Universidad del Desarrollo
Universidad de Excelencia

Football Analytics - Ranking de Jugadores

POR: Ana Harringthon y William Marín

Capstone project presentado a la Facultad de Ingeniería de la Universidad del Desarrollo para optar al grado académico de Magíster en Data Science.

PROFESORES GUÍA:

Loreto Bravo
Hugo Contreras

Diciembre 2022
Santiago de Chile

Tabla de Contenido

Resumen

La evaluación de jugadores de fútbol considerando múltiples características de juego y su comparación con otros jugadores de roles similares, resulta una tarea compleja hoy en día. Este trabajo tiene como objetivo principal la construcción de un ranking de jugadores de fútbol de nacionalidad chilena por cada posición del campo. Para esto, se construyeron modelos de predicción binario con bases en técnicas estadísticas y del Machine Learning como son el Support Vector Machine y la Regresión logística. Se calculó el coeficiente de cada variable estudiada, más de 54 variables de entrada en este caso, y se tomaron como pesos o ponderaciones, que, a su vez, se multiplicaron por el promedio que cada jugador tenía en cada una de estas variables. El resultado final resultó de la sumatoria de todas las variables y del ordenamiento según el resultado.

La base de datos que se usó fue aportada por la plataforma Wyscout, que registra cada uno de los números que cada partido y jugador en una plantilla completa y que puede obtenerse a través de suscripción. Antes de la construcción de los modelos, se hizo una minuciosa limpieza de datos donde algunas posiciones y variables fueron suprimidas. Entre los resultados destacan figuras importantes del fútbol chileno de la última década, como son Alexis Sanchez, Eduardo Vargas, Arturo Vidal, Charles Aránguiz, Paulo Diaz, Leonardo Gil, entre otros.

1. Introducción

La estadística ha sido una herramienta útil en el deporte. A través de esta, entrenadores de selecciones nacionales y propietarios de clubes han cuantificado el desempeño tanto del equipo como de sus integrantes. Además, periodistas y medios de comunicación construyen sus reseñas o comentarios con este fundamental recurso, los mismo hacen los cazatalentos para el fichaje de los nuevos jugadores. Sin embargo, el gran impulso de la tecnología ha permitido un almacenamiento cada vez mayor de datos, de manera casi simultánea, y lo que es más revolucionario, el análisis de esos grandes volúmenes de información, lo que se ha denominado como la era del Big Data. El deporte es un área que no ha sido ignorada por esta revolución digital.

Actualmente, en el fútbol existen numerosas herramientas que tienen como objetivo principal recoger el mayor número de datos posibles de todos los eventos que ocurren durante un partido, con precisión exacta de acuerdo con el tiempo, lugar del campo y del jugador o jugadores involucrados, algunas de ellas son OptaSport, WyScout o MediaCoach [1]. No obstante, y a pesar de todos estos recursos, no existe un método integral, que incluya todas las estadísticas o características registradas en un partido y que permita clasificar a los jugadores, a diferencia de otras disciplinas deportivas por equipos como es el Basketball o Hockey sobre hielo. Los recursos actuales son unidimensionales [2], ya que son métricas fundamentadas en un único aspecto del juego, como son los tiros al arco, pases, duelos, entre otras, que evalúan una característica muy específica del jugador.

Wyscout es una de las plataformas más utilizadas en el mundo del fútbol para análisis de partidos, rendimiento y exploración de múltiples estadísticas recolectadas desde el 2004 cuando fue fundada. La plataforma provee información sobre todos y cada uno de los jugadores, competencias y partidos [3]. Es posible obtener acceso a la información recolectada por la empresa contratando una suscripción anual que, dependiendo del paquete, proveerá más o menos privilegios.

Con el uso de la información de esta última plataforma se construyó un conjunto de rankings por cada posición del campo entre jugadores de nacionalidad chilena, independientemente del origen de la liga donde participan. En primer lugar, se

implementó un sistema de predicción binario para la victoria, fundamentado en Support Vector Machine o Regresión Logística, y luego, los coeficientes resultantes de cada variable se multiplicaron por el promedio sobre noventa minutos de cada jugador. El ranking fue producto de la sumatoria de los pesos de features y su ordenamiento. Posteriormente, se hace una discusión de los resultados y se compara con la información que ha sido reportada por algunos medios de comunicación dedicados al deporte. Como se describe en el trabajo, los resultados son paralelos a lo señalado por algunos expertos. De esta manera, se tiene un sistema, más amplio e integral, capaz de tomar en cuenta otras aristas de un partido para la evaluación de los jugadores, a diferencia de los sistemas unidimensionales ya existentes.

2. Trabajo Relacionado

La principal fuente de información usada para la realización de este trabajo proviene del científico italiano Lucca Pappalardo, del Instituto de Ciencias de la Información y Tecnología de Pisa, titulada: PlayeRank: data-driven performance evaluation and player ranking in soccer via a machine learning approach. Junto a sus colaboradores, Pappalardo se plantea la interrogante de crear un framework integral que comparara a todos los jugadores de fútbol de las principales ligas del mundo, independientemente de su posición en el campo del juego, y los ordena en un ranking de mayor a menor según su desempeño. En la metodología se utilizó una base de datos 76 variables aportada por la empresa Wyscout, especialista en datos para la industria del fútbol, y que contenía toda la información de partidos jugados y de cada jugador. Para obtener el ranking, Pappalardo hace una fase de entrenamiento del modelo, donde construye un sistema de clasificación con base en el soporte de máquinas vectoriales, en la cual se extrae el peso (w) de cada uno de los features. Luego, realiza una fase de Rating por jugador, resultado del producto escalar entre el peso de los features (w) y el valor del feature por cada jugador (x), extraído de una evaluación previa del performance. Para el ranking, se obtiene el ordenamiento según la sumatoria del paso previo y se construye por cada una de las posiciones. Paralelo al ranking, este grupo de investigadores también predijo la posición del campo jugada según los datos mediante técnicas de aprendizaje no supervisado [2].

Previo al trabajo innovador de Lucca Pappalardo, sólo estudios con análisis unidimensionales se habían publicado en la literatura, entre ellos, el trabajo de Duch en 2008, quien propuso la métrica de Flow Centrality (FC), que se define como la fracción de veces que un jugador interviene en los pases que culminan en un gol. Sin embargo, los mismos autores consideran que es solo aplicable efectivamente a jugadores del mediocampo o delanteros, con poca utilidad en el resto de las posiciones [4]. Para el año 2012, Brooks y colaboradores desarrollan el Pass Shot Value (PSV), una métrica que estima la importancia de un pase para generar un gol. Este grupo de investigadores representan el pase como un vector de 360 variables que describe la cercanía en la zona del juego con el origen y destino del pase. Aunque para su implementación se usaron técnicas de Machine Learning para predecir si un pase determinado puede culminar como

un gol, y a pesar de que se construyó un ranking de jugadores de La Liga mediante este método, la técnica de PSV está fuertemente sesgada y sobreestima a las posiciones de juego que son ofensivas [5].

Otros intentos previos para calcular el rating, y con base en metodologías usadas en otros deportes por equipo, han intentado clasificar clubes y jugadores de fútbol. Un ejemplo son las investigaciones hechas por Elo para TrueSkill, cuyo ranking se fundamentó en la tasa de victorias/derrotas y en la estimación de la fortaleza del oponente. A diferencia de este último, la ventaja del rating construido por Pappalardo radica en el uso de features obtenidos en amplias fuentes de información específica de partidos y jugadores en varios años, esto se traduce en un modelo más integral y que considera el aporte de cada integrante del equipo para la victoria [6]

Es incuestionable todas las ganancias en marketing que genera el fútbol y las astronómicas cifras de dinero que ganan las principales figuras de este deporte. Por esta razón, algunos trabajos que relacionan performance con el valor del mercado han sido publicados. Stanojevic y Gyarmati usaron datos de jugadores y partidos para inferir la relación entre el desempeño, el valor monetario y su apreciación por la fanaticada. Estos autores encontraron grandes discrepancias entre el valor real y estimado en el mercado, y que fueron atribuidos principalmente a lesiones y a factores relacionados a la salud e integridad física de los jugadores [7]

A pesar de la inmensa popularidad del fútbol, y su amplia difusión en todo el planeta, el estudio de sus estadísticas, y el uso de los datos provenientes de partidos para la construcción de modelos predictivos, no es común como sí lo es en otros deportes por equipo como el Baseball, Hockey (Scoring Impact Metric) y Basketball (Performance Efficiency Rating) [8]. Recientemente, el interés de usar la información almacenada ha ido creciendo; sin embargo, aún no hay un modelo multidimensional para la clasificación de futbolistas. La motivación de Pappalardo fue romper este paradigma y construir un modelo, lo más integral posible, que pudiera ser una buena aproximación a la correcta para la clasificación. Los resultados obtenidos por el grupo de científicos italianos coinciden con la opinión de muchos expertos y periodistas deportivos. Por esta razón, como científicos de datos en formación, nos parece el más completo, y además, surgió la

motivación del uso de esta metodología para el estudio de las ligas de primera división del fútbol chileno. De este modo, se estaría contribuyendo al estudio riguroso del deporte del país y se abren las puertas para otros estudios que pudieran aumentar el rendimiento de los deportes en equipos en el país.

3. Objetivos

3.1. **Objetivo General:** Construir un ranking de jugadores chilenos de fútbol por posición según su desempeño en los partidos disputados.

3.2. Objetivos Específicos:

- Implementar un modelo de clasificación binario para la predicción de la victoria con base en la información obtenida en la plataforma WyScout.
- Identificar qué variables de un partido de fútbol tienen mayor importancia para determinar la victoria o derrota del equipo.
- Discutir los resultados obtenidos y compararlos con la opinión de expertos o periodistas.

4. Plan de Trabajo

TAREAS	S1	S2	S3	S4	S5	S6	S7	S8	S9	S10	S11	S12	S13	S14	S15	S16
Definir Proyecto(Objetivo, potencial)	■	■														
Revisión de Referencias / Trabajos previos	■	■														
Elaboración de breve informe y planificación	■	■														
Entrega Etapa 1	■	■	■													
Definición de features en comun entre datos de jugadores y partidos		■	■													
Pre-procesamiento y limpieza de datos		■	■													
Exploracion de datos			■	■	■											
Resultados Preliminares				■	■	■										
Entrega Etapa 2				■	■	■										
Desarrollo / Ajustes Proyecto						■	■	■	■	■	■	■	■			
Elaboración trabajo escrito										■	■	■	■			
Entrega Etapa 3.1													■	■		
Presentacion Oral													■	■	■	■

SEMANA	INTERVALO
S1	(5-sep al 11-sep)
S2	(12-sep al 18-sep)
S3	(19-sep al 25-sep)
S4	(26-sep al 2-oct)
S5	(5-oct al 9-oct)
S6	(10-oct al 16-oct)
S7	(17-oct al 23-oct)
S8	(24-oct al 30-oct)
S9	(24-oct al 30-oct)
S10	(31-oct al 6-nov)
S11	(7-nov al 13-nov)
S12	(14-nov al 20-nov)
S13	(21-nov al 27-nov)
S14	(28-nov al 4-dic)
S15	(4-dic al 11-dic)
S16	(12-dic al 18-dic)

5. Datos y Metodología

5.1. Datos

Se utilizaron varios sets de datos contenidos en WyScout, donde se almacena la información de todos los partidos de las últimas temporadas de fútbol de las diferentes ligas de Chile. Igualmente, hay registro de las estadísticas por temporada de cada jugador de nacionalidad chilena, independientemente del país donde compite. Esta variedad de datos permite la construcción de modelos que permiten aproximar la calidad del jugador y su importancia en el resultado de la victoria, así como la construcción de un ranking por cada una de las posiciones.

Cada fuente (de jugadores y de partido de los equipos chilenos) contiene más de 100 atributos, los cuales serán comparados a fin de seleccionar los que se encuentran en ambas fuentes con el objetivo de poder hacer una evaluación de los jugadores con base en los datos de partidos.

En primer lugar, se realizó un levantamiento de los campos en común en cada set de datos encontrando 69 atributos en común, o con posibilidad de calcular según a otros campos dentro del mismo set de datos. Los atributos encontrados se muestran a continuación:

1. Tiros
2. Tiros a la portería
3. % Tiros a la portería
4. Tiros libres
5. Pases
6. Pases logrados
7. Pases fallidos
8. % Pases logrados
9. Duelos
10. Duelos ganados
11. Duelos perdidos
12. % Duelos ganados
13. Córneres

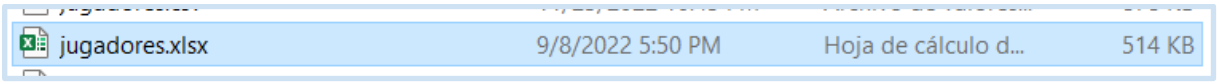
14. Penaltis
15. Penaltis marcados
16. Penaltis fallados
17. % Penaltis marcados
18. Centros
19. Centros precisos
20. % Centros precisos
21. Centros sin precisión
22. Pases en profundidad completados
23. Toques en el área de penalti
24. Duelos ofensivos
25. Duelos ofensivos ganados
26. Duelos ofensivos perdidos
27. % Duelos ofensivos ganados
28. Duelos defensivos
29. Duelos defensivos ganados
30. Duelos defensivos perdidos
31. % Duelos defensivos ganados
32. Duelos aéreos
33. Duelos aéreos ganados
34. Duelos aéreos perdidos
35. % Duelos aéreos ganados
36. Interceptaciones
37. Faltas
38. Tarjetas amarillas
39. Tarjetas rojas
40. Pases hacia adelante
41. Pases hacia adelante logrados
42. Pases hacia adelante fallidos
43. % Pases hacia adelante logrados
44. Pases hacia atrás
45. Pases hacia atrás logrados
46. Pases hacia atrás fallidos

47. % Pases hacia atrás logrados
48. Pases laterales
49. Pases laterales logrados
50. Pases laterales fallidos
51. % Pases laterales logrados
52. Pases largos
53. Pases largos logrados
54. Pases largos fallidos
55. % Pases largos logrados
56. Pases en el último tercio
57. Pases en el último tercio logrados
58. Pases en el último tercio fallidos
59. % Pases en el último tercio logrados
60. Pases progresivos
61. Pases progresivos precisos
62. Pases progresivos sin precision
63. % Pases progresivos precisos
64. Desmarques
65. Desmarques logrados
66. Desmarques fallidos
67. % Desmarques logrados
68. Lanzamiento largo %
69. Longitud media pases

Una vez identificados los atributos en común, se procedió a realizar el procesamiento y limpieza de datos en cada fuente de información: partidos y jugadores. La finalidad de esto es obtener una entrada de datos al modelo limpia y organizada, además de construir variables necesarias para el desarrollo de este estudio y seleccionar los atributos en común en ambas fuentes.

5.1.1. Set de datos Jugadores

El set de datos de jugadores corresponde a un solo archivo Excel consolidado con las estadísticas de los jugadores chilenos provenientes de la plataforma Wyscout, se guardó este único archivo en un dataframe para su análisis. Este contiene 743 jugadores con distintas variables sobre su performance e información general sobre el equipo donde juega, posición, entre otros. En la imagen 1 se muestra archivo fuente y algunos campos contenidos en dicho archivo.



A	B	C	D	E	F	G
Jugador	Equipo	Equipo durante el período seleccionado	Posición específica	Edad	Valor de mercado	Vencimiento contrato
B. Brereton Díaz	Blackburn Rovers	Blackburn Rovers	CF, LAMF, LWF	23	$ 16,000,000	2023-06-30
G. Maripán	Monaco	Monaco	RCB, LCB	28	$ 15,000,000	2024-06-30
P. Díaz	River Plate	River Plate	RCB, LCB	27	$ 7,000,000	2024-12-31
E. Pulgar	Galatasaray	Galatasaray	DMF, RDMF, LDMF	28	$ 6,500,000	2023-06-30
D. Valdés	América	América	AMF, CF, LCMF	28	$ 6,000,000	2024-12-31
S. Vegas	Monterrey	Monterrey	LCB, LB	25	$ 6,000,000	
V. Dávila	León	León	CF, AMF	24	$ 5,000,000	2024-12-31
C. Baeza	Toluca	Toluca	RCMF, LCMF, RDMF	28	$ 4,500,000	2024-12-31
F. Mora	Portland Timbers	Portland Timbers	CF	28	$ 4,000,000	2024-12-31
N. Díaz	Mazatlán	Mazatlán	LB, LCB	22	$ 3,000,000	2024-12-31

Imagen 1. Fuente de datos de Jugadores.

