2018, *Retos, 34,* xxx-xxx
© Copyright: Federación Española de Asociaciones de Docentes de Educación Física (FEADEF) *ISSN: Edición impresa: 1579-1726. Edición Web: 1988-2041 (www.retos.org)*

Aplicación de métodos de aprendizaje automático en el análisis y la predicción de resultados deportivos

Application of machine learning methods for analyzing and predicting sport outcomes

César Soto-Valero
Universidad Central «Marta Abreu» de Las Villas (Cuba)

Resumen. El aprendizaje automático es una herramienta muy útil para el análisis de la gran cantidad de datos que se manejan en el deporte moderno. En la actualidad, este tipo de métodos se han convertido en un ámbito de investigación con grandes perspectivas de aplicación. En el presente trabajo se realiza una revisión del estado del arte sobre los principales métodos de aprendizaje automático empleados en el análisis cuantitativo de datos deportivos. En particular, se plantean las posibilidades que ofrecen estos métodos para dar solución a dos de los problemas más complejos en el deporte: el análisis del desempeño deportivo y la predicción de resultados competitivos. Además, se estudian las ventajas que ofrece el uso del aprendizaje automático para el análisis de los mercados deportivos y se propone una metodología para su aplicación como parte del proceso de toma de decisiones en el caso de las apuestas deportivas. La aplicación de esta teoría contribuye al desarrollo del análisis de datos deportivos, lo cual trae consigo una mejor comprensión del funcionamiento de las diferentes disciplinas deportivas y potencia el desarrollo técnico-táctico en el deporte.
Palabras clave: aprendizaje automático, datos deportivos, análisis cuantitativo, desempeño deportivo, predicción de resultados competitivos

Abstract. Machine learning is a very useful tool for studying the vast amount of data that is being constantly generated in modern sport. At present, machine learning has become a field of research with a wide range of perspectives and applications. In this work, we perform a review on the machine learning methods used for analyzing quantitative sport data. In particular, we focus on the advantages offered by machine learning methods to solve two of the most complex problems in sports: the analysis of performance and the prediction of competitive outcomes. We study the machine learning methods used for analyzing sport markets. Furthermore, we propose a methodology for the application of machine learning methods for the decision making process during sports betting. The application of this theory contributes significantly to the development of the analysis of sports data, which ensures a better understanding of the different sport disciplines and enhances technical-tactical development in sport.
Key words: machine learning, sport data sets, quantitative analysis, sport performance, game outcome prediction.

Introducción

Los logros del deporte moderno han estado determinados por diversos factores. No cabe duda que uno de los más importantes ha sido la aplicación efectiva de los avances de la ciencia y la tecnología en las distintas disciplinas deportivas. En la actualidad, los problemas existentes en el deporte se abordan utilizando métodos de investigación científica. Este proceder ha ido descartando cada vez más la empírea y la espontaneidad en las investigaciones. Hoy existe una mayor comprensión de que no es suficiente conocer la realidad observable con vista a solucionar los problemas prácticos de la actividad física y deportiva, sino que se hace necesario además describir, comprender, interpretar, explicar teóricamente o predecir para transformar esa realidad; todo lo cual requiere de la utilización de métodos y medios especiales de conocimiento.

En el contexto deportivo actual, el proceso de toma de decisiones es de vital importancia para la obtención de buenos resultados competitivos. La trascendencia de dichas decisiones está directamente relacionada con su complejidad, por lo que casi siempre se requiere del conocimiento de expertos o especialistas a la hora de elaborar las diferentes estrategias deportivas. Han sido muchos los avances obtenidos en el desarrollo y aplicación de nuevos métodos para el análisis cuantitativo de datos deportivos, sobre todo utilizando técnicas estadísticas. Esta es un área de la ciencia en constante desarrollo debido a que enlaza varios aspectos claves del análisis técnico-táctico en el deporte.

El análisis del rendimiento competitivo ha ganado importancia en las últimas décadas, siendo esta la principal forma para medir y evaluar la actuación de los deportistas. La Metodología Observacional (Aguera, Blanco, Mendo, & Losada, 2015) favorece en buena medida el estudio de las observaciones sistemáticas obtenidas, fundamentalmente mediante el análisis de datos jugada a jugada. En la actualidad, el principal problema no radica en la obtención de los datos deportivos, sino en cómo obtener información útil a partir de los mismos (Pueo & Jimenez-Olmedo).

