Universidad del Desarrollo
Facultad de Ingeniería

# JUGAMOS COMO NUNCA, PERDIMOS COMO SIEMPRE - ¿SIEMPRE GANA EL MEJOR?

Contraste entre las sumatorias de valoraciones de rendimientos individuales de los equipos y el resultado final de un partido de la Primera División del fútbol chileno

POR: BASTIAN GUZMÁN SAGREDO Y MATTHIAS CLEIN ESPINOZA

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

PROFESOR GUÍA:

Dra. Loreto Bravo Celedón

DICIEMBRE 2023

SANTIAGO

“Jugamos como nunca, perdimos como siempre” Alfredo Di Stéfano (1926-2014)

AGRADECIMIENTOS

Queremos expresar nuestro más sincero agradecimiento a nuestras familias, cuyo apoyo y consejos han sido fundamentales para lograr este objetivo. Agradecer también su comprensión y paciencia para aceptar largas jornadas de estudio y de dedicación horaria.

No queremos dejar la oportunidad de agradecer a nuestro grupo de estudio formado en el primer trimestre del programa con el que tuvimos el privilegio de trabajar en las diferentes actividades de las asignaturas. Muchas gracias a Sebastián, Patricia, Fabián, Jose y Leandro por su dedicación, colaboración y puntos de vista que fueron valiosos a lo largo del Magíster e indirectamente enriquecieron este trabajo.

Finalmente agradecemos a nuestra profesora guía, la Dra. Loreto Bravo, por los consejos y comentarios y por gestionar el acceso a los datos para la realización de este trabajo.

TABLA DE CONTENIDO

# Resumen

El fútbol es un deporte colectivo en el cual gana el equipo que le marca más goles al rival. Si bien esto es simple de entender, no es sencillo conseguirlo ni explicarlo pues los equipos marcan goles a través de jugadas colectivas o de acciones individuales que pueden generarse a partir de diversas situaciones y en diferentes contextos durante un partido, como tiros de media distancia, lanzamientos penales, centros al área y jugadas asociadas, entre otras, por lo que dependen directamente de todas las acciones involucradas en el juego como por ejemplo pases, remates, faltas y lanzamientos libres.

Al existir tantas variables involucradas, cada partido de fútbol puede generar una gran cantidad de información. En la actualidad, los avances tecnológicos permiten recopilar y almacenar estos grandes volúmenes, por ejemplo, en el Mundial de Qatar 2022 se generaron cerca de 15.000 data points por partido (FIFA, 2022), lo que equivale aproximadamente a ocho millones de datos por encuentro. Plataformas como Wyscout, Opta, InStat y StatsBomb recogen y publican más de cien métricas distintas de los partidos y los jugadores.

El siguiente trabajo presenta el análisis de los datos de todos los partidos del fútbol chileno de Primera División jugados entre los años 2017 y 2023, obtenidos desde Wyscout, para el desarrollo de modelos predictivos utilizando distintos algoritmos de machine learning, que determinen si un equipo puede ganar o no un encuentro a partir de las métricas grupales. Los pesos o la importancia de las características de dichos modelos se utilizan posteriormente para evaluar y calificar el rendimiento individual de los jugadores involucrados en un partido, con el fin de obtener una valoración final por equipo y contrastar con el resultado final del encuentro para saber si un equipo que tiene mejores rankings individuales es finalmente el que termina imponiéndose en el marcador y llevándose la victoria, o dicho de otra manera: si gana el mejor.

# 1. Introducción

El fútbol es sin duda el deporte más popular en Chile (Iturrieta, 2018). Pese a que en la actualidad este no pasa por su mejor momento ni deportiva ni administrativamente y que tanto los campeonatos locales como la selección nacional se han devaluado en los últimos años, es el deporte que moviliza más gente en el país tanto en atención mediática como en asistencia a los eventos.

La principal competición del país es el campeonato nacional de Primera División, que generalmente cuenta con 16 equipos que disputan treinta partidos anuales, en modalidad todos contra todos con partidos de ida y vuelta, para dirimir un campeón, los clasificados a competencias internacionales (Copa Libertadores y Copa Sudamericana) y los dos descendidos a la Primera B (segunda categoría del fútbol chileno).

Tradicionalmente son los equipos más grandes y con mayor presupuesto los que compiten por el título de campeón. Equipos como Colo-Colo, Universidad de Chile, Universidad Católica, Unión Española, Audax Italiano y Huachipato comúnmente están en la disputa de los primeros lugares, pero últimamente han existido equipos con menores presupuestos, jugadores claves y comodidades, como Cobresal, Palestino, Unión La Calera, Ñublense y Curicó Unido, que han terminado el torneo como campeones o clasificados a copas internacionales.

Lo anterior plantea la interrogante de cuál es la fórmula o clave para poder ser campeón o posicionarse en los primeros lugares del campeonato. Las reglas de la competición y la lógica indican que evidentemente los equipos que más partidos ganan son los que quedan mejor posicionados en la tabla, por lo tanto lo importante sería conocer cuáles son las características que debe tener un equipo o las acciones que más influyen dentro de un partido para que este logre vencer al rival, lógicamente descartando el anotar más goles que el rival. De manera inmediata se podría pensar que los equipos que tengan mejores jugadores (mejores salarios, mejores trayectorias, mayor repercusión mediática, entre otros) son los que ganarán, pero ha habido variados ejemplos a lo largo de los años que demuestran que no es un factor que logre asegurar resultados. Uno de ellos es el

Paris Saint-Germain, equipo francés de la Ligue 1 que en agosto de 2021 hace oficial la contratación del argentino Lionel Messi, uno de los mejores jugadores de la historia, que llegaba a integrarse a un plantel plagado de estrellas mundiales de la talla de Neymar, Kylian Mbappé, Ángel Di María, Marco Verrati, Sergio Ramos, Marquinhos, Achraf Hakimi y Keylor Navas. Al contrario de lo que se hubiera pensado, el PSG al final de la temporada sólo consiguió ganar la Ligue 1 (campeonato de Primera División de Francia) y fue eliminado en octavos de final de la UEFA Champions League, su principal objetivo, tras caer ante el Real Madrid.

Ahora si nos centramos en el juego, debemos necesariamente analizar el estilo y las estadísticas de los equipos participantes de un torneo en particular, pues no será lo mismo conseguir un triunfo en un partido de la Premier League de Inglaterra o en la Primera División chilena. Ante esto el objetivo de este trabajo fue analizar los datos de los últimos siete campeonatos nacionales en su totalidad, los cuales se obtuvieron desde la plataforma Wyscout, para a través de modelos de machine learning, como Regresión Logística, Support Vector Machine (SVM) y Árboles de Decisión, determinar cuáles son las acciones que más influyen para que un equipo gane un partido. La importancia de estas acciones permitió a su vez evaluar el rendimiento individual de cada jugador que disputa un partido para asignarle una puntuación a partir de las acciones que realizó en el encuentro. Finalmente se compararon las sumatorias de puntuaciones de ambos equipos para contrastarlas con el resultado final y determinar si en un partido de la Primera División chilena efectivamente se imponen los equipos que tienen mejor desempeño.

# 2. Trabajo Relacionado

En el mundo de los deportes, la generación de grandes volúmenes de datos de diversas índoles es una constante. Estos datos abarcan desde acciones durante el juego y geolocalización hasta los resultados de los partidos. En los últimos años, el procesamiento y análisis de estos datos ha cobrado una importancia creciente, contribuyendo a la optimización de procesos, desarrollo de estrategias y un mejor entendimiento de los rivales.

En particular, en el análisis deportivo aplicado al fútbol, la utilización de modelos estadísticos y datos detallados para evaluar el rendimiento de los jugadores se ha vuelto un foco de interés significativo. Este interés se ha intensificado gracias al empleo de tecnologías avanzadas y métodos de medición precisos. Un ejemplo claro de esta evolución se observó durante el Mundial de Qatar 2022, donde el registro de datos experimentó un aumento notable, pasando de un promedio de entre 2.000 y 2.500 eventos por partido a alrededor de 15.000 eventos (Infobae, 2022). Este incremento en la recopilación de datos contribuyó a mejorar la equidad y transparencia en la competición. La necesidad de comprender de manera más profunda la naturaleza y dinámica del juego ha impulsado la investigación en este campo (Dufour et al., 2017). Un ejemplo notable de lo mencionado anteriormente es el modelo clasificador de jugadores de Brooks et al. (2016), el cual emplea aprendizaje automático para evaluar la clasificación de jugadores basado completamente en el valor de los pases completados.

El trabajo de Lucca Pappalardo y colaboradores es particularmente relevante para nuestra investigación ya que en su estudio llamado "PlayeRank: data-driven performance evaluation and player ranking in soccer via a machine learning approach", (Pappalardo et al, 2019), diseña e implementa un framework de rankeo de jugadores tomando como base cuatro temporadas de catorce competiciones nacionales (entre ellas Inglaterra, España, Italia, Alemania y Francia), la UEFA Champions League y Europa

League, la Eurocopa 2016 y la Copa del Mundo 2018, lo que significa 19.619 partidos disputados, 21.361 jugadores involucrados y 31.496.332 eventos en total. Este framework es bautizado como PlayeRank y utiliza el modelo Linear Support Vector Machine (LSVC) para clasificar el resultado de un partido dadas ciertas variables de entrada.

El modelo también permite conocer los pesos de dichas variables, lo que se traduce en la influencia que estas tienen en el resultado final del partido de fútbol. Estas variables son seleccionadas de tal manera que también estén presentes en los datos individuales que se registran para cada jugador durante un partido. De esta manera, ponderando cada una de las variables individuales de un jugador por el peso de la variable que entrega el modelo, se puede obtener la valoración de dicho jugador para un partido específico.
El enfoque de Pappalardo ofrece una evaluación más holística y precisa del desempeño de los jugadores. Este enfoque multidimensional es crucial para nuestra investigación, ya que buscamos aplicar y adaptar el modelo Playerank en el contexto específico de la liga chilena de fútbol, abordando tanto el desempeño individual como el colectivo de los jugadores desde la temporadas 2017 hasta la 2023.

Nuestro estudio se alinea con la tendencia creciente de aplicar análisis basados en datos en el fútbol, como se ha observado en otros deportes. Ejemplos en hockey y baseball muestran la implementación exitosa de métricas de rendimiento como Scoring Impact y Performance Efficiency Rating, respectivamente (Schulte & Zhao, 2017; Baumer & Zimbalist, 2014). Estas métricas, junto con el marco conceptual de PlayeRank, nos proporcionan una base sólida para explorar enfoques innovadores en la evaluación del rendimiento de jugadores en la liga chilena, contribuyendo significativamente al estudio y optimización del rendimiento en este deporte.

# 3. Hipótesis y Objetivos

## 3.1. Hipótesis

La sumatoria de las valoraciones de desempeño individual de los jugadores de un equipo durante un partido de fútbol está correlacionada directamente con el resultado final del compromiso, esto quiere decir, que un equipo ganará un encuentro en el caso de que la suma de valoraciones de sus jugadores sea mayor a la del rival.

## 3.2. Objetivos

Objetivo General

- Comparar la sumatoria de valoraciones individuales de los jugadores de dos equipos en un partido de fútbol del campeonato chileno de Primera División y contrastar con el resultado final del encuentro.

Objetivos Específicos

- Caracterizar el campeonato chileno de Primera División a partir de los datos de sus últimas siete temporadas.
- Desarrollar modelos basados en PlayeRank utilizando distintos algoritmos de machine learning, entrenados con los datos de los partidos disputados en el Campeonato Chileno de Primera División entre los años 2017 y 2023.
- Definir los partidos y jugadores a analizar.
- Calificar el rendimiento individual de los jugadores.

# 4. Datos y Metodología

## 4.1. Datos

Para realizar este trabajo fue necesario contar con los datos de todos los partidos del Campeonato de Primera División del fútbol chileno disputados entre los años 2017 y 2023. Con ellos se implementaron modelos de clasificación a partir de distintos algoritmos de machine learning, los cuales permiten determinar si un equipo resultará vencedor en un partido dadas ciertas variables de entrada, como por ejemplo pases, remates, duelos, entre otras. La plataforma Wyscout permite la descarga de un archivo en formato Excel con todos los partidos disputados por un equipo en un año específico, por lo tanto se tienen 115 archivos en formato Excel pues en siete años hubo siete campeonatos: cinco de ellos con 16 equipos, uno con 17 equipos y uno con 18 equipos. La estructura de cada archivo se compone de filas que representan cada partido y columnas que indican la fecha en la que se disputó el encuentro, la descripción del partido (equipos involucrados y resultado), la competición a la que corresponde el partido y 105 variables propias del juego. El listado completo de equipos y las temporadas en que participan en la Primera División se presenta en el Anexo 8.1.

A continuación se presenta un resumen de los últimos siete campeonatos de Primera División, para entregar contexto de sus diferencias en cuanto a cantidad de partidos jugados y equipos participantes por año.

### 4.1.1. Campeonato Nacional 2017

El Campeonato Nacional 2017, también conocido como Torneo de Transición 2017 por ser el paso previo al retorno a los torneos jugados en año calendario (enero a diciembre) en reemplazo del calendario europeo (de agosto a mayo), se jugó entre el 29 de julio y el 10 de diciembre de 2017. Contó con 16 equipos participantes que jugaron 15 partidos cada uno, bajo la modalidad todos contra todos. En total durante el campeonato se

jugaron 122 partidos: 120 correspondientes al campeonato regular y 2 partidos a una llave de ida y vuelta entre Universidad de Concepción y Unión Española disputados los días 14 y 20 de diciembre respectivamente, para definir el cuarto cupo a Copa Libertadores 2018. En este campeonato hubo sólo un equipo descendido, Santiago Wanderers, que perdió la llave de Promoción ante Unión La Calera en lanzamientos penales. El campeón fue Colo-Colo, equipo que consiguió en ese entonces su estrella número 32.

### 4.1.2. Campeonato Nacional 2018

El campeonato Nacional 2018 marca la vuelta a los torneos largos jugados en año calendario. Se disputó entre el 3 de febrero y el 2 de diciembre de 2018 con 16 equipos cuya novedad era el retorno de Unión La Calera a Primera División luego de descender a mediados del año 2016. Cada equipo jugó 30 partidos bajo la modalidad todos contra todos y en partidos de ida y vuelta, contando el torneo con un total de 240 partidos disputados. Universidad Católica fue el equipo campeón consiguiendo su estrella número 13 y los equipos descendidos fueron Deportes Temuco y San Luis de Quillota.

### 4.1.3. Campeonato Nacional 2019

El Campeonato Nacional 2019 se jugó entre el 15 de febrero y el 22 de noviembre de 2019. Estuvo compuesto por 16 equipos y era originalmente un campeonato de 30 fechas, pero debido al estallido social que vivió el país a partir del 18 de octubre de 2019 sólo se disputaron íntegramente 24 fechas más dos partidos de la fecha 25. Los cambios con respecto al 2018 son los descendidos Deportes Temuco y San Luis de Quillota y los ascendidos Coquimbo Unido y Cobresal. En total se disputaron 194 partidos de los cuales sólo se cuenta con los datos de 193 pues Wyscout no registra los datos del partido entre Unión La Calera y Deportes Iquique disputado el 22 de noviembre de 2019 en el estadio Bicentenario de La Florida, pues este fue suspendido al minuto 65 del compromiso por el ingreso de manifestantes a la cancha (Tobar, 2019). Con motivo de la

suspensión del campeonato, no hubo descensos  pero sí se entregó el reconocimiento como campeón a Universidad Católica, que se convirtió en bicampeón del torneo nacional.

### 4.1.4. Campeonato Nacional 2020

Pasaron 63 días para que volviera a disputarse un encuentro de fútbol de primera división en el país. Originalmente pensado para disputarse entre enero y diciembre, el Campeonato Nacional 2020 se jugó entre el 24 de enero de 2020 y el 15 de febrero de 2021. De los campeonatos analizados es el único que tiene partidos en dos años distintos, producto de la pandemia de COVID-19 que obligó a paralizar el campeonato entre el 16 de marzo y el 29 de agosto de 2020 y a jugar cerca del 80% del campeonato sin espectadores en las tribunas. En total se disputaron 306 partidos, considerando la participación de 18 equipos pues por la suspensión del Campeonato 2019 no hubo descensos pero sí ascendieron Santiago Wanderers y La Serena. Se disputó con modalidad todos contra todos con ida y vuelta, jugándose 34 partidos por equipo salvo Colo-Colo y Universidad Católica cuyo partido disputado el 16 de febrero de 2020 fue suspendido momentáneamente a los 20’ por la caída de fuegos artificiales a la cancha y por manifestaciones en las tribunas, y luego suspendido definitivamente por el árbitro Piero Maza a los 76’ al caer petardos cerca del jugador Nicolás Blandi (As.com, 2020). Wyscout no registra los datos de partidos suspendidos, por lo que para términos del análisis el campeonato tuvo 305 partidos disputados al que se le suma el partido por la permanencia en Primera División que enfrentó a Colo-Colo y Universidad de Concepción el día 17 de febrero de 2021 en el Estadio Fiscal de Talca, a puertas cerradas por motivos de la cuarentena total de la ciudad, encuentro que terminaría ganando Colo-Colo por 1-0 con anotación de Pablo Solari a los 19’. El campeón de este torneo fue nuevamente Universidad Católica que conseguía de esta manera su estrella número 15 que los convertía por primera vez en su historia en tricampeones. Por otra parte los equipos descendidos fueron Deportes Iquique, Universidad de Concepción y Coquimbo Unido.