En primera instancia, se tiene un campo llamado “Posición específica”, que define las posiciones que pudiera tener cada jugador, siendo la primera de la lista su posición principal. Para efectos de este estudio, se consideró esta posición principal para la construcción de los distintos rankings.

En la figura 1, se puede observar las 22 posiciones definidas por la plataforma Wyscout. Para simplificar el ranking resultante del estudio, se construyeron 6 grandes grupos que describen de forma general el rol de cada jugador en el campo:

1. Delantero → CF
2. Defensa → CB, RCB,RB,RWB,LCB,LB,LWB
3. Extremo → RW, RWF, LW, LWF
4. Mediocampo → CMF,RCMF, LCMF
5. Mediocampo ofensivo → AMF, RAMF, LAMF
6. Mediocampo defensivo → DMF, RDMF, LDMF

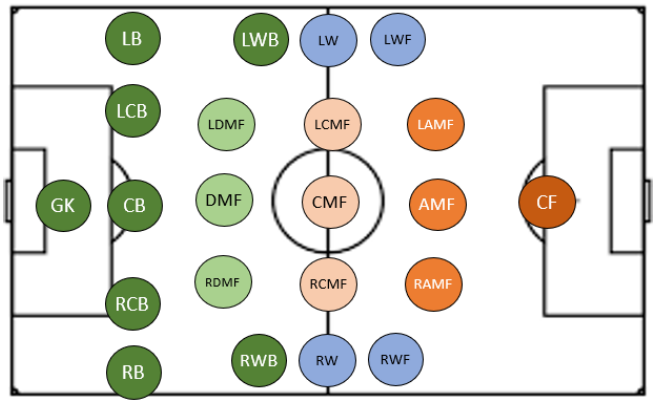


Figura 1. Posiciones en el campo de juego Wyscout.

El resultado final de esta nueva agrupación se puede observar en la imagen 2, donde se muestra el jugador, las posiciones que es capaz de jugar y los 2 nuevos campos mencionados anteriormente donde se toma como posición principal la primera posición de la lista y el campo que define en qué agrupación se encuentra dicha posición.

	Jugador	Posición específica	Pocision_Principal	posicion_gen
0	B. Brereton Díaz	CF, LAMF, LWF	CF	Delantero
1	G. Maripán	RCB, LCB	RCB	Defensa
2	P. Díaz	RCB, LCB	RCB	Defensa
3	E. Pulgar	DMF, RDMF, LDMF	DMF	Mediocampo defensivo
4	D. Valdés	AMF, CF, LCMF	AMF	Mediocampo Ofensivo
...	...	...	...	...
738	J. Beausejour	LB, LWB	LB	Defensa
739	H. Suazo	CF, AMF	CF	Delantero
740	G. Villegas	LB, LCB	LB	Defensa
741	A. Sanhueza	DMF, LCMF, RCMF	DMF	Mediocampo defensivo
742	M. Riquero	RCMF, LCMF, DMF	RCMF	Mediocampo

687 rows × 4 columns

Imagen 2. Definición de posición principal y su agrupación.

Debido a los atributos que existen en común entre el dataset de jugadores, se excluyeron de la evaluación la posición de Arquero, esto debido a que en el set de datos de partidos no se encuentran atributos para esta posición tan específica y, por lo tanto, no es posible entregar una evaluación certera. En consecuencia, el set de datos de Jugadores se redujo a 687 jugadores.

Posterior a la definición de la posición principal para cada jugador, y su categorización en los 6 grandes grupos, se seleccionaron los campos o features en común entre partidos y jugadores. Inicialmente, se identificaron los atributos y se renombraron con el nombre de la variable en el dataset de partidos tal como se muestra en la imagen

3, con el fin de tener las variables identificadas de igual forma en cada fuente y así, realizar una evaluación de forma clara y expedita en etapas futuras.

```
df_jug_match.rename(columns={
'Goles/90' : 'Goles',
'xG/90' : 'xG',
'Remates/90' : 'Tiros',
'Tiros a la portería, %' : '% Tiros a la portería',
'Pases/90' : 'Pases',
'Precisión pases, %' : '% Pases logrados',
'Duelos/90' : 'Duelos',
'Duelos ganados, %' : '% Duelos ganados',
'Córneres/90' : 'Córneres',
'Tiros libres/90' : 'Tiros libres',
'Penaltis a favor' : 'Penaltis',
'Penaltis realizados, %' : '% Penaltis marcados',
'Centros/90' : 'Centros',
'Precisión centros, %' : '% Centros precisos',
'Toques en el área de penalti/90' : 'Toques en el área de penalti',
'Duelos atacantes/90' : 'Duelos ofensivos',
'Duelos atacantes ganados, %' : '% Duelos ofensivos ganados',
'Goles recibidos/90' : 'Goles recibidos',
'Duelos defensivos/90' : 'Duelos defensivos',
'Duelos defensivos ganados, %' : '% Duelos defensivos ganados',
'Duelos aéreos en los 90' : 'Duelos aéreos',
```

Imagen 3. Renombramiento de variables en dataset de jugadores.

Una vez identificados y renombrados los atributos en común del set de datos de jugadores con los atributos de los partidos, se validó si existían nulos, encontrando que existen nulos para varios atributos. Para resolver este problema, se procedió a reemplazar dichos nulos por cero ya que en el caso que el atributo sea un valor nulo no aportará ningún punto en la evaluación que se genere posterior a la construcción del modelo y pesos resultantes para cada atributo. En la imagen 4, se muestran los nulos existentes en el dataframe original y su corrección.

```
df_jug_match.info(verbose =True,null_counts=True)
```

```
Output exceeds the size limit. Open the full output data_
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 743 entries, 0 to 742
Data columns (total 46 columns):
 #   Column                              Non-Null Count
---  ------                              --------------
 0   Jugador                             743 non-null
 1   Equipo                              732 non-null
 2   posicion_gen                        743 non-null
 3   Goles                               743 non-null
 4   xG                                  737 non-null
 5   Tiros                               737 non-null
 6   % Tiros a la portería               737 non-null
 7   Pases                               737 non-null
 8   % Pases logrados                    737 non-null
 9   Duelos                              737 non-null
 10  % Duelos ganados                    737 non-null
 11  Córneres                            737 non-null
 12  Tiros libres                        682 non-null
 13  Penaltis                            743 non-null
```

```
Output exceeds the size limit. Open the full output data_
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 743 entries, 0 to 742
Data columns (total 46 columns):
 #   Column                              Non-Null Count
---  ------                              --------------
 0   Jugador                             743 non-null
 1   Equipo                              743 non-null
 2   posicion_gen                        743 non-null
 3   Goles                               743 non-null
 4   xG                                  743 non-null
 5   Tiros                               743 non-null
 6   % Tiros a la portería               743 non-null
 7   Pases                               743 non-null
 8   % Pases logrados                    743 non-null
 9   Duelos                              743 non-null
 10  % Duelos ganados                    743 non-null
 11  Córneres                            743 non-null
 12  Tiros libres                        743 non-null
 13  Penaltis                            743 non-null
```

Imagen 4. Procesamiento de nulos en dataframe de jugadores.

Adicionalmente, se realizó el cálculo de algunas variables que no se encuentran de forma directa en el set de datos tal como los atributos de falla, entre otros, que pueden ser calculados a partir de campos que sí vienen de forma directa en el archivo tal como se muestra en la imagen 5.

```
df_jug_match['Tiros a la portería'] = df_jug_match['Tiros']*df_jug_match['% Tiros a la p
df_jug_match['Pases logrados'] = df_jug_match['Pases']*df_jug_match['% Pases logrados']/
df_jug_match['Duelos ganados'] = df_jug_match['Duelos']*df_jug_match['% Duelos ganados']
df_jug_match['Penaltis marcados'] = df_jug_match['Penaltis']*df_jug_match['% Penaltis ma
df_jug_match['Centros precisos'] = df_jug_match['Centros']*df_jug_match['% Centros preci
df_jug_match['Pases en profundidad completados'] = df_jug_match['Pases en profundidad/90
df_jug_match['Duelos ofensivos ganados'] = df_jug_match['Duelos ofensivos']*df_jug_match
df_jug_match['Duelos defensivos ganados'] = df_jug_match['Duelos defensivos']*df_jug_mat
df_jug_match['Duelos aéreos ganados'] = df_jug_match['Duelos aéreos']*df_jug_match['% Du
df_jug_match['Pases hacia adelante logrados'] = df_jug_match['Pases hacia adelante']*df_
df_jug_match['Pases hacia atrás logrados'] = df_jug_match['Pases hacia atrás']*df_jug_ma
df_jug_match['Pases laterales logrados'] = df_jug_match['Pases laterales']*df_jug_match[
df_jug_match['Pases largos logrados'] = df_jug_match['Pases largos']*df_jug_match['% Pas
df_jug_match['Pases en el último tercio logrados'] = df_jug_match['Pases en el último te
df_jug_match['Pases progresivos precisos'] = df_jug_match['Pases progresivos']*df_jug_ma
df_jug_match['Desmarques logrados'] = df_jug_match['Desmarques']*df_jug_match['% Desmarq
df_jug_match['Lanzamiento largo %'] = df_jug_match['Pases largos']*100/df_jug_match['Pas
df_jug_match['Pases fallidos'] = df_jug_match['Pases']-df_jug_match['Pases logrados']
df_jug_match['Duelos perdidos'] = df_jug_match['Duelos']-df_jug_match['Duelos ganados']
```

Imagen 5. Campos calculados en dataframe de jugadores.

En resumen, se obtiene un set de datos con 687 filas y 78 atributos que describen al jugador, los cuales serán analizados con el fin de conocer si todos ellos deben ser considerados en las próximas etapas del proyecto que involucran la evaluación del jugador. La limpieza y procesamiento de datos realizados para este set de datos se muestra en la figura 2 de forma gráfica.

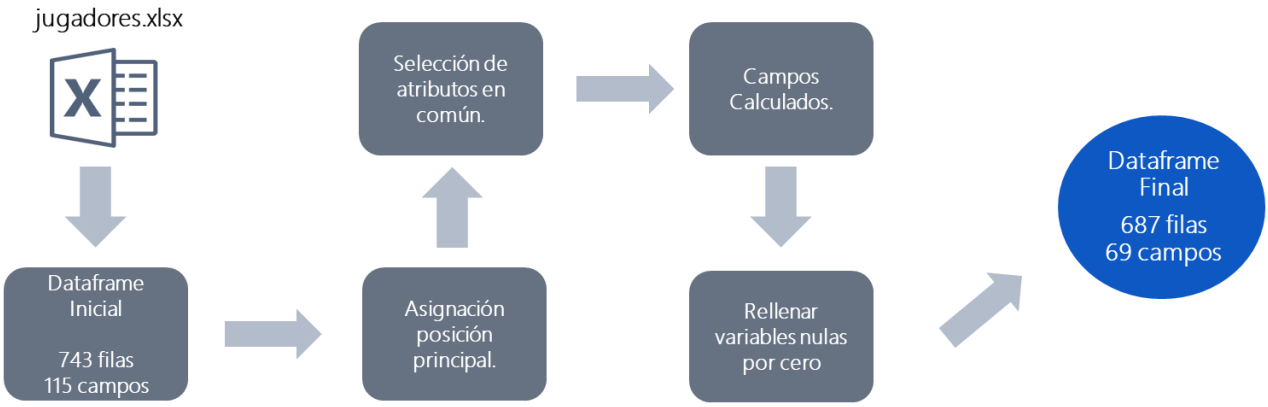


Figura 2. Fases de limpieza de datos en dataframe de jugadores.

5.1.2. Set de datos Partidos

Para el set de partidos, los datos provienen de múltiples archivos Excel extraídos de la plataforma Wyscout, los cuales contienen todos los partidos de un equipo en particular y el año según indique el nombre del archivo tal como se muestra en la imagen 6. Dentro de cada archivo se encuentran múltiples atributos que describen lo acontecido en el juego y los resultados finales.

Imagen 6. Fuente datos de partidos.

En total, se tienen 122 archivos con información de partidos entre los años 2017 y 2022, los cuales serán consolidados en un dataframe para proceder con la limpieza y procesamiento de datos. Una vez consolidados todos los archivos en un solo dataframe, en la imagen 7, se puede observar que en total se tienen 2862 partidos.

```
df_p_match.info()

Output exceeds the size limit. Open the full output data in
<class 'pandas.core.frame.DataFrame'>
Int64Index: 2862 entries, 2 to 20
Data columns (total 73 columns):
 #   Column                          Non-Null Count
---  ------
 0   Goles                           2862 non-null
 1   xG                              2862 non-null
 2   Tiros                           2862 non-null
 3   Tiros a la portería             2862 non-null
 4   % Tiros a la portería           2862 non-null
 5   Pases                           2862 non-null
 6   Pases logrados                  2862 non-null
 7   % Pases logrados                2862 non-null
 8   Duelos                          2862 non-null
 9   Duelos ganados                  2862 non-null
 10  % Duelos ganados                2862 non-null
 11  Córneres                        2862 non-null
```

Imagen 7. Dataframe de partidos.

Al igual que el dataset de jugadores se generan los atributos de no éxito o fallas en los partidos como se muestra en la imagen 8 y validar si existen nulos en el set de datos. Una vez calculados los campos y realizada la validación, se encuentra que no existen

valores nulos en ninguno de los atributos por lo no fue necesario ejecutar acciones al respecto.

```
df_p_match['Pases fallidos'] = df_p_match['Pases']-df_p_match['Pases logrados']
df_p_match['Duelos perdidos'] = df_p_match['Duelos']-df_p_match['Duelos ganados']
df_p_match['Penaltis fallidos'] = df_p_match['Penaltis']-df_p_match['Penaltis marcados']
df_p_match['Centros sin precision'] = df_p_match['Centros']-df_p_match['Centros precisos']
df_p_match['Duelos ofensivos perdidos'] = df_p_match['Duelos ofensivos']-df_p_match['Duelos ofensivos ganados']
df_p_match['Duelos defensivos perdidos'] = df_p_match['Duelos defensivos']-df_p_match['Duelos defensivos ganados']
df_p_match['Duelos aéreos perdidos'] = df_p_match['Duelos aéreos']-df_p_match['Duelos aéreos ganados']
df_p_match['Pases hacia adelante fallidos'] = df_p_match['Pases hacia adelante']-df_p_match['Pases hacia adelante logrados
df_p_match['Pases hacia atrás fallidos'] = df_p_match['Pases hacia atrás']-df_p_match['Pases hacia atrás logrados']
df_p_match['Pases laterales fallidos'] = df_p_match['Pases laterales']-df_p_match['Pases laterales logrados']
df_p_match['Pases largos fallidos'] = df_p_match['Pases largos']-df_p_match['Pases largos logrados']
df_p_match['Pases en el último tercio fallidos'] = df_p_match['Pases en el último tercio']-df_p_match['Pases en el último
df_p_match['Pases progresivos sin precision'] = df_p_match['Pases progresivos']-df_p_match['Pases progresivos precisos']
df_p_match['Desmarques fallidos'] = df_p_match['Desmarques']-df_p_match['Desmarques logrados']
df_p_match.head()
```

Imagen 8. Campos calculados en dataframe de partidos.

Por otro lado, en el caso del set de datos de partidos es necesario construir una nueva variable que identifique si el partido fue ganado o perdido. Esto con el fin de generar la variable target del modelo que determinará la importancia o los pesos que se utilizarán en cada variable para la evaluación de cada jugador.

