UNIVERSIDAD DE CONCEPCIÓN

FACULTAD DE INGENIERÍA

DEPARTAMENTO DE INGENIERÍA CIVIL INDUSTRIAL

MODELO DE MACHINE LEARNING PARA EVALUACIÓN DEL RENDIMIENTO DE JUGADORES DE FÚTBOL

POR

Cristóbal Orlando Burdiles Gutiérrez

Memoria de Título presentada a la Facultad de Ingeniería de la Universidad de Concepción para optar al título profesional de Ingeniero Civil Industrial

Profesor Guía

Carlos Contreras Bolton

Profesor Co-Guía

Pablo Galaz Cares

Diciembre 2022

Concepción (Chile)

© 2022 Cristóbal Orlando Burdiles Gutiérrez

Dedicatoria

A aquellos quienes aún creen en el amor, la libertad y la búsqueda de la verdad.

Agradecimientos

A todas las personas que alguna vez coincidí en la vida.

Resumen

La evaluación del desempeño de forma objetiva no es una tarea fácil, puesto que generalmente se realiza por personas, las cuales tienen sus propios sesgos y juicios personales. Este trabajo presenta una herramienta para evaluar el rendimiento deportivo de futbolistas mediante modelos de machine learning basados en datos estadísticos de la liga chilena de fútbol profesional. Se presenta una escala de evaluación que asigna una nota a cada jugador para cada partido disputado. La nota se calcula a partir de un modelo de regresión lineal donde los predictores varían de acuerdo a la posición del campo de juego donde se desempeña el jugador. La regresión se escoge como modelo para estimar luego de comparar su carácter predictivo, mediante la raíz del error cuadrático medio (RMSE), con otros modelos como Red Neuronal (2,98), Support Vector Machine (5,57) y Random Forest (2,44). Si bien su estimación no es precisa (4,42), se escoge debido a su interpretabilidad y amplio uso en la literatura. El modelo permite comparar el rendimiento entre distintos jugadores, tanto para un partido en específico como a lo largo del torneo. Como resultado se obtienen las variables que son significativas, para cada posición y explicar el rendimiento del jugador. Se concluye que existen variables relevantes para evaluar independiente de la posición del campo de juego, mientras que existen otras que solo son relevantes para una posición en específico.

Palabras Claves – Rendimiento deportivo, regresión lineal, evaluación del desempeño.

Abstract

Objective performance appraisal is not an easy task, since it is usually done by people, who have their own personal biases and judgments. This work presents a tool to evaluate the sports performance of soccer players through machine learning models based on statistical data from the Chilean professional soccer league. An evaluation scale is presented that assigns a grade to each player for each match played. The score is calculated from a linear regression model where the predictors vary according to the position of the playing field where the player plays. The regression is chosen as the model to estimate after comparing its predictive character, through the root mean square error (RMSE), with other models such as Neural Network (2,98), Support Vector Machine (5,57) and Random Forest (2,44). Although its estimate is not the most precise (4,42), it is chosen due to its interpretability and wide use in the literature. The model allows to compare the performance between different players, both for a specific match and throughout the tournament. As a result, the variables that are significant, for each position, in explaining the player's performance are obtained. It is concluded that there are variables that are relevant to evaluate regardless of the position of the playing field, while there are others that are only relevant for a specific position.

Keywords – sports performance, linear regression, performance evaluation

Contenido

Lista de Tablas

Lista de Figuras

1. Introducción

En este capítulo se presenta una introducción general a las evaluaciones de rendimiento en el deporte. Además, se presenta el objetivo general, los objetivos específicos y la organización del documento.

1.1 Introducción General

Toda empresa buscar crecer, mejorar y eventualmente superar a la competencia. Para esto necesita empleados productivos y capaces. Una forma de medir la productividad es con evaluaciones del desempeño. Actualmente las empresas disponen de algún componente de evaluación de desempeño con el objetivo de implementar cambios o mejoras si es necesario. La industria deportiva no es la excepción, y el desempeño de un deportista de nivel profesional se puede medir por factores técnicos, físicos, sicológicos, incluso personales.

Una particularidad de la industria deportiva es que sus planteles se renuevan con frecuencia. Existen dos periodos durante el año, el mercado de pases de verano e invierno, donde los equipos pueden realizar contrataciones de otros jugadores, ya sea que se encuentran libres (sin contrato) o jugando en algún otro club (FIFA, 2022). La alta rotación de los planteles motiva a evaluar el desempeño no solo los jugadores del propio club, sino también de los rivales o de mercados alcanzables en término económicos. En Chile, los principales mercados para contratar futbolistas son: Chile, Argentina, Uruguay, Paraguay, Perú, Venezuela, entre otros (Transfermarkt, 2022).

El fútbol está tan impregnado en la cultura popular chilena que tiene un sinfín de opiniones mayoritariamente subjetivas, cada una influida por la forma y cursiva de juego del espectador. Es aquí donde surge el problema evaluar el rendimiento de los jugadores de forma objetiva, la valoración de los jugadores y equipos se realiza mayormente de forma subjetiva, sin establecer criterios claros de comparación. El acercamiento que propone este trabajo para dar respuesta a la problemática es realizar la evaluación a partir de los datos.

En los datos existe información que corresponde principalmente a pruebas físicas, como los obtenidos por un club en la evaluación de exámenes médicos antes de ser oficialmente contratado por el club, con el objetivo de asegurarse que el jugador llega en condiciones de salud mínimas. Incluso, existen clínicas en Chile como la Clínica Alemana, que ofrecen evaluaciones del rendimiento deportivo a partir de los datos y desde un punto de vista fisiológico. De manera de detectar los ámbitos susceptibles de mejora para optimizar los resultados obtenidos (Fernández, 2016).

Adicionalmente, existen datos sobre el número de acciones, porcentajes, esperanza o datos personales de los jugadores para un cierto partido. Si este tipo de datos se trabajan de forma agrupada mediante técnicas estadísticas y de ciencia de datos, se pueden obtener hallazgos interesantes sobre el rendimiento de un jugador.

Por tanto, en este estudio se aborda el problema de la subjetividad en la evaluación del rendimiento y para abordarlo, se propone establecer una escala de evaluación a partir de modelos de machine learning para comparar el rendimiento de los jugadores del campeonato nacional de Chile de forma objetiva.

1.2 Objetivo general

Proponer un modelo para evaluar el rendimiento de futbolistas en base a variables observables.

1.3 Objetivos específicos

- Revisar la literatura con las distintas formas de evaluar el rendimiento deportivo
- Seleccionar las variables significativas que explican el rendimiento de los jugadores de fútbol, para cada posición en el campo de juego, mediante un modelo.
- Comparar el carácter predictivo del modelo propuesto con otros modelos.
- Evaluar rendimiento de ciertos jugadores.

1.4 Organización del documento

El documento sigue la siguiente estructura. En el Capítulo 2 se presenta el problema, con revisión de la literatura. Luego, en el Capítulo 3 se define la metodología propuesta. En el Capítulo 4 se presentan los resultados del trabajo y la discusión. Finalmente, el Capítulo 5 presenta las conclusiones.

2. Problema de la evaluación del rendimiento

En este capítulo se define la problemática surgida al evaluar rendimiento. Definiéndolo desde un enfoque organizacional y cómo el problema afecta a la organización. Además, las limitaciones para abordar el problema.

2.1 Introducción

Como se menciona, cada empresa tiene sus propias formas de evaluar el desempeño de sus empleados, pues no todas persiguen los mismos objetivos y pueden llegar a tener visiones muy distintas de lo que significa un buen desempeño. Surgen problemas relacionados qué tan objetiva puede ser la evaluación del desempeño. Considerando factores como el exceso de subjetividad al evaluar o enfocarse principalmente en el desempeño del último periodo. Generando una evaluación de desempeño sesgada, que podría eventualmente, por ejemplo, generar conflictos con los mismos empleados.

En el caso particular del fútbol profesional, al menos en Chile, el proceso de scouting se desarrolla a partir de tres puntos claves:

- Situación contractual: Un jugador en condición de libre es preferible a otro jugador con contrato vigente en otro club. Debido a que para contratarlo requiere un costo de transferencia menor.
- Evaluación cualitativa a través de videos.
- Reporte cuantitativo básico, que incluye algunos datos como partidos jugados, minutos jugados, goles convertidos, asistencias realizadas o alguna otra métrica de particular interés para esa posición.

El grado de importancia de cada uno de los puntos depende principalmente de las personas que realizan el análisis. Pero, también influida por factores como el proyecto deportivo del club y filosofía de juego, generando un problema de sesgo en la evaluación.

2.2 Problemática

El principal problema que aparece cuando se desea evaluar jugadores son los sesgos y juicios personales condicionados por la emocionalidad en el momento de evaluar. Por tanto, una forma de abordar el problema es establecer una escala de evaluación propia a partir de modelos de machine learning para comparar el rendimiento de los jugadores del campeonato nacional de Chile de forma objetiva. Con el fin de minimizar sesgos y juicios personales. El supuesto es que las variables que explican el rendimiento de un jugador dependen de su posición en el campo de juego.