### 4.1.5. Campeonato Nacional 2021

El Campeonato Nacional 2021 trae consigo la vuelta del público al estadio luego de casi un año y medio sin espectadores producto de la pandemia del COVID-19. Disputado entre el 27 de marzo y el 5 de diciembre, fue un torneo donde participaron 17 equipos tras los descensos de Deportes Iquique, Coquimbo Unido y Universidad de Concepción y los ascensos de Ñublense y Deportes Melipilla. Al igual que en años anteriores se disputa en la modalidad todos contra todos con partidos de ida y vuelta, jugándose 272 partidos en total, de los cuales Wyscout registra sólo 271 pues no se encuentran los datos del partido entre Universidad Católica y Universidad de Chile disputado el 7 de noviembre y que tuvo como ganador al equipo cruzado por uno a cero con anotación de Fernando Zampedri a los 47'. El campeón del torneo fue nuevamente Universidad Católica consiguiendo su estrella número 16 y convirtiéndose en tetracampeón del campeonato chileno, marca que sólo había conseguido Colo-Colo al haber ganado los cuatro torneos disputados en los años 2006 y 2007. Los equipos descendidos en esta edición fueron Melipilla y Santiago Wanderers.

### 4.1.6. Campeonato Nacional 2022

El Campeonato Nacional 2022 se jugó entre el 4 de febrero y el 6 de noviembre de 2022 con la participación de 16 equipos tras los descensos de Deportes Melipilla y Santiago Wanderers y el ascenso de Coquimbo Unido. Se jugó en modalidad todos contra todos en partidos de ida y vuelta por lo que debían disputarse en total 240 partidos, de los cuales Wyscout cuenta con registros de 239 pues el partido que debían jugar Antofagasta y Palestino el día 15 de octubre de 2022 fue suspendido porque el conjunto nortino no contaba con la autorización municipal para disputar su encuentro en el Estadio Calvo y Bascuñán (TNT Sports, 2022). Esta situación provocó que se diera como ganador a Palestino por un marcador de 3-0, resultado que complicó al equipo nortino en la tabla de posiciones y provocaría tiempo después su descenso junto a La Serena. El campeón

de esta edición fue Colo-Colo que le sacó una ventaja de 11 puntos a su escolta Ñublense, obteniendo así su estrella número 33.

### 4.1.7. Campeonato Nacional 2023

A la fecha de redacción de este trabajo, el Campeonato Nacional 2023 todavía se encuentra en desarrollo. Comenzó el 20 de enero y está presupuestado que termine durante el mes de diciembre. Cuenta también con 16 equipos y los cambios con respecto al 2022 son los descensos de La Serena y Antofagasta y los ascensos de Magallanes y Copiapó. Hasta el momento se han disputado 27 fechas, lo que significa 216 partidos jugados. El líder momentáneo del torneo es Cobresal con 52 puntos, seguido de cerca por Huachipato con 50 y Colo-Colo con 48 quedando tres fechas por jugar, es decir, nueve puntos en disputa.

A modo de síntesis la Tabla 1 presenta la cantidad de partidos considerados por año además de la sumatoria total de encuentros que se usarán para analizar el campeonato nacional de Primera División y entrenar los modelos de predicción.

| Campeonato | Cantidad de partidos |
|---|---|
| 2017 | 122 |
| 2018 | 240 |
| 2019 | 193 |
| 2020 | 306 |
| 2021 | 271 |
| 2022 | 239 |
| 2023 | 216 |
| Total | 1.587 partidos |

**Tabla 1:** Cantidad de partidos para cada campeonato de la Primera División Chilena.

### 4.1.8. Jugadores

Para la realización de este trabajo también es necesario contar con los datos del desempeño individual de cada uno de los jugadores presentes en el listado del Anexo 8.2, en los partidos que estos disputaron los años 2019, 2022 y 2023. Dichos jugadores fueron seleccionados con el fin de poder analizar cuatro partidos elegidos bajo el criterio de ser de alta importancia y/o "clásicos": Colo-Colo vs Magallanes (Supercopa 2023), Universidad Católica vs Unión Española (Fecha 17 Campeonato 2022), Colo-Colo vs Universidad de Chile (Superclásico 2019) y Universidad Católica vs Colo-Colo (Supercopa 2022). Wyscout permite, para cada jugador, la descarga de un archivo en formato Excel con sus datos de rendimiento individual en todos los partidos que ha disputado, por lo tanto se tendrán 129 archivos en formato Excel cuya estructura se compone de filas que representan cada partido y columnas que indican la fecha en la que se disputó el encuentro, la descripción del partido (equipos involucrados y resultado), la competición a la que corresponde el partido y 69 variables propias del rendimiento individual del jugador durante el encuentro.

## 4.2. Metodología

En este apartado se presenta en detalle la metodología utilizada para llevar a cabo el trabajo. Para exponer de una forma general todas las actividades realizadas se presenta la Figura 1, que incluye un diagrama con las actividades y el flujo de avance del trabajo.

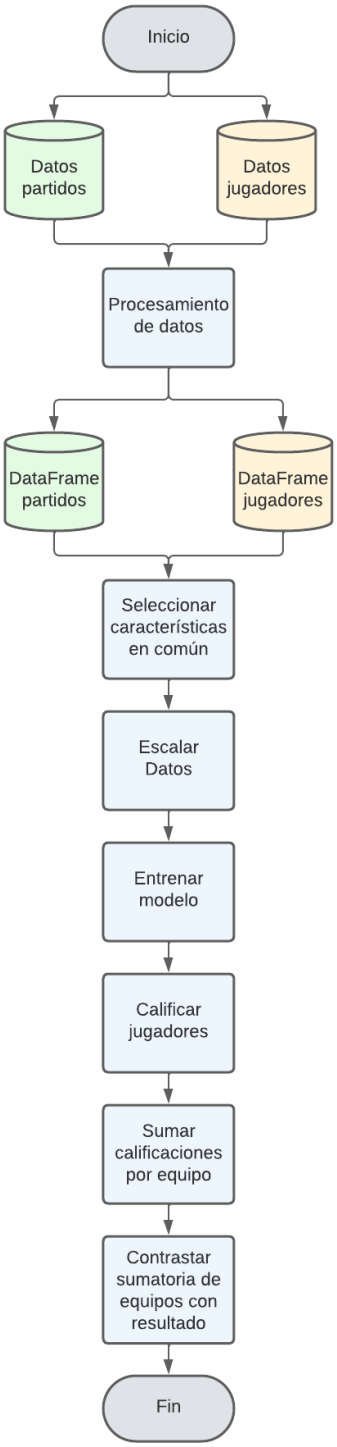



**Figura 1:** Diagrama de flujo con las principales actividades realizadas.

### 4.2.1. Partidos

A partir de los datos de los partidos de cada equipo se generó un único dataset que contiene todos los encuentros disputados en el Campeonato Nacional de Primera División entre los años 2017 y 2023. Este conjunto de datos fue limpiado y procesado, lo que permitió realizar un análisis exploratorio en donde se presentan las principales características de los partidos del fútbol chileno. Luego se seleccionaron las variables que se utilizaron para entrenar los modelos de clasificación, pues estas variables también deben estar presentes en los datos que corresponden al desempeño individual de un jugador durante un partido. Finalmente, antes de entrenar los modelos se normalizaron los datos con el fin de obtener mejores resultados dada la gran diferencia de escala entre variables.

Como se mencionó anteriormente se contó con 115 archivos en formato Excel, cada uno de ellos con los datos de los partidos jugados por un equipo en un año. Estos archivos a su vez cuentan con 108 columnas de características, donde las primeras tres son la fecha del partido, la información general (equipos involucrados y resultado) y la competición a la que corresponde el partido. Las restantes 105 columnas presentan datos de las acciones del equipo en un encuentro, como goles, pases, remates, duelos, tiros libres, posesión, entre otros.

Al cargar los datos de los archivos en un DataFrame de Pandas (biblioteca de Python para análisis de datos) fue posible detectar que muchas de sus columnas no tenían un nombre asignado, lo que se debía a que en muchas de las columnas se asocian nombres que relacionan más de una columna, por ejemplo “Tiros / a la portería” es el título de una columna de la cual se debe asumir el nombre de otras dos, por lo que finalmente los nombres de las tres columnas implicadas deberían ser “Tiros”, “Tiros a la portería” y “% Tiros a la portería”. En total existían 56 columnas que no tenían un nombre asignado y que Pandas renombró automáticamente como “Unnamed: X” siendo X el número de la columna.

Ante esto se generó una función capaz de detectar todas las columnas que llevan por nombre “Unnamed” y relacionarlas con las columnas inmediatamente anteriores para poder generar correctamente sus nombres de manera automática. La función renombra correctamente la mayoría de las columnas, siendo las únicas excepciones las columnas cuyos nombres se desprenden de la columna 52 “Entradas al área de penalti (carreras / pases cruzados)” las cuales finalmente se modificaron de manera manual. Luego, sólo bastó con corregir los títulos originales de las columnas por sólo la primera parte de estos. Una vez realizado esto, el DataFrame que contiene todos los partidos jugados por un equipo en un año estuvo listo para ser utilizado.

Se propuso consolidar la información de todos los partidos, equipos y años en un solo DataFrame para simplificar el procesamiento de datos. Se implementó un script capaz de iterar desde el año 2017 hasta el 2023 y que en cada iteración obtiene las rutas de los archivos correspondientes al año. Luego itera por cada archivo, carga la información del archivo en un DataFrame temporal, reemplaza el nombre de las columnas, con la metodología presentada anteriormente y a través de expresiones regulares extrae detalles clave, como el nombre de los equipos y el resultado. Luego, crea un nuevo DataFrame llamado “matches_df”, en el que integra y concatena los datos procesados de cada archivo. Este DataFrame final compuesto de 4111 registros y 112 columnas, representa una compilación completa de la información de todos los partidos, con detalles como el nombre de los equipos locales y visitantes, así como el equipo analizado en cada partido.

Una vez obtenido el DataFrame general, se procedió a enriquecer su información, esto quiere decir que se agregaron nuevas columnas con datos que podían aportar al análisis exploratorio y al entrenamientos de los modelos de predicción. Se crearon nuevas columnas que corresponden a la cantidad de pases, tiros, centros y duelos (en todas sus formas) errados. Estos valores se pueden inferir desde los mismos datos pero se añadieron explícitamente para utilizarlos posteriormente como características de los modelos de predicción que se implementaron.

Además se agregaron tres columnas extra: goles marcados por el equipo local, goles marcados por el equipo visitante y el total de goles marcados. Esto simplificó la caracterización del campeonato chileno realizada a través de un análisis exploratorio, el cual se presenta en la sección de Resultados.

Finalmente, fue necesario enriquecer el DataFrame con una columna que indica explícitamente el resultado que obtuvo el equipo analizado. Esta información se puede obtener a partir del cruce de columnas "Partido" y "Goles" y sirvió como variable objetivo para entrenar los modelos de predicción. Luego de todas estas inserciones, el DataFrame contó en total con 127 columnas.

Como se mencionó anteriormente el DataFrame "matches_df" cuenta con 4111 registros y 127 columnas, pero no todos sus registros corresponden a partidos del campeonato chileno de Primera División, que es en el cual se enfoca este trabajo, por lo tanto se debió realizar un filtro. El detalle de la composición del DataFrame en cuanto a partidos por competición se presenta en la Tabla 2. El DataFrame filtrado que finalmente se utilizó para realizar el trabajo cuenta con 3174 registros y 127 columnas. Cabe destacar que estos 3174 registros corresponden a las estadísticas grupales de los dos equipos involucrados en cada uno de los partidos analizados, es decir, corresponden a los 1587 partidos descritos en la Tabla 1.

| Competición | Cantidad de registros |
|---|---|
| Primera División | 3174 |
| Copa Chile | 373 |
| Primera B | 271 |
| Copa Sudamericana | 140 |
| Copa Libertadores | 115 |
| Amistosos Internacionales | 19 |
| Supercopa de Chile | 11 |
| Playoffs | 4 |
| Torneos de Verano | 4 |
| Total | 4.111 registros |

**Tabla 2:** Cantidad de registros por competición en el DataFrame “matches_df”.

### 4.2.2. Jugadores

Por su parte, el procesamiento de los archivos que contienen los datos de los jugadores se realizó de la manera descrita para los archivos de los partidos, esto quiere decir que se generó un DataFrame general con todos los partidos de todos los jugadores y se reemplazaron los nombres de las columnas que estaban incompletas. Además, para lograr una correspondencia completa se debió editar ligeramente algunos títulos de las columnas. Esto, si bien podría no haberse realizado, facilitó el proceso final de calificación, pues ambos DataFrames contaban con los mismos nombres en sus columnas. Los cambios realizados fueron: de Pases en profundidad logrados a Pases en profundidad completados, de Tiros logrados a Tiros a la portería y de Carreras en profundidad a Entradas al área de penalti con carreras.

El DataFrame de los jugadores también fue enriquecido con nuevas columnas, en este caso principalmente con las que corresponden a acciones erradas considerando pases, tiros, centros y duelos, las cuales se calcularon como la diferencia entre el total y los completados.

Una vez generado el DataFrame de los jugadores bastó con filtrar por partido para contar con los datos de todos los jugadores involucrados en un encuentro determinado. Se decidió generar un archivo para cada partido para respaldar la información.

### 4.2.3. Modelos de clasificación

Para generar una relación entre los datos de los partidos y los datos individuales de los jugadores se determinaron las características que estos tienen en común. En total se seleccionaron 46 características, presentadas en el Anexo 8.3, para entrenar los distintos modelos y para posteriormente, a partir de los pesos o importancia de cada característica, ponderar las acciones individuales de los jugadores para obtener una valoración de su rendimiento en un partido. A las 46 características se le sumó la columna "Resultado", la cual, como se mencionó anteriormente, será utilizada como variable objetivo, por lo tanto fueron 47 columnas las que se almacenaron en un nuevo archivo llamado "model_df.csv".

Antes de comenzar con el entrenamiento de los modelos se debió escalar los valores de las características, debido principalmente a la significativa diferencia de escalas de valores que existe entre ellas. Por ejemplo, el valor de la variable pases puede estar entre 139 y 686, a diferencia de tiros a la portería que puede ir de 0 a 14 o balones recuperados que puede ir de 43 a 136. Para realizar esto se utilizó Robust Scaler, un método de escala proporcionado por la librería Scikit-learn en Python que resta la mediana y escala los datos de acuerdo con el rango intercuartílico, lo que entrega como resultados datos escalados menos sensibles a los valores atípicos. Es importante destacar que los parámetros del escalador debieron ser almacenados para posteriormente escalar

los datos de los jugadores utilizando los mismos valores. Los parámetros del escalador se almacenaron en un archivo llamado “scaler.pkl”.

Una vez escalados los datos fue momento de separar el DataFrame en dos conjuntos. Uno de entrenamiento con el cual se entrenaron los modelos de predicción, y un conjunto de prueba con el cual se testeó la capacidad predictiva de los modelos. Esta acción se realizó utilizando la función train_test_split de Scikit-learn, configurando su parámetro de tamaño del conjunto de prueba en un 20% del total y el parámetro random_state en un valor fijo para lograr que el experimento sea reproducible (entregar los mismos resultados independiente del momento de ejecución). La función train_test_split entregó como resultado cuatro listas de Python, correspondientes al vector “X” de entrenamiento, el vector “X” de prueba, el vector “y” de entrenamiento y el vector “y” de prueba respectivamente.

Para la realización de este trabajo se utilizaron tres algoritmos de aprendizaje supervisado: Regresión Logística, Support Vector Machine (SVM) y Random Forest con el fin de realizar una comparación de los resultados obtenidos con cada uno de ellos.