Debido a que la idea principal es armar un modelo de clasificación binaria, se decide descartar todos los partidos que dieron como resultado un empate. En consecuencia, el set de datos se reduce a 2091 partidos categorizados con un nuevo campo llamado 'flag_triunfo' que indicará uno (1) cuando el partido sea ganado y cero (0) cuando implique una derrota. La limpieza y procesamiento de datos realizados para este set de datos de partidos se muestra en la figura 3:

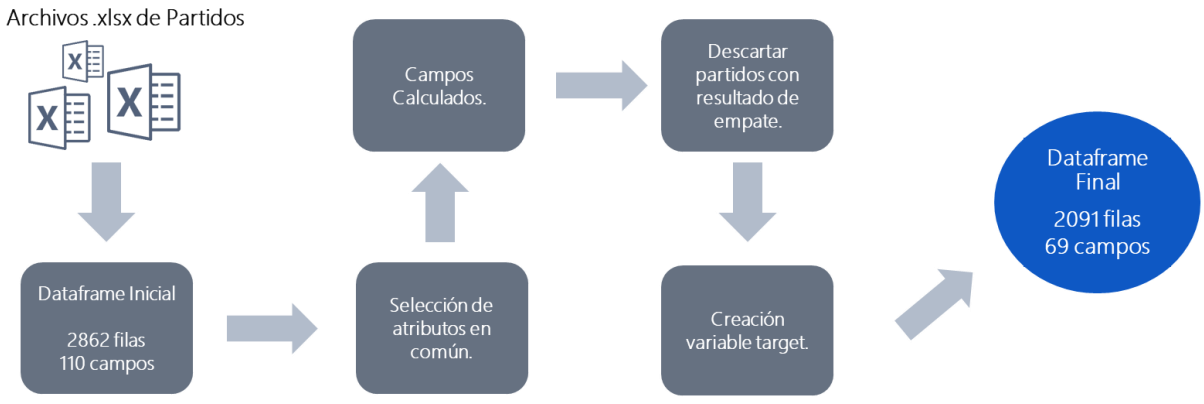


Figura 3. Fases de limpieza de datos en dataframe de jugadores.

5.2. Metodología

Como se mencionó previamente, el objetivo principal del proyecto es asignar un puntaje a los jugadores para posteriormente construir un ranking por posiciones. Este objetivo se cumplirá por medio de la implementación de un modelo de clasificación, construido con base en la ponderación de la importancia de cada feature o variable en la predicción de la victoria, luego, este peso o coeficiente será multiplicado al promedio sobre noventa minutos de cada jugador para el feature determinado y, posteriormente; se hará una sumatoria de todos estos. El ranking será calculado según el resultado de cada jugador. Es importante destacar que la mayoría de los partidos corresponden a la liga de primera división de Chile, sin embargo; otras ligas son consideradas, aunque con menor relevancia. En la figura 4, se muestra un diagrama que resume todas las etapas que contempla el proyecto desde las fuentes hasta el producto final.

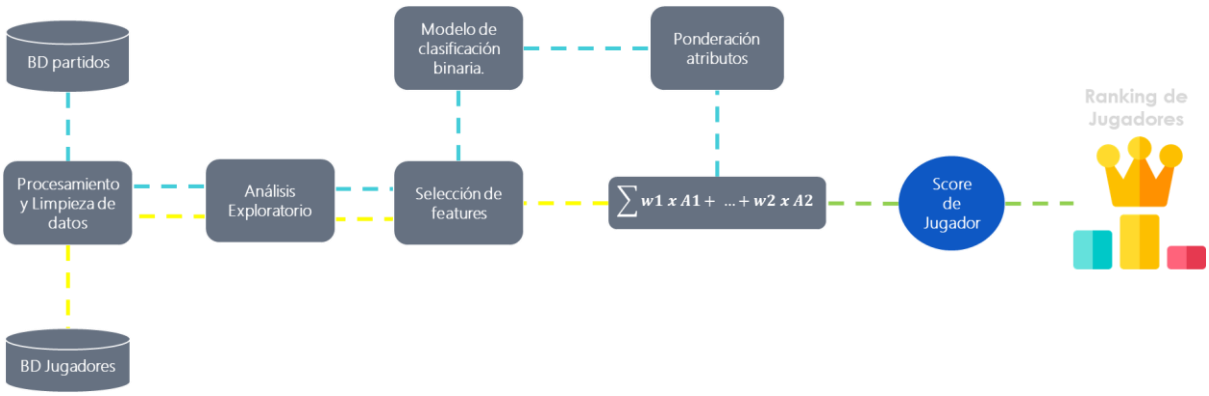


Figura 4. End to End flujo de datos del proyecto.

Para la selección de features, se analizó las variables del dataset y en muchos casos se encuentran medidas cuantitativas y porcentajes construidos con base en estas variables, por lo tanto, se seleccionan las variables base(cuantitativas) como entrada del modelo lo cual implica un total de 54 variables, obteniendo como resultado la siguiente lista de variables de entrada observadas en la figura 9.

```
features_names = ['Tiros','Tiros a la portería',
       'Pases','Pases logrados','Pases fallidos',
       'Duelos','Duelos ganados','Duelos perdidos',
       'Córneres', 'Tiros libres',
```

```
        'Penaltis','Penaltis marcados','Penaltis fallidos',
        'Centros','Centros precisos','Centros sin precision',
        'Pases en profundidad completados',
        'Toques en el área de penalti',
        'Duelos ofensivos',’Duelos ofensivos ganados',
        'Duelos ofensivos perdidos',
        'Duelos defensivos','Duelos defensivos ganados',
      'Duelos defensivos perdidos',
      'Duelos aéreos','Duelos aéreos ganados','Duelos aéreos perdidos',
      'Interceptaciones', 'Faltas',
      'Tarjetas amarillas', 'Tarjetas rojas',
      'Pases hacia adelante','Pases hacia adelante logrados',
      'Pases hacia adelante fallidos',
      'Pases hacia atrás','Pases hacia atrás logrados',
      'Pases hacia atrás fallidos',
      'Pases laterales','Pases laterales logrados',
      'Pases laterales fallidos',
      'Pases largos', 'Pases largos logrados',
      'Pases largos fallidos'
      'Pases en el último tercio',
      'Pases en el último tercio logrados',
      'Pases en el último tercio fallidos',
      'Pases progresivos',
      'Pases progresivos precisos',
      'Pases progresivos sin precision',
      'Desmarques','Desmarques logrados','Desmarques fallidos',
      'Lanzamiento largo %', 'Longitud media pases']
```

Imagen 9. Características seleccionadas para entrada del modelo

Una vez definidas las variables de entrada del modelo, se procedió a estandarizar dichas variables debido a que algunas de ellas tienen grandes diferencias entre sus rangos de valores, esto podría provocar que el modelo le de mayor importancia a ciertas variables que a otras que a su vez, podrían tener mucho más valor. Basado en trabajos anteriores, inicialmente se seleccionó utilizar un escalador estándar tradicional (Standard Scaler de la librería Scikit-learn), sin embargo; al investigar sobre este punto se encontró que dicho escalador podría ser susceptible a outliers.

Observando nuestras variables en el análisis exploratorio realizado existen outliers en todas las variables, por lo que se procedió a utilizar un escalador robusto (Robust

Scaler), el cual opera en base a rangos de cuantiles, por defecto el rango intercuartil (IQR), solucionando que el proceso de estandarización se vea afectada por outliers.

5.2.1 Herramientas de clasificación a utilizar

Para crear el modelo de predicción se usaron dos métodos ampliamente difundidos, tanto en la estadística analítica como en la ciencia de datos: Support Vector Machine SVM (Soporte de Máquinas Vectoriales) y la Regresión Logística. El primero de estos, es un modelo de aprendizaje supervisado que se asocia a algoritmos que analizan los datos para implementar un sistema de clasificación y de análisis de regresión. Creado por Vladimir Vapnik en los años noventa en los laboratorios de AT&T y se fundamenta como un clasificador binario lineal no probabilístico, es considerado uno de los métodos más robustos [9]. Por otro lado, la regresión logística es una herramienta estadística que hace un análisis de regresión para predecir el resultado de un variable categórica en función de otras variables predictoras o independientes [10][11]. La finalidad de usar dos modelos es hacer comparaciones entre estos, según los resultados obtenidos en cada uno. De esta manera se podrá tener mayor riqueza de información para la construcción de un mejor análisis final.

5.2.2 Métricas de evaluación

Para validar el performance de los modelos a desarrollar, se considerarán las siguientes métricas:

- **Exactitud o Accuracy:** Corresponde al porcentaje de observaciones que fueron predichas correctamente. No se recomienda su uso en datos desbalanceados ya que el modelo podría ser muy bueno prediciendo una de las 2 clases mas no ambas lo cual no es lo idóneo. Por esta razón, se analizará en etapas futuras si corresponde ocupar esta métrica en el estudio.
- **Precisión o Precision:** Corresponde al porcentaje de observaciones positivas que fueron clasificadas correctamente. Es decir, la fracción de predicciones positivas verdaderas de todas las predicciones positivas.

Muestra si el clasificador es capaz de diferenciar una clase de todas las demás

- **Exhaustividad o Recall:** Corresponde a la medida que cuantifica como el modelo identifica los verdaderos positivos. Muestra si el clasificador es capaz de detectar una clase dada en absoluto.

6. Resultados

6.1. Análisis Exploratorio

6.1.1. Set de datos jugadores

Uno de los campos más relevantes del set de datos de jugadores corresponde al rol que ocupa en la cancha, validando en que posición principal se encuentra cada jugador. En la figura 5, se puede observar que la mayor parte de los jugadores son defensas seguidos de delanteros y la posición que menos jugadores tiene corresponde a medio campo defensivo y extremos. Por lo tanto, será más difícil destacar en posiciones con mayor número de jugadores como en el caso de los jugadores con rol de defensa, que en los roles que existen menos jugadores para evaluar.

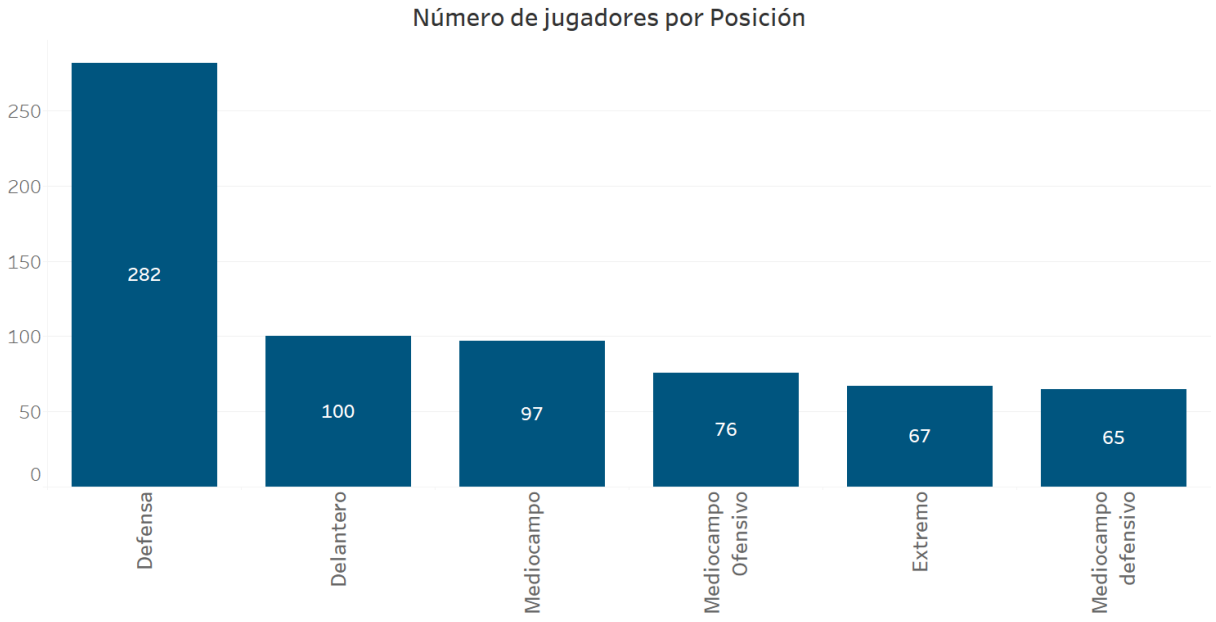


Figura 5. Distribución de posición principal en set de datos de jugadores.

Otro aspecto importante que destacar dentro del dataset de jugadores es la variable de “Minutos jugados”, esto debido a que esta variable puede dar una referencia de si el jugador es un jugador activo, es titular o ha jugado muy pocos minutos durante el año. La intención principal de considerar esta variable al momento de la evaluación es poder obtener un ranking de jugadores con las mismas características y evitar que jugadores que jueguen muy poco y tengan buenas estadísticas, se encuentren por encima de jugadores que juegan toda la temporada.

Con base en esta variable de Minutos jugados, se creó una nueva variable llamada Partidos jugados aproximados la cual corresponde a los minutos divididos entre 90 minutos de un partido, esto con el fin de categorizar a los jugadores por partidos jugados conociendo que las competiciones de primera división tienen alrededor de 30 fechas y de esta forma comparar jugadores que jueguen cantidad de partidos similares.

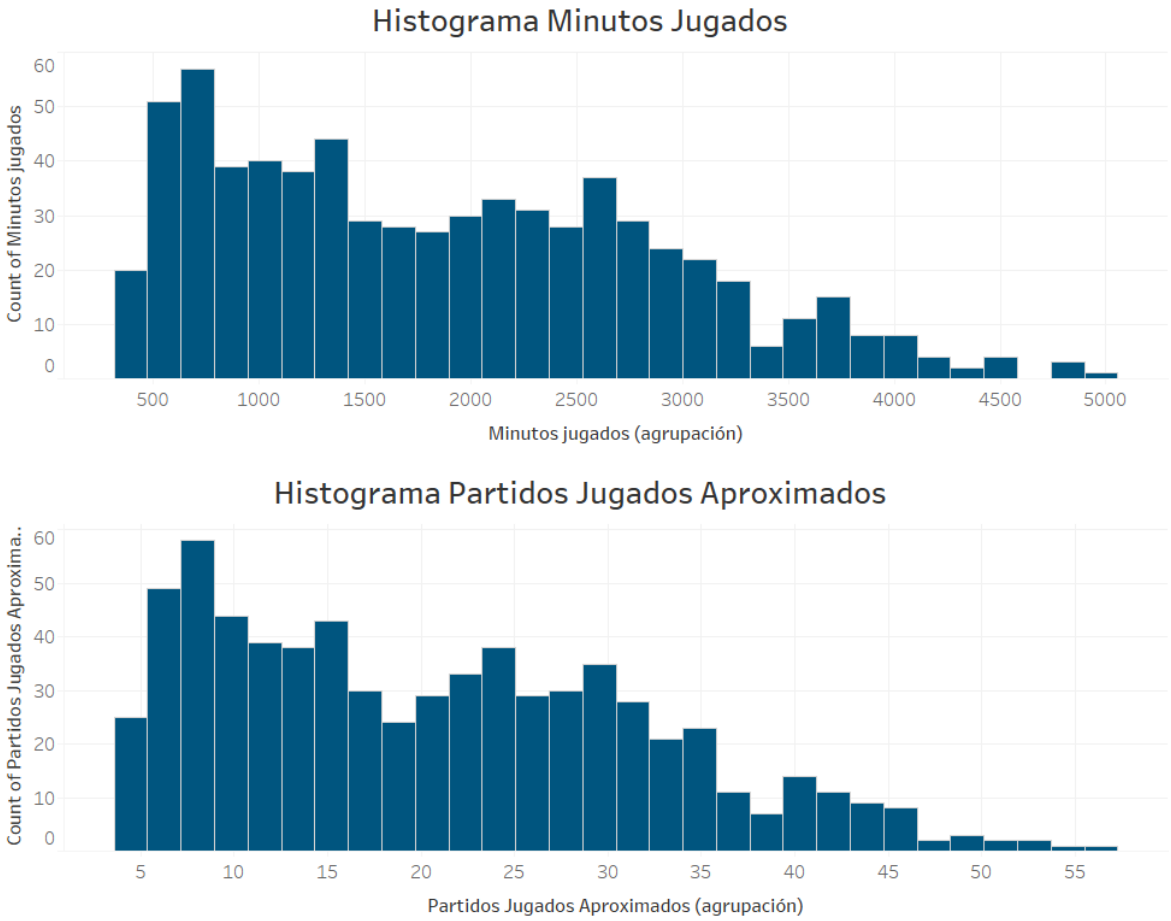


Figura 6. Histogramas de minutos y partidos aproximados jugados.

Según los histogramas de la figura 6, se definen 3 categorías para comparar a los jugadores en la etapa de evaluación:

- Jugadores con menos de 15 partidos jugados.
- Jugadores entre 15 y 30 partidos jugados.
- Jugadores con más de 30 partidos jugados.