La gestión automática de la información y la inteligencia artificial son disciplinas que están estrechamente ligadas a los avances en las tecnologías de la informática y las comunicaciones. Una de las ramas de la inteligencia artificial que ha alcanzado mayor popularidad en estos tiempos es el aprendizaje automático. Este constituye un campo de investigación muy joven en el ámbito deportivo (Baraniuk, 2015), por los que sus potencialidades de aplicación requieren de un estudio más profundo y abarcador en dicho contexto.

El aprendizaje automático se ocupa del desarrollo de sistemas computacionales diseñados con el propósito de aprender y adaptarse a partir de los datos, sin la necesidad de introducir explícitamente el nuevo conocimiento adquirido. El auge del aprendizaje automático se debe a su gran aplicabilidad, puesto que prácticamente todo conocimiento es susceptible de ser aprendido e interpretado. Resulta ineludible el hecho de que los métodos de aprendizaje automático constituyen una herramienta de gran utilidad en diversas ramas de la ciencia. Por ejemplo, en bioinformática para el diagnóstico médico o la clasificación de secuencias de ADN, en economía para el análisis del mercado de valores o la detección de fraudes en tarjetas de crédito, o en el reconocimiento del habla y del lenguaje escrito en campos como la teoría de juegos y la robótica, entre otros (Bishop, 2006). En este sentido, las ciencias del deporte resultan ser un campo de aplicación relativamente novedoso (Alderson, 2015).

El aprendizaje automático ofrece varias ventajas para el análisis de los datos deportivos respecto a los métodos estadísticos tradicionales. Su aplicación ha sido posible fundamentalmente gracias a los avances en las tecnologías de la informática y las comunicaciones, principalmente en las mejoras en el proceso de adquisición y manejo de los datos (Barshan & Yüksek, 2014; Hua, Lai, You, & Cheng, 2015). Los métodos de aprendizaje automático han demostrado su eficacia en tareas tales como la evaluación del rendimiento deportivo o la predicción de resultados competitivos en varios deportes colectivos tales como el fútbol o el baloncesto (Schumaker, Solieman, & Chen, 2010; Van Haaren, Ben Shitrit, Davis, & Fua, 2016).

El presente trabajo tiene como objetivo exponer los aspectos fundamentales del aprendizaje automático enfocado al análisis de datos deportivos. En especial, se abordan dos problemas fundamentales en este campo: el análisis del desempeño deportivo y la predicción de resultados competitivos. Además, se describen los principales métodos utilizados en la literatura, así como los pasos necesarios para llevar a cabo la búsqueda de conocimiento útil con el propósito de favorecer la toma de decisiones en el deporte.

Fecha recepción: 30-05-17. Fecha de aceptación: 28-08-17
César Soto Valero
cesarsotovalero@gmail.com

Estado actual del tema

En esta sección se presentan los fundamentos del análisis cuantitativo de datos deportivos. Se realiza una introducción a los principales conceptos del aprendizaje automático como una alternativa novedosa respecto a las técnicas estadísticas empleadas tradicionales en las ciencias del deporte. Consecuentemente, se exponen las ventajas que presenta el uso de técnicas de aprendizaje automático en el contexto deportivo, sobre todo como apoyo en la toma de decisiones.

Análisis cuantitativo de datos deportivos

En la actualidad, las ciencias del deporte han despertado un gran interés en toda la comunidad científica internacional. En particular, la ciencia de la computación aplicada al deporte es un área de investigación multidisciplinaria y relativamente novedosa. Su objetivo consiste en combinar los aspectos teóricos y prácticos, así como los métodos pertenecientes al área de la informática y la actividad física, para impulsar los avances en la teoría y práctica del deporte moderno (Link & Lames, 2009).

El estudio de los aspectos técnico-tácticos constituye la base del análisis del rendimiento deportivo. La Figura 1 muestra la interrelación que se establece entre estas disciplinas y el análisis de datos deportivos. Como se puede apreciar, el objetivo general es la mejora del rendimiento competitivo, el cual se sustenta en un adecuado proceso de toma de decisiones. En este sentido, el análisis cuantitativo de datos deportivos constituye un eje fundamental, ya que enlaza varios aspectos claves del análisis técnico-táctico, tales como el desempeño competitivo o la estrategia competitiva.

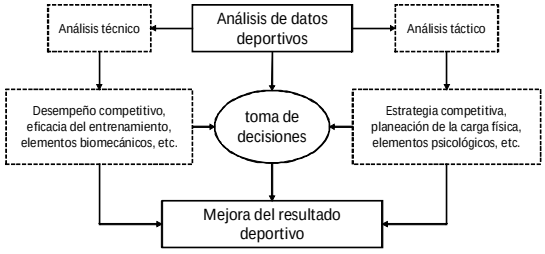

Figura 1: Aspectos del análisis técnico-táctico como sustento a las toma de decisiones en el deporte.

