risti
Revista Ibérica de Sistemas e Tecnologias de Informação
Iberian Journal of Information Systems and Technologies

Recebido/Submission: 17/10/2022
Aceitação/Acceptance: 14/12/2022

Técnicas y algoritmos para predecir el resultado de los partidos de fútbol utilizando la minería de datos, una revisión de la literatura

Antonio Araujo-Ahon[1], Brayan Cardenas-Mayta[2], Orlando Iparraguirre-Villanueva[3], Joselyn Zapata-Paulini[4], Michael Cabanillas-Carbonell[5]

aaraujoa@autonoma.edu.pe; bcardenas@autonoma.edu.pe; oiparraguirre@ieee.org; 70994337@continental.edu.pe; mcabanillas@ieee.org

[1,2,3] Facultad de Ingeniería y Arquitectura, Universidad Autónoma del Perú, 15842, Lima, Peru.

[4] Escuela de Posgrado, Universidad Continental, 15301, Lima, Peru.

[5] Facultad de Ingeniería, Universidad Privada del Norte, 15306, Lima, Peru

Pages: 245-263

Resumen: El resultado de un deporte se ha convertido en una necesidad para los competidores, así como para los fanáticos que siguen a sus equipos favoritos. Sin embargo, la predicción de los resultados de un partido de fútbol (PSMR) es muy variada debido a los diversos modelos existentes. La investigación es una revisión sistemática de la literatura (SLR) basada en manuscritos publicados en IEEE Xplore, Scopus, Science Direct y Springer. Se utilizó la metodología Prisma para el análisis y sistematización. El objetivo de esta investigación es ofrecer una guía para haciendo uso de técnicas de machine learning (ML). Los resultados mostraron que las técnicas de ML más utilizadas son el aprendizaje supervisado (SL) y el aprendizaje no supervisado (UL) y el algoritmo de ML más frecuente para predecir el resultado de un partido de fútbol es Random Forest (RF), teniendo en cuenta su gran contribución en la precisión de la predicción. Además, tras el estudio se propone un modelo novedoso y eficiente para predecir el resultado de los partidos de fútbol, apoyado con Data Mining (DM) y centrado en ML.

Palabras-clave: Fútbol, algoritmo, aprendizaje automático, predicciones.

Techniques and algorithms to predict the outcome of soccer matches using data mining, a review of the literature

Abstract: The outcome of a sport has become a necessity for competitors as well as for fans following their favorite teams. However, the prediction of the outcome of a soccer match (PSMR) is very varied due to the various existing models. The research is a systematic literature review (SLR) based on manuscripts published in IEEE Xplore, Scopus, Science Direct, and Springer. Prisma methodology was used for analysis and systematization. The objective of this research is to provide a guide for using machine learning (ML) techniques. The results showed that the most frequently used ML techniques are supervised learning (SL) and unsupervised

learning (UL) and the most frequent ML algorithm for predicting the outcome of a soccer match is Random Forest (RF), considering its great contribution in prediction accuracy. In addition, a novel and efficient model for predicting the outcome of soccer matches, supported with Data Mining (DM) and focused on ML, is proposed after the study.

Keywords: Soccer, algorithm, machine learning, predictions.

1. Introducción

Con el paso del tiempo, el deporte ha formado parte de nuestro día a día, atrayendo a miles de espectadores que, como aficionados a diferentes equipos, siempre han estado interesados en conocer el resultado de cada partido de nuestros equipos favoritos. La tecnología en este deporte nos proporciona una visión amplia del fútbol, con el propósito de poder interpretar los datos, así como ampliar la información identificada en los partidos (Phatak et al., 2022).

En un partido de fútbol, muchas veces el equipo local es considerado como el equipo favorito, acostumbrado a su campo, clima y aficionados, demostrando ser factores relevantes, pero, no muchas veces, el territorio del equipo local es el factor principal, la estrategia, la velocidad, el regate, la potencia de tiro, el control del balón, etc. Se trata de datos que se consideran en el juego individual y en el juego colectivo, pertenecientes a las principales estadísticas que se consideran para el resultado de un partido. Todo esto afecta a un partido, y el equipo visitante puede demostrar ser superior al equipo local (Andrzejewski et al., 2022; Smithies et al., 2021; Talattinis et al., 2019).

Los algoritmos actuales son diversos y muy potentes, capaces de procesar y analizar grandes volúmenes de información. El desarrollo de algoritmos de ML permite analizar y medir el rendimiento de los atletas, no solo físicamente, sino también en atletas electrónicos, siendo la mentalidad un factor importante.(Beal et al., 2020; Muller et al., 2020). El uso de algoritmos tiene una base de pronóstico positiva, acertando el resultado en muchas ocasiones, utilizado por las casas de apuestas a nivel internacional, sin embargo, aún queda mucho camino por recorrer para crear un algoritmo con perfecta precisión en el resultado de los partidos de fútbol (Almulla & Alam, 2020; S. Zhang, 2016)

Sobre la base de lo que se ha descrito anteriormente, el objetivo de este artículo de revisión es analizar y aprender sobre las técnicas de ML y los algoritmos de ML para predecir el resultado de los partidos de fútbol. Proporcionar un resumen analítico basado en investigaciones internacionales en los últimos 4 años. Además de propone un modelo para predecir de manera mas eficiente el resultado de los partidos de fútbol, apoyado con Data Mining (DM) y centrado en ML, en base a las experiencias vividas en anteriores investigaciones.

El artículo está organizado de la siguiente manera. En la sección 2 se describen los principales trabajos relacionados. En la sección 3 se presenta la metodología. La Sección 4 describe y discute los resultados y la discusión. En la sección 5 se presenta el modelo propuesto. Por último, en la sección 6 se presentan las conclusiones.

2. Trabajos relacionados

Los autores, al proponer un modelo de ML centrado en la Premier League inglesa, analizando cinco temporadas desde 2014-2015 hasta 2018-2019, donde el objetivo de su investigación fue identificar las características que serían las más influyentes en los partidos con múltiples clases (local, empate, visitante) y clases binarias (local, visitante). Asimismo, en ellos seleccionaron solo 37 características relevantes e hicieron uso de ML donde trabajaron tres modelos de aprendizaje: Naive Bayes (NB). CHIRP e Hyper Pipes, tomando muestras desde el año 2016 hasta el año 2019, a las que analizaron dieciséis conjuntos, con el objetivo de demostrar cómo estos tres modelos pueden llegar a ser simples, precisos y eficientes en la predicción de los resultados de los partidos de bádminton (Raju et al., 2020; Sharma et al., 2021).