En la figura 7 se puede observar cómo se distribuyen los jugadores por Partidos aproximados jugados, teniendo el mayor número de jugadores en el tramo de 15 a 30 partidos y el menor número de ellos se encuentran en el tramo de más de 30 partidos, que podrían ser jugadores que son titulares en su liga actual y participan en otras competiciones adicionales como las convocatorias de los partidos de la selección nacional.

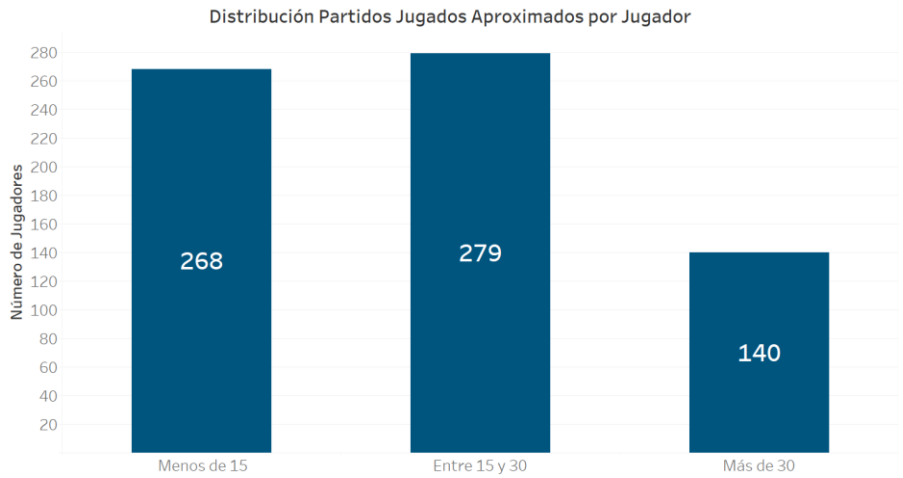


Figura 7. Distribución de Jugadores por intervalos de partidos aproximados.

6.1.2. Set de datos partidos

Tal como indica la etapa de limpieza y procesamiento de datos, el dataset de partidos está compuesto por 2091 partidos, la mayoría de ellos corresponde a partidos de la primera división A de Chile, alrededor de 75% de los partidos corresponden a esta competencia tal como se observa en la figura 8, lo cual implica que los jugadores serán evaluados mayormente bajo los estándares o estadísticas de la misma.

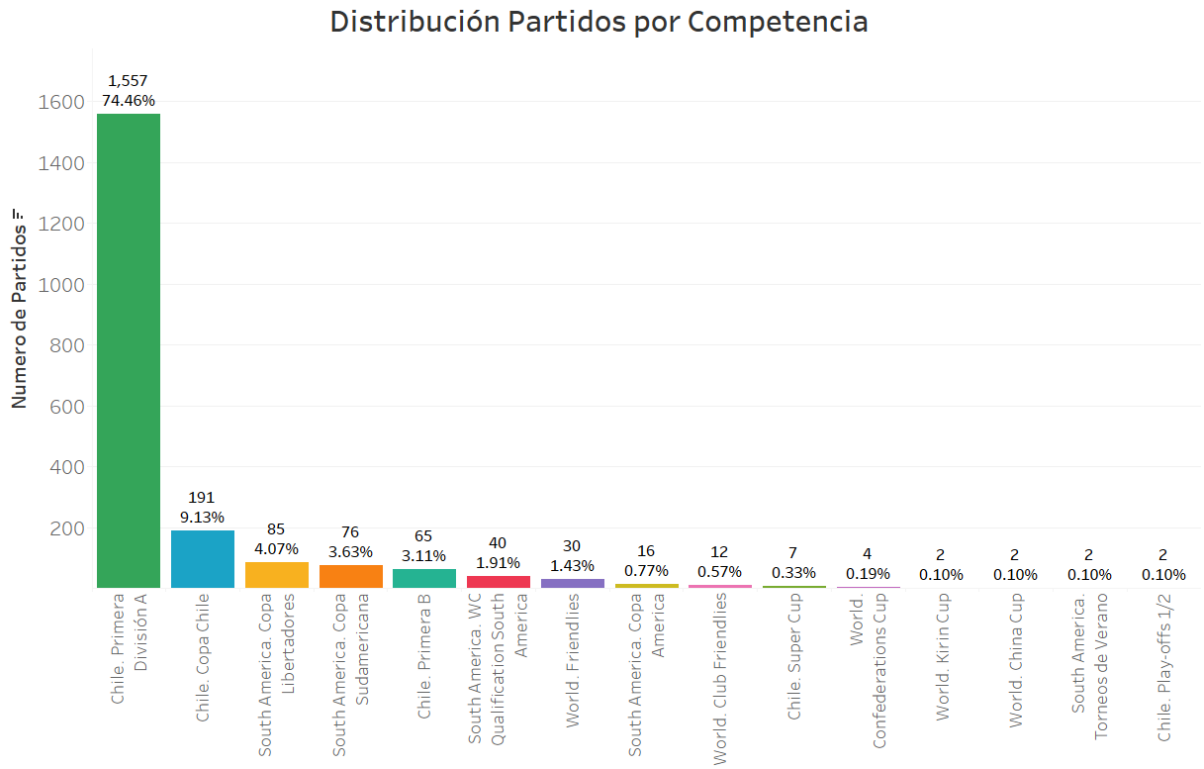


Figura 8. Distribución de partidos por competencia.

Otro aspecto importante del dataset de partidos, es conocer cómo están distribuidos los triunfos y derrotas ya que esto indicará si se requiere realizar algún procesamiento adicional en la etapa de entrenamiento del modelo en caso de que no exista una cantidad de registros balanceada. Al observar la figura 9, se puede evidenciar que existe una muestra de partidos balanceada ya que alrededor de 50% de los datos representa una u otra opción.

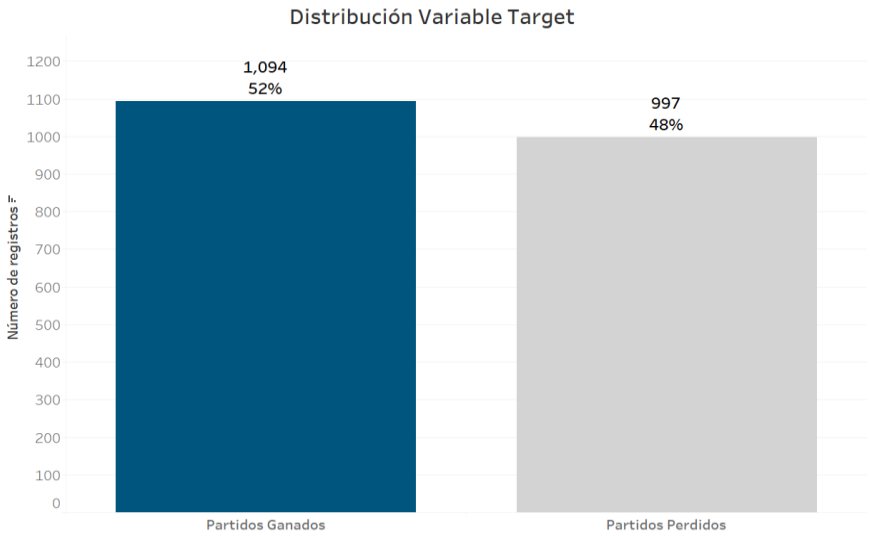


Figura 9. Distribución variable target.

Analizando las variables que tenemos disponibles y cómo se comportan de acuerdo con la variable 'flag triunfo', se encontró que existen variables con distribuciones muy similares entre sí y otras variables donde se puede observar una diferencia entre ganar o perder un partido. En la figura 10, se muestra lo mencionado anteriormente.

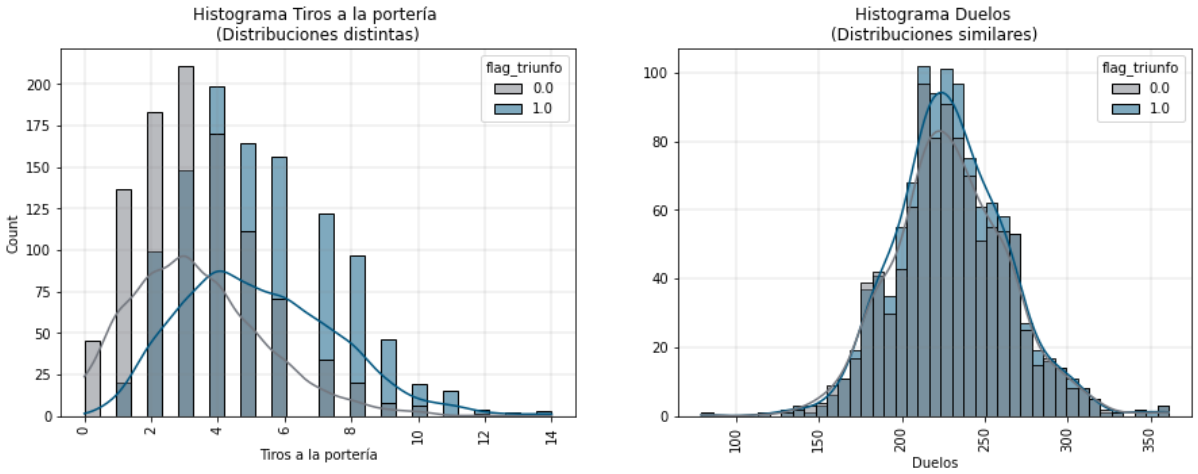


Figura 10. Variables con distribución distinta y distribución similar

Dada esta situación, se procedió a realizar una prueba de Kolmogorov-Smirnov para verificar si las distribuciones de las variables son iguales o distintas entre sí con el fin de analizar las variables que sí tengan una diferencia en la distribución, la hipótesis nula de esta prueba estadística indicará que las distribuciones son iguales, considerando un nivel de significancia de 0.005. Realizando este test estadístico se obtuvieron que las siguientes variables, rechazan la hipótesis nula, es decir, presentan una distribución distinta en el caso de un escenario de triunfo o derrota, por ende, se esperaría que alguna de ellas tenga impacto en el modelo de clasificación binaria que se construirá en las siguientes etapas:

1. Tiros
2. Tiros a la portería
3. % Tiros a la portería
4. Tiros libres
5. Pases
6. Pases logrados
7. Pases fallidos
8. % Duelos ganados

9. Penaltis
10. Penaltis marcados
11. % Penaltis marcados
12. Centros
13. Centros precisos
14. Centros sin precisión
15. Pases en profundidad completados
16. Toques en el área de penalti
17. % Duelos aéreos ganados
18. Interceptaciones
19. Faltas
20. Tarjetas rojas
21. Pases hacia adelante
22. Pases hacia adelante logrados
23. Pases laterales
24. Pases laterales logrados
25. Pases laterales fallidos
26. Pases en el último tercio
27. Pases en el último tercio logrados
28. Pases en el último tercio fallidos
29. Pases progresivos precisos
30. % Pases progresivos precisos
31. Desmarques
32. Desmarques logrados
33. % Desmarques logrados

Según el listado anterior, se analizan las distribuciones con el fin de observar cómo se comporta cada variable de acuerdo con un resultado de triunfo o derrota. En el caso de las variables relacionadas con los tiros observadas en la figura 11, se observa que para los tiros totales y los tiros a la portería los diagramas de caja tienden a estar desplazados hacia arriba en el caso de triunfo. Esto puede indicar que en un partido el equipo ganador ataca con mayor frecuencia, en particular en los tiros que son hacia la portería, la visualización de esta variable muestra que en el diagrama de caja del ganador tiene mayor

cantidad y los bigotes se desplazan hacia arriba. Por otro lado, en la variable tiros libres ocurre lo contrario, en los partidos perdidos los equipos parecen tener una mayor cantidad de tiros libres.

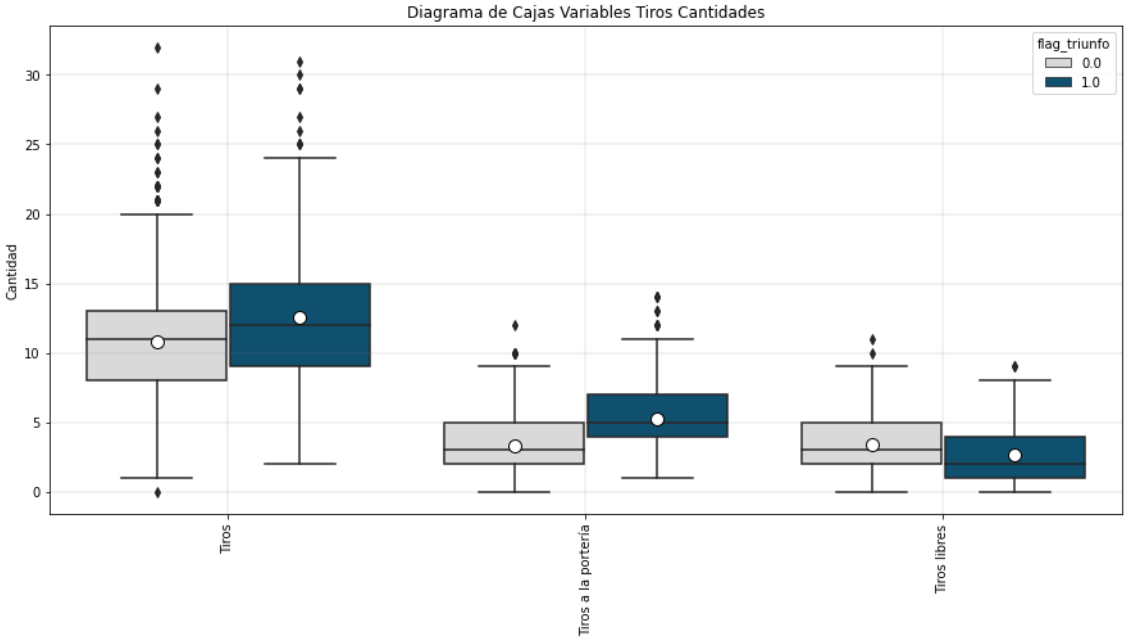


Figura 11. Diagrama de cajas variables Tiros.

En algunas variables con distribución que favorecen a los equipos vencidos, como es el caso de los pases ganados, donde el diagrama está desplazado hacia arriba tal como se muestra en la figura 12, esto podría sugerir que los equipos derrotados suelen tener mayor número de pases que los ganadores, sin embargo, parece ser no significativa ya que las cantidades parecen ser similares, sin embargo; se requieren de pruebas estadísticas para cuantificar si la diferencia es significativa.

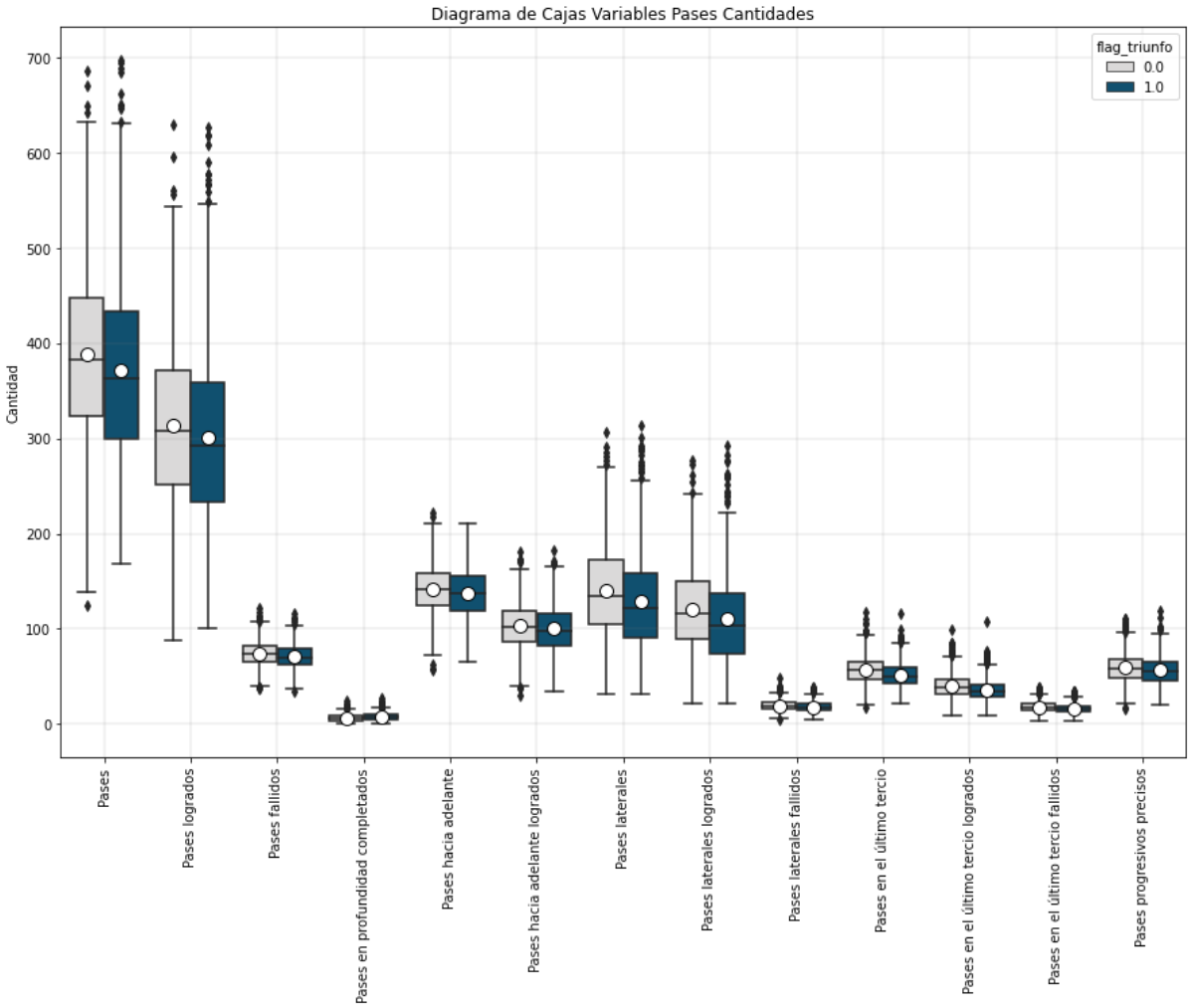


Figura 12. Diagrama de cajas variables Pases.