Una de las limitaciones que presenta este trabajo, son los datos para entrenar modelos ya que solamente son accesibles a través de empresas especializadas. Extraerlos de forma propia no es factible por alto costo. La ventaja es que la extracción de datos se realiza para todos los partidos de un cierto campeonato y para todos los jugadores que tuvieron alguna participación en este, por lo que el riesgo de pérdida de información es casi nulo.

2.3 Revisión de literatura

En Seirul-lo (2009) establecen ciertos criterios a considerar en cualquier evaluación del rendimiento de deportes de equipos. En general, el rendimiento está muy relacionado al contexto en el que el jugador es evaluado. Por lo tanto, una evaluación desarrollada durante un partido oficial tiene mayor valor que una desarrollada en un entrenamiento a puertas cerradas.

Rodríguez (2013) construye un programa para la evaluación del rendimiento de los deportistas para facilitar y orientar a todos los entrenadores, preparadores físicos o cualquier persona inmersa dentro del proceso de entrenamiento deportivo en la evaluación cuantificable del rendimiento deportivo. El programa está diseñado y fundamentando en la teoría y metodología del entrenamiento deportivo de tal manera que permite el control de las variables a entrenar en el proceso deportivo.

Molina et al. (2013) examinan el miedo a la evaluación negativa y la autoestima como posibles factores modulares de la caída del rendimiento deportivo asociado a la presión psicológica. Los participantes del estudio con elevado nivel de miedo a la evaluación negativa experimentaron una caída significativa en el rendimiento deportivo durante la condición de alta presión. El estudio proporciona evidencia sobre la implicación del miedo a la evaluación negativa y la autoestima en el campo de la psicología de la actividad física y el deporte.

En González-Neira et al. (2014) se analiza la ingesta nutricional y la composición corporal, comprobando su relación con el rendimiento deportivo, en 17 jugadores del equipo semiprofesional de Torrelodones C.F de Madrid. El estudio concluye que la alimentación fue inadecuada en las jugadoras, no correspondiendo la ingesta de nutrientes con sus requerimientos. Por la importancia que la nutrición juega en la competición y rendimiento deportivo, se recomienda seguir trabajando para lograr una recomendación adecuada.

En García-López et al. (2018) desarrollan y validan una herramienta de evaluación del rendimiento de juego (GPET: Game Performance Evaluation Tool). GPET mide la toma de decisiones y la ejecución de acciones técnico tácticas en deportes de invasión. Los deportes de invasión son aquellos que enfrentan dos equipos y el objetivo es invadir el terreno del equipo contrincante, con el fin de alcanzar la meta con el móvil (aro, pelota, etc.) (Marín, 2016). Se concluye que la herramienta es adecuada para los fines relacionados con la evaluación del comportamiento técnico y táctico de atacantes con y sin balón.

En Ursino et al. (2019) se estudia el rendimiento deportivo en los artículos empíricos que incluyen variables psicológicas. El estudio concluye que el rendimiento deportivo depende de la disciplina deportiva, por tanto, cada disciplina debe establecer sus propios criterios de evaluación del rendimiento y estos no se pueden replicar de un deporte a otro.

En Chile, también existen trabajos relacionados a la analítica aplicada al fútbol profesional. A pesar de que su enfoque no está 100% orientado a la evaluación del rendimiento, esta es igualmente necesaria. A partir de esto, Galaz (2020) identifica qué jugadores aportan a obtener un mejor rendimiento colectivo y Mena (2021) realiza un método de simulación a través de inferencia

Bayesiana para medir el impacto de la inclusión de un futbolista en un equipo de la Premier League de Inglaterra.

3. Metodología

En este capítulo se presenta la metodología aplicada a la evaluación del rendimiento de jugadores, incluyendo los supuestos generales, bajo el fundamento futbolístico correspondiente. En primer lugar, se presenta la metodología de forma general, que se divide en tres etapas principales. Posteriormente, el desglose de las etapas.

3.1 Metodología general para evaluación de jugadores

La evaluación se realiza mediante una escala creada en base a modelos estadísticos y de aprendizaje automático. Estos buscan determinar las principales variables que explican el rendimiento de los jugadores, de acuerdo a las distintas posiciones en el terreno de juego. A partir de esto, se asignan notas de forma individual para cada fecha, las cuales permiten comparar el rendimiento del jugador a lo largo de las 34 fechas que dura el campeonato, o bien entre distintos jugadores.

La metodología se explica a partir de una posición particular en el terreno de juego y luego se presenta un resumen para el resto de las posiciones. La estructura es la siguiente: 1) se abordan los datos y como obtenerlos. 2) se explica cómo se trabaja con los datos y su limpieza. 3) se comparan distintos modelos y se selecciona uno en particular, y 4) se establecen las métricas para la evaluación del rendimiento.

3.2 Datos

Existen diversas empresas proveedores de datos asociada al deporte, cómo OptaSports (www.statsperform.com), Wyscout (www.wyscout.com), Be Soccer (es.besoccer.com), 365 Scores (www.365scores.com), InStat (www.football.instatscout.com). En particular, InStat es capaz de ofrecer una variable, *InStat Index*, que representa la evaluación del rendimiento del jugador.

InStat Index se calcula mediante un algoritmo automático que considera la contribución del jugador al éxito del equipo, la importancia de sus acciones, el nivel del oponente y el nivel del campeonato en el que juegan. Cada parámetro tiene un factor que cambia según el número de acciones y eventos en el partido. Hay un conjunto único de parámetros clave para cada posición (12 a 14 factores). El peso de los factores de acción difiere según la posición del jugador. Para calcular el *InStat Index*, el jugador debe pasar un tiempo en el campo y realizar un número mínimo de acciones.

Los datos utilizados en este trabajo corresponden a la cantidad de acciones específicas que el jugador realiza durante un partido, denominados datos de nivel 2. Esta colección fue proporcionada por el Instituto de Sistemas Complejos de Ingeniería (Universidad de Chile).

Las variables se agrupan en cinco diferentes categorías: generales, ofensivas, pérdidas de balón y recuperaciones, duelos, personales y XG rates, detallados en la Tabla 1.

Tabla 1. Variables agrupadas por categoría.

Generales	**Ofensivas**	**Pérdidas de balón y recuperaciones**	**Duelos**	**Personales**	**XG Rates**
Matches played	Goals	Lost balls	Challenges	Nationality	xG (Expected goals)
Minutes played	Assists	Lost balls in own half	Challenges won	Team	xG per shot
Starting lineup appearances	Chances	Ball recoveries	Challenges won, %	National team	Expected assists
Substitute out	Chances successful	Ball recoveries in opponent's half	Defensive challenges	Age	xG conversion
Substitutes in	Chances, % of conversion	Fouls	Defensive challenges won	Height	xG with a player on
Total actions	Chances created	Fouls suffered	Challenges in defence won, %	Weight	Opponent's xG with a player on
Successful actions	Shots	Tackles	Attacking challenges	Foot	Net xG (xG player on - opp. team's xG)
Successful actions, %	Shots on target	Ball interceptions	Attacking challenges won	National team (last match date, mm.yy)	Defensive xG (xG of shots made by guarded player)
Offsides	Shots on target, %	Free ball pick ups	Challenges in attack won, %	Youth national team (last match date, mm.yy)	Defensive xG per shot
Yellow cards	Shots wide		Air challenges	Position	
Red cards	Blocked shots		Air challenges won	Nombre	
InStat Index	Shots on post / bar		Air challenges won, %	Dorsal	
Fecha	Penalty			ID	
	Penalties scored				
	Penalty kicks scored, %				
	Passes				
	Accurate passes, %				
	Key passes				
	Key passes accurate				
	Crosses				
	Crosses accurate				
	Accurate crosses, %				
	Dribbles				
	Dribbles successful				
	Successful dribbles, %				
	Tackles successful				
	Tackles won, %				

Es importante diferenciar las variables positivas y negativas. Las positivas son las que benefician al equipo, por ejemplo: *goals*, *assists*, *passes*. Las negativas, son las que perjudican al equipo, por

ejemplo: *lost balls in own half/90*, *fouls*, *yellow cards*, *red cards*. Son menores que las positivas, es relevante hacer la distinción al momento de considerar un grupo diverso de variables, un valor más alto de la variable no necesariamente representa un rendimiento mejor.

La Tabla 2 resume las distintas posiciones que un jugador de campo puede utilizar. Cada posición corresponde a un sector del campo de juego donde el jugador principalmente desarrolla su juego, como se observa en la Figura 1.

Tabla 2. Posiciones.

Posición	Inglés	Español
LD	Left Defender	Defensa izquierdo
CD	Central Defender	Defensa central
RD	Right Defender	Defensa derecho
DM	Defensive Midfielder	Mediocampista defensivo
LM	Left Midfielder	Mediocampista izquierdo
CM	Central Midfielder	Mediocampista central
RM	Right Midfielder	Mediocampista derecho
F	Forward	Delantero

Figura 1. Posiciones del campo de juego.