La cantidad de datos disponibles en casi todos los ámbitos deportivos ha crecido de forma vertiginosa en los últimos años. Dichos datos pueden ser obtenidos de varias maneras: a partir de mediciones individuales de los atletas en juegos y eventos, producto del trabajo e investigación de los propios entrenadores, o a través de la búsqueda y el análisis llevado a cabo por los cazatalentos. En la actualidad, el principal reto no radica en cómo obtener los datos, sino en identificar cuál es la información más relevante y cómo generar conocimiento útil a partir de la misma (Fister Jr, Ljubiè, Suganthan, Perc, & Fister, 2015).

El objetivo final de cualquier deporte consiste en alcanzar la victoria frente al contrario. El primer problema presente en la preparación para la obtención de este resultado reside en identificar correctamente las métricas de desempeño, para poder mejorarlas progresivamente. Muchas de las medidas utilizadas actualmente pueden ser irrelevantes en la práctica, o peor aún, pueden resultar inadecuadas en diversas circunstancias y arrojar resultados erróneos. Otro aspecto importante consiste en identificar patrones relevantes en los datos colectados. Por ejemplo, la búsqueda de tendencias y patrones de comportamiento de determinados contrarios, la identificación del comienzo de una posible temporada de bajo rendimiento a través del monitoreo de medidas de desempeño o la realización de predicciones deportivas utilizando datos históricos.

Por medio de la búsqueda de información y el manejo del conocimiento inherente a las mediciones deportivas los directores de equipos y analistas deportivos tienen la posibilidad de asegurar una importante ventaja competitiva frente a sus rivales. Dicho conocimiento, una vez validado en la práctica, puede ser aplicado en los más disímiles niveles de toda la organización deportiva, desde la mejora del desempeño individual de los jugadores (Hua et al., 2015), elaborando hipótesis y probándolas mediante el uso de pruebas estadísticas (Jeff & John, 2011) utilizando técnicas de decisión en tiempo real, en la predicción del desempeño y la identificación de talentos o para determinar cuál jugador tiene la mayor importancia en el equipo (B. S. Baumer, Jensen, & Matthews, 2015).

Antes de la utilización de técnicas estadísticas para el análisis de datos, las organizaciones deportivas dependían casi exclusivamente de la experiencia humana. Se debía asumir el supuesto de que los expertos (entrenadores, directores de equipos, cazatalentos, etc.) son realmente capaces de convertir los datos de los que disponen en conocimiento útil y veraz. Sin embargo, con el significativo incremento en la cantidad de datos colectados se ha hecho evidente la necesidad de encontrar métodos que, en la práctica, arrojen una mayor cantidad de información útil de forma más eficiente. El principal propósito en todos los casos consiste en minimizar el margen de error de las decisiones que se tomen atendiendo a las características de cada deporte específico.

El uso de la computación ha venido a ser una herramienta fundamental en todos los dominios de aplicación de las ciencias del deporte, especialmente en el análisis cuantitativo de datos deportivos. En general, se pueden identificar las siguientes áreas de investigación en este campo: (1) adquisición y pre-procesamiento de los datos; (2) representación de la información y análisis descriptivo; (3) modelación de bases de datos y sistemas expertos y (4) simulación.

Las técnicas estadísticas se han usado tradicionalmente para resolver los problemas inherentes al manejo de datos deportivos. Estas se aplican sobre todo para realizar comparaciones entre poblaciones que han sido objeto de determinados planes de entrenamiento o para distinguir patrones significativos en eventos deportivos (Soto-Valero & González-Castellanos, 2015). Las técnicas tradicionales han resultado ser sumamente útiles, permitiendo a los investigadores y directores de equipos deportivos evaluar hipótesis y realizar predicciones a partir de datos de juegos reales. Sin embargo, la estadística por sí misma no es capaz de explicar relaciones más complejas entre los datos, lo cual es el propósito de la minería de datos y en especial de los métodos de aprendizaje automático (Piatetsky, 2016). En la siguiente sección se abordan con más detalle los elementos fundamentales de este novedoso enfoque.

Aprendizaje automático de datos deportivos

Desde la aparición de las computadoras estas han sido capaces de resolver problemas muy complejos para el hombre, sin embargo, aún no poseen la habilidad de aprender por sí solas. A pesar de esto, el desarrollo de la inteligencia artificial ha propuesto una gran cantidad de algoritmos que intentan imitar esta habilidad. Estos algoritmos han demostrado ser especialmente efectivos para ciertos tipos de problemas (González-Ruiz, Gómez-Gallego, Pastrana-Brincones, & Hernández-Mendo, 2015; Hagenbuchner, Cliff, Trost, Van Tuc, & Peoples, 2015).