Observando las variables con medida porcentual con distribución distinta representadas en la figura 13, se puede apreciar que para variables como % de Duelos ganados, % de Duelos aéreos ganados, % Desmarques logrados y % Tiros a la portería tienden a ser mayor cuando se trata de la variable triunfo, lo cual resulta bastante lógico de acuerdo con la naturaleza de estas variables. Es importante destacar que en el caso de la variable Tiros a la portería, las distribuciones se separan notoriamente tal como se observó con la misma variable desde un punto de vista cuantitativo. En el caso de % de pases progresivos precisos también se observa el mismo comportamiento que las variables de pases que se analizaron por cantidades, la distribución de un escenario de derrota suele estar desplazada hacia arriba con respecto al resultado de triunfo, sin embargo, las medias se encuentran muy cerca.

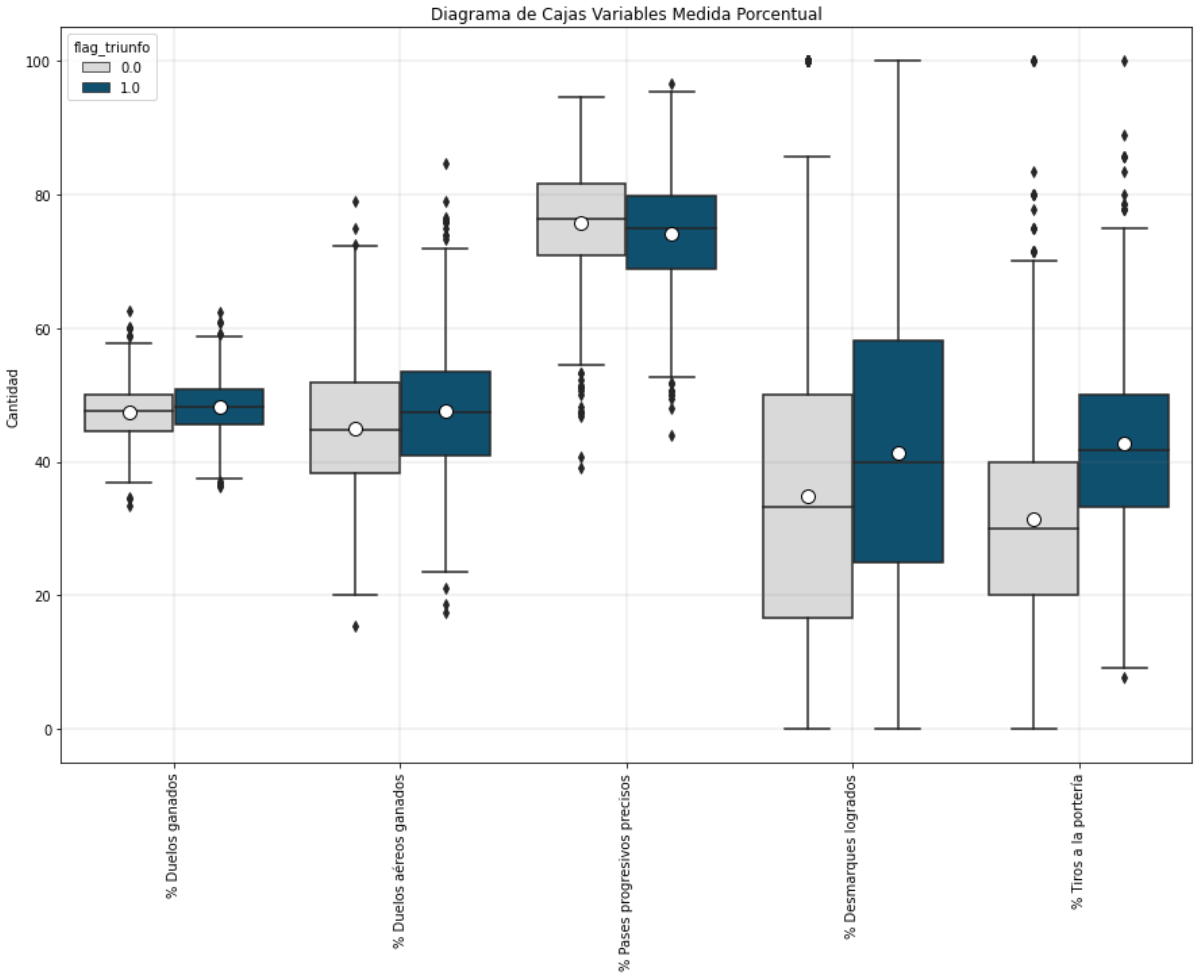


Figura 13. Diagrama de cajas variables porcentuales.

Otras variables que obtuvieron como resultado una distribución distinta para ambos escenarios, son los penaltis, centros, toques en el área de penalti, interceptaciones, faltas, tarjetas rojas y desmarques observados en la figura 14. En cuanto a las variables relacionadas con centros, se observa que la distribución tiende a un mayor número cuando se trata de un escenario de derrota tanto para centros precisos como no precisos, ocurre lo contrario para las demás variables como los toques en el área de penalti donde en un escenario de triunfo este valor suele ser mayor, al igual que las interceptaciones donde se observa la media entre distribuciones más separada con respecto a otras variables.

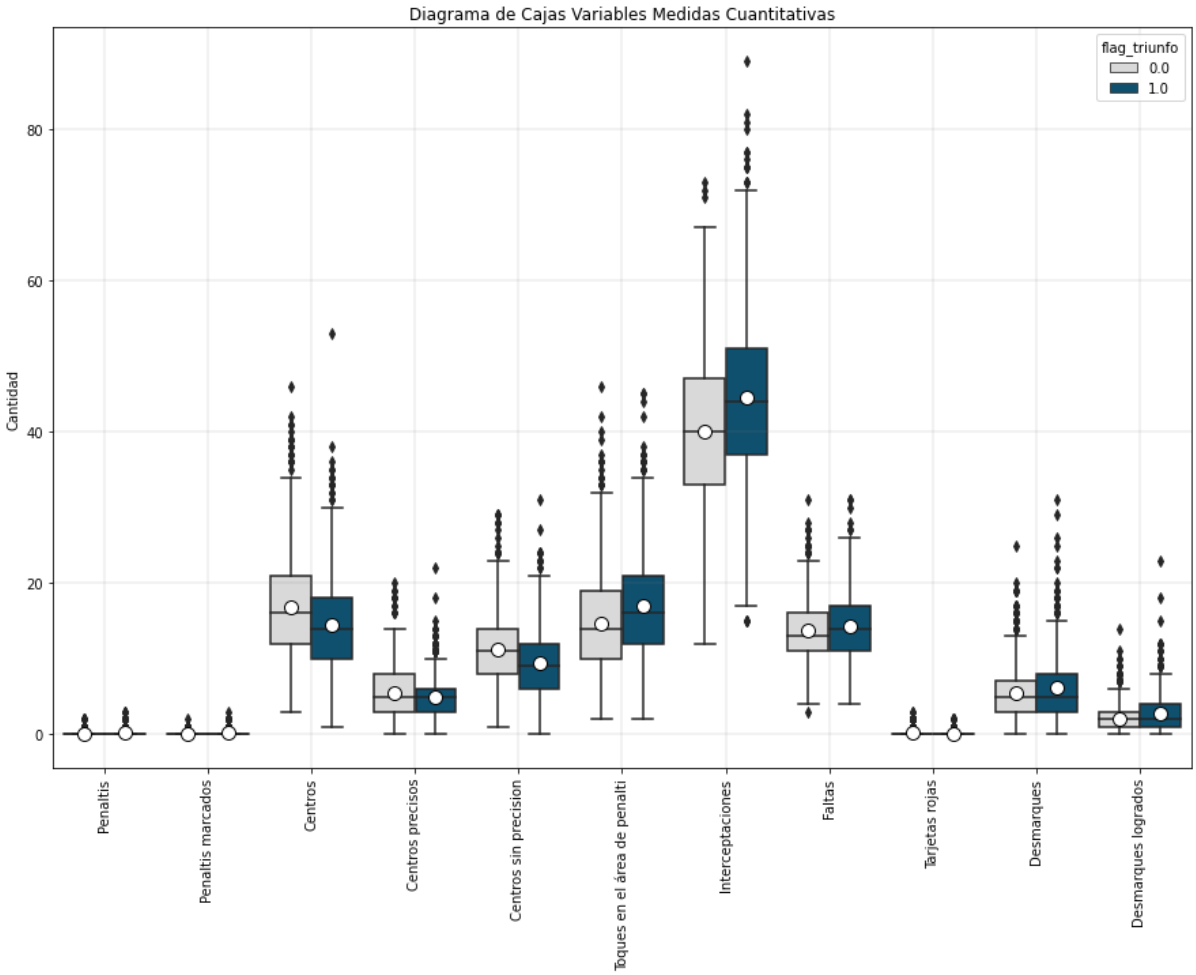


Figura 14. Diagrama de cajas, otras variables cuantitativas.

6.2. Modelos

6.2.1. Modelo SVM

El primero de los modelos fue construido con Support Vector Machine como herramienta estadística, previo a optimización de parámetros el modelo arrojó un accuracy de 0,79, debido a que en etapas previas se observó que los datos se encuentran balanceados por lo que es posible tomar en consideración este valor el cual indica que en 79% de los casos el modelo predijo correctamente.

Por otro lado, la precisión y la exhaustividad igualmente resultaron sobre el 75%. En el caso de la precisión, se puede observar que de todas las oportunidades que el

modelo categoriza a un equipo como ganador, lo hará correctamente en un 78% de los casos, y cuando hace lo propio con la derrota, lo hará en un 80% de los casos.

Observando Exhaustividad, o Recall, se obtuvo que de todas las victorias o derrotas, el modelo identifica correctamente el 76% de las victorias y el 82% de las derrotas. Es importante tener clara la distinción de estos indicadores para así entender mejor la capacidad predictiva del modelo. Las métricas mencionadas previamente pueden ser observadas en la imagen 10.

	precision	recall	f1-score	support
0.0	0.80	0.76	0.78	308
1.0	0.78	0.81	0.80	320
accuracy			0.79	628
macro avg	0.79	0.79	0.79	628
weighted avg	0.79	0.79	0.79	628

Imagen 10. Métricas modelo SVM

El modelo de SVM, previo a optimización arrojó que los primeros tres feature más importantes para alcanzar la victoria tal como se muestra en la figura 15, los cuales corresponden a: Tiros a la portería, pases logrados y pases en general. Al contrario, las variables con menor coeficiente fueron: Tarjetas Rojas, pases laterales logrados y pases laterales en general. Estos parecen ser resultados razonables, pues un equipo que más ofensivo y que logre mayor cantidad de pases parece tener mayores chances de lograr la victoria, por otra parte, los pases laterales y su valor negativo puede estar asociados a que son movimientos defensivos, hechos para prevenir la anotación del equipo contrario.

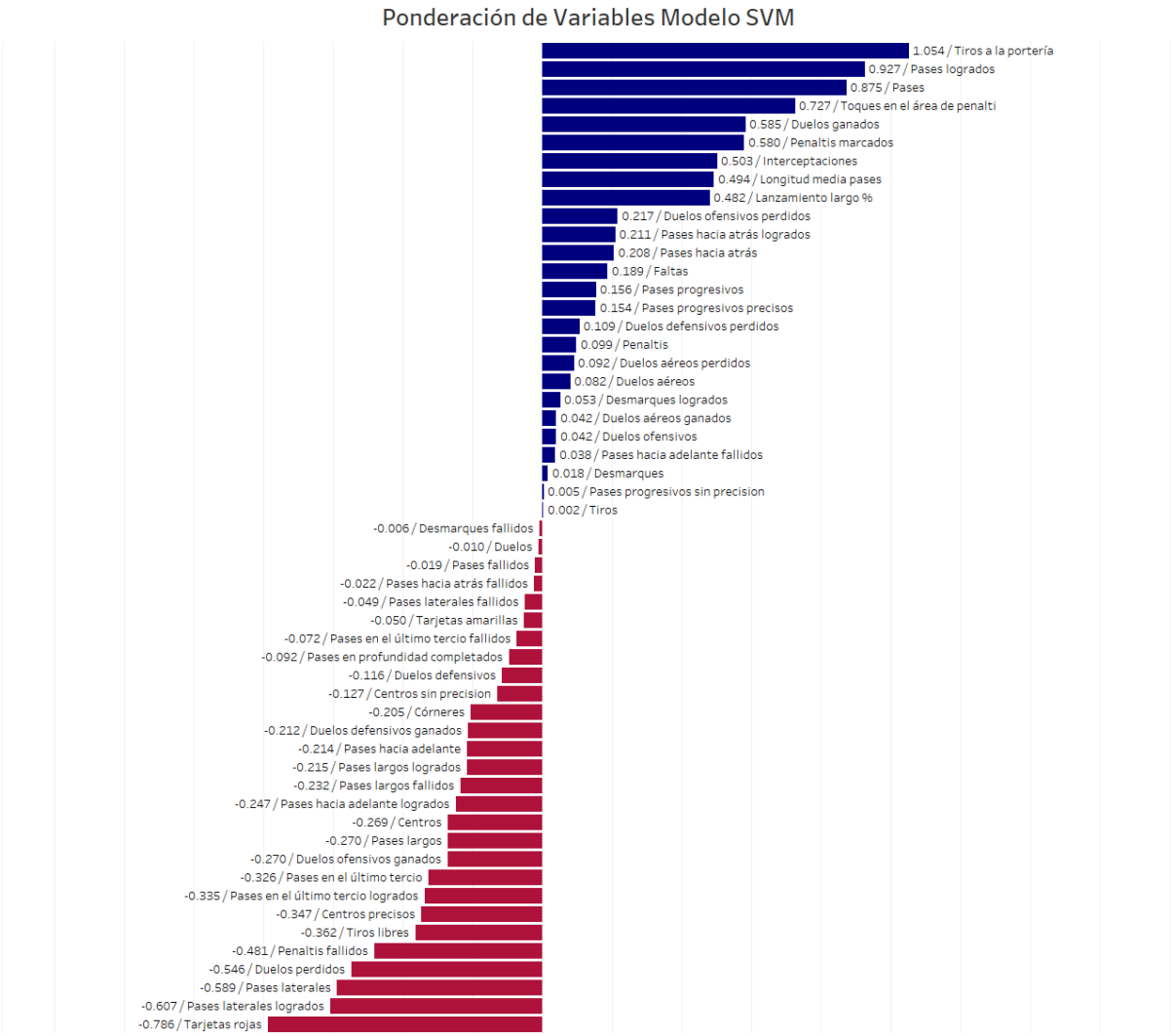


Figura 15. Ponderación realizada por Modelo SVM.

Posteriormente, se optimiza el modelo de SVM con la función GridSearch de la librería sklearn. En este, se obtienen valores similares de Accuracy y Precisión, el único parámetro que cambió fue la exhaustividad, o recall, de los valores predichos como victoria tal como se observa en la imagen 11. De todas las victorias, este modelo es capaz de identificar el 82% versus el 81% en su par sin optimización. Esta diferencia probablemente no sea significativa, por lo tanto, ambos ejemplos pueden ser usados para la construcción de un sistema de predicción eficaz.

	precision	recall	f1-score	support
0.0	0.80	0.76	0.78	308
1.0	0.78	0.82	0.80	320
accuracy			0.79	628
macro avg	0.79	0.79	0.79	628
weighted avg	0.79	0.79	0.79	628

Imagen 11. Métricas modelo SVM optimizado.

6.2.2. Regresión Logística

El segundo de los modelos implementados se trató de una regresión logística que también arrojó resultados similares. En líneas generales todas las métricas arrojaron valores superiores al 75% al igual que el modelo basado en Support Vector Machine, sin embargo, resultaron valores menores con respecto a este modelo tal como se puede observar en la imagen 12 donde algunas métricas disminuyeron entre 1 y 2 puntos porcentuales.

	precision	recall	f1-score	support
0.0	0.79	0.76	0.77	308
1.0	0.77	0.80	0.79	320
accuracy			0.78	628
macro avg	0.78	0.78	0.78	628
weighted avg	0.78	0.78	0.78	628

Imagen 12. Métricas modelo Regresión Logística.

En el caso de las ponderaciones para este modelo, se encontró entre las más importantes, los tiros a la portería, pases logrados y toques en el área de penalti, lo cual coincide con dos de las variables que se obtuvieron con el modelo SVM, sin embargo; los pases en este modelo fueron desplazados al cuarto lugar siendo más importante los toques en el área de penalti. Por otro lado, entre las variables con mayor valor negativo se encuentran las tarjetas rojas, pases laterales logrados y pases laterales en general coincidiendo con los resultados del modelo SVM. Adicionalmente, se puede observar que la regresión logística otorga una mayor magnitud a la ponderación de las variables. El

detalle de las ponderaciones obtenidas por este modelo puede ser visualizado en la figura 16.

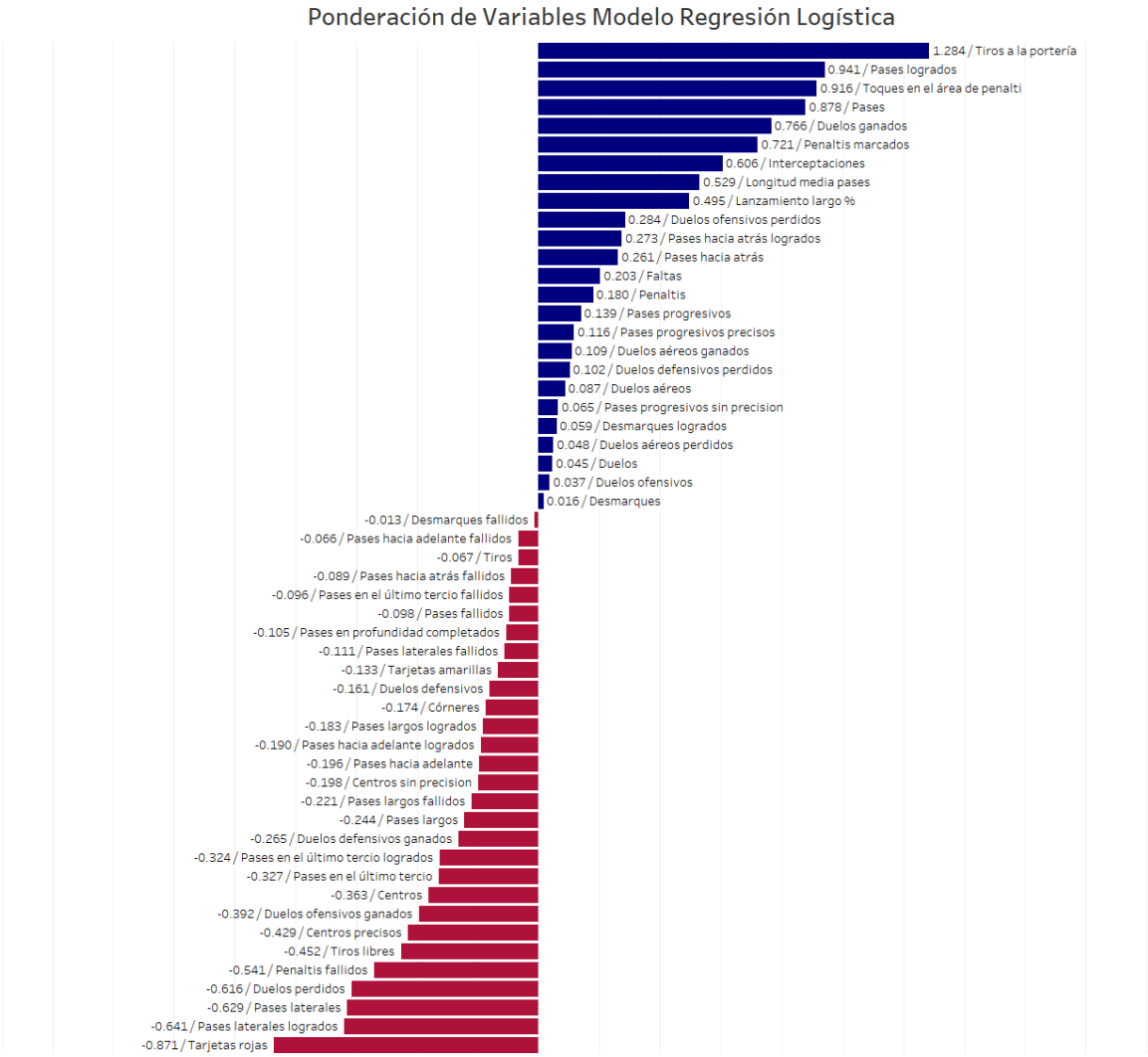


Figura 16. Ponderación realizada por Modelo de Regresión Logística.

Al igual que para el modelo de SVM, se procedió a realizar una búsqueda de hiperparametros para optimizar el rendimiento del modelo. En líneas generales, las métricas nuevamente no varían mucho tal como se observa en la imagen 13, se mantiene un accuracy de 78%, se mejora la precisión para el resultado de triunfo y la exhaustividad en el caso de las derrotas, sin embargo; se obtiene un modelo con métricas muy similares al original.

	precision	recall	f1-score	support
0.0	0.77	0.78	0.78	308
1.0	0.79	0.78	0.78	320
accuracy			0.78	628
macro avg	0.78	0.78	0.78	628
weighted avg	0.78	0.78	0.78	628

Imagen 13. Métricas modelo Regresión Logística Optimizada.

En los modelos realizados con regresión, tanto con optimización como sin este recurso, se aprecia que los valores de accuracy, precisión y recall (exhaustividad) fueron similares. Por lo tanto, ambos casos son buenos para establecer el modelo de predicción de victorias y derrotas, esto demuestra que, en este caso, optimizar el modelo no cambia los resultados de manera significativa. Igualmente, como se aprecia en las visualizaciones de los features con sus respectivos coeficientes o pesos, el top 3 de de ambos valores, tanto de la punta como de la cola, son similares. Estos coinciden con el modelo previo hecho con SVM. En conclusión, ambas técnicas tienen resultados similares ya que tantos los pesos como el orden de los features resultantes son parecidos.

6.3. Ranking de Jugadores

6.3.1. Visión General

Para visualizar los puntajes obtenidos para cada jugador se utilizarán los pesos obtenidos del modelo SVM, ya que, aunque presentó métricas muy similares, fueron superiores a la regresión logística. Adicionalmente para obtener valores del 0 al 1 se ocupará una estandarización del score obtenido por medio de la técnica Min Max Scaler.

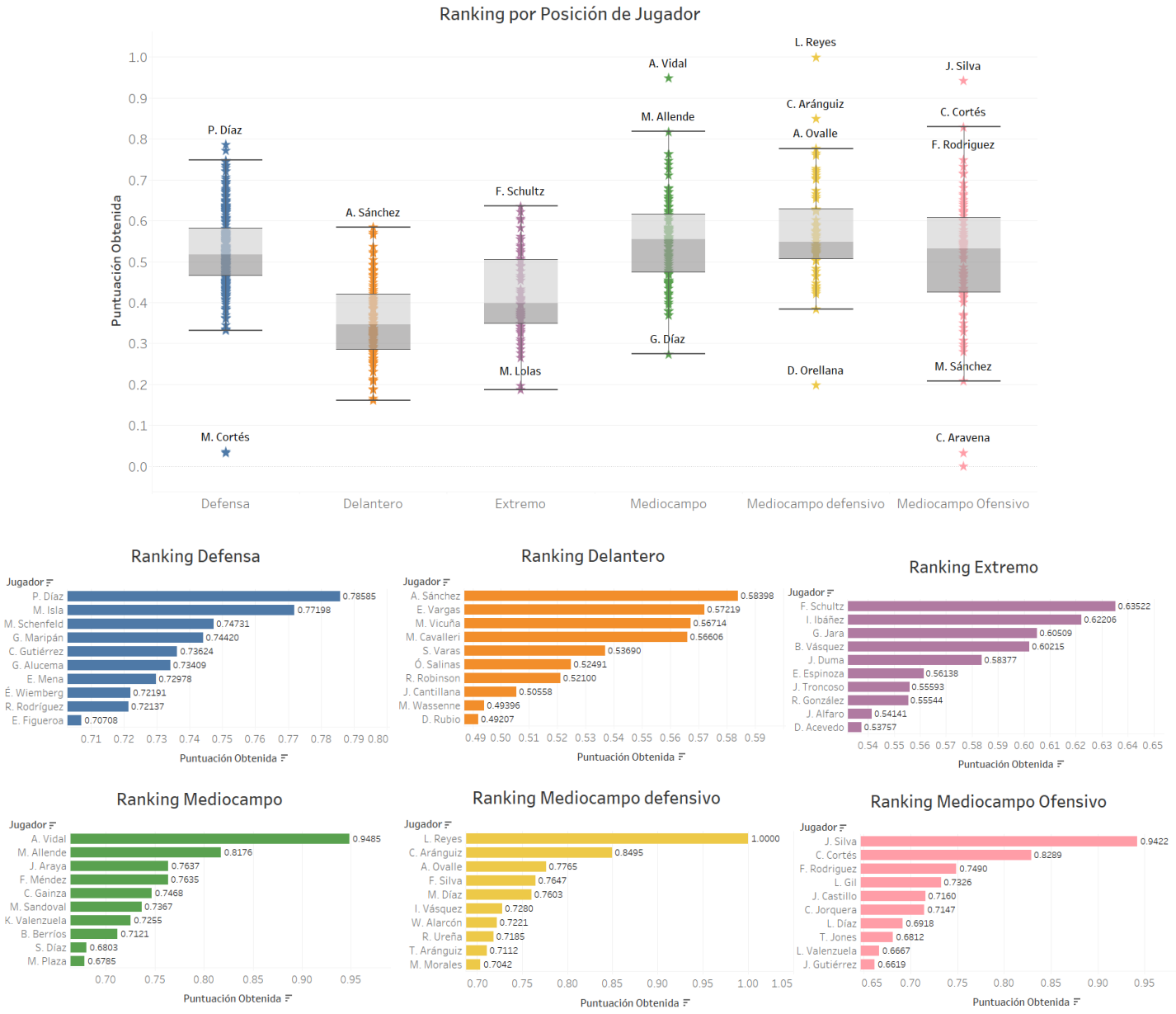


Figura 17. Ranking de Jugadores por posición.

Esta primera gráfica (figura 17), que corresponde a un diagrama de cajas con sus respectivos bigotes y outliers, demuestra un ranking de jugadores por cada una de las principales posiciones. Igualmente, se evidencia en todos los diagramas que existen muy pocos outliers, los cuales a simple vista podrían tratarse de jugadores diferenciales en rendimiento cuando se trata de una puntuación muy alta o jugadores que poseen muy pocas estadísticas debido a pocos partidos jugados en el caso de las puntuaciones muy bajas. En la posición defensa, por ejemplo, se tiene que el jugador Paulo Díaz es un outlier superior y M. Cortés uno inferior. En la posición de delantero resultó el popular Alexis Sánchez como outlier superior. Igualmente, algunos jugadores ex seleccionados nacionales como Arturo Vidal o el mediocampista de la Bundesliga, Charles Aránguiz, aparecen como outliers en sus respectivas posiciones. Cada una de estas posiciones será discutida con detalle a continuación.

6.3.2. Análisis por posición

6.3.2.1. Posición Defensa

Al estudiar los datos por tiempo jugado, como se evidencia en la figura 18, en la posición defensa se evidencia como outlier un nombre conocido, el destacado defensa chileno Paulo Díaz, estrella del River Plate. Según este resultado, Diaz es capaz de mantener un nivel altísimo como defensa por más de 30 partidos en la primera división argentina, algunos medios lo han posicionado como una de las estrellas de la "era Gallardo", período de 8 años en el cual el equipo bonaerense obtuvo 14 títulos, entre ellos la Copa de Libertadores [12]. Entre los 15 a 30 partidos destaca el defensa Maximiliano Schenfeld, un jugador sólido en la liga estadounidense [13]. Por último, Gino Alucemam del recién ascendido Magallanes, y con una amplia trayectoria pasada en el Everton, se alza como un defensa importante en menos de 15 partidos jugados.

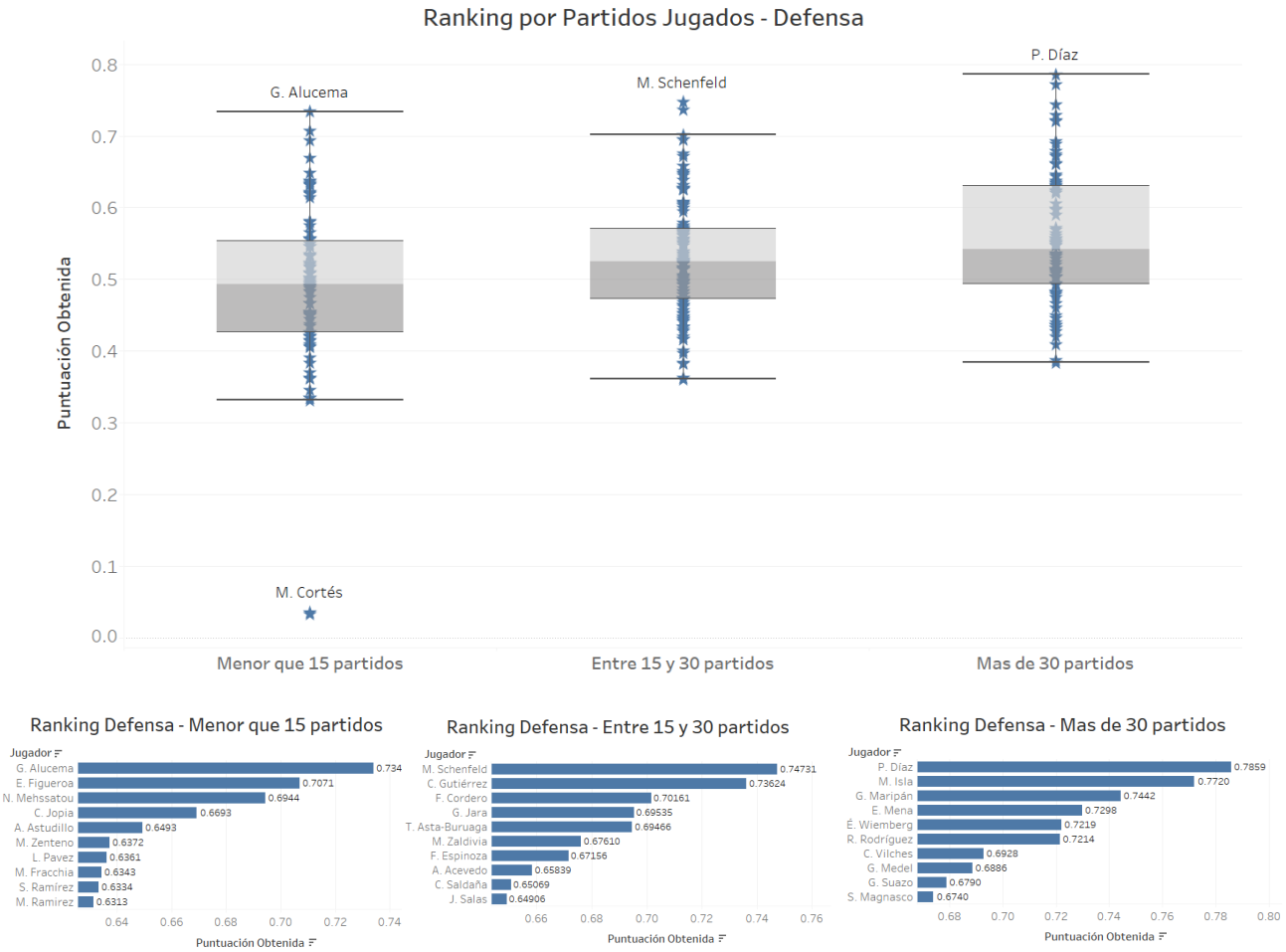


Figura 18. Ranking de Defensa por partidos aproximados jugados.

6.3.2.2. Posición Delantero

En la posición de delantero el ranking obtenido muestra figuras conocidas no solo en Chile, sino también en el fútbol regional y mundial (ver figura 19). Entre los jugadores como más de 30 partidos disputados destaca el oriundo de Renca Eduardo Vargas, figura principal de la generación de dorada del 2015 y 2016, segundo jugador con mayor cantidad de partidos en la historia de La Roja, su segundo goleador (nueve menos que Alexis Sánchez), máximo goleador en la Historia de la Universidad de Chile en torneos internacionales y artífice de su triple corona en la década pasada; actualmente número 10 del Atlético Mineiro, equipo de la ciudad de Belo Horizonte en Brasil [14][15]. En la subcategoría entre 15 a 30 partidos destaca la estrella del fútbol chileno Alexis Sánchez, el niño maravilla de Tocopilla, máximo goleador en la historia de La Roja, exponente principal de la generación dorada que llevó a Chile conseguir el doble campeonato

continental, y el primer jugador en la historia en marcar un triplete en la Premier League, en la Liga de España y en la Serie A, en su dilatada carrera ha pasado por los más prestigiosos equipos europeos: Barcelona FC, Inter Milan, Arsenal FC, Manchester United. En Sudamérica hizo lo propio en Colo Colo y en el River Plate de Buenos Aires [16][17]. Entre los delanteros con menos de 15 partidos jugados destacan las jóvenes promesas Matías Cavalleri de Unión La Calera y Manuel Vicuña del más reciente campeón de la Copa Chile Deportes Magallanes. Otros destacados delanteros que aparecen en los rankings son Luis Jiménez, Nicolás Guerra, entre otros.

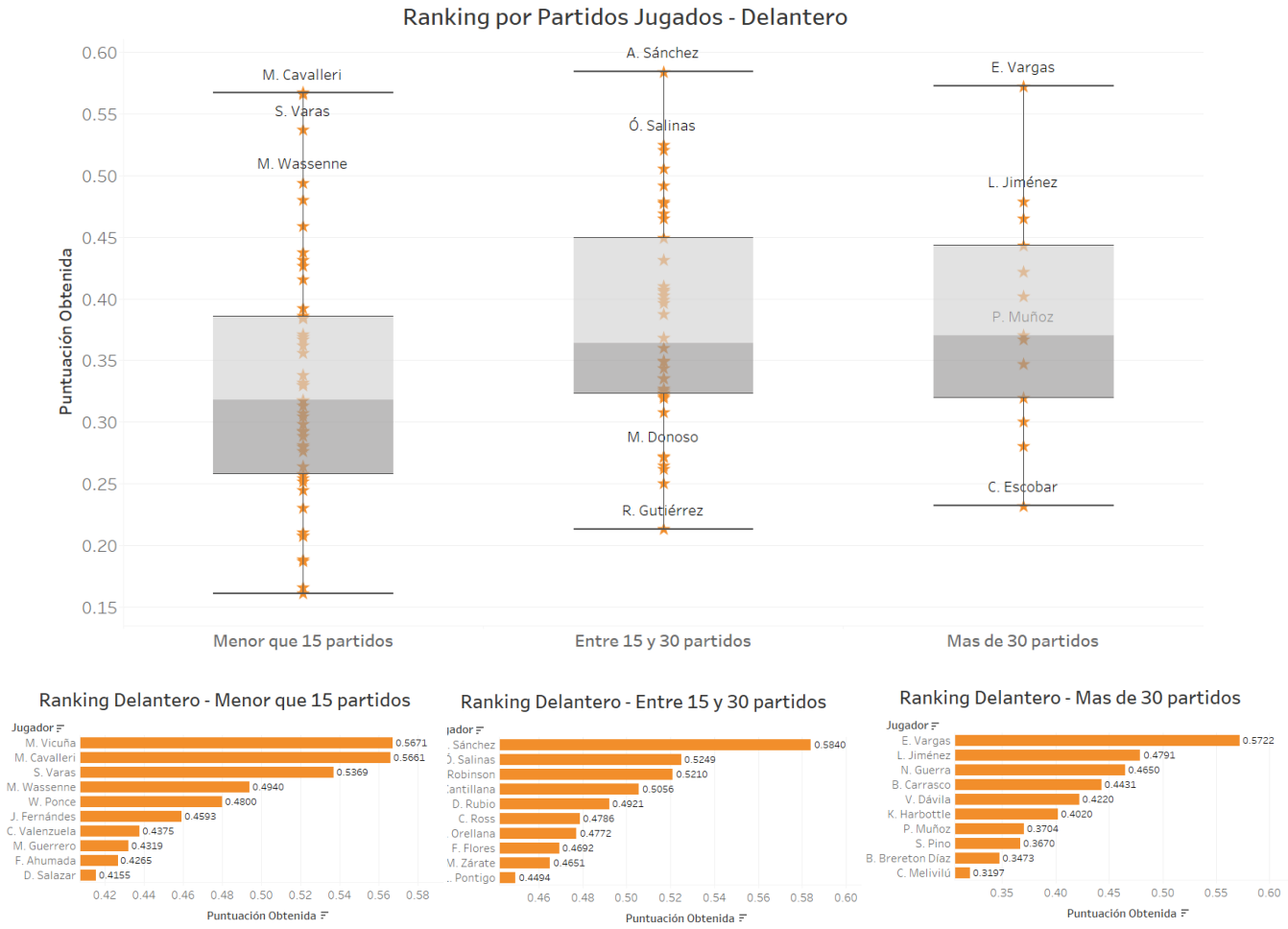


Figura 19. Ranking de Delantero por partidos aproximados Jugados.

6.3.2.3. Posición Extremo

Respecto a los extremos o laterales chilenos observados en la figura 20 destaca, con más de 30 partidos, el jugador Julián Alfaro, ex U de Chile integrante del recién ascendido Deportes Magallanes, quien ha sido catalogado como una de las principales piezas en la notable mejoría del equipo en el último año [18]. Entre los 15 y 30 partidos, el lateral Franz Schulz, del Santiago Wanderers, se ubica como el único outlier superior, este jugador ha tenido una carrera destacada en primera A y B, y es recordado por haber conformado la plantilla de jugadores de la Universidad de Chile a finales de la década pasada. Con menos de 15 partidos, destaca el joven jugador Ignacio Ibañez de la Unión Española.

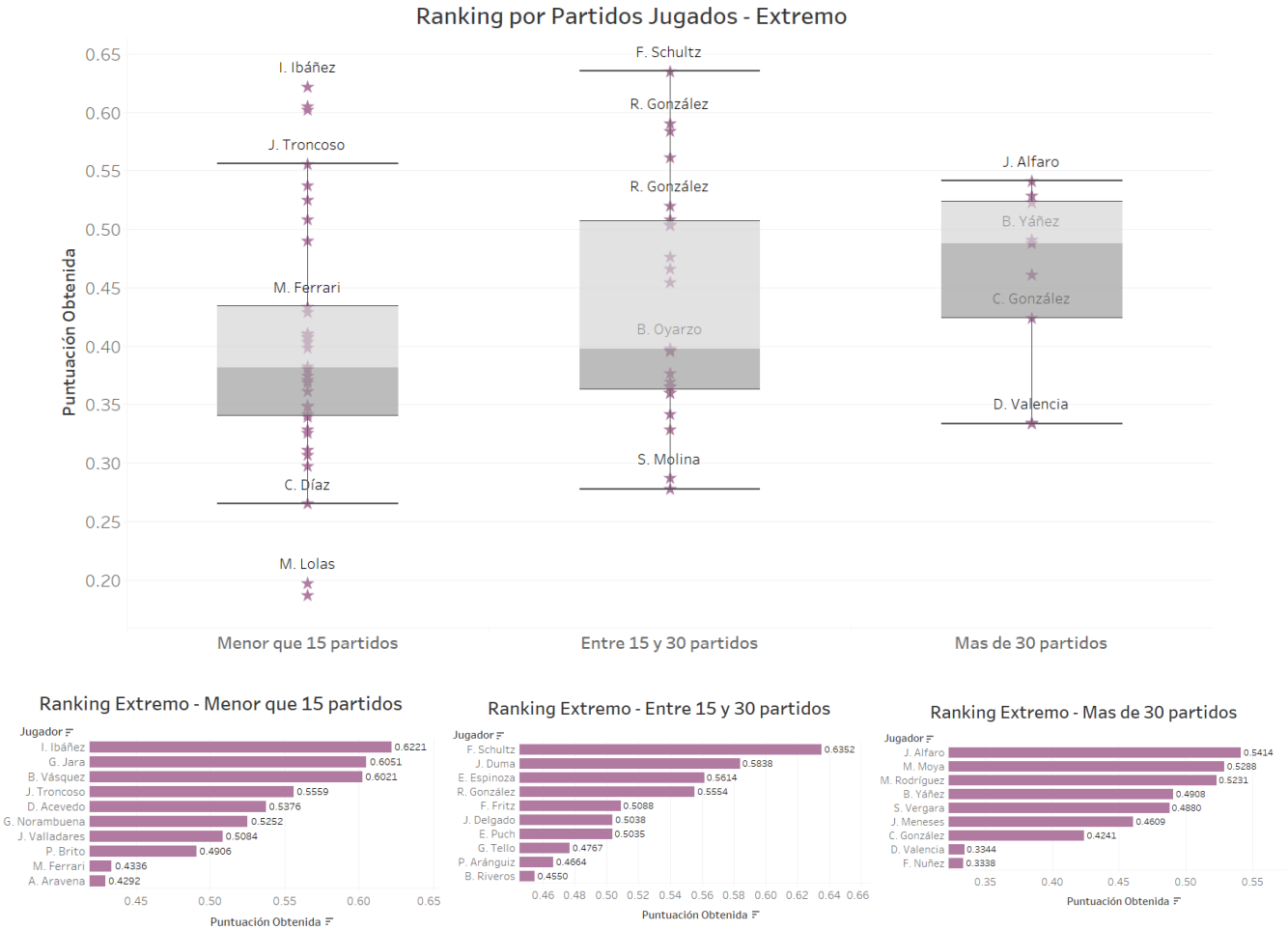


Figura 20. Ranking de Extremo por partidos aproximados jugados.

6.3.2.4. Posición Mediocampo

La gráfica previa evidencia que entre los mediocampistas destacan nombres tan populares como Arturo Vidal "El Rey", quien por muy lejos es el outlier superior entre los jugadores con 15 a 30 partidos disputados, la dilatada carrera de Vidal se ha desarrollada en equipos de élite del continente europeo como son el Bayern Munich, Leverkusen, Juventus, Barcelona, Inter de Milán y actualmente en el Flamengo de Río de Janeiro; es considerado uno de los mejores mediocampistas de la historia de la Bundesliga y el mundo [19][20]. En ese mismo apartado, pero dentro de la liga de clubes chilenos, destaca el mediocampo del Ñublense Mariano Rivera. Con más de 30 partidos, el mediocampo estrella de la liga sudafricana Marcelo Allende destaca como el outlier superior, con estos hallazgos se reafirma el protagonismo de algunos talentos chilenos en otras ligas mundiales. En outliers inferiores, destaca el ex Colo Colo Yashir Pinto, quien actualmente es contratado por el equipo Khon Kaen United de la Liga de Tailandia.

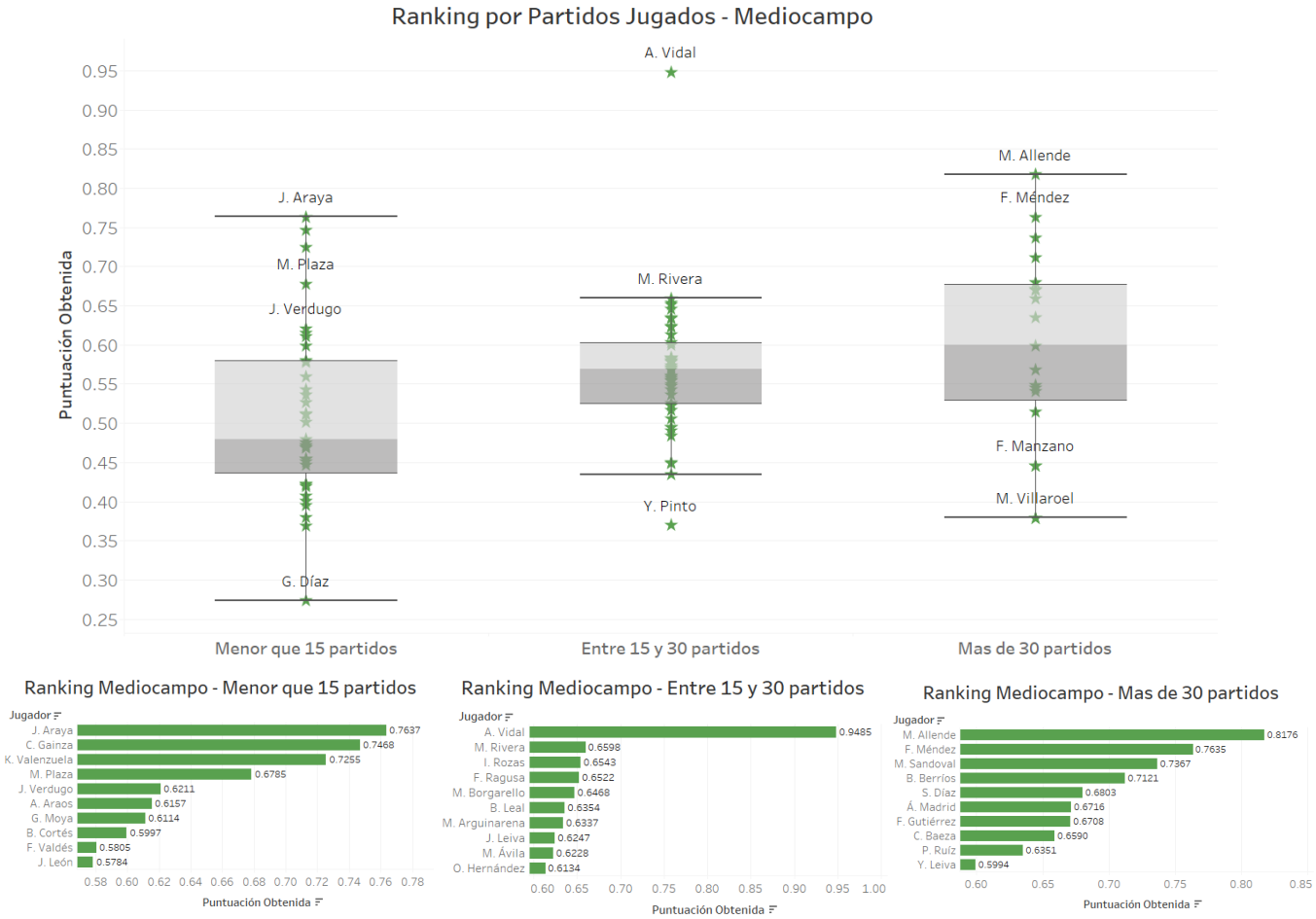


Figura 21. Ranking de Mediocampo por partidos aproximados jugados.