Una limitación que tienen los datos es no considerar la posición de portero, debido a que participa en acciones diferentes a la de los jugadores de campo.

3.3 Carga y Limpieza de Datos

Los datos se encuentran en formato Excel, con una planilla para cada fecha, las columnas representan las variables y la filas a los jugadores. El campeonato nacional consta de 34 fechas, se utilizan 34 planillas. Estas se agrupan en un mismo dataframe, el cual es el objeto de trabajo que contiene todos

los datos. Este objeto tiene la misma estructura de la planilla, en forma de columnas y filas, pero con una columna extra con la del registro.

La primera etapa es la limpieza de datos y corresponde a la validación y corrección de datos, ajustando formatos y transformando variables. Así, esta etapa considera los siguientes pasos:

- Se renombran algunas columnas y los datos faltantes se remplazan con datos nulos.
- Se crea un diccionario con un *player_id* para cada jugador del set de datos.
- Se eliminan variables que no aportan en explicar el rendimiento de un jugador o no son factibles de trabajar mediante los modelos. Por lo tanto, se realizan las siguientes acciones:
    - Variables de porcentajes quedan como número decimal. Por ej: 86% -> 0.86.
    - Se eliminan variables no numéricas cómo Nombre, Nacionalidad, Pie Hábil, Equipo.
    - Se eliminan variables binarias que responden a si el jugador fue titular o suplente.
    - Se elimina filas que no contengan registro de *Instat Index*.
    - Se analizan qué variable les corresponde el valor 0 o valor nulo cuando no existe registro.
    - Se realiza un ajuste de ciertas variables cómo Passes, Crosses, Challenges, por cada 90 minutos, para no castigar la nota de jugadores que tuvieron menos minutos.
- Se elige la posición con mayor número de registros (CM) para trabajar y entrenar modelos.

La segunda etapa considera la imputación de datos. La idea es mantener variables que pueden ser relevantes, a pesar de que existan jugadores para los cuales no existe un valor. La imputación es realizada a variables que contengan una proporción de datos nulos menor a 30% de todo el dataframe. El resto de las variables son eliminadas, ya que el modelo pierde significancia estadística al trabajar con muchos datos estimados. La imputación se realiza mediante una regresión lineal, puesto que es una forma simple y confiable de estimar el valor de un parámetro en relación a otros. Los predictores son el resto de variables cuyos valores son conocidos. Así, las variables a imputar son las siguientes:

Tabla 3. Variables imputadas por posición.

Posición	Variables significativas
LD	Opponent's xG with a player on Height
CD	Opponent's xG with a player on Height
RD	Opponent's xG with a player on Height
DM	Opponent's xG with a player on Height Weight
LM	Height Opponent's xG with a player on
CM	Height Opponent's xG with a player on Weight
RM	Height

	Opponent's xG with a player on
F	Height Opponent's xG with a player on xG (Expected goals) xG per shot Weight

La tercera etapa considera la eliminación de variables que presentan multicolinealidad, de forma manual, esto para determinar de mejor forma los efectos de las características individuales (predictores) sobre la variable respuesta. La eliminación se realiza si la variable cumple alguno de los siguientes criterios:

- Tiene una alta correlación (>0.7) *(Rowntree, 1984)* con una cantidad considerable (>5) de variables, por ejemplo, *challenges* (7) y *succesful actions* (6). Un número alto de variables correlacionadas indica que la variable no aporta nueva información al modelo y una solución es eliminarla para estimar de mejor forma el efecto del resto de variables. *(Neter et al, 1990)*.
- Puede ser representada por otra variable estadísticamente muy correlacionada. Por ejemplo: *dribbles* y *dribbles succesful* hace referencia al mismo tipo de acontecimiento, por lo que se elimina la primera y solo se trabaja con la segunda.

3.4 Modelos utilizados

Primero, se realiza una Regresión Lineal Múltiple *(Walpole, 2012)*, que retorna las variables significativas para la variable explicada *InStat Index*, con un 95% de confianza. Para evaluar el carácter predictivo de este modelo se compara con otros modelos como Support Vector Regression, que es una variación del Support Vector Machine aplicada a la predicción *(James et al., 2021)*, y otros modelos clásicos utilizados dentro de la ciencia de datos cómo Random Forest *(James et al., 2021)* y Red Neuronal *(James et al., 2021)*. Como criterio de comparación se utiliza la Raíz del Error Cuadrático Medio (RMSE), debido a que es uno de los criterios más populares dentro de la predicción *(Kenney et al. 1962)*. Este valor se calcula para los modelos considerando, por un lado, todas las variables como predictores, y por el otro, al considerar como predictores solamente las variables significativas. Con el fin de comparar la variación en el carácter predictivo del modelo al considerar un subconjunto de variables.

3.5 Evaluación del rendimiento de jugadores

Para realizar la evaluación del rendimiento de los jugadores, primero, se extrae una muestra aleatoria de 11 jugadores *(team)* para evaluar su rendimiento. Luego, a la muestra seleccionada previamente *(team)*, se aplica la regresión lineal para cada posición, y se obtiene una estimación del InStat Index. Posteriormente, se le realiza una transformación Min-max (Han et al., 2022), que corresponde a una estandarización de datos a un pequeño intervalo específico. La transformación se realiza para dejar

las notas en escala de 1 a 10, donde 1 representa la nota de peor rendimiento, y 10 representa un rendimiento excelente, con el objetivo de obtener un rango más fácilmente interpretable para cualquier persona. Finalmente, se obtiene una evaluación del rendimiento de acuerdo con la escala transformada.

4. Resultados y discusión

En este capítulo se presentan los resultados de la evaluación del rendimiento, que consiste en la selección e interpretación de variables y el tratamiento de los datos necesarios. Además, la comparación del rendimiento de uno y más jugadores en un partido específico y durante el torneo.

4.1 Datos de la limpieza

La limpieza de datos se compone de tres etapas principales. En primer lugar, una limpieza general eliminando filas y columnas sin aporte a los modelos. En segundo lugar, una imputación de datos en variables donde existen valores faltantes. Finalmente, se eliminan variables que presentan multicolinealidad.

4.1.1 Limpieza general

En la primera etapa, el set de datos pasa de dimensiones 7548×85 a 7138×77, es decir, las filas se reducen en un 5,43% y columnas en un 9,41%.

4.1.2 Imputación de Datos

Para la selección de las variables a imputar se utiliza el porcentaje de datos faltantes por variable, que se presenta en la Figura 2.

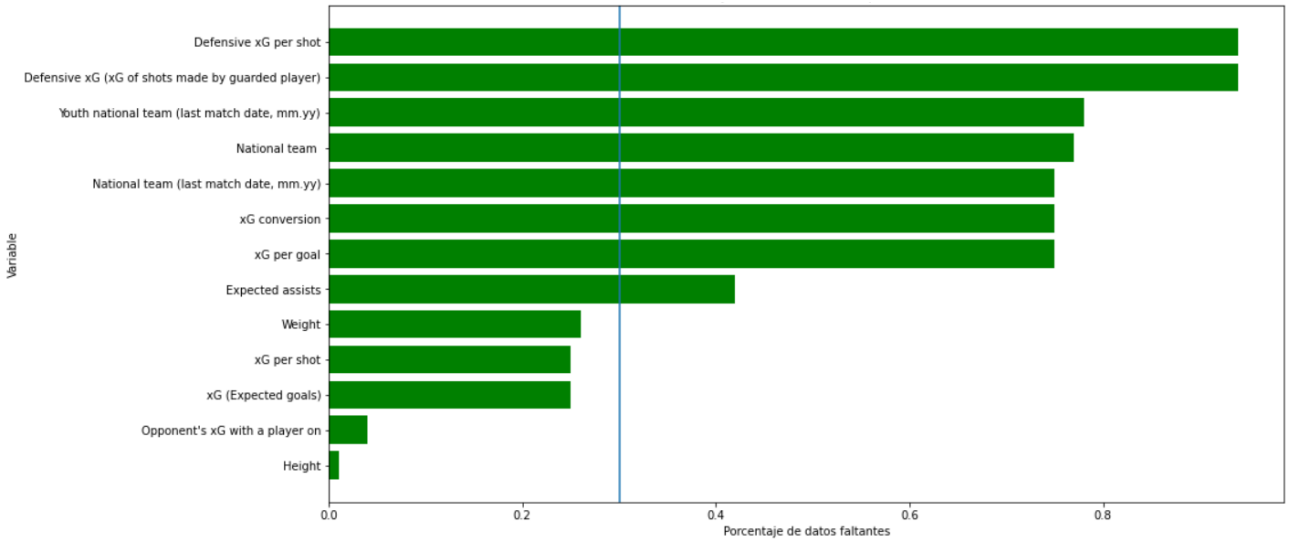

Figura 2. Porcentaje de datos nulos por variable (CM).