Por otro lado, al presentar un enfoque basado en conjuntos heterogéneos de clasificadores, entrenaron un modelo de ML al que se recogieron seis atributos condicionales y una decisión (match result), con el objetivo de demostrar si a diferencia de los clásicos conjuntos clasificadores con clasificadores heterogéneos mejoran la predicción de los resultados de fútbol en la Bun-league. Del mismo modo, en el que intentaron predecir por rangos de jugadores utilizando algoritmos de clasificación de ML, utilizaron el método de conjunto para combinar varios algoritmos en los que tenían como objetivo crear un conjunto de datos que los identificara cuáles son los problemas de los jugadores en función de sus habilidades y reconociera las nuevas promesas del fútbol iraní (Feng, 2017; Kozak & Głowania, 2021; Maanijou & Mirroshandel, 2019).

En ellos se centran en la Premier League inglesa (EPL) utilizando la técnica de redes neuronales profundas, comparando con algoritmos ML como support vector machine (SVM), NB y la RF cuyo objetivo principal era proponer su clasificador de perceptrones multicapa (MLP) con un conjunto de funciones utilizando el método de retención cuya finalidad es dividir los conjuntos, y poder entrenarlos para analizar el rendimiento en los resultados de los partidos (Rudrapal et al., 2020). Asimismo, en la que proponen un modelo ML para clasificar los diferentes manuscritos deportivos haciendo uso del algoritmo que propusieron Modificado Cortical (CA), buscaron ver si su algoritmo entrenado podía extraer los datos objetivos y subjetivos de los 1000 manuscritos deportivos seleccionados cuya finalidad de su objetivo principal era comparar su modelo con los modelos del algoritmo SVM y su variante LMSVM, para ver el grado de reducción en la validación de las características encontradas. Del mismo modo, en el que buscan comparar el rendimiento proporcionado por el perovskita en equipos de entrenamiento deportivo de jugadores utilizando predicciones con redes neuronales, dividiendo 126 conjuntos en 2 grupos entrenándolos y probándolos unas 199 veces, utilizaron el algoritmo genético y el algoritmo SVM centrándose en dos tipos de reconocimiento de patrones con el objetivo de ver qué algoritmos con sus métodos les daban los resultados con valores estándar a la predicción. Asimismo, en la que propusieron predecir partidas de ajedrez en las que construyeron un modelo con redes neuronales LSTM, para compararlo con otros métodos de clasificación como entrada de bits y entrada algebraica, buscaron seleccionar el método que debía ser entrenado por redes neuronales porque pretendían llegar a ser altamente precisos aplicando grandes enfoques de conjuntos de

datos con registros del año 2017 al 2018 (Drezewski & Wator, 2021; Hajj et al., 2019; L. Zhang & Li, 2022).

3. Metodología

En la presente investigación se utiliza la metodología PRISMA para buscar investigación internacional adecuada para nuestra investigación. También se realiza un análisis bibliométrico para tener una descripción detallada de los tipos de algoritmos que se utilizan para la predicción de resultados en los partidos de fútbol. Finalmente, se analizan los manuscritos para tener a nuestra disposición las técnicas utilizadas para la predicción de resultados de partidos de fútbol. Como resultado, se puede utilizar para informar futuras investigaciones y estudios en PSMR u otros deportes para lograr una predicción de mayor precisión porcentual del resultado del partido.

El estudio SLR Related Manuscripts presenta una evaluación de la investigación internacional sobre el tema de la predicción de resultados de partidos de fútbol utilizando una metodología precisa y comprobable basada en el enfoque PRISMA, que consta de cuatro pasos que se describen a continuación.

- Identificación de manuscritos relevantes para el tema.
- Exclusión de la selección de texto completo.
- Análisis de elegibilidad
- Inclusión de manuscritos finales para ser analizados en detalle

Además, se incluye un análisis a través de un mapa bibliométrico para encontrar relaciones entre palabras comunes en la investigación. Se evaluó el número de veces que se repiten las palabras, el número de palabras más frecuentes y la frecuencia de estas palabras en los artículos finales del estudio. Según PRISMA, esta sección está estructurada por: 1) preguntas de investigación, 2) estrategias de búsqueda, 3) criterios de inclusión y exclusión, y 4) selección final de manuscritos.

3.1. Preguntas de investigación

En esta sección se presentan las preguntas de investigación (RQ) del estudio con respecto a nuestro objetivo principal de investigación. Las preguntas de investigación propuestas se muestran en la Tabla 1, que son las siguientes:

RQ1	¿Qué tipos de técnicas de ML se utilizan más para predecir el resultado de los partidos de fútbol?
RQ2	¿Qué tipos de algoritmos de ML se utilizan más para predecir el resultado de los partidos de fútbol?
RQ3	¿Qué tipos de técnicas de minería de datos (DM) se utilizan más para predecir el resultado de los partidos de fútbol?

Tabla 1 – Preguntas de investigación

3.2. Estrategias de búsqueda

Se requirió un análisis cualitativo sistemático para categorizar los datos según diferentes temas de análisis. Los manuscritos de 2019 a 2022 se exploraron en cuatro bases de datos virtuales (Scopus, IEEE Xplore, Science Direct y Springer). Se consideró que los años de 2019 a 2022 tenían a nuestra disposición manuscritos relevantes centrados en algoritmos para predecir un resultado en los deportes y mejorar la predicción de los resultados de los partidos en el fútbol. La cadena de búsqueda y las combinaciones que se utilizaron fueron "Algorithm AND Predict AND Outcome OR Results AND Data AND Mining And Sports". Esto asegura que el algoritmo para predecir el resultado de los partidos de fútbol estará ampliamente cubierto por disciplinas de técnicas de ML, algoritmos de ML y técnicas de DM.

3.3. Criterios de inclusión y exclusión

Para el estudio SLR se aplicaron los siguientes criterios de inclusión y exclusión, como se muestra en la Tabla 2.

Criterios de inclusión
Manuscritos relacionados con ML
Manuscritos relacionados con algoritmos
Manuscritos relacionados con la predicción del rendimiento deportivo
Manuscritos relacionados con DM
En los últimos 4 años
Criterios de exclusión
Manuscritos no relacionados con el tema
Libros, Documentos cortos
Manuscritos duplicados
Manuscritos SLR

Tabla 2 – Criterios de inclusión y exclusión

El proceso de búsqueda y selección se muestra en la Figura 1. La búsqueda dio como resultado un total de 1225 manuscritos identificados a través de las bases de datos, de los cuales se identificaron 8 duplicados, luego se excluyeron un total de 1068 manuscritos según los criterios. Esto dio lugar a 149 manuscritos para su evaluación. Otros 98 manuscritos fueron eliminados en base a los siguientes criterios: se excluyeron aquellos que no estaban relacionados con el tema, no reportaron los resultados medidos, no tuvieron una lectura abierta y libre para el usuario, lo que nos dio un total de 51 manuscritos para revisión.

Los 51 manuscritos incluidos en la revisión de metaanálisis fueron del motor de búsqueda IEEE Xplore con 34 manuscritos, Springer con 8, Scopus con 7 y Science Direct con el menor número, con solo 2.

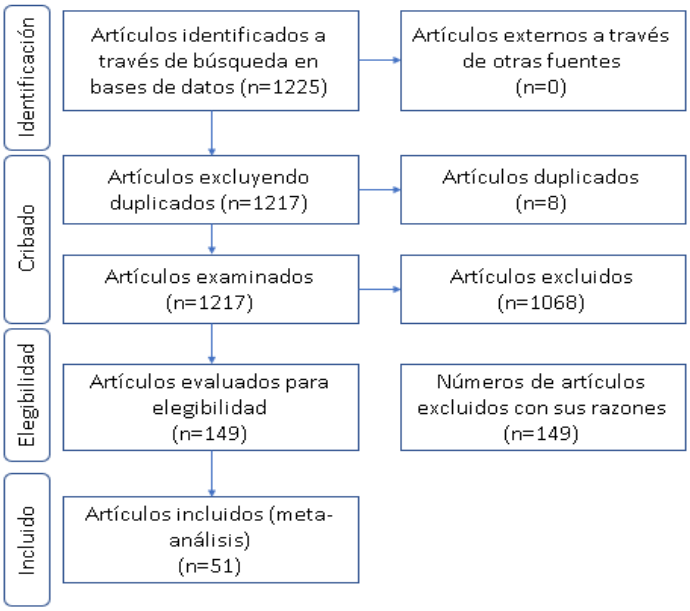

Figura 1 – Diagrama de prisma

Finalmente, estos 51 manuscritos se dividieron en categorías, estudios relacionados con las técnicas de ML para la previsión; estudios relacionados con algoritmos de ML y técnicas de DM para la predicción.

4. Resultados y discusión

Esta sección presenta el análisis bibliométrico y el análisis detallado de trabajos anteriores. La primera parte presenta las relaciones entre las palabras comunes relacionadas con los algoritmos para PSMR, utilizando DM y métricas. La segunda parte busca encontrar la brecha científica entre los trabajos analizados en este estudio para desarrollar un nuevo modelo para predecir con precisión los resultados de coincidencia a través de técnicas de ML.

4.1. Análisis bibliométrico

VOSviewer, una red de visualización bibliométrica, se utilizó para encontrar terminología común relacionada con PSMR utilizando DM y algoritmos en los 51 manuscritos analizados. Esta herramienta fue de gran ayuda, permitiendo la visualización de información gráfica, visualizando las asociaciones de términos clave asociados a PSMR. Además, ayudó a identificar técnicas, algoritmos de DM y predicciones de resultados de partidos de fútbol en grupos. La Figura 3 presenta el mapa de red que muestra las relaciones entre la terminología más utilizada y cómo se asocia. El nodo más grande representa la terminología más utilizada en los

manuscritos, y su tamaño representa el número de veces que estas palabras aparecen en los artículos analizados. El VOSviewer agrupa la terminología en grupos de acuerdo con su importancia entre sí (Munir et al., 2022).

El análisis se realizó con respecto al título y al resumen, utilizando un método de conteo numérico de 117 palabras clave examinadas con un umbral mínimo de 1 ocurrencia, resultando en 117 términos, como se muestra en la Figura 2. Los nodos más grandes que representan cada clúster en el mapa de red se determinan como: clúster de aprendizaje automático (amarillo), clúster DM (azul claro), clúster de análisis deportivo (verde), clúster de predicción (marrón), neuronal clúster de red (rojo) y clúster de aprendizaje profundo (lead).

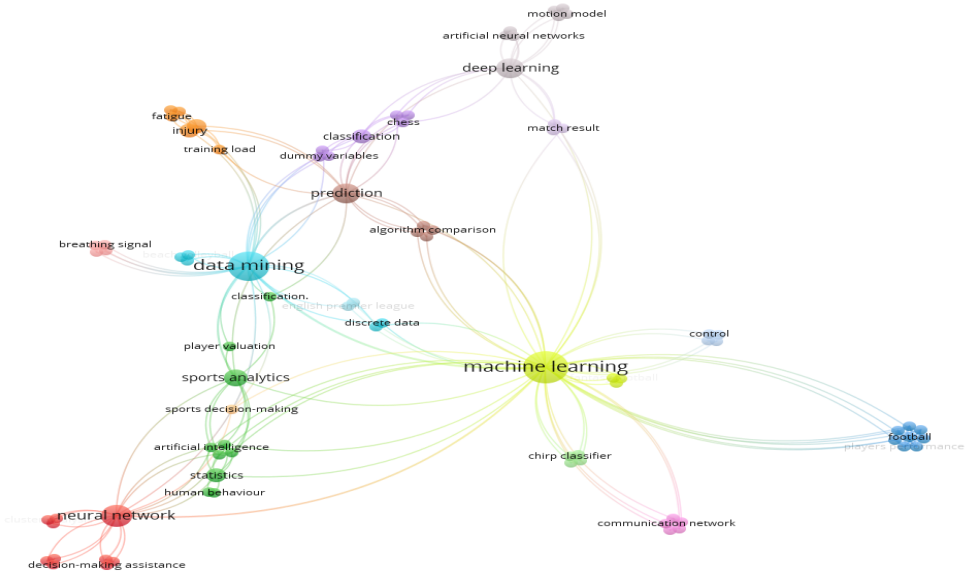

Figura 2 – Mapa bibliométrico con las relaciones de los términos más utilizados.

Mirando el mapa bibliométrico de la Figura 2, podemos cuantificar 16 cúmulos que están conectados entre sí; por ejemplo, los términos de “aprendizaje automático” se asocian con “fútbol” en el mismo clúster amarillo, también se asocian con el clúster “análisis deportivo”, inteligencia artificial” y “estadísticas” en el clúster verde, también está conectado con “predicciones” y “comparación de algoritmos” en el clúster marrón; por otro lado, también se asocia con “red neuronal” en el clúster rojo, se asocia con “aprendizaje profundo” en el clúster principal. Además, la “clasificación” en el clúster púrpura se asocia con la “minería de datos” en el clúster azul claro, también se asocia con “predicciones” en el clúster marrón y el “aprendizaje profundo” en el clúster principal, el “resultado de coincidencia” está conectado en el clúster amarillo. Finalmente, mediante el análisis del mapa de red en la Figura 2 fue posible identificar los términos más relevantes en cada uno de los clusters, como se muestra en la Figura 3.

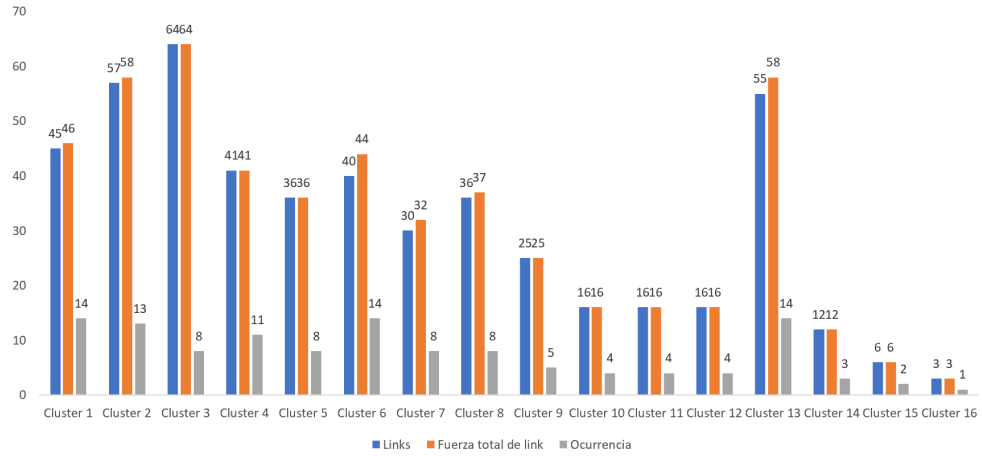

Figura 3 – Enlaces, fuerza total del enlace y ocurrencia.

Los principales hallazgos con respecto a las técnicas de ML se muestran en la Tabla 4. Además, la Figura 4 muestra los principales algoritmos de ML utilizados para predecir los resultados de los partidos de fútbol. Del mismo modo, en (Elmiligi & Saad, 2022; Fialho et al., 2019) concluido que técnicas como UL y SL son las más utilizadas para el procesamiento de datos en DM.

Uno de los mejores trabajos de ML se presentó en donde los autores proponen un modelo para predecir partidos de fútbol, para lo cual utilizaron algoritmos ML, SVM, Regresión Bayesiana y LR. Para probar los diferentes algoritmos, utilizaron la variable de resultado a tiempo completo (home win, draw, away win) en forma de un vector compuesto por 3 valores binarios ([1,0,0,0], [0,1,0], [0,0,1]), para verificar si su modelo puede alcanzar una buena precisión, realizaron pruebas con los algoritmos mencionados. En su lugar, propone hacer uso de técnicas de ML como SVM, LR, Classifier, NB, Decision Tree (DT) y AdaBoost, aplicando ingeniería de características, buscaron probar si las variables (goles del equipo local, goles recibidos y porcentaje de victorias) serían precisas en base a su modelo en análisis estadístico con datos entre las temporadas 2014-2015 y 2018-2019 (Elmiligi & Saad, 2022; Fialho et al., 2019).

Los factores que pueden cambiar el resultado de un juego son múltiples, por lo que existen muchos modelos que buscan predecir con precisión a través del análisis de las diversas características dentro de los datos. Por ejemplo, en los factores que afectan a un partido (el terreno de juego, el clima, el campo, el lugar, el equipo), se deben analizar en conjuntos de datos para procesarlos y extraer valores destacados para una predicción. Corroboraron con el uso de algoritmos de ML ya que la precisión NB aborda entre 81.63%, el SVM abordó 83.67%, con DT fue de 87.75%, el RF obtuvo 83.67% como también en el RL 95.92% de precisión concluyendo que es importante definir los factores de rendimiento. Por otro lado, al evaluar los factores (combinación, ligas, posición, categoría) proponiendo una medición de rendimiento presentando una puntuación máxima del 90%, se evaluaron diferentes métodos (filtro, BR, RF) donde destacaron las

técnicas clásicas de ML con un 75% como las más acertadas en una predicción de partido de fútbol (Barot et al., 2020; Nsolo et al., 2019).

4.1. Análisis de manuscritos

Para la búsqueda (4) se utilizaron bases de datos virtuales, se recuperaron un total de mil doscientos veinticinco (1225) documentos. La base de datos Science Direct tiene el mayor número de documentos recuperados, quinientos cincuenta y siete (557) en el año 2021, con el mayor número de manuscritos publicados, como se presenta en la Tabla 3.

No	Base de datos	Búsqueda inicial	Categorización por año de publicación				
			2022	2021	2020	2019	Resultado final
1	Ciencia Directa	557	-	1	-	1	2
2	IEEEXplore	423	7	13	7	7	34
3	Salmer	223	4	3	1	-	8
4	Scopus	22	-	3	-	4	7
Resumen		1225	11	20	8	12	51

Tabla 3 – Análisis de manuscritos relacionados con PSMR

En la revisión, se encontraron 8 duplicados en los 1225 documentos recuperados. Un total de 22 documentos fueron recuperados en Scopus, de los cuales 15 fueron descartados en base a los criterios de exclusión. En Science Direct, se recuperaron 557 documentos, aunque 555 se descartaron en función de los criterios de exclusión. El número total de documentos recuperados en Springer es de 223 y 215 fueron descartados. En IEEE Xplore, se recuperaron 423 documentos y se descartaron 339 en función de los criterios de exclusión.

RQ1 provocó la búsqueda de fuentes de datos y tipos de técnicas de ML para predecir el resultado de los partidos de fútbol. Nuestra revisión de manuscritos nos permitió extraer y analizar técnicas de ML como: SL y UL. La Tabla 4 muestra los manuscritos que utilizaron SL y UL como técnicas de ML para predecir el resultado de los partidos de fútbol. Se encontraron 9 manuscritos usando SL y el mismo número de manuscritos usando UL. Se ha observado que las dos técnicas de ML, mencionadas anteriormente, son las más utilizadas para predecir el resultado de los partidos de fútbol.

Técnicas de ML	Manuscritos
Sl	(Barot et al., 2020; Elmiligi & Saad, 2022; Hatharasinghe & Poravi, 2019; LeTu, 2022; Mahbub et al., 2021; Nsolo et al., 2019; H. Wang et al., 2021; W.-Y. Wang et al., 2021; Weeraddana & Premaratne, 2021)
Colmena	(Bafna & Saini, 2019; Fialho et al., 2019; Guan & Wang, 2022; Meddegoda et al., 2021; Theiner et al., 2022; Tyran & Chomatek, 2021; T. Wang, 2022; Yao, 2019; Y. Zhang et al., 2022)

Tabla 4 – Referencias manuscritas asociadas con técnicas de predicción de ML

Al abordar RQ2, examinamos los algoritmos aplicables a la predicción del resultado de los partidos de fútbol. Con este objetivo, se analizaron los manuscritos asociados al algoritmo RF, DT, BR, clasificación, NB, LR, entre otros, como se muestra en la Tabla 5 y Figura 5, con el algoritmo Random Forest se encontró en 5 manuscritos relacionados que se aplicaban a la predicción del resultado de los partidos de fútbol. Con algoritmo de clasificación y Naive Bayes, ambos con 3 manuscritos respectivamente para predecir el resultado de los partidos de fútbol. Para ello se encontró algoritmo árbol de decisión y regresión logística, 2 en ambos manuscritos relacionados con la predicción del resultado de los partidos de fútbol. Solo se encontró un artículo donde se aplicó el algoritmo de regresión lineal. Por último, los algoritmos que se encuentran en el apartado "otros" son métodos que no son muy comunes o son variaciones, mejoras a los algoritmos tradicionales, con el objetivo de obtener resultados mucho más eficientes.

Algoritmos de ML	**Cantidad**	**Artículos**
Ramdom Forest	5	(Barot et al., 2020; Kumar et al., 2021; Mahbub et al., 2021; Mundhe et al., 2021; Thenmozhi et al., 2019)
Tree Decision	2	(Hewko et al., 2019; Ur et al., 2019)
Lineal Regression	1	(Yao, 2019)
Classification	3	(Apostolou & Tjortjis, 2019; Balasundaram et al., 2020; Fujita et al., 2019)
Naive Bayes	3	(Kumar et al., 2021; Ur et al., 2019; Weeraddana & Premaratne, 2021)
Logistic Regression	2	(Barot et al., 2020; Elmiligi & Saad, 2022)
Others	11	(Azeman et al., 2021; Hatharasinghe & Poravi, 2019; LeTu, 2022; Nsolo et al., 2019; Pantzalis & Tjortjis, 2020; Shi et al., 2022; Victor et al., 2021; H. Wang et al., 2021; T. Wang, 2022; W.-Y. Wang et al., 2021; Yuan et al., 2021)

Tabla 5 – Estudios relacionados con Algoritmos de ML

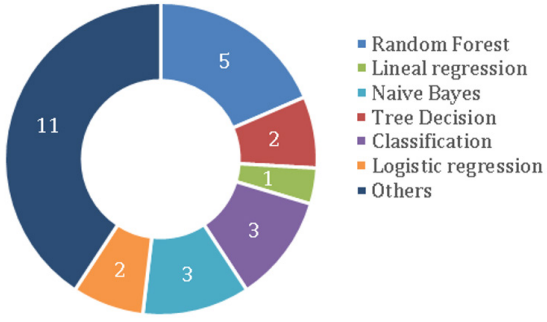

Figura 4 – Algoritmos ml para predecir el resultado de los partidos de fútbol.

El estudio también analizó las técnicas de DM para predecir el resultado de los partidos de fútbol en el ámbito de nuestro RQ3. Los resultados se muestran en la Tabla 6, donde podemos apreciar las técnicas de DM que se utilizaron para la extracción y procesamiento de datos para predecir el resultado de los partidos de fútbol. Con la técnica de clasificación se encontraron 14 manuscritos que utilizaron esta técnica en los manuscritos analizados; con las técnicas de agrupamiento, se encontraron 6 manuscritos, respectivamente. En el apartado de otros, encontramos técnicas que no son muy comunes o variaciones para el análisis de datos.

Técnicas de Mineria de Datos (DM)		**Cantidad**	**Artículos**
Clasificación		14	(Ahmadi et al., 2020; Apostolou & Tjortjis, 2019; Balasundaram et al., 2020; Can Yücebaş, 2022; Fialho et al., 2019; Fujita et al., 2019; Gu et al., 2021; Meddegoda et al., 2021; Pramanik et al., 2022; Song et al., 2019; Subbaraj et al., 2020; Tyran & Chomatek, 2021; Yao, 2019; Yin & Cui, 2021)
Agrupamiento		6	(Amarasena et al., 2019; Bafna & Saini, 2019; Ma, 2022; Rajesh et al., 2020; D. Wang, 2022; Y. Zhang et al., 2022)
Otros	*Modelo Estadístico*	1	(Chun et al., 2021)
	Modelo estadístico de Markow	1	(Song et al., 2019)

Tabla 6 – Estudios relacionados con las Técnicas de Minería de Datos (DM) para predecir el resultado de los partidos de fútbol.

5. Modo Propuesto

Después de haber analizado los manuscritos relacionados con la predicción de los resultados de los partidos de fútbol y de haber encontrado en la mayor parte de las investigaciones el uso de diferentes modelos para predecir el resultado de los partidos de fútbol, en esta sección, se propone un modelo novedoso y eficiente para predecir el resultado de los partidos de fútbol, apoyado con DM y centrado en ML. El modelo propuesto ha utilizado varias investigaciones de ciencia de datos, como es el caso en este trabajo, el modelo propuesto presenta 6 fases, que se presentan de la siguiente manera: 1) recolección de datos; 2) preprocesamiento de datos; 3) técnicas de evaluación; 4) construcción de modelos; 5) modelo de entrenamiento; 6) pruebas de modelos.

1. Recogida de datos: en esta primera fase se obtiene el conjunto de datos para iniciar el preprocesamiento.
2. Preprocesamiento de datos: el conjunto de datos está compuesto por varias entidades que identifican los atributos que no serán relevantes para el modo.
3. Técnica de evaluación: La técnica de evaluación para el conjunto de datos es donde debemos comparar y seleccionar el modelo en el ML. Para la comprensión y la aplicación, la técnica de validación cruzada K-Fold es una buena opción, ya que su conjunto de datos se divide en k parte al azar y cada modelo se entrena y prueba k veces, donde se determina tomando la media de las medidas de precisión.

$$CV = \frac{1}{k}\sum_{i=1}^{k} A_i$$

4. Construcción de modelos: Para lograr el análisis, se pueden utilizar hasta cuatro algoritmos de ML supervisados, rf, SVM, LR y NB. Para construir el modelo requerido.
5. Entrenamiento del modelo: El modelo se entrena con un corpus, donde se muestra el resultado debido al entrenamiento, buscando una mayor precisión y disminuyendo los márgenes de error para predecir el resultado de un partido de fútbol.
6. Pruebas de modelos: La exactitud de la predicción, la precisión y la tasa de error se verifican para comparar los algoritmos de ML.

La representación gráfica del modelo propuesto se muestra en la Figura 6 a continuación.

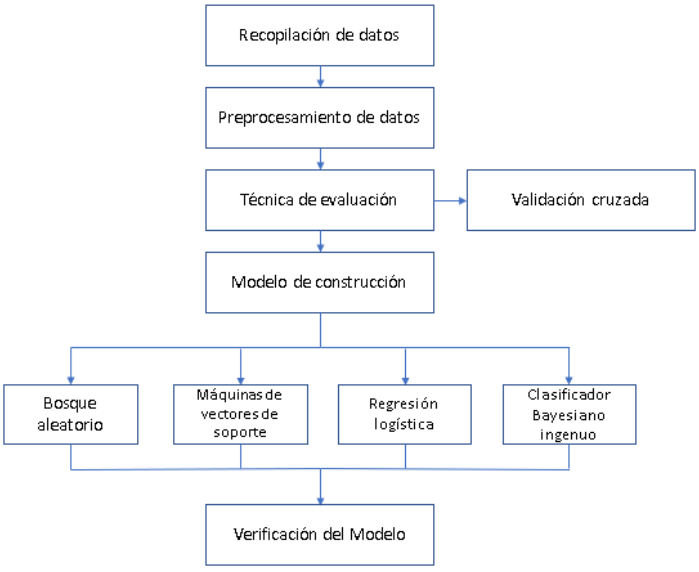

Figura 6 – Modelo propuesto

6. Conclusiones

En el presente trabajo, se seleccionaron 51 manuscritos de diversas fuentes de datos, países y años que fueron revisados y promovidos como referencias para predecir el resultado de los partidos de fútbol utilizando DM, técnicas y algoritmos de ML. Esta investigación encontró oportunidades para futuras investigaciones en la predicción del resultado de los partidos de fútbol basadas en la revisión y discusión de la investigación existente a través de sus problemas, metodologías y usos de datos abordados.

Respondiendo rq1. (¿Qué tipos de técnicas de ML son las más utilizadas para predecir el resultado de los partidos de fútbol?), podemos afirmar que las técnicas de ML más utilizadas para predecir el resultado de los partidos de fútbol son SL y UL, como se indica en la Tabla 4.

Respondiendo rq2. (¿Qué tipos de algoritmos de ML son los más utilizados para predecir el resultado de los partidos de fútbol?), podemos señalar que el algoritmo de RF es el que tiene un alto porcentaje de precisión, además de ser el más utilizado en diferentes investigaciones, como se muestra en la Tabla 5 y Figura 4.

Respondiendo rq3. (¿Qué tipos de técnica de DM son los más utilizados para predecir el resultado de los partidos de fútbol?), podemos asegurar que la técnica de DM más utilizada para la extracción y procesamiento de datos es la técnica de clasificación, ampliamente utilizada en las diversas investigaciones, como se muestra en la Tabla 6. Finalmente, los resultados de este estudio ayudarán a los investigadores a predecir los resultados del fútbol o el deporte. El uso de técnicas y algoritmos de ML ayudará a muchas personas en la predicción para obtener el resultado de los partidos. El estudio propone algunas referencias para futuras investigaciones basadas en manuscritos revisados sobre técnicas y algoritmos de ML, técnicas de DM. Las oportunidades para un mayor estudio de investigación se pueden realizar con un mayor uso de datos, especialmente para un sistema para predecir el resultado de los partidos de fútbol y proponer acciones recomendadas para poder predecir el resultado de los partidos.

Referencias

Ahmadi, A. H., Noori, A., & Teimourpour, B. (2020). Social Network Analysis of Passes and Communication Graph in Football by mining Frequent Subgraphs. *2020 6th International Conference on Web Research (ICWR)*, 1–7. https://doi.org/10.1109/ICWR49608.2020.9122303

Almulla, J., & Alam, T. (2020). Machine Learning Models Reveal Key Performance Metrics of Football Players to Win Matches in Qatar Stars League. *IEEE Access*, *8*, 213695–213705. https://doi.org/10.1109/ACCESS.2020.3038601

Amarasena, P. T., Kumara, B. T. G. S., & Jointion, S. (2019). Data Mining Approach for Identifying Suitable Sport for Beginners. *2019 International Research Conference on Smart Computing and Systems Engineering (SCSE)*, 57–62. https://doi.org/10.23919/SCSE.2019.8842785

Andrzejewski, M., Oliva-Lozano, J. M., Chmura, P., Chmura, J., Czarniecki, S., Kowalczuk, E., Rokita, A., Muyor, J. M., & Konefał, M. (2022). Analysis of team success based on match technical and running performance in a professional soccer league. *BMC Sports Science, Medicine and Rehabilitation*, *14*(1), 82. https://doi.org/10.1186/S13102-022-00473-7

Apostolou, K., & Tjortjis, C. (2019). Sports Analytics algorithms for performance prediction. *2019 10th International Conference on Information, Intelligence, Systems and Applications (IISA)*, 1–4. https://doi.org/10.1109/IISA.2019.8900754

Azeman, A. A., Mustapha, A., Razali, N., Nanthaamomphong, A., & Abd Wahab, M. H. (2021). Prediction of Football Matches Results: Decision Forest against Neural Networks. *2021 18th International Conference on Electrical Engineering/ Electronics, Computer, Telecommunications and Information Technology (ECTI-CON)*, 1032–1035. https://doi.org/10.1109/ECTI-CON51831.2021.9454789

Bafna, P. B., & Saini, J. R. (2019). Identification of Significant Challenges in the Sports Domain using Clustering and Feature Selection Techniques. *2019 9th International Conference on Emerging Trends in Engineering and Technology - Signal and Information Processing (ICETET-SIP-19)*, 1–5. https://doi.org/10.1109/ICETET-SIP-1946815.2019.9092011

Balasundaram, A., Ashokkumar, S., Jayashree, D., & Magesh Kumar, S. (2020). Data mining based Classification of Players in Game of Cricket. *2020 International Conference on Smart Electronics and Communication (ICOSEC)*, 271–275. https://doi.org/10.1109/ICOSEC49089.2020.9215413

Barot, H., Kothari, A., Bide, P., Ahir, B., & Kankaria, R. (2020). Analysis and Prediction for the Indian Premier League. *2020 International Conference for Emerging Technology (INCET)*, 1–7. https://doi.org/10.1109/INCET49848.2020.9153972

Beal, R., Norman, T. J., & Ramchurn, S. D. (2020). Optimising Daily Fantasy Sports Teams with Artificial Intelligence. *International Journal of Computer Science in Sport*, *19*(2), 21–35. https://doi.org/10.2478/ijcss-2020-0008

Can Yücebaş, S. (2022). A deep learning analysis for the effect of individual player performances on match results. *Neural Computing and Applications*, *34*, 12967–12984. https://link.springer.com/article/10.1007/s00521-022-07178-5

Chun, S., Son, C.-H., & Choo, H. (2021). Inter-dependent LSTM: Baseball Game Prediction with Starting and Finishing Lineups. *2021 15th International Conference on Ubiquitous Information Management and Communication (IMCOM)*, 1–4. https://doi.org/10.1109/IMCOM51814.2021.9377370

Drezewski, R., & Wator, G. (2021). Chess as sequential data in a chess match outcome prediction using deep learning with various chessboard representations. *Procedia ComputerScience*,*192*,1760–1769.https://doi.org/10.1016/J.PROCS.2021.08.180/CHESS_AS_SEQUENTIAL_DATA_IN_A_CHESS_MATCH_OUTCOME_PREDICTION_USING_DEEP_LEARNING_WITH_VARIOUS_CHESSBOARD_REPRESENTATIONS.PDF

Elmiligi, H., & Saad, S. (2022). Predicting the Outcome of Soccer Matches Using Machine Learning and Statistical Analysis. *2022 IEEE 12th Annual Computing and Communication Workshop and Conference (CCWC)*, 1–8. https://doi.org/10.1109/CCWC54503.2022.9720896

Feng, S. (2017). The Application of the Improved Chaos Algorithm in the Sports Performance Prediction. *Proceedings of the 2017 International Conference on Sports, Arts, Education and Management Engineering (SAEME 2017)*. https://doi.org/10.2991/saeme-17.2017.117

Fialho, G., Manhães, A., & Teixeira, J. P. (2019). Predicting Sports Results with Artificial Intelligence – A Proposal Framework for Soccer Games. *Procedia Computer Science*, *164*, 131–136. https://doi.org/10.1016/j.procs.2019.12.164

Fujita, K., Hori, A., Maki, S., Isono, M., Sugahara, T., & Kato, C. (2019). A Study on Attractive Elements Characteristic of Paralympic Sports. *2019 8th International Congress on Advanced Applied Informatics (IIAI-AAI)*, 985–988. https://doi.org/10.1109/IIAI-AAI.2019.00197

Gu, Y., Zhou, W., Yuan, Z., & Xu, W. (2021). Intelligent analysis framework of sports training intensity based on breathing signal detection algorithm. *2021 5th International Conference on Electronics, Communication and Aerospace Technology (ICECA)*, 1279–1283. https://doi.org/10.1109/ICECA52323.2021.9675940

Guan, S., & Wang, X. (2022). Optimization analysis of football match prediction model based on neural network. *Neural Computing and Applications*, *34*(4), 2525–2541. https://doi.org/10.1007/s00521-021-05930-x

Hajj, N., Rizk, Y., & Awad, M. (2019). A subjectivity classification framework for sports articles using improved cortical algorithms. *Neural Computing and Applications*, *31*(11), 8069–8085. https://doi.org/10.1007/S00521-018-3549-3

Hatharasinghe, M. M., & Poravi, G. (2019). Data Mining and Machine Learning in Cricket Match Outcome Prediction: Missing Links. *2019 IEEE 5th International Conference for Convergence in Technology (I2CT)*, 1–4. https://doi.org/10.1109/I2CT45611.2019.9033698

Hewko, J., Sullivan, R., Reige, S., & El-Hajj, M. (2019). Data Mining in The NBA: An Applied Approach. *2019 IEEE 10th Annual Ubiquitous Computing, Electronics & Mobile Communication Conference (UEMCON)*, 426–432. https://doi.org/10.1109/UEMCON47517.2019.8993074

Kozak, J., & Głowania, S. (2021). Heterogeneous ensembles of classifiers in predicting Bundesliga football results. *Procedia Computer Science*, *192*, 1573–1582. https://doi.org/10.1016/J.PROCS.2021.08.161

Kumar, Y., Sharma, H., & Pal, R. (2021). Popularity Measuring and Prediction Mining of IPL Team Using Machine Learning. *2021 9th International Conference on Reliability, Infocom Technologies and Optimization (Trends and Future Directions) (ICRITO)*, 1–5. https://doi.org/10.1109/ICRITO51393.2021.9596405

LeTu, G. R. (2022). A Machine Learning Framework for Predicting Sports Results Based on Multi-Frame Mining. *2022 4th International Conference on Smart Systems and Inventive Technology (ICSSIT)*, 810–813. https://doi.org/10.1109/ICSSIT53264.2022.9716296

Ma, H. (2022). Complex Software Ecology and AI Fusion Multi-Sensing Data Mining and Cluster Analysis of Online Sports Courses for College Students. *2022 6th International Conference on Trends in Electronics and Informatics (ICOEI)*, 1735–1738. https://doi.org/10.1109/ICOEI53556.2022.9776926

Maanijou, R., & Mirroshandel, S. A. (2019). Introducing an expert system for prediction of soccer player ranking using ensemble learning. *Neural Computing and Applications*, *31*(12), 9157–9174. https://doi.org/10.1007/S00521-019-04036-9

Mahbub, Md. K., Miah, Md. A. M., Islam, S. Md. S., Sorna, S., Hossain, S., & Biswas, M. (2021). Best Eleven Forecast for Bangladesh Cricket Team with Machine Learning Techniques. *2021 5th International Conference on Electrical Engineering and Information Communication Technology (ICEEICT)*, 1–6. https://doi.org/10.1109/ICEEICT53905.2021.9667862

Meddegoda, A., Kumara, B. T. G. S., & Kuhaneswaran, B. (2021). Neural Network Based Approach for Identifying Suitable Sport for Beginners. *2021 International Conference on Data Analytics for Business and Industry (ICDABI)*, 409–412. https://doi.org/10.1109/ICDABI53623.2021.9655882

Muller, S., Ghawi, R., & Pfeffer, J. (2020). Using Communication Networks to Predict Team Performance in Massively Multiplayer Online Games. *Proceedings of the 2020 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining, ASONAM 2020*, 353–360. https://doi.org/10.1109/ASONAM49781.2020.9381481

Mundhe, E., Jain, I., & Shah, S. (2021). Live Cricket Score Prediction Web Application using Machine Learning. *2021 International Conference on Smart Generation Computing, Communication and Networking (SMART GENCON)*, 1–6. https://doi.org/10.1109/SMARTGENCON51891.2021.9645855

Munir, F., Yani, Y. M., Nizmi, Y. E., & Suyastri, C. (2022). State of The Art Para-Diplomacy: A Systematic Mapping Studies and a Bibliometric Analysis VOS Viewer in Scopus Database. *Academic Journal of Interdisciplinary Studies*, *11*(2), 129–141. https://doi.org/10.36941/AJIS-2022-0040

Nsolo, E., Lambrix, P., & Carlsson, N. (2019). Player valuation in European football. *Lecture Notes in Computer Science (Including Subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, *11330 LNAI*, 42–54. https://doi.org/10.1007/978-3-030-17274-9_4

Pantzalis, V. C., & Tjortjis, C. (2020). Sports Analytics for Football League Table and Player Performance Prediction. *2020 11th International Conference on Information, Intelligence, Systems and Applications (IISA*, 1–8. https://doi.org/10.1109/IISA50023.2020.9284352

Phatak, A. A., Mehta, S., Wieland, F. G., Jamil, M., Connor, M., Bassek, M., & Memmert, D. (2022). Context is key: normalization as a novel approach to sport specific preprocessing of KPI's for match analysis in soccer. *Scientific Reports*, *12*(1). https://doi.org/10.1038/S41598-022-05089-Y

Pramanik, Md. A., Hasan Suzan, Md. M., Biswas, A. A., Rahman, M. Z., & Kalaiarasi, A. (2022). Performance Analysis of Classification Algorithms for Outcome Prediction of T20 Cricket Tournament Matches. *2022 International Conference on Computer Communication and Informatics (ICCCI)*, 01–07. https://doi.org/10.1109/ICCCI54379.2022.9740867

Rajesh, P., Bharadwaj, Alam, M., & Tahernezhadi, M. (2020). A Data Science Approach to Football Team Player Selection. *2020 IEEE International Conference on Electro Information Technology (EIT)*, 175–183. https://doi.org/10.1109/EIT48999.2020.9208331

Raju, M. A., Mia, M. S., Sayed, M. A., & Riaz Uddin, M. (2020). Predicting the Outcome of English Premier League Matches using Machine Learning. *2020 2nd International Conference on Sustainable Technologies for Industry 4.0, STI 2020*. https://doi.org/10.1109/STI50764.2020.9350327

Rudrapal, D., Boro, S., Srivastava, J., & Singh, S. (2020). *A Deep Learning Approach to Predict Football Match Result* (H. S. Behera, J. Nayak, B. Naik, & D. Pelusi, Eds.; Vol. 990, pp. 93–99). Springer Singapore. https://doi.org/10.1007/978-981-13-8676-3_9

Sharma, M., Monika, Kumar, N., & Kumar, P. (2021). Badminton match outcome prediction model using Naïve Bayes and Feature Weighting technique. *Journal of Ambient Intelligence and Humanized Computing*, *12*(8), 8441–8455. https://doi.org/10.1007/S12652-020-02578-8

Shi, F., Marchwica, P., Gamboa Higuera, J. C., Jamieson, M., Javan, M., & Siva, P. (2022). Self-Supervised Shape Alignment for Sports Field Registration. *2022 IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, 3768–3777. https://doi.org/10.1109/WACV51458.2022.00382

Smithies, T. D., Campbell, M. J., Ramsbottom, N., & Toth, A. J. (2021). A Random Forest approach to identify metrics that best predict match outcome and player ranking in the esport Rocket League. *Scientific Reports 2021 11:1*, *11*(1), 1–12. https://doi.org/10.1038/s41598-021-98879-9

Song, W., Xu, M., & Dolma, Y. (2019). Design and Implementation of Beach Sports Big Data Analysis System Based on Computer Technology. *Journal of Coastal Research*, *94*(sp1), 327. https://doi.org/10.2112/SI94-067.1

Subbaraj, S., Thiagarajan, R., & Rengaraj, M. (2020). Multi-objective league championship algorithm for real-time task scheduling. *Neural Computing and Applications Volume*, *32*, 5093–5104. https://link.springer.com/article/10.1007/s00521-018-3950-y

Talattinis, K., Kyriakides, G., Kapantai, E., & Stephanides, G. (2019). Forecasting Soccer Outcome Using Cost-Sensitive Models Oriented to Investment Opportunities. *International Journal of Computer Science in Sport*, *18*(1), 93–114. https://doi.org/10.2478/IJCSS-2019-0006

Theiner, J., Gritz, W., Muller-Budack, E., Rein, R., Memmert, D., & Ewerth, R. (2022). Extraction of Positional Player Data from Broadcast Soccer Videos. *2022 IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, 1463–1473. https://doi.org/10.1109/WACV51458.2022.00153

Thenmozhi, D., Mirunalini, P., Jaisakthi, S. M., Vasudevan, S., Veeramani Kannan, V., & Sagubar Sadiq, S. (2019). MoneyBall - Data Mining on Cricket Dataset. *2019 International Conference on Computational Intelligence in Data Science (ICCIDS)*, 1–5. https://doi.org/10.1109/ICCIDS.2019.8862065

Tyran, J., & Chomatek, L. (2021). Influence of outliers in MOBA games winner prediction. *Procedia Computer Science*, *192*, 1973–1981. https://doi.org/10.1016/j.procs.2021.08.203

Ur, J., Craner, M., & El-Hajj, R. (2019). What Makes a National Football League Team Successful? an Analysis of Play by Play Data. *2019 IEEE 10th Annual Ubiquitous Computing, Electronics & Mobile Communication Conference (UEMCON)*, 419–425. https://doi.org/10.1109/UEMCON47517.2019.8993072

Victor, B., Nibali, A., He, Z., & Carey, D. L. (2021). Enhancing trajectory prediction using sparse outputs: application to team sports. *Neural Computing and Applications*, *33*(18), 11951–11962. https://doi.org/10.1007/S00521-021-05888-W

Wang, D. (2022). Clustering and Evolutionary System Analysis of Data Mining Algorithms in the Field of Football. *2022 IEEE Asia-Pacific Conference on Image Processing, Electronics and Computers (IPEC)*, 1341–1344. https://doi.org/10.1109/IPEC54454.2022.9777487

Wang, H., Dong, C., & Fu, Y. (2021). Optimization analysis of sport pattern driven by machine learning and multi-agent. *Neural Computing and Applications*, *33*(4), 1067–1077. https://doi.org/10.1007/S00521-020-05022-2

Wang, T. (2022). Sports training auxiliary decision support system based on neural network algorithm. *Neural Computing and Applications*. https://doi.org/10.1007/S00521-022-07137-0

Wang, W.-Y., Chan, T.-F., Yang, H.-K., Wang, C.-C., Fan, Y.-C., & Peng, W.-C. (2021). Exploring the Long Short-Term Dependencies to Infer Shot Influence in Badminton Matches. *2021 IEEE International Conference on Data Mining (ICDM)*, 1397–1402. https://doi.org/10.1109/ICDM51629.2021.00178

Weeraddana, N., & Premaratne, S. (2021). Unique approach for cricket match outcome prediction using Xgboost algorithms. *Journal of Theoretical and Applied Information Technology*, *99*(9), 2162–2173.

Yao, A. (2019). Comparing Neural and Regression Models to Predict NBA Team Records. *IOS Press Ebooks*, *30*, 421–428. https://doi.org/10.3233/FAIA190206

Yin, Z., & Cui, W. (2021). Outlier data mining model for sports data analysis. *Journal of Intelligent & Fuzzy Systems*, *40*(2), 2733–2742. https://doi.org/10.3233/JIFS-189315

Yuan, C., Yang, Y., & Liu, Y. (2021). Sports decision-making model based on data mining and neural network. *Neural Computing and Applications*, *33*(9), 3911–3924. https://doi.org/10.1007/S00521-020-05445-X

Zhang, L., & Li, N. (2022). Material analysis and big data monitoring of sports training equipment based on machine learning algorithm. *Neural Computing and Applications*, *34*(4), 2749–2763. https://doi.org/10.1007/S00521-021-05852-8

Zhang, S. (2016). The prediction of the sport performance based on the IGSA. *RISTI - Revista Iberica de Sistemas e Tecnologias de Informacao*, *E11*(11), 313–321.

Zhang, Y., Hou, X., & Xu, S. (2022). Neural network in sports cluster analysis. *Neural Computing and Applications*, *34*, 3301–3309. https://link.springer.com/article/10.1007/s00521-020-05585-0