6.3.2.5. Posición Medio Campo Defensivo

Em cuando al ranking de mediocampistas defensivos observado en la figura 22, demostró que Charles Aránguiz del Bayern Leverkusen, y ex seleccionado nacional, es el chileno más destacado en esta posición con más de 30 partidos jugados, sus números positivos y su incuestionable talento lo llevaron a capitanear a este equipo germánico hasta mediados del 2021. Durante su carrera Aránguiz ha brillado en equipos como la Universidad de Chile, Cobreloa, Colo Colo, Porto Alegre de Brasil, además; parte de la selección dorada que consiguió el bicampeonato del continente [21][22]. Por otro lado, el joven Williams Alercón de Unión La Calera y Lorenzo Reyes, ex seleccionado y actual mediocampista del Ñublense, hacen lo propio en los subgrupos de 15 a 30 minutos y menos de 15 minutos, respectivamente. En este último grupo también destaca Adolfo

Ovalle Jr, de la plantilla del Tacoma Defiance del Estado de Washington en los Estados Unidos.

Figura 22. Ranking de Defensas por partidos aproximados jugados.

6.3.2.6. Posición Medio Campo Ofensivo

Por último, en el Ranking de mediocampistas ofensivos (ver figura 23), se obtuvo a figuras destacadas como es Leonardo Gil, figura estelar del Colo Colo, campeón más reciente del campeonato de la primera A [23]. Entre los 15 a 30 partidos destaca el experimentado jugador ex U de Chile y Palestinos César Cortés, uno de los protagonistas del ascenso del Deportivo Magallanes a primera A y de la victoria del equipo magallánico en la más reciente Copa Chile [24]. Por último, con menos de 15 juegos disputados está José Silva, ya experimentado mediocampo que conforma la planta del Lautaro Buin de la

primera B, y que ha sido parte de equipos como Universidad de Chile, Everton, Cobreloa y Deportivo Curicó.

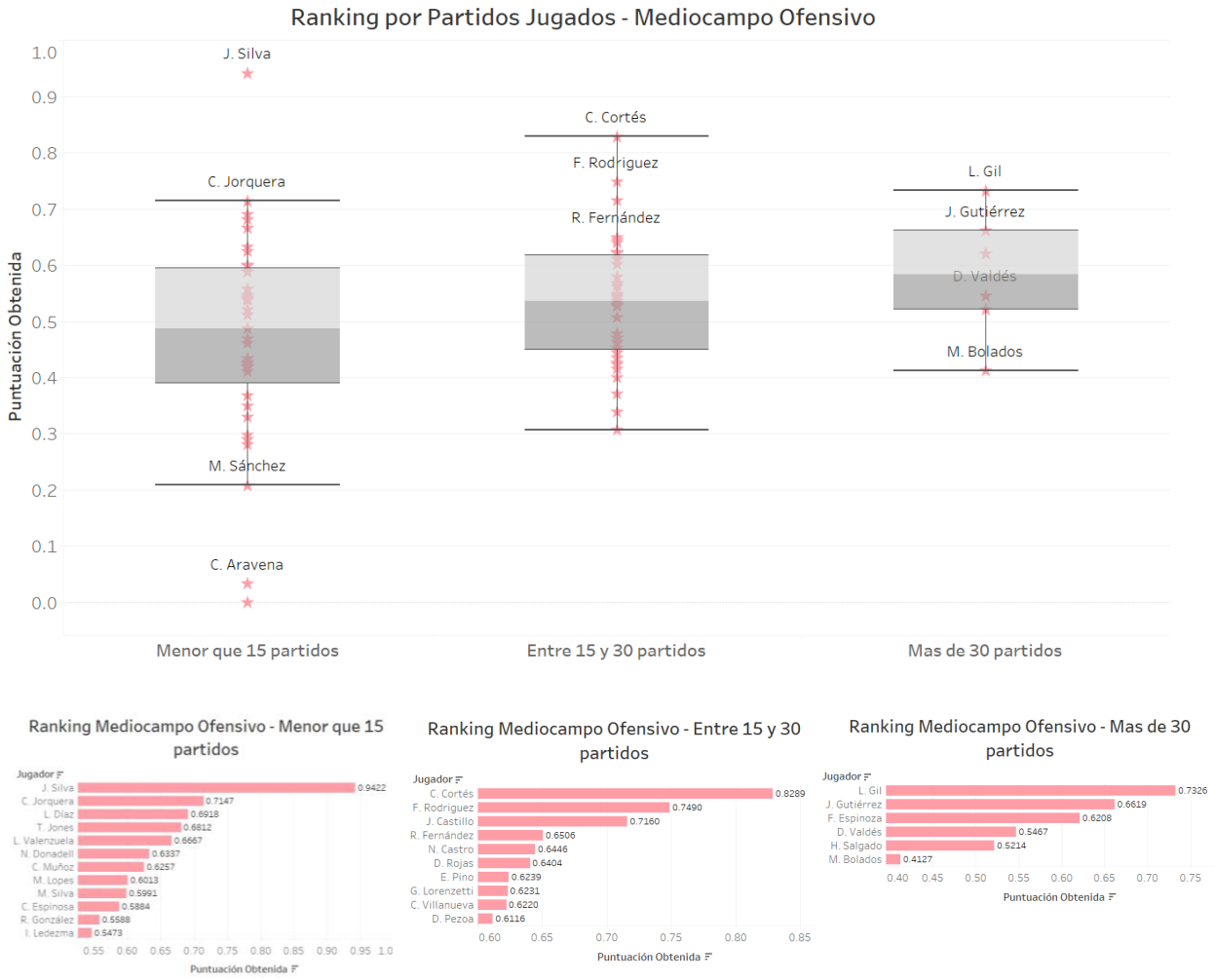


Figura 23. Ranking de Mediocampo ofensivo por partidos aproximados jugados.

7. Conclusiones

En este trabajo se construyó un modelo, con uso de técnicas estadísticas y de machine learning; como son el Support Vector Machine y la Regresión Logística, para predecir las victorias en los partidos de fútbol de las distintas divisiones de la liga chilena. Con base en estos resultados, se construyó un ranking de jugadores por posición según su desempeño. En ambos modelos, SVM y de Regresión logística, se obtuvo el coeficiente de cada una de las 54 variables estudiadas y que se establecieron como inputs o features para la predicción. Estos coeficientes fueron tomados como pesos, que posteriormente fueron multiplicados por el promedio de estas variables que cada jugador chileno tenía en las últimas temporadas. Finalmente, se hizo una sumatoria de todas las variables por futbolista y se construyó el ranking de mayor a menor según el resultado de la adición.

Este trabajo mostró resultados interesantes y acordes a la opinión de periodistas expertos de diversos medios de comunicación escritos. Se pueden citar algunos ejemplos, en el rubro de defensa destaca Paulo Díaz, quien ha tenido resultados brillantes como parte de la plantilla del River Plate en la primera división argentina. Entre los delanteros aparecen los históricos, y muy queridos por el público, Alexis Sánchez y Eduardo Vargas, parte de la generación dorada que llevó al país al bicampeonato continental en 2015 y 2016. La carrera de Sánchez ha sido forjada en los principales clubes europeos y la de Vargas entre los más destacados equipos de Sudamérica, Estados Unidos y Europa.

En la posición de mediocampo el ranking obtenido clasificó como principal figura a Arturo Vidal, quien destaca en ligas mundiales de élite, y parte igualmente de la generación de oro, o Mariano Rivera, quien hace lo propio en el Ñublense de la liga nacional. En medio campo defensivo el nombre de Charles Aránguiz, pieza clave del Bayern Leverkusen, excapitán de ese equipo y también parte de la generación que logró la gloria en la década pasada.

Posterior a un minucioso trabajo de limpieza y exploración de datos se construyó un modelo predictivo con más de un 75% de accuracy y que además fue la base para la construcción de diversos rankings que coincidieron con la opinión general de la fanaticada y de periodistas expertos. La diferencia de este trabajo, con base en la metodología usada por Luca Pappalardo pero adaptada al medio futbolístico del país y diferente al resto de modelos conocidos y usados en Europa, está en que puede ser

tomado en cuenta como un modelo integral para el estudio del fútbol, en particular por entrenadores, cazatalentos y dueños de clubes que trabajan para mejorar el rendimiento de sus organizaciones. De esta manera, este deporte puede acercarse a otros como el Hockey, Basketball y el Baseball, en los cuales el uso de estadísticas y de bases de datos son más frecuentes para el estudio del equipo y la toma de decisiones en contrataciones o cambios estructurales. Lo antes expuesto, una vez más, acentúa la importancia de la ciencia de datos para comprender otras ramas del conocimiento tan diferentes y particulares como son los deportes.

8. Trabajos futuros

Para la construcción de este trabajo se usó una base de datos acotada, debido al tamaño de la liga chilena y su cantidad escasa de equipos. Se recomendaría, para trabajos futuros, el ampliar el número de datos con incluir otras ligas latinoamericanas para hacer del modelo un instrumento más amplio y generalizado, que también pueda construir rankings de otros países o de copas regionales como es la Copa Libertadores.

9. Limitaciones

Una vez más, lo limitado y localizado de los datos impide la construcción de un modelo más generalizado. Haber construido un ranking con base en el desempeño del jugador por partido y no por temporada, como se ha hecho en otros países [2], hubiera resultado en un modelo y ranking más preciso. Sin embargo, el trabajo es una buena aproximación y los resultados corresponden con lo escrito por medios de comunicación expertos.

Referencias Bibliográficas

1. Big Data en el fútbol: Herramientas y funcionamiento [Internet]. KeepCoding Tech School. 2021 [citado el 25 de septiembre de 2022]. Disponible en: https://keepcoding.io/blog/funcionamiento-del-big-data-en-el-futbol
2. Pappalardo L, Cintia P, Ferragina P, Massucco E, Pedreschi D, Giannotti F. PlayeRank: Data-driven performance evaluation and player ranking in soccer via a machine learning approach. ACM Trans Intell Syst Technol [Internet]. 2019;10(5):1–27. Disponible en: http://dx.doi.org/10.1145/3343172
3. Professional Football Platform for football analysis [Internet]. Wyscout. [citado el 25 de septiembre de 2022]. Disponible en: https://wyscout.com/
4. Jordi Duch, Joshua S. Waitzman, and Luís A. Nunes Amaral. 2010. Quantifying the Performance of Individual Players in a Team Activity. PLOS ONE 5, 6 (2010), 1–7. https://doi.org/10.1371/journal.pone.0010937
5. Joel Brooks, Matthew Kerr, and John Guttag. 2016. Developing a Data-Driven Player Ranking in Soccer Using Predictive Model Weights. In Procs of the 22nd ACM SIGKDD Intl Conf on Knowledge Discovery and Data Mining. 49–55.
6. Ralf Herbrich, Tom Minka, and Thore Graepel. 2006. TrueSkillTM: A Bayesian Skill Rating System. In Procs of the 19th Intl Conf on Neural Information Processing Systems. 569–576.
7. R. Stanojevic and L. Gyarmati. 2016. Towards Data-Driven Football Player Assessment. In Procs of the IEEE 16th Intl Conf on Data Mining. 167–172.
8. Oliver Shulte and Zeyu Zhao. 2017. Apples-to-Apples: Clustering and Ranking NHLPlayers Using Location Information and Scoring Impact. In MIT Sloan Sports Analytics Conference. Hynes Convention Center, Boston, MA, USA.
9. Cortes, Corinna; Vapnik, Vladimir (1995). "Support-vector networks" (PDF). Machine Learning. 20 (3): 273–297
10. Tolles, Juliana; Meurer, William J (2016). "Logistic Regression Relating Patient Characteristics to Outcomes". JAMA. 316 (5): 533–4. doi:10.1001/jama.2016.7653. ISSN 0098-7484. OCLC 6823603312. PMID 27483067.

11. Hosmer, David W.; Lemeshow, Stanley (2000). Applied Logistic Regression (2nd ed.). Wiley. ISBN 978-0-471-35632-5.
12. Paulo Díaz integra una prestigiosa lista de 19 jugadores en la era Gallardo. TNT Sports Digital [Internet]. 2022 [citado 3 diciembre 2022];. Disponible en: https://tntsports.cl/internacional/Paulo-Diaz-integra-una-prestigiosa-lista-de-19-jugadores-en-la-era-Gallardo-20220520-0011.html
13. El sueño americano de Maxi. El Ovallino [Internet]. 2018 [citado 3 diciembre 2022];. Disponible en: http://www.elovallino.cl/entrevista/social/pasaporte-ovallino/sueno-americano-maxi
14. Ramirez D. Eduardo Vargas: el héroe de Universidad de Chile en la obtención de la Copa Sudamericana 2011 [Internet]. adnradio. 2021 [citado 3 diciembre 2022]. Disponible en: https://www.adnradio.cl/futbol/2021/12/13/eduardo-vargas-el-heroe-de-la-u-en-la-copa-sudamericana-2011.html
15. Miranda G. EDUARDO VARGAS: EL BIGOLEADOR DE COPA AMÉRICA QUE SE CONSAGRÓ COMO SEGUNDO ANOTADOR HISTÓRICO DE LA ROJA [Internet]. adnradio. 2022 [citado 3 diciembre 2022]. Disponible en: https://www.adnradio.cl/la-roja-dorada/2022/03/30/eduardo-vargas-el-bigoleador-de-copa-america-que-se-consagro-como-segundo-goleador-historico-de-la-generacion-dorada.html
16. Sólo dos jugadores en la historia: La estadística que sólo ostentan Alexis Sánchez y Cristiano Ronaldo [Internet]. 24Horas.cl TVN. 2020 [citado 3 diciembre 2022]. Disponible en: https://www.24horas.cl/deportes/futbol-internacional/solo-dos-jugadores-en-la-historia-la-estadistica-que-solo-ostentan-alexis-sanchez-y-cristiano-ronaldo-3836394
17. Juan Pablo R. Alexis Sánchez brilla y rompe un increíble récord [Internet]. adnradio. 2022 [citado 3 diciembre 2022]. Disponible en: https://tntsports.cl/nacional/Alexis-Sanchez-hace-historia-y-rompe-un-increible-record-20221012-0020.html
18. Jerez M. De la U a campaña brillante en la B: “Quiero dejar una huella”. As.com [Internet]. 2022 [citado 3 diciembre 2022];. Disponible en: https://chile.as.com/futbol/de-la-u-a-campana-perfecta-en-la-b-quiero-dejar-una-huella-n/

19. Xavi: «Vidal es uno de los mejores mediocampistas de la década» [Internet]. Prensa Fútbol. 2021 [citado 3 diciembre 2022]. Disponible en: https://www.prensafutbol.cl/408241-xavi-vidal-es-uno-de-los-mejores-mediocampistas-de-la-decada/
20. Esta es la historia del mediocampista chileno Arturo Vidal en la Bundesliga [Internet]. Bundesliga. 2018 [citado 3 diciembre 2022]. Disponible en: https://www.bundesliga.com/es/bundesliga/noticias/arturo-vidal-historia-liga-alemana-bayern-munich-bayer-leverkusen-barcelona-9477
21. El perfil de la "generación dorada" del fútbol chileno [Internet]. Al aire libre. 2017 [citado 3 diciembre 2022]. Disponible en: https://www.alairelibre.cl/noticias/deportes/futbol/seleccion-chilena/el-perfil-de-la-generacion-dorada-del-futbol-chileno/2017-07-02/225850.html
22. Charles Aránguiz es el nuevo capitán de Bayer Leverkusen [Internet]. ESPN. 2020 [citado 3 diciembre 2022]. Disponible en: https://www.espn.cl/futbol/alemania/nota/_/id/7382912/charles-aranguiz-es-el-nuevo-capitan-de-bayer-leverkusen
23. El nuevo 'clasiquero' de Colo Colo [Internet]. 24Horas.cl TVN. 2022 [citado 3 diciembre 2022]. Disponible en: https://chile.as.com/futbol/el-nuevo-clasiquero-de-colo-colo-n/
24. Miranda G. "Capitán 2023": César Cortés renovó su vínculo con Magallanes para la próxima temporada [Internet]. adnradio.cl. 2022 [citado 6 diciembre 2022]. Disponible en: https://www.adnradio.cl/futbol/2022/12/06/capitan-2023-cesar-cortes-renovo-su-vinculo-con-magallanes-para-la-proxima-temporada.html