Predicción del potencial en el Fútbol

Universidad de O'Higgins

felipe.cabezas@pregrado.uoh.cl 1st Felipe Cabezas *Instituto de Ciencias de la Ingeniería*
Ingeniería Civil Industrial
Rancagua, Chile
nicolas.herrera@pregrado.uoh.cl 2nd Nicolás Herrera *Instituto de Ciencias de la Ingeniería*
Ingeniería Civil Industrial
Rancagua, Chile
alvaro.lara@pregrado.uoh.cl 3rd Alvaro Lara *Instituto de Ciencias de la Ingeniería*
Ingeniería Civil Industrial
Rancagua, Chile

Abstract*—El fútbol es el deporte más popular y económicamente significativo en el mundo, afectando tanto a la sociedad como a la economía de los países. Con la disponibilidad actual de grandes cantidades de datos, se presenta una oportunidad para utilizar herramientas de minería de datos para evaluar el rendimiento y potencial de los jugadores. Este análisis es valioso para los clubes, ya que permite estimar el valor económico de los jugadores basándose en atributos personales y futbolísticos. En este estudio, se utilizó un conjunto de datos de aproximadamente 17,000 jugadores extraídos de SoFIFA.com para la temporada 2018/2019. Este dataset incluye una variedad de atributos como edad, estatura, peso, posición, habilidades y características financieras. El objetivo fue clasificar a los jugadores en tres categorías de potencial: bajo, medio y alto, utilizando tanto algoritmos de aprendizaje supervisado (Árbol de decisión, Support Vector Machine, Redes Neuronales) como un algoritmo de aprendizaje no supervisado (K-means). Para determinar el número óptimo de clusters en el algoritmo K-means, se utilizó el método del codo, el cual indicó que tres clusters eran adecuados. Las métricas de evaluación para los modelos supervisados incluyeron accuracy, precision, recall y F1 score. Los resultados mostraron que el modelo de Árbol de decisión tuvo la mayor precisión (0.961571), seguido por Redes Neuronales (0.955444) y SVM (0.947925). En base a las métricas se utilizó el algoritmo de aprendizaje supervisado Árbol de Decisión en la implementación. Se clasificaron 477 jugadores como de potencial "bajo", 16,066 como de potencial "medio" y 1,411 como de potencial "alto". Esta información es crucial en la industria del fútbol para realizar scouting masivo e identificar jugadores rentables en términos de rendimiento y valor económico.

***Index Terms*—Fútbol, potencial, aprendizaje, datos, algoritmo.**

I. Introducción

El fútbol es el deporte más popular del mundo [1] y uno de los deportes que más dinero mueve, teniendo un gran impacto no solo en la sociedad sino que también en la economía de los países. Con la disponibilidad de datos que se tiene hoy en día, resulta una gran oportunidad el usar herramientas de minería de datos que nos permita determinar tanto el rendimiento como el potencial de cada jugador, siendo esto último una oportunidad para los clubes de determinar qué tanto rédito económico se le puede sacar a un jugador de fútbol. Ejemplo de esto, Kylian Mbappé, que a su corta edad y gracias a su gran potencial es poseedor de uno de los salarios más grandes en la historia del fútbol [2]. Esto se puede lograr en base a sus atributos tanto personales (edad, estatura, peso, etc.) como netamente futbolísticos (pase, definición, recuperación de balones, etc.). Por lo mismo, trabajaremos la base de datos de SoFIFA.com [3] que corresponde a la base de datos más grande del mundo sobre atributos de jugadores de fútbol en todo el mundo. Esta incluye atributos tanto físicos como futbolísticos de un aproximado de 17000 jugadores de fútbol.

Dado el avance de la tecnología en los tiempos actuales, el análisis de "big data" en los servicios de scouting y la formación de jugadores es algo cada vez más frecuente [4]. A veces se prefieren métodos más tradicionales, como el simple observamiento de jugadores mientras juegan o realizan trabajos físicos bajo ojos de personas expertas, pero es innegable que de esta forma no se puede analizar de forma rápida y robusta el rendimiento de cada jugador, mucho menos de la forma que se puede llegar a conseguir con métodos de aprendizaje automático, donde se analizan una gran cantidad de atributos de miles de jugadores de fútbol alrededor de todo el mundo. Por ello, haciendo uso de 4 algoritmos (Árbol de Decisión, SVM, Redes Neuronales y K-means) buscamos clasificar a los jugadores dentro de un rango de potencial basando en sus atributos personales. Información que es considerada útil para la toma de decisiones tanto en el ámbito económico como en el formativo de un club.