En el caso de la posición CM, se observa como las variables *weight*, *opponent's XG with a player on* y *height* están bajo el límite del 30% de datos nulos. Por tanto, estas variables son imputadas y el resto

de variables son eliminadas. Las variables imputadas por cada posición son similares, excepto en *forward*, donde se imputan variables del tipo XG Rates como *xG (expected goals)* y *xG per shot*.

4.1.3 Multicolinealidad

Para la segunda etapa, se analizan las variables con una mayor cantidad de variables con alta correlación, como se presenta en la Figura 3. Se observa como las variables *totals actions* (7), *challenges* (6) y *succesful actions* (6), entre otras, presentan una alta correlación con un número considerable de variables, lo que podría indicar multicolinealidad. Para ver esto en detalle se utiliza la matriz de correlación de Pearson, que mide el grado de dependencia lineal entre dos variables aleatorias cuantitativas *(Pearson, 1909)*, representada en la Figura 4. Se puede notar que la correlación se debe a que existen eventos que se incluyen en dos o más variables. Por ejemplo, el evento pase está considerado como parte de *passes* y *total actions*, o un duelo defensivo está considerado dentro de *defensive challenges* y *challenges*.

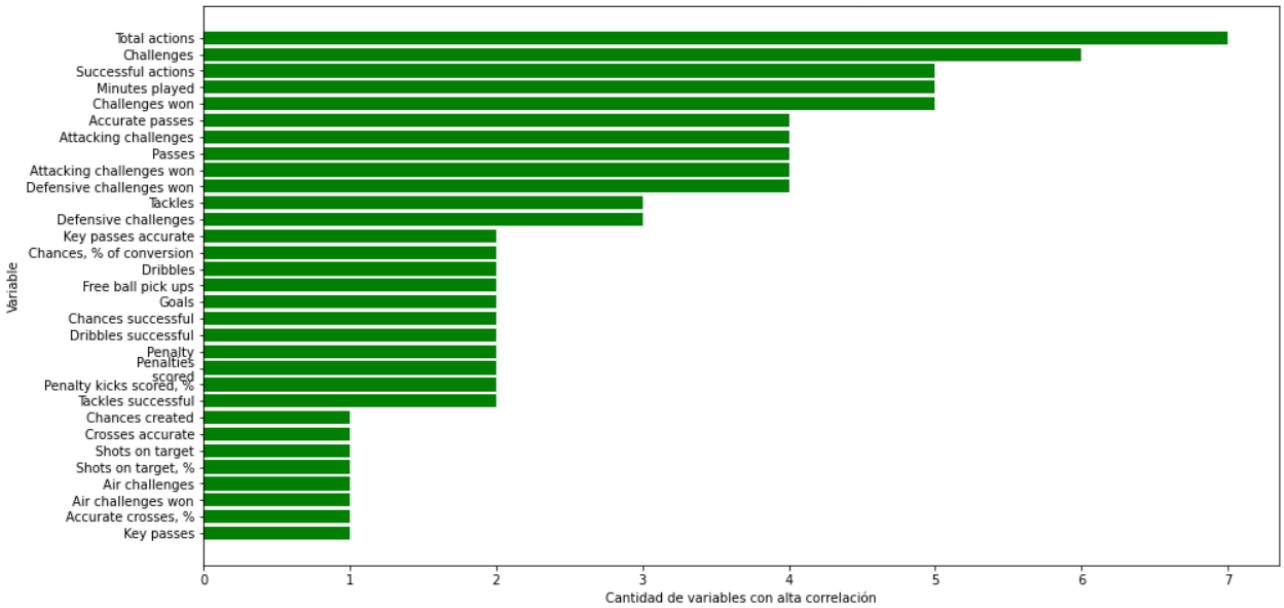

Figura 3. Cantidad de variables con alta correlación por variable.

En la Figura 4 se observa que la variable *total actions* presenta una alta correlación con todas las otras variables por lo que se elimina. *passes* y *accurate passes* tienen una correlación cercana a uno por lo que solo se trabaja con *accurate passes* y *passes* se elimina. Las variables eliminadas para cada posición, por cumplir algún criterio de multicolinealidad, se resumen en la Tabla 4. Estas variables cumplen con el criterio de multicolinealidad debido a que son variables que representan una combinación lineal de otras. Por ejemplo, *challenges* es una combinación lineal de *attacking challenges* y *defensive challenges*, por lo que al eliminar *challenges*, se puede analizar de mejor forma el efecto de las variables específicas que lo componen.

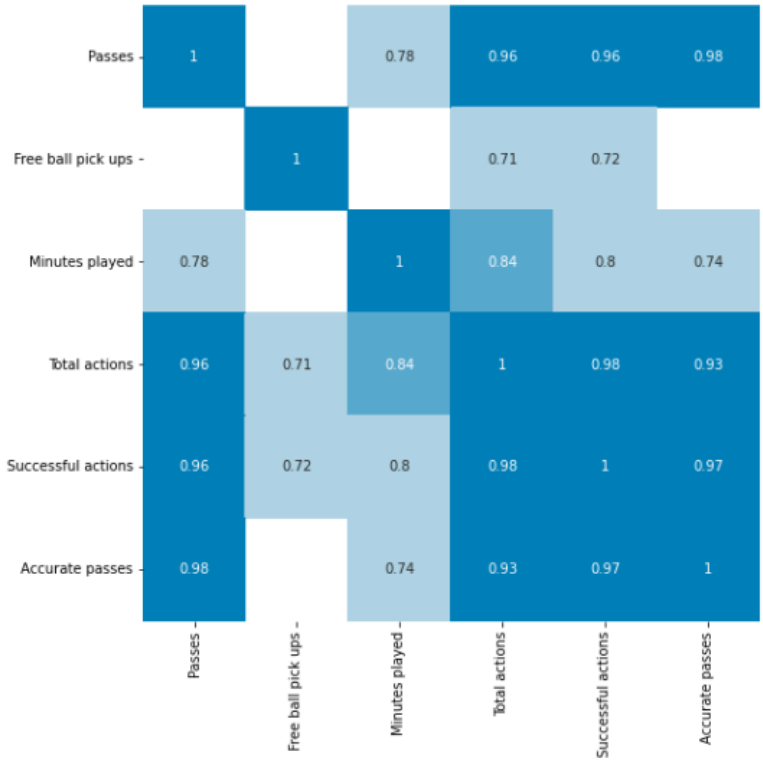

Figura 4. Matriz de correlaciones entre variables (CM).

Tabla 4. Variables eliminadas por multicolinealidad.

Posición	**Variables significativas**
LD	Challenges Succesful actions Total actions
CD	ID
RD	ID Total actions Challenges
DM	ID Total actions Challenges Succesful actions
LM	ID Total actions Challenges Succesful actions
CM	Succesful actions Challenges Challenges won Dribbles
RM	ID Total actions

	Challenges
F	ID
	Total actions
	Challenges

4.2 Resultados del entrenamiento y comparación

En el entrenamiento, se realizan dos estimaciones del *InStat Index* con distintos modelos. La primera estimación considera todas las variables como predictores, y la segunda considera solo las variables significativas, obtenidas a partir de la regresión lineal. Esta estimación es comparada con el valor real de la variable respuesta, al utilizar como criterio el RMSE. En la Figura 5 se observa que el valor del RMSE del Random Forest y Red Neuronal son los más cercanos a 0, lo que representa una precisión mejor que la Regresión Lineal. El modelo con menor precisión es el Support Vector Regresión. A pesar de que la Regresión Lineal no resulta ser el modelo más preciso, se elige por sobre los demás debido a la facilidad de interpretación de sus resultados, puesto que les otorga coeficientes a las variables predichas, lo cual es útil para comparar la influencia entre las distintas variables sobre la variable respuesta. Esta es una ventaja en comparación al resto de modelos como la Red Neuronal, la cual funciona en base a pesos y capas y se convierte en una caja negra que no permite realizar una interpretación.

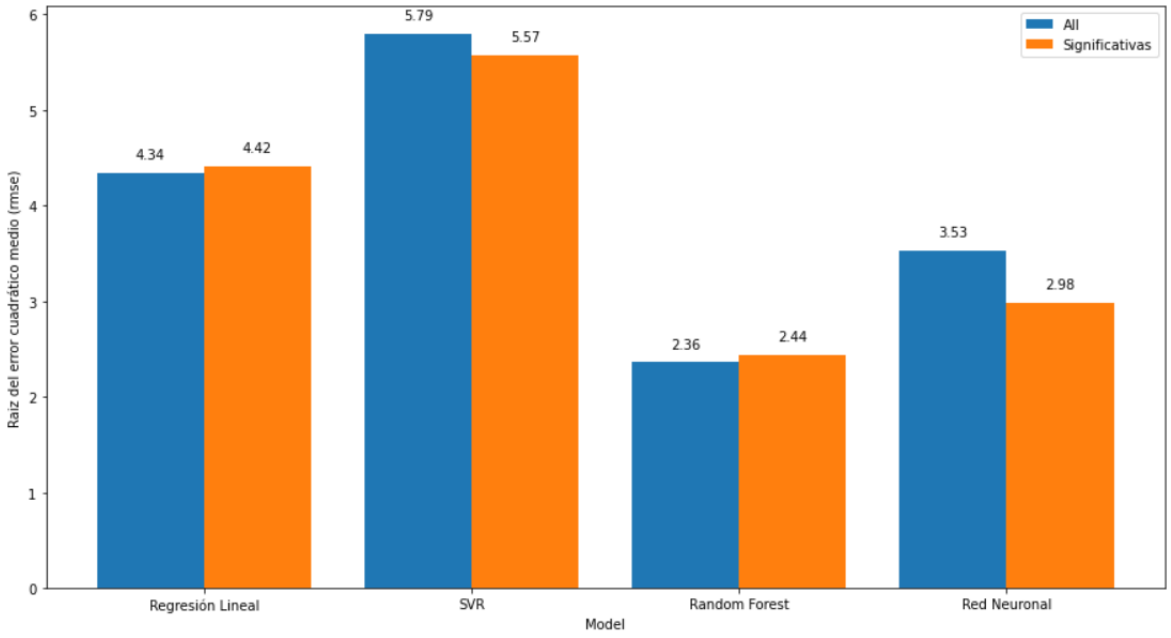

Figura 5. Errores de los modelos.

4.3 Regresión lineal

En la Tabla 5 se muestran los resultados de la regresión lineal para la posición CM, incluyendo estadísticos de importancia como el valor-*p* y el intervalo de confianza para el coeficiente de cada variable. Se observa que las variables *assists* y *chances successful* son las que tienen un mayor impacto en la nota debido al valor de su coeficiente. Otras variables como *defensive challenges* y *penalties scored* tienen un impacto negativo en la nota debido al signo de su coeficiente. Las variables significativas para el resto de posiciones, ordenadas de acuerdo a su coeficiente, se resumen en la Tabla 6. Así, en la Tabla 6 se puede observar que, si bien existen variables que son significativas en explicar el rendimiento para varias posiciones, como *goals*, *assists* y *passes*, existen otras que son específicas para ciertas posiciones cómo *penalty kicks scored, %* (*Forward*) y *air challenges won* (*Right Midfielder* y *Forward*).

Tabla 5. Regresión Lineal para posición CM.

	coef	**std err**	**t**	**P>|t|**	**[0.025**	**0.975]**
const	184,72	0,85	218,41	0,00	183,06	186,38
Ball interceptions	1,28	0,25	5,05	0,00	0,78	1,77
Assists	17,60	1,53	11,50	0,00	14,60	20,60
Chances successful	20,87	1,64	12,71	0,00	17,65	24,10
Shots on target	8,19	1,22	6,69	0,00	5,79	10,59
Shots on target, %	-6,92	2,17	-3,19	0,00	-11,17	-2,67
Penalties scored	-9.66	4,30	-2,24	0,03	-18,10	-1,21
Defensive challenges	-1,62	0,23	-7,13	0,00	-2,06	-1,17
Defensive challenges won	4,48	0,40	11,27	0,00	3,70	5,26
Attacking challenges won	0,95	0,28	3,43	0,00	0,41	1,50
Dribbles successful	4,65	0,41	11,29	0,00	3,84	5,46

Tabla 6. Variables significativas por posición.

Posición	**LD**	**CD**	**RD**	**DM**	**LM**	**CM**	**RM**	**F**
Variables significativas	Chances successful (22,40) Key passes accurate (16,33) Defensive Challenges won (6,55) Defensive challenges (-3,58) Crosses accurate (2,53) Accurate passes (1,69) Passes (-1,11)	Goals (21,86) Red cards (-14,15) Chances, % of conversión (11,83) Assists (10,22) Chances created (9,84) Tackles won, % (6,59) Defensive Challenges won (5,98) Air Challenges won, % (5,71) Defensive Challenges (-4,02) Challenges won (1,20) Tackles (-0,83)	Assists (17,68) Key passes accurate (11,78) Succesful dribbles, % (10,47) Shots (6,78) Defensive Challenges won (5,33) Accurate crosses, % (4,88) Defensive Challenges (-3,04) Accurate passes (1,63) Passes (-1,08) Tackles	Goals (26,74) Assists (10,31) Chances created (5,81) Key passes (4,76) Challenges won (4,21) Succesful dribbles, % (3,29) Shots (2,59) Defensive Challenges won (2,50) Challenges (-2,21) Accurate passes (2,03) Passes	Red cards (-20,22) Goals (16,76) Chances, % of conversión (12,92) Assists (11,20) Successful dribbles, % (8,29) Key passes (3,97) Chances created (3,13) Crosses accurate (3,12) Shots on target (2,51)	Chances successful (20,87) Assists (17,60) Penalties Scored (-9,66) Shots on target (8,19) Shots on target, % (-6,92) Dribbles successful (4,65) Defensive challenges won (4,48) Defensive challenges (-1,62)	Goals (26,35) Red cards (-22,24) Assists (11,23) Succesful dribbles, % (8,09) Chances created (4,40) Key passes (4,00) Shots (2,89) Challenges won (1,97) Air challenges won (-1,70) Accurate passes (1,29)	Penalties scored (29,12) Penalty kicks scored, % (-26,63) Assists (13,42) Shots on target, % (12,87) Goals (12,44) Chances, % of conversión (11,30) Penalty (-10,74) Red cards (-10,05)

	Ball interceptions (0,53)	(-0,77)	(-1,65) Ball interceptions (1,44) Defensive Challenges (-0,95) Free ball pick ups (0,42) Height (-0,29)	Dribbles successful (2,46) Shots (2,14) Chances (-1,80) Accurate passes (1,46) Challenges won (1,17) Passes (-1,06)	Ball interceptions (1,28) Attacking challenges won (0,95)	Dribbles (1,13) Challenges (-1,09) Attacking challenges won (1,08) Defensive challenges won (0,89) Passes (-0,89) Ball interceptions (0,73) Height (-0,25)	Chances successful (9,33) Key passes accurate (8,26) Shots on post / bar (5,72) Shots (3,79) Key passes (2,52) Challenges won (1,70) Tackles (1,26) Lost balls in own half (0,97) Attacking Air Challenges won (-0,81) Weight (-0,10)

4.4 Transformación de escala

Esta etapa consiste en la transformación de escala y el objetivo es escalar los datos en un pequeño intervalo específico para interpretar, en este caso entre 1 y 10. Para esto, es necesario conocer el rango de valores entre el cual oscila la variable *InStat Index*. La distribución entre los valores mínimos y máximos para la variable, en cada posición a lo largo de las 34 fechas, se representa en la Figura 6. Al observar la Figura 6 y comparar las posiciones F y RD, la diferencia entre el promedio de notas mínimas y máximas es muy superior en los delanteros que en los laterales derechos. Dado que la posición de delantero es relevante en el resultado de un partido. Por lo tanto, la evaluación se escala de acuerdo con el promedio entre todas las posiciones para un único valor mínimo y máximo. Es de interés analizar los valores intermedios idealmente con variabilidad para identificar patrones desconocidos.

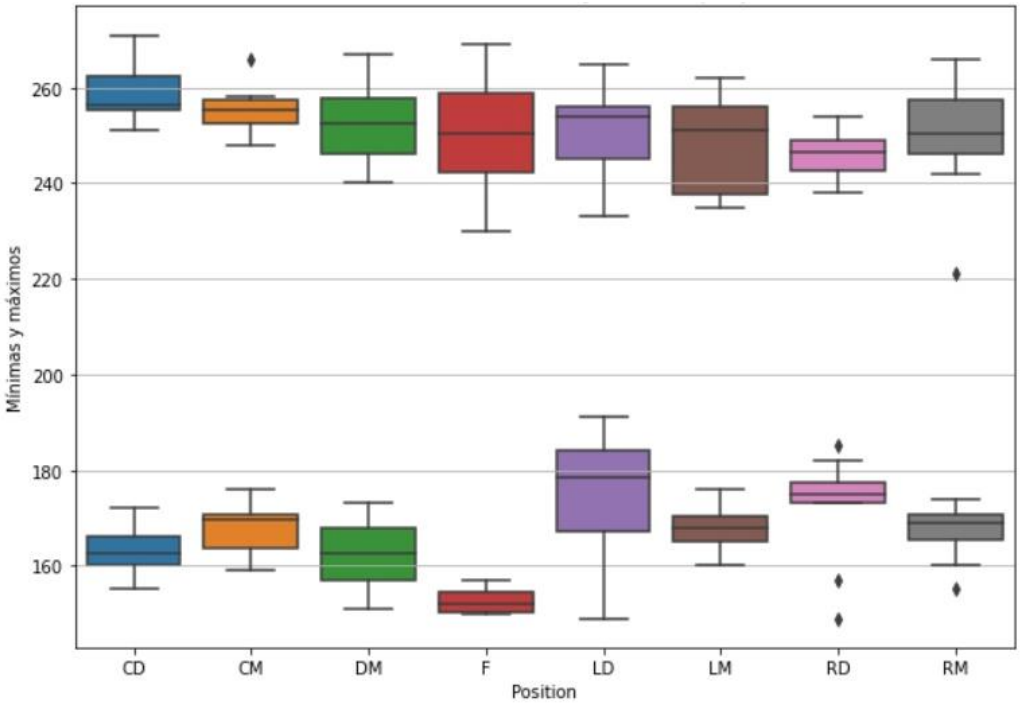

Figura 6. Distribución notas mínimas y máximas por posición.

La transformación es representada mediante la fórmula de la Ecuación (1).

$$Nota = \frac{\widehat{InStat\ Index} - \bar{x}_{mínimo}}{\bar{x}_{máximo} - \bar{x}_{mínimo}} * 9 + 1 \quad (1)$$

Donde:

$\bar{x}_{máximo} = \frac{\sum_{i=1}^{n} \bar{x}_{máximo,i}}{n}$

$\bar{x}_{mínimo} = \frac{\sum_{i=1}^{n} \bar{x}_{máximo,i}}{n}$

$n =$ Cantidad de posiciones.

4.5 Resultados de la evaluación del rendimiento

A continuación, se obtiene una evaluación para cada jugador de *team* a partir de las variables significativas para cada posición luego de realizar la transformación de escala. Ver Tabla 7, la primera y segunda columna son el ID y nombre el jugador. La tercera columna es la fecha en la cual el jugador es evaluado y la cuarta columna el equipo al cual el jugador pertenece. La quinta columna es la posición del jugador, mientras que la sexta y séptima columna representan el valor del *InStat Index* y la nota del jugador, ordenadas de mejor a peor.

Tabla 7. Team.

player_id	*nombre*	*fecha*	*equipo*	*position*	*InStat Index*	*nota*
62	G. Suazo	28	Colo Colo	LD	272	9,75
158	N. Guerra	3	U. de Chile	LM	249	7,62
122	J. C. Gaete	20	Cobresal	LM	234	5,91
353	J. Villagra	13	U. Española	CD	220	4,86
230	J. Barroilhet	26	Curicó Unido	CD	182	4,59
91	A. Vilches	10	U. La Calera	F	196	4,48
423	I. Fernández	18	Antofagasta	LD	198	4,02
356	B. Vejar	31	Palestino	RD	186	3,88
86	W. Gama	10	Santiago Wanderers	RM	189	3,64
339	V. Pizarro	31	Colo Colo	DM	203	3,48
136	J. Flores	6	Antofagasta	CD	171	3,04

En la Figura 7 se compara el rendimiento de dos jugadores, J. Barroilhet (Curicó Unido) y J. Villagran (U. Española), a partir de las variables significativas de acuerdo a su posición (CD). Si bien J. Villagra fue superior en la mayoría de las categorías, esto no se refleja de forma tan clara en su nota (4,86 vs 4,59). Esto podría significar que las variables en las que fue superior no tienen relevancia en la nota como *Tackles won, %*, donde fue superior Jordan Barroilhet. En la Figura 8 compara el rendimiento de todos los jugadores de *team* y se observa que el mejor rendimiento fue de G. Suazo (9.54) y N. Guerra (9.52). Mientras que el peor rendimiento lo tuvo J. Barroihet (2.42) y J. Flores (1.35).

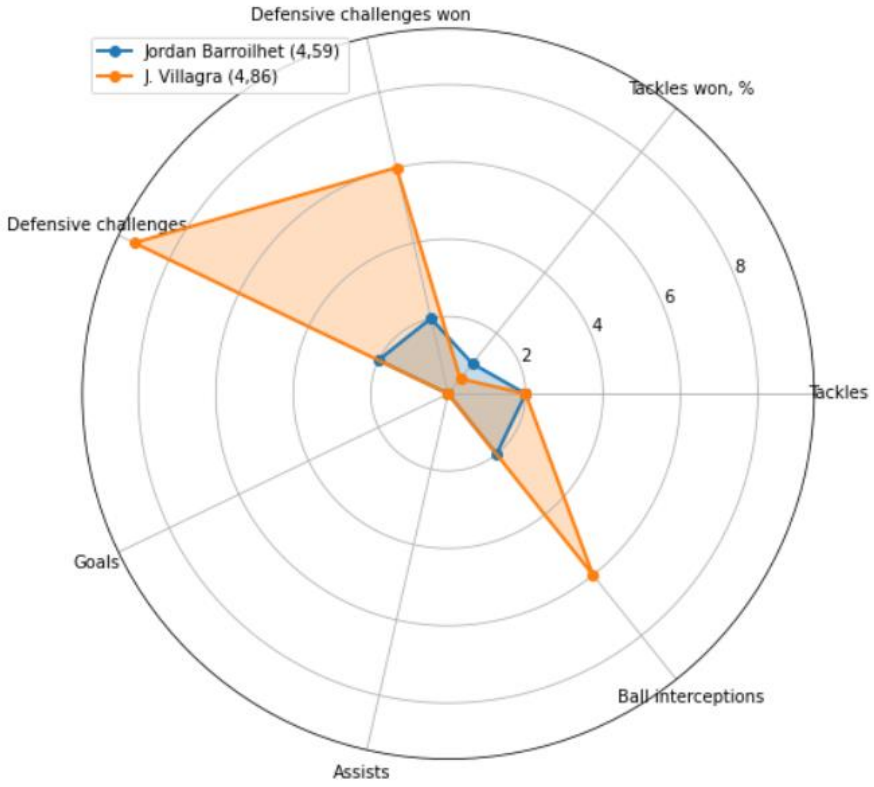

Figura 7. Comparación jugadores.

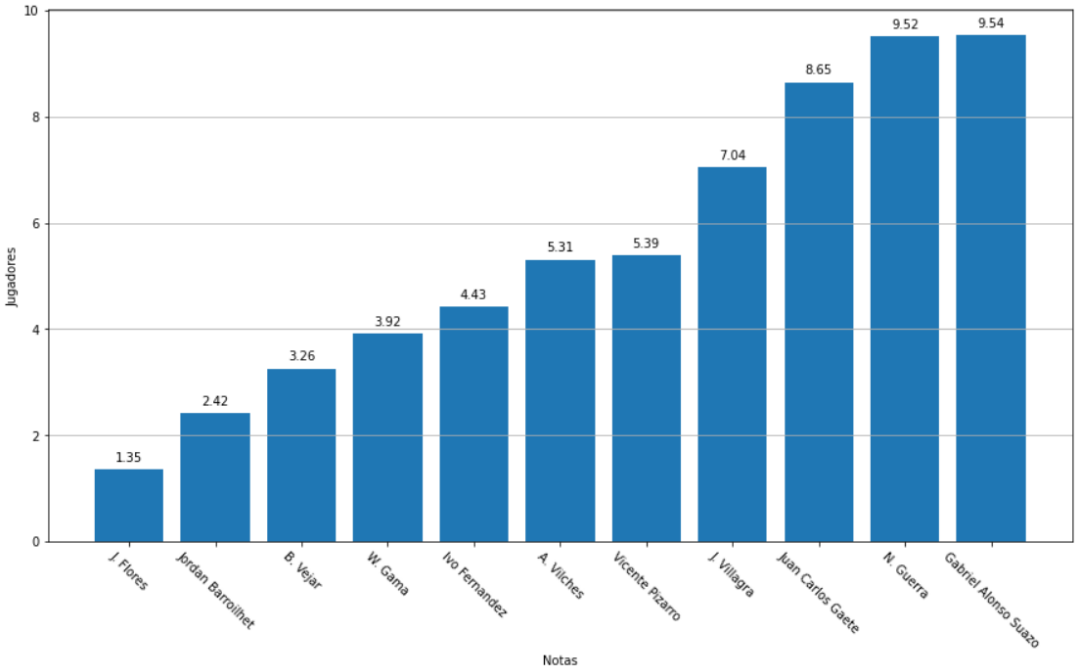

Figura 8. Rendimiento *team*.

En la Figura 9 se representa la distribución de las notas para las distintas fechas para los jugadores J. Villagra y V. Espinoza. Es importante notar que no existe un valor para todas las fechas, puesto que los jugadores no necesariamente juegan todos los partidos del campeonato. Esta separación es evidente en V. Espinoza, quién tiene nota en la fecha 1 y luego no vuelve a tener nota hasta la fecha 15, esto indica que estuvo ausente durante varios partidos. Eso podría ser signo de una lesión, lo cual también es importante al evaluar el rendimiento general durante un campeonato.

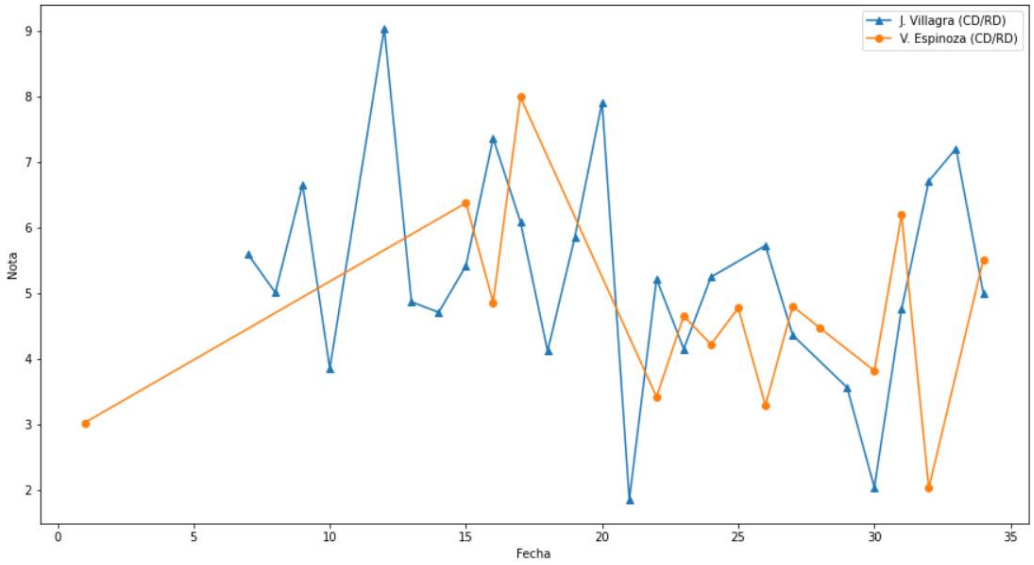

Figura 9. Distribución de notas por fecha.

Si bien la nota se basa a partir del InStat Index, estas tienen dos principales diferencias. Por un lado, la nota se calcula a partir de un subconjunto de variables que podría ser modificado si el evaluador así lo estima conveniente, lo cual le entrega flexibilidad al modelo. Por otro lado, las notas están en una escala donde los valores intermedios tienen una interpretación más simple, donde las notas entre 6-8 se podrían calificar como un rendimiento bueno, pero no excelente, y las notas entre 2-4 como un rendimiento malo, pero no horrible.

4.6 Discusión general

La discusión finalmente se centra en si las variables escogidas por el modelo son realmente relevantes en explicar el rendimiento del jugador. Para esto se requiere una interpretación humana con conocimiento futbolístico capaz de discernir la importancia de las variables de acuerdo a la posición del campo. Por ejemplo, evaluar los defensas por la cantidad de goles que hacen sería contraintuitivo, por lo que igualmente se requiere el juicio humano capaz de aplicar el criterio para la selección final de variables.

La ventaja de este modelo es que se puede adaptar a otras variables del juego que sean consideradas de importancia y no hayan sido capturadas por el modelo. Un jugador se podría evaluar de acuerdo a distintas variables a lo largo de un torneo, si, por ejemplo, juega en dos posiciones distintas de forma regular.

Además, es importante distinguir que existen diversas variables que influyen de forma negativa en el desempeño del jugador, como por ejemplo los balones perdidos en propio campo. Así, la interpretación de las variables más allá del valor numérico, permite una tolerancia a los errores variables. Debido a que, por ejemplo, un técnico puede considerar que perder el balón el propio campo que termina en gol del rival es un error fatal. Mientras que otro puede considerar que no es tan relevante si durante el resto del partido mantiene un buen nivel.

La utilidad de la herramienta depende fuertemente de la disponibilidad de datos con la que se cuente, puesto que, si solamente se trabaja con datos internos de un propio club, se estará limitando el análisis de la evaluación a este club, mas no a sus rivales. En cambio, si se dispone de una de base de datos amplia para trabajar, como los utilizados por las empresas proveedoras de datos, se puede analizar, por ejemplo, el rendimiento del rival antes de cada partido y así preparar de mejor forma la táctica a utilizar, o realizar seguimiento a jugadores de otras ligas con lo que existe un cierto interés de contratación al finalizar la temporada.

Así, el valor que se puede obtener de esta herramienta también depende de las necesidades que posea el equipo, sus puntos fuertes y puntos débiles, el modelo de juego, la planificación del equipo, etc, todo esto en tiempo real pues la necesidades y ajustes tácticos pueden ir cambiando semana a semana. Es importante mencionar que el proceso de scouting lo pueden realizar diferentes personas, ya sea el scouter quién tiene connotaciones orientada a captación de talento, por lo tanto, está más inserto en la secretaría técnica de los equipos; o bien por un analista, quién está encargado de estudiar el fútbol en cada una de sus fases, desglosándolo e identificando de manera pormenorizada las características y patrones en los equipos (Pérez, 2019), por lo que está más relacionado al cuerpo técnico del equipo. Aquí se puede diferenciar el analista táctico, técnico, de rendimiento físico y de datos. Es este último quién le podría sacar mayor valor a esta herramienta.

En cuanto a los modelos utilizados, es necesario que estos puedan tener una interpretación más allá del valor numérico. El Random Forest, por ejemplo, trabaja en base la importancia de la característica (feature importance) que se calcula como la disminución de la impureza del nodo ponderada por la probabilidad de alcanzar ese nodo *(Breiman, 2001)*. Cuanto mayor sea el valor, más importante será la característica. A partir de estos valores se podría analizar los distintos pesos de cada variable sobre el valor final de la regresión (rendimiento).

5. Conclusiones

El presente capítulo presenta las conclusiones extraídas a partir de las comparaciones sobre el desempeño de los distintos métodos para predecir y la interpretación futbolística de los resultados.

5.1 Conclusiones generales

El trabajo realizado presenta una escala de evaluación de rendimiento para jugadores de fútbol que se realiza a partir del torneo chileno de primera división durante el año 2021. El modelo propuesto, a partir de la Regresión Lineal, presenta un buen desempeño comparado con otros modelos como la Red Neuronal. La interpretación de la regresión lineal es un aporte para entender la valorización de un jugador en términos de su rendimiento deportivo. Es imprescindible que el modelo a utilizar posea la flexibilidad de adaptarse a diversas visiones estratégicas y tácticas de juego.

En cuanto a la interpretación de las variables seleccionadas, existen algunas relevantes independiente de la posición del jugador. Son las típicas dentro del fútbol, como *goals* y *assists*, y no representan un hallazgo interesante en el marco de este trabajo. En cambio, existen otras variables relevantes para un buen rendimiento dependiendo de la posición, y representan un hallazgo relevante dado que permiten establecer criterios de evaluación específicos, como *ball interceptions* (DM y CM) y *defensive challenges won* (CD y RD).

La ventaja que presenta este modelo es que permite comparar jugadores dentro de un mismo club, y con jugadores de otros clubes incluso en distintas ligas en base a una escala pareja para todos, lo cual cumple con el objetivo de minimizar el sesgo y juicio personal. Esto representa un sustento técnico para la toma de decisión sobre qué jugadores contratar o despedir, y ser capaces de poder identificar talento antes que el resto.

5.2 Recomendaciones

Una posible mejora a la metodología es diferenciar los jugadores que juegan de titular con los que juegan de suplente, puesto que los primeros, por lo general, juegan todos los minutos de un partido (90 minutos). Mientras que los segundos, por lo general, ingresan en el segundo tiempo (20-30 minutos). Otra recomendación es agregar un ponderador que tome en cuenta la dificultad del rival.

Se recomienda interpretar los resultados más allá del valor numérico, puesto que los modelos no pueden capturar todas las características relevantes en el fútbol. Para esto es necesaria la interpretación humana global de los resultados. Eso facilita que sea entendido desde un punto de vista táctico y estratégico, lo cual es de real aporte a un equipo de fútbol, más allá de la estadística y matemática.

5.3 Trabajos futuros

Una mejora sugerida al modelo podría ser diferenciar a los jugadores que juegan de titular con los que entran de suplente, tomando en cuanto factores como, que en el segundo tiempo los jugadores están más agotados física y mentalmente, o las jugadas pueden terminar siendo más decisivas en el resultado del partido. Otra mejora podría incorporar a los porteros, estableciendo sus propias variables relevantes para evaluar su rendimiento y poder compararlas por el resto de jugadores de campo.

En cuanto a los modelos, se recomienda profundizar en la interpretación de la predicción, por ejemplo, en el Random Forest se podría realizar el análisis en base a la importancia de la característica (feature importance). Además, se podría utilizar algunos modelos que sean capaces de mejorar la precisión de la predicción de la variable predicha, como podría ser un algoritmo de árboles de decisión *(Koning et al., 2019)* o de Redes neuronales profundas *(Weidman, 2019)*. Se sugiere probar también modelos de aprendizaje no supervisado, es decir, que no tienen una variable respuesta. Uno de los más populares es la clusterización *(James et al., 2021)*, para agrupar, por ejemplo, jugadores con características similares (edad, posición, altura, peso, pie hábil) y realizar el análisis por separado.

7. Referencias

Breiman, Leo (2001). “Random Forests.” Machine Learning 45 (1). Springer: 5-32.

Fernandez, G. (2016). Evaluación del rendimiento deportivo: medir para mejorar. Clínica Alemana. Recuperado de www.clinicaalemana.cl

Galaz, P. (2020). Datazul: Un primer caso de Analytics aplicado al fútbol profesional en Chile, Tesis de Magíster, Universidad de Chile.

García-López, L. M., González-Víllora, S., Gutiérrez, D., & Serra, J. (2013). Development and validation of the Game Performance Evaluation Tool (GPET) in soccer. *Revista Euroamericana de Ciencas del Deporte*. 2(1), 89-99.

GLOBAL TRANSFER REPORT 2021. (2022). *FIFA.*

González-Neira, M., San Mauro-Martin, I.,García-Angulo, B., Fajardo, D. & Garicano-Villar, E. (2014). Valoración nutricional, evaluación de la composición corporal y su relación con el rendimiento deportivo en un equipo de fútbol femenino. *Revista Española de Nutrición Humana y Dietética*. 19(1). 36-48

Han, J., Pei, J., & Tong, H. (2022). Data Mining: Concepts and Techniques (4th ed.). *Morgan Kaufmann Publishers*.

James, G., Witten, D., Hastie, T., & Tibshirani, R. (2021). An Introduction to Statistical Learning: With Applications in R (2nd 2021 ed.). *Springer*.

Kenney, J.F. and Keeping, E.S (1962). Mathematics of Statistics, Pt. 1, 3rd ed. Princeton, *NJ: Van Nostrand*, 59-60.

Koning, M., & Smith, C. (2017a). Decision Trees and Random Forests. *Van Duuren Media*.

Marín, G. (2016). Juegos de Invasión. *Publicaciones Didácticas*, *66*, 117–125.

Mena, S. (2021). Impacto de los futbolistas de la “Premier League 2017-2018” en la probabilidad de salir campeón a través de un método de simulación con Inferencia Bayesiana, Tesis de Magíster, Universidad de Chile.

Molina, J., Chorot, P., Valiente, R. M., & Sandín, B. (2014). Miedo a la evaluación negativa, autoestima y presión psicológica: Efectos sobre el rendimiento deportivo en adolescentes. *Cuadernos de Psicología del Deporte*, 14(3), 57–66.

Neter, J., Wasserman, W. y M. H. Kutner (1990), Applied Linear Statistical Models, 3a edn., M.A: Irwin.

Pearson, K. (1909). Determination of the Coefficient of Correlation. *Science*, 30(757), 23–25.

Pérez, D. (2019). Scouting, Scout, Analista, Ojeador. . . *Objetivo Analista.*

Primera División de Chile - Fichajes 2022. (s. f.-b). *Transfermarkt.* https://www.transfermarkt.es/primera-division-de-chile/transfers/wettbewerb/CLPD/plus

Rodríguez, G. (2013). Programa de evaluación del rendimiento deportivo. *EFDeportes.com, Revista Digital*. Buenos Aires, Año 17, Nº 178.

Rowntree, Derek. (1984). *Introducción a la estadística: un enfoque no matemático.* Bogotá: Norma

Seirul-lo, F. (2009). Una línea de trabajo distinta. *Revista de Entrenamiento Deportivo*, 23(4): 13-18.

Ursino, D., Abal, F., Cirami, L. & Barrios, R. (2019). La evaluación del rendimiento deportivo en psicología del deporte: Una revisión sistemática. Anuario de Investigaciones, Universidad de Buenos Aires, vol. XXVI, pp. 413-425

Walpole, R. E. (2012). Probabilidad y estadística para ingeniería y ciencias (9.a ed.). *Pearson Education*. 443-506.

Weidman, S. (2019). Deep Learning from Scratch: Building with Python from First Principles. *O'Reilly Media*.

8. Anexos

Tabla 8. Traducción/Explicación por variable.

Variable	Traducción/Explicación
InStat Index	Índice InStat
Matches played	Partidos jugados
Position	Posición
Minutes played	Minutos jugados
Starting lineup appearances	Aparición en el equipo titular
Substitute out	Salida por sustitución
Substitutes in	Entrada por sustitución
Goals	Goles
Assits	Asistencias
Chances	Ocasiones
Chances successful	Ocasiones exitosas
Chaces, % of conversión	Ocasiones exitosas/Ocasiones
Chances created	Ocasiones creadas
Fouls	Faltas cometidas
Fouls suffered	Faltas recibidas
Offsides	Fuera de juego
Yellow cards	Tarjetas amarillas
Red cards	Tarjetas rojas
Total actions	Acciones totales
Succesful actions	Acciones exitosas
Successful actions, %	Acciones exitosas/Acciones totales
Shots	Tiros
Shots on target, %	Porcentaje de tiros a puerta
Shots wide	Tiros fuera
Blocked shot	Tiros bloqueados
Shots on post/bar	Tiros al poste
Penalty	Penales
Penalty scored	Penales anotados
Passes	Pases
Accurate passes	Pases precisos
Accurate passes, %	Pases precisos/Pases
Key passes	Pases clave
Key passes accurate	Pases clave precisos
Crosses	Cruces
Crosses accurate	Cruces precisos
Accurate crosses, %	Cruces precisos/Cruces
Lost balls	Balones perdidos
Lost balls in own half	Balones perdidos en propia mitad
Ball recoveries	Balones recuperados
Ball recoveries in opponent’s half	Balones recuperados en la mitad del oponente
xG (Expected goals)	Goles esperados
Expected assists	Asistencias esperadas
xG per shot	Goles esperados por tiro
xG per goal	Goles esperados por gol
xG conversión	Goles esperados por conversión
xG with a player on	Goles esperados con un jugador encima
Opponent’s xG with a player on	Goles esperados del rival con un jugador encima
Net xG (xG player on - opp. team's xG)	Goles esperados netos
Defensive xG (xG of shots made)	Goles esperados defensivos
Defensive xG per shot	Goles esperados por tiro defensivo
Challenges	Duelos
Challenges won	Duelos ganados
Challenges won, %	Duelos ganados/Duelos
Defensive challenges	Duelos defensivos
Defensive challenges won	Duelos defensivos ganados

Challenges in defence won	Duelos defensivos ganados/Duelos defensivos
Attacking challenges	Duelos ofensivos
Attacking challenges won	Duelos ofensivos ganados
Challenges in attack won, %	Duelos ofensivos ganados/Duelos ofensivos
Air challenges	Duelos aéreos
Air challenges won	Duelos aéreos ganados
Air challenges won, %	Duelos aéreos ganados/Duelos aéreos
Dribbles	Regates
Dribbles successful	Regates exitosos
Succesful dribbles, %	Regates exitosos/Regates
Tackles	Entradas
Tackles successful	Entradas exitosas
Tackles won, %	Entradas exitosas/Entradas
Ball interceptions	Intercepciones de balón
Free ball pick ups	Recogida de balones libres
Nationality	Nacionalidad
Team	Equipo
National team	Equipo nacional
Age	Edad
Height	Altura
Weight	Peso
Foot	Pie hábil
National team (last match date)	Fecha del último partido por el equipo nacional
Youth national team (last match date)	Fecha del último partido por el equipo nacional jóven

UNIVERSIDAD DE CONCEPCION – FACULTAD DE INGENIERIA
RESUMEN DE MEMORIA DE TITULO

Departamento de Ingeniería			
Título		MODELO DE MACHINE LEARNING PARA EVALUACIÓN DEL RENDIMIENTO DE JUGADORES DE FÚTBOL	
Nombre Memorista		CRISTÓBAL ORLANDO BURDILES GUTIÉRREZ	
Modalidad	PRESENCIAL	Profesor(es) Patrocinante	
Concepto			
Calificación			
Fecha	16/12/2022	Ingeniero Supervisor	Institución
Comisión (Nombre y Firma)			
LORENA PRADENAS			
Resumen			

La evaluación del desempeño de forma objetiva no es una tarea fácil, puesto que generalmente se realiza por personas, las cuales tienen sus propios sesgos y juicios personales. Este trabajo presenta una herramienta para evaluar el rendimiento deportivo de futbolistas mediante modelos de machine learning basados en datos estadísticos de la liga chilena de fútbol profesional. Se presenta una escala de evaluación que asigna una nota a cada jugador para cada partido disputado. La nota se calcula a partir de un modelo de regresión lineal donde los predictores varían de acuerdo con la posición del campo de juego donde se desempeña el jugador. La regresión se escoge como modelo para estimar luego de comparar su carácter predictivo, mediante la raíz del error cuadrático medio (RMSE), con otros modelos como Red Neuronal (2,98), Support Vector Machine (5,57) y Random Forest (2,44). Si bien su estimación no es precisa (4,42), se escoge debido a su interpretabilidad y amplio uso en la literatura. El modelo permite comparar el rendimiento entre distintos jugadores, tanto para un partido en específico como a lo largo del torneo. Como resultado se obtienen las variables que son significativas, para cada posición y explicar el rendimiento del jugador. Se concluye que existen variables relevantes para evaluar independiente de la posición del campo de juego, mientras que existen otras que solo son relevantes para una posición en específico.