Universidad del Desarrollo
Facultad de Ingeniería

¿El 11 ideal?

Un Enfoque Data - Driven para la conformación de un Plantel.

POR: J. SEBASTIAN BECERRA

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

PROFESOR GUÍA:

Dra. Loreto Bravo

Dr(c). Hugo Contreras

12-2022

SANTIAGO

AGRADECIMIENTO

Agradezco profundamente a mi profesora guía Loreto Bravo, a mi familia, a Tenny y a todos los que hicieron posible este proyecto con sus valiosos comentarios.

TABLA DE CONTENIDO

Resumen

El problema de evaluar el rendimiento de los futbolistas está despertando el interés de muchas empresas y de la comunidad científica, gracias a la disponibilidad de datos masivos que capturan todos los eventos generados durante un partido (p. ej., entradas, pases, tiros, etc.). Desafortunadamente, no existe una métrica consolidada y ampliamente aceptada para medir la calidad del desempeño en todas sus facetas. En este artículo, diseñamos e implementamos un ranking de jugadores basado en datos que ofrece una evaluación justificada de como distintos atributos contribuyen a aumentar la probabilidad de victoria de un equipo.

Usando dos conjuntos de datos, el primero con información a nivel de equipo desde el año 2017 al 2022 para la primera división del futbol chileno determinamos el desempeño de los equipos clasificándolos en victoria y no victoria. El segundo set de datos contiene información de todos los jugadores chilenos en competencia. Basándonos en el criterio experto, desarrollados un procedimiento de co-ocurrencia donde detectamos los atributos comunes existentes en ambos sets de datos. Con estos atributos comunes entrenamos un modelo que nos permite obtener el ponderador para cada atributo y con ellos construir el ranking por jugador.

Finalmente clasificamos a cada jugador de acuerdo con su rol en el equipo: Arquero, Defensa, Lateral Derecho, Lateral Izquierdo, Mediocampista Defensivo, Mediocampista, Mediocampista Ofensivo y Delantero. Con esto extendemos el ranking por posición y podemos conformar un plantel que maximizaría la probabilidad de victoria de un equipo.

1. Introducción

La revolución de los datos y los enfoques data-driven han revolucionado las principales industrias y el deporte no es la excepción. Cada día es más relevante poder contar con herramientas basadas en datos que permitan apoyar el criterio experto y la toma de decisiones. En este sentido, poder cuantificar el aporte de cada integrante al desempeño final de un equipo, evaluar contrataciones y poder estimar el potencial impacto que un nuevo jugador tendría en el funcionamiento del equipo es uno de los desafíos centrales de este mercado (Bornn et al. 2018, Gerrard 2017 y Pappalardo 2019). Junto con esto, el problema de la evaluación basada en datos del rendimiento de distintos equipos y jugadores está ganando interés en la comunidad científica, gracias a la disponibilidad de grandes volúmenes de datos generados por tecnologías de detección como los llamados soccer-logs que detallan todos los eventos espacio-temporales relacionados con los jugadores durante un partido (pases, faltas, tiros, fintas, etc.).

Por otro lado, clasificar a los jugadores significa definir una relación de orden entre ellos con respecto a alguna medida de su desempeño en una secuencia de partidos. Esta es una tarea compleja, ya que no existe una definición objetiva y compartida de la calidad del desempeño, que es un concepto inherentemente multidimensional. En la literatura se han propuestos distintos enfoques y algoritmos de clasificación que en general adolecen de varias limitaciones como las que describiremos a continuación.

La más conocida es enfocarse en pocas dimensiones del juego, en el sentido de que proponen métricas que evalúan el rendimiento del jugador centrándose en un solo aspecto (pases o tiros) olvidándose de otros atributos tanto o más importantes o del aporte colectivo al equipo. Es muy probable que los veedores enfoquen su búsqueda de jugadores en un conjunto mucho más amplio de atributos, hecho que en la actualidad es totalmente abordable dada la gran disponibilidad de información. Por ejemplo, nos interesa tanto las capacidades ofensivas, como defensivas, efectividad de los pases, precesión e incluso el manejo del temperamento que puede ser aproximado como el número de tarjetas rojas y/o amarillas recibidas. Con esto, resulta imperativo un enfoque que permita explorar y realizar una evaluación integral del rendimiento basada en un amplio set de atributos recolectados.

Otro problema a considerar es que los entrenadores necesitan jugadores en distintas posiciones o roles, que tienen capacidades y habilidades específicas y no iguales dentro del equipo. Por ejemplo, no son iguales los atributos de un delantero que de un defensa o mediocampista por lo que comparar jugadores que cumplan la misma posición es un desafío. Esto es especialmente notorio cuando evaluamos la posición de delantero. Es probable que estos jugadores tengan una mayor probabilidad de anotar un gol y eso estaría altamente correlacionado con el desempeño del equipo y podría sesgar la decisión de contratación de un entrenador.

En tercer lugar, sin un gran número de datos sistematizados y estandarizados, los enfoques existentes en la literatura tienden a guiar su decisión en juicios personales basados principalmente en la interpretación de alguna métrica simplista como popularidad, valor de mercado, goles marcados, etc (Brooks et. al 2016, Stein et. al 2016 y Torgler et. al 2007). En este sentido, sería importante estimar la bondad de los algoritmos de clasificación y evaluación del desempeño de manera cuantitativa y exhaustiva, a través de grandes conjuntos de datos construidos y evaluados con la ayuda de expertos humanos.

En este trabajo, proponemos una medida en dos pasos. La primera consiste en medir el rendimiento de un equipo en base al resultado final del partido, es decir, clasificamos si el equipo gana con un 1 o con un 0 si no gana (pierde o empata). Una vez clasificado el rendimiento del equipo buscamos los principales atributos que explicarían este desempeño entrenando distintos algoritmos. Esto procedimiento nos permitirá encontrar cuanto aporta (ponderador) cada atributo al desempeño del equipo. Una vez encontrado estos ponderadores explotamos otro set de datos relativo al desempeño de los jugadores. A través de un proceso de co-ocurrencia buscamos los atributos en ambas bases de datos y aplicamos los mismos ponderadores a los atributos de los jugadores para encontrar un ranking. Finalmente, este ranking resumirá el desempeño de cada jugador. Este enfoque innovador basado en datos va más allá de los goles marcados, la popularidad o el resultado de un partido. Será capaz de entregar una clasificación de jugadores que aportará a los veedores y analistas una visión más amplia, ahorrará recursos y democratizará las oportunidades de carrera a los jugadores talentosos.

Este trabajo se configura de la siguiente manera. En la sección dos se presenta una breve descripción de los principales trabajados relacionados. En la sección tres se plantea el objetivo e hipótesis, En la sección cuatro describimos los datos y metodología utilizadas. En la sección cinco presentamos los principales resultados y en la seis concluimos.

2. Trabajos Relacionados

Como imaginamos el problema de evaluar el desempeño de los jugadores no sólo es relevante para el futbol, otros deportes se encuentran más avanzados y han formalizado sus métricas de desempeño. Por ejemplo, para el hockey, Schulte y Zhao 2017 propusieron la métrica *Scoring Impact* para clasificar a los jugadores según las probabilidades que un equipo tenia de marcar el próximo gol. En el baloncesto, la calificación de eficiencia en el rendimiento[1] es una métrica ampliamente utilizada para evaluar a los jugadores mediante la implementación de registros estadísticos (por ejemplo, pases completados, tiros logrados). En el béisbol, Baumer y Zimbalist 2014 han propuesto distintas métricas estadísticas para evaluar el desempeño de jugadores y equipos.

[1] https://www.basketball-reference.com/about/per.html

El principal impulso a este tipo de trabajos ha sido la gran recolección de datos obtenidos. Esta habilidad ha cambiado la forma de decidir en muchas industrias dentro de ellas, el futbol. Sin embargo, esto ha debido seguir un proceso ordenado y sistematizado para entender cuáles son las estadísticas realmente importantes de recolectar. En este sentido, varios trabajos han sido referentes tanto a nivel de equipo (Cintia et. al 2015, Lucey 2013 y Pappalardo 2017) como a nivel de individual (Brooks et. al 2016, Duch et. al 2010 y Nsolo et. al 2018).

Por otro lado, la conducción de estrategias y decisiones basados en datos (cultura data-driven) ha sido el otro motor de estas investigaciones. Es así cómo se han propuesto muchas métricas para capturar aspectos específicos del rendimiento del fútbol (por ejemplo, goles esperados, precisión del pase, etc.), pero sólo unos pocos enfoques evalúan la calidad del rendimiento de un jugador de manera sistémica. La métrica de centralidad de flujo (*FC*) propuesta por Duch et al 2010, es uno de los primeros intentos. Esta métrica se define como la fracción de veces que un jugador interviene en una secuencia de pases que terminan con un tiro a portería. Al estar basada meramente en la centralidad del pase, como los propios autores destacan en el artículo, la métrica *FC* tiene sentido principalmente para mediocampistas y delanteros. Brooks et al 2016 desarrollan el *Pass Shot Value (PSV)*, una métrica para estimar la importancia de un pase para generar un tiro. Representan un pase como un vector de 360 atributos que describen la vecindad de una zona de campo al origen y destino del pase. Luego, utilizan un modelo de aprendizaje automático supervisado para predecir si un pase dado resulta en un tiro o no.

Finalmente, este trabajo se basa en el marco conceptual propuesto por Pappalardo et. al 2019, llamado PlayeRank. En este trabajo los autores, proponen una evaluación multidimensional y consciente de la posición de cada jugador en equipo y de la calidad del desempeño en el partido. PlayeRank es un marco que implementa los eventos descritos por soccer-logs para evaluar la calidad del desempeño de un jugador y el rol de un jugador en un partido. A diferencia de los otros trabajos como *FC* y *PSV*, estos autores que carecen de una validación adecuada con expertos en el campo, probaron el marco con un conjunto de datos etiquetados humanamente creado específicamente para evaluar el rendimiento de los jugadores entregando luces sobre los patrones estadísticos que caracterizan el su desempeño.

3. Hipótesis y Objetivos

El objetivo de este trabajo es desarrollar un esquema basado en datos para la conformación de un equipo de futbol. Para esto seguimos tres pasos.

El primero es explotar un extenso set de datos que resume el desempeño de un equipo y con el construir una variable que permita clasificar sí el equipo obtiene un buen rendimiento, es decir, si gana el partido o si obtiene un mal rendimiento, es decir, pierde o empata.

Una vez establecido esto, el segundo paso es determinar qué atributos aportan de manera significativa a que el equipo aumente su probabilidad de ganar y cuáles atributos aportan de manera negativa. Esta relación se puede estimar a través de distintos modelos de machine learning, clasificación, aprendizaje no supervisado y aprendizaje supervisado. Específicamente en nuestro caso estimaremos un clasificador de vectores de soporte lineal (LSVC).

Tercero, una vez estimado el modelo y obtenido los ponderadores podemos crear un ranking de jugadores de futbol donde su puntaje sea directamente proporcional a la probabilidad de que el equipo obtenga un mejor rendimiento.

Finalmente, el cuarto paso es clasificar a los jugadores por posición para poder construir un equipo basado solo en criterios estadísticos.

4. Datos y Metodología

4.1. Datos

Utilizamos una base de datos masiva obtenida desde wysocut (clic aquí). Wyscout es una plataforma de futbol especialmente desarrollada para profesionales. En ella podemos encontrar datos, estadísticas, videos y distintas herramientas que permitirán desarrollar un enfoque basado en datos para entregar distintas soluciones. Específicamente, esta plataforma internacional contiene análisis de partidos, recogiendo estadísticas de rendimiento para equipos y jugadores con gran nivel de detalle.

Los datos utilizados para este trabajo corresponden a campeonato nacional chileno de primera división comprendido entre el 2017 al 2022. Con esto se construye un panel que contiene información con atributos por equipo. Si un equipo desciende, los años que está en segunda división no es incluido en la base de datos. Por otro lado, un equipo asciende a primera división este es incluido en el panel. Es importante destacar, que también se generan nuevos atributos para cada equipo en base a la información disponible. Estos atributos son el desempeño del equipo, si juega de local o visita, la formación del equipo

y el tiempo que duró la formación con que el equipo empezó el partido. Específicamente, contamos con una base de datos con 116 atributos para cada uno de los equipos, los cuales son presentados en la tabla 1. De estos atributos removemos el número de goles, ***'Goles'*** marcado por cada jugador y la probabilidad del marcar un gol '***xG***' dado que, son variables estrechamente relacionadas entre sí y producirían un sesgo y sobre ajuste en el modelo. Además, en base al criterio experto consultado, ambos atributos explicarían directamente el desempeño de cualquier equipo obviando otras dimensiones relevantes a la hora de conformar un equipo. Por otro lado, estas variables sesgarían los resultados hacia jugadores en posición de delanteros, donde su objetivo principal es anotar un gol. Es importante mencionar, que el desempeño de un equipo será cuantificado a través del resultado del partido. Construiremos una variable con un 1 cuando el equipo gana y con un 0 cuando el equipo empata o pierde.