Instituto de Ciencias de la Ingeniería: Universidad de O'Higgins

II. Materiales y Métodos

A. Descripción de los datos

Este conjunto de datos ofrece a detalle las características de un aproximado de 17000 jugadores en todo el mundo, extraídos de SoFIFA.com correspondiente a la temporada 2018/2019. Abarca una gran cantidad de atributos por cada jugador, incluyendo sus nombres, nacionalidades, edad, peso, estatura, etc. Además de otras características relacionadas a su juego como posición, definición, pase, velocidad, poder de disparo, etc. Algunos de los atributos del dataset son los siguientes:

- Edad del jugador.
- Altura del jugador en centímetros.
- Peso del jugador en kilogramos.
- Posiciones en las que puede jugar el jugador.
- Nacionalidad del jugador.
- Calificación general del jugador en FIFA.
- Valoración potencial del jugador en FIFA.
- Valor de mercado del jugador en euros.
- Salario semanal del jugador en euros.
- Pie preferido del jugador.
- Calificación de reputación internacional de 1 a 5.
- Calificación del pie más débil del jugador de 1 a 5.
- Calificación de movimientos de habilidad de 1 a 5.

Para el preprocesamiento de los datos que nos permitirá aplicar los algoritmos de aprendizaje se prosiguió de la siguiente manera:

1) Se rellenaron los datos faltantes con la media de cada columna.
2) Se codificaron los atributos categóricos (nacionalidad, posición, pie preferido, etc.) como valores numéricos para que puedan ser procesados dentro de los algoritmos.
3) Se llevó a cabo la discretización de la variable objetivo (potencial) de tal forma que quedó dividida en tres clases (bajo, medio, alto)
4) Se definió la variable objetivo y se excluyeron los atributos que no son relevantes.
5) Se dividieron los datos en conjuntos de prueba (20 por ciento) y entrenamiento (80 por ciento)
6) Se normalizaron los datos.

Para el análisis del problema en cuestión, no se considerarán los atributos nombre, nombre completo y fecha de nacimiento debido a que son identificadores únicos de cada jugador y no proporcionan información útil para predecir el potencial de un jugador. El tamaño del conjunto de entrenamiento es de 14363 datos y del conjunto de prueba 3591 datos.

B. Algoritmos a utilizar

Los algoritmos de aprendizaje supervisado y no supervisado que utilizaremos serán los siguientes:

- Árboles de decisión: Este algoritmo funciona bajo la estrategia de dividir de la forma más óptima posible los puntos dentro del árbol. Este proceso de división se repite de forma recursiva de arriba hacia abajo hasta que todos o la mayoría de los registros se hayan clasificado bajo etiquetas de clase específicas [5].
- Support Vector Machine (SVM): El objetivo del algoritmo SVM es encontrar un hiperplano que separe de la mejor forma posible dos clases diferentes de puntos de datos, es decir, el hiperplano con el margen más amplio entre las dos clases. Este concepto de hiperplano de separación en el que se basan los SVMs no se generaliza de forma natural para más de dos clases [6]. Para este caso, dado que es un problema multiclase, se usará SVM One vs One.
- Redes Neuronales: Una red neuronal es un método de la inteligencia artificial que enseña a las computadoras a procesar datos de una manera que está inspirada en la forma en que lo hace el cerebro humano. Se trata de un tipo de proceso de machine learning llamado aprendizaje profundo, que utiliza los nodos o las neuronas interconectados en una estructura de capas que se parece al cerebro humano. Crea un sistema adaptable que las computadoras utilizan para aprender de sus errores y mejorar continuamente. De esta forma, las redes neuronales pueden ser útiles para resolver problemas como, por ejemplo, la clasificación de jugadores de fútbol según su potencial [7].
- K-Means: K-means es un algoritmo de clasificación no supervisada (clusterización) que agrupa objetos en k grupos basándose en sus características. El agrupamiento se realiza minimizando la suma de distancias entre cada objeto y el centroide de su grupo o cluster. Se suele usar la distancia euclidiana [8].

C. Métricas utilizadas para la evaluación.

Las métricas usadas para evaluar los algoritmos de aprendizaje supervisado serán:

- Matriz de confusión: También conocida como matriz de error, es una tabla resumida que se utiliza para evaluar el rendimiento de un modelo de clasificación. El número de predicciones correctas e incorrectas se resume con los valores de conteo y se desglosa por cada clase [9].
- Verdadero Positivo (TP): Resultado en el que el modelo predice correctamente la clase positiva.
- Verdadero Negativo (TN): Resultado donde el modelo predice correctamente la clase negativa.
- Falso Positivo (FP): También llamado error de tipo 1, resultado donde el modelo predice incorrectamente la clase positiva cuando en realidad es negativa.
- Falso Negativo (FN): También llamado error de tipo 2, un resultado en el que el modelo predice incorrectamente la clase negativa cuando en realidad es positiva [9].
- Accuracy: Se define la exactitud (accuracy en inglés) como la proporción entre las predicciones correctas (suma de verdaderos positivos y verdaderos negativos) y las predicciones totales [10]. Su fórmula es la siguiente:
  Accuracy = (TP + TN) / (TP + FP + TN + FN) (1)
- Precision: esta métrica expresa cuántos de los resultados positivos que se han predicho son verdaderamente positivos [11]. Su formula es:
  Precision = TP / (TP + FP) (2)
- Recall: La métrica de recall, también conocida como el ratio de verdaderos positivos, es utilizada para saber cuántos valores positivos son correctamente clasificados [12]. Su formula es:
  Recall = TP / (TP + FN) (3)
- F1 score: Esta métrica combina el precision y el recall, para obtener un valor mucho más objetivo. Se define como la media armónica entre la sensibilidad (recall) y el valor predictivo positivo (precisión), lo que permite balancear ambos aspectos en una sola medida [12]. Su formula es:
  F1 score = 2 ·((precision · recall) / (precision + recall)) (4)

Las métricas para evaluar los algoritmos de aprendizaje no supervisado serán:

- Método de la silueta: El análisis de la silueta mide la calidad del agrupamiento o clustering. Mide la distancia de separación entre los clusters y nos indica cómo de cerca está cada punto de un cluster a puntos de los clusters vecinos. Esta medida de distancia se encuentra en el rango [-1, 1], de tal modo que un valor alto indica un buen clustering [13].
- Método del codo: La idea detrás de este método es bastante sencilla. Identificar el número de clústeres para el que se observa un cambio significativo en la tasa de disminución de la varianza intra-cluster (también conocido como suma total de las distancias al cuadrado) [14]. De esta forma, cuando analizamos gráficamente el método del codo, el número óptimo de clusters será donde la función tienda a volverse "lineal"

III. Resultados

Haciendo un análisis preliminar de los datos respecto a los atributos edad y potencial, se denota de la Fig. 1 que la edad sigue una distribución asimétrica positiva. Mientras que, con respecto al potencial, esta sigue una distribución normal teniendo en cuenta que la medida de potencial va desde cero a cien, tal como se muestra en la Fig. 2.

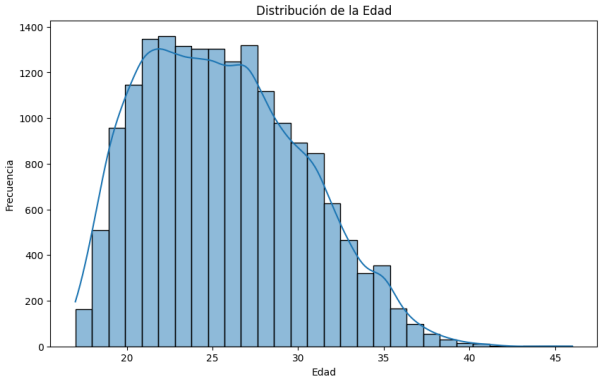

Fig. 1. Distribución de la edad de los jugadores.