El aprendizaje automático es un campo multidisciplinario cuyo objetivo es desarrollar programas de computadora que mejoren su funcionamiento en ciertas tareas a partir de la experiencia (Mitchell, 1997). La minería de datos ha contribuido al desarrollo del aprendizaje automático ya que sus métodos han sido ampliamente utilizados en el descubrimiento de información valiosa a partir de datos almacenados (Witten, Frank, & Hall, 2011). Con frecuencia el campo de aplicación del aprendizaje automático se solapa con el de la estadística, ya que las dos disciplinas se basan en el análisis de datos, por lo que resulta difícil establecer una línea divisoria entre ambas. No obstante, el aprendizaje automático se centra más en el estudio de la complejidad computacional de los problemas y la descripción de los resultados obtenidos. Muchos problemas son de clase NP-Hard (Shalev-Shwartz & Ben-David, 2014), por lo que gran parte de la investigación realizada en esta rama de la ciencia se enfoca en el diseño de soluciones eficientes para esos problemas.

De forma más concreta, se trata de crear algoritmos capaces de generalizar comportamientos a partir de información suministrada en forma de instancias o ejemplos de aprendizaje. Tales ejemplos sirven

como entrenamiento, para que luego el algoritmo pueda enfrentarse a nuevos datos. Los algoritmos de aprendizaje automático construyen un modelo de aprendizaje a partir de los ejemplos y lo usan para hacer predicciones o describir patrones, en lugar de seguir instrucciones estáticas estrictas como cualquier otro programa de computadora.

En este caso, las instancias están conformadas por variables o atributos, los cuales pueden tomar tanto valores numéricos como nominales. Esta representación de instancia es una abstracción de los objetos, pudiéndose ignorar otras características que no son representadas por los atributos.

Existen varias formas de adquirir el conocimiento necesario, una puede ser directamente a partir del humano, o a partir de problemas resueltos previamente. Los datos que se le proporcionan como entrada al sistema permiten que el algoritmo de aprendizaje pueda extraer de ellos la información necesaria para enfrentarse a nuevos datos y realizar la función para la cual fue diseñado. Según el resultado que se desea obtener a través del sistema, existen un grupo de categorías en las cuales se engloban las tareas de aprendizaje automático. El Cuadro 1 describe algunas de las más importantes.

Cuadro 1
Principales tareas del aprendizaje automático.

Método	Descripción
Clasificación	La entrada es dividida en dos o más clases. El sistema debe producir un modelo capaz de asignarle a una nueva entrada una o más de estas clases.
Regresión	Es también una tarea supervisada, similar a la clasificación, pero la salida obtenida es un valor numérico continuo en lugar de uno discreto.
Agrupamiento	El conjunto de entrada es dividido en grupos. A diferencia de la clasificación los grupos no son conocidos de antemano.
Reglas de asociación	Se trata de encontrar reglas, generalmente del tipo condicional, que relacionen a los datos.

Los métodos de clasificación y regresión son ejemplos típicos del paradigma de aprendizaje supervisado, mientras el agrupamiento y descubrimiento de reglas pertenecen al aprendizaje no supervisado (Shalev-Shwartz & Ben-David, 2014). En el primero, cada una de las instancias de entrenamiento está asociada con una salida, ya sea numérica o nominal, denominada etiqueta de clase; mientras que en el segundo no tiene en cuenta la clase con que han sido etiquetadas o no las instancias.

La minería de datos, y en especial los métodos de aprendizaje automático, difieren de las técnicas estadísticas ya que tienen la capacidad de hacer generalizaciones mucho más complejas a partir de situaciones, lo cual permite llevar a cabo predicciones más precisas y revelar relaciones ocultas en los datos. Esta forma de obtención del conocimiento es sumamente útil, ya que puede ser usada para justificar la toma de decisiones de expertos deportivos en determinado momento del juego, o también puede ser usada por atletas de manera independientemente como método de auto evaluación con independencia del criterio especializado (Haghighat, Rastegari, & Nourafza, 2013).