Tabla 1: Atributos a Nivel de Equipo

```
Datos_Partidos =['Fecha', 'Competición', 'Duración', 'Seleccionar esquema', 'Goles', 'xG', 'Tiros', 'Tiros a Porteria', '%Tiros a portería', 'Pases',
 'Pases Logrados', '% de pases logrados', 'Posesión del balón, %', 'Balones Perdidos', 'Balones Perdidos Bajos', 'Balones Perdidos Medios',
 'Balones Perdidos Altos', 'Balones Recuperados', 'Balones Recuperados Bajos', 'Balones Recuperados Medios', 'Balones Recuperados Altos',
 'Duelos', 'Duelos Ganados', '% de duelos ganados', 'Tiros Fuera del Area', 'Tiros Fuera del Area a la Porteria',
 'Porcentaje Tiros Fuera del Area a la Porteria', 'Ataques posicionales', 'Ataques posicionales con remate', 'Ataques posicionales con remate %',
 'Contraataques', 'Contraataques con Remate', 'Contraataques con Remate %', 'Jugadas a balón parado', 'Jugadas a balón parado con Remate',
 'Jugadas a balón parado con Remate %', 'Córneres', 'Córneres con remate con Remate', 'Córneres con remate con Remate %', 'Tiros libres',
 'Tiros Libres con remate con Remate', 'Tiros Libres con remate con Remate %', 'Penaltis', 'Penaltis marcados', '% de Penaltis marcados',
 'Centros', 'Centros precisos', '% de Centros precisos', 'Pases cruzados en profundidad completados', 'Pases en profundidad completados',
 'Entradas al Area', 'Carreras', 'Pases Cruzados', 'Toques en el área de penalti', 'Duelos ofensivos', 'Duelos ofensivos ganados',
 '% de Duelos ofensivos ganados', 'Fuera de juego', 'Goles recibidos', 'Tiros en Contra', 'Tiros en contra a la portería',
 'Tiros en contra a la portería %', 'Duelos defensivos', 'Duelos defensivos ganados', '% de Duelos defensivos ganados', 'Duelos aereos',
 'Duelos aéreos ganados', '% Duelos aéreos ganados', 'Entradas a ras de suelo', 'Entradas a ras de suelo logradas',
 'Entradas a ras de suelo logradas %', 'Interceptaciones', 'Despejes', 'Faltas', 'Tarjetas amarillas', 'Tarjetas rojas', 'Pases hacia adelante',
 'Pases hacia adelante logrados', '% de Pases hacia adelante logrados', 'Pases hacia atrás', 'Pases hacia atrás logrados',
 '% de Pases hacia atrás logrados', 'Pases laterales', 'Pases laterales logrados', '% de Pases laterales logrados', 'Pases largos',
 'Pases largos logrados', '% de Pases largos logrados', 'Pases en el último tercio', 'Pases en el último tercio logrados',
 '% de Pases en el último tercio logrados', 'Pases progresivos', 'Pases progresivos precisos', '% de Pases progresivos precisos',
 'Desmarques', 'Desmarques logrados', '% de Desmarques logrados', 'Saques laterales', 'Saques laterales logrados', 'Saques laterales logrados %',
 'Saques de meta', 'Intensidad de paso', 'Promedio pases por posesión del balón', 'Lanzamiento largo %', 'Distancia media de tiro',
 'Longitud media pases', 'PPDA', 'Equipo', 'Marcador', 'Local', 'Visita', 'Resultado', 'Resultado_Partido', 'Formacion', 'Duracion Formacion']
```

Nota: Atributos registrados para cada equipo del campeonato chileno. Marcador, Local, Visita, Resultado, Formación, Duración formación creado por los autores.
Fuente: Elaboración propia en base a wyscout.

Por otro lado, contamos con una base de datos que contine el registro de los todos los jugadores chilenos en competición. Esta base de datos consiste en un set de 144 atributos para cada jugador que en general son el promedio de su rendimiento como futbolista. Es importante destacar que se cuenta además con la posición del jugador, el equipo donde milita, el valor de mercado y la situación contractual. Todos los atributos de los jugadores son presentados en la tabla 2.

La posición de cada jugador es especialmente relevante ya que servirá de control para la creación del ranking. Es importante destacar que esta base de datos contiene jugadores de distintas ligas, distintos equipos y que un jugador puede contar con múltiples posiciones dependiendo de sus propias características. Que un jugador tenga múltiples posiciones o roles significa un desafío adicional. En este sentido clasificaremos a cada jugador de acuerdo con la posición principal informada wysocut. Sumado a esto, resumimos el número de roles o posiciones a 7 clasificándolas de acuerdo con el criterio experto. La tabla 3 muestra el diccionario construido para dicha clasificación

Tabla 2: Atributos a Nivel de Jugador

```
Datos_Jugadores = ['Jugador', 'Equipo', 'Equipo durante el período seleccionado', 'Posición específica', 'Edad', 'Valor de mercado',
'Vencimiento contrato', 'Partidos jugados', 'Minutos jugados', 'Goles', 'xG', 'Asistencias', 'xA', 'Duelos/90',
'Duelos ganados, %', 'País de nacimiento', 'Pasaporte', 'Pie', 'Altura', 'Peso', 'En préstamo',
'Acciones defensivas realizadas/90','Duelos defensivos/90', 'Duelos defensivos ganados, %', 'Duelos aéreos en los 90',
'Duelos aéreos ganados, %', 'Entradas/90','Posesión conquistada después de una entrada', 'Tiros interceptados/90',
'Interceptaciones/90','Posesión conquistada después de una interceptación', 'Faltas/90', 'Tarjetas amarillas',
'Tarjetas amarillas/90','Tarjetas rojas', 'Tarjetas rojas/90', 'Acciones de ataque exitosas/90', 'Goles/90',
'Goles (excepto los penaltis)','Goles, excepto los penaltis/90', 'xG/90', 'Goles de cabeza', 'Goles de cabeza/90',
'Remates', 'Remates/90', 'Tiros a la portería, %', 'Goles hechos, %', 'Asistencias/90', 'Centros/90',
'Precisión centros, %', 'Centros desde la banda izquierda/90', 'Precisión centros desde la banda izquierda, %',
'Centros desde la banda derecha/90', 'Precisión centros desde la banda derecha, %', 'Centros al área pequeña/90',
'Regates/90', 'Regates realizados, %', 'Duelos atacantes/90', 'Duelos atacantes ganados, %',
'Toques en el área de penalti/90', 'Carreras en progresión/90', 'Aceleraciones/90', 'Pases recibidos /90',
'Pases largos recibidos/90', 'Faltas recibidas/90', 'Pases/90', 'Precisión pases, %','Pases hacia adelante/90',
'Precisión pases hacia adelante, %', 'Pases hacia atrás/90', 'Precision pases hacia atrás, %','Pases laterales/90',
'Precisión pases laterales, %', 'Pases cortos / medios /90', 'Precisión pases cortos / medios, %','Pases largos/90',
'Precisión pases largos, %', 'Longitud media pases, m', 'Longitud media pases largos, m', 'xA/90','Asistencias/90.1',
'Second assists/90', 'Third assists/90', 'Desmarques/90', 'Precisión desmarques, %', 'Jugadas claves/90',
'Pases en el último tercio/90', 'Precisión pases en el último tercio, %', 'Pases al área de penalti/90',
'Pases hacía el área pequeña, %', 'Pases en profundidad/90', 'Precisión pases en profundidad, %', 'Ataque en profundidad/90',
'Centros desde el último tercio/90', 'Pases progresivos/90', 'Precisión pases progresivos, %', 'Goles recibidos',
'Goles recibidos/90', 'Remates en contra', 'Remates en contra/90', 'Porterías imbatidas en los 90', 'Paradas, %',
'xG en contra', 'xG en contra/90', 'Goles evitados', 'Goles evitados/90', 'Pases hacía atrás recibidos del arquero/90',
'Salidas/90', 'Duelos aéreos en los 90.1', 'Tiros libres/90', 'Tiros libres directos/90', 'Tiros libres directos, %',
'Córneres/90', 'Penaltis a favor', 'Penaltis realizados, %']
```

Nota: Atributos registrados para cada jugador del chileno.
Fuente: Elaboración propia en base a wyscout.

Tabla 3: Clasificación de los Roles o Posiciones de Jugadores en Base al Criterio Experto

```
posicion = {'GK': 'Arquero', 'AMF': 'Mediocampo ofensivo',
            'RAMF': 'Mediocampo ofensivo', 'LAMF': 'Mediocampo ofensivo',
            'CF': 'Delantero', 'CB': 'Defensa', 'RCB': 'Defensa',
            'RB': 'Defensa', 'LCB': 'Defensa', 'LB': 'Defensa',
            'RW': 'Lateral derecho', 'RWF': 'Lateral derecho',
            'RWB': 'Lateral derecho', 'LW': 'Lateral izquierdo',
            'LWF': 'Lateral izquierdo','LWB': 'Lateral izquierdo',
            'CMF': 'Mediocampo','RCMF': 'Mediocampo',
            'LCMF': 'Mediocampo','DMF': 'Mediocampo defensivo',
            'RDMF': 'Mediocampo defensivo','LDMF': 'Mediocampo defensivo'
}
```

Nota: Clasificación de los roles de un jugador de acuerdo a la posición principal informada por wyscout y al criterio experto.
Fuente: Elaboración propia en base a wyscout.

Una vez descritas los sets de datos que serán utilizados en nuestro análisis exploraremos las estadísticas descriptivas de los principales atributos.

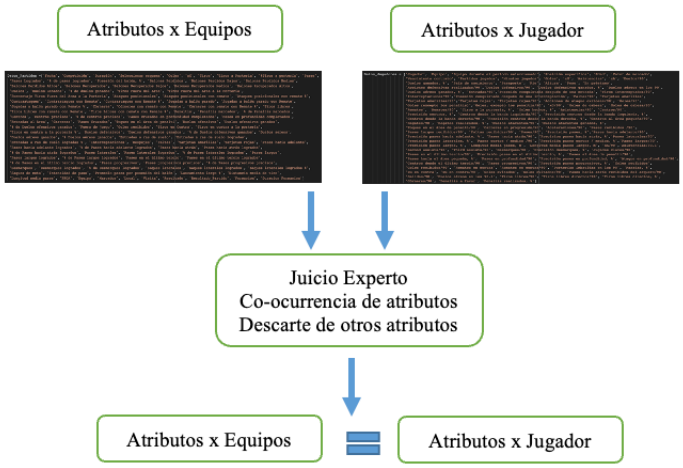
4.2. Metodología

La metodología planteada en este trabajo apuntará a responder una de las principales interrogantes a las que se enfrenta la industria del futbol profesional, que es saber cómo evaluar de forma objetiva el desempeño de distintos jugadores, ya sea por posición, de manera absoluta y/o colectiva. Con esto en mente construiremos un ranking de los jugadores profesionales chilenos que se basará en atributos estadísticos que contribuirán resultado final de un partido.

Como se describió en la sección anterior, contamos con un set de atributos a nivel de equipos y con un set de atributos a nivel de jugador. El principal desafío es encontrar una co-ocurrencia en ambas bases de datos, es decir, tener los mismos a atributos. Para realizar este procedimiento de co-ocurrencia nos basamos en el criterio experto para clasificar y etiquetar los atributos. En definitiva, el experto decidió que atributos eran co-ocurrentes

en ambas bases de datos y descarto los demás. La figura 1 resume este proceso. Este proceso es fundamental porque se entrenará y estimará el modelo sólo con estos atributos.

Figura 1: Co-ocurrencia entre Atributos por Equipo y por Jugador



Nota: La figura 1 describe el proceso de co-ocurrencia entre los atributos por equipo y por jugador. Este proceso es fundamental porque se entrenará y estimará el modelo sólo con estos atributos.
Fuente: Elaboración propia en base a wyscout.

Una vez determinados los atributos a utilizar con el procedimiento de co-ocurrencia estableceremos el desempeño de un equipo a través del resultado de cada partido que clasificaremos como victoria y no victoria. Luego con los atributos ya seleccionados entrenaremos un modelo para determinar el peso relativo de cada uno en el desempeño del equipo. Finalmente, el ranking de jugadores será una función de estos pesos y los atributos de cada jugador. A continuación, detallaremos el proceso metodológico que será la base de nuestro análisis. Este proceso se dividirá en cuatro partes: 1) Extracción de Atributos, 2) Problema de Clasificación, 3) Ponderación de los Atributos y 4) Ranking del Jugador.

4.2.1. Extracción de Atributos.

Para un partido p se tiene un set de atributos:

$$A_e^p = [x_1, x_2, \ldots, x_n]$$

donde x_i es un atributo que resume un aspecto especifico de la conducta del equipo e en el partido p. Por ejemplo,

$$A_{CC}^5 = [centros_{CC}^5, \%\ duelos\ ganados_{CC}^5, \ldots, Tarjetas\ Rojas_{CC}^5]$$

4.2.2. Problema de Clasificación.

Dentro de todos los atributos presentes en el punto anterior podemos construir el resultado de partido como la diferencia absoluta entre los goles marcados y los goles recibidos en el partido.

Así, el resultado de un partido p para un equipo e será R_e^p , donde $R_e^p = 1$ indica una victoria del equipo e en el partido p y $R_e^p = 0$ indica no victoria (es decir, una derrota o un empate).

$$R_e^p = \begin{cases} 1 \; si \; el \; equipo \; e \; gana \; el \; partido \; p \\ 0 \; en \; otro \; caso \end{cases}$$

Con esto, resolvemos el problema de clasificación entre el desempeño de un equipo R_e^p y los atributos del partido A_e^p. Pappalardo y Cintia (2017) demostraron que este problema de clasificación es significativo, porque existe una fuerte relación entre el vector de atributos, A_e^p del equipo y el resultado del partido R_e^p. Como se observa en la figura 2, la distribución del resultado de los partidos parece equilibrada con un 40% de victorias y un 60% de no victorias (derrotas más empates). Este resultado es de especial relevancia a la hora de estimación de cualquier algoritmo de clasificación ya que permite asegurar un resultado estable e insesgado.

Figura 2: Distribución del Resultado de los Partidos

Nota: La figura 2 muestra el resultado de clasificar el desempeño de cada equipo a través del resultado del partido, victoria o no victoria. Podemos observar que existe un 40% de victorias y un 60% de no victorias. Esta distribución de los resultados aseguraría una buena estimación del modelo.
Fuente: Elaboración propia en base a wyscout.

4.2.3. Ponderación de los Atributos

Una vez resuelto el problema de desempeño del equipo, podemos establecer una relación entre R_e^p y A_e^p y obtener un vector con las ponderaciones $\boldsymbol{w}$. Este vector $\boldsymbol{w}$ será aquel que encuentra el mejor resultado. Es decir, obtendremos $\boldsymbol{w}$ de la siguiente manera**:**

$$R_e^p = f(A_e^p|\boldsymbol{w})$$

Donde $\boldsymbol{w} = [w_1, w_2, \ldots, w_n]$ es el vector de ponderadores w_i que acompaña a cada x_i. Finalmente, estos ponderadores cuantifican la influencia de los distintos atributos x_i en el resultado del partido. Para obtener $\boldsymbol{w}$, usamos un clasificador de vectores de soporte lineal (LSVC).

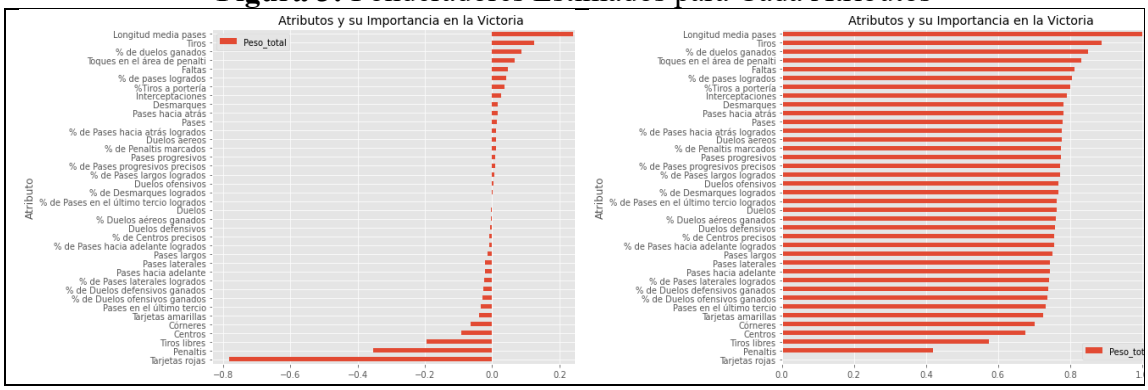
4.2.4. Atributos por jugador

Como se discutió en la sección de datos, contamos con dos bases de datos relevantes para este trabajo: 1) La base de atributos por equipo y 2) la base de atributos por jugador. Si bien ambos sets de datos vienen del mismo proveedor la información contenida en cada uno es distinta. Sin embargo, en base a un trabajo de clasificación manual y de conocimiento experto se realiza un procedimiento de co-ocurrencia en ambas bases de datos y así contar con la misma información.

Después de proceso de ingeniera inversa quedamos finalmente con 38 atributos por equipo y jugador. La figura dos muestra los atributos y los pesos estimados correspondientes al desempeño del equipo. Es importante destacar que se utilizaron otros métodos de estimación y distintas muestras de entrenamiento y los resultados no cambiaron significativamente.

Encontramos que las funciones basadas en asistencia son las más importantes, seguidas por los tiros, el porcentaje de duelos ganados y los pases dentro del área. Por el contrario, recibir una tarjeta roja/amarilla tiene un fuerte peso negativo, junto con los cobros de penales y tiros libres en contra. Es interesante notar que, aunque estas elecciones son bastante naturales para quienes son expertos en evaluaciones de jugadores de fútbol, el modelo implementado los derivó automáticamente con solo mirar los registros masivos de fútbol en nuestro conjunto de datos. Finalmente, siguiendo a Pappalardo, et al. (2019), escalamos estos ponderadores entre 0 y 1 para garantizar que todos los atributos están expresados en la misma escala.

Figura 3: Ponderadores Estimados para Cada Atributos



Nota: La figura 3 muestra el peso estimado para cada atributo usando Clasificador de vectores de soporte lineal (LSVC). Es importante notar que estos 38 atributos son aquellos que se encuentran disponible tanto en la base de datos de equipo como de jugadores
Fuente: Elaboración propia en base a datos de wyscout.

5. Resultados

5.1. Ranking Jugador

Una vez escalados los ponderadores obtenidos a través de LSVC y ejecutado el procedimiento de co-ocurrencia entre la base de datos de equipos y jugadores utilizaremos la misma lógica para evaluar su desempeño. Es decir, los mismos atributos utilizados para la evaluación del rendimiento de un equipo se utilizan para evaluar el desempeño de un jugador. Este desempeño se evaluará a través de un ranking calculado como el producto escalar entre los ponderadores obtenidos para cada atributo y los atributos correspondientes a cada jugador j. Formalmente:

$$r(j) = \sum_{i=1}^{n} w_i \times x_i^j$$

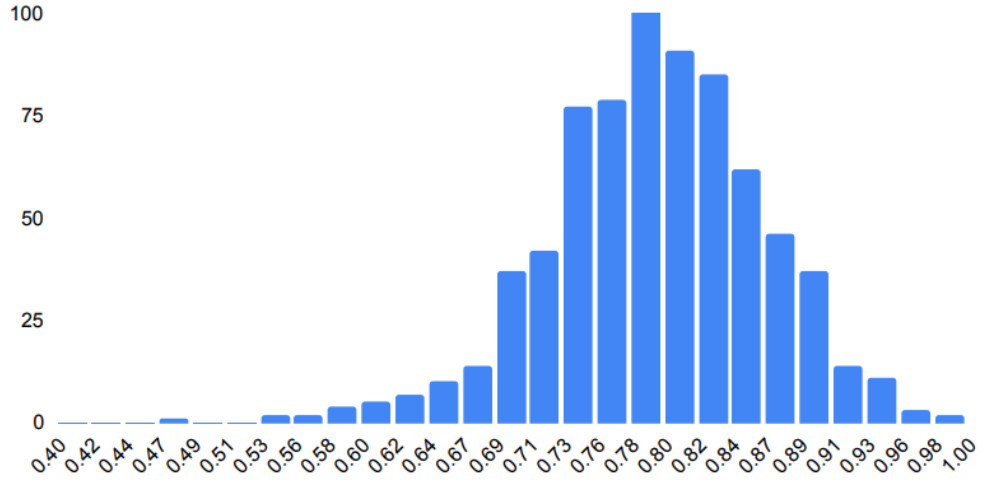
El ranking r de un jugador j será la suma ponderada de los atributos de cada jugador x_i y los pesos obtenidos en la fase anterior $\boldsymbol{w}$. Finalmente se proponen dos medidas para interpretar los resultados. La primera es un ranking agregado y escalado entre 0 y 1 para todos los jugadores (figura 4). La segunda medida es un ranking escalado entre 0 y 1 pero relativo a la posición de cada jugador. De esta forma podremos obtener el mejor jugador por cada posición dado que vamos a estandarizar por la media y varianza relativa a cada rol.

La figura 4 muestra la distribución del ranking agregado. Podemos observar una distribución normal centrada en $\mu = 0.80$ con $\sigma = 0.07$, lo que aseguraría que un 95% de los jugadores se encuentra en el intervalo $[\mu \pm 2\sigma]$. Este resultado es coherente con lo encontrado en la literatura (Papalardo y Candia, 2017) donde el desempeño de los

jugadores se encuentra dentro de un rango esperado, sin embargo, existen jugadores que tienen atributos y actuaciones por sobre lo esperado que es precisamente lo encontrado en este trabajo. Siguiendo este análisis, los mejores jugadores se encontrarían en la cola derecha de la distribución. Así, existen 20 jugadores a la derecha de este intervalo, es decir, $r(j) > 0.93$. En la tabla 4, mostramos los 20 jugadores que se encuentran en la cola derecha de la distribución ordenados de mayor a menor junto con el equipo al que pertenecen la posición en la que juegan dentro del equipo y el valor de mercado. Existe una relación positiva pero débil entre el ranking y el valor de mercado, Este efecto será abordado en futuras investigaciones.

Otro hecho destacable de estos resultados es la diferencia existente entre posiciones. El mejor desempeño lo tienen los mediocampistas defensivos, $\mu_{MD} = 0.836$, seguidos por los mediocampistas, $\mu_M = 0.818$, los defensas con una media de $\mu_{\mathrm{D}} = 0.815$ y los mediocampistas ofensivos, $\mu_{MO} = 0.806$ aunque esta diferencia no parecer ser estadísticamente significativa entre mediocampo y defensas. Un hecho que llama poderosamente la atención es que los delanteros tienen un desempeño inferior a los defensas y mediocampistas con una media de $\mu_F = 0.757$ comparable al desempeño encontrados para los laterales derechos e izquierdos $\mu_{LD} = 0.752$ y $\mu_{LI} = 0.769$ respectivamente. Por último, en base a esta metodología era esperable encontrar que los arqueros tienen el peor desempeño $\mu_A = 0.713$.

Figura 4: Histograma Ranking Agregado



Nota: La figura 4 muestra la distribución del ranking agregado. Podemos observar una distribución normal centrada en $\mu = 0.80$ con $\sigma = 0.07$. Los mejores jugadores se encuentran en la cola derecha. Dado que la distribución es aproximadamente normal el intervalo $[\mu \pm 2\sigma]$ contiene el 95% de los jugadores. Existen 16 jugadores a la derecha este intervalo, es decir, *r(j)>0.93.*
Fuente: Elaboración propia en base a datos obtenidos por wyscout.

Tabla 4: Ranking de los mejores 20 jugadores.

Ranking	Edad	Jugador	Equipo	Posición	Valor de mercado
1	30	L. Reyes	Ñublense	Posicion	500000
2	22	F. Méndez	Unión Española	Mediocampo defensivo	2000000
3	37	C. Cortés	Magallanes	Mediocampo	100000
4	29	C. Sepúlveda	Huachipato	Delantero	550000
5	21	W. Alarcón	Unión La Calera	Mediocampo defensivo	300000
6	32	J. Castillo	Santiago Wanderers	Mediocampo defensivo	125000
7	25	N. Ramírez	Huachipato	Delantero	375000
8	28	D. Valdés	América	Defensa	6000000
9	22	A. Ríos	Cobreloa	Delantero	100000
10	23	M. Allende	Torque	Defensa	1600000
11	27	J. Abrigo	Coquimbo Unido	Mediocampo	700000
12	25	J. Méndez	Puerto Montt	Delantero	200000
13	28	N. Vargas	Ñublense	Defensa	300000
14	32	E. Vargas	Atlético Mineiro	Defensa	3000000
15	30	A. Césped	Unión San Felipe	Delantero	125000
16	29	J. Cantillana	PSIS Semarang	Mediocampo	300000
17	35	A. Vidal	Internazionale	Delantero	2500000
18	36	G. Jara	Coquimbo Unido	Mediocampo	200000
19	28	B. Ampuero	Universidad Católica	Defensa	650000
20	28	N. Maturana	Cobreloa	Defensa	250000

Nota: La tabla 4 muestra el ranking de los mejores 20 jugadores junto con el equipo al que pertenecen la posición en la que juegan dentro del equipo y el valor de mercado. Existe una relación positiva pero débil entre el ranking y el valor de mercado, Este efecto será abordado en futuras investigaciones.
Fuente: Elaboración propia en base a wyscout.

Tabla 5: Estadísticas descriptivas del Ranking agregado

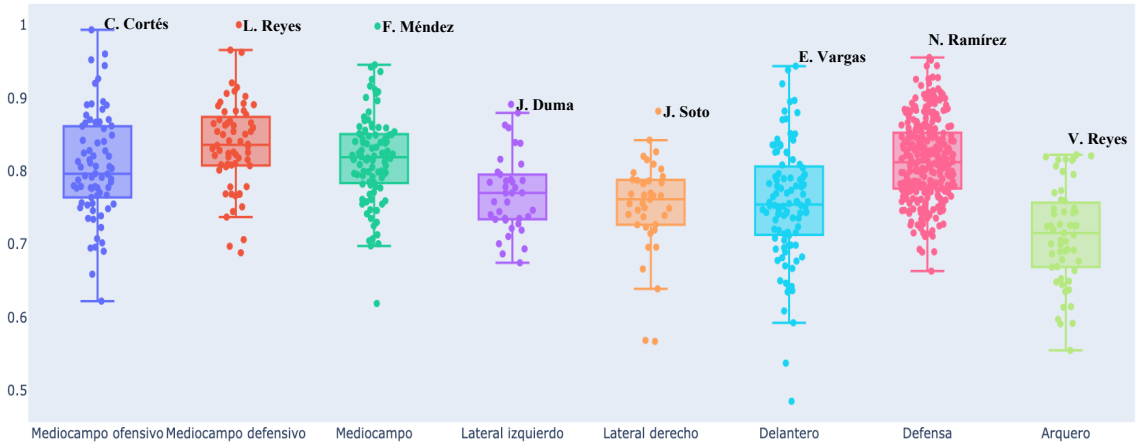
Posicion	count	mean	std	min	25%	50%	75%	max
Arquero	55.0	0.713002	0.067554	0.554915	0.668961	0.715426	0.754293	0.822379
Defensa	269.0	0.814815	0.056281	0.663225	0.776203	0.812176	0.852358	0.955313
Delantero	100.0	0.757415	0.079757	0.485361	0.713395	0.754297	0.805307	0.943408
Lateral derecho	39.0	0.752182	0.063914	0.567170	0.726656	0.761396	0.787779	0.881768
Lateral izquierdo	38.0	0.769596	0.053175	0.674584	0.734333	0.770106	0.793576	0.891185
Mediocampo	97.0	0.817802	0.061025	0.618970	0.784058	0.818937	0.850257	0.998100
Mediocampo defensivo	64.0	0.836611	0.061294	0.688334	0.807918	0.835918	0.872079	1.000000
Mediocampo ofensivo	74.0	0.806587	0.071440	0.622207	0.764521	0.796325	0.860535	0.993010
Total	736.0	0.795216	0.072239	0.485361	0.749662	0.797143	0.842066	1.000000

Nota: La tabla 5 muestra las estadísticas descriptivas del ranking agregado construido para los jugadores chilenos. Además, de clasifican por posición encontrado diferencias significativas entre estas. Los arqueros, en promedio tienen el peor desempeño en ranking $\mu_A = 0.71$ mientras que mediocampistas defensivos $\mu_{MD} = 0.84$, mediocampistas $\mu_M = 0.818$, defensas de $\mu_D = 0.815$ y mediocampistas ofensivos $\mu_{MO} = 0.81$ muestran los mejores resultados. El desempeño de los delanteros $\mu_F = 0.76$ se encuentra en línea con el de los laterales derecho $\mu_{LD} = 0.75$ $\mu_{LI} = 0.77$ respectivamente.
Fuente: Elaboración propia en base a wyscout.

Finalmente, la figura 5 resume el principal resultado de esta investigación. Lo primero es mostrar explícitamente la distribución del ranking agregado por posición. Esta figura confirma los resultados presentados con anterioridad. Además, es posible observar que, en la mayoría de las posiciones, excepto arquero y defensas, existen desempeños sobre y bajo lo esperado. Siguiendo la estructura de la figura 5 el mejor desempeño de los

mediocampistas ofensivos lo obtiene Cesar Cortés de Magallanes. Dentro de los mediocampistas defensivos lo obtiene Lorenzo Reyes de Ñublense, para los mediocampistas el mejor es Felipe Méndez de Unión Española, dentro de los laterales izquierdos la mejor puntuación la obtiene Juan Duma de Barnechea. El mejor lateral derecho es Juan Soto de Cobreloa. Según nuestro modelo el mejor delantero es Eduardo Vargas de Atlético Mineiro. Dentro de los defensas existe menos dispersión y una alta concentración del ranking en torno a la media, donde el mejor defensa es Nicolas Ramírez de Huachipato. Finalmente, bajo esta metodología el mejor arquero es Vicente Reyes de Atalanta United 2.

Figura 5: Ranking Agregado por Posición



Nota: La figura 5, muestra los resultados del ranking agregado para cada uno de los futbolistas chilenos clasificados por posición. Se confirman los resultados encontrados en las secciones anteriores, como, por ejemplo, el mejor desempeño promedio de los mediocampistas y más bajo para los laterales, delanteros y arquero. De esta figura también es posible deducir la distribución de jugadores para cada posición y detectar los de mejor desempeño. Siguiendo la estructura de la figura 5 el mejor desempeño de los mediocampistas ofensivos lo obtiene Cesar Cortés de Magallanes, de los mediocampistas defensivos lo obtiene Lorenzo Reyes de Ñublense, para los mediocampistas el mejor es Felipe Méndez de Unión Española, dentro de los laterales izquierdos la mejor puntuación la obtiene Juan Duma de Barnechea. El mejor lateral derecho es Juan Soto de Cobreloa. Según nuestro modelo el mejor delantero es Eduardo Vargas de Atlético Mineiro. Dentro de los defensas existe menos dispersión y una alta concentración del ranking en torno a la media, donde el mejor defensa es Nicolas Ramírez de Huachipato. Finalmente, bajo esta metodología el mejor arquero es Vicente Reyes de Atalanta United 2. Para más detalle visitar https://sebabecerra.github.io/Futbol-Ranking/
Fuente: Elaboración propia en base a wyscout.

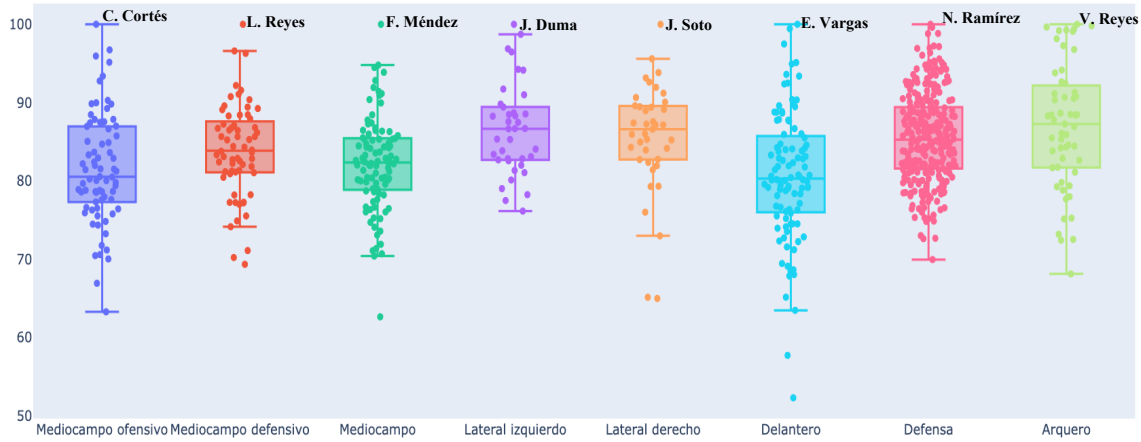
5.2. Estandarización del Ranking por Posición

Dada la naturaleza distinta de las posiciones, donde existe un entrenamiento y especialización diferente, además de un sesgo natural de los propios jugadores y/o entrenadores hacia distintos roles es natural corregir el ranking por las propiedades estadísticas de cada posición. De esta manera, realizaremos la siguiente estandarización:

$$r_j^{posicion} = 100 \times \left[\frac{r_j - \mu^{posicion}}{\sigma^{posicion}} \right]$$

Donde $r_j^{posicion}$ será el ranking por posición para cada jugador j estandarizado por la media ($\mu^{posicion}$) y la desviación estándar ($\sigma^{posicion}$) correspondiente. Con esto es posible de forma simple realizar la conformación de un equipo dependiendo de la alineación que decida el entrenador, buscando en el ranking por posición los mejores defensas, delanteros, laterales, mediocampista y arqueros. La Figura 6 muestra este resultado.

Figura 6: Ranking Estandarizado por Posición



Nota: La figura 6, muestra los resultados del ranking estandarizado por posición para cada uno de los futbolistas chilenos. Para más detalle visitar https://sebabecerra.github.io/Futbol-Ranking/
Fuente: Elaboración propia en base a wyscout.

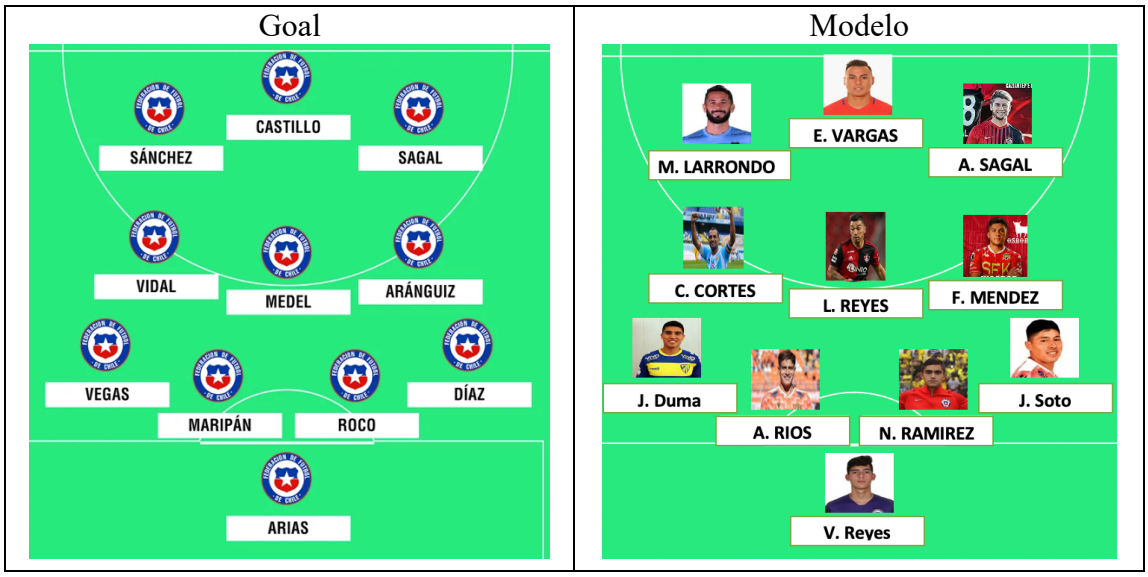
5.3. Diseño del equipo en base a una formación definida exante

Una vez que es posible determinar un ranking por posición podemos seleccionar el o los mejores jugadores requeridos para la conformación de un equipo. Por ejemplo, el portal goal.com apostó por cual sería la alineación de Chile para el mundial de Qatar 2022. Siguiendo esta lógica decidimos construir nuestro 11 ideal seleccionando los mejores jugadores de acuerdo con el esquema táctico presentado por Goal un 4-3-3. La figura 6 muestra la comparativa entre la alineación entregada por Goal y la encontrada a través de nuestra metodología.

La comparación entre una selección chilena realizada en base al criterio de los expertos y a la obtenida directamente del ranking propuesto en este trabajo es realmente interesante y llama a la discusión. Lo primero que salta a la vista es lo diametralmente distintas de estas dos propuestas. En la selección en base a criterios expertos se encuentran los jugadores más conocidos, de renombre mundial y con una alta cotización en el mercado. Por otra parte, la selección chilena, basada en nuestro metodología y basada sólo en datos propone para esta alineación jugadores con altos atributos pero que en la mayoría son desconocidos, juegan en el medio local y tienen un valor de mercado más bajo. Un hecho podría explicar estos resultados, es que nuestra metodología busca, en los jugadores,

atributos que aportan al resultado como equipo, de ahí la importancia, por ejemplo, de los pases bien realizados. En este sentido nuestro modelo capturaría el aporte de estos atributos al desempeño global del equipo descartando otros de carácter individual como por ejemplo la velocidad o la edad del jugador. Finalmente, es importante decir que este modelo no busca desplazar al criterio experto, sino que buscar complementarlo y aportar en busca de jugadores de alto impacto, pero de bajo valor de mercado que potenciarían a los equipos.

Figura 7: Selección Chilena que jugaría en Qatar 2022 según Goal. vs la predicha por el Modelo.



Nota: La figura 7 muestra la comparación entre una selección chilena realizada en base al criterios de los experto y a la obtenida directamente del ranking propuesto en este trabajo. Es interesante notar lo distintas de estas dos propuestas. En la selección en base a criterios expertos se encuentran los jugadores más conocidos, de renombre mundial y con una alta cotización en el mercado. Por otra parte, el modelo basado sólo en datos propone para esta alineación jugadores con altos atributos pero que en la mayoría son desconocidos, juegan en el medio local y tienen un valor de mercado más bajo. Es importante decir que este modelo no busca desplazar al criterio experto sino que buscar complementarlo y aportar en busca de jugadores de alto impacto pero bajo valor de mercado que potenciarían a los equipos.
Fuente: Elaboración propia.

6. Conclusiones

La revolución de los datos y los enfoques data-driven han revolucionado las principales industrias y el deporte no es la excepción. Cada día es más relevante poder contar con herramientas basadas en datos que permitan apoyar el criterio experto y la toma de decisiones. En este sentido, poder cuantificar el aporte de cada integrante al desempeño final de un equipo, evaluar contrataciones y poder estimar el potencial impacto que un nuevo jugador tendría en el funcionamiento del equipo es uno de los desafíos centrales de este mercado (Bornn et al. 2018, Gerrard 2017 y Pappalardo 2019).

El objetivo de este trabajo fue desarrollar un esquema basado en datos para la conformación de un plantel de futbol. Para esto seguimos tres pasos. El primero es explorar un extenso set de datos que resume el desempeño de equipo. Con el construir una variable que permita clasificar si el equipo obtiene un buen desempeño, es decir, gana un partido o si obtiene un mal desempeño, es decir, pierde o empata.

Una vez establecido esto, el segundo paso fue determinar los atributos que aportan de manera significativa y positiva a que el equipo aumente su probabilidad de ganar y cuales atributos aportan de manera negativa a este atributo. Esta relación se puede estimar a través de distintos modelos de machine learning, clasificación, aprendizaje no supervisado y aprendizaje supervisado. Específicamente en nuestro caso estimaremos un clasificador de vectores de soporte lineal (LSVC).

Tercero, una vez estimado el modelo y obtenido los ponderadores podemos crear un ranking de jugadores de futbol donde su puntaje sea directamente proporcional a la probabilidad de que el equipo obtenga un buen desempeño.

Finalmente, una vez desarrollados los tres pasos construimos un ranking de jugadores basado enteramente en propiedades estadísticas. Esto nos dio la posibilidad de proponer un plantel de 11 jugadores, donde cada uno de ellos representa al mejor jugador en la posición descrita. Si un equipo necesita más de un jugador en una posición simplemente se sigue el orden del ranking para selección al segundo mejor y así sucesivamente.

Este enfoque basado en datos puede parecer controversial, porque selecciona jugadores desconocidos o con bajo valor de mercado, por ejemplo. Sin embargo, esto puede ser una gran herramienta de democratización para los jóvenes talentos. Esto también puede ser un gran beneficio para los clubes que se aventuren a probar un enfoque basado en datos, ya que podrían descubrir grandes jugadores a bajo costo que le dan un gran rendimiento al equipo y que después pueden significar importantes negocios para el club.

A pesar de los alentadores resultados obtenidos debemos mencionar algunas limitaciones encontradas a la metodología. Primero, debe interpretarse con cuidado los resultados para los arqueros, ya que, por su rol especifico podrían ser relevantes otros atributos. Segundo,

existen diferencias significativas entre las ligas donde compiten cada uno de estos jugadores. Es un efecto importante que por ahora no hemos considerado y que para un trabajo futuro controlaremos.

Finalmente, este trabajo tiene múltiples extensiones, desde aplicarlo a diferentes deportes, sociabilizarlo con los expertos para ver que tanto sentido tiene la selección de jugadores hecha por el modelo. Otra extensión relevante de este trabajo seria considerar variables o atributos que capturen el desempeño colectivo del equipo, ya que, en el deporte, y especialmente en el futbol es bien sabio que …

"Ningún jugador es tan bueno como todos juntos"
Alfredo Di Stéfano

Bibliografía

1. Joel Brooks, Matthew Kerr, and John Guttag. 2016. Developing a data-driven player ranking in soccer using predictive model weights. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. 49–55.

2. Paolo Cintia, Fosca Giannotti, Luca Pappalardo, Dino Pedreschi, and Marco Malvaldi. 2015. The harsh rule of the goals: Data-driven performance indicators for football teams. In *Proceedings of the IEEE International Conference on Data Science and Advanced Analytics*.
DOI: https://doi.org/10.1109/DSAA.2015.7344823

3. Jordi Duch, Joshua S. Waitzman, and Luís A. Nunes Amaral. 2010. Quantifying the performance of individual players in a team activity. PLOS ONE 5, 6 (2010), 1–7.
DOI: https://doi.org/10.1371/journal.pone.0010937

4. Patrick Lucey, Dean Oliver, Peter Carr, Joe Roth, and Iain Matthews. 2013. Assessing team strategy using spatiotemporal data. In *Proceedings of the 19th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. 1366–1374.
DOI: https://doi.org/10.1145/2487575.2488191

5. Edward Nsolo, Patrick Lambrix, and Niklas Carlsson. 2018. Player valuation in European football. In *Proceedings of the Machine Learning and Data Mining for Sports Analytics workshop (MLSA'18)*.

6. Luca Pappalardo and Paolo Cintia. 2017. Quantifying the relation between performance and success in soccer. *Adv. Complex Syst.* 20, 4 (2017).
DOI: https://doi.org/10.1142/S021952591750014X

7. Luca Pappalardo, Paolo Cintia, Paolo Ferragina, Emanuele Massucco, Dino Pedreschi, and Fosca Giannotti. 2019. PlayeRank: Data-driven Performance Evaluation and Player Ranking in Soccer via a Machine Learning Approach. ACM Trans. Intell. Syst. Technol. 10, 5, Article 59 (September 2019), 27 pages.
DOI: https://doi.org/10.1145/3343172

8. Oliver Shulte and Zeyu Zhao. 2017. Apples-to-apples: Clustering and ranking NHL players using location information and scoring impact. In *Proceedings of the MIT Sloan Sports Analytics Conference*.