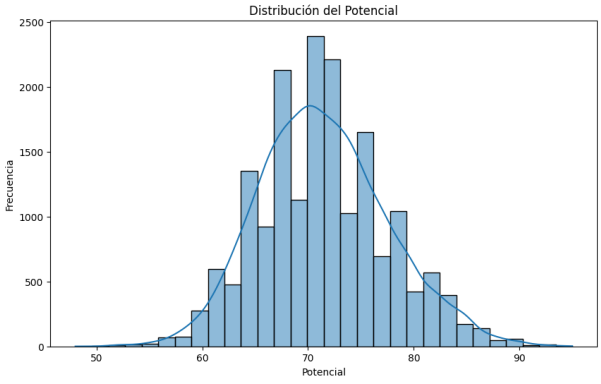

Fig. 2. Distribución del atributo Potencial de los jugadores.

Por su parte, la Figura 3 muestra la relación entre la calificación general ("Overall rating") y el valor de mercado de los jugadores. Este gráfico ilustra cómo el valor de mercado de los jugadores de fútbol está correlacionado con su calificación general. Se puede observar una tendencia positiva entre ambas variables. A medida que la calificación general de los jugadores aumenta, también tiende a aumentar su valor de mercado. Es probable que el gráfico muestre una dispersión de puntos que reflejan esta tendencia, con algunos puntos que podrían ser outliers indicando jugadores con un valor de mercado excepcionalmente alto o bajo para su calificación general.

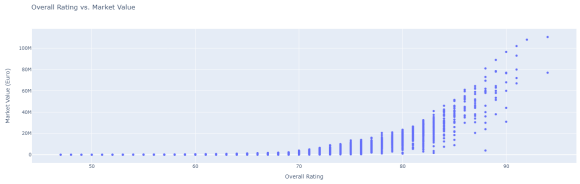

Fig. 3. Relación entre Overall rating y Valor de mercado de los jugadores.

A. Método de aprendizaje supervisado

Para los algoritmos de Árbol de decisión, SVM y Redes neuronales se obtuvieron las siguientes matrices de confusión:

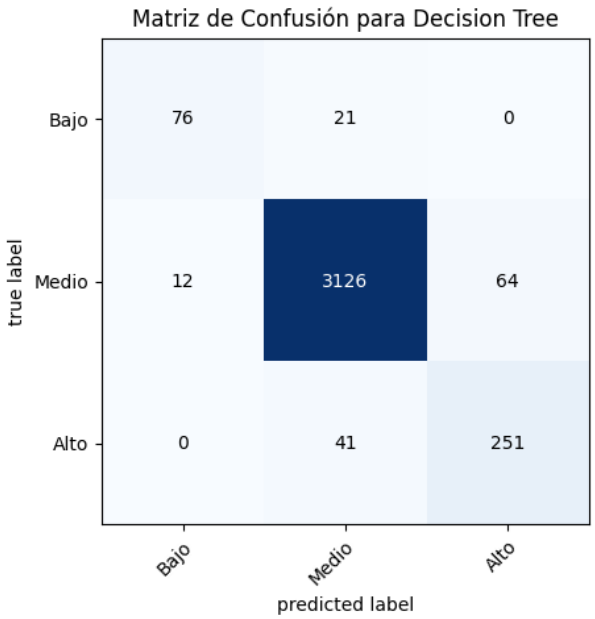

Fig. 4. Matriz de confusión para Árbol de decisión.

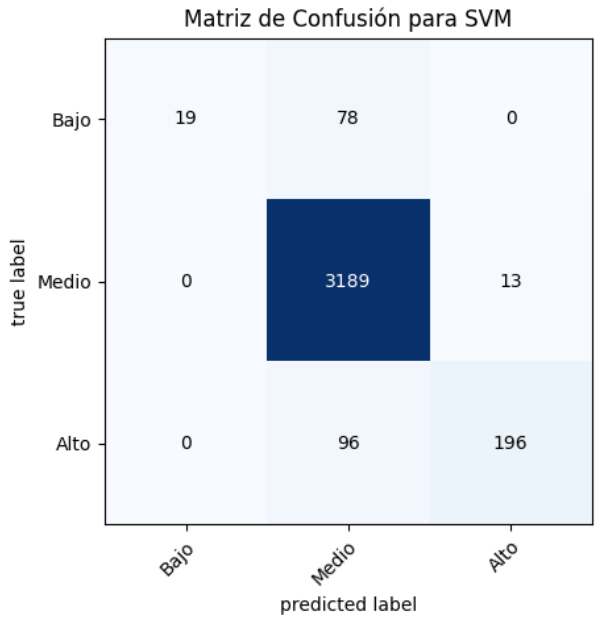

Fig. 5. Matriz de confusión para SVM.

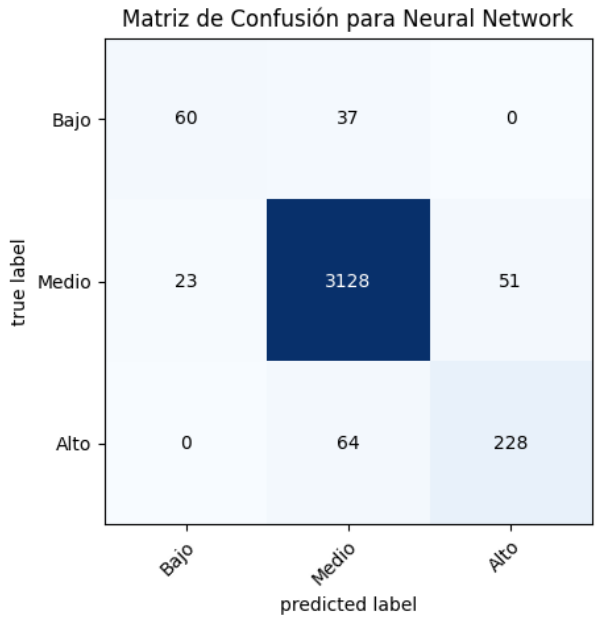

Fig. 6. Matriz de confusión para Redes neuronales.

En base a las matrices de confusión, podemos ver que, en general, los 3 algoritmos predijeron de manera precisa los jugadores con un potencial "medio", sin embargo, los algoritmos de SVM (Fig. 5) y Redes neuronales (Fig 6) tuvieron problemas para predecir las clases "bajo" y "alto". El algoritmo de Árbol de decisión (Fig. 4), en cambio, predijo de manera aceptable estas dos clases.

En la siguiente tabla (Tabla I) se muestra la comparación de las distintas métricas usadas para evaluar los métodos de aprendizaje supervisado:

De la Tabla I, se refleja que el Árbol de Decisión tuvo un desempeño ligeramente superior en comparación a SVM y

TABLE I
MÉTRICAS DE LOS MODELOS DE APRENDIZAJE SUPERVISADO.

Modelo	Accuracy	Precision	Recall	F1 Score
Árbol de Decisión	0.961571	0.962454	0.961570	0.961859
SVM	0.947925	0.948807	0.947925	0.938746
Redes Neuronales	0.955444	0.954219	0.955444	0.954681

Redes Neuronales en cada una de las métricas a evaluar. En base a esto, se elige el algoritmo de Árbol de Decisión para predecir el potencial de los jugadores, el cual al implementarlo dentro del dataset se obtienen las siguientes clasificaciones (Tabla II):

TABLE II
CANTIDAD Y PORCENTAJE DE JUGADORES POR RANGO AL IMPLEMENTAR EL MODELO DE ÁRBOL DE DECISIÓN.

Rango	Número de Jugadores	% Número de Jugadores
Alto	1411	7.858
Medio	16066	89.484
Bajo	477	2.656

B. Método de aprendizaje no supervisado

Seleccionamos el algoritmo de aprendizaje no supervisado K-means, dada la gran cantidad de datos con la que estamos trabajando. Para determinar el número óptimo de clusters, se obtuvo el siguiente gráfico del método del codo:

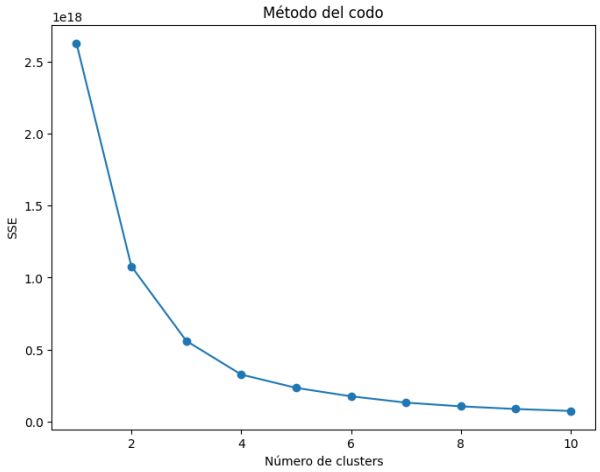

Fig. 7. Método del codo para K-means.

De la Fig. 7 se denota que el número óptimo de clusters es tres, pues es en este punto donde la curva tiende a volverse "lineal". Adicionalmente, se usó el método de la silueta, de donde se obtuvo una puntuación de 0.8199 para k=3, indicando que efectivamente tres es el número de clusters indicado.

De esta forma, se generó el siguiente gráfico (Fig. 8), donde el cluster 2 representa los jugadores con potencial bajo/medio-bajo (16410 jugadores en este cluster), el cluster 0 los jugadores con potencial medio/medio-alto (1399 jugadores en este cluster) y el cluster 1 los jugadores con potencial alto (145 jugadores en este cluster).

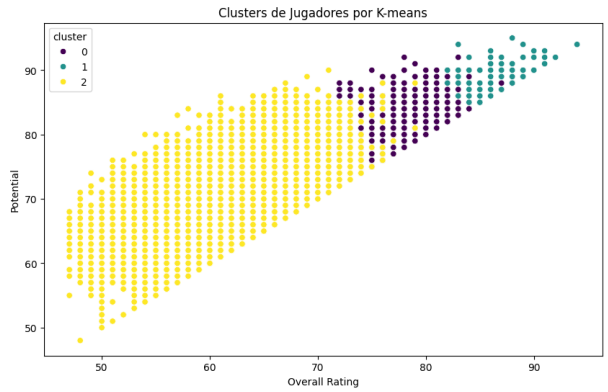

Fig. 8. Clusterización de jugadores mediante K-means.

IV. Conclusión

Mediante la aplicación de 3 algoritmos de aprendizaje supervisado (Árbol de decisión, SVM y Redes neuronales) y 1 de aprendizaje no supervisado (K-means), hemos logrado predecir el potencial de más de 17000 jugadores de fútbol de todo el mundo. Analizando gráficamente mediante las matrices de confusión y las métricas definidas para evaluar los modelos, se determinó que el algoritmo de aprendizaje supervisado más preciso y, en consecuencia, aquel que será definido para ser aplicado dentro del dataset es Árbol de decisión. De esta forma, dentro del conjunto de datos, se predijeron como jugadores con un potencial "bajo" a 477 jugadores, a 16066 con un potencial "medio" y a 1411 con un potencial "alto". En cuanto al algoritmo de aprendizaje no supervisado seleccionado (K-means), se determinó, mediante el método del codo y el método de la silueta, que el número óptimo de clusters era 3. De esta forma, se clasificaron a 16.410 jugadores con un potencial bajo/medio-bajo, a 1399 jugadores con un potencial medio/medio-alto y a 145 jugadores con un potencial alto.

Tener esta información puede ser de gran utilidad dentro de la industria del fútbol, tanto para clubes y Scouts (cazatalentos), como para las estrategias de negociación. Por un lado las relaciones identificadas y las metodologias usadas en este trabajo pueden ser utilizadas por clubes de fútbol y scouts para identificar jugadores con buena relación calidad/-precio. Jugadores con alta calificación general pero con valor de mercado relativamente bajo podrían representar buenas oportunidades de inversión. Por otro lado, los agentes de jugadores pueden usar esta información para negociar contratos y transferencias, demostrando el valor del jugador basado en su calificación general. De esta forma se podría hacer un scouting masivo y así llegar a lograr identificar jugadores que pueden ser altamente rentables para comprar, pensando en el rendimiento que pueden desarrollar dentro del equipo o el beneficio económico que puede significar en futuras ventas. Esto toma aún más relevancia si vemos las enormes sumas de dinero que se pagan hoy en día por jugadores jóvenes con alto potencial, con traspasos que superan los 100 millones de euros y récords de ventas que se rompen cada año dentro de las ligas más importantes del mundo. Un ejemplo claro de esto es el club Independiente del Valle, el cual gracias a la formación y captación de jugadores con un alto potencial han logrado llegar a la élite del fútbol, logrando ganar campeonatos nacionales, disputar la final de la Copa Libertadores y ganar la Copa Sudamericana. Todo esto hubiera sido imposible si no fuera por su enorme capacidad de identificar el potencial de cada jugador, logrando que el club tenga un gran poder económico y un gran prestigio por su fútbol formativo [15].

References

[1] "El favorito de los aficionados: Aumenta la popularidad mundial del fútbol," Nielsen, Junio 2018. [Online]. Disponible en: https://www.nielsen.com/es/insights/2018/fan-favorite-the-global-popularity-of-football-is-rising/. [Accedido: 9-Jul-2024].

[2] "Cuál será el sueldo de Mbappé en Real Madrid y cómo quedó en el ranking entre compañeros," TyC Sports, 4-Jun-2024. [Online]. Disponible en: https://www.tycsports.com/espana/la-liga/cual-sera-el-sueldo-de-mbappe-en-real-madrid-id588900.html. [Accedido: 9-Jul-2024].

[3] M. Ahmed, "Football players data," Kaggle, 2021. [Online]. Disponible en: https://www.kaggle.com/datasets/maso0dahmed/football-players-data/data. [Accedido: 30-Jun-2024].

[4] S. Navarro, "Big Data en el fútbol: Herramientas y aplicaciones," KeepCoding, 17-Apr-2024. [Online]. Disponible en: https://keepcoding.io/blog/funcionamiento-del-big-data-en-el-futbol/. [Accedido: 9-Jul-2024].

[5] "¿Qué es un árbol de decisión?," IBM. [Online]. Disponible en: https://www.ibm.com/es-es/topics/decision-trees. [Accedido: 8-Jul-2024].

[6] J. Amat, "Máquinas de Vector Soporte (Support Vector Machines, SVMs)," Cienciadedatos.net, Abril 2017. [Online]. Disponible en: https://cienciadedatos.net/documentos/

[7] "¿Qué es una red neuronal? - Explicación de las redes neuronales artificiales," AWS. [Online]. Disponible en: https://aws.amazon.com/es/what-is/neural-network/. [Accedido: 8-Jul-2024].

[8] Universidad de Oviedo, "kmeans," Unioviedo.es. [Online]. Disponible en: https://www.unioviedo.es/compnum/laboratorios_py/kmeans/kmeans. [Accedido: 8-Jul-2024].

[9] A. Lee. "Comprensión de la Matriz de Confusión y Cómo Implementarla en Python". DataSource.ai. Accedido el 9 de julio de 2024. [En línea]. Disponible: https://www.datasource.ai/es/data-science-articles/comprension-de-la-matriz-de-confusion-y-como-implementarla-en-python

[10] "Exactitud — Interactive Chaos". Home page — Interactive Chaos. Accedido el 9 de julio de 2024. [En línea]. Disponible: https://interactivechaos.com/es/manual/tutorial-de-machine-learning/exactitud

[11] "¿Cómo aprovechar el rendimiento de la matriz de confusión?" Formación en ciencia de datos — Data-

Scientest.com. Accedido el 9 de julio de 2024. [En línea]. Disponible: https://datascientest.com/es/matriz-de-confusion

[12] R. Díaz. "Métricas de Clasificación". Accedido el 9 de julio de 2024. [En línea]. Disponible: https://www.themachinelearners.com/metricas-de-clasificacion

[13] Á. Gonzalo, "Segmentación utilizando K-means en Python," Machine Learning para todos, 8-Mar-2019. [Online]. Disponible en: https://machinelearningparatodos.com/segmentacion-utilizando-k-means-en-python/. [Accedido: 9-Jul-2024].

[14] D. Rodríguez, "Método del codo (Elbow method) para seleccionar el número óptimo de clústeres en K-means," Analytics Lane, 9-Jun-2023. [Online]. Disponible en: https://www.analyticslane.com/2023/06/09/metodo-del-codo-elbow-method-para-seleccionar-el-numero-optimo-de-clusteres-en-k-means/. [Accedido: 9-Jul-2024].

[15] "Independiente del Valle y un modelo totalmente exitoso que ya suma 6 coronas," ESPN, 1-Mar-2023. [Online]. Disponible en: https://www.espn.cl/futbol/ecuador/nota/_/id/11610784.[Accedido: 9-Jul-2024].