Sevilla and the Science of Soccer's Summer Transfer Window *https://www.nytimes.com/2019/08/02/sports/transfer-window-sevilla-monchi.html* [Consulta: 10/08/2020]