Los métodos de aprendizaje automático son un complemento a las herramientas del análisis estadístico tradicional en el deporte. Esto posibilita, entre otras cosas, dar respuesta a un gran número de interrogantes. La computación se ha convertido en una parte importante de este proceso de decisión y análisis, los entrenadores de primer nivel en el mundo usan actualmente diversas técnicas de aprendizaje automático y simulación para la planificación de estrategias completas antes y durante las temporadas competitivas (Min, Kim, Choe, Eom, & McKay, 2008; O'Reilly & Knight, 2007).

El uso de métodos de aprendizaje automático en este contexto ofrece varias ventajas, una de ellas es que se evita la influencia de factores humanos subjetivos. Esto se debe a que las decisiones pueden ser tomadas sin prejuicios (De Marchi, 2011). Un ejemplo podría ser un director de equipo que se sienta atraído especialmente hacia los atributos de desempeño de un jugador determinado, ignorando buena parte de sus debilidades. Mediante la eliminación de este tipo de sesgos humanos en el proceso de toma de decisiones se hace posible dirigir de forma más efectiva, lo cual redunda en una mejora en la organización y el rendimiento competitivo.

El Cuadro 1 muestra una selección de trabajos publicados en la literatura que son representativos de la aplicación del aprendizaje automático en el contexto deportivo. Como se observa, el campo de aplicación de los diferentes métodos abarca un variado número de deportes y de técnicas, siendo las tareas de clasificación y regresión las más empleadas. En general, podemos decir que sus principales usos han sido en el análisis del desempeño deportivo, la predicción de resultados competitivos en deportes tanto colectivos como individuales y para estudios macro-económicos y de mercado.

Cuadro 2
Ejemplos de aplicación de métodos de aprendizaje automático para la solución de problemas en el contexto deportivo (Fuente: SCOPUS, WEB OF SCIENCE y SCIELO).

Método	Autor/es	Deporte
Clasificación	(Hamilton et al., 2014)	Béisbol
	(Davoodi & Khanteymoori, 2010)	Equitación
	(Delen, Cogdell, & Kasap, 2012)	Bolos
	(Rama Iyer & Sharda, 2009)	Cricket
Regresión	(Shao, 2009)	Gimnasia
	(Demens, 2015)	Hockey
	(Lock & Nettleton, 2014)	Fútbol
	(Jelinek, Kelarev, Robinson, Stranieri, & Cornforth, 2014)	Fútbol
Agrupamiento	(Ofoghi, Zeleznikow, MacMahon, & Dwyer, 2010)	Ciclismo
	(Soto-Valero, 2017)	Fútbol
Reglas de asociación	(Sun, Yu, & Zhao, 2010)	Tenis
	(Soto-Valero, Pérez-Morales, González-Castellanos, & de la Celda Brovkina, 2016)	Polo acuático

Materiales y métodos

En esta sección se lleva a cabo una descripción detallada sobre dos de las principales aplicaciones que ha tenido el aprendizaje automático de datos deportivos: el análisis del desempeño deportivo y la predicción de resultados competitivos. Se propone un sistema general para el monitoreo del desempeño deportivo, así como un esquema para la conformación de una instancia de aprendizaje orientada a la predicción de resultados competitivo. Además, por el auge que ha tenido en el contexto actual, se ofrece una introducción al estudio de los mercados deportivos utilizando métodos de aprendizaje automático, haciendo especial énfasis en el mercado de apuestas deportivas.

Análisis del desempeño deportivo

Uno de los objetivos fundamentales en el deporte, especialmente en los de alto rendimiento, es la mejora del desempeño competitivo para la mejora de los resultados deportivos. Desde el punto de vista del análisis cuantitativo de los datos, el principal propósito es generar información útil a partir de las mediciones realizadas a los deportistas de forma periódica en entrenamientos y competencias oficiales.

La Figura 3 presenta un esquema general para la creación de un sistema que posibilite el monitoreo de los parámetros técnico-tácticos en el deporte a partir del uso de las nuevas tecnologías de la informática y las comunicaciones. Como se puede apreciar, la interacción directa entre los atletas, analistas deportivos y entrenadores es una parte esencial del proceso de toma de decisiones.

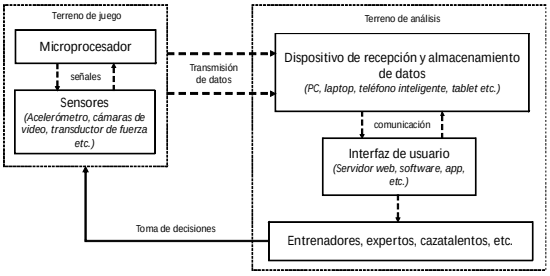

Figura 2: Esquema general para la implementación de un sistema de monitoreo del desempeño deportivo.