#### 4.2.3.1. Regresión Logística

La regresión logística es un algoritmo de machine learning utilizado para la clasificación y predicción de variables categóricas o binarias. El objetivo principal de la regresión logística es encontrar el mejor modelo para describir la relación entre la característica binaria de interés (variable dependiente) y un conjunto de variables independientes (predictores). La regresión logística genera las probabilidades de ocurrencia de un evento específico.

Para utilizar este modelo, es necesario importar la clase LogisticRegression desde el módulo linear_model de Scikit-learn. Luego, es necesario instanciar la clase y utilizar el método fit, entregando como parámetros los vectores “X” e “y” de entrenamiento.

#### 4.2.3.2. Support Vector Classification

Es un algoritmo de aprendizaje supervisado utilizado tanto en problemas de clasificación como de regresión, este tiene como objetivo encontrar el hiperplano que mejor separa los datos en clases distintas en un espacio multidimensional, esto quiere decir que el algoritmo de SVM busca encontrar un plano que tenga el margen máximo, es decir, la distancia máxima entre los puntos de datos de ambas clases. Maximizar la distancia del margen proporciona cierto refuerzo para que los puntos futuros de datos puedan ser clasificados con más confianza.

Para utilizar este modelo, es necesario importar la clase SVC (Support Vector Classification) desde el módulo svm de Scikit-learn. Luego, es necesario instanciar la clase y definir su parámetro kernel como “linear” cuando se quiere implementar un clasificador binario. Finalmente se utiliza el método fit para entrenar el modelo, entregando como parámetros los vectores “X” e “y” de entrenamiento.

#### 4.2.3.3. Random Forest

Es un algoritmo de machine learning basado en el ensamblaje de árboles de decisión. Utiliza una técnica para construir múltiples árboles de decisión y combinar sus resultados. El objetivo principal de un modelo de Random Forest es combinar múltiples árboles de decisión para resolver un problema. Cada árbol individual en el bosque aleatorio genera una predicción de clase y la clase con más votos se convierte en la predicción del modelo. Los bosques aleatorios corrigen la tendencia de los árboles de decisión a sobreajustar a su conjunto de entrenamiento. Al combinar muchos árboles, este método tiende a reducir la varianza y mejorar la precisión de las predicciones.

Para utilizar este modelo, es necesario importar la clase RandomForestClassifier desde el módulo ensemble de Scikit-learn. Luego, es necesario instanciar la clase, definir el número de estimadores (cantidad de árboles) y el random state para hacer reproducible

el experimento. Al igual que con los modelos anteriores, se utiliza el método fit para entrenar el modelo a partir de los conjuntos “X” e “y” de entrenamiento.

Todos los modelos implementados fueron almacenados en archivos con extensión .pkl la cual permite serializar las estructuras de datos de Python de manera sencilla. La finalidad de guardar los modelos es su posterior uso en el proceso de calificación de las actuaciones individuales de los jugadores, pues es necesario conocer los pesos o importancia de las características para poder ponderar sus datos.

Cada modelo fue utilizado para realizar una prueba con el conjunto de datos de test. Para esto fue necesario utilizar el método predict, incluido en cada uno de ellos, entregando como parámetro el vector “X” de prueba. El resultado de esas predicciones se guardó en una variable para posteriormente con ayuda de la función accuracy_score incluida en el módulo metrics de Scikit-learn comparar los resultados entregados por el modelo con los resultados reales almacenados en el vector y de prueba. La función accuracy_score entrega como resultado un porcentaje que indica la cantidad de predicciones realizadas correctamente sobre la cantidad total de predicciones.

### 4.2.4. Calificación

Una vez realizado lo anteriormente descrito, se seleccionaron cuatro partidos con alta relevancia del fútbol chileno, como Superclásicos, Clásicos y Finales de Copa, para calificar a todos los jugadores que disputaron dichos encuentros. Cada equipo tiene en promedio 14 jugadores involucrados (11 titulares más 3 suplentes) por partido, por lo que el análisis de un encuentro contó con cerca de 30 calificaciones individuales.

Para comenzar con la calificación de jugadores lo primero fue cargar los archivos correspondientes a los partidos a analizar y almacenar la información en un DataFrame. Luego, para facilitar el procedimiento de puntuación, fue necesario eliminar ciertas

columnas del DataFrame como “Partido”, “Jugador”, “Equipo”, “Posición específica” y “Minutos jugados”, pues estas no son consideradas dentro de los modelos.

Una vez cargados los datos de los jugadores involucrados en el partido a analizar, se cargó también el escalador almacenado en el archivo scaler.pkl cuya implementación se presentó anteriormente. Utilizando el método transform del escalador se escalaron los datos de los jugadores para que estos estén en el mismo rango que las variables que se utilizaron para entrenar el modelo. Ya escalados los valores, se cargaron los modelos desde su archivo .pkl para comenzar con la calificación de los jugadores. A partir de cada modelo se obtuvieron los coeficientes o la importancia de las características y luego se realizó una sumatoria de los productos entre los coeficientes y los datos del jugador, siendo el resultado final el PlayeRank del jugador. Una vez obtenidos los puntajes individuales de todos los jugadores se escalaron los valores utilizando el escalador MinMax, el cual realiza un escalamiento, transformando cada valor al rango entre 0 y 1, conservando la distribución relativa de los datos originales y considerando los valores mínimos y máximos presentes en el conjunto. Hecho esto, se procedió a sumar los valores por equipo para finalmente comparar si el equipo que mayor puntuación tuvo efectivamente resultó vencedor del encuentro, esto con el fin de determinar si existe una correlación entre la sumatoria de las calificaciones individuales de un equipo y el resultado final de un partido de fútbol.

El proceso descrito anteriormente se realizó para cada uno de los cuatro partidos seleccionados y utilizando los tres algoritmos para finalmente realizar una comparación de los resultados donde se considerará la diferencia entre las calificaciones y si finalmente estas se condicen con el resultado final del partido.

# 5. Resultados

En este apartado se presentan los resultados del trabajo al aplicar la metodología explicada en el apartado anterior. En primer lugar se presenta un análisis exploratorio de los datos de los partidos, con el fin de caracterizar el campeonato chileno de Primera División. Luego, ya teniendo ese contexto, se presentan los resultados de los modelos de aprendizaje supervisado que se implementaron, poniendo especial énfasis en la importancia de las características al realizar una clasificación. Posteriormente se realiza una comparación de los resultados de todos los modelos tanto para clasificación binaria como multiclase con la finalidad de seleccionar el de mejor desempeño. Finalmente, a partir de la importancia de características del modelo seleccionado se realiza la calificación de desempeños individuales de los jugadores presentes en los cuatro partidos seleccionados: Colo-Colo vs Magallanes (Supercopa 2023), Unión Española vs Universidad Católica (Clásico de Independencia 2022), Colo-Colo vs Universidad de Chile (Superclásico 2019) y Universidad Católica vs Colo-Colo (Supercopa 2022). Con las valoraciones individuales de los jugadores se pudo obtener el valor total del equipo y contrastar con el resultado final del encuentro.

## 5.1. Análisis exploratorio

El conjunto de datos analizado fue el de partidos de primera división, que como se mencionó en apartados anteriores, cuenta con 3174 registros los que corresponden a 1587 partidos disputados entre los años 2017 y 2023. Cuenta también con 127 columnas (características) de distinta naturaleza y por lo tanto con distintos tipos de datos, como objetos (cadenas de texto) y valores numéricos enteros y reales. Como la información fue extraída directamente desde la plataforma Wyscout, el conjunto no presenta valores nulos en ninguno de sus registros, por lo que no es necesario realizar descarte de registros o imputación de datos.

Lo primero que se puede obtener a partir de los datos es la distribución de los partidos a través de los años, lo que se presenta en la Figura 2.

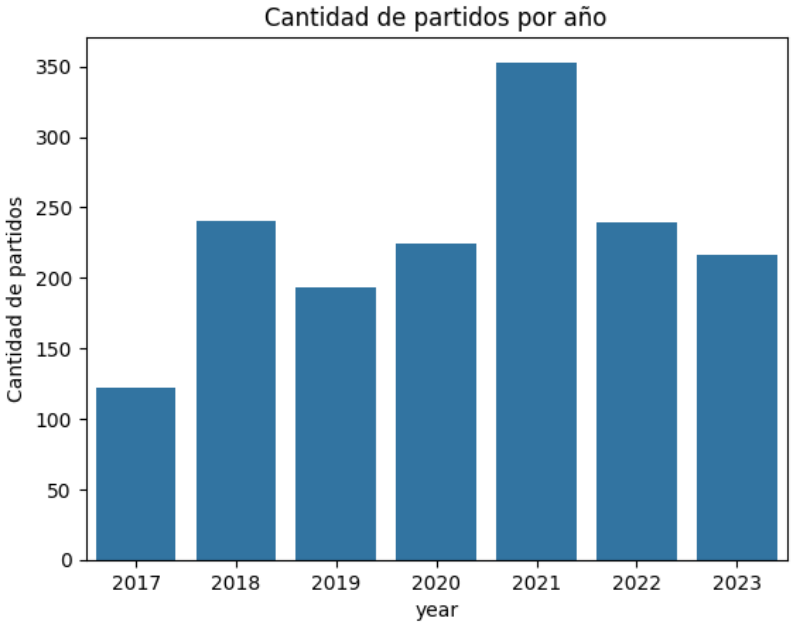



**Figura 2:** Cantidad de partidos del campeonato de Primera División disputados en los últimos 7 años

La notoria diferencia de valores del año 2017 se debe a que en dicho año se disputó un torneo de transición, es decir, un torneo corto de 15 fechas a diferencia de las 30 que comúnmente se juegan en el año. Con respecto al año 2021, la gran cantidad de partidos se debe a la pandemia de COVID-19 que obligó a paralizar el campeonato 2020. Esto generó que el calendario tuviera que acomodarse y terminar de disputar el campeonato en el verano del 2021.

En los últimos siete años han sido 24 equipos los que han participado en al menos un campeonato de Primera División. Es sabido que cada año se produce un recambio de equipos al producirse descensos y ascensos desde la Primera B. Los 24 equipos que son parte del análisis de este trabajo se presentan en el anexo 8.1 y la cantidad de partidos que disputa cada uno de los clubes se presenta en la Figura 3.

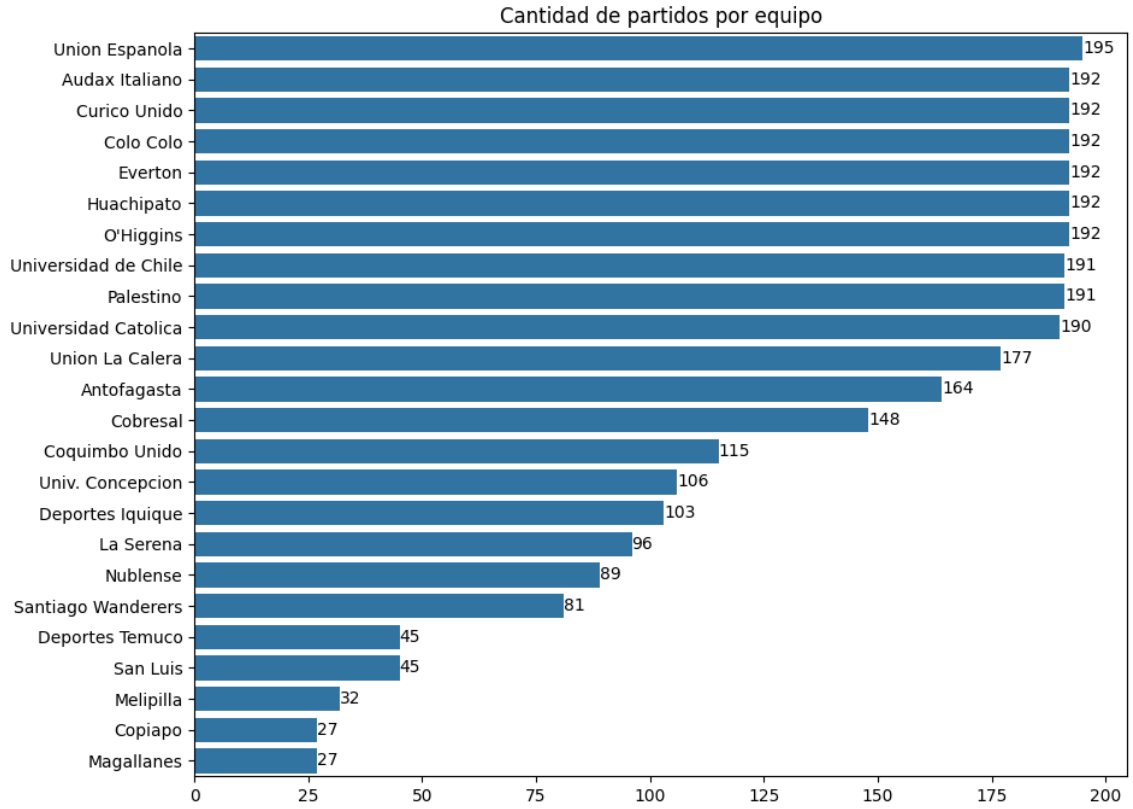



**Figura 3:** Cantidad de partidos en Primera División de todos los equipos participantes en los últimos siete años.

A partir de la Figura 3 se puede apreciar claramente los diez equipos que no han descendido en los últimos siete años: Unión Española, Audax Italiano, Curicó Unido, Colo-Colo, Everton, Huachipato, O'Higgins, Universidad de Chile, Palestino y Universidad Católica. La diferencia entre ellos en cuanto a la cantidad de partidos se debe principalmente a la disputa de partidos especiales y/o fechas suspendidas, lo que fue explicado en el apartado de Datos. Además se aprecia la baja cantidad de partidos disputados por Copiapó y Magallanes, ambos equipos ascendidos a finales de 2022.

### 5.1.1. Cantidad de goles marcados.

El histograma de la Figura 4 presenta la distribución de la cantidad de goles por partido en el Campeonato Nacional. Es relevante destacar que en el 74% de los partidos analizados, los goles totales marcados en un partido fueron entre 0 y 3. Esta concentración en el rango más bajo de goles supone que los partidos suelen ser reñidos, posiblemente defensivos y/o con poca elaboración de jugadas ofensivas.

Por otro lado, los marcadores con una mayor cantidad de goles totales marcados por partido son menos comunes, lo que se observa en la disminución progresiva de la frecuencia a medida que aumenta el número de goles por partido. En particular, los partidos con más de 3 goles representan sólo el 26% del total, lo que indica que las ocasiones en las que se genera una goleada son relativamente escasas.

También se puede observar que en sólo un partido se han marcado un total de 9 goles, siendo este el máximo presente en el conjunto de datos. El partido en cuestión es el que disputaron O'Higgins y Antofagasta el día 28 de julio de 2019 en el estadio El Teniente de Rancagua, donde el conjunto nortino se impuso en condición de visitante por seis goles a tres.

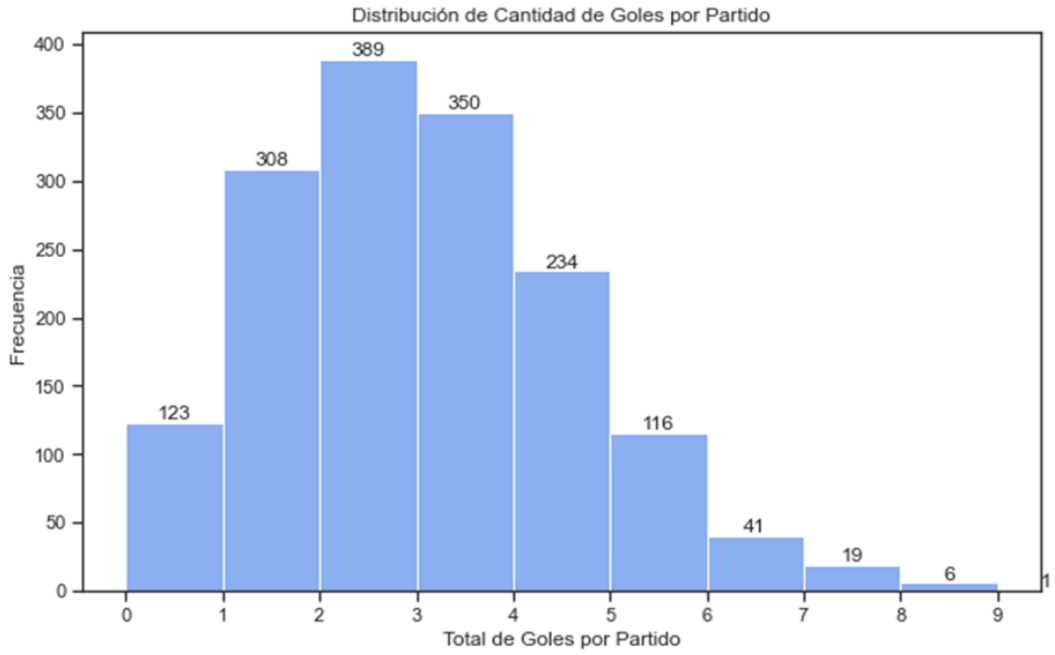



**Figura 4:** Distribución de la cantidad de goles por partido en el campeonato nacional de Primera División.

### 5.1.2.Cantidad de minutos jugados.

Al analizar los 1.587 partidos de la primera división de la liga chilena, se determina que la duración promedio de los encuentros es de 97 minutos, lo que se presenta en la Figura 5, y con lo presentado en la Figura 6 se refuerza esta observación pues muestra una tendencia creciente hacia partidos más largos, especialmente en el 2023. Este incremento en la duración de los juegos coincide con cambios reglamentarios de la FIFA (Águila, 2023), diseñados para aumentar el tiempo efectivo de juego, el cual, según estudios, promedia sólo 60 minutos por partido. Las modificaciones incluyen medidas como otorgar tiempos de descuentos más prolongados y adicionar el tiempo detenido al cronómetro oficial. Estos ajustes buscan mitigar las pérdidas de tiempo, frecuentemente atribuidas a tácticas como simulaciones, para asegurar que el tiempo de juego sea lo más fiel posible a los 90 minutos estipulados.

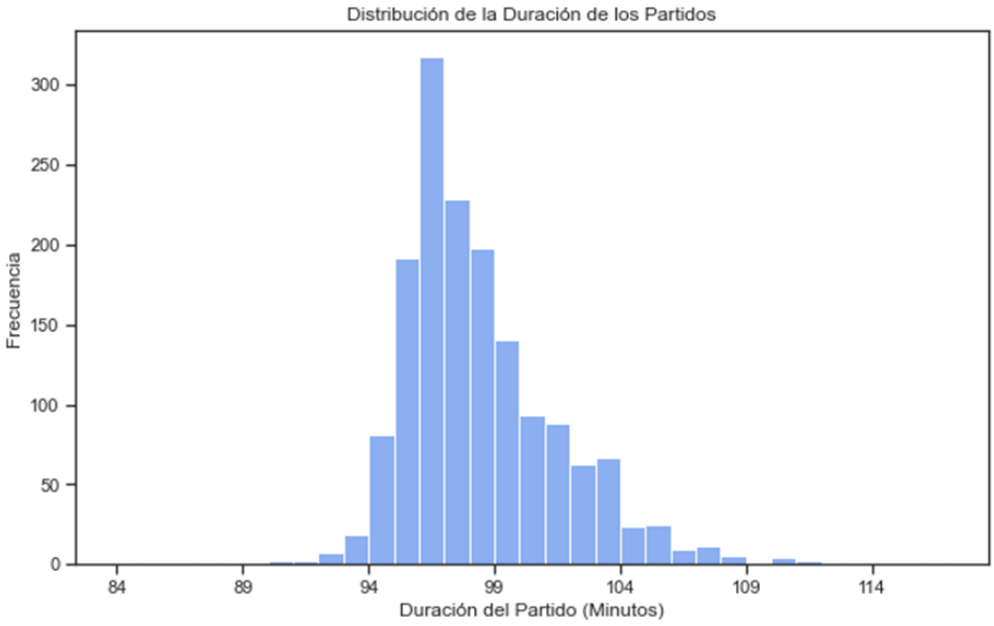



**Figura 5:** Distribución de los minutos jugados por partido.

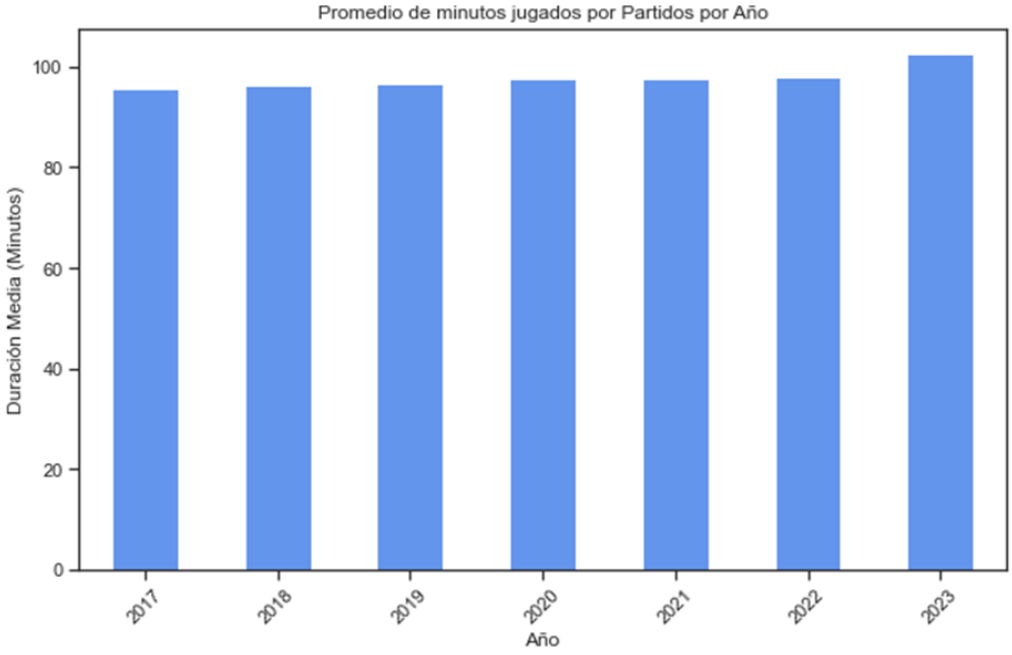



**Figura 6:** Duración promedio de los partidos por año.

### 5.1.3. Cantidad de tiros.

Al analizar las variables de los tiros se obtiene la Figura 7, la cual muestra una distribución que detalla que la mayoría de los equipos rematan entre 5 y 15 veces por partido, siendo 11 el promedio, sugiriendo esto un enfoque táctico centrado en la posesión y en la calidad de las oportunidades más que en la cantidad. De la misma forma se puede observar en la Figura 8 que a lo largo de los años existe una fluctuación en el volumen de tiros, alza notable en 2021, debido a la alta concentración de partidos. Otro punto importante para destacar es la relación entre tiros y tiros a la portería detallado en la Figura 9 en donde se aprecia una tendencia consistente ya que no todos los tiros resultan en intentos directos al arco, lo que subraya la importancia de la precisión sobre la simple acumulación de intentos. Este análisis sugiere que mientras los equipos buscan maximizar su producción ofensiva, la precisión en el ataque es un factor clave para el éxito.

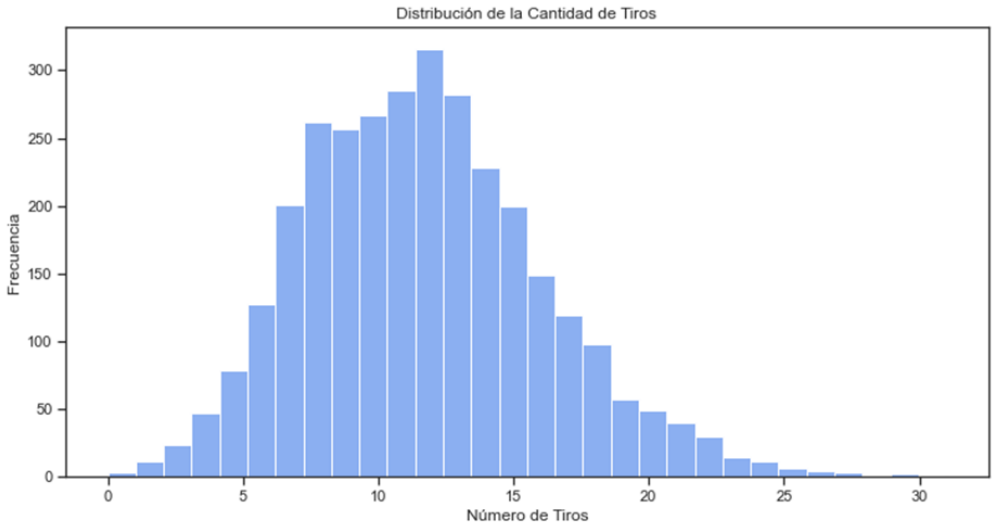



**Figura 7:** Distribución de la cantidad de tiros por partido.

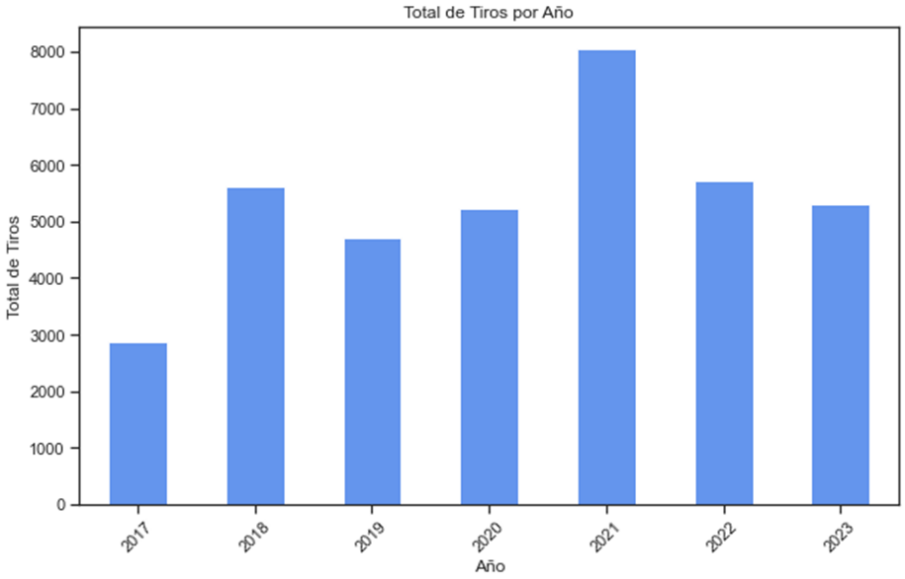



**Figura 8:** Total de tiros realizados por año.


**Figura 9:** Relación entre tiros y tiros a la portería.

### 5.1.4. Cantidad de pases.

En relación al análisis de los pases del campeonato nacional, se aprecia una tendencia clara de los equipos a realizar entre 300 y 400 pases por partido, siendo 374 la media, lo cual evidencia que en gran parte los equipos del fútbol chileno poseen un juego centrado en la posesión y la construcción de jugadas colectivas. La distribución de la cantidad de pases por partido se presenta en la Figura 10.

En la Figura 11 se aprecia que existe una alta concentración de pases logrados, lo que se traduce en que los pases que se realizan en su gran mayoría terminan siendo pases logrados en vez de balones perdidos.

Finalmente, la relación que existe entre los pases y los pases logrados detallada en la Figura 12, indica que existe un bajo porcentaje de pases errados, lo que resalta que existe una estrategia de juego que prioriza la eficiencia y la efectividad en la entrega del balón. La correlación significativa que existe entre pases y pases logrados se traduce directamente en una alta posesión del balón y una alta tasa de pases efectivos.

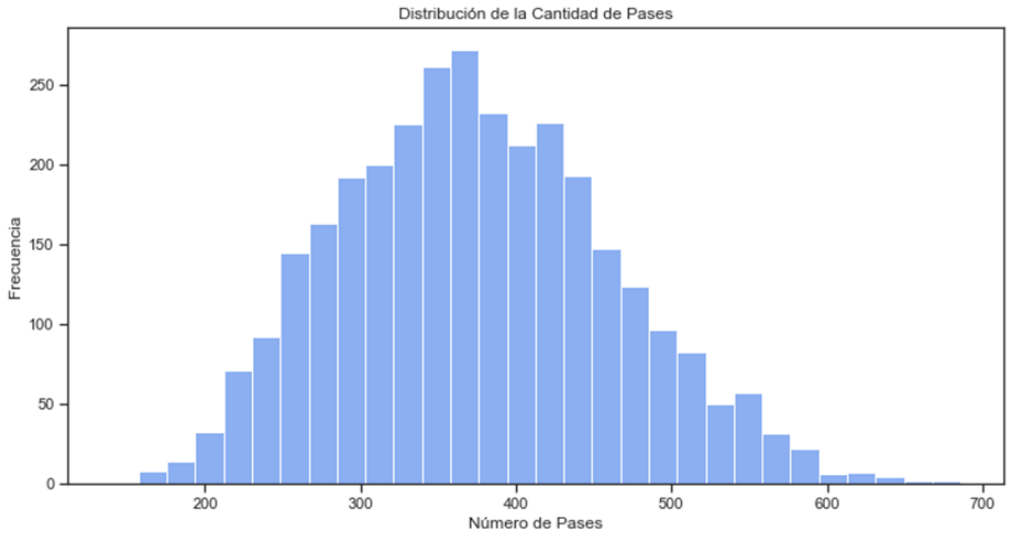



**Figura 10:** Distribución de la cantidad de pases por equipo en un partido.

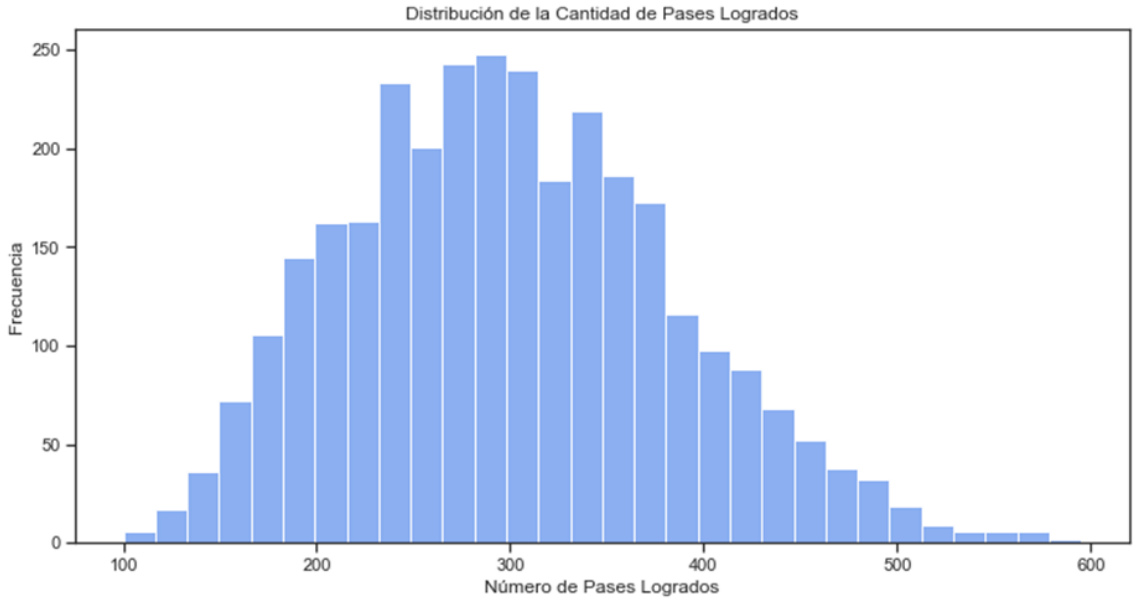



**Figura 11:** Distribución de pases logrados por equipo en un partido.

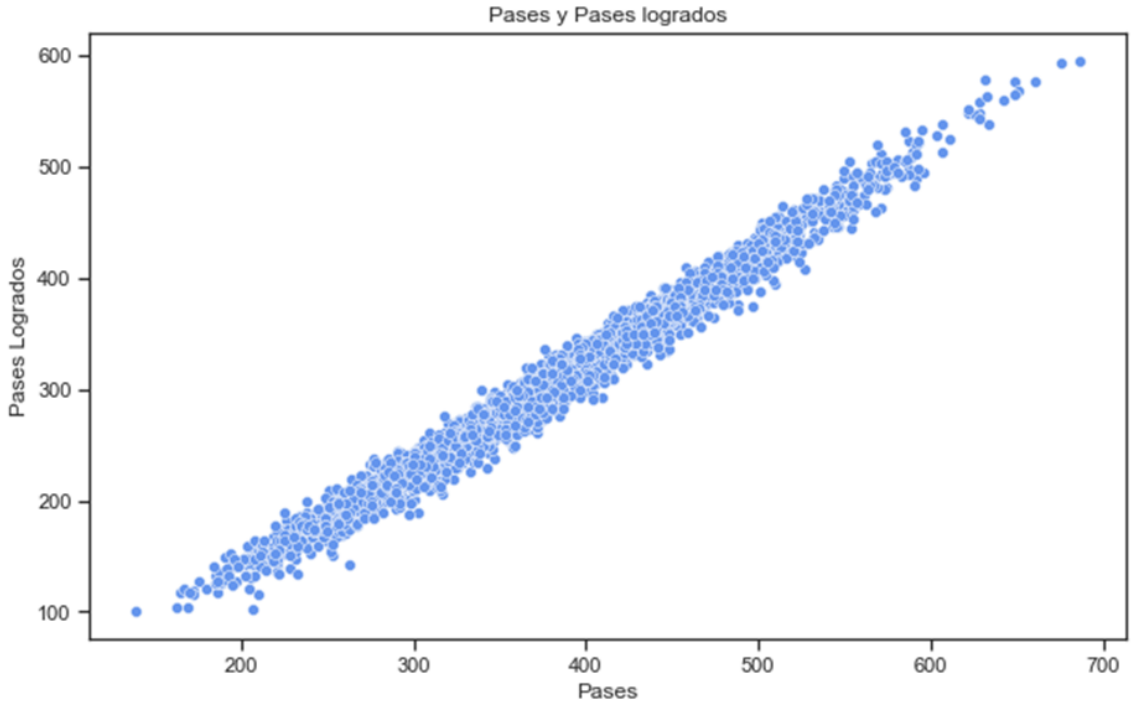



**Figura 12:** Relación entre pases y pases logrados.

### 5.1.5. Cantidad de duelos.

El histograma presentado en la Figura 13 muestra la cantidad total de duelos realizados por un equipo en un partido. La distribución parece aproximarse a una distribución normal, donde la mayoría de los partidos presentan un número de duelos en un rango entre 200 y 250, siendo 232 el promedio de duelos realizados por partido. Esto indica que constantemente existe una lucha por disputar el balón, lo que puede dar indicios de un estilo de juego dinámico y posiblemente agresivo, lo que no significa que toda disputa de balón sea sancionada como falta o con una tarjeta amarilla, ya que en el fútbol al ser un deporte de contacto se está expuesto a roces.

Por otra parte, los gráficos de dispersión detallados en la Figura 14 muestra un gráfico de dispersión que representa la relación entre el número total de duelos, los duelos defensivos y los duelos efectivos. Ambos gráficos muestran una amplia dispersión de puntos, lo que sugiere que tener más duelos no necesariamente conduce a ganar un mayor número de duelos ofensivos o defensivos. Este patrón sugiere que la calidad del compromiso en los duelos, que puede depender de la técnica individual, la táctica del equipo y las circunstancias del partido es más crítica que la cantidad de duelos en los que un equipo participa.

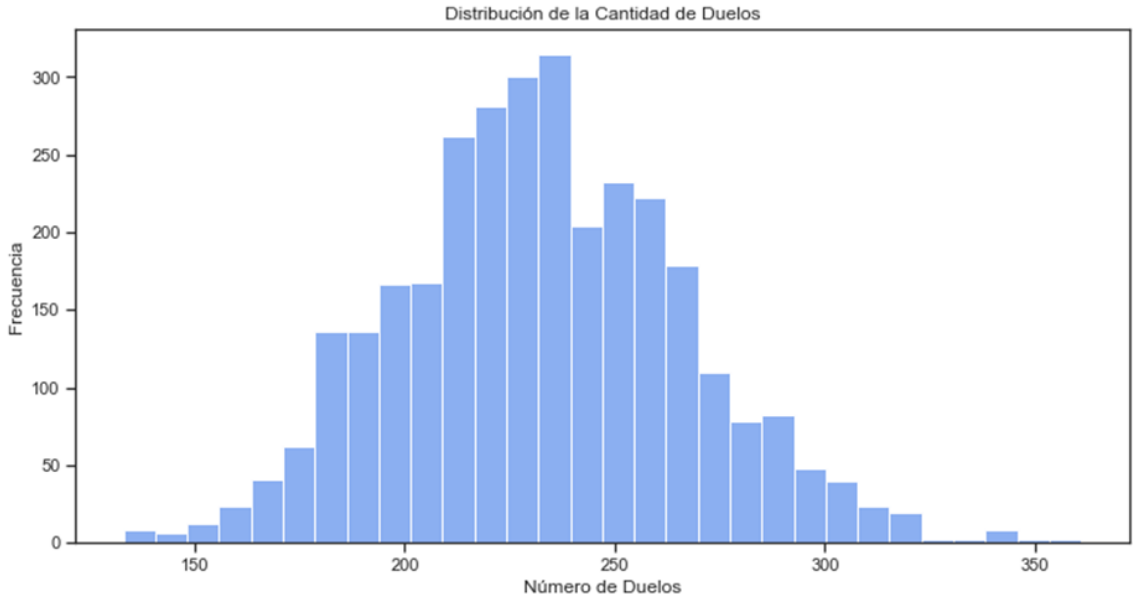



**Figura 13:** Distribución de la cantidad de duelos por equipo en un partido.

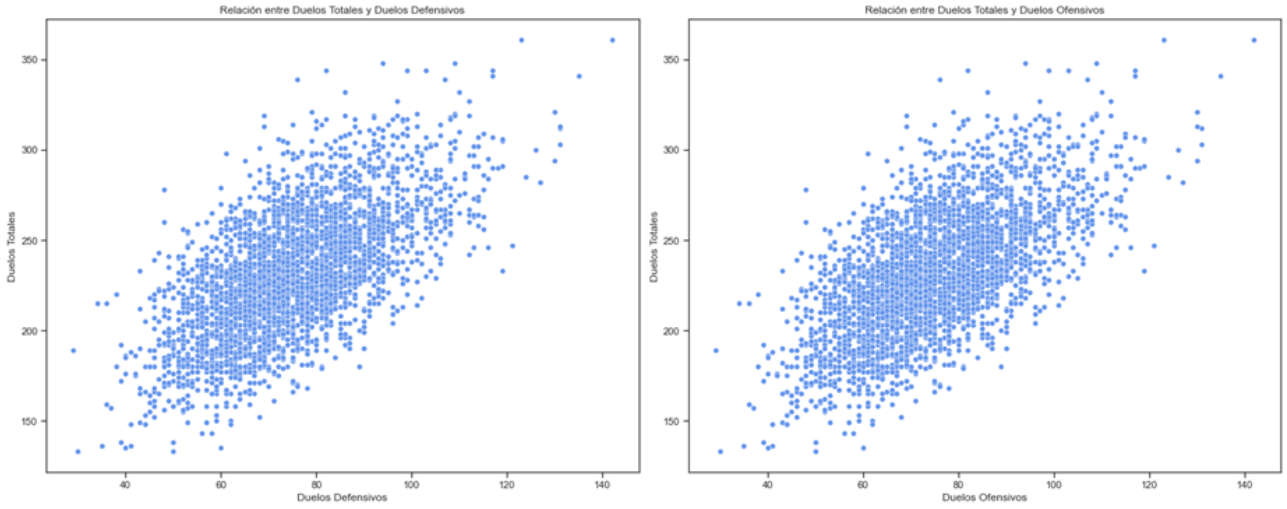



**Figura 14:** Relación entre duelos totales y duelos defensivos y ofensivos.

### 5.1.6. Cantidad de faltas y tarjetas.

El histograma presentado en la Figura 15 representa la distribución de la cantidad de faltas cometidas en los partidos del campeonato nacional de Primera División del fútbol chileno. Este histograma, tiene una forma simétrica y acampanada, la cual sugiere que existe una tendencia general en cuanto a la cantidad de faltas que suelen cometerse en un partido, lo que se traduce en que en la mayoría de los partidos se comenten faltas que están cerca del promedio, siendo 13 el promedio de faltas que se comenten por partido. Por otro lado, las desviaciones de la norma podrían estar influidas por varios factores, como partidos particularmente disputados, diferencias en las tácticas de los equipos, o la influencia de las condiciones del juego.

El gráfico de barras presentado en la Figura 16, detalla que existe un aumento progresivo de la cantidad de tarjetas amarillas esto debido a que desde el año 2020 en adelante se comenzó a hacer uso del VAR (Arbifup, 2020), lo cual cambió en gran medida la forma en la que se sancionan tanto las faltas como la sanción de jugadas confusas en las que el árbitro no determina correctamente quien fue el responsable, incluso esto afecta a jugadas que están fuera de la mirada del balón.

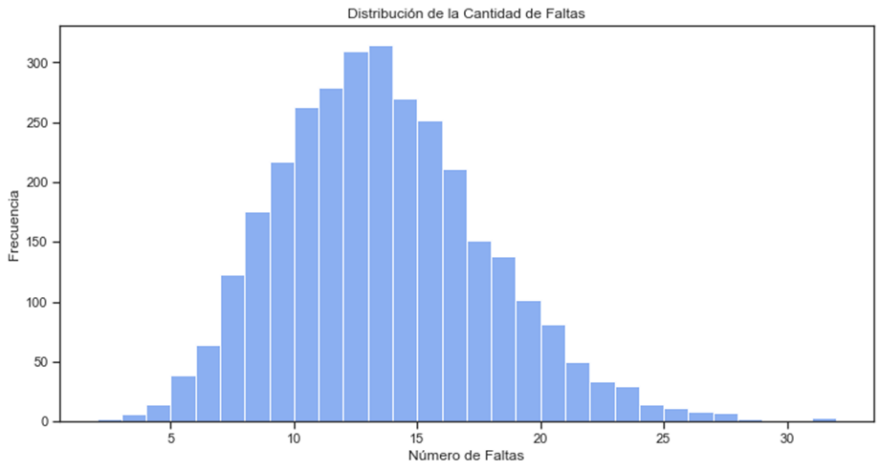



**Figura 15:** Distribución de la cantidad de faltas cometidas por un equipo en un partido.

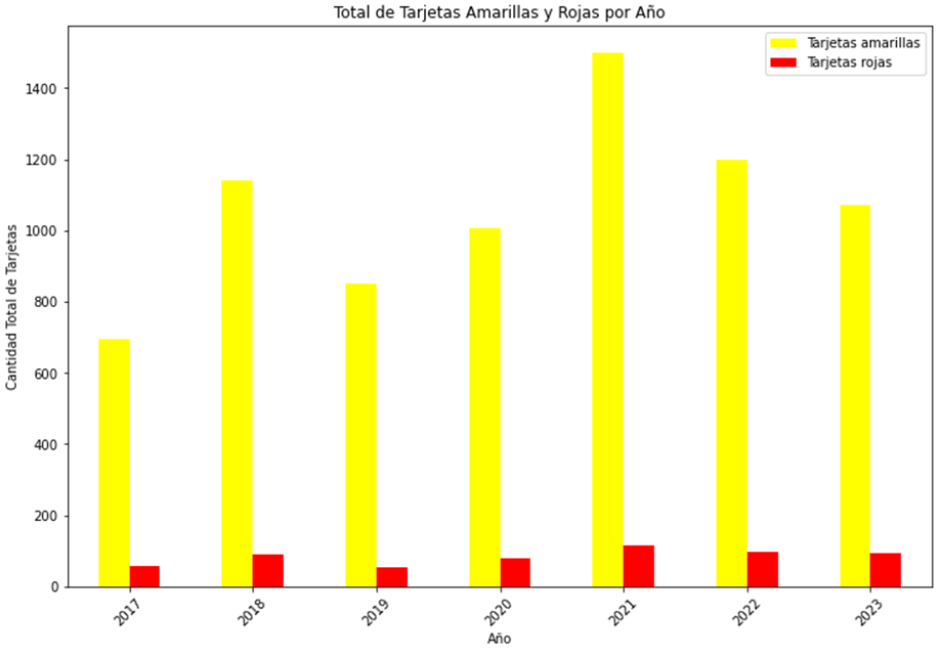



**Figura 16:** Cantidad total de tarjetas amarillas y rojas por año.

## 5.2. Entrenamiento de modelos

A continuación se presentan los resultados de los modelos predictivos entrenados a partir de los datos de los partidos del campeonato chileno de Primera División jugados entre los años 2017 y 2023. Se utilizaron los tres modelos descritos en la metodología: Regresión Lineal, Support Vector Machine y Random Forest, todos ellos entrenados a partir de las 46 características del dataset con el fin de predecir una variable objetivo denominada resultado la cual es un valor binario que es igual a cero cuando el equipo analizado pierde o empata el partido, e igual a uno cuando lo gana. Para entrenar el modelo se utilizaron 2539 registros (80% del conjunto de datos) y para realizar pruebas se utilizaron los restantes 635 registros. Con el fin de comparar, análogamente se realiza el entrenamiento de los mismos tres modelos pero con una variable objetivo de naturaleza multiclase, es decir, con un valor 0 cuando el equipo pierde, 1 cuando empata y 2 cuando gana. Los resultados de dichos modelos se presentan en la Tabla 3, y su implementación es similar a lo descrito en el apartado de metodología.

Cabe destacar que en la regresión logística, los coeficientes o "pesos" se asocian con cada característica y reflejan la contribución relativa de esa característica a la variable objetivo. Los valores negativos o positivos de los coeficientes indican si una característica tiene una influencia negativa o positiva en el resultado. Por su parte el Random Forest no presenta coeficientes sino que la importancia de las características, las cuales no tienen valores negativos. Esta métrica se calcula mediante la disminución en la precisión del modelo cuando se quita una característica específica durante la construcción del árbol. La importancia de las características se normaliza de manera que la suma de todas sea 1, lo que significa que se evalúa la contribución relativa de cada característica en lugar de sus efectos positivos o negativos en la predicción.

### 5.2.1. Modelos de clasificación

#### 5.2.1.1. Regresión Logística

Se entrenó un modelo de regresión logística con penalización L2 y un máximo de 100 iteraciones para converger, ambos valores por defecto de la clase LogisticRegression de Scikit-learn. Una vez entrenado el modelo se obtuvo un arreglo de 46 elementos, correspondientes a los coeficientes o "pesos" del modelo para cada una de las 46 características con las que fue entrenado.

En la Figura 17 se presentan las 5 características que más y menos influyen en determinar si un partido se gana o no. Se aprecia que con diferencia, la acción más relevante en la toma de decisión es el tiro a portería, lo cual es lógico pues es la forma en que se consigue anotar goles que es lo que finalmente determina cuál equipo gana un encuentro. El modelo también establece la importancia de acciones defensivas u orientadas al juego de posesión como lo son las interceptaciones, los despejes y los pases hacia atrás, lo que se condice también con lo observado en el análisis exploratorio en cuanto a la cantidad y precisión de pases, pues, en general los pases hacia atrás son pases

seguros y con una alta tasa de acierto. En lo que respecta a pesos negativos los resultados son los esperados pues el principal es recibir una tarjeta roja, lo que atenta contra las posibilidades de conseguir una victoria pues el equipo queda con un jugador menos. El resto son acciones que se relacionan con la pérdida del balón y duelos, además de pases y tiros errados. Los valores de los coeficientes para las 46 características del modelo se presentan en la gráfica del Anexo 8.4.

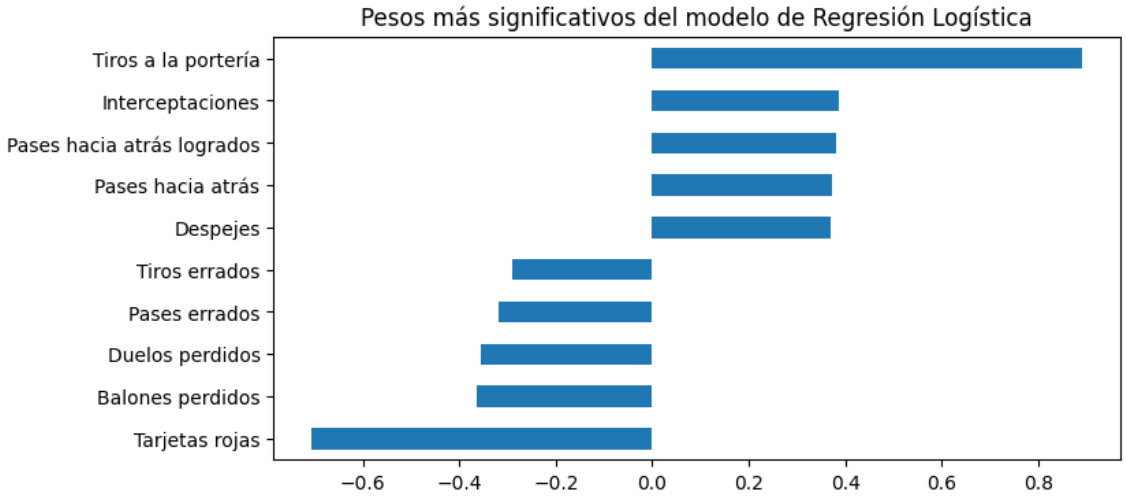



**Figura 17:** Características más relevantes para determinar el resultado de un partido con el modelo de regresión logística.

Finalmente se evaluó el modelo realizando predicciones a partir del conjunto de pruebas (635 registros) y comparando los resultados de las predicciones con los valores reales de dicho conjunto utilizando la función accuracy_score. La precisión del modelo fue de un 76.4% con un recall de 88% para la clase 0 (perder o empatar) y un 54% para la clase 1 (ganar), lo que indica que el modelo es más certero prediciendo cuando un equipo no gana.

### 5.2.1.2. Support Vector Classification

En lo que respecta al uso de Support Vector Machine, se entrenó un modelo utilizando la clase SVC (Support Vector Classification) de Scikit-learn, la cual corresponde a un modelo de clasificación basado en SVM. Se indicó un kernel linear para que el hiperplano que separa las clases fuera una línea recta y se obtuvieron los 46 coeficientes del modelo, siendo muy similares a los obtenidos con la regresión logística. La Figura 18 presenta los pesos más significativos del modelo SVC, destacando que las diferencias con la regresión logística son que en este modelo se valora más las entradas al área de penalti con carreras que la cantidad total de pases para atrás y se penalizan más los pases en el último tercio que la cantidad total de pases errados. Esto último puede atribuirse a que los equipos que realizan muchos pases en el último tercio en general no terminan las jugadas en remate y puede que pierdan la posesión del balón o retrocedan, alejándose de la zona de remate. Los valores de los coeficientes para las 46 características del modelo se presentan en la gráfica del Anexo 8.5.

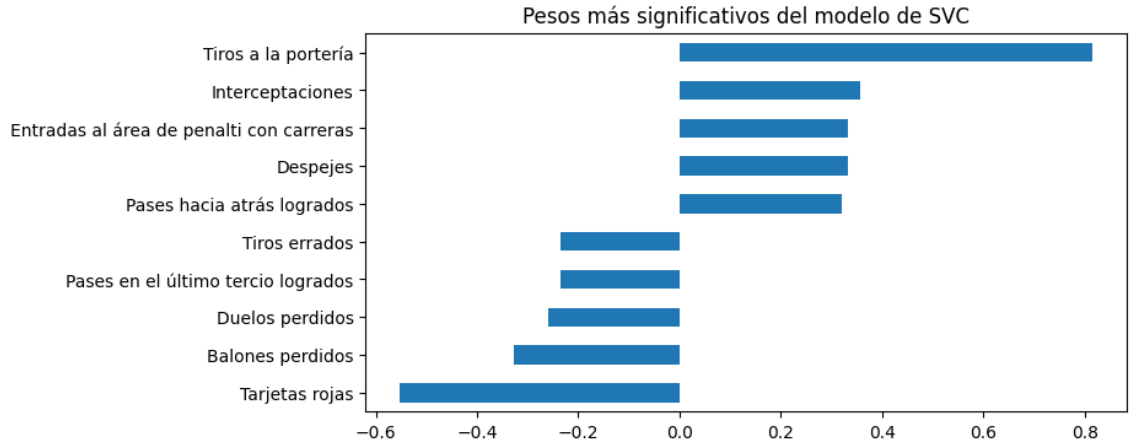



**Figura 18:** Características más relevantes para determinar el resultado de un partido con el modelo de Support Vector Classification.

Con este modelo se realiza la misma evaluación descrita anteriormente, obteniendo una precisión del 76.5%, sólo un 0.1% superior al de regresión logística y con los mismos porcentajes de recall, es decir, 88% para la clase 0 (perder o empatar) y un 54% para la clase 1 (ganar), indicando nuevamente que el modelo es más certero prediciendo cuando un equipo no gana.

#### 5.2.1.3. Random Forest

En el entrenamiento del modelo de Random Forest se utilizó la clase RandomForestClassifier definiendo un número de estimadores igual a 200, lo que significa que el bosque aleatorio está compuesto por 200 árboles de decisión. Como se explicó al inicio de este apartado, el Random Forest no tiene coeficientes, sino que cada una de sus características tiene una importancia que es un valor entre 0 y 1. La Figura 19 presenta las características con importancias más y menos significativas, las cuales en general coinciden con los otros modelos. Destacan entre las menos significativas la aparición de las tarjetas amarillas como factor relevante y los fuera de juego. En las más significativas se mantienen los tiros, despejes e interceptaciones pero aparecen los pases totales en el último tercio, lo que resulta contradictorio con respecto a lo obtenido en el modelo de SVC que catalogaba a los pases logrados en el último tercio como uno de los pesos que afectaba las posibilidades de conseguir una victoria. La explicación puede ser que los equipos que en total intentan más pases son los que juegan en campo contrario y terminan en remates al arco, en cambio los que intentan y completan los pases tienen mayor posesión pero no finalizan la jugada. Los valores de la importancia de las 46 características del modelo se presentan en la gráfica del Anexo 8.6.

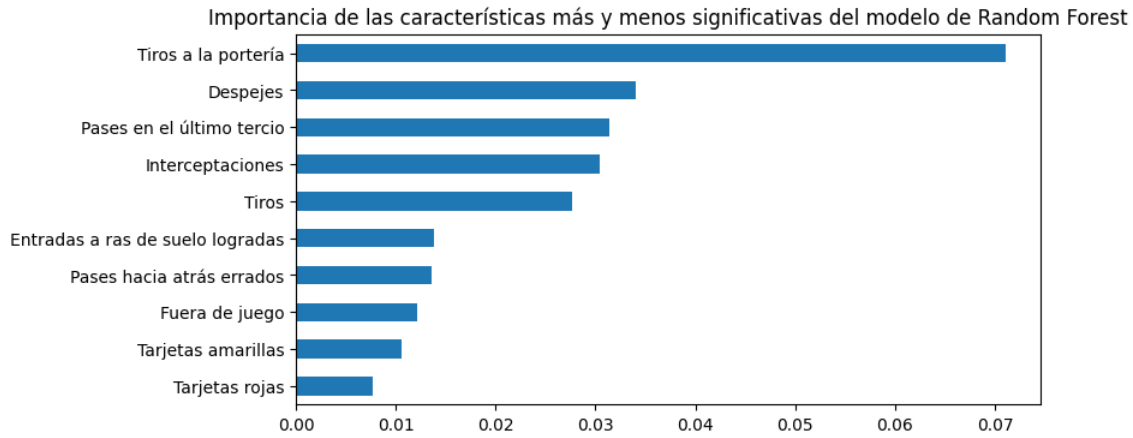



**Figura 19:** Importancia de las características más y menos significativas para determinar el resultado de un partido con el modelo de Random Forest.

La evaluación de este modelo entregó como resultado una precisión del 75.4%, es decir aproximadamente 1% por debajo de la obtenida con los dos modelos anteriores. El recall para la clase 0 (perder o empatar) fue de 92% y para la clase 1 (ganar) de un 44%, ratificando claramente que el modelo es más certero prediciendo cuando un equipo no gana.

### 5.2.2. Comparación de modelos

A continuación se presenta la comparación de porcentajes de accuracy de los modelos implementados. Se presenta también una columna con los resultados de los mismos modelos pero entrenados con valores objetivo multiclase, los cuales tienen un valor igual a 0 si el equipo perdió, igual a 1 si el equipo empató o igual a 2 si el equipo ganó. Se aprecia que la precisión de estos modelos en promedio es de un 55%, muy por debajo de los resultados obtenidos con la variable objetivo binaria, lo que puede producirse por una baja cantidad de datos por clase, en total 1145 para ganado/perdido y 884 para empate de los cuales se utiliza un 80% para entrenamiento y un 20% para pruebas.

| Modelo | Binaria | Multiclase |
|---|---|---|
| Regresión Logística | 0.764 | 0.557 |
| SVC | 0.765 | 0.565 |
| Random Forest | 0.754 | 0.513 |

**Tabla 3:** Tabla comparativa de la precisión de los distintos modelos entrenados.

De la Tabla 3 se puede desprender que el modelo con mejor precisión es el SVC, por lo tanto fue el elegido para el proceso de calificar los rendimientos individuales.

## 5.3. Calificación de rendimientos individuales

A continuación se presenta el análisis de los cuatro partidos seleccionados. Para cada partido se obtuvo la valoración individual de los jugadores de ambos equipos con el fin de contrastar las sumatorias de estos valores con el resultado final del encuentro.

### 5.3.1. Colo-Colo vs Magallanes, Supercopa 2023

Final de la Supercopa de Chile, la cual enfrenta al campeón de la Copa Chile con el campeón del Torneo Nacional. Partido disputado el 15 de enero de 2023 en el Estadio Sausalito de Viña del Mar, enfrentó a Colo-Colo y Magallanes, culminando en una emocionante definición a penales. Colo-Colo abrió el marcador a los 22 minutos con gol de Marcos Bolados, mientras que Felipe Flores igualó para Magallanes a los 26 minutos. El partido duró en total 96 minutos,  en tiempo regular antes de ir a la tanda de penales. Leonardo Gil, volante de Colo-Colo fue expulsado por doble tarjeta amarilla al minuto 91. Brayan Cortés, Maximiliano Falcón y Ramiro González de Colo Colo, así como Felipe Flores y Carlos Villanueva de Magallanes, recibieron tarjetas amarillas. Al final, Magallanes se coronó campeón tras imponerse en la tanda de penales por 4 a 3. Durante el encuentro se contabilizaron 905 pases y ambos equipos registraron 3 disparos a puerta.

Al realizar el análisis de este encuentro, se observó una participación notable de ambos equipos, con Colo-Colo desplegando 14 jugadores y Magallanes 15. La suma total de las valoraciones de Colo-Colo alcanzó el valor de 6.018, con los jugadores titulares sumando 4.850. En contraste, Magallanes acumuló una puntuación total de 7.101, y sus titulares aportaron 5.440 a este total. Las calificaciones individuales de los jugadores titulares de ambos equipos se presentan en la Figura 20.

Este análisis revela que, a nivel individual, los jugadores de Magallanes exhibieron un rendimiento superior al de Colo-Colo. Aunque el partido culminó en un empate y se definió en penales, los datos sugieren que Magallanes mostró una mayor efectividad en términos de rendimiento individual.

Un aspecto crucial en este análisis es la ponderación de diferentes variables. Las tarjetas rojas, por ejemplo, tienen una influencia negativa significativa en la valoración de los jugadores. Esto se evidencia en el caso de Leonardo Gil de Colo-Colo, cuya tarjeta roja lo llevó a obtener una valoración de 0 (siendo el peor evaluado en una escala de 0 a 1), no contribuyendo al total del equipo. En contraposición, el jugador más destacado de Magallanes no fue Felipe Flores, autor del gol del empate, sino Manuel Vicuña. Este jugador, con 1 tiro a portería, 10 pases exitosos hacia atrás y 6 intervenciones, obtuvo una calificación individual de 1 (máximo valor del conjunto), demostrando su contribución crucial al rendimiento del equipo.

Este enfoque detallado resalta la importancia de analizar individualmente el desempeño de los jugadores y cómo las acciones específicas dentro del campo influyen significativamente en las valoraciones finales, ofreciendo una comprensión más profunda del impacto de cada jugador en el resultado general del equipo.


**Figura 20:** Calificaciones individuales de las oncenas titulares de Colo-Colo y Magallanes en la Supercopa 2023.

### 5.3.2. Unión Española vs Universidad Católica, Clásico de Independencia 2022

En los últimos años los partidos entre Unión Española y Universidad Católica han ido adquiriendo un carácter de clásico. Son partidos entre rivales directos que compiten por ser campeón y que han sido generalmente partidos muy disputados en cuanto a la táctica y la intensidad del juego. Este partido ha sido coloquialmente bautizado como el Clásico de Independencia, dada la ubicación de los estadios de ambos clubes en dicha comuna de Santiago entre los años 1940 y 1960.

El partido jugado en el Estadio Santa Laura el 19 de octubre de 2022, fue un encuentro futbolístico que duró 98 minutos. Fernando Zampedri abrió la cuenta para la Universidad Católica a los 23’ del primer tiempo, luego a los 52’ Unión Española logra empatar el partido con tanto de Leandro Garate, y finalmente a los 76’ fue el jugador argentino Yamil Asad el que marcó el tanto que le dio la victoria al conjunto cruzado. En el aspecto disciplinario, Miguel Pinto e Ignacio Jara de Unión Española recibieron tarjetas amarillas, y Manuel Fernández fue expulsado por doble tarjeta amarilla a los 66’. Del lado de Universidad Católica, Clemente Montes, Luciano Aued y Matías Dituro vieron tarjetas amarillas, mientras que Raimundo Rebolledo fue expulsado con roja directa a los 63’. El partido contó con un total de 658 pases y en lo que respecta a tiros a la portería, Unión Española registró 6 tiros al arco, mientras que Universidad Católica 3.

El análisis de desempeño individual de los jugadores juega un rol crucial en la interpretación de los resultados finales. En este sentido, Universidad Católica demostró una superioridad cuantificable sobre Unión Española, con un total acumulado de 9.149, en contraste con los 8.828 de su contrincante. Esta discrepancia se hace más evidente al considerar únicamente a los jugadores titulares, donde Universidad Católica suma 7.180, superando significativamente los 5.800 de Unión Española. Las valoraciones individuales de las oncenas titulares de Unión Española y Universidad Católica se presentan en la Figura 21.

Al profundizar en las actuaciones individuales, se destaca la figura de Fernando Zampedri de Universidad Católica, quien alcanza la cima del ranking con una valoración perfecta de 1 (en una escala de 0 a 1). Su aporte incluye 2 tiros a portería, 3 pases exitosos hacia atrás y 3 intervenciones clave. Por otro lado, Raimundo Rebolledo, a pesar de su participación activa, se sitúa en el extremo inferior del espectro con un ranking de 0.44, atribuible a una tarjeta roja y la pérdida de 8 duelos y 7 balones.

En el campo de Unión Española, la situación es variada. Mario Larenas registra el ranking más bajo con 0, a causa de 17 balones perdidos y 9 duelos fallidos. En contraste,

Leonardo Garate emerge como el jugador más destacado, con una valoración de 1, gracias a sus 2 tiros a portería y 10 pases exitosos hacia atrás.

Es crucial resaltar la importancia de cada variable en la valoración final. Las tarjetas rojas, por ejemplo, ejercen un impacto negativo considerable, como se observa en el caso de Rebolledo. Asimismo, otros factores, como la pérdida de balones y duelos, pueden influir significativamente en la valoración, como se evidencia en el caso de Mario Larenas. Estos elementos resaltan cómo ciertas acciones en el campo de juego tienen una repercusión notable en las valoraciones individuales.

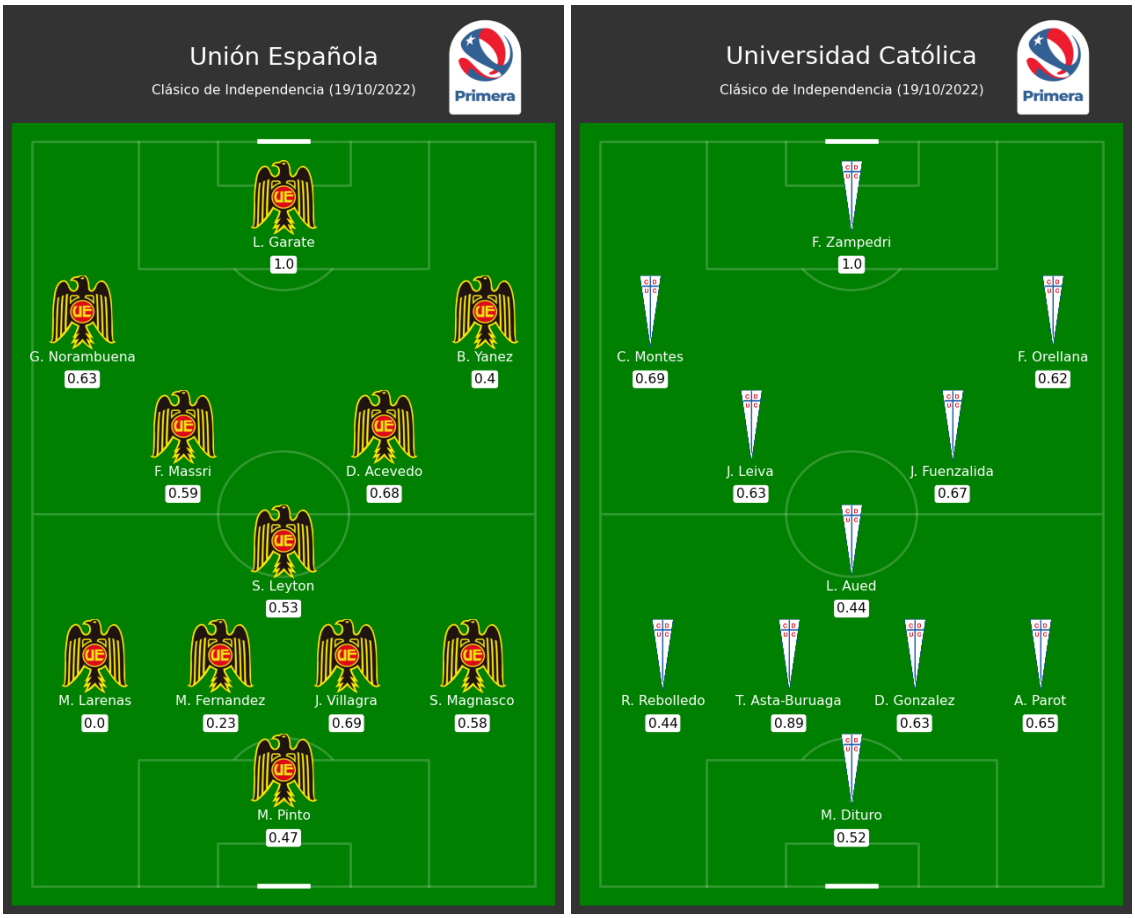



**Figura 21:** Calificaciones individuales de las oncenas titulares de Unión Española y Universidad Católica en el Clásico de Independencia 2022.

### 5.3.3. Colo-Colo vs Universidad de Chile, Superclásico 2019

"Superclásico" del fútbol chileno, jugado el 5 de octubre de 2019 entre Colo-Colo y Universidad de Chile en el Estadio Monumental David Arellano. Este encuentro destacó por el gol 216 de Esteban Paredes, que lo convirtió en el máximo goleador de la historia de la Primera División de Chile. El partido, que duró 96 minutos, finalizó con una emocionante victoria para Colo-Colo por 3 goles a 2. Los goles de Colo-Colo fueron anotados por Gabriel Suazo a los 49', Esteban Paredes a los 64', y Julio Barroso a los 93', mientras que para la Universidad de Chile marcaron Gonzalo Espinoza a los 11' mediante lanzamiento penal y Ángelo Henríquez a los 75'. Camilo Moya de la Universidad de Chile fue expulsado por doble tarjeta amarilla a los 85', mientras que Leandro Benegas y Gonzalo Espinoza de la Universidad de Chile, y Branco Provoste e Ivan Rossi de Colo-Colo, fueron amonestados con tarjeta amarilla. El partido contó con un total 714 pases, y para Colo-Colo se registraron 6 remates al arco, mientras que para la Universidad de Chile se registraron 5 remates a la portería.

Al realizar el análisis de este encuentro deportivo se puede observar que Colo-Colo utilizó 14 jugadores, acumulando una valoración total de 7.592 de los cuales los jugadores titulares aportaron 6.250. Por su parte, Universidad de Chile también presentó 14 jugadores, con una sumatoria total de 6.884 y una contribución de 5.450 por parte de los jugadores titulares. Estas estadísticas revelan que Colo-Colo superó a Universidad de Chile en rendimiento individual, tanto en la alineación total como en la de los titulares, reflejándose esto en el resultado final del partido. La valoración individual de las oncenas titulares de ambos equipos se presenta en la Figura 22.

Entre los jugadores destacados de Colo-Colo, se encuentran Pablo Mouche y Esteban Paredes. Pablo Mouche con una valoración de 0.85 aportó 1 tiro a portería, 10 pases logrados hacia atrás y 2 intercepciones, mientras que Esteban Paredes, con una valoración de 0.79 contribuyó con 2 tiros a portería, 2 pases logrados hacia atrás y 2

intercepciones. Estas actuaciones ejemplifican cómo acciones específicas en el campo pueden influir notablemente en las valoraciones de los jugadores y, por consiguiente, en el desempeño global del equipo.

En contraste, Camilo Moya de Universidad de Chile tuvo un rendimiento desfavorable, reflejado en su valoración de 0 (en una escala de 0 a 1). Esta baja puntuación se debió a la acumulación de 2 tarjetas amarillas, posteriormente roja y la pérdida de 10 balones. Esto ilustra cómo las acciones negativas, en particular las tarjetas, tienen un impacto negativo significativo en el modelo de valoración y pueden afectar profundamente las valoraciones individuales y, por lo tanto, la sumatoria total del equipo.


**Figura 22:** Calificaciones individuales de las oncenas titulares de Colo-Colo y Universidad de Chile en el Superclásico 2019.

### 5.3.4. Universidad Católica vs Colo-Colo, Supercopa 2022

Final de la Supercopa de Chile 2022, partido que se llevó a cabo el 23 de enero de 2022 en el Estadio Alcaldesa Ester Roa Rebolledo de Concepción, enfrentando a Universidad Católica (campeón Primera División 2021) y Colo-Colo (campeón Copa Chile 2021). Colo-Colo dominó el partido desde el inicio, marcando el primer gol a los 18 minutos por medio de Gabriel Costa. La intensidad del juego se mantuvo alta, con Universidad Católica buscando el empate, pero sin éxito. En la segunda mitad, a los 67 minutos, Esteban Pavez aumentó la ventaja para Colo-Colo con un potente disparo. A pesar de los esfuerzos de Universidad Católica, el partido concluyó con un marcador de 2-0 a favor de Colo-Colo. Durante el encuentro, 9 jugadores recibieron tarjeta amarilla, 4 para Colo-Colo y 5 para Universidad Católica, se realizaron un total de 662 pases y 6 remates a puerta, de los cuales Colo-Colo realizó 5.

En este encuentro Colo-Colo utilizó un total de 14 jugadores (11 jugadores titulares más 3 cambios), que en conjunto lograron una valoración total de 5.004 mientras que los jugadores titulares sumaron 4.110 en sus valoraciones. Por su parte, Universidad Católica, que también utilizó a 14 jugadores, alcanzó una sumatoria total de 2.986 mientras que los titulares obtuvieron una sumatoria de 2.380. Estas cifras reflejan un rendimiento individual superior de los jugadores de Colo-Colo, tanto en la alineación general como entre los titulares, además del escaso aporte de los suplentes que ingresaron a la cancha. La valoración individual de las oncenas titulares de ambos equipos se presenta en la Figura 23.

Un elemento destacado del partido fue la actuación dispar de los delanteros de ambos equipos. Juan Martín Lucero de Colo-Colo exhibió un rendimiento sobresaliente, con una valoración perfecta de 1. Su desempeño incluyó 2 tiros a portería, 9 pases logrados hacia atrás, 4 Interceptaciones y 2 despejes, manteniéndose libre de tarjetas amarillas.

Por su parte, Fernando Zampedri de Universidad Católica tuvo un rendimiento discreto, con una valoración de 0.21. A pesar de sus esfuerzos, incluyendo 2 pases logrados hacia atrás, 1 interceptación y 1 despeje, el recibir una tarjeta amarilla y la ausencia de tiros a portería influyeron negativamente en su valoración.

Esta disparidad en el rendimiento de los delanteros es ilustrativa de la dinámica general del partido. Mientras Juan Martin Lucero jugó un papel clave en el éxito de su equipo, la actuación menos determinante de Fernando Zampedri y otros jugadores de Universidad Católica se vio reflejada en la baja sumatoria total de valoraciones.

Es importante destacar el cómo determinadas acciones y estadísticas (como los tiros a portería, las tarjetas y las intercepciones) impactan significativamente en la valoración de un jugador. En este caso, la ausencia de tarjetas amarillas y el aporte tanto ofensivo como defensivo de Juan Martin Lucero fueron determinantes para su alta valoración, mientras que la tarjeta amarilla de Fernando Zampedri tuvo un efecto negativo en su puntuación.


**Figura 23:** Calificaciones individuales de las oncenas titulares de Universidad Católica y Colo-Colo en la Supercopa 2022.

# 6. Conclusiones

Este trabajo se enfoca en implementar un modelo de valoración de rendimiento individual de jugadores en partidos de la Primera División del fútbol chileno basado en el framework Playerank de Pappalardo, utilizando técnicas estadísticas y modelos de machine learning, como lo son la Regresión Logística, SVM y Random Forest. La implementación de los modelos permitió obtener la valoración individual de los jugadores que disputaron un partido de fútbol, con las cuales luego se hizo una sumatoria para contrastar si esta sumatoria por equipo está directamente correlacionada con el resultado final del partido, en otras palabras, un equipo ganará un encuentro si la suma de las valoraciones individuales de sus jugadores supera a la del rival.

Una vez realizado el trabajo de limpieza, exploración y escalado de los datos, se implementaron modelos de clasificación que consideran 46 variables de interés y que tienen cerca de un 77% de precisión. Al entrenar los modelos, las 46 variables tienen un coeficiente de "peso" o una importancia, los cuales se utilizaron para ponderar las acciones individuales de los futbolistas durante un encuentro y con ello calificar su rendimiento. Algunas de las variables más significativas fueron tiros a la portería, despejes, balones perdidos y tarjetas rojas.

Este trabajo tuvo resultados interesantes, ya que se aprecia que la sumatoria individual de cada jugador en este caso, si afecta en el resultado del partido, ejemplo de esto se puede ver en el partido disputado entre Colo-Colo y Universidad de Chile, partido que ganó Colo-Colo por 3 goles a 2, siendo un factor importante en el resultado la tarjeta roja que recibió Camilo Moya a los 85' (lo que genera que su calificación sea de 0), lo que dejó a Universidad de Chile con 10 jugadores en cancha, permitiendo que los albos marcaron el 3 a 2 a los 93'.

Los modelos de machine learning implementados demostraron una capacidad notable para capturar la esencia de lo que contribuye a una victoria en el fútbol, reflejando con

precisión las valoraciones de los jugadores y su impacto en el resultado del partido. Este enfoque innovador resalta la importancia y la aplicabilidad de la ciencia de datos en el análisis deportivo, proporcionando conocimiento valioso que puede ser utilizado por entrenadores, cazatalentos y directivos de clubes para optimizar estrategias y tomar decisiones informadas.

Como trabajo futuro se propone:

- **Análisis de Copa Chile:** Podría implementarse la misma metodología presentada en este trabajo pero utilizando los datos de los partidos correspondientes a la Copa Chile, campeonato que disputan equipos de variadas divisiones del país y que entrega como premios un cupo en Copa Libertadores y la posibilidad de jugar la Supercopa ante el campeón de Primera División. Este torneo se caracteriza por disputarse en llaves de eliminación directa con partidos de ida y vuelta, por lo que probablemente existirán diferencias en el estilo de juego de los equipos.

- **Proyección de jugadores sub-21:** Sería interesante realizar este mismo análisis para jugadores chilenos con proyección, con el fin de realizar un seguimiento y poder potenciarlos para que eventualmente puedan jugar en la selección chilena adulta, algo similar a lo realizado por José Sulantay en el año 2007, donde en el Mundial Sub 20 de Canadá obtuvo el tercer lugar, gracias a la participación de jugadores como Mauricio Isla, Gary Medel, Arturo Vidal y Alexis Sánchez.

- **Expansión a otras Ligas:** Ampliar el análisis para incluir otras ligas latinoamericanas y comparar sus resultados con la liga chilena. Esto permitiría saber cómo se encuentra la liga chilena en comparación a otras ligas de la región, lo que ayudaría a explicar el porqué los equipos chilenos no avanzan en copas internacionales.

Este trabajo se centró en el campeonato nacional de Primera División, lo cual limita la generalización de los hallazgos a otras ligas. Además, la metodología se basó en la sumatoria de valoraciones por partido en lugar de por temporada, lo que podría afectar la precisión del modelo y de la calificación en ciertas circunstancias. También se puede mencionar como limitación la cantidad de archivos que demandaría un análisis en profundidad y valoración de los jugadores de todos los partidos de un campeonato o de la historia de la Primera División, lo que sería un trabajo que tomaría tiempo (producto de la necesidad de descargar los archivos de manera manual) pero que podría aportar más información para validar la hipótesis planteada en este trabajo. A pesar de estas limitaciones, los resultados obtenidos son consistentes y reflejan fielmente la realidad del fútbol chileno, lo que valida la efectividad del enfoque utilizado.

# 7. Bibliografía


- Pappalardo, L., Cintia, P., Ferragina, P., Massucco, E., Pedreschi, D., & Giannotti, F. (2019). PlayeRank. *ACM Transactions on Intelligent Systems and Technology*, *10*(5), 1-27. https://doi.org/10.1145/3343172
- Harrington, A., Marín, W. (2022). *Football Analytics - Ranking de jugadores*. https://repositorio.udd.cl/items/bcce9e2a-fc8b-499e-9764-1c9cb113b4d4
- Nina, M. (2022). Aproximación a determinantes sobre el resultado de un partido de fútbol y ranking de jugadores del Campeonato Nacional Chileno: predicción, clasificación y ranking. https://repositorio.udd.cl/items/13da7619-cb32-4e33-ad5c-718877b06057
- Becerra, J. (2022). *¿El 11 ideal?: un enfoque data - driven para la conformación de un plantel.* https://repositorio.udd.cl/items/5659ab91-6885-4c60-b6dc-b7a67e898c57
- Infobae. (2022, 25 octubre). Qatar 2022: Cuál será el relevante rol de los científicos de datos. Infobae. https://www.infobae.com/inhouse/2022/10/25/qatar-2022-cual-sera-el-relevante-rol-de-los-cientificos-de-datos/
- Dufour, M., Phillips, J., & Ernwein, V. (2017). What makes the difference? Analysis of the 2014 World Cup. Journal of Human Sport and Exercise, 12(3). https://doi.org/10.14198/jhse.2017.123.06
- Brooks, J. D., Kerr, M., & Guttag, J. V. (2016). Developing a Data-Driven Player Ranking in Soccer Using Predictive Model Weights. the 22nd ACM SIGKDD International Conference. https://doi.org/10.1145/2939672.2939695
- Schulte, O., Zhao, Z., & Javan, M. (2017). Apples-to-Apples : Clustering and Ranking NHL Players Using Location Information and Scoring Impact.
- Baumer, B., & Zimbalist, A. (2014). Quantifying Market Inefficiencies in the Baseball Players' Market. *Eastern Economic Journal*, *40*(4), 488–498. http://www.jstor.org/stable/24693687

- Iturrieta Ortiz, J. (2018).El rol socio cultural del fútbol en el Chile de la segunda mitad del siglo XX: el caso de la campaña de Colo Colo en la Copa Libertadores de 1973. Disponible en https://repositorio.uchile.cl/handle/2250/168752
- Tobar, D. (2019, 22 noviembre). La cronología de la mañana más difícil del fútbol chileno. Diario AS. https://chile.as.com/chile/2019/11/22/futbol/1574430214_000799.html
- As.com. (2020, 16 febrero). Colo Colo 0 - U. Católica 2: Goles, resumen y resultado, clásico suspendido. Diario AS. https://chile.as.com/chile/2020/02/16/futbol/1581878196_587098.html
- TNT Sports. (2022, 16 octubre). Partido entre Deportes Antofagasta y Palestino fue suspendido. TNT Sports. https://tntsports.cl/nacional/Partido-entre-Deportes-Antofagasta-y-Palestino-fue-suspendido--20221015-0011.html
- FIFA. (2022, 19 noviembre). FIFA unveils Technical Study Group for FIFA World Cup Qatar 2022TM. (s. f.). https://www.fifa.com/technical/media-releases/fifa-unveils-technical-study-group-for-fifa-world-cup-qatar-2022-tm
- Águila, I. (2023, 3 marzo). Más tiempo efectivo, modificación al offside y regla anti-Dibu: los cambios que planea la IFAB para revolucionar el fútbol. La Tercera. https://www.latercera.com/el-deportivo/noticia/mas-tiempo-efectivo-modificacion-al-offside-y-regla-anti-dibu-los-cambios-que-planea-la-ifab-para-revolucionar-el-futbol/5GNSHF3XLJF3PFEIJ4EGDLVT7I/
- Arbifup. (2020, 8 junio). 10 claves del VAR y su aplicación en el fútbol chileno. https://arbifup.cl/10-claves-del-var-y-su-aplicacion-en-el-futbol-chileno/

# 8. Anexos

## 8.1. Listado de equipos por temporadas

| Nº | Equipo | Temporadas |
|---|---|---|
| 1 | Colo-Colo | 2017, 2018, 2019, 2020, 2021, 2022, 2023 |
| 2 | Universidad de Chile | 2017, 2018, 2019, 2020, 2021, 2022, 2023 |
| 3 | Universidad Católica | 2017, 2018, 2019, 2020, 2021, 2022, 2023 |
| 4 | Unión Española | 2017, 2018, 2019, 2020, 2021, 2022, 2023 |
| 5 | Audax Italiano | 2017, 2018, 2019, 2020, 2021, 2022, 2023 |
| 6 | Curicó Unido | 2017, 2018, 2019, 2020, 2021, 2022, 2023 |
| 7 | Everton | 2017, 2018, 2019, 2020, 2021, 2022, 2023 |
| 8 | Huachipato | 2017, 2018, 2019, 2020, 2021, 2022, 2023 |
| 9 | O'Higgins | 2017, 2018, 2019, 2020, 2021, 2022, 2023 |
| 10 | Palestino | 2017, 2018, 2019, 2020, 2021, 2022, 2023 |
| 11 | Antofagasta | 2017, 2018, 2019, 2020, 2021, 2022 |
| 12 | Unión La Calera | 2018, 2019, 2020, 2021, 2022, 2023 |
| 13 | Cobresal | 2019, 2020, 2021, 2022, 2023 |
| 14 | Deportes Iquique | 2017, 2018, 2019, 2020 |
| 15 | Universidad de Concepción | 2017, 2018, 2019, 2020 |
| 16 | Coquimbo Unido | 2019, 2020, 2022, 2023 |
| 17 | Santiago Wanderers | 2017, 2020, 2021 |
| 18 | La Serena | 2020, 2021, 2022 |
| 19 | Ñublense | 2021, 2022, 2023 |
| 20 | Deportes Temuco | 2017, 2018 |
| 21 | San Luis de Quillota | 2017, 2018 |
| 22 | Melipilla | 2021 |
| 23 | Copiapó | 2023 |
| 24 | Magallanes | 2023 |

## 8.2. Listado de jugadores

| Nº | Jugador | Equipo |
|---|---|---|
| 1 | Diego Tapia | Magallanes |
| 2 | Felipe Espinoza | Magallanes |
| 3 | Fernando Piñero | Magallanes |
| 4 | Albert Acevedo | Magallanes |
| 5 | Matías Vásquez | Magallanes |
| 6 | Iván Vásquez | Magallanes |
| 7 | César Cortés | Magallanes |
| 8 | Tomás Aránguiz | Magallanes |
| 9 | Yorman Zapata | Magallanes |
| 10 | Felipe Flores | Magallanes |
| 11 | Julián Alfaro | Magallanes |
| 12 | Carlos Villanueva | Magallanes |
| 13 | Nicolás Nuñez | Magallanes |
| 14 | Rodrigo Díaz | Magallanes |
| 15 | Manuel Vicuña | Magallanes |
| 16 | Thomas Jones | Magallanes |
| 17 | Gastón Rodríguez | Magallanes |
| 18 | Marcelo Filla | Magallanes |
| 19 | Alfred Canales | Magallanes |
| 20 | Javier Quiroz | Magallanes |
| 21 | Miguel Pinto | Unión Española |
| 22 | Jonathan Villagra | Unión Española |
| 23 | Manuel Fernández | Unión Española |
| 24 | Thomas Galdames | Unión Española |
| 25 | Mario Larenas | Unión Española |
| 26 | Felipe Massri | Unión Española |
| 27 | Gonzalo Espinoza | Unión Española |

| | | |
|---|---|---|
| 28 | Rodrigo Piñeiro | Unión Española |
| 29 | Bryan Rabello | Unión Española |
| 30 | Bastián Yañez | Unión Española |
| 31 | Leandro Garate | Unión Española |
| 32 | Augusto Barrios | Unión Española |
| 33 | Diego Acevedo | Unión Española |
| 34 | Vicente Conelli | Unión Española |
| 35 | Sebastián Leyton | Unión Española |
| 36 | Stefano Magnasco | Unión Española |
| 37 | Gabriel Norambuena | Unión Española |
| 38 | Ignacio Jara | Unión Española |
| 39 | Ignacio Ibañez | Unión Española |
| 40 | Bryan Cortés | Colo-Colo |
| 41 | Bruno Gutiérrez | Colo-Colo |
| 42 | Maximiliano Falcón | Colo-Colo |
| 43 | Ramiro González | Colo-Colo |
| 44 | Erick Wiemberg | Colo-Colo |
| 45 | César Fuentes | Colo-Colo |
| 46 | Esteban Pavez | Colo-Colo |
| 47 | Leonardo Gil | Colo-Colo |
| 48 | Marco Rojas | Colo-Colo |
| 49 | Marcos Bolados | Colo-Colo |
| 50 | Agustin Bouzat | Colo-Colo |
| 51 | Matías Moya | Colo-Colo |
| 52 | Jordhy Thompson | Colo-Colo |
| 53 | Leandro Benegas | Colo-Colo |
| 54 | Óscar Opazo | Colo-Colo |
| 55 | Emiliano Amor | Colo-Colo |
| 56 | Gabriel Suazo | Colo-Colo |
| 57 | Pablo Solari | Colo-Colo |

| | | |
|---|---|---|
| 58 | Gabriel Costa | Colo-Colo |
| 59 | Juan Martín Lucero | Colo-Colo |
| 60 | Jeyson Rojas | Colo-Colo |
| 61 | Matías Zaldivia | Colo-Colo |
| 62 | Carlo Villanueva | Colo-Colo |
| 63 | Christian Santos | Colo-Colo |
| 64 | Felipe Campos | Colo-Colo |
| 65 | Julio Barroso | Colo-Colo |
| 66 | Juan Manuel Insaurralde | Colo-Colo |
| 67 | Branco Provoste | Colo-Colo |
| 68 | Iván Rossi | Colo-Colo |
| 69 | Esteban Paredes | Colo-Colo |
| 70 | Pablo Mouche | Colo-Colo |
| 71 | Jaime Valdés | Colo-Colo |
| 72 | Javier Parraguez | Colo-Colo |
| 73 | Hernán Galíndez | Universidad de Chile |
| 74 | Yonathan Andía | Universidad de Chile |
| 75 | Bastián Tapia | Universidad de Chile |
| 76 | Ignacio Tapia | Universidad de Chile |
| 77 | Marcelo Morales | Universidad de Chile |
| 78 | Felipe Seymour | Universidad de Chile |
| 79 | Álvaro Brun | Universidad de Chile |
| 80 | Darío Osorio | Universidad de Chile |
| 81 | Junior Fernandes | Universidad de Chile |
| 82 | Cristián Palacios | Universidad de Chile |
| 83 | Ronnie Fernández | Universidad de Chile |
| 84 | Jeisson Vargas | Universidad de Chile |
| 85 | Luis Felipe Gallegos | Universidad de Chile |
| 86 | Pablo Aránguiz | Universidad de Chile |
| 87 | Camilo Moya | Universidad de Chile |

| | | |
|---|---|---|
| 88 | José María Carrasco | Universidad de Chile |
| 89 | Israel Poblete | Universidad de Chile |
| 90 | Lucas Assadi | Universidad de Chile |
| 91 | José Castro | Universidad de Chile |
| 92 | Cristóbal Muñoz | Universidad de Chile |
| 93 | Fernando De Paul | Universidad de Chile |
| 94 | Matías Rodríguez | Universidad de Chile |
| 95 | Osvaldo González | Universidad de Chile |
| 96 | Lucas Aveldaño | Universidad de Chile |
| 97 | Jean Beausejour | Universidad de Chile |
| 98 | Nicolás Oroz | Universidad de Chile |
| 99 | Leonardo Fernández | Universidad de Chile |
| 100 | Nicolás Guerra | Universidad de Chile |
| 101 | Ángelo Henríquez | Universidad de Chile |
| 102 | Rafael Caroca | Universidad de Chile |
| 103 | Rodrigo Echeverría | Universidad de Chile |
| 104 | Sebastián Pérez | Universidad Católica |
| 105 | Branco Ampuero | Universidad Católica |
| 106 | Nehuén Paz | Universidad Católica |
| 107 | Alfonso Parot | Universidad Católica |
| 108 | Raimundo Rebolledo | Universidad Católica |
| 109 | Felipe Gutiérrez | Universidad Católica |
| 110 | Juan Leiva | Universidad Católica |
| 111 | Cristian Cuevas | Universidad Católica |
| 112 | Diego Buonanotte | Universidad Católica |
| 113 | Gonzalo Tapia | Universidad Católica |
| 114 | Fernando Zampedri | Universidad Católica |
| 115 | Marcelino Nuñez | Universidad Católica |
| 116 | Lucas Melano | Universidad Católica |
| 117 | Yamil Assad | Universidad Católica |

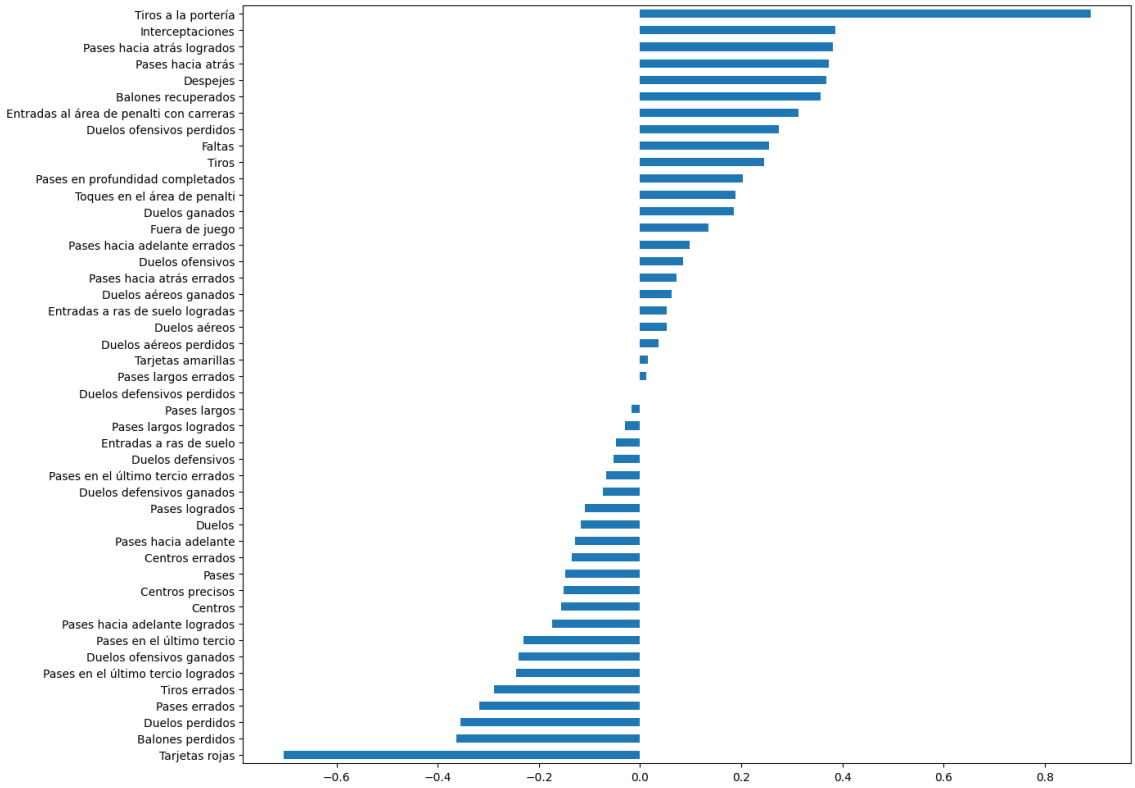
| | | |
|---|---|---|
| 118 | José Pedro Fuenzalida | Universidad Católica |
| 119 | Tomás Asta-Buruaga | Universidad Católica |
| 120 | Matías Dituro | Universidad Católica |
| 121 | Daniel González | Universidad Católica |
| 122 | Luciano Aued | Universidad Católica |
| 123 | Clemente Montes | Universidad Católica |
| 124 | Fabián Orellana | Universidad Católica |
| 125 | Germán Lanaro | Universidad Católica |
| 126 | Aaron Astudillo | Universidad Católica |
| 127 | Valber Huerta | Universidad Católica |
| 128 | Sebastián Galani | Universidad Católica |
| 129 | Diego Valencia | Universidad Católica |

## 8.3. Listado de características de los modelos

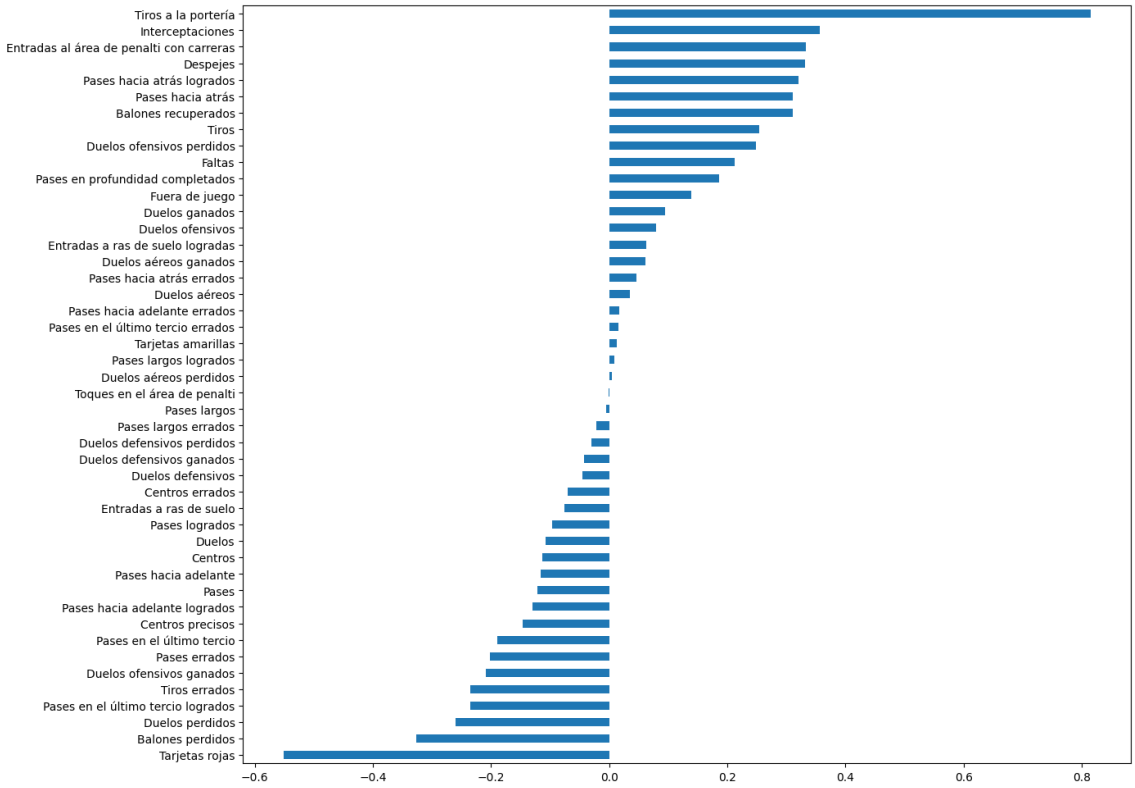
| Nº | Características de los modelos |
|---|---|
| 1 | Pases |
| 2 | Pases logrados |
| 3 | Pases errados |
| 4 | Pases largos |
| 5 | Pases largos logrados |
| 6 | Pases largos errados |
| 7 | Pases en el último tercio |
| 8 | Pases en el último tercio logrados |
| 9 | Pases en el último tercio errados |
| 10 | Pases hacia adelante |
| 11 | Pases hacia adelante logrados |
| 12 | Pases hacia adelante errados |
| 13 | Pases hacia atrás |
| 14 | Pases hacia atrás logrados |
| 15 | Pases hacia atrás errados |
| 16 | Pases en profundidad completados |
| 17 | Toques en el área de penalti |
| 18 | Centros |
| 19 | Centros precisos |
| 20 | Centros errados |
| 21 | Despejes |
| 22 | Tiros |
| 23 | Tiros a la portería |
| 24 | Tiros errados |
| 25 | Duelos |
| 26 | Duelos ganados |
| 27 | Duelos perdidos |
| 28 | Duelos aéreos |

| | |
|---|---|
| 29 | Duelos aéreos ganados |
| 30 | Duelos aéreos perdidos |
| 31 | Interceptaciones |
| 32 | Duelos defensivos |
| 33 | Duelos defensivos ganados |
| 34 | Duelos defensivos perdidos |
| 35 | Entradas a ras de suelo |
| 36 | Entradas a ras de suelo logradas |
| 37 | Faltas |
| 38 | Duelos ofensivos |
| 39 | Duelos ofensivos ganados |
| 40 | Duelos ofensivos perdidos |
| 41 | Balones perdidos |
| 42 | Balones recuperados |
| 43 | Entradas al área de penalti carreras |
| 44 | Fuera de juego |
| 45 | Tarjetas amarillas |
| 46 | Tarjetas rojas |

## 8.4. Coeficientes del modelo de Regresión Logística



## 8.5. Coeficientes del modelo de SVC



## 8.6. Importancia de características del modelo de Random Forest