

UNIVERSIDAD TÉCNICA FEDERICO SANTA MARÍA
DEPARTAMENTO DE INFORMÁTICA
SANTIAGO - CHILE



“MODELO PREDICTIVO DE FÚTBOL CHILENO, SU
CONTEXTO NACIONAL Y LATINOAMERICANO BASADO
EN *POWER RANKING* HISTÓRICOS.”

JEREMMY MOYA GUAJARDO

MEMORIA PARA OPTAR AL TÍTULO DE
INGENIERO CIVIL EN INFORMÁTICA

Profesor Guía: José Luis Martí Lara
Profesor Correferente: Mauricio Olivares Faúndez

Julio - 2024

DEDICATORIA

Dedicado a todos quienes me acompañaron durante mi periodo universitario: amigos desde el colegio y desde la universidad; profesores del departamento de informática entregados a sus estudiantes; y a mi familia: mi madre, mi pareja, mi tío y mi abuela.

AGRADECIMIENTOS

Agradezco a mi profesor guía, José Luis Martí Lara, quien, a pesar de sus responsabilidades como jefe de carrera, orienta a todos sus memoristas con tiempo y atención dedicados.

Agradezco también a David Saldivia, quien mantiene la página web *golessonamores.cl*, desde la cual se extrajeron todos los registros del fútbol chileno profesional e internacional hasta el año 2019.

RESUMEN

Resumen—

Este trabajo de título se centra en la aplicación de ciencia de datos de forma predictiva al fútbol en Chile y Latinoamérica, a pesar de la disponibilidad limitada de datos históricos, basándose en la creación de *power rankings* históricos. El objetivo es proporcionar una herramienta efectiva para la toma de decisiones estratégicas por parte de equipos. Se desarrollará un modelo predictivo dinámico para la temporada de fútbol chileno 2024. Además, se visualizará la evolución histórica del fútbol chileno y se unificarán los *power rankings* en el contexto latinoamericano para contar con información más precisa y aportar en la toma de decisiones futuras en el contexto internacional.

Palabras Clave—

Ciencia de datos, Modelo predictivo, *Power ranking*, Visualización de datos, Fútbol chileno.

ABSTRACT

Abstract—

This thesis focuses on the application of data science in a predictive manner to football in Chile and Latin America, despite the limited availability of historical data. It is based on the creation of historical power rankings. The objective is to provide an effective tool for strategic decision-making by teams. A dynamic predictive model for the 2024 Chilean football season will be developed. Additionally, the historical evolution of Chilean football will be visualized, and the power rankings will be unified in the Latin American context to provide more accurate information and contribute to future decision-making in the international context.

Keywords— Data Science, Predictive Model, Power Ranking, Data Visualization, Chilean Football

ÍNDICE DE CONTENIDOS

RESUMEN	IV
ABSTRACT	IV
ÍNDICE DE FIGURAS	VII
ÍNDICE DE TABLAS	VII
INTRODUCCIÓN	1
CAPÍTULO 1: DEFINICIÓN DEL PROBLEMA	3
1.1 Aporte de la ciencia de datos en el fútbol	4
1.2 Objetivos	5
1.2.1 Objetivo general	5
1.2.2 Objetivos específicos	5
1.3 Alcances y limitaciones	6
CAPÍTULO 2: MARCO CONCEPTUAL	7
2.1 Ciencia de datos	7
2.2 Modelos predictivos	7
2.3 Metodologías de investigación	11
2.3.1 Metodología de investigación cuantitativa	11
2.3.2 Metodología CRISP-DM	14
2.4 <i>Power rankings</i>	14
2.4.1 Sistemas a partir de grafos	15
2.4.2 Sistema 'Elo'	16
2.5 Métodos de construcción de <i>power rankings</i> orientados al fútbol	16
2.5.1 Modelo FIFA 'SUM'	16
2.5.2 <i>Ranking</i> Mundial Femenino FIFA	18
2.5.3 Modelo 'ELO ++'	21
2.5.4 Método de clasificación basado en redes	24
2.5.5 Método de <i>World Football Elo Ratings</i> (eloratings.net)	26
2.5.6 Método 'Opta'	28
2.6 Funciones de predicción	29
2.6.1 Valor esperado de modelo FIFA 'SUM'	30
2.6.2 Valor esperado de <i>ranking</i> mundial femenino FIFA y eloratings.net	30
2.6.3 Predicción 'ELO++' adaptada	30
2.7 Predicción de empates	31
2.7.1 Victoria - Derrota	31
2.7.2 Rao y Kupper	32
2.7.3 Davidson	33
2.8 Métodos de proyección	34
2.8.1 Simulación Montecarlo	34

2.8.2	Puntos esperados	34
CAPÍTULO 3: PROPUESTA DE SOLUCIÓN		37
3.1	Comprensión del negocio	37
3.2	Comprensión de datos	37
3.3	Preparación de datos	39
3.4	Modelado	40
3.4.1	Resultado real	40
3.4.2	Resultado esperado	40
3.4.3	Variación de puntos	41
3.4.4	Encuentros entre divisiones o países	42
3.4.5	Predicción de resultados	43
3.4.6	Proyección de temporada	43
3.5	Evaluación	43
3.5.1	Definición de hiperparámetros	43
3.5.2	Medición de errores	44
3.6	Despliegue	47
CAPÍTULO 4: EVALUACIÓN Y DESPLIEGUE		48
4.1	Evaluación	48
4.2	Despliegue	53
4.2.1	<i>Power ranking</i> histórico nacional	53
4.3	Proyección de temporada actual	60
4.3.1	Proyección de Primera	61
4.3.2	Proyección de Primera B	63
4.4	<i>Power ranking</i> histórico internacional	66
CAPÍTULO 5: CONCLUSIONES		73
ANEXOS		77
REFERENCIAS BIBLIOGRÁFICAS		78

ÍNDICE DE FIGURAS

1	Diagrama de la función ideal hipotética	8
2	Fórmula de calculo de puntos Elo de Opta	28
3	Probabilidades de resultados respecto a diferencia de puntos	41
4	Índice de probabilidades de resultados respecto a σ	49
5	<i>Ranked probability score</i> medio respecto a σ	51
6	<i>Brier score</i> medio respecto a σ	52
7	<i>Power ranking</i> histórico nacional de equipos en las 3 primeras divisiones en 2023.	53
8	<i>Power ranking</i> histórico nacional de equipos de primera A en 2023.	55
9	<i>Power ranking</i> histórico nacional de equipos de primera B en 2023.	56
10	<i>Power ranking</i> histórico nacional de equipos de segunda en 2023.	57
11	<i>Power ranking</i> histórico nacional de equipos que alguna vez fueron top 3 <i>power ranking</i>	58
12	<i>Power ranking</i> histórico nacional de las 3 primeras divisiones.	60
13	Evolución de puntos esperados de Primera 2024.	61
14	Evolución de puntos esperados de Primera B 2024.	63
15	<i>Power ranking</i> histórico internacional de equipos participantes de competencias internacionales en 2023.	66
16	<i>Power ranking</i> histórico internacional de equipos participantes en copa libertadores en 2023.	68
17	<i>Power ranking</i> histórico internacional de equipos participantes en copa sudamericana en 2023.	69
18	<i>Power ranking</i> histórico internacional de equipos participantes en recopa sudamericana en 2023.	70
19	<i>Power ranking</i> histórico equipos participantes de recopa sudamericana.	71
20	<i>Power ranking</i> histórico nacional de países.	71
21	<i>Power ranking</i> histórico nacional de equipos en las 3 primeras divisiones en 2023 con $\sigma = 100$	77
22	<i>Power ranking</i> histórico nacional de equipos en las 3 primeras divisiones en 2023 con $\sigma = 800$	77

ÍNDICE DE TABLAS

1	Tabla de resultado real basado en goles anotados y diferencia de goles	19
2	Matriz de casos por rango de ejemplo.	45
3	Matriz de resultados por rango de ejemplo.	45
4	Matriz de porcentajes de ocurrencia por rango de ejemplo.	46
5	Matriz de índices de probabilidades por rango de ejemplo.	46
6	Matriz de índices de probabilidades ponderada de ejemplo.	46
7	Calidad de algoritmo por σ asignado.	48
8	Matriz de casos por rango con $\sigma = 380$	49

9	Matriz de resultados por rango con $\sigma = 380$	50
10	Matriz de porcentajes de ocurrencia por rango con $\sigma = 380$	50
11	Matriz de índices de probabilidades de resultados con $\sigma = 380$	50
12	Matriz de índices de probabilidades de resultados ponderados con $\sigma = 380$. .	51
13	Tabla de <i>power ranking</i> nacional de equipos en las tres primeras divisiones en 2023.	54
14	Matriz de probabilidades de posiciones finales de Primera 2024.	62
15	Matriz de probabilidades de posiciones finales de Primera B 2024.	65
16	Tabla de <i>power ranking</i> internacional de equipos participantes en competi- ciones CONMEBOL en 2023.	67

INTRODUCCIÓN

El fútbol es el deporte más visto y jugado en el mundo, su popularidad, impacto y complejidad lo ha llevado a ser objeto de estudio, y el área de la ciencia de datos ha sido protagonista en estos estudios experimentando avances significativos a nivel global, especialmente es su más alto nivel, proporcionando herramientas sofisticadas que mejoran la toma de decisiones tanto a nivel colectivo como individual. Sin embargo, el fútbol chileno está lejos de formar parte del alto nivel del fútbol, pues la adopción de estos avances se ha rezagado, ubicando a la región en un nivel inferior en términos de acceso y aplicación de datos detallados de equipos y jugadores. Este rezago limita la implementación de modelos predictivos avanzados, dificultando el aprovechamiento pleno de las ventajas que la ciencia de datos ofrece en otros lugares de mayor desarrollo futbolístico.

A pesar de esta limitante, en Chile se dispone de un dato muy valioso y a veces infravalorado, que son los resultados de los partidos históricos desde el comienzo de su profesionalismo en 1933. Si bien no aporta mucha información en cuanto al desarrollo detallado de los partidos ni del rendimiento individual de los jugadores, la interpretación de los equipos como una unidad toma más relevancia a la hora de analizar más de 60.000 resultados a lo largo de la historia del fútbol chileno; además, de ser prácticamente el único dato que permite análisis a nivel histórico, pues son muchos los partidos que no tienen más registro que este.

Este dato se vuelve especialmente útil a la hora de cuantificar y comparar la calidad de los equipos mediante *power rankings*, es decir, definir un orden de equipos según su rendimiento, pues estos se basan en los resultados de partidos de los distintos equipos. Y lo más importante, es la utilidad de los *power rankings* para la predicción de resultados, de forma que los datos limitados e históricos permitan la aplicación de modelos predictivos que, realizados en base a *power rankings* de calidad, posibilitan obtener predicciones de buen rendimiento a pesar de no utilizar datos detallados de cada equipo o jugador.

Mediante dichos elementos se propone una herramienta que muestre la realidad de los clubes chilenos en un *ranking* nacional y su evolución, el panorama internacional a comparación de clubes de países latinoamericanos en un *ranking* internacional y proyecciones de la temporada actual con las probabilidades de posiciones finales de los equipos según esta se desarrolla.

La memoria se compone de 5 capítulos; Definición del problema, donde se expone la motivación de la memoria, que busca solucionar y los objetivos. Marco conceptual, se describen las bases teóricas de alternativas a problemas similares. Propuesta de solución, se consolida una propuesta conformada por bases vistas el capítulo anterior, nuevos aportes y enfoques

al problema. Validación de resultados, justificar los parámetros utilizados y los resultados a partir de estos. Conclusiones, donde se responde a la completitud de los objetivos planteados inicialmente, los aportes de la metodología y trabajos a futuro.

CAPÍTULO 1

DEFINICIÓN DEL PROBLEMA

En el primer capítulo se describe el problema, cuál es, de donde surge, por qué es de interés su solución, mediante qué objetivos se busca resolver y qué alcances y limitaciones conlleva.

En el mundo del fútbol la toma de decisiones estratégicas es constante, por lo que se requiere de una amplia gama de herramientas que respalden estratégicamente la gestión de los equipos. La ciencia de datos proporciona muchas de estas herramientas, que abarcan desde registro de estadísticas simples (goles, asistencias, etc), pasando por seguimiento posicional de jugadores y equipos en general, hasta modelos predictivos de encuentros y proyección de jugadores.

La complejidad predictiva del fútbol radica en que inicialmente, cada equipo se compone de 11 jugadores individuales interactuando entre sí, pero la comprensión del equipo como una entidad cohesionada, más allá de la suma de los 11 individuos; aborda aspectos como las posibilidades que tiene un plantel de más de 20 jugadores, el cuerpo técnico, las dinámicas técnicas y culturales que caracterizan a un equipo, e incluso factores más intangibles como la calidad de local y visitante, jerarquías establecidas de equipos o lo que se conoce como 'ganar por la camiseta' entre otros aspectos propios del folclore futbolístico. Esto es fundamental ya que si juegan los mismos 11 jugadores cambiando su contexto en cuanto a los factores mencionados, es esperable que no se generen los mismos resultados.

Este enfoque integral de los equipos es aún más crítico en el contexto latinoamericano y chileno, donde la disponibilidad limitada de datos individuales de alta calidad contrasta con la abundancia de información en Europa, donde la ciencia de datos de fútbol está altamente profesionalizada y proporciona variedad de datos, tanto por equipos como por jugador. A pesar de estas diferencias, la existencia de datos históricos de resultados de partidos en Chile, y en toda América Latina, proporciona una base sólida para desarrollar predicciones y *rankings* confiables, pudiendo cuantificar y dimensionar la calidad de los equipos según sus resultados, considerando factores que la tabla de posiciones común no toma en cuenta como la dificultad de los encuentros, apoyándose de la interpretación del rendimiento del equipo como una unidad en lugar de centrarse únicamente en datos individuales.

Latinoamérica posee un vasto registro de datos históricos de fútbol que abarca muchas décadas, lo que permite no solo análisis contemporáneos, sino también un profundo estudio de la evolución a lo largo del tiempo. Además, esta riqueza de datos permite comparaciones a nivel internacional, dejando al alcance la respuesta a la incógnita sobre el desempeño relativo de los equipos chilenos en comparación con sus homólogos latinoamericanos en

diferentes países.

Esta memoria busca resolver la necesidad de contar con herramientas efectivas para la toma de decisiones estratégicas en el fútbol chileno y latinoamericano, a pesar de las limitaciones en el acceso a datos detallados de jugadores. La creación de modelos predictivos y *power rankings* basados en datos históricos de resultados se presenta como una solución viable para abordar este desafío y proporcionar información precisa y confiable para los equipos chilenos.

1.1. Aporte de la ciencia de datos en el fútbol

El fútbol se profesionalizó por primera vez en Inglaterra en 1888, y si bien no se puede dar una fecha con certeza, la ciencia de datos ha estado presente desde los comienzos del profesionalismo, al principio con aplicaciones básicas y manuales de estadísticas simples por parte de entrenadores y aficionados de equipos, registrando goles, asistencias y alineaciones para identificar tendencias y patrones.

El uso de estadísticas más avanzadas se hizo posible en la década de 1950, registrando datos más detallados como pases intentados y completados, tiros al arco, tiros de esquina, duelos individuales etc. De esta etapa se tiene el registro de la 'historia de Charles Reep', un fanático inglés quien, motivado por el mal ataque realizado por su equipo, comenzó a tomar notas con lápiz y libreta de pases, tiros y contraataques con dimensiones direccionales, es decir, lugar y dirección de los eventos y no solo contabilizar la ocurrencia de estos [Tippett, 2019]. Este es el tipo de estadísticas utilizadas de cara al público hoy en día, lo cual es preocupante desde el punto de vista analítico de los aficionados, pues el análisis objetivo que se tiene a disposición es el que nació en los años 50 y que el único avance que se tiene de este es su magnificación a distintos equipos, competencias y países, y visualización.

La década de 1980 se destaca por la digitalización de los registros que empezaron a almacenarse en bases de datos, facilitando la manipulación y registro de datos y permitiendo realizar análisis más profundos.

A comienzos de este milenio el principal avance es la recopilación de datos en tiempo real y avance en datos posicionales y direccionales. Estos avances se logran principalmente por la empresa Opta, que ha liderado la ciencia de datos en el deporte desde entonces.

En la década del 2010 se vive el boom de los modelos predictivos con *machine learning* que, junto con la variedad y cantidad de datos que se manejaba ya en esos años, hacen la mezcla

perfecta para explorar en modelos predictivos basados en distintos criterios y parámetros.

Actualmente, gracias a la tecnología *Wearable* y de *Tracking*, es decir los sensores en jugadores y balón para hacer seguimiento de posición, acciones, entre otros, se está incorporando a las estadísticas ya existentes una tercera dimensión de altura, para saber la elevación del balón en pases largos, tiros al arco, etc. En el caso de los jugadores, hacer seguimiento de sus capacidades físicas para administrar sus tiempos de juegos y esfuerzos según su cuerpo se los permita, y así lograr el más alto rendimiento minimizando el riesgo a lesiones.

1.2. Objetivos

El deporte ha sido históricamente guiado principalmente por sensaciones e intuición, especialmente el fútbol. Sin embargo la evolución de la tecnología y el manejo de datos dejan a disposición la cuantificación de factores que involucran a los distintos deportes. Este enfoque ha sido criticado dado que el fútbol, específicamente, es bastante dinámico y con estructuras muy distintas partido a partido, además de que se cree que los datos buscan dejar de lado la guía de la intuición. Pero al contrario de esto último, los datos buscan complementar estas guías, ayudando a confirmar lo que se cree, o ver distintos factores que no se tenían en cuenta en el comienzo, reabriendo debates de distintas índoles. De esta forma se definen los objetivos de la memoria.

1.2.1. Objetivo general

Diseñar una herramienta efectiva para la toma de decisiones estratégicas por parte de equipos chilenos de fútbol profesional para que, haciendo uso de ciencia de datos y de *power rankings*, aporte con información estratégica que ayude a un mejor rendimiento futuro de los mismos.

1.2.2. Objetivos específicos

Para cumplir este objetivo general se definen tres objetivos específicos a cumplir en esta memoria.

- **Diseño y creación de un modelo predictivo dinámico de temporada en el fútbol chileno y aplicación de este en la temporada 2024**

Para hacer seguimiento de la proyección a fin de temporada del equipo, proporcionando información de valor al club y la temporada actual.

- **Visualización de la evolución histórica del fútbol chileno a través del tiempo**

Para hacer seguimiento del rendimiento de los equipos en el contexto de la liga y comprender su evolución en el tiempo.

- **Unificación de *power rankings* en el Fútbol Latinoamericano**

Para comparar equipos de distintos países en un *power ranking* latinoamericano a nivel de clubes y a nivel país, para analizar la realidad internacional de distintos clubes respecto a los locales y proporcionar información más precisa para competencias internacionales.

1.3. Alcances y limitaciones

El alcance del trabajo es la localización a nivel país que implica el uso de data histórica de distintas competiciones y niveles, y como esto repercute en el desarrollo actual de competiciones profesionales.

En el aspecto internacional, el alcance es también a nivel histórico y comparativo de nivel entre países, pero se ve limitado a las apariciones intermitentes de equipos en competiciones internacionales, lo que implica poca representatividad para estos caso, pero para equipos con apariciones recurrentes no es el caso.

CAPÍTULO 2

MARCO CONCEPTUAL

Para comprender el desarrollo de este trabajo es necesario comprender las bases teóricas utilizadas en problemas similares y sobre las que se sustenta, que se explican en este capítulo.

2.1. Ciencia de datos

La ciencia de datos es el área interdisciplinaria compuesta por la informática, estadística e inteligencia de negocios, que estudia datos para extraer información significativa para la entidad o negocio en cuestión.

La informática proporciona las herramientas para la manipulación de datos, dígame su obtención, almacenamiento y procesamiento de grandes volúmenes de datos de forma eficiente, tarea que sería muy tediosa e inviable si se realizara a mano.

El área estadística aporta el conocimiento necesario para saber qué procesamiento realizar con los datos según su naturaleza, y cómo entender los resultados de estos procesos, identificar patrones y tendencias de forma que las conclusiones sean estadísticamente sólidas.

La inteligencia de negocios se refiere al conocimiento del medio en el que se esté trabajando para visualizar lo obtenido y presentarlo efectivamente, para que las personas encargadas de tomar decisiones puedan recibir una síntesis de todo el trabajo realizado por la ciencia de datos de forma comprensiva.

2.2. Modelos predictivos

Los modelos predictivos son herramientas que juntan la informática con la matemática y la estadística, identificando patrones para hacer predicciones o estimaciones a través de las componentes que influyen en el resultado esperado. Conociendo las componentes que influyen en un resultado en concreto se puede estimar valores futuros se pueden emplear algoritmos que determinen las relaciones entre las componentes estudiadas.

La predicción de valores a través de componentes se basa en la máxima de la función

óptima o ideal, que plantea que para todo problema de predicción existe una función hipotética que predice el valor objetivo de forma exacta. Esta función hipotética debiera contar con todas las variables que influyen en el valor a predecir, estas variables están bien cuantificadas (es decir, variables no matemáticas se le asignan valores matemáticos) y las relaciones entre estas variables son las correctas, de modo que el valor de la variable a predecir es el correcto (Ver figura 1). Sin embargo esta función ideal es hipotética, y los modelos a definir aspiran a acercarse lo más posible a esta función óptima; pero muchas veces sin éxito, ya que hay muchas predicciones que dependen de variables a las cuales no se tiene acceso y algunas son no matemáticas y su cuantificación no es trivial.

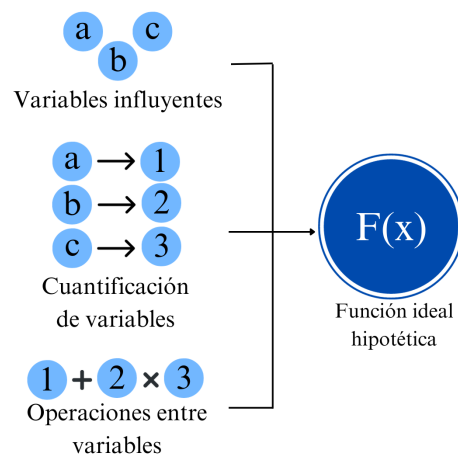


Figura 1: Diagrama de la función ideal hipotética
Fuente: Elaboración propia.

La creación de modelos predictivos que buscan acercarse al modelo ideal se puede realizar principalmente de tres formas presentadas a continuación.

1. Métodos tradicionales o heurísticos:

Para definir modelos a través de estos métodos se necesita conocimiento del medio para determinar reglas y heurísticas (reglas y valores arbitrarios) determinadas por humanos que al conocer el dominio, pueden identificar dichas reglas y patrones. Sin embargo, los patrones no siempre son claros o eficientes, por lo que estos métodos están limitados en este aspecto, además de que la definición de un modelo toma más tiempo y estudios que otros métodos.

2. Métodos de aprendizaje automatizado (*machine learning*):

Los métodos de *machine learning* tienen un enfoque más automatizado, ya que no depende de definiciones de algoritmos predictivos manuales, si no que se utilizan algoritmos como regresiones, maquinas de soporte vectorial, bosques aleatorios, entre

otros, algunos de ellos tienen capacidad generativa, que a partir de datos de entrenamiento este 'aprende' a generar una predicción. Este enfoque es muy útil para problemas de complejidades no lineales y de altas dimensiones, área que no cubre con efectividad los métodos heurísticos. La desventaja con la que cuentan estos métodos es que dependen de la disposición de grandes volúmenes de datos y variados de forma que sean representativos de los distintos perfiles de datos que existan en el problema.

3. Métodos de aprendizaje profundo (*deep learning*) o redes neuronales artificiales:

El aprendizaje profundo nace con la intención de simular la estructura del cerebro humano y sus distintas capas de neuronas. En la actualidad se ha asumido que esta idea ya no debe ser la finalidad puesto que el funcionamiento de las neuronas es difícil de replicar en la computación; sin embargo, el enfoque de distintas capas ha demostrado ser eficiente. Para definir un modelo con métodos de aprendizaje profundo se debe construir la estructura de la red neuronal artificial con parámetros como la cantidad de neuronas o nodos por capas, cantidad de capas y parámetros relacionados con los tipos de nodos, de forma que al proporcionarle un conjunto de datos de entrenamiento, la red neuronal va 'calibrando' parámetros de las neuronas y así, que mejorando su capacidad de producir una salida exitosa. Nuevamente estos métodos dependen de que exista un gran volumen de datos, pero proporcionan capacidades generativas de alta calidad, como generación de imágenes y de texto.

La calidad de los modelos predictivos se mide por distintas métricas, las cuales lo evalúan desde distintas perspectivas. Algunas de las métricas más comunes son:

1. Precisión:

Métrica muy útil, especialmente en problemas de clasificación, que mide la proporción de predicciones positivas correctas entre todas las predicciones positivas (correctas e incorrectas) [analyticsvidhya.com, 2023b].

$$Precision = \frac{VerdaderosPositivos}{VerdaderosPositivos + FalsosPositivos} \quad (1)$$

2. Recall:

Similar a la precisión, en este caso se define la proporción de predicciones positivas correctas entre todos los casos positivos reales [analyticsvidhya.com, 2023b].

$$Recall = \frac{VerdaderosPositivos}{VerdaderosPositivos + FalsosNegativos} \quad (2)$$

3. Exactitud:

Describe la proporción de predicciones correctas (verdaderos positivos y verdaderos negativos) respecto al total de predicciones [analyticsvidhya.com, 2023b].

$$Accuracy = \frac{VerdaderosPositivos + VerdaderosNegativos}{TotalPredicciones} \quad (3)$$

4. Error cuadrático medio (ECM):

Mide el promedio de los errores entre los valores predichos y los valores reales. Un ECM más cercano a 0 indica un mejor ajuste del modelo [ncl.ac.uk (Newcastle University), 2023].

$$ECM = \frac{\sum_{i=1}^n (ValorReal - ValorPredicho)^2}{n} \quad (4)$$

con n siendo el número total de observaciones.

$$R^2 = \frac{\sum_{i=1}^n (ValorReal - ValorPredicho)^2}{\sum_{i=1}^n (ValorReal - ValorMedio)^2} \quad (5)$$

donde n es el número total de observaciones.

5. Binary cross entropy

Métrica que mide la discrepancia entre las predicciones de un modelo y los resultados reales de un evento. Para cada evento, se calcula la probabilidad predicha de los distintos resultados, y se compara con el resultado real del evento; mientras más cercano a 0 sea el resultado, más cercanas al resultado real son las predicciones del modelo. La *binary cross entropy* se calcula como [analyticsvidhya.com, 2023a].

$$CE = \frac{-1}{N} \sum_{i=1}^n (R_{real} \times \log_{10} R_{predicho} + (1 - R_{real}) \times \log_{10}(1 - R_{predicho})) \quad (6)$$

6. Brier score:

Métrica de que mide la distancia entre la predicción del modelo y si el escenario en cuestión ocurrió o no para cada uno de los escenarios posibles (la ocurrencia de un escenario se representa con 1 si es el resultado real y si no, con 0) [Wheatcroft, 2021].

$$Brier = \sum_{i=1}^r (PrediccionEscenario - EscenarioOcurrido)^2 \quad (7)$$

donde r son los distintos escenarios posibles

7. Ranked Probability Score:

RPC está relacionado con el *Brier Score*, pero compara la función de distribución acumulada de la predicción y la salida real [Wheatcroft, 2021].

$$RPC = \sum_{i=1}^{r-1} \sum_{j=1}^i (PrediccionEscenario - EscenarioOcurrido)^2 \quad (8)$$

donde r son los distintos escenarios posibles

Estas son algunas de las métricas más comunes para determinar la calidad de un modelo predictivo, cada una apropiada dentro de su propio contexto y complementarias entre sí.

2.3. Metodologías de investigación

La investigación es un proceso mediante el que se genera conocimiento de la realidad para explicarla, comprenderla y transformarla acorde a las necesidades materiales y socioculturales que cambian constantemente. Se comprende también como un proceso creativo regido por reglas de validez y aceptabilidad compartidas por la comunidad científica, buscando resolver problemas produciendo conocimientos nuevos.

A continuación se exponen tres metodologías de investigación aplicables en el contexto del problema.

2.3.1. Metodología de investigación cuantitativa

La metodología de investigación se puede separar en dos metodologías dependiendo de la naturaleza del problema, cualitativa y cuantitativa. La metodología cualitativa busca captar la realidad 'a través de los ojos' de la gente que se estudia, donde el investigador induce las propiedades del problema a partir de la forma como orientan e interpretan el mundo los individuos examinados. Por otro lado, la metodología cuantitativa suele comenzar de bases teóricas aceptadas por la comunidad científica, en las cuales se basa para plantear hipótesis sobre relaciones esperadas del problema.

El método cuantitativo realiza una recolección de información cuantitativa con conceptos empíricos medibles y analiza la información recopilada para determinar el grado de significancia de las variables. La estructura del método cuantitativo consta de 5 fases; conceptual, planificación y diseño, empírica, analítica y fase de difusión. Las fases de la estructura constan con una serie de pasos, los cuales se señalan a continuación [Monje, 2011]:

1. **Fase 1, Conceptual:** El comienzo de la investigación consta en ordenar y establecer las inquietudes, preguntas y antecedentes del tema.

- Paso 1: Formulación y delimitación del problema.

El punto de partida de la investigación, se determina lo que se investiga y los objetivos del mismo.

- Paso 2: Revisión de la literatura.

La búsqueda de información que fundamente lo investigado es necesaria para entender el contexto del área, lo que se ha encontrado, temas o métodos similares, junto con sus aportes y dificultades.

- Paso 3: Construcción del marco teórico.

Una vez revisada la literatura, se debe organizar y elaborar el marco conceptual que define las bases conceptuales y teóricas que permiten comprender la investigación propuesta.

- Paso 4: Formulación de hipótesis.

Se establece una hipótesis que anticipa una probable explicación del tema investigado, planteando respuestas preliminares del mismo.

2. Fase 2, Planeación y diseño: Es la segunda fase, donde se toman decisiones en cuanto a métodos y estrategias a emplear, y se plantea la recolección de datos necesarios.

- Paso 5: Selección de un diseño de investigación.

La metodología de investigación determina las estrategias y procedimientos a seguir para dar respuesta al problema planteado. El método a elegir se debe adecuar a la naturaleza de la investigación, permitiendo ser flexible dado factores menores como la intencionalidad, recursos y disponibilidad de tiempo.

- Paso 6: Identificación de la población a estudiar.

El universo o población a estudiar se refiere a los objetos, sujetos o unidades que comparten la característica a estudiar.

- Paso 7: Selección de métodos e instrumentos.

Se debe exponer los métodos para observar o medir las variables de la investigación, definir su significado y traducirlos a indicadores empíricos en base a los conceptos definidos anteriormente.

- Paso 8: Diseño del plan de muestreo.

Dependiendo del problema, los datos recolectados corresponden a una muestra y no a toda la población, de forma que la muestra sea lo suficientemente representativa del universo reduciendo los costos de manejar todos los datos. De esta forma, se debe planear como obtener la muestra manteniendo una representatividad del universo.

- Paso 9: Término y revisión del plan de investigación.

Realizados los pasos anteriores, de todas formas deben ser revisados y perfeccionados antes de ponerse en ejecución.

- Paso 10: Realización del estudio piloto y las revisiones.

Finalmente, antes de la ejecución, es recomendable hacer un estudio piloto para someter a prueba la viabilidad final del estudio principal mediante una prueba de pequeña escala.

3. **Fase 3, Empírica:** En las dos fases previas se llevaron a cabo todas las tareas de planificación que permiten la ejecución práctica de lo planteado en la presente fase, recolectando los datos y dejándolos a disposición del posterior análisis.

- Paso 11: Recolección de datos.

La obtención de los datos se realiza basado en lo establecido en la fase 2, tanto en metodología, instrumentos y muestra a recolectar.

- Paso 12: Preparación de los datos para análisis.

Los datos ya recolectados deben ser analizados para identificar y corregir posibles incongruencias, omisiones o errores, de forma que los datos finales sean de calidad y estandarizados en caso de que el caso lo amerite.

4. **Fase 4, Analítica:** Una vez con los datos a disposición, deben ser analizados de forma tal que aporten a esclarecer el problema planteado anteriormente, permitiendo sacar conclusiones al respecto.

- Paso 13: Análisis de datos.

La disposición de los datos por si solo no aportan respuestas a la pregunta de investigación. Estos deben ser procesados y analizados ordenada y coherentemente para identificar patrones y relaciones, descomponiendo sus partes y estudiadas minuciosamente según corresponda y luego sintetizadas.

- Paso 14: Interpretación de resultados.

La interpretación aborda la comparación de los resultados del análisis con los objetivos planteados de la investigación, señalando las características de las distribuciones y relaciones del proceso de análisis. También se establecen comparaciones o relaciones con resultados obtenidos de trabajos similares.

5. **Fase 5, Difusión:** Finalmente se tiene la divulgación, de forma que solo en cuanto se dan a conocer los resultados de la investigación se estará contribuyendo al conocimiento existente sobre el tema tratado.

- Paso 15: Comunicación de las observaciones.

La comunicación de la investigación contempla aspectos como antecedentes, problema planteado, objetivos, diseño metodológico empleado, resultados, dificultades y limitaciones de datos y sugerencias para futuros estudios. Esto se puede realizar en disertaciones, artículos, conferencias, libros, etc.

- Paso 16: Aplicación de las observaciones.

La finalidad de la investigación es contribuir a la transformación de la realidad en sus diferentes manifestaciones. Sin distinción de la importancia de los hallazgos de la investigación, si no se aplican en el mundo real, la investigación termina siendo un ejercicio sin mayor aporte.

2.3.2. Metodología CRISP-DM

En el área de minería de datos y la ciencia de datos, la metodología CRISP-DM (*CRoss-Industry Standard Process for Data Mining*) ha sido un referente desde su origen a mediados de los años 90'. CRISP-DM ha sido adoptada como un estándar para el desarrollo de proyectos de minería de datos por su estructura sencilla y definida, pero adaptable a distintos negocios, contextos y necesidades. Dicha estructura es la siguiente:

1. **Fase 1, Comprensión del negocio:** En esta primera instancia se establecen los objetivos y necesidades del proyecto desde la perspectiva del negocio, y qué herramientas dispone la minería de datos para lograr estos objetivos.
2. **Fase 2, Comprensión de datos:** Esta fase comprende la recopilación y exploración de datos, aplicando una evaluación inicial de la calidad de estos, los problemas y limitantes que implican y finalmente, se determinan las variables relevantes para el análisis.
3. **Fase 3, Preparación de datos:** Consta de la manipulación inicial de datos, la limpieza, transformación e integración de estos con tal de disponer de un formato adecuado para la explotación de la información.
4. **Fase 4, Modelado:** Se define un modelo para utilizar los datos con la finalidad de cumplir los objetivos planteados inicialmente; seleccionando, ajustando y complementando distintos algoritmos de modelado adecuados para el contexto del proyecto.
5. **Fase 5, Evaluación:** Se evalúa el modelo planteado para determinar si es adecuado y eficaz para el cumplimiento de objetivos, utilizando métricas de evaluación para determinar el ajuste adecuado del modelo.
6. **Fase 6, Despliegue:** Finalmente, se implementa el modelo a nivel organización, documentando hallazgos y comunicando los resultados, y estableciendo un plan de mantenimiento continua del modelo.

2.4. Power rankings

Los *power ranking* son una herramienta para cuantificar el poder o nivel (*power*) de una persona, equipo o entidad en general, respecto a otros pares de un mismo entono

(liga, competición, etc) en una tabla de clasificación, donde los mejores se posicionan al comiendo de la tabla, y mientras va bajando el nivel, se posicionan más abajo en la tabla. La puntuación asignada a cada equipo o persona varía según el método utilizado, y suele variar según se realizan enfrentamientos y dependiendo del método utilizado, los *rankings* tienen una capacidad predictiva de enfrentamientos entre integrantes de la tabla [Pradhan y Abdourazakou, 2020].

Este tipo de tablas resulta más efectiva que las tablas de posicionamiento tradicionales que asigna puntos por victoria o empate dependiendo del deporte, ya que considera factores como la dificultad de cada enfrentamiento, de forma que se premia mejor vencer a un equipo de la parte alta de la tabla que a uno de la parte baja de la tabla, y en el caso de la derrota, se castiga más el perder con equipos de la parte baja de la tablas que perder contra equipos de la parte alta de la tabla. Dependiendo del método de construcción del *ranking* se pueden considerar más factores como condiciones de local o visita, importancia de la competición o encuentro, factores temporales para dar más valor a los encuentros recientes, entre muchos otros según la finalidad y enfoque de cada método.

Los *power rankings* se han utilizado en variedad de deportes como fútbol, fútbol americano, béisbol, baloncesto, hockey, tenis, automovilismo, entre otros. También se ha implementado en distintos deportes electrónicos y en disciplinas como el ajedrez, póker y go, entre otros. En estas distintas disciplinas se diseñan los métodos de puntuación considerando las distintas reglas y características de cada una, y de la misma forma, su capacidad predictiva también es propia, pero el factor más común es la diferencia de posición en el *ranking* y el rendimiento de cada individuo en la posición en la que se encuentra (tipo de campo, escenario y variantes dependientes de la competencia) ([Pradhan y Abdourazakou, 2020], [Burer, 2012]).

La construcción de un *power ranking* tiene distintos enfoques que son más apropiados según el problema y contexto en cuestión, algunos de estos son los sistemas a partir de grafos y el sistema 'Elo'.

2.4.1. Sistemas a partir de grafos

El enfoque de construcción de *power ranking* utilizando grafos busca cuantificar los resultados de encuentros indirectos, de forma que si un jugador/equipo A gana a B , que resultado tiene indirectamente un encuentro entre A y C sin haber disputado dicho encuentro, pero si uno entre B y C . Los resultados indirectos disminuyen su peso según aumenta la distancia del encuentro indirecto.

Este planteamiento es beneficioso en contextos donde no todos los integrantes de una competición o del *power ranking* juegan entre si o si la cantidad de encuentros es muy baja. [Katz, 1953]

2.4.2. Sistema 'Elo'

Una forma de armar un *power ranking* es el sistema 'Elo', un sistema de clasificación creada en la década de 1960 por el físico y maestro de ajedrez húngaro Árpád Emrick Élő (1903-1992) para clasificar a los jugadores de ajedrez según su habilidad relativa.

La puntuación de habilidad mínima de un jugador es de 0, y si bien no se define un máximo, se considera como un máximo teórico de 3000 dada la forma de crecimiento que tiene el sistema. Dentro de la tabla se definen distintos niveles para categorizar a los jugadores, por ejemplo, los jugadores de más de 2000 puntos son considerados maestros, o expertos si se prefiere. [Szczecinski y Djebbi, 2020]

El sistema de clasificación 'Elo' ha sido adoptado por múltiples disciplinas en las que se enfrentan 2 personas, grupos o fuerzas por encuentro, y al día de hoy es de los mejores métodos para clasificar la habilidad de jugadores o equipos de distintas competiciones, cada una con ligeras adaptaciones, según corresponda. [Ebtekar y Liu, 2021]

2.5. Métodos de construcción de *power rankings* orientados al fútbol

Actualmente existen distintos modelos de generación de *ranking* orientados al fútbol, todos con sus pros y contras. Se presentan, a continuación, las estudiadas para el desarrollo de la propuesta.

2.5.1. Modelo FIFA 'SUM'

La FIFA desarrolló en 2018 un nuevo modelo para clasificar a las selecciones internacionales masculinas. La necesidad de un modelo más actualizado para las selecciones masculinas era crítica dado que el método anterior que se tenía había sido demostrado como ineficiente [McHale et al., 2007].

El modelo 'SUM' de FIFA se basa en el sistema 'Elo', y la suma o resta de puntos de cada equipo involucrado en un encuentro basado en el rendimiento relativo de cada equipo (basado en los puntos de estos), la importancia del encuentro y el resultado esperado previo

al encuentro dada la diferencia de puntos de los equipos, entendiendo que un equipo es mejor que otro mientras más puntos tenga.

La fórmula de 'SUM' para el cálculo de puntuación dado un partido es la siguiente

$$P = P_{\text{before}} + I \times (W - W_e) \quad (9)$$

donde:

- P son los puntos del equipo posterior al encuentro.
- P_{before} son los puntos del equipo previo al encuentro.
- I es la importancia del encuentro, dada por los siguientes valores:
 - $I = 05$ para partidos amistosos fuera de fechas FIFA.
 - $I = 10$ para partidos amistosos durante fechas FIFA.
 - $I = 15$ para partidos de fase de grupos de la *UEFA Nations League* (competencia europea).
 - $I = 25$ para partidos de *Play-off* y finales de la *UEFA Nations League*.
 - $I = 25$ para partidos clasificatorios de todas las confederaciones para el Mundial de fútbol.
 - $I = 35$ para partidos previos a los cuartos de final en competiciones de copa de confederaciones (Copa América, Eurocopa, etc.).
 - $I = 40$ para partidos a partir de cuartos de final (cuartos, semis y final) en competiciones de copa de confederaciones.
 - $I = 50$ para partidos de la Copa del Mundo previos a los cuartos de final.
 - $I = 60$ para partidos de la Copa del Mundo a partir de los cuartos de final.
- W es el resultado del partido, con los siguientes valores:
 - $W = 1$ si el equipo en cuestión gana.
 - $W = 0,5$ si el equipo en cuestión empata.
 - $W = 0$ si el equipo en cuestión pierde.

- W_e es el valor esperado del encuentro dado por la fórmula:

$$W_e = 1 / (10^{-\frac{dr}{600}} + 1)$$

donde dr Es la diferencia de *ranking* que se calcula como:

$$dr = P_{\text{before del equipo en cuestión}} - P_{\text{before del equipo rival}}$$

Este modelo tiene dos consideraciones, para los partidos eliminatorios, donde el equipo perdedor no tiene reducción de puntos de forma que se protegen los puntos de equipos que lograron avanzar a fases eliminatorias. La segunda consideración aplica a los partidos definidos por penales, el W asignado a los equipos es:

- $W = 0,75$ Para el equipo ganador.
- $W = 0,5$ Para el equipo perdedor.

Para más detalle, consultar la web oficial de FIFA donde se explica el sistema en más profundidad [FIFA, 2018].

Este método tiene ventajas como:

- Distingue de importancia de partidos.
- Relativiza la dificultad de un partido por las diferencias de nivel.
- Considera la cronología de los encuentros, importando el orden en que se juegan los partidos.
- Es simple de entender y ejecutar.

Sin embargo, también tiene desventajas:

- Fuera de su ejecución cronológica, la temporalidad de los eventos no influye en el cálculo, donde tienen más importancia los partidos más recientes.
- No distingue la calidad de local de la de visita de los equipos.
- No considera la contundencia del resultado (goles marcados, diferencia de goles, etc.)
- Es orientada a selecciones, lo que es una desventaja a la hora de orientar el trabajo en clubes. Sin embargo, esto tiene impacto más que nada en la cuantificación de la importancia de los partidos.

2.5.2. **Ranking Mundial Femenino FIFA**

El modelo de clasificación de selecciones femeninas de FIFA fue implementado en el 2003, y fue la base para el modelo 'SUM' de FIFA ya expuesto, por lo que son modelos muy parecidos, sin embargo hay ciertas diferencias que resultan interesantes.

La fórmula de clasificación mundial de selecciones femeninas para el cálculo de puntuación dado un partido es el siguiente

$$P = P_{\text{before}} + I \times (W - W_e) \quad (10)$$

donde:

- P son los puntos del equipo posterior al encuentro.
- P_{before} son los puntos del equipo previo al encuentro.
- I es la importancia del encuentro, dada por los siguientes valores:
 - $I = 15$ para partidos amistosos.
 - $I = 30$ para partidos amistosos entre dos equipos top 10 de todas las selecciones femeninas.
 - $I = 30$ para partidos de clasificación de competencias continentales.
 - $I = 45$ para finales de competencias continentales.
 - $I = 45$ para partidos de clasificación de los juegos olímpicos
 - $I = 45$ para partidos clasificatorios a la copa del mundo femenina.
 - $I = 60$ para partidos de los juegos olímpicos femeninos.
 - $I = 60$ para partidos de la Copa del Mundo femenina.
- W es el valor del resultado real del partido, con los valores de la tabla 1:

	Goal difference						
	0	1	2	3	4	5	6/+
Goals scored	Actual result value (%)						
0	47	15	8	4	3	2	1
1	50	16 *	8.9	4.8	3.7	2.6	1.5
2	51	17	9.8	5.6	4.4	3.2	2
3	52	18 **	10.7	6.4	5.1	3.8 ***	2.5
4	52.5	19	11.6	7.2	5.8	4.4	3
5/+	53	20	12.5	8	6.5	5	3.5

Tabla 1: Tabla de resultado real basado en goles anotados y diferencia de goles

Fuente: FIFA Women's World Ranking

(<https://digitalhub.fifa.com/m/3d9cb1decbbb2ac7/original/Womens-World-Ranking-Procedures.pdf>).

El valor del resultado real, o *Actual result value*, se representa como un porcentaje extrapolado a partir de la diferencia de goles del partido (*Goal difference*) y el número

de goles anotados por el equipo perdedor (en la tabla *Goals scored*). Esto significa que el equipo ganador es recompensado por el valor restante del obtenido en la tabla, excepto en los casos de empate (columna con *Goal difference* con valor 0), donde ambos equipo tienen el mismo valor del resultado real.

Por ejemplo: El equipo A pierde 1.2 (*) por lo que recibe 16 % del 100 % de los puntos y el equipo rival recibe el 84 % restante. De forma similar, el equipo que pierde 3-4 (**) recibe 18 % de los puntos disponibles y el otro equipo recibe el 82 % restante. Por último, el equipo que pierde 3-8 (***) recibe 3,8 % de los puntos.

- W_e es el valor esperado del encuentro dado por la fórmula:

$$W_e = 1/(10^{\frac{x}{2}} + 1)$$

donde x es la diferencia de puntos escalada entre los equipos que se calcula como:

$$x = \frac{P_{\text{before del equipo en cuestión}} - P_{\text{before del equipo rival}}}{-200}$$

En este caso, -200 funciona como un factor de escalado.

Esta formula es similar a la del método 'SUM', pero tiene ciertas diferencias. Sin embargo estas diferencias son de notación, por lo que si se escribe la fórmula del resultado esperado del método del fútbol femenino de forma similar a la del método 'SUM' resulta:

$$W_e = 1/(10^{-\frac{dr}{400}} + 1)$$

donde se observa que la única diferencia se encuentra en el factor de escalado, que en este caso es de 400 (con esta notación) y en el método 'SUM' es de 600.

Una consideración que tiene este método es que en caso de que un equipo sea local, se le suman 100 puntos para representar la ventaja de campo, ya que históricamente los equipos locales ganan el 66 % de los partidos y el visitante el 34 %, y al sumar los 100 puntos, en caso de enfrentarse dos equipos de igual puntuación inicialmente, el resultado esperado para el equipo local (habiendo sumado los 100 puntos) es del 64 % [FIFA, 2003].

Este método tiene ventajas como:

- Distingue de importancia de partidos.
- Relativiza la dificultad de un partido por las diferencias de nivel.
- Considera la cronología de los encuentros, importando el orden en que se juegan los partidos.
- La contundencia de resultado influye.

- Representa el resultado no como solo 3 valores posibles (victoria, empate, derrota) si no como un valor continuo que depende del resultado.
- Considera la calidad de local o visita de los equipos.
- Es simple de entender y ejecutar.

Sin embargo, también tiene desventajas:

- Fuera de su ejecución cronológica, no se considera la temporalidad de los eventos, donde tienen más importancia los partidos más recientes.
- Es orientada a selecciones, lo que es una desventaja a la hora de orientar el trabajo en clubes. Sin embargo, esto tiene impacto más que nada en la cuantificación de la importancia de los partidos.

2.5.3. Modelo 'ELO ++'

El modelo 'ELO ++' nace de una competencia de *Kaggle* para clasificar a jugadores de ajedrez, siendo este modelo el ganador de la competencia. En el artículo [Lasek *et al.*, 2013] se realiza una adaptación del modelo para aplicarlo al contexto del fútbol de selecciones masculinas; se consideran dos resultados posibles de un partido, la victoria de un equipo o de otro, es decir, no se consideran empates como tal, pero se puede interpretar uno cuando la probabilidad de ganar de ambos equipos es similar.

Este modelo se define mediante distintas fórmulas.

■ Función de predicción

Se define el resultado esperado de un partido como

$$P_{ij} = 1 / (1 + e^{-(r_i + h - r_j)})$$

donde

- P_{ij} es la probabilidad de que el equipo i gane al enfrentarse con el equipo j
- r_i es la clasificación (*rating*) del equipo i
- h es un hiperparámetro a definir para describir la ventaja del equipo local, cuando el partido se juega en territorio neutro, $h = 0$.

De esta forma, la probabilidad de que el equipo j se calcula como $P_{ji} = 1 - P_{ij}$.

■ Escalado de tiempo

En este modelo, cada encuentro tiene peso dada la antigüedad del partido; dicho peso se define

$$W_{ij} = \left(\frac{1 + t - t_{min}}{1 + t_{max} - t_{min}} \right)^2$$

donde

- W_{ij} es el peso de cierto enfrentamiento entre el equipo i y el equipo j .
- t_{min} se refiere al número de mes más bajo (antiguo) dentro de los datos a revisar.
- t_{max} se refiere al número de mes más alto (actual) dentro de los datos a revisar.
- t se refiere al número de mes en que ocurrió el encuentro en cuestión.

Los valores que puede tomar W_{ij} se encuentran en el intervalo $]0, 1]$, es mayor según aumenta t , y viceversa. Esto implica que mientras más actual es el partido, más peso tiene para el cálculo del *ranking*.

■ Vecinos

El modelo 'ELO++' aborda un aspecto muy particular de las clasificaciones de selecciones, que es ajustar la clasificación de un equipo respecto al nivel de sus rivales, de forma que si en una confederación un equipo es superior al resto pero, en general, esa confederación no es mejor que una de alto nivel, el equipo que destaca en dicha confederación no podrá ganar demasiados puntos dado que el nivel medio de sus rivales es muy bajo. De la misma forma, un equipo que tiende a perder en una confederación de alto nivel, evita perder puntos exageradamente dado que el nivel promedio de sus rivales es más alto.

El cálculo del nivel promedio de los rivales de un equipo es sencillo.

$$a_i = \frac{\sum_{k \in N_i} W_{ik} r_k}{\sum_{k \in N_i} W_{ik}}$$

donde

- a_i es la puntuación promedio de los rivales de i .
- N_i es el conjunto de rivales de i con los que se ha enfrentado mutuamente en el rango de tiempo en que se evalúa la clasificación.

■ Cálculo de clasificación

Finalmente, para el cálculo de clasificación de cada equipo se considera que inicialmente todos los equipos tienen una puntuación de 0; además se define una cantidad de P iteraciones. Durante cada iteración se escanea todo el conjunto de datos (partidos) de forma aleatoria y realizando las siguientes actualizaciones.

$$r_i \leftarrow r_i - \eta[W_{ij}(S_{ij} - P_{ij})P_{ij}(1 - P_{ij}) + \lambda \frac{1}{|N_i|}(r_i - a_i)]$$

$$r_j \leftarrow r_j - \eta[W_{ji}(S_{ji} - P_{ji})P_{ji}(1 - P_{ji}) + \lambda \frac{1}{|N_j|}(r_j - a_j)]$$

donde

- η es la tasa de aprendizaje (*learning rate*) que se define como

$$\eta = \sqrt{\frac{1 + 0,1\dot{P}}{p + 0,1\dot{P}}}$$

Donde p es el número de iteración actual.

- λ es un hiper-parámetro que determina la importancia de la regularización basado en los vecinos con los que un equipo se enfrenta.
- h es el ya mencionado hiper-parámetro que determina la ventaja del equipo local.

El valor de a_i se calcula después de cada iteración.

El método 'ELO++' itera repetidamente sobre todo el conjunto de datos repetidas veces para determinar la clasificación de los distintos competidores involucrados.

Este método tiene ventajas como:

- Relativiza la dificultad de un partido por las diferencias de nivel.
- Considera el tiempo en que se jugaron los partidos, dando más importancia a los más recientes.
- Distingue entre la calidad de local o visita de los equipos.
- Ideal para un entorno en el que no todos los equipos juegan regularmente entre todos, es decir, cada equipo juega en su 'nicho' que sería su confederación.

Sin embargo, también tiene desventajas:

- No considera la importancia de los partidos (dado que inicialmente se hizo para el ajedrez y en su contexto no tiene mayor implicancia).
- Es bastante compleja de entender en comparación a otros métodos.

- Su complejidad computacional es muy alta, sobre todo si se eligen rangos de tiempo de estudio muy largos.
- Está sujeta a hiper-parámetros que pueden variar en el tiempo o sobre ajustar el modelo en casos más concretos.

2.5.4. Método de clasificación basado en redes

Este método deriva del análisis de grafos, donde los nodos son los equipos y las conexiones son los encuentros entre los equipos conectados. Es una modificación deportiva de Katz [Katz, 1953], que asigna pesos a los distintos nodos de un grafo; en el caso de la adaptación propuesta el 2005 en el artículo [Park y Newman, 2005], se define la matriz de adyacencia del grafo A y una entrada i, j (a_{ij}) de la matriz que describe la cantidad de victorias del equipo j sobre el equipo i , de un total de n equipos. En el caso de empates, se representa como mitad victoria y mitad derrota.

Se define también una puntuación de victoria y una de derrota para cada equipo. La puntuación de victoria considera el número de victorias directas e indirectas del equipo; las victorias indirectas se ven reducidas por un factor α a definir.

Para un equipo i , se calcula la cantidad de victorias directas como

$$\text{Victorias directas de } i = \sum_j a_{ji}; \forall i, j \in \{1, 2, \dots, n\} \wedge j \neq i$$

y el número de victorias indirectas de distancia 2 se calcula como

$$\text{Victorias indirectas de distancia 2 de } i = \sum_{k,j} a_{kj}a_{ji}; \forall i, j, k \in \{1, 2, \dots, n\} \wedge j \neq k \neq i$$

y así según aumenta la distancia de los equipos.

Se puede calcular el factor de victoria w_i para el equipo i con un factor de descuento α^{k-1} , donde k es la distancia de los equipos.

$$\begin{aligned}
 w_i &= \sum_j a_{ij} + \alpha \sum_{j,k} a_{jk} a_{ji} + \alpha^2 \sum_{j,k,l} a_{lk} a_{kj} a_{ji} + \dots \\
 &= \sum_j (1 + \alpha \sum_k a_{kj} + \alpha^2 \sum_{l,k} a_{lk} a_{kj} + \dots) a_{ji} \\
 &= \sum_j (1 + \alpha w_j) a_{ji} \\
 w_i &= d_i^{in} + \alpha \sum_j (A^T)_{ij} w_j
 \end{aligned}$$

donde

- d_i^{in} es la cantidad de vectores apuntando a i , es decir, la cantidad de victorias directas de i
- A^T es la matriz transpuesta de la matriz de adyacencia ya definida A

La forma matricial del factor de victoria resulta en

$$w = (I - \alpha A^T)^{-1} d^{in}$$

donde

- d^{out} es el vector donde la posición i -ésima es la cantidad de vectores que apuntan desde i

Y la forma matricial del factor análogo de derrota l resulta en

$$l = (I - \alpha A^T)^{-1} d^{out}$$

donde

- d^{out} es el vector donde la posición i -ésima es la cantidad de vectores que apuntan desde i

Finalmente, la clasificación (*rating*) se determina por

$$r = w - l$$

Para determinar el valor óptimo de α se iteran valores en el rango de 0 a 0.95 con saltos de 0.05, en un conjunto de datos realizando predicciones y midiendo la precisión de cada iteración, definiendo el α que mejor precisión da en las predicciones. Este método da un enfoque distinto a los otros vistos, ya que no está basado en 'Elo', esto proporciona una serie de ventajas.

- Permite comparar equipos que no han jugado juntos a través de equipos 'intermedios'.
- La lógica es simple, considerando victorias o derrotas entre equipos.
- La implementación para equipos de una sola liga es similar a la de equipos de distintas ligas y países. Lo único a considerar es la agregación de equipos.

De todos modos, el este método presenta desventajas.

- No considera la cronología ni temporalidad de los partidos.
- La contundencia del resultado de los partidos no influye.
- No asigna dificultad de los partidos, ya que asigna la clasificación de los partidos al final.
- No distingue entre calidad de local o visita de los equipos.
- Todos los partidos tienen la misma importancia.

2.5.5. Método de *World Football Elo Ratings* (eloratings.net)

La página web 'elorating.net' proporciona clasificaciones de fútbol para selecciones nacionales basadas en el sistema de clasificación Elo. También proporciona estadísticas y clasificaciones actualizadas para las selecciones nacionales de fútbol de todo el mundo, y han estado implementado esta metodología en el hockey sobre hielo. Para más información sobre la clasificación actualizada por 'elorating.net', consultar directamente su página [elorating.net, 2023].

La fórmula utilizada en la web es

$$R_n = R_o + K(W - W_e)$$

donde

- R_n es la clasificación de un determinado equipo luego de un partido.

- R_o es la clasificación del equipo previo al encuentro.
- K es la importancia del encuentro. El valor de K está dado por
 - 60 para finales de la copa del mundo.
 - 50 para finales de campeonatos de confederaciones.
 - 40 para clasificatorias mundiales, continentales y cualquier tipo de torneo mayor.
 - 30 para cualquier otro tipo de torneo.
 - 20 para partidos amistosos.

K se ajusta según la diferencia de goles en el resultado del partido de forma que

- Incrementa la mitad si la diferencia de gol es de 2 goles
 - Incrementa en $\frac{3}{4}$ si el partido se define por 3 goles de diferencia
 - Y para diferencias de 4 goles en adelante, incrementa en $\frac{3}{4} + \frac{(N-3)}{8}$
- W es el resultado del partido desde la perspectiva del equipo del cual se está calculando su clasificación, siendo **1** en caso de victoria, **0.5** en caso de empate y **0** en caso de derrota.
 - W_e es el resultado esperado que se calcula con la formula

$$W_e = 1 / (10^{\frac{-dr}{400}} + 1)$$

donde dr es la diferencia de clasificación de los equipos, dando 100 puntos extra al equipo que tenga calidad de local.

Este método es bastante similar al método 'SUM' de la FIFA, con ciertas diferencias que presentan puntos a favor como:

- Distingue la importancia de partidos.
- Relativiza la dificultad de un partido por las diferencias de nivel.
- Considera la cronología de los encuentros.
- Distingue entre calidad de local o visita de los equipos.
- La contundencia del resultado influye.
- Simple de entender y ejecutar.

Aún así hay ciertas contras:

- A pesar de considerar cronología, no considera temporalidad.
- Es orientada a selecciones.

2.5.6. Método 'Opta'

Opta es la empresa líder en datos deportivos, proporcionando información bastante detallada orientada a equipos. Públicamente proporciona información como goles esperados de partidos, probabilidades de terminar en cada lugar de la tabla de los equipos de las principales ligas, y un *power ranking* unificado de más de 13.500 equipos de 183 países.

El página web de Opta [Opta.com, 2023] explica y describe su método, el cual se basa en el método 'Elo'; sin embargo, el método es explicado de forma general, mostrando sus componentes, pero no su cuantificación.



Figura 2: Fórmula de cálculo de puntos Elo de Opta

Fuente: Opta (<https://theanalyst.com/eu/2023/01/power-rankings-your-club-ranked/>).

Los distintos elementos de esta fórmula son actualizados basados en ciertas condiciones del encuentro realizado:

- Puntuación dentro de la liga (*Within-league team rating*):
Es la puntuación del equipo base, comparada con los equipos de su misma liga. Esta puntuación es actualizada todos los partidos, independiente de que el rival no sea de la misma liga, país o continente.
- Puntuación de la liga (*League rating*):
Es la puntuación de la liga en cuestión, abarcando todos sus equipos. Se actualiza solo en partidos entre equipos del mismo país, pero distinta liga.
- Puntuación del país (*Country rating*):
Es la puntuación del país, incluyendo todas sus divisiones. Esta puntuación se actualiza en solo en partidos entre equipos de distintos países, pero del mismo continente.
- Puntuación del continente (*Continent rating*):
Es la puntuación del continente, englobando a los países que lo conforman. Se actualiza en partidos entre clubes de distintos continentes.

Opta no informa sobre cómo calcula la actualización de puntuaciones, pero sí de ciertas bases de este cálculo que son comunes entre los otros métodos basados en 'Elo':

- El intercambio de puntos es equitativo en ambos equipos, es decir, si el equipo 1 gana 20 puntos, el equipo 2 pierde 20 puntos.
- Si el equipo de mayor puntuación previa al encuentro gana, el intercambio de puntuación es menor.
- Si el equipo con menor puntuación gana, el intercambio de puntos es mayor.
- Si el partido termina en empate, el equipo con menor puntuación ganará una pequeña cantidad de puntos.
- Mientras mayor es el margen de victoria, más puntos son intercambiados.

Finalmente, Opta plantea el *power ranking* desde 0 a 100 aplicando una transformación min-max a las puntuaciones.

Este método tiene ventajas que los otros métodos no abarca a pesar de no exponer los cálculos explícitos:

- Tiene una actualización cronológica del puntaje.
- Posee un enfoque global, independiente de la cantidad de ligas, países y continentes.
- Permite complementar los otros métodos vistos, sobre todo los basados en el sistema 'ELO', ya que el cálculo de Opta también es basado en 'ELO'.
- Permite comparar ligas, países y continentes como una sola unidad, y no solo entre equipos.

En cuanto a las desventajas, es difícil exponerlas ya que no se sabe en detalle que considera en el cálculo o no, como la calidad de local, temporalidad, etc. Por lo que su única desventaja explícita es que su explicación sirve más como un enfoque complementario a otros métodos más que un método directo por si mismo.

2.6. Funciones de predicción

Los *power ranking* permiten cuantificar la calidad de los equipos, y las diferencias entre ellos, pero no predecir directamente la probabilidad de un resultado previo a un encuentro. Es por esto que se necesita de una función para predecir las probabilidades de que gane el equipo A, el equipo B o que haya un empate. A continuación se exponen tres funciones de predicción que utilizan los métodos vistos.

2.6.1. Valor esperado de modelo FIFA 'SUM'

Hay modelos como el FIFA 'SUM', que si bien no cuentan con una fórmula de predicción como tal, sí tienen una fórmula de valor esperado que funcionan como una probabilidad de que gane un equipo u otro. El principal problema de trabajar con los valores esperados como probabilidades de predicción es que son predicciones binarias, es decir, asigna probabilidades a que gane un equipo u otro, pero no al empate, que es un resultado fundamental en el fútbol. Aún así cumplen con la función de predicción, y se puede trabajar el caso del empate con técnicas que se presentarán más adelante.

Como se presentó en la sección 2.5.1, la fórmula de valor esperado de un encuentro del método FIFA 'SUM' es.

$$W_e = 1 / (10^{-\frac{dr}{600}} + 1)$$

donde dr Es la diferencia de *ranking* que se calcula como:

$$dr = P_{\text{before del equipo en cuestión}} - P_{\text{before del equipo rival}}$$

2.6.2. Valor esperado de *ranking* mundial femenino FIFA y eloratings.net

De forma similar al valor esperado del modelo FIFA 'SUM' (sección 2.5.1), los métodos expuestos en *ranking* mundial femenino FIFA (sección 2.5.2) y eloratings.net (sección 2.5.5) también cuentan con la misma fórmula de valor esperado con las mismas implicancias que la respectiva fórmula del modelo FIFA 'SUM'.

$$W_e = 1 / (10^{-\frac{dr}{400}} + 1)$$

Donde dr es la diferencia de clasificación de los equipos, dando 100 puntos extra al equipo que tenga calidad de local.

La estructura de la fórmula es la misma que la del modelo FIFA 'SUM' con dos diferencias: el valor de escalado en este caso es de 400 (en modelo FIFA 'SUM' es de 600) y al equipo con calidad de local se le suma 100 puntos a su *ranking* para abordar la ventaja de local que se observa en el fútbol.

2.6.3. Predicción 'ELO++' adaptada

El modelo 'ELO++' (sección 2.5.3) tiene una función de predicción explícita, que funciona a su vez para determinar el valor esperado de un resultado, esta es:

$$P_{ij} = 1/(1 + e^{-(r_i+h-r_j)})$$

donde

- P_{ij} es la probabilidad de que el equipo i gane al enfrentarse con el equipo j
- r_i es la clasificación (*rating*) del equipo i
- h es un hiperparámetro a definir para describir la ventaja del equipo local, cuando el partido se juega en territorio neutro, $h = 0$.

La fórmula es muy similar a las ya expuestas, con la diferencia de que en vez de elevar 10, se eleva e y que no existe valor de escalado, lo cual hace que esta fórmula de predicción funcione solo dentro del modelo 'ELO++' dadas las dimensiones que maneja.

2.7. Predicción de empates

El empate es un resultado muy común en el fútbol, por lo que reducir la predicción solo a la victoria de un equipo o de otro no es viable. Existen variadas formas de predecir el empate basado en predicciones inicialmente binarias.

2.7.1. Victoria - Derrota

Una forma bastante directa y sencilla es entender el empate como lo mismo de una victoria seguida de una derrota dado el mismo encuentro de forma independiente, de forma que si la función de predicción tiene la forma dada en [Lasek *et al.*, 2013] presentada a continuación:

$$P(s|r_A, r_B) = \frac{(e^{a(r_A-r_B)+h_{(A \text{ es local})}})^s}{1 + e^{a(r_A-r_B)+h_{(A \text{ es local})}}}$$

donde

- $P(s|r_A, r_B)$ es la probabilidad de que se de el resultado s dada las puntuaciones de *ranking* r_A, r_B
- a es un hiperparámetro a definir el valor de escalado de la diferencia de puntuación.
- h es un hiperparámetro a definir para describir la ventaja del equipo local, cuando el partido se juega en territorio neutro, $h = 0$. Este valor siempre considera al equipo A como posible local, si este no es local, B tampoco lo es.

- s es el resultado en cuestión, 1 si se refiere a la victoria de A , 0 si se refiere a la victoria de B .

Dado que los valores a y h son regulables, se puede establecer la misma forma, pero en vez de utilizar e , utilizar 10, resultando en.

$$P(s|r_A, r_B) = \frac{(10^{a(r_A-r_B)+h_{(A \text{ es local})}})^s}{1 + 10^{a(r_A-r_B)+h_{(A \text{ es local})}}}$$

manteniendo las mismas definiciones, por lo tanto, para obtener la probabilidad de obtener una victoria seguido de una derrota de forma independiente, se tiene:

$$\frac{(10^{a(r_A-r_B)+h})}{1 + 10^{a(r_A-r_B)+h}} \times \frac{1}{1 + 10^{a(r_A-r_B)+h}}$$

$$\frac{(10^{a(r_A-r_B)+h})}{(1 + 10^{a(r_A-r_B)+h})^2}$$

$$\sqrt{\frac{(10^{a(r_A-r_B)+h})}{(1 + 10^{a(r_A-r_B)+h})^2}}$$

$$\frac{(10^{a(r_A-r_B)+h})^{0,5}}{1 + 10^{a(r_A-r_B)+h}}$$

De esta forma, el empate se puede definir como $s = 0,5$

El problema con esta técnica es que en las ligas de fútbol actualmente otorgan 3 puntos por victoria, 1 por empate y 0 por derrota, antiguamente se daban 2 puntos por victoria, pero para dar más importancia al estilo de juego ofensivo, y promover el buscar la victoria en vez de conformarse con el empate, que con esa distribución de puntos 2 empates tenían el mismo valor que 1 victoria; la cantidad de puntos por victoria pasó a 3. Por lo que la lógica detrás de esta técnica apunta a lo que la regla de los 3 puntos busca evitar o castigar.

2.7.2. Rao y Kupper

En el trabajo '*Understanding Draws in Elo Rating Algorithm*' [Szczecinski y Djebbi, 2020] se expone brevemente la técnica de Rao y Kupper en su trabajo [Rao y Kupper, 1967]. La idea se reduce en que dada cualquier función que asigne probabilidades binarias de resultado

(Victoria de un equipo o del otro) a partir de una diferencia de puntuación, se reste la diferencia en un cierto valor a definir en el calculo de la victoria del local, y en caso de la victoria del visitante, y la probabilidad de empate será el porcentaje que falta de la suma de las 2 probabilidades iniciales para lograr el 100 %.

Dada una función $F(z)$ para calcular la probabilidad de cierto resultado a partir de la diferencia de puntos z (Local - Visita) y F_H, F_D, F_A la probabilidad de victoria del local, de empate y victoria del visitante respectivamente, el cálculo de probabilidades inicialmente (en el caso de resultados binarios) se obtiene la expresión:

$$F_H(z) = F(z), F_A(z) = F(-z)$$

Aplicando la técnica de Rao y Kupper y considerando los empates, se tiene.

$$\begin{aligned} F_H(z) &= F(z - z_0), F_A(z) = F(-z - z_0) \\ F_D(z) &= 1 - F_H(z) - F_A(z) = F(z + z_0) - F(z - z_0) \end{aligned}$$

2.7.3. Davidson

La técnica Davidson expuesta en el trabajo '*Understanding Draws in Elo Rating Algorithm*' [Szczecinski y Djebbi, 2020] y explicada más en detalle en '*On Extending the Bradley-Terry Model to Accommodate Ties in Paired Comparison Experiments*' [Davidson, 1970], se basa en las fórmulas de probabilidades estándar que pueden ser reescritas de la siguiente forma:

$$F(z) = 1/(1 + 10^{\frac{-z}{\sigma}}) = 10^{\frac{z}{2\sigma}} / (10^{\frac{z}{2\sigma}} + 10^{\frac{-z}{2\sigma}})$$

A esto se le agrega un factor K que es directamente proporcional a la frecuencia de empates, resultando las formulas de predicción tanto de victoria del local, empate y victoria del visitante.

$$\begin{aligned} F_H(z) &= 10^{\frac{z}{2\sigma}} / (10^{\frac{z}{2\sigma}} + 10^{\frac{-z}{2\sigma}} + K) \\ F_A(z) &= 10^{\frac{-z}{2\sigma}} / (10^{\frac{z}{2\sigma}} + 10^{\frac{-z}{2\sigma}} + K) \\ F_D(z) &= K \sqrt{F_H(z) F_A(z)} = K / (10^{\frac{z}{2\sigma}} + 10^{\frac{-z}{2\sigma}} + K) \end{aligned}$$

El valor de K se debe definir y σ es el factor de escalado.

2.8. Métodos de proyección

Las predicciones que se busca abordar no son solo de eventos inmediatos, también se abordan a largo plazo para poder proyectarse en el tiempo y evaluar los resultados más probables.

2.8.1. Simulación Montecarlo

La simulación de Montecarlo es una técnica estadística que se utiliza para modelar la probabilidad de resultados en situaciones que involucran cierto grado de incertidumbre. Se basa en la generación de múltiples muestras aleatorias para modelar el comportamiento de un sistema. La idea fundamental es utilizar la aleatoriedad para simular el comportamiento de un sistema en el que intervienen múltiples variables aleatorias.

Los pasos para realizar una simulación Montecarlo son.

1. Definición de variables aleatorias

Se identifican las variables aleatorias involucradas en el sistema en cuestión. Las variables pueden representar parámetros físicos, financieros, temporales, entre otros dependiendo del contexto del sistema.

2. Asignación de distribuciones de probabilidad

Las variables aleatorias identificadas tienen una distribución de probabilidad asignada que describen su comportamiento.

3. Generación de muestras aleatorias

Mediante algoritmos de generación de números aleatorios se generan múltiples muestras para cada variable aleatoria.

4. Evaluación del modelo

Con las muestras aleatorias generadas, se evalúa el modelo que representa el sistema, realizando los cálculos o simulaciones correspondientes [Murtha, 1997].

5. Análisis de resultados

Los resultados obtenidos de las muestras aleatorias se analizan para así comprender la probabilidad de los diferentes escenarios posibles.

2.8.2. Puntos esperados

dada la base matemática del valor esperado donde.

$$xE = P(E) \times E \quad (11)$$

Donde

- xE Es el valor esperado de E .
- $P(E)$ Es la probabilidad de que ocurra E .
- E Es el valor numérico de E .

De esta manera, el valor esperado xE representa el promedio ponderado de un conjunto de datos, concepto que está estrechamente relacionado con la esperanza matemática, un término acuñado por la teoría de la probabilidad.

Esto llevado a un contexto deportivo, más en concreto el fútbol desde el punto de vista de uno de los dos equipos enfrentados, se tiene.

$$xV = P(V) \times 3 \quad (12)$$

$$xE = P(E) \times 1 \quad (13)$$

$$xD = P(D) \times 0 \quad (14)$$

donde

- xV Es el valor esperado de victoria.
- $P(V)$ Es la probabilidad de victoria.
- xE Es el valor esperado de empate.
- $P(E)$ Es la probabilidad de empate.
- xD Es el valor esperado de derrota.
- $P(D)$ Es la probabilidad de derrota.

De esta forma, los puntos esperados de un equipo en un encuentro se calculan como:

$$xP = xV + xE + xD \quad (15)$$

Si bien es imposible conseguir valores decimales de puntos en un partido, el valor esperado aporta una capa adicional de información y sofisticación al análisis de desempeño de los equipos. Por ejemplo, un equipo que obtiene 1.5 puntos esperados en un partido no significa que obtendrá 1 o 2 puntos exactamente, sino que refleja la probabilidad ponderada de esos resultados. Esto se extrapola a los puntos esperados de una sucesión de partidos, de forma que si faltan 10 encuentros por realizarse para un equipo, la suma de los puntos esperados de los 10 encuentros restantes aportan una proyección de los puntos ponderados que se espera que consiga el equipo aproximadamente.

CAPÍTULO 3

PROPUESTA DE SOLUCIÓN

En el presente capítulo se propone un modelo que se basa en algunos de los modelos, inicialmente independientes, presentados en el capítulo 2, unificándolos y adaptándolos al contexto de fútbol de clubes a nivel nacional y latinoamericano.

La definición y aplicación del modelo sigue la metodología CRISP-DM presentada en 2.3.2, y es aplicada según su estructura a continuación.

3.1. Comprensión del negocio

Para el presente trabajo, la comprensión del negocio se comprende como la comprensión del fútbol, su cuantificación de rendimiento y capacidad predictiva en este contexto. Los cuales fueron expuestos a detalle en el capítulo 1 - Definición del problema.

A modo de breve recapitulación, los objetivos definidos para la memoria son:

1. Diseño y creación de un modelo predictivo dinámico de temporada en el fútbol chileno y aplicación de este en la temporada 2024
2. Visualización de la evolución histórica del Fútbol Chileno a través del Tiempo
3. Unificación de *Power Rankings* en el Fútbol Latinoamericano

Las herramientas a disposición por parte de la minería de datos que aplican para el cumplimiento de los objetivos fueron explicadas en el capítulo 2 - Marco conceptual.

3.2. Comprensión de datos

La extracción de datos se realizó desde dos fuentes la principal es **golesonamores.cl**, un portal web que almacena resultados históricos principalmente del fútbol chileno desde el comienzo de su profesionalismo en 1933 y de competiciones internacionales latinoamericanas. Actualmente está actualizada hasta el año 2019, por esto, todo lo previo a ese año es extraído de esta fuente.

Para los encuentros entre 2020 a la actualidad, se obtienen a partir de registros de **wikipedia.com** en la entrada controlada del campeonato chileno, los datos de esta entrada provienen de **campeonatochileno.cl** y **soccerway.com**, en caso de las competiciones internacionales, los datos provienen de **conmebol.com**.

Los datos en ambas fuentes se presentan mediante tablas, que si bien no tienen una función directa de exportación, el formato tabulado permite una extracción manual sencilla de los datos crudos a hojas de cálculo, en este caso, Google Sheets.

La completitud y años abarcados por competición se expone a continuación:

- **Primera:** Desde 1933 hasta 2024. Se consideran todos los partidos, fase regular, ligüilla, definición de ascenso y definición de título.
- **Primera B:** Desde 1952 hasta 2024. Se consideran todos los partidos, fase regular, ligüilla, definición de ascenso y definición de título.
- **Segunda:** Desde 2012 hasta 2023. Se consideran todos los partidos, fase regular, ligüilla, definición de ascenso y definición de título.
- **Copa Chile:** Desde 1933 hasta 2023. Se consideran todos los partidos, fases preliminares, regionales, grupos, llaves directas.
- **Supercopa:** Desde 2013 hasta 2023. Competición de partido único, se consideran todos.
- **Copa CONMEBOL Libertadores:** Desde 1960 hasta 2023. Se consideran todos los partidos, fases preliminares, grupos, llaves directas.
- **Supercopa Sudamericana:** Desde 1988 hasta 1997. Se consideran todos los partidos, fases preliminares, grupos, llaves directas.
- **Copa Merconorte:** Desde 1998 hasta 2001. Se consideran todos los partidos, fases preliminares, grupos, llaves directas.
- **Copa Mercosur:** Desde 1998 hasta 2001. Se consideran todos los partidos, fases preliminares, grupos, llaves directas.
- **Copa CONMEBOL Sudamericana:** Desde 2002 hasta 2023. Se consideran todos los partidos, fases preliminares, grupos, llaves directas.
- **Recopa Sudamericana:** Desde 1989 hasta 2023. Competición de partido único o ida y vuelta de un enfrentamiento, se consideran todos los partidos.

De estos registros disponen de variables como: jornada, fecha, equipos, resultado, competencia, fase de la competencia. Algunos de los datos extraídos de **golesonamores.cl** contienen información del estadio, asistencia de público y arbitraje. Existen registros puntuales principalmente entre los años 40' y 60' donde no se registra una fecha exacta, solo el mes en que se dio el encuentro.

Las variables utilizadas son: jornada, fecha, equipos, resultado, A local (indicador booleano que indica si el primer equipo es local en el encuentro, o es partido en territorio neutro), competición, división de los equipos, país de los equipos.

3.3. Preparación de datos

Ya extraídos los datos crudos de sus respectivas fuentes, fueron sometidos a los siguientes preprocesamientos:

- Estandarización de formato.

El orden de datos es distinto entre fuentes, ya sea en orden de columnas o diferencias de las mismas y son estandarizados a un formato único con las columnas referentes a las variables a utilizar.

- Revisión de fechas faltantes.

Como fue mencionado, hay años en que se pierde registro de la fecha exacta de los encuentros, y se indica solo el mes, o solo algunos partidos tienen el registro. Para completar la información, se realiza un proceso manual que asigna la misma fecha a encuentros de la misma jornada, y entre jornadas se aplica una diferencia de siete días entre encuentro. La finalidad de la fecha es establecer un orden entre los partidos, por lo que mientras los partidos estén en el orden correspondiente es una aplicación correcta, de esta forma la fecha exacta no es necesaria.

- Diferencias en nombres de equipos.

Al trabajar con dos fuentes de datos, hay ciertas diferencias en los nombres de los equipos, una fuente utiliza abreviaturas que la otra no, una puede referirse a equipos por su nombre completo y la otra fuente un nombre más común y breve (ejemplo: Independiente de Avellaneda - Independiente). Estos casos son identificados manualmente y consolidados a un solo nombre.

- Asignación de divisiones.

Para las competiciones nacionales, determinar la división de equipos por liga es sencillo, todos los equipos de una liga son parte de esa misma división, sin embargo, para copa chile se debe cruzar dicha información ya que se enfrentan equipos de distintas

divisiones, e incluso de divisiones no mapeadas por si solas, de esta forma se define el cruce de divisiones para cada año determinando en qué liga participa un equipo ese mismo año.

- **Asignación de países.**

Para las competencias internacionales se debe identificar a qué país corresponde cada equipo, para esto se obtienen todos los equipos que con aparición en competencias internacionales y se cruza con una tabla que indica a qué país pertenece cada uno.

3.4. Modelado

Ya los datos preparados para ser explotados, se propone un modelo sustentado en algunos de los modelos independientes entre sí expuestos en el capítulo 2 - Marco conceptual, consolidándolos en uno solo con consideraciones extra.

La definición del modelo comprende distintas componentes desglosadas a continuación.

3.4.1. Resultado real

Para asociar un valor al resultado real de un partido disputado se utiliza la tabla definida en 2.5.2 con tal de dar valor a la contundencia de un resultado, de forma que se favorezca una victoria más abultada que una ajustada, o distinguir entre un empate sin goles de un empate con 4 goles por equipo, y en los encuentros muy dispares, donde se tiene un resultado esperado del equipo dominante muy alto, a pesar de la victoria se 'castigue' si es un 1 - 0.

Así, el valor de resultado real de un encuentro depende únicamente del marcador del encuentro.

3.4.2. Resultado esperado

El resultado esperado a utilizar es el definido en 2.7.3, ya que aborda la predicción de empate como un resultado más y no como una reinterpretación de una combinación de empate derrota como se hace en 2.7.1, y es relativa a la diferencia de puntos del encuentro con un factor de empate ajustable K , a diferencia de 2.7.2, que toma el empate como una resta a la diferencia de puntos en la función de probabilidad de victoria del equipo 'A' y de 'B'.

El valor de resultado esperado depende de 3 factores:

- z : Diferencia de puntos entre los equipos enfrentados
- σ : Factor de escalado que 'relativiza' el valor de z
- K : Factor de empate directamente proporcional a la probabilidad de empate.

El comportamiento de las distintas probabilidades dado un cambio relativo de puntos se muestra en la figura 3.

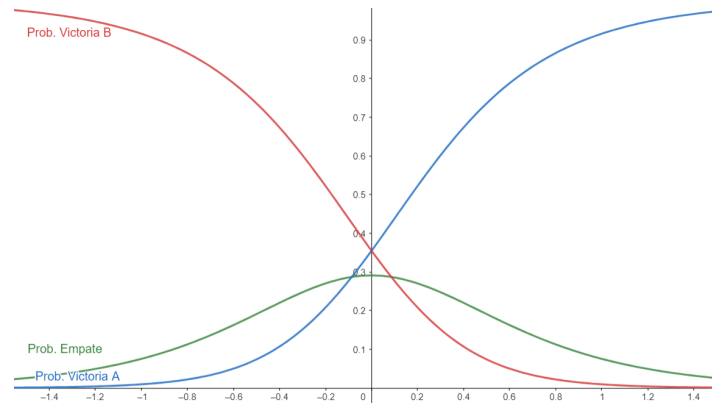


Figura 3: Probabilidades de resultados respecto a diferencia de puntos
Fuente: Elaboración propia.

3.4.3. Variación de puntos

Para la variación de puntos dado un encuentro entre dos equipos se utiliza como base la estructura utilizada en los métodos 'FIFA SUM' (2.5.1), '*ranking* mundial femenino FIFA' (2.5.2) y 'elratings.net' (2.5.5), que es el cálculo fundamental de puntos del sistema 'Elo', resultando en la siguiente expresión de variación de puntos:

$$\text{Puntos nuevos} = \text{Puntos previos} + (\text{Resultado real} - \text{Resultado esperado})$$

Para los encuentros donde el equipo A es local, se le suma un factor de localía h que busca representar la ventaja de jugar de local. En caso de que el encuentro sea en campo neutral, no se aplica dicho factor h .

A diferencia de los métodos base de esta expresión, no se utiliza el factor I , importancia del encuentro, ya que los valores utilizados en estos modelos son respecto a competiciones a nivel selección, por lo que llevarlo a nivel clubes resulta poco intuitivo relativizar la importancia de competiciones y fases con distinto formato que las internacionales. Además de esto, en el trabajo '*FIFA ranking: Evaluation and path forward*' [Szczecinski y Roatis, 2022] se

recomienda eliminar la importancia de los partidos ya que no se aporta ninguna evidencia que sustente los valores dados, y se plantea que puede llegar a ser hasta contraproducente, este estudio se hizo con las importancias de partido a nivel selección, por lo que escalarlo a nivel clubes se vuelve aún más complejo.

En el mencionado trabajo [Szczecinski y Roatis, 2022] se hacen otras dos sugerencias para trabajos de *power ranking*: en caso de que un encuentro se defina por penales, el resultado se considera empate; en fases de eliminación no se aplica la regla de protección que indica que el equipo perdedor no pierde puntos. Dichas sugerencias son aplicadas en el presente trabajo.

3.4.4. Encuentros entre divisiones o países

Tanto en el ámbito nacional como internacional existen escenarios donde los equipos enfrentados no son de la misma división o país, a nivel nacional esto solo ocurre en Supercopa y Copa Chile, especialmente desde el año 2022 en que se incluyeron equipos de divisiones amateur de Tercera B y campeones regionales. A nivel internacional, ocurre la mayoría de ocasiones los equipos enfrentados son de distinto país.

Para abarcar estos casos y determinar como influye el pertenecer a distintas divisiones o países se utiliza la lógica utilizada por Opta, planteada en 2.5.6, que adaptada a los casos abarcados por este trabajo implica lo siguiente:

- **Encuentros nacionales entre divisiones**

El puntaje de cada equipo es el propio de sí mismo dentro de su liga más la puntuación de la división en la que se encuentra. Una vez dado el encuentro, los puntos nuevos de cada equipo se calculan normalmente, y la misma variación aplica a la división de cada equipo.

- **Encuentros internacionales entre países**

Similar a encuentros entre divisiones, cada equipo tiene su puntaje propio, y al enfrentarse entre clubes de distintos países se les suma la puntuación de su país. Luego del encuentro se calcula la nueva puntuación del equipo con normalidad, y la misma variación de puntos aplica a su país.

En caso de que el encuentro ocurra entre equipos de la misma división o país según el contexto, se utiliza el puntaje propio de cada equipo, sin influir el de su división o país.

3.4.5. Predicción de resultados

El cálculo de predicción de resultados es el Davidson, definido en 2.7.3, por las mismas razones expuestas en 3.4.2, además de mantener consistencia en el modelo.

3.4.6. Proyección de temporada

Para la proyección de una temporada activa se utilizan ambos enfoques presentados en 2.8:

'Puntos ponderados' (2.8.2), de forma que se comparen los puntos reales obtenidos hasta la fecha por los equipos, con los puntos esperados de cada uno, utilizando los partidos que quedan por jugar y los valores de resultado esperado definidos en 3.4.2, para ver la variación de posiciones en la tabla que se espera mediante este método.

'Montecarlo' (2.8.1), utilizando la tabla real hasta la fecha, simular cierta cantidad de escenarios donde se disputan el resto de partidos por jugar en la temporada, y con el paso de las simulaciones contabilizar las veces en que cada equipo termina a final de temporada en cada posición, para así calcular la probabilidad de que ocurra cada escenario dividiendo por la cantidad de simulaciones totales.

3.5. Evaluación

A continuación se define cómo se realiza la evaluación del modelo, la aplicación de la misma se realiza en el Capítulo 4.

3.5.1. Definición de hiperparámetros

En la sección 3.4 quedaron por definir tres hiperparámetros: σ (factor de escalado), h (factor de localía) y K (factor de empate). Estos tres hiperparámetros influyen en los cálculos de valor esperado y predicción de resultado, y estos cálculos se utilizan para determinar los valores a definir.

Al utilizar un conjunto considerable y no sesgado de registros, las diferencias de puntos se distribuyen de forma pareja en el sentido de que el ponderado de ventaja de A ($z > 0$) tiende a ser igual al ponderado de ventaja de B ($z < 0$) por distribución normal de grandes número y los formatos de ida y vuelta en la mayoría de los partidos, por lo que se puede considerar

un $z = 0$.

Utilizando el mismo conjunto de registros, se calcula el porcentaje de ocurrencia de cada resultado (Victoria A, Empate, Victoria B), y se utiliza como la probabilidad de que ocurra cada uno de los resultados, resultando la siguiente expresión:

$$F_H(0) \approx (\% \text{ Ocurrencia Victoria A}) \approx 10^{\frac{h}{2\sigma}} / (10^{\frac{h}{2\sigma}} + 10^{\frac{-h}{2\sigma}} + K)$$

$$F_A(0) \approx (\% \text{ Ocurrencia Empate}) \approx 10^{\frac{-h}{2\sigma}} / (10^{\frac{h}{2\sigma}} + 10^{\frac{-h}{2\sigma}} + K)$$

$$F_D(0) \approx (\% \text{ Ocurrencia Victoria B}) \approx K / (10^{\frac{h}{2\sigma}} + 10^{\frac{-h}{2\sigma}} + K)$$

Si bien se tienen tres ecuaciones y tres incógnitas, las ecuaciones no son linealmente independientes, lo que impide aplicar un sistema de ecuaciones, por lo que el hiperparámetro σ se define como un valor a iterar, de forma que solo queden como incógnitas de un sistema los valores de h y K , el sistema en cuestión es el siguiente:

$$(\% \text{ Ocurrencia Victoria A}) \approx 10^{\frac{h}{2\sigma}} / (10^{\frac{h}{2\sigma}} + 10^{\frac{-h}{2\sigma}} + K)$$

$$(\% \text{ Ocurrencia Empate}) \approx K / (10^{\frac{h}{2\sigma}} + 10^{\frac{-h}{2\sigma}} + K)$$

Iterando el valor de σ se puede validar el modelo mediante las siguientes tres métricas de evaluación:

3.5.2. Medición de errores

■ Matriz de probabilidades de resultados por rango

Esta métrica se propone como una evaluación de modelo con un enfoque diferente al de las métricas ya expuestas, pues parte de la premisa de que si un evento tiene un 50 % de probabilidades de ocurrir, se espera que si se repite el escenario 100 veces, el evento debería ocurrir aproximadamente 50 veces, o lo que es igual, si dado un escenario, un evento tiene 30 % de probabilidades de ocurrir, si el escenario se repite 1000 veces, se espera que ocurra el evento aproximadamente 300 veces de las 1000.

De forma que se busca favorecer a que esto ocurra, y no a predecir con más probabilidad el resultado real, dado que en el contexto del fútbol, no siempre se cumple el resultado con mayor probabilidad, hay veces en que se le da un 10 % de probabilidades de ganar al Sheriff cuando se enfrenta por *Champions League* al Real Madrid, y el Sheriff gana, eso no significa que la predicción sea mala, si no que si se repite el partido 10 veces, se espera que gane el Sheriff solo gane una vez, y empaten o pierdan las otras 9.

Esta perspectiva es dejada de lado por las otras métricas, ya que son más utilizadas en problemas de clasificación donde el resultado es solo uno y hay que determinar con que certeza lo es, por ejemplo, dada una foto, un modelo debe dar un porcentaje de certeza de que la foto es de un perro, pero la foto ya de por sí es de un perro, por lo que si se repite infinitamente la predicción, la foto siempre será de un perro.

La métrica propuesta consta en construir una matriz de 3 filas, una por cada resultado (Victoria equipo A, empate y victoria de equipo B) y 10 columnas, una por cada grupo de probabilidades, en este caso se hace utilizan grupos de rango 10 %, e inicialmente se registran las probabilidades asignadas a cada escenario previo a un conjunto de partidos, un caso de ejemplo se presenta en la tabla 2.

Matriz de casos por rango										
	0-10	10-20	20-30	30-40	40-50	50-60	60-70	70-80	80-90	90-100
Victoria A	300	700	1000	1300	1700	1400	800	600	400	300
Empate	100	400	4000	3000	1000					
Victoria B	800	1400	2600	1600	800	600	400	300		

Tabla 2: Matriz de casos por rango de ejemplo.
Fuente: Elaboración propia.

En dicha matriz, cada partido predicho suma uno en cada fila, sumando cada fila el total de partidos.

Luego, por cada partido predicho se arma otra matriz que indica la ocurrencia de un resultado en el rango predicho de este, por ejemplo, en la tabla 3 se registra la ocurrencia de los eventos 'predichos' en la tabla 2.

Matriz de resultados por rango										
	0-10	10-20	20-30	30-40	40-50	50-60	60-70	70-80	80-90	90-100
Victoria A	55	130	90	600	560	700	380	490	190	230
Empate	20	120	680	250	340					
Victoria B	730	600	800	500	300	410	50	275		

Tabla 3: Matriz de resultados por rango de ejemplo.
Fuente: Elaboración propia.

En este caso, cada partido suma uno en solo una casilla de toda la matriz, si por ejemplo, se predijo que victoria B tenía un 33 % de probabilidades y ocurre, se suma uno a la casilla de fila 'Victoria B' y columna '30 - 40', sumando así toda la matriz la totalidad de resultados.

Ahora, calculando el porcentaje de ocurrencia de cada evento y su respectivo rango se obtiene la matriz de porcentajes de ocurrencia presentada en la tabla 4.

Matriz de porcentajes de ocurrencia por rango										
	0-10	10-20	20-30	30-40	40-50	50-60	60-70	70-80	80-90	90-100
Victoria A	18,3%	18,6%	9,0%	46,2%	32,9%	50,0%	47,5%	81,7%	47,5%	76,7%
Empate	20,0%	30,0%	17,0%	8,3%	34,0%					
Victoria B	91,3%	42,9%	30,8%	31,3%	37,5%	68,3%	12,5%	91,7%		

Tabla 4: Matriz de porcentajes de ocurrencia por rango de ejemplo.

Fuente: Elaboración propia.

Y para calcular el índice de cada celda, se calcula la distancia absoluta entre la probabilidad de ocurrencia obtenida y la marca de clase del rango respectivo, obteniendo la matriz de índices de probabilidades por rango de la tabla 5.

Matriz de índices de probabilidades por rango										
	0-10	10-20	20-30	30-40	40-50	50-60	60-70	70-80	80-90	90-100
Victoria A	0,1333	0,0357	0,1600	0,1115	0,1206	0,0500	0,1750	0,0667	0,3750	0,1833
Empate	0,1500	0,1500	0,0800	0,2667	0,1100					
Victoria B	0,8625	0,2786	0,0577	0,0375	0,0750	0,1333	0,5250	0,1667		

Tabla 5: Matriz de índices de probabilidades por rango de ejemplo.

Fuente: Elaboración propia.

Para condensar esta matriz en un solo número, se hace un promedio ponderado de los índices de cada resultado posible, donde los pesos son los casos que se predijo cada escenario en su rango respectivo. En el ejemplo trabajado, los índices ponderados por resultados se exponen en la tabla 6

Índices de probabilidades ponderado	
	índice ponderado
Victoria A	0,1212
Empate	0,1535
Victoria B	0,1988

Tabla 6: Matriz de índices de probabilidades ponderada de ejemplo.

Fuente: Elaboración propia.

Finalmente, promediando estos tres índices ponderados se obtiene el índice de probabilidades de resultados por rango, que en el ejemplo desarrollado, resulta en **0,1578**.

■ Brier score medio

Se utiliza *brier score* presentado en el capítulo 2.2, punto 6, pero calculando su promedio, de forma que el calculo es:

$$BrierScoreMedio = \frac{\sum_{i=1}^n (BrierScore)}{n} \quad (16)$$

- **Ranked Probability Score medio**

Similar al *brier score* medio, se utiliza *Ranked Probability Score* presentado en el capítulo 2.2, punto 7, pero calculando su promedio, de forma que el calculo es:

$$RankedProbabilityScoreMedio = \frac{\sum_{i=1}^n (RankedProbabilityScore)}{n} \quad (17)$$

3.6. Despliegue

La fase de despliegue consta en la aplicación y ejecución de lo definido en el 3.4, y se desarrolla a continuación en el capítulo 4 junto a la ejecución de la evaluación.

CAPÍTULO 4

EVALUACIÓN Y DESPLIEGUE

Para la evaluación de la propuesta se determinarán los valores de los parámetros σ , h y K y a partir de estos, obtener los resultados asociados a los objetivos específicos planteados.

4.1. Evaluación

Como se planteó en 3.5, se debe definir un σ para a partir de este definir h y K mediante el sistema

$$F_A(0) \approx 0,4794 \approx 10^{\frac{h}{2\sigma}} / (10^{\frac{h}{2\sigma}} + 10^{\frac{h}{2\sigma}} + K)$$

$$F_D(0) \approx 0,2802 \approx K / (10^{\frac{h}{2\sigma}} + 10^{\frac{h}{2\sigma}} + K)$$

Iterando distintos σ , se obtienen los valores presentados en la tabla 7 de las métricas 'Matriz de probabilidades de resultados por rango', 'Ranked probability Score medio' y Brier score medio'.

sigma	h	k	índice de probabilidades de resultados	Ranked Probability Score Medio	Brier Score Medio
10	2,99829	0,82554	0,02124	0,45177	0,5296
50	14,99144	0,82554	0,01792	0,44925	0,5190
100	29,98288	0,82554	0,01685	0,44537	0,5097
200	59,96576	0,82554	0,01609	0,44171	0,5072
300	89,94865	0,82554	0,01522	0,43484	0,5015
350	104,94009	0,82554	0,01438	0,43073	0,4986
380	113,93495	0,82554	0,01384	0,42994	0,4995
400	119,93153	0,82554	0,01407	0,42945	0,5002
500	149,91441	0,82554	0,01457	0,43052	0,5033
600	179,89729	0,82554	0,01497	0,43378	0,5063
800	239,86305	0,82554	0,01545	0,43742	0,5101
1.000	299,82882	0,82554	0,01574	0,43921	0,5124

Tabla 7: Calidad de algoritmo por σ asignado.
Fuente: Elaboración propia.

La intención es minimizar las tres mediciones de error. En el caso del 'índice de probabilidad-

des de resultados' se observa que de los valores iterados, el que da mejor índice de probabilidades es $\sigma = 380$, y se trabajará con este a continuación.



Figura 4: Índice de probabilidades de resultados respecto a σ .

Fuente: Elaboración propia.

Con este σ , se obtiene la matriz de casos por rango presentada en la tabla 8.

Matriz de casos por rango										
	0-10	10-20	20-30	30-40	40-50	50-60	60-70	70-80	80-90	90-100
Victoria A				19	1212	8462	3106	483		
Empate				13282						
Victoria B		1680	10222	1347	33					

Tabla 8: Matriz de casos por rango con $\sigma = 380$.

Fuente: Elaboración propia.

Esta matriz indica que de los 13282 partidos, en 8462 se predijo una probabilidad de victoria del local entre un 40 y un 50 %, de la misma forma, en todos los partidos la probabilidad de empate se predijo entre un 20 y un 30 % y en 10222 se proyectaba una probabilidad de victoria de B entre un 20 y un 30 %.

De todas predicciones por rango y su respectiva ocurrencia se obtiene la matriz de resultados por rango mostrada en la tabla 9

Matriz de resultados por rango										
	0-10	10-20	20-30	30-40	40-50	50-60	60-70	70-80	80-90	90-100
Victoria A				5	404	3808	1670	301		
Empate				3536						
Victoria B			294	2706	550	18				

Tabla 9: Matriz de resultados por rango con $\sigma = 380$.

Fuente: Elaboración propia.

La matriz de resultados por rango indica que de los 8462 partidos que se predijo entre un 40 y un 50 % de probabilidades de victoria de A, en 3808 el resultado fue la victoria de A. De todos los partidos, en 3536 el resultado fue el empate y de los 10222 que se proyectaba una probabilidad de victoria de B entre un 20 y un 30 %, en 2706 el resultado fue la victoria de B.

Calculando el porcentaje de ocurrencia de cada resultado y sus rangos, se obtiene lo mostrado en la tabla 10.

Matriz de porcentajes de ocurrencia por rango										
	0-10	10-20	20-30	30-40	40-50	50-60	60-70	70-80	80-90	90-100
Victoria A			26,3157	33,3333	45,0011	53,7669	62,3188			
Empate			26,5472							
Victoria B		17,5000	26,4723	40,8314	54,5454					

Tabla 10: Matriz de porcentajes de ocurrencia por rango con $\sigma = 380$.

Fuente: Elaboración propia.

Y comparando la diferencia absoluta entre los porcentajes obtenidos y la marca de clase del rango que se encuentran, se obtiene la matriz de índices de probabilidades por rango mostrado en la tabla 11.

Matriz de índices de probabilidades por rango										
	0-10	10-20	20-30	30-40	40-50	50-60	60-70	70-80	80-90	90-100
Victoria A			0,0131	0,0161	0,0011	0,0123	0,0268			
Empate			0,0154							
Victoria B		0,0250	0,0146	0,0584	0,0954					

Tabla 11: Matriz de índices de probabilidades de resultados con $\sigma = 380$.

Fuente: Elaboración propia.

Los dos de los tres peores índices obtenidos, en concreto Victoria A 60-70 y Victoria B 40-50, son rangos con menos de 500 casos de representatividad, 483 y 33 respectivamente, por lo que en el índice de probabilidades de resultados ponderado influyen poco, a diferencia del índice de Victoria A, que es de 0,0011 y representa a 8462 escenarios de los 13282 totales.

Ponderando los índices por resultado se tienen los valores presentados en la tabla 12.

Índices de probabilidades ponderado	
	índice ponderado
Victoria A	0,0060
Empate	0,0154
Victoria B	0,0206

Tabla 12: Matriz de índices de probabilidades de resultados ponderados con $\sigma = 380$.
Fuente: Elaboración propia.

Lo que indica que el modelo predice mejor la victoria de A que el empate o Victoria B.

Finalmente, el promedio de estos índices resulta en el índice de probabilidades de resultados final con $\sigma = 380$ es aproximadamente 0,01384.

Los valores de *Ranked probability score* medio según el valor de σ se muestra gráficamente en la figura 5

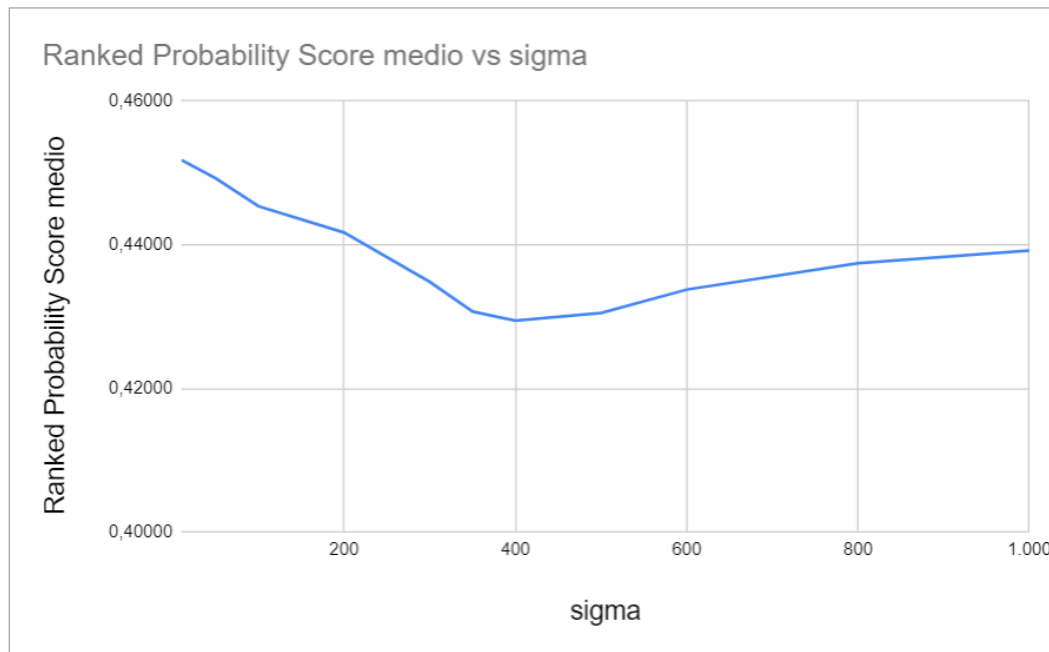


Figura 5: *Ranked probability score* medio respecto a σ .
Fuente: Elaboración propia.

Las tendencias de esta métrica es similar a la obtenida del índice de probabilidades de resultados en la figura 4, obteniendo su mínimo en valores similares.

La misma visualización para *Brier score* medio se presenta en la figura

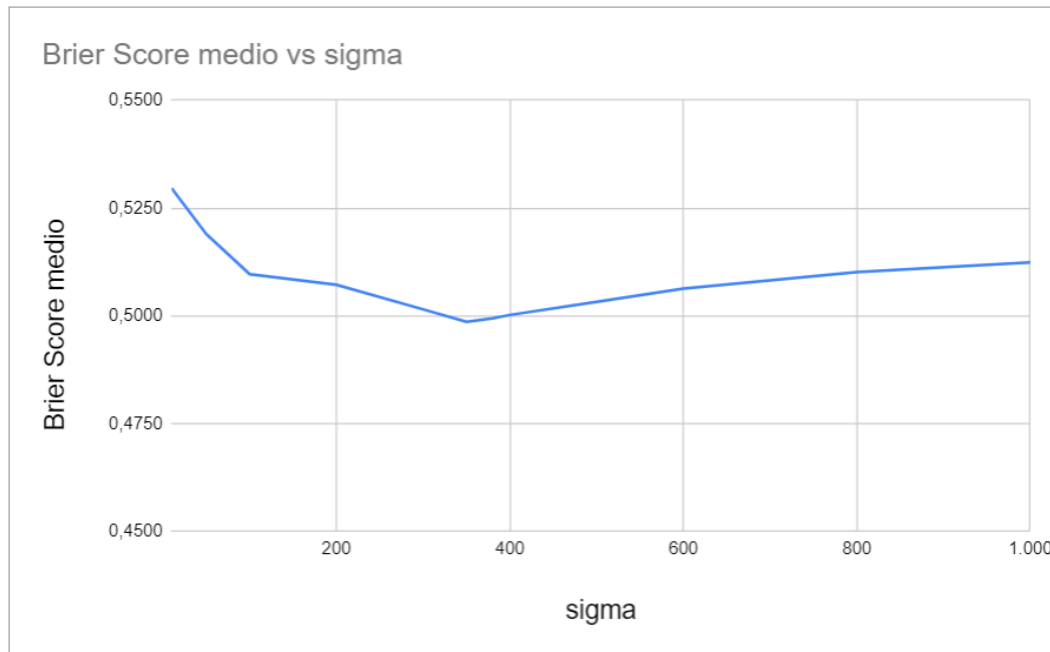


Figura 6: *Brier score* medio respecto a σ .
Fuente: Elaboración propia.

La tendencia se repite, variando el punto mínimo exacto, pero las distribuciones se asemejan entre las métricas al igual que el rango aproximado con mejores resultados.

Con lo expuesto, se trabaja a continuación la fase de despliegue 4.2 con $\sigma = 380$, que implica un valor de $h = 113,93495$ y $k = 0,82554$.

Con estos valores se obtiene el evolutivo de *power ranking nacional* de la figura 7. A modo de comparación y comparativo según la variación del hiperparámetro σ , en la figura anexa 21 se muestra el mismo evolutivo pero utilizando un $\sigma = 100$. Se observa que las variaciones de puntos son más drásticas y volátiles. Por otro lado, el evolutivo obtenido utilizando $\sigma = 800$ se observa en la figura anexa 22, en este caso los cambios de puntos son leves y conservadores.

Esto último indica que el valor de σ controla que tanta importancia se le da al rendimiento en el corto plazo o a lo largo del tiempo, siendo que al disminuir el valor del hiperparámetro, el *power ranking* sea un indicativo del rendimiento del equipo los últimos encuentros; y al aumentar su valor, el *power ranking* tienda a mostrar el rendimiento mantenido en el tiempo del equipo, haciendo que la historia pese más que la actualidad.

MODELO PREDICTIVO DE FÚTBOL CHILENO, SU CONTEXTO NACIONAL Y LATINOAMERICANO BASADO EN
POWER RANKING HISTÓRICOS.

Posición	Equipo	División	Puntos Power Ranking	Clasificación
1	Colo-Colo	Primera	1143,03	100,00
2	Universidad Católica	Primera	1125,40	90,98
3	Universidad de Chile	Primera	1101,88	78,95
4	Cobreloa	Primera B	1066,12	60,66
5	Unión Española	Primera	1058,36	56,70
6	O'Higgins	Primera	1032,20	43,32
7	Audax Italiano	Primera	1028,96	41,66
8	Everton	Primera	1027,42	40,87
9	Huachipato	Primera	1026,27	40,29
10	Cobresal	Primera	1023,62	38,93
11	Palestino	Primera	1022,83	38,53
12	Universidad de Concepción	Primera B	1014,25	34,14
13	Santiago Wanderers	Primera B	1011,23	32,59
14	Deportes Antofagasta	Primera B	1010,94	32,45
15	Deportes Temuco	Primera B	1007,00	30,43
16	San Antonio Unido	Segunda	1006,85	30,36
17	Deportes Limache	Segunda	1006,12	29,98
18	Deportes Concepción	Segunda	1003,18	28,48
19	Deportes Iquique	Primera B	1000,54	27,13
20	Deportes Puerto Montt	Primera B	998,82	26,25
21	Coquimbo Unido	Primera	998,79	26,23
22	Lautaro de Buin	Segunda	997,46	25,55
23	Fernández Vial	Segunda	997,14	25,39
24	Deportes Rengo	Segunda	996,78	25,21
25	Real San Joaquín	Segunda	996,71	25,17
26	Curicó Unido	Primera	996,39	25,01
27	Deportes Recoleta	Primera B	995,58	24,59
28	Trasandino	Segunda	995,25	24,43
29	San Marcos de Arica	Primera B	994,15	23,86
30	Barnechea	Primera B	993,00	23,27
31	Santiago Morning	Primera B	992,44	22,99
32	Nublense	Primera	992,38	22,96
33	Deportes Melipilla	Segunda	992,26	22,90
34	Deportes La Serena	Primera B	992,17	22,85
35	Provincial Osorno	Segunda	991,70	22,61
36	Rangers	Primera B	990,80	22,15
37	Deportes Santa Cruz	Primera B	989,85	21,66
38	San Luis de Quillota	Primera B	988,87	21,16
39	Deportes Copiapó	Primera	988,25	20,85
40	Magallanes	Primera	987,53	20,47
41	Deportes Valdivia	Segunda	987,50	20,46
42	Unión La Calera	Primera	986,46	19,93
43	General Velásquez	Segunda	985,02	19,19
44	Unión San Felipe	Primera B	983,46	18,39
45	Iberia	Segunda	959,31	6,04
46	Deportes Linares	Segunda	947,49	0,00

Tabla 13: Tabla de *power ranking* nacional de equipos en las tres primeras divisiones en 2023.

Fuente: Elaboración propia.

Para facilitar la lectura de datos y minimizar el solapamiento de nombres de los equipos, se grafica por separado los equipos que en 2023 estaban en primera A, primera B y segunda.

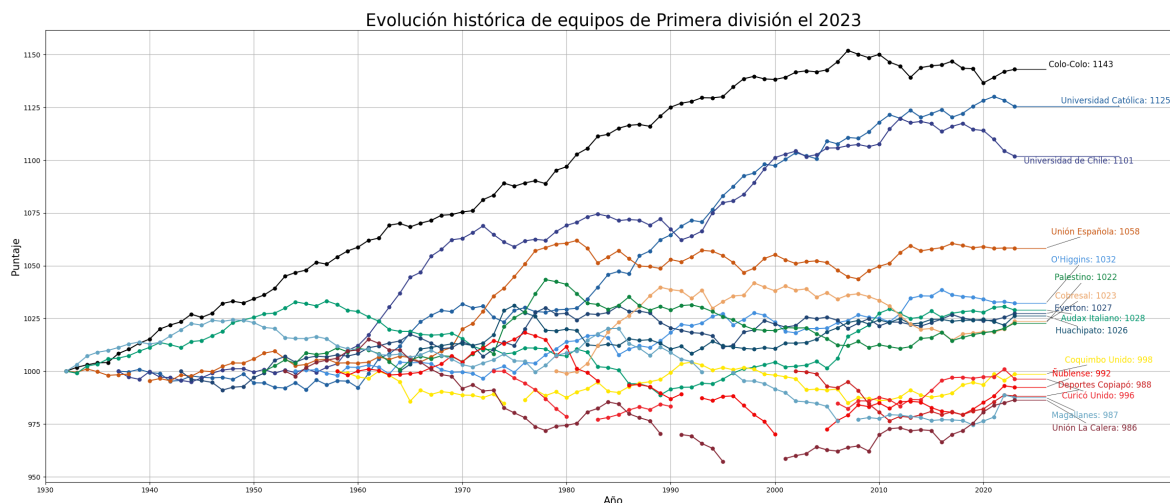


Figura 8: *Power ranking* histórico nacional de equipos de primera A en 2023.
Fuente: Elaboración propia.

La evolución histórica del *power ranking* de equipos en primera destaca primero por la diferencia de Colo-Colo, U. Católica y U. de Chile del resto de equipos, además de identificar en qué momento de la historia los equipos ganaron su 'estatus' que mantuvieron en el tiempo.

En el caso de Colo-Colo ha sido un crecimiento constante a lo largo de la historia, destacando procesos especialmente buenos con aumentos considerables en pocos años, como también bajas en el rendimiento al comienzo de la década del 2010 y en el 2020.

La Universidad Católica tuvo unos primeros años sutiles hasta mediados de la década de los 60', donde vieron un aumento en el rendimiento que los posicionó de buena forma hasta los años 80' donde comenzó el crecimiento constante hasta los últimos años, destacando el final de la década del 2010, donde lograron el tricampeonato del torneo chileno.

La Universidad de Chile también tuvo unos primeros años sutiles hasta los años 60' donde mostraron un aumento explosivo de rendimiento, que los dejó en una posición que mantuvieron hasta los años 90, donde sufrieron una bajada de rendimiento que les significó el descenso de la máxima categoría chilena para luego comenzar su mejora hasta la década del 2010, donde ganaron el título internacional de la Copa Sudamericana, desde entonces el rendimiento se mantuvo, pero sufriendo una baja significativa los últimos años.

Unión española también mostró inicios normales, sin embargo en la década del 70' mostraron una mejora de rendimiento tal que los posicionó hasta el día de hoy en valores similares

en cuanto a *power ranking*.

Destaca también la existencia de 2 grupos de equipos más, la conformada por O'higgins, Palestino, Cobresal, Everton, Audax Italiano y Huachipato, equipos regulares en el campeonato. El segundo equipo conformado por Coquimbo Unido, Ñublense, Deportes Copiapó, Curicó Unido, Magallanes y Unión la Calera destaca por ser equipos que ascendieron a la primera categoría del fútbol chileno en años más recientes.

Dentro de este segundo grupo destacan procesos como el de Ñublense y Unión la Calera, que mostraron mejoras explosivas de rendimiento que significó el ascenso a la Primera División, disputando lugares altos de la tabla ciertos años y en otros manteniéndose en la división.

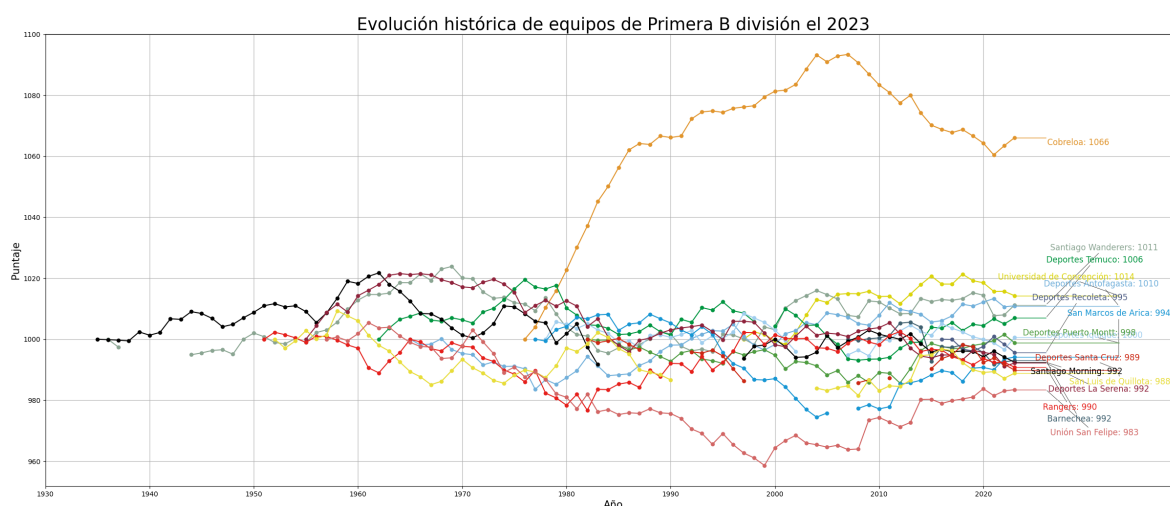


Figura 9: *Power ranking* histórico nacional de equipos de primera B en 2023.

Fuente: Elaboración propia.

Lo que más destaca es la evolución del *power ranking* de equipos de la Primera B es la distancia entre Cobreloa y el resto de equipos. Cobreloa desde su fundación en 1977 estuvo en un ascenso hasta mediados de los 2000, donde alcanzan su punto más alto para luego entrar en una bajada de rendimiento que se vio reflejada en la tabla y su eventual descenso el 2015. Cobreloa asciende a la Primera división el mismo 2023, y el año previo quedó segundo en la tabla y perdió el partido definitorio para ascender, pero su rendimiento a lo largo del año fue sobresaliente. Estos 2 últimos años derivan en un aumento de puntos en el *power ranking*, frenando la tendencia negativa que se traía hasta entonces.

Lo siguiente que destaca es la agrupación que existe entre los otros equipos, que en general mantienen rendimientos estables, con sus respectivas subidas y bajadas con el paso de los años, pero viéndose cierta tendencia a la baja de estos equipos por los años.

80. Dentro de estos, el equipo más veterano es Santiago Morning con inicios en ascenso pero luego cayendo en los años 80 donde se les pierde el rastro en lo que concierne a los datos estudiados, ya que en 1982 descienden de Primera a Primera B y el año siguiente descienden a lo que ese entonces era la Tercera división (No considerada en los datos de estudio) hasta 1996, donde ascienden de Tercera división a Primera B.

Otro equipos que destaca del resto, es Unión San Felipe, que si bien los últimos años a estado acercándose al rendimiento agrupado del resto de equipos, estuvo muchos años por debajo del rendimiento de ese entomces del resto de equipos. Estos años oscuros del equipo desvían la atención del año más importante de San Felipe, 1971 son campeones de la Primera División chilena, y no solo eso, el año previo fueron campeones de la Primera B, siendo el único equipo en ser campeón de Primera B y Primera de forma consecutiva. lamentablemente, los años posteriores fueron complejor para Unión San Felipe, pelenado el descenso hasta la última fecha en 1973 y en 1974 descendiendo, donde los años siguientes no fueron fáciles.

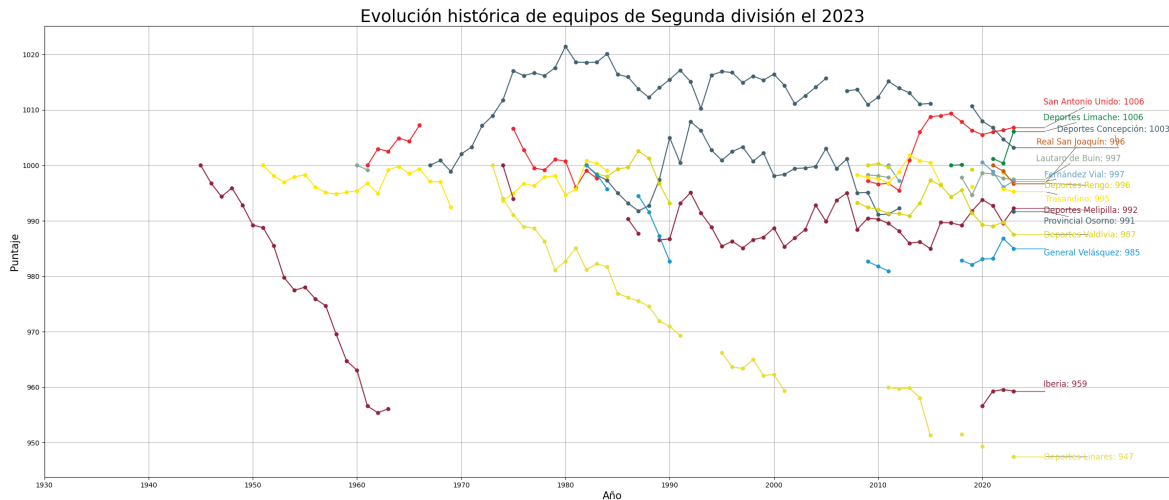


Figura 10: *Power ranking* histórico nacional de equipos de segunda en 2023.
Fuente: Elaboración propia.

El evolutivo de puntos del **power ranking** de equipos participantes de Segunda división en 2023 es distinto a los vistos hasta el momento, está compuesto de equipos que debutaron en el profesionalismo en años más recientes o de equipos veteranos cuyo rendimiento fue a la baja con los años. Esto sumado con la pérdida de seguimiento por descensos y ascensos con la división inmediatamente menor a la ahora Segunda división profesional.

El equipo que ha logrado un mejor rendimiento de estos es Deportes Concepción, que formó parte de la Primera división en los años 70, luego descendiendo y ascendiendo intermitentemente hasta 2006, donde por problemas financieros se le impone un castigo que impidió la participación del equipo en competiciones de la ANFP. Al año siguiente

vuelven a participar en la Primera División, pero los problemas administrativos continuaron hasta un descenso por secretaría a Primera B en 2008. Los problemas siguieron hasta llegar a la Tercera B en 2018 y logrando el regreso al profesionalismo en 2020 (Se registran partidos en 2019 dado que Deportes Concepción fue invitado a participar en Copa Chile dicho año).

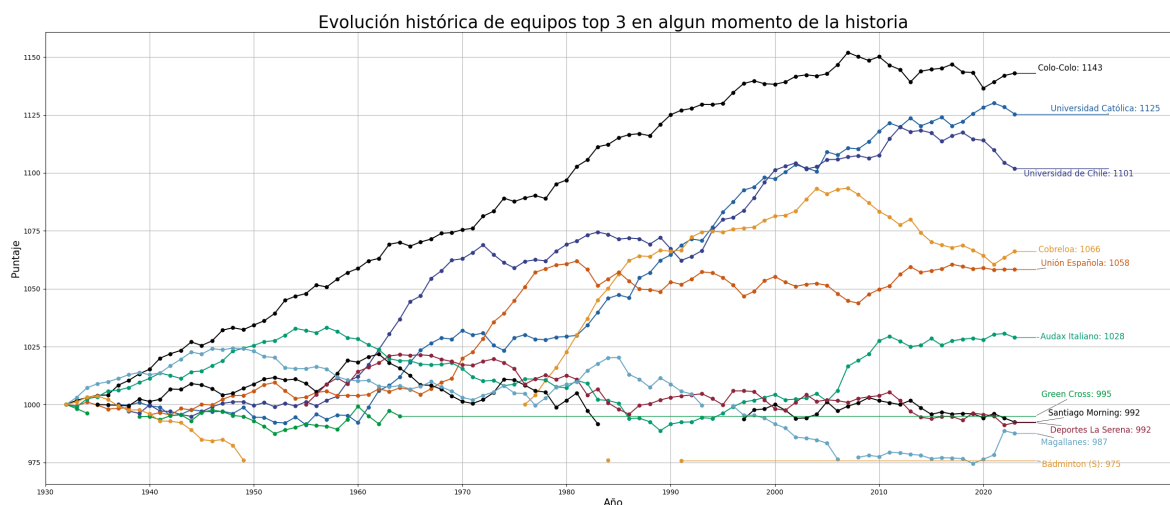


Figura 11: *Power ranking* histórico nacional de equipos que alguna vez fueron top 3 *power ranking*.

Fuente: Elaboración propia.

Los equipos mostrados en el gráfico son todos los que algún año terminaron dentro de los 3 primeros puestos en puntos del *power ranking* a nivel nacional, cualquier valoración que implique competencias internacionales no es considerada en el análisis.

Lo más destacado es el dominio de Colo-Colo desde 1940 hasta la actualidad, el equipo que más campeonatos de primera división tiene (33), ha sido campeón de la misma todas las décadas y el único que nunca ha descendido de esta. Siendo también más ganador de la Copa Chile (14).

Hasta 2023, el segundo equipo con más puntos en el *power ranking* nacional es Universidad Católica que a mediados de la década del 2010 logró separarse de la universidad de Chile, ya que fueron grandes años para U. Católica y años oscuros para U. de Chile. En dicha década U. Católica logra el Tetracampeonato en los años 2018, 2019, 2020, 2021. Siendo el Segundo equipo en lograrlo después de Colo-Colo en los torneos Apertura 2006, Clausura 2006, Apertura 2007 y Clausura 2007, y el primero en lograrlo en torneos largos. Es el tercer equipo con más campeonatos de Primera división (18) y el segundo que más veces ha sido subcampeón del mismo.

Universidad de Chile es el segundo equipo con más campeonatos de Primera división (18), teniendo 6 de estos entre los años 1959 y 1969, años donde ocurre su primer aumento de puntos considerable en el *power ranking*, el segundo ocurre en los años 90', década donde gana 3 campeonatos.

Cobreloa tuvo un breve periodo donde fue de los 3 equipos con más puntos en el *power ranking* entre los años 1986 y 1993. Cobreloa es el cuarto equipo con más campeonatos de primera división (8), 4 de ellos obtenidos en la década del 80' y 3 de ellos en 2003 y 2004 (años de puntos más altos en el *power ranking*).

Entre 1973 y 1985 Unión española se posicionó como tercer equipo con más puntos en el *power ranking*, entre estos años fue campeón 3 veces de sus 7 en total. y su último año como campeón fue el 2013, donde muestra uno de sus mejores años en cuanto a puntos en el *power ranking*.

Audax Italiano fue el tercer equipo mejor valorado en el *power ranking* desde los comienzos del profesionalismo chileno hasta 1949, años donde gana 3 de sus 4 campeonatos, y pasa a ser el segundo mejor valorado hasta 1962, en este periodo solo fue campeón una vez (1957). Desde entonces ha tenido un paso discreto hasta mediados de la década del 2000, donde tuvo una clara mejora que lamentablemente no se tradujo en campeonatos.

Green Cross y Bádminton aparecieron entre los 3 mejores en los primeros años del profesionalismo chileno de forma efímera, luego ambos equipos fueron en bajada de rendimiento hasta sus respectivas disoluciones. Green Cross fue disuelto en 1965 y fusionado con Club Deportes Temuco. Bádminton fue disuelto 2 veces, una vez en 1949, donde para evitar el descenso se fusionó con Ferroviarios, formando el Club Deportivo Ferrobádminton, dicha fusión fue disuelta en 1969 y sufrió una nueva fusión en 1973, formando a Curicó Unido.

Santiago Morning fue el tercer equipo con mejor puntuación en el *power ranking* por un breve periodo de 4 años entre 1958 y 1961, años en los que venían de un aumento de rendimiento que no se vió reflejado en campeonatos de Primera división.

Deportes la Serena aparece en esta visualización dado un periodo puntual en entre 1963 y 1965, años donde terminó en posiciones positivas en la tabla de primera división pero que no pasaron de un tercer lugar en 1963.

Magallanes fue el primer equipo campeón de Primera división (1933) y primer tricampeón de la competición (1933, 1934, 1935), lo que se ve reflejado en el *power ranking* siendo el único equipo que tuvo el primer lugar en puntos fuera de Colo-Colo. Esto hasta 1940, desde entonces se mantuvo en el top 3 hasta 1958, teniendo un repunte en los 80' que luego fue en bajada hasta 2022, donde ascienden a Primera división luego de 36 años en una campaña de 22 victorias en Primera B y salir campeones de Copa Chile Y Super Copa. Este singular año se refleja como un aumento muy considerable en puntos del *power ranking*.

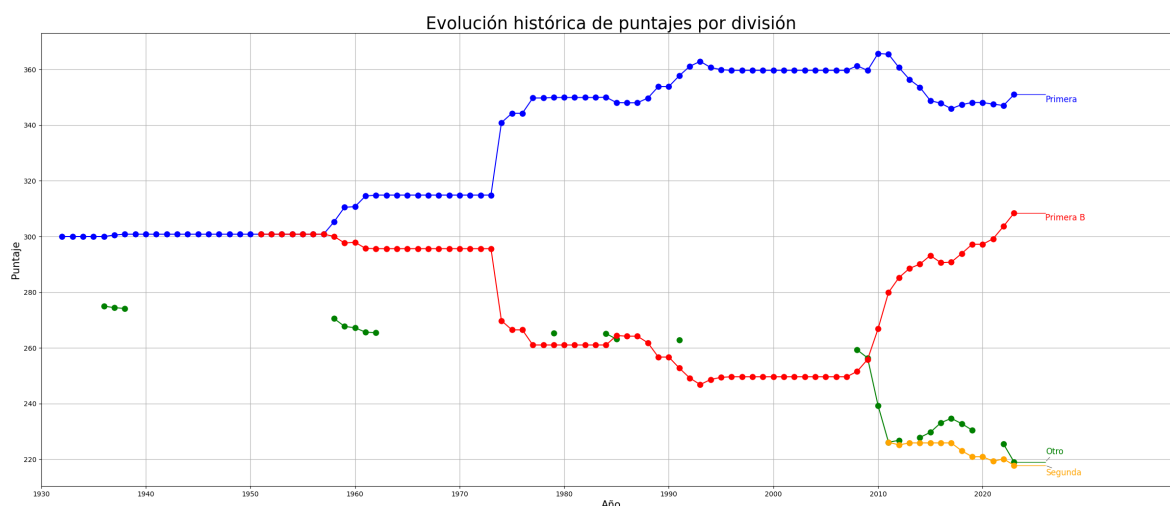


Figura 12: *Power ranking* histórico nacional de las 3 primeras divisiones.
Fuente: Elaboración propia.

Al englobar el rendimiento por divisiones a la hora de enfrentarse entre sí, se ve una clara brecha entre Primera y Primera B que ha ido disminuyendo en los últimos años. La brecha entre Primera B y categorías inferiores es aún más considerable y parece estar en aumento.

4.3. Proyección de temporada actual

La proyección de temporada consta en la proyección de las divisiones Primera y Primera B en la temporada actual 2024 una vez finalizada la primera rueda de ambas competencias (15 partidos). Cada temporada es proyectada por 2 métodos: puntos esperados y simulación montecarlo.

4.3.1. Proyección de Primera

El evolutivo desde los puntos actuales de los equipos, con la fecha 15 ya jugada hasta los puntos esperados de los equipos de Primera división se muestran en la figura 13 a continuación.

Evolución de puntos esperados Primera

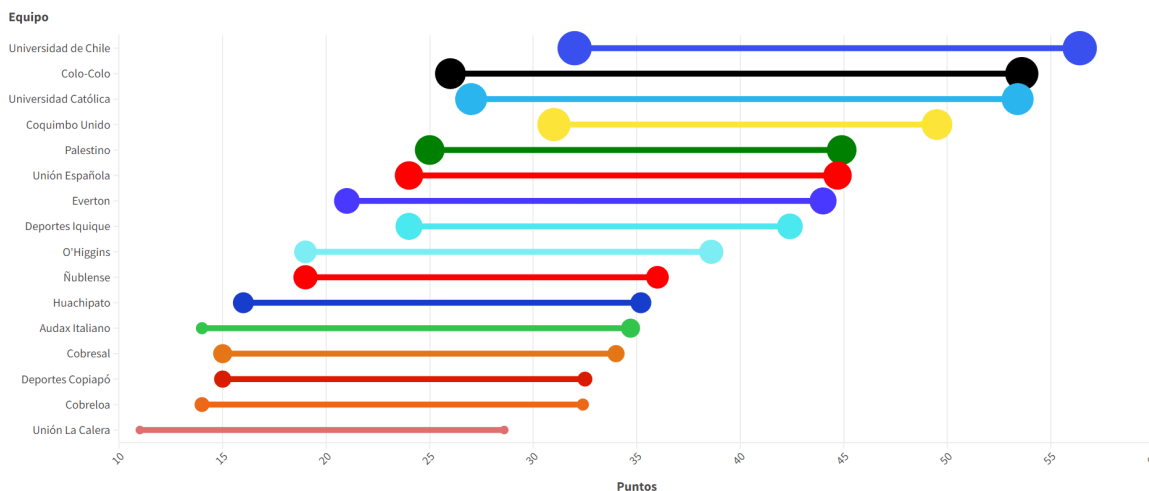


Figura 13: Evolución de puntos esperados de Primera 2024.

En cuanto a los primeros puestos se espera que Universidad de Chile mantenga el primer lugar con una puntuación esperada final de casi 3 puntos de diferencia de su perseguidor, a pesar de que se espera que de los puntos por jugar obtenga casi 3 puntos menos que el equipo que le sigue en la tabla esperada.

Se espera también que Universidad Católica y Colo-Colo terminen muy justos en puntos para definir el segundo y tercer lugar.

Coquimbo Unido, que actualmente ocupa el segundo lugar. se espera que termine cuarto, aproximadamente 4 puntos esperados por debajo del tercer lugar.

Del quinto al octavo lugar se espera que no varíen los equipos, pero las posiciones tienen posibilidades de variar, ya que los puntos esperados no tienen mucha diferencia entre estos equipos. Luego al comparar con el noveno equipo esperado, hay una brecha de aproximadamente 4 puntos esperados finales con el octavo.

Se espera que Audax Italiano salga de la zona de descenso, y que el penúltimo lugar se

dispute entre Cobresal y Deportes Copiapó, mientras que Unión La Calera se espera que se mantenga en el último lugar con una brecha de puntos esperados final de aproximadamente 4 puntos.

Al ejecutar una simulación montecarlo de 10000 simulaciones, a partir de la tabla actual de puntos y los partidos por disputar se obtiene la matriz de probabilidades de posiciones finales de la Primera división.

Pos real	Equipo	Pos final															
		1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16
1	Universidad de Chile	46%	24%	15%	8%	4%	2%										
2	Coquimbo Unido	10%	16%	20%	20%	14%	9%	6%	3%	2%							
3	Universidad Católica	19%	24%	21%	15%	9%	6%	3%	2%								
4	Colo-Colo	21%	24%	20%	15%	9%	6%	1%									
5	Palestino	1%	4%	7%	12%	16%	15%	15%	11%	8%	5%	4%	2%	1%			
6	Unión Española	1%	4%	7%	10%	15%	16%	14%	11%	8%	6%	4%	2%	1%			
7	Deportes Iquique		2%	4%	7%	11%	15%	15%	14%	11%	8%	5%	3%	2%	1%		
8	Everton		3%	5%	8%	13%	15%	15%	13%	10%	7%	4%	3%	2%	1%		
9	Ñublense					2%	3%	5%	8%	10%	13%	12%	12%	11%	8%	8%	4%
10	O'Higgins				2%	3%	6%	8%	11%	13%	13%	12%	10%	9%	7%	5%	2%
11	Huachipato					1%	2%	4%	6%	9%	11%	12%	13%	13%	11%	10%	6%
12	Cobresal							1%	3%	4%	6%	9%	11%	13%	16%	19%	17%
13	Deportes Copiapó							1%	3%	4%	6%	9%	11%	14%	16%	18%	16%
14	Cobreloa					1%	2%	4%	6%	8%	10%	12%	13%	12%	12%	11%	7%
15	Audax Italiano						2%	3%	6%	7%	10%	11%	13%	13%	13%	12%	9%
16	Unión La Calera									2%	2%	4%	6%	8%	13%	20%	43%

Tabla 14: Matriz de probabilidades de posiciones finales de Primera 2024.

Universidad de Chile se proyecta como el campeón más probable y con muchas probabilidades de asegurar quedar dentro de los 3 mejores en la tabla. El siguiente equipo con más probabilidades de ser campeón es Colo-Colo a pesar de ser actualmente cuarto en la tabla de posiciones, luego le sigue Universidad Católica y por último Coquimbo Unido, que a pesar de ser segundo actualmente, se le proyecta como posiciones más probables la de tercero y cuarto.

El siguiente bloque de equipos es el de Palestino, Unión Española, Deportes Iquique y Everton, que se disputarán los puestos desde el quinto al octavo, lo que implica asegurar un cupo en Copa Sudamericana (hasta el séptimo puesto).

Del noveno al décimo primer puesto se proyectan como mayores candidatos Ñublense, O'Higgins y Huachipato.

Cobresal y Deportes Copiapó a pesar de ser décimo segundo y décimo terceros en la tabla de clasificaciones, se les proyecta en la lucha de los últimos 4 puestos.

Cobreloa y Audax Italiano a pesar de estar antepenúltimo y penúltimo, tienen probabilidades de terminar en variedad puestos superiores a los actuales.

Por último, Unión La Calera tiene altas probabilidades de mantenerse último o penúltimo (puestos de descenso).

4.3.2. Proyección de Primera B

En cuanto al evolutivo de los puntos actuales de los equipos, con la fecha 15 ya jugada respecto a los puntos esperados de los equipos de Primera B se muestran en la figura 14 a continuación.

Evolución de puntos esperados Primera B

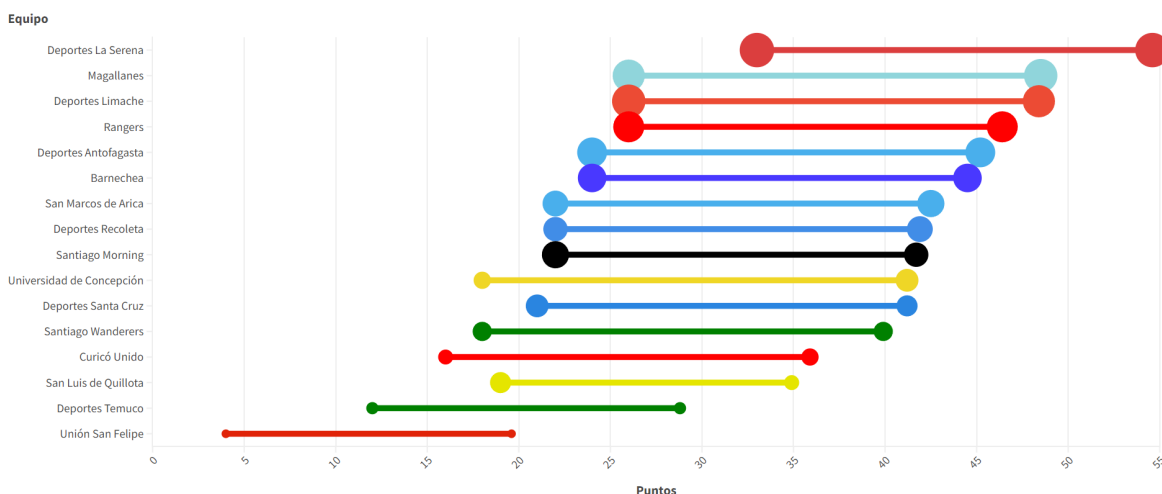


Figura 14: Evolución de puntos esperados de Primera B 2024.

Se espera que Deportes La Serena termine con aproximadamente 6 puntos de diferencia del segundo lugar, principalmente por la buena posición en la que termina la primera rueda. Deportes La Serena con 7 puntos de diferencia de los 3 perseguidores.

Del segundo al cuarto lugar se espera una competencia apretada entre Rangers, Deportes Limache y Magallanes por obtener el pase directo al partido definitorio de ascenso (Segundo lugar) o en su defecto obtener una buena posición en los *Play Off* de ascenso (del tercer al octavo lugar).

El quinto y sexto lugar se espera que se definan entre Deportes Antofagasta y Barnechea. Luego del séptimo al décimo lugar se espera que se defina entre San Marcos de Arica, Deportes Recoleta, Santiago Morning y Universidad de Concepción, que se espera que escale 3 posiciones.

Del décimo primer al décimo tercer lugar se espera que se dispute entre Deportes Santa Cruz, Santiago Wanderers y Curicó Unido.

San Luis de Quillota proyecta una bajada importante, dejándolo en una posición esperada de décimo cuarto, pero distanciado bastante en puntos de Deportes Temuco y Unión San Felipe. Este último tuvo una mala primera rueda, sumando solo 4 puntos, proyectándolo como el último lugar y único descendido.

Al ejecutar una simulación montecarlo de 10000 simulaciones, a partir de la tabla actual de puntos y los partidos por disputar se obtiene la matriz de probabilidades de posiciones finales de la Primera B.

Pos real	Equipo	Pos final															
		1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16
1	Deportes La Serena	59%	18%	10%	5%	3%	2%	1%									
2	Deportes Limache	9%	15%	14%	12%	10%	9%	8%	7%	5%	4%	3%	2%	1%			
3	Magallanes	8%	13%	13%	12%	11%	10%	8%	7%	5%	5%	4%	3%	2%			
4	Rangers	8%	14%	14%	12%	11%	9%	9%	7%	6%	5%	3%	2%	1%			
5	Deportes Antofagasta	5%	11%	12%	12%	10%	11%	8%	7%	7%	6%	5%	3%	2%	1%		
6	Barnechea	5%	9%	10%	11%	11%	10%	9%	9%	7%	6%	5%	4%	3%	2%		
7	Santiago Morning	2%	4%	6%	7%	8%	8%	9%	9%	10%	9%	9%	8%	7%	4%	2%	
8	San Marcos de Arica	2%	4%	6%	7%	8%	9%	8%	9%	10%	9%	8%	7%	6%	4%	2%	
9	Deportes Recoleta	2%	4%	6%	7%	8%	8%	9%	9%	9%	9%	9%	8%	6%	5%	2%	
10	Deportes Santa Cruz	1%	3%	5%	5%	6%	7%	8%	9%	9%	10%	10%	9%	8%	6%	3%	
11	San Luis de Quillota			2%	2%	3%	4%	5%	6%	7%	9%	11%	13%	14%	14%	8%	1%
12	Santiago Wanderers		2%	3%	4%	5%	6%	7%	8%	9%	10%	10%	11%	10%	10%	5%	
13	Universidad de Concepción		2%	2%	3%	4%	5%	7%	7%	9%	10%	10%	12%	12%	10%	6%	
14	Curicó Unido				1%	2%	2%	3%	4%	5%	8%	9%	11%	15%	20%	16%	3%
15	Deportes Temuco								1%	2%	2%	4%	6%	10%	19%	43%	11%
16	Unión San Felipe													1%	3%	12%	82%

Tabla 15: Matriz de probabilidades de posiciones finales de Primera B 2024.

Se proyecta a Deportes La Serena como el máximo candidato al ascenso directo con casi 60 % de probabilidades. La disputa por el segundo lugar está aún por definirse, ya que hay 5 equipos con probabilidades similares y relativamente altas de terminar segundo, y fuera de esos equipos, hay muchos con pocas pero con probabilidades.

Los últimos 2 puestos están bastante definidos, dejando a Unión San Felipe con más del 80 % de probabilidades de terminar último y Deportes Temuco con más del 40 % de probabilidades de terminar penúltimo. Actualmente el descenso en la Primera B solo aplica al último lugar de forma directa.

En cuanto al resto de la tabla todos tienen probabilidades repartidas en distintas posiciones, por lo que hay que esperar a que se acoten los caso posibles con el paso de las jornadas

MODELO PREDICTIVO DE FÚTBOL CHILENO, SU CONTEXTO NACIONAL Y LATINOAMERICANO BASADO EN
POWER RANKING HISTÓRICOS.

Posición	Equipo	País	Puntos Power Ranking	Clasificación
1	Palmeiras	Brasil	1024,50	100,00
2	Boca Juniors	Argentina	1024,38	99,72
3	River Plate	Argentina	1023,41	97,51
4	Flamengo	Brasil	1018,04	85,23
5	Nacional (U)	Uruguay	1013,96	75,90
6	Atlético Nacional	Colombia	1013,75	75,40
7	Olimpia	Paraguay	1013,29	74,35
8	São Paulo	Brasil	1012,81	73,26
9	Bolívar	Bolivia	1009,76	66,28
10	Independiente del Valle	Ecuador	1009,26	65,15
11	Peñarol	Uruguay	1008,56	63,53
12	Estudiantes de La Plata	Argentina	1008,24	62,80
13	Libertad	Paraguay	1008,16	62,62
14	Internacional	Brasil	1006,34	58,46
15	Fluminense	Brasil	1005,88	57,41
16	Liga de Quito	Ecuador	1005,87	57,39
17	Santos	Brasil	1005,59	56,73
18	Atlético Mineiro	Brasil	1005,37	56,24
19	Colo-Colo	Chile	1004,77	54,86
20	Corinthians	Brasil	1004,65	54,60
21	Racing Club	Argentina	1003,77	52,58
22	Cerro Porteño	Paraguay	1003,32	51,56
23	Audax Italiano	Chile	1001,57	47,54
24	Atlético Paranaense	Brasil	1001,56	47,53
25	Bragantino	Brasil	1001,54	47,47
26	Millonarios	Colombia	1001,29	46,90
27	Defensa y Justicia	Argentina	1001,21	46,72
28	Newell's Old Boys	Argentina	1001,20	46,71
29	Universidad Católica	Chile	1000,81	45,81
30	Cienciano	Perú	1000,79	45,76
31	Fortaleza	Brasil	1000,62	45,37
32	El Nacional	Ecuador	1000,59	45,30
33	Junior	Colombia	1000,54	45,20
34	Deportivo Táchira	Venezuela	1000,40	44,87
35	Boston River	Uruguay	1000,07	44,11
36	San Lorenzo	Argentina	1000,00	43,96
37	Deportivo Pereira	Colombia	999,88	43,68
38	Botafogo	Brasil	999,88	43,67
39	Sportivo Ameliano	Paraguay	999,71	43,28
40	Deportivo Maldonado	Uruguay	999,71	43,28
41	General Caballero	Paraguay	999,58	42,99
42	Deportes Tolima	Colombia	999,53	42,88
43	Magallanes	Chile	999,47	42,75
44	Zamora	Venezuela	999,45	42,70

Posición	Equipo	País	Puntos Power Ranking	Clasificación
45	Barcelona	Ecuador	999,38	42,55
46	Curicó Unido	Chile	999,36	42,48
47	Santa Fe	Colombia	999,33	42,42
48	Ñublense	Chile	999,27	42,27
49	Metropolitanos	Venezuela	999,26	42,27
50	Guaraní	Paraguay	999,04	41,75
51	Atlético Palmaflor	Bolivia	998,93	41,50
52	Delfín	Ecuador	998,91	41,47
53	Independiente Medellín	Colombia	998,83	41,28
54	Liverpool	Uruguay	998,78	41,16
55	Palestino	Chile	998,67	40,92
56	Goiás	Brasil	998,64	40,85
57	Patronato	Argentina	998,60	40,75
58	Águilas Doradas	Colombia	998,58	40,70
59	Deportivo Binacional	Perú	998,49	40,49
60	Carabobo	Venezuela	998,42	40,33
61	Emelec	Ecuador	998,19	39,82
62	Nacional Potosí	Bolivia	998,16	39,74
63	Puerto Cabello	Venezuela	998,12	39,65
64	Sport Huancayo	Perú	997,95	39,27
65	Universidad César Vallejo	Perú	997,94	39,23
66	Cobresal	Chile	997,88	39,09
67	Monagas	Venezuela	997,74	38,79
68	The Strongest	Bolivia	997,67	38,62
69	Gimnasia y Esgrima La Plata	Argentina	997,56	38,38
70	Always Ready	Bolivia	997,48	38,19
71	Tacuare	Paraguay	997,45	38,11
72	Aucas	Ecuador	997,18	37,50
73	América Mineiro	Brasil	996,94	36,96
74	Huracán	Argentina	996,72	36,45
75	Caracas	Venezuela	996,61	36,20
76	Tigre	Argentina	996,44	35,80
77	Melgar	Perú	996,39	35,70
78	Deportivo Cuenca	Ecuador	996,37	35,65
79	Argentinos Juniors	Argentina	995,72	34,17
80	Guabirá	Bolivia	995,47	33,59
81	Universitario	Perú	995,29	33,18
82	Defensor Sporting	Uruguay	994,94	32,37
83	Sporting Cristal	Perú	994,39	31,12
84	Danubio	Uruguay	993,29	28,60
85	Blooming	Bolivia	993,22	28,44
86	Estudiantes de Mérida	Venezuela	991,78	25,15
87	Oriente Petrolero	Bolivia	984,02	7,39
88	Alianza Lima	Perú	980,79	0,00

Tabla 16: Tabla de *power ranking* internacional de equipos participantes en competencias CONMEBOL en 2023.

Fuente: Elaboración propia.

Para facilitar la lectura de datos y minimizar el solapamiento de nombres de los equipos, se grafica por separado los equipos que en 2023 participaron en Copa Libertadores, Copa Sudamericana y Recopa Sudamericana en la figura 16.

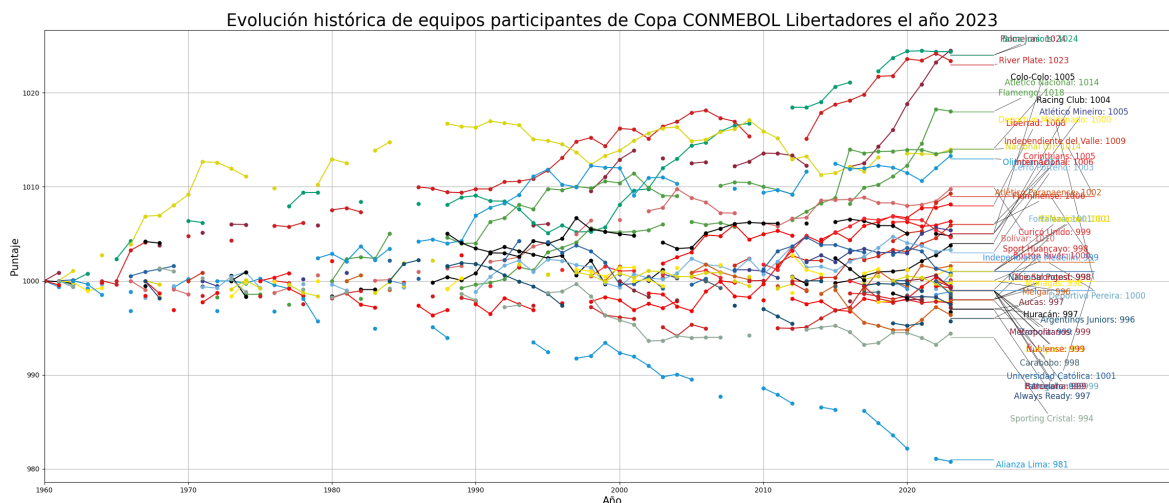


Figura 16: *Power ranking* histórico internacional de equipos participantes en copa libertadores en 2023.

Fuente: Elaboración propia.

Una diferencia fundamental con el *power ranking* nacional es la interminencia de los equipos, ya que no todos los equipos clasifican todos los años a competencias CONMEBOL, aún así hay equipos que tienen participaciones consecutivas que permiten un análisis de su rendimiento internacional.

Los equipos participantes de la Copa CONMEBOL Libertadores el 2023 con mejores puntuaciones son Boca Juniors, el segundo equipo con más Copas Libertadores (6) y el equipo con más apariciones en finales. Palmeiras, 3 veces campeón de la competencia. River Plate, el cuarto equipo con más Copas Libertadores (4). Flamengo, Nacional de Montevideo y Olimpia, cada uno campeón 3 veces de la competencia y Atlético nacional, dos veces campeón. De estos equipos, en su mayoría siguen en ascenso de rendimiento a pesar de su crecimiento histórico.

La mayoría de equipos se acumulan en puntos similares y al final aparece Alianza Lima, equipo en baja de rendimiento internacional de forma histórica, pero manteniendo su participación de forma consecutiva.

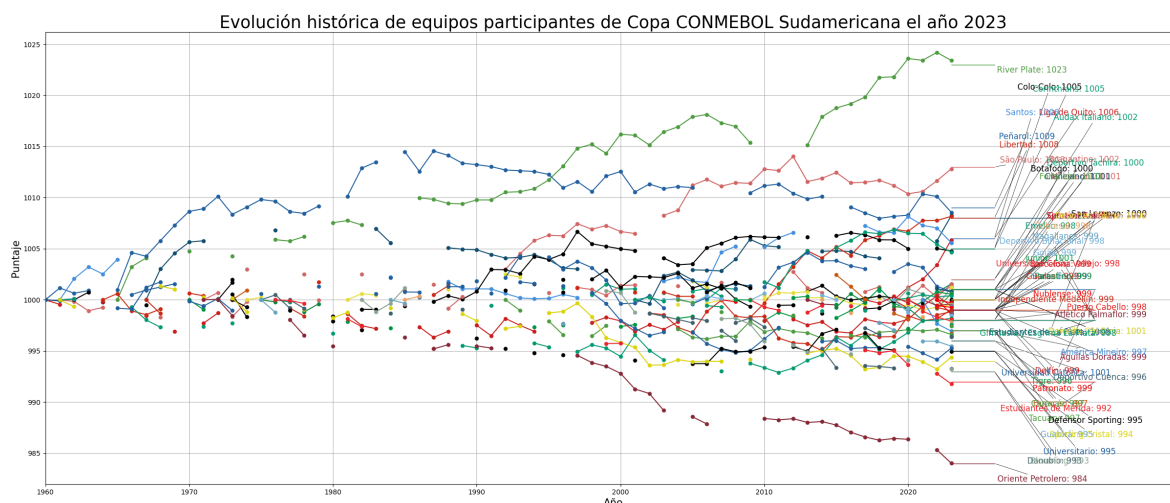


Figura 17: *Power ranking* histórico internacional de equipos participantes en copa sudamericana en 2023.

Fuente: Elaboración propia.

En el evolutivo de puntos de equipos participantes de Copa Sudamericana en 2023 (figura 17) se observan equipos ya presentes en el evolutivo de la Copa Libertadores, esto se da por el formato de las competencias donde, los terceros lugares de la fase de grupos de la Copa Libertadores pasan a participar en la Copa Sudamericana. El equipo con mejor clasificación en el *power ranking* es River Plate que participó en la Copa Sudamericana. Luego se encuentra Sao Paulo, que desde el 1992 solo ha estado ausente un año de competencias CONMEBOL. Y luego está Peñarol, equipo que ha ido a la baja, pero en los años 60' tuvo un crecimiento explosivo y en los 80' tuvo su máximo rendimiento de puntos del *power ranking*, en estas décadas Peñarol ganó sus 5 Libertadores.

Como es tendencia en los gráficos evolutivos de *power ranking*, se agrupan la mayoría de equipos en puntuaciones promedio, y de forma similar al evolutivo de la Copa Libertadores, el equipo con peor rendimiento, Oriente Petrolero, viene a la baja de forma histórica a pesar de apariciones en años consecutivos.



Figura 18: *Power ranking* histórico internacional de equipos participantes en recopa sudamericana en 2023.

Fuente: Elaboración propia.

Los equipos de la Recopa de 2023 (figura 17) son el campeón de la Libertadores (Flamengo) y el campeón de la Sudamericana (Independiente el Valle) el 2022. Ambos equipos destacan por tener un gran aumento de puntos en el *power ranking* durante los últimos años, indicativo de procesos exitosos que dieron resultados, especialmente el caso de Independiente del Valle, que tuvo su debut internacional el 2012 y 10 años después es campeón de la Copa Sudamericana.

Fuera de eso, es clara la diferencia de puntos entre el campeón de la Libertadores y de la Sudamericana.

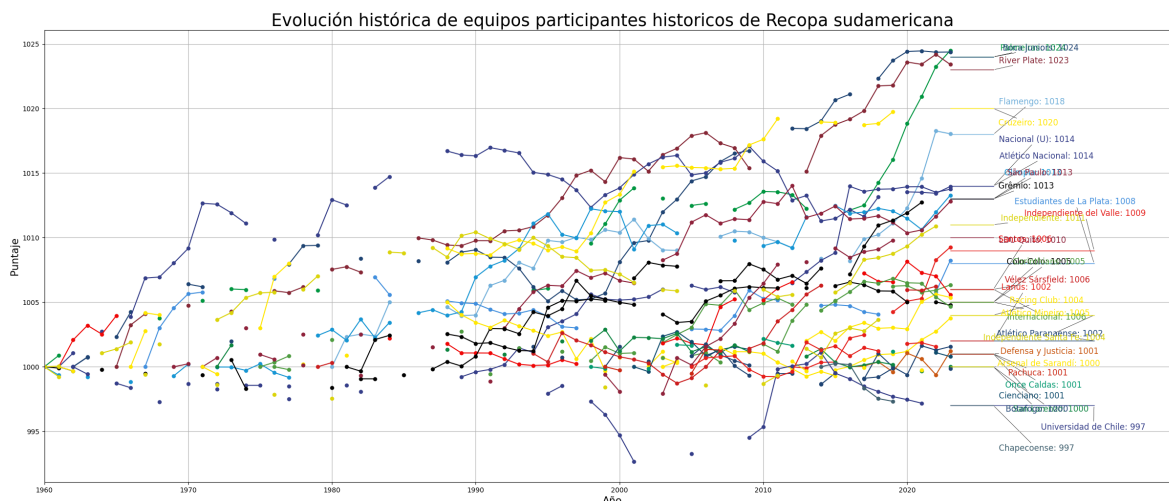


Figura 19: *Power ranking* histórico equipos participantes de recopa sudamericana.
Fuente: Elaboración propia.

Al observar solo equipos participantes en algún momento de Recopa en la figura 19 se observan a todos los equipos que alguna vez fueron campeones de Libertadores y/o Sudamericana. Estos equipos muestran una tendencia positiva o de mantención.

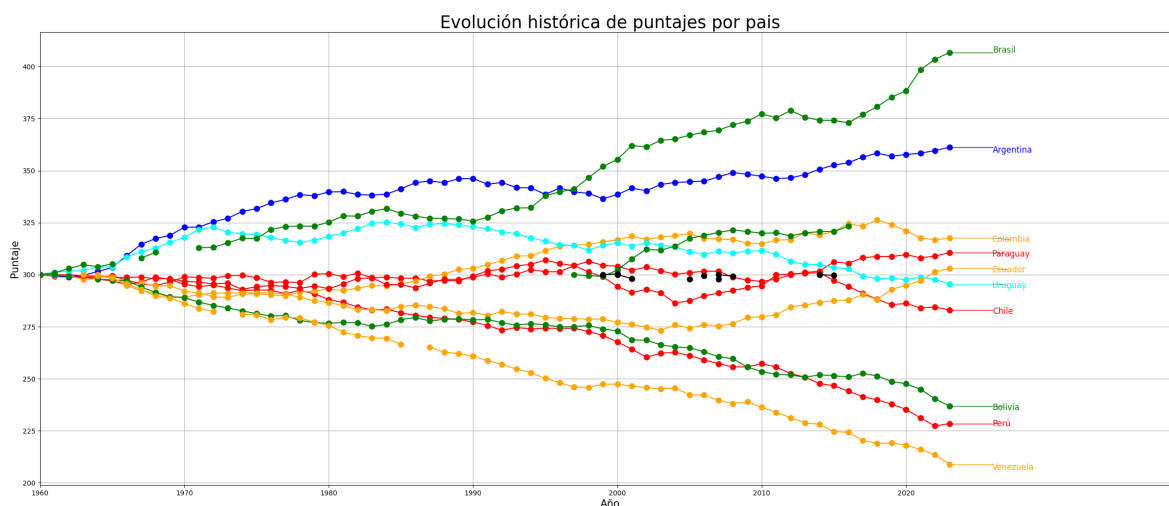


Figura 20: *Power ranking* histórico nacional de países.
Fuente: Elaboración propia.

El rendimiento de países en el ámbito internacional, de la figura 20, tiene tendencias similares a los evolutivos de los equipos, destacan en este caso 2 países, Brasil y Argentina, se acumula un grupo de rendimiento promedio y otro que va a la baja. Entre fines de 1990 y mediados de la década del 2010 por cambios de formato de las competiciones CONMEBOL se tiene registro de países CONCACAF que fueron invitados a la competición en

su momento, pero no son considerados en el análisis.

Brasil, si bien destaca por sobre el resto de equipos no es el país con más Copas Libertadores ni Copas Sudamericanas, son el segundo país en cantidad de títulos en ambas competencias, solo superados por Argentina.

Actualmente Colombia se posiciona tercero en el *power ranking*, y es el cuarto equipo con más Libertadores y Sudamericanas.

El cuarto equipo en el *power ranking* es Paraguay, pero este se ganó el cuarto lugar en los últimos años donde mostró una mejora de rendimiento, hasta entonces mostraba una tendencia neutra, su última Libertadores de las 3 que tiene fue en 2002.

Quinto se ubica actualmente Ecuador, que viene en una clara alza desde mediados de la década del 2000, donde ganó su única Libertadores el 2008 y desde fines de la década del 2010 hasta la actualidad ha ganado una Libertadores y dos Sudamericanas. Actualmente Ecuador está en un profundo proceso de mejora del fútbol a nivel país que ha demostrado resultados y es valioso estudiar.

Uruguay se ubica en el sexto lugar a pesar de ser el tercer equipo con más Libertadores. Sin embargo, la última copa Internacional de clubes levantada por Uruguay fue en 1988, época en la que se codeaba con Argentina y Brasil en los primeros puestos del *power ranking*, tendiendo luego a la baja a nivel de clubes de la cual no se ha podido recuperar. De la misma forma que Ecuador es un buen ejemplo de un proceso, Uruguay debe ser objeto de estudio para entender esta caída y como evitar esta baja de rendimiento.

Séptimo se encuentra Chile que mostraba rendimientos promedio a nivel internacional hasta mediados de la década del 2010, donde empezó una baja de rendimiento preocupante para el fútbol chileno con la cual aún batalla.

Finalmente, Bolivia, Perú y Venezuela muestran tendencias a la baja de forma histórica, de forma similar a Alianza Lima (Perú) y Oriente petrolero (Bolivia).

CAPÍTULO 5

CONCLUSIONES

El presente trabajo presentó y aplicó un método que construye *power rankings* a nivel nacional e internacional que cuantifica el rendimiento de los equipos, considerando el estado actual e histórico de los equipos, siendo identificables mejoras o deficiencias de estos a lo largo de la historia tanto chilena como latinoamericana a nivel clubes. Mediante estos *power ranking* se realizan predicciones a futuro, proyectando los escenarios posibles de cara al fin de una temporada en desarrollo, ofreciendo así una herramienta que complementa la toma de decisiones y seguimiento de objetivos de los equipos del fútbol chileno profesional.

Los objetivos planteados en la sección 1.2.2 son cumplidos, aportando al análisis tanto a nivel ejecutivo de equipos de fútbol como a nivel aficionado.

- **Diseño y creación de un modelo predictivo dinámico de temporada en el fútbol chileno y aplicación de este en la temporada 2024**

El modelo planteado y creado predice la temporada actual 2024 desde dos perspectivas: puntos esperados en base al rendimiento del equipos y los partidos por jugar del mismo y simulación montecarlo, que simulando un gran número de veces los encuentros que quedan por jugar, con las probabilidades calculadas de cada uno, se obtienen las probabilidades que tiene cada equipo de terminar en cada una de las posiciones de la tabla. La predicción es dinámica ya que con el paso de jornadas, los puntos esperados y probabilidades cambian.

Desde el punto de vista del negocio, permite evaluar las opciones que tiene el equipo de cara a fin de temporada adaptado al calendario por jugar de todos los equipos y distintas combinaciones y posibilidades, analizando si los objetivos planteados se espera que sean cumplidos o hay que mejorar, e identificar a los equipos que realmente compiten por la misma posición final para preparar con más énfasis encuentros directos con estos.

Para el aficionado, esto representa una forma de medir las posibilidades que tiene su equipo a fin de temporada fecha a fecha, como también identificar qué equipos son rivales directos y poner especial atención a esos encuentros.

- **Visualización de la evolución histórica del Fútbol Chileno a través del Tiempo**

El modelo permite visualizar un histograma del *power ranking* calculado desde los comienzos del profesionalismo hasta la actualidad, exhibiendo etapas de distintos

equipos, ya sean exitosas, promedio o fracasos.

Lo que permite identificar los procesos por los que ha pasado el equipo a lo largo de la historia, e identificar si actualmente la situación es similar, o compararse con otros equipos, que dado un contexto similar logran mejorar el rendimiento, para estudiar a fondo la situación y evaluar si se pueden replicar buenas prácticas o evitar las malas.

A ojos del espectador, el *power ranking* histórico nacional permite identificar la jerarquía de los equipos actualmente y en qué punto de la historia se establecieron, como también visualizar el estado de la liga chilena en tiempos en los que no pudo presentarse, incentivando así la curiosidad por la historia de los clubes y respectivas glorias.

■ **Unificación de *Power Rankings* en el Fútbol Latinoamericano**

El *power ranking* del modelo basado solo en encuentros internacionales logra representar el nivel de equipos que son recurrentes en estas competiciones, pero si un equipo tiene apariciones esporádicas o puntuales a nivel internacional su puntuación del *power ranking* no será representativo de su rendimiento. Sin embargo, al abstraer el *ranking* a nivel países si se puede extraer información continua en el tiempo, pudiendo ver la realidad chilena a nivel internacional de clubes tanto ahora como a lo largo de la historia, y comparar la gran brecha que existe entre países como Brasil y Argentina con el resto de países.

A la hora de preparar un club, este *power ranking* permite dimensionar la calidad del rival en caso de que tenga presencia continua en competiciones internacionales, y en caso de no ser así, considerar el nivel de su país como un representante, identificando la complejidad de grupos de las competiciones internacionales y planificar mejor los partidos de estas.

Para el fanático, este *power ranking* facilita el análisis de equipos a los que no hace seguimiento normalmente, pero ahora se enfrentan a equipos chilenos a nivel internacional, cuales vienen en procesos exitosos y cuales van en caída, como también identificar jerarquías entre países a nivel clubes de forma más clara.

La metodología diseñada se ha expuesto de forma clara y explícita, tanto en lo metodológico como en los cálculos realizados, exponiendo ejemplos y resultados y facilitando la replicación de la misma en distintos contextos.

Se aporta la definición de la matriz de probabilidades de resultados por rango en la sección 3.5.2, una métrica orientada a predicción de eventos con distintos escenarios posibles, abordada específicamente al contexto futbolístico con tres escenarios posibles, pero aplicable con su respectiva adaptación a otro tipo de predicciones.

El trabajo expuesto deja abiertas las puertas de futuros trabajos, tales como la aplicación de predicciones en competencias con formato de copas, para el fútbol chileno significaría la proyección de la Copa Chile o liguillas de definición, a nivel latinoamericano permite predecir los distintos escenarios de la Copa CONMEBOL Libertadores y Copa CONMEBOL Sudamericana.

De cara a los *power ranking* y predicciones a nivel latinoamericano, se puede trabajar en el registro histórico de encuentros de las distintas ligas latinoamericanas que participan en competencias CONMEBOL como se hizo para Chile en este trabajo, para así cubrir la problemática de apariciones esporádicas o escasas de equipos, para tener una mejor representación del nivel de estos.

Un tercer trabajo futuro es la aplicación y evaluación dedicada de la métrica propuesta de 'matriz de probabilidades de resultados por rango', tanto para el fútbol como en otros contextos, comparándola con modelos predictivos ya establecidos y matizando con otro tipo de métricas.

Haciendo retrospectiva ahora de cómo la carrera de ingeniería civil informática en la universidad técnica Federico Santa María ha aportado al desarrollo de la presente memoria, a nivel general ha sido fundamental el desarrollo del pensamiento computacional y sus cuatro pilares:

- Descomposición, que permite descomponer problemas, y abordarlos como una colección de problemas pequeños y manejables.
- Reconocimiento de patrones, identificando similitudes y diferencias en datos y procesos, para aplicar soluciones a problemas parecidos y aplicar patrones manejables por un programa computacional.
- Abstracción, enfocándose en los detalles realmente influyentes, para llegar a la raíz del problema y pensar en la solución de forma que sea aplicable en código, dejando de lado razonamientos subjetivos que un *software* no puede manejar.
- Diseño de algoritmos, para extraer el problema a un algoritmo, identificando correctamente los pasos, su orden y variables de forma objetiva y cuantitativa.

Este pensamiento es desarrollado desde distintos enfoques en la mayoría de las asignaturas, de estas, las que han influenciado más directamente en aspectos técnicos para el desarrollo de esta memoria son: Inteligencia artificial, algoritmos y complejidad, estadística computacional y visualización de datos.

Sin embargo, el enfoque de la carrera en ciencia de datos es bastante secundario, tanto en asignaturas de la malla de carrera como en variedad de electivos, principalmente por la variedad de ramos siendo que se dispone de profesores muy buenos en estas áreas, y en el proyecto más grande previo al desarrollo de memoria que hace alusión al 'desarrollo de proyecto de informática' está sesgado a solo el desarrollo de software y videojuegos, dejando sin espacio de desarrollo y exposición de trabajos de ciencia de datos.

Por último destacar la gran disposición y entrega del departamento de informática hacia los estudiantes, tanto en el apoyo de procesos académicos como en la cercanía hacia los estudiantes, cualidad de la que no todos departamentos de carrera destacan.

Figura 21: *Power ranking* histórico nacional de equipos en las 3 primeras divisiones en 2023 con $\sigma = 100$.

Fuente: Elaboración propia.

Figura 22: *Power ranking* histórico nacional de equipos en las 3 primeras divisiones en 2023 con $\sigma = 800$.

Fuente: Elaboración propia.

REFERENCIAS BIBLIOGRÁFICAS

- [analyticsvidhya.com, 2023a] analyticsvidhya.com (2023a). Binary cross entropy/log loss for binary classification. <https://www.analyticsvidhya.com/blog/2021/03/binary-cross-entropy-log-loss-for-binary-classification/>.
- [analyticsvidhya.com, 2023b] analyticsvidhya.com (2023b). Metrics to evaluate your classification model to take the right decisions. <https://www.analyticsvidhya.com/blog/2021/07/metrics-to-evaluate-your-classification-model-to-take-the-right-decisions/>.
- [Burer, 2012] Burer, S. (2012). Robust rankings for college football. *Journal of Quantitative Analysis in Sports*, 8(2).
- [Davidson, 1970] Davidson, R. R. (1970). On extending the bradley-terry model to accommodate ties in paired comparison experiments. *Journal of the American Statistical Association*, 65(329):317–328.
- [Ebtekar y Liu, 2021] Ebtekar, A. y Liu, P. (2021). Elo-mmr: A rating system for massive multiplayer competitions. En *Proceedings of the Web Conference 2021*, pp. 1772–1784.
- [elorating.net, 2023] elorating.net (2023). Elo ratings. <https://elorating.net>.
- [FIFA, 2003] FIFA (2003). Women's ranking procedures. <https://www.fifa.com/fifa-world-ranking/procedure-women>. Accessed: 3rd December, 2023.
- [FIFA, 2018] FIFA (2018). Fifa men's world ranking procedur. <https://www.fifa.com/fifa-world-ranking/procedure-men>. Accedido el 29 de noviembre de 2023.
- [Katz, 1953] Katz, L. (1953). A new status index derived from sociometric analysis. *Psychometrika*, 18(1):39–43.
- [Lasek et al., 2013] Lasek, J., Szlávik, Z., y Bhulai, S. (2013). The predictive power of ranking systems in association football. *International Journal of Applied Pattern Recognition*, 1(1):27–46.
- [McHale et al., 2007] McHale, Ian and Davies, Stephen and others (2007). Statistical analysis of the effectiveness of the fifa world rankings. *Statistical thinking in sports*, 15:77–90.
- [Monje, 2011] Monje, C. (2011). Guía didáctica metodología de la investigación. *Universidad Surcolombiana*.
- [Murtha, 1997] Murtha, J. A. (1997). Monte carlo simulation: its status and future. *Journal of Petroleum Technology*, 49(04):361–373.

- [ncl.ac.uk (Newcastle University), 2023] ncl.ac.uk (Newcastle University) (2023). Coefficient of determination, r-squared. <https://www.ncl.ac.uk/webtemplate/ask-assets/external/maths-resources/statistics/regression-and-correlation/coefficient-of-determination-r-squared.html>.
- [Opta.com, 2023] Opta.com (2023). Opta. <https://theanalyst.com/eu/2023/01/power-rankings-your-club-ranked/>.
- [Park y Newman, 2005] Park, J. y Newman, M. E. (2005). A network-based ranking system for us college football. *Journal of Statistical Mechanics: Theory and Experiment*, 2005(10):P10014.
- [Pradhan y Abdourazakou, 2020] Pradhan, S. y Abdourazakou, Y. (2020). “power ranking” professional circuit esports teams using multi-criteria decision-making (mcdm). *Journal of Sports Analytics*, 6(1):61–73.
- [Rao y Kupper, 1967] Rao, P. y Kupper, L. L. (1967). Ties in paired-comparison experiments: A generalization of the bradley-terry model. *Journal of the American Statistical Association*, pp. 194–204.
- [Szczecinski y Djebbi, 2020] Szczecinski, L. y Djebbi, A. (2020). Understanding draws in elo rating algorithm. *Journal of Quantitative Analysis in Sports*, 16(3):211–220.
- [Szczecinski y Roatis, 2022] Szczecinski, L. y Roatis, I.-I. (2022). Fifa ranking: Evaluation and path forward. *Journal of Sports Analytics*, 8:1–20.
- [Tippett, 2019] Tippett, J. (2019). *The expected goals philosophy: A game-changing way of analysing football*. wydawca nieznany.
- [Wheatcroft, 2021] Wheatcroft, E. (2021). Evaluating probabilistic forecasts of football matches: the case against the ranked probability score. *Journal of Quantitative Analysis in Sports*, 17(4):273–287.