Los métodos de aprendizaje automático han resultado ser sumamente útiles para el procesamiento y análisis de los datos colectados en el terreno de juego. Los dispositivos creados para la medición y el control de los atletas posibilitan un reflujo de información constante. Estas técnicas han permitido la identificación de aspectos esenciales del desempeño competitivo, generando todo un conjunto de nuevas métricas para su comparación, así como planes de entrenamiento más eficientes en varios deportes. Por ejemplo, en Edelmann-Nusser, Hohmann, and Henneberg (2002) proponen un modelo basado en redes neuronales

artificiales para el análisis del desempeño de nadadores olímpicos, mientras que Morgan, Williams, and Barnes (2013) utilizan árboles de decisión en la identificación de atributos importantes durante la interacción uno contra uno en el hockey.

Predicción de resultados competitivos

Una de las principales aplicaciones que han tenido los métodos de aprendizaje automático en el contexto deportivo, y en especial las técnicas de clasificación y regresión, es en la predicción de resultados competitivos (Schumaker et al., 2010). El objetivo de la predicción en este caso consiste en la obtención de alguna ventaja, ya sea desde el punto de vista de la competencia o también financiera, frente los rivales. En la actualidad, las apuestas deportivas y el mercado del deporte en general han despertado un gran interés en este ámbito, sobre todo en el caso de los deportes colectivos profesionales.

La predicción en el deporte tiene características especiales que la distinguen de otras áreas de investigación. Por ejemplo, en el deporte cabe señalar su carácter eminentemente competitivo, así como el dinamismo presente, el cual requiere de una actualización constante de los modelos de predicción empleados. Esto hace posible en muchos casos generar resultados no solo a partir de datos propios de los jugadores objeto de estudio, sino también sobre la base de otros elementos y variables del contrario (Knottenbelt, Spanias, & Madurska, 2012).

Existen marcadas diferencias entre la predicción del desempeño en deportes individuales y colectivos, siendo estos últimos los más estudiados desde el punto de vista predictivo por ser los más complejos y en los que intervienen un mayor número de variables. La predicción en deportes individuales tiene como punto de partida el análisis del rendimiento individual del deportista. Los modelos para la selección de talentos, el ajuste de variables para la evaluación de jugadores y la prevención de lesiones son algunos ejemplos de aplicación (Rama Iyer & Sharda, 2009).

Por otro lado, los deportes colectivos se caracterizan por la interacción entre los miembros del equipo. En este caso, la cuantificación de los resultados individuales debe tener en cuenta el aporte realizado por cada jugador a la victoria de todo el equipo. La planificación de las estrategias tiene una base en la predicción del resultado, así como el análisis del contrario. De entre todos los deportes colectivos, el fútbol, el baloncesto y el béisbol han sido objeto de un mayor número de investigaciones utilizando técnicas de aprendizaje automático (Yadav, Sharma, Gautam, Bathla, & Jindal, 2017).

La Figura 4 propone la conformación de una instancia para la predicción de resultados competitivos utilizando aprendizaje supervisado. Dicha instancia corresponde a un juego entre dos equipos, el local A y el visitante B. Los atributos del encuentro están dados por las variables generales o IDs (ej. fecha, terreno de juego, clima, etc.) y las estadísticas numéricas de desempeño previas al encuentro correspondiente a ambos equipos (modeladas como una resta de variables de rendimiento del equipo A respecto al B). El resultado del encuentro está determinado por la etiqueta de la instancia y puede ser visto como un problema de clasificación binaria atendiendo a la victoria o derrota del equipo local o como un problema de regresión en el que se tiene en cuenta el margen de diferencia del resultado final del encuentro.

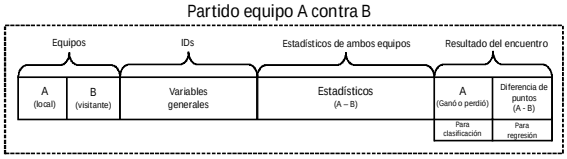

Figura 3: Representación de una instancia para la predicción del resultado de un juego entre los equipos A y B utilizando aprendizaje automático supervisado.

Desde el punto de vista predictivo, la ventaja de jugar en casa o HA ha recibido una atención especial. Esta está determinada por la asociación entre el porcentaje de victorias conseguidas por el equipo local respecto al equipo visitante (Courneya & Carron, 1992; Pollard, 1986). Las causas que podrían explicar el fenómeno son inconclusas y tienen una explicación multifactorial (Pollard & Pollard, 2005). En deportes de equipo, se ha reportado que el HA puede ser de alrededor del 60% (Jamieson, 2010), habiéndose particularizado estudios más rigurosos en varios deportes colectivos específicos tales como el fútbol (Gomez, Pollard, & Luis-Pascual, 2011; Pic & Castellano, 2016).

Varios han sido los sistemas computacionales desarrollados con fines predictivos, cada uno una con mayor o menor complejidad y fiabilidad (Bustamante & Burillo, 2016). Algunos de estos sistemas están basados en métodos de aprendizaje automático, entre ellos cabe destacar: *Advanced Scout* (Bhandari et al.), sistema desarrollado por IBM para automatizar todo el proceso de minería de datos en juegos de baloncesto de la NBA; *Digital Scout*, herramienta de software muy utilizada en la realización de una amplia variedad análisis estadísticos y evaluaciones de jugadas en deportes como fútbol, voleibol y baloncesto; e *Inside Edge*, el cual es un sistema pionero en el estudio y recopilación de datos de béisbol, siendo actualmente un líder mundial en exploración y análisis de datos en este deporte.

Estudio de los mercados deportivos

En la actualidad, el deporte ha pasado de ser una simple manifestación social, destinada a la contemplación y práctica de actividades recreativas para la búsqueda de cierto entretenimiento o satisfacción personal, a ser considerado como un bien cuya producción, consumo, financiación y gestión responde a criterios de racionalidad económica bien definidos (B. Baumer & Zimbalist, 2014). Por un lado, el deporte profesional ha abierto a la economía nuevos y rentables mercados, así como distintas oportunidades de empleo hasta hace poco tiempo desconocidas. Por otro, la economía ha dotado al deporte de una estructura de pensamiento diferente a la hora de tomar decisiones, valorar relaciones institucionales o evaluar sus consecuencias en el plano material. Se ha pasado de esta manera de una situación caracterizada por una tradicional ausencia de lo económico, a otra en la que las relaciones ideológicas y de valor, de cooperación, de transferencia o de regulación entre el deporte y la economía se han ido haciendo cada vez más estrechas.

Una de las características de los mercados deportivos que los distinguen de otras ramas de la economía es que en este caso las empresas, bien sean los clubes en los deportes de equipo o los deportistas en el caso de los deportes individuales, necesitan de la competencia para maximizar sus beneficios, no pudiendo aspirar a monopolizar el mercado. Esto refuerza el hecho de que el deporte sea considerado ante todo como espectáculo competitivo. En este sentido, el mercado ha propiciado un lento pero continuo avance de las distintas disciplinas deportivas.

Desde el propio surgimiento del deporte el interés por las apuestas ha resultado ser una parte importante la cual ha potenciado significativamente su desarrollo (Woodland & Woodland, 1994). El mercado electrónico de apuestas deportivas ha despertado un gran auge en los últimos años. En el caso de algunos deportes colectivos de Estados Unidos y Europa tales como el fútbol, béisbol o baloncesto, las ganancias estimadas resultan ser billonarias (Paul & Weinbach, 2009; Sauer, Waller, & Hakes, 2010; Spann & Skiera, 2009). El Cuadro 3 describe los tres modelos fundamentales empleados por las agencias que manejan apuestas deportivas.

Cuadro 3
Descripción de los tres modelos fundamentales de apuestas deportivas.

Modelo	**Descripción**
Money-line	El resultado de la apuesta está determinado exclusivamente por el equipo ganador.
Over-under	La apuesta tiene en cuenta la diferencia de los puntos obtenidos entre los equipos rivales.
In-line	La apuesta se lleva a cabo durante el transcurso de la competición.

En el contexto del aprendizaje automático las apuestas deportivas se han abordado como un problema de predicción de resultados competitivos. La modelación de este problema se realiza, por lo general, mediante el uso de métodos de aprendizaje supervisado. El modelo *Money-line* ha sido el más estudiado debido a que constituye un problema clásico de clasificación binaria. Por otro lado, el modelo *Over-under* se ha abortado usando métodos de regresión. La Figura 4 presenta un

esquema general de aplicación de métodos de aprendizaje automático para la toma de decisiones en estos dos modelos de apuestas deportivas.

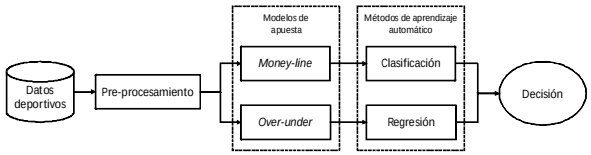

Figura 4: Esquema general para la utilización de métodos de aprendizaje automático durante la toma de decisiones en apuestas deportivas de tipo *Money-line* y *Over-under*.

El modelo de apuestas *In-line* representa un desafío mayor para la predicción. Esto se debe fundamentalmente a la necesidad de una constante actualización de los datos de aprendizaje, lo cual lleva aparejado un cambio de las decisiones tomadas a cada momento. Este modelo de apuestas ha sido abordado utilizado fundamentalmente técnicas de simulación y aprendizaje reforzado.

Discusión

En la actualidad, no cabe duda que tanto los directores de equipos profesionales, entrenadores o cazatalentos emplean activamente los últimos avances en materia de tecnologías de la informática y comunicaciones. Esto ha provocado un incremento sin precedentes de la cantidad de datos generados durante las competencias o entrenamientos deportivos, así como de un mayor interés y motivación en el desarrollo de modelos que permitan interpretar dichos datos y ofrecer información valiosa para cada deporte específico.

El análisis cuantitativo de datos deportivos constituye un elemento esencial para el desarrollo y mejora de los resultados competitivos, constituyendo una herramienta imprescindible en el proceso de toma de decisiones. El uso de métodos de aprendizaje automático en el contexto deportivo brinda importantes ventajas en comparación con las técnicas estadísticas utilizadas tradicionalmente.

Por la relevancia del tema, en este trabajo se llevó a cabo un estudio del análisis de los mercados deportivos desde el punto de vista cuantitativo. Particularmente, se describieron los modelos de apuestas deportivas más populares. En este sentido, se presentó un esquema general que posibilita el empleo de métodos de aprendizaje automático en la toma de decisiones para los tipos de apuestas *Money-line* y *Over-under*.

A partir del estudio realizado y como resultado de esta investigación se recomienda el empleo de los siguientes pasos generales para llevar a cabo el análisis cuantitativo de datos deportivos usando métodos de aprendizaje automático: (1) identificación, recopilación y almacenamiento de los datos relevantes en el terreno de juego; (2) preprocesamiento de la información y creación de las instancias de datos para el aprendizaje; (3) identificación del método de aprendizaje idóneo para la resolución del problema planteado; (4) ejecución del algoritmo propuesto y creación del modelo; (5) validación experimental de los resultados obtenidos; (6) toma de decisiones en el terreno de juego. La Figura 5 resume gráficamente el modelo propuesto.

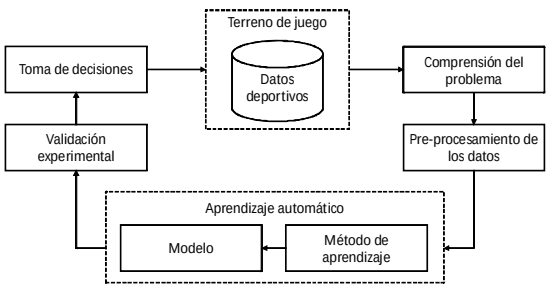

Figura 5: Modelo general para el análisis de datos deportivos utilizando métodos de aprendizaje automático.

omo se puede observar en la Figura 5, el análisis de datos deportivos utilizando métodos de aprendizaje automático es un proceso cíclico. Se parte de un conjunto de datos recopilados, por general en el terreno de juego, y la definición de un problema. A continuación, se lleva a cabo un pre-procesamiento de los datos con el propósito de modelar problema y seleccionar un método de aprendizaje automático adecuado. Una vez definido el modelo, es preciso validarlo experimentalmente antes de proceder a la toma de decisiones en el terreno de juego. De esta manera, se produce un constante reflujo de información, modelación y análisis de los datos el cual es preciso actualizar de manera constante.

Conclusiones

El análisis cuantitativo de datos deportivos constituye un elemento esencial para el desarrollo y mejora del rendimiento competitivo en el deporte. El uso de métodos de aprendizaje automático en el contexto deportivo brinda importantes ventajas para su análisis cuantitativo, en comparación con las técnicas estadísticas tradicionales. En este trabajo se realizó una descripción de los principales aspectos que sustentan la aplicación de técnicas de aprendizaje automático en el ámbito deportivo. En particular, se abordaron dos de los problemas más representativos de este campo: el análisis del desempeño deportivo y la predicción de resultados competitivos. Respecto al primero, se presentó un esquema general para la implementación de un sistema para el monitoreo de los parámetros técnico-tácticos en el deporte. Respecto al segundo, se describieron los modelos fundamentales de aprendizaje automático que han sido propuestos y se conformó la estructura de una instancia de aprendizaje supervisado orientada a la predicción de resultados competitivos. Por otro lado, se realizó un estudio sobre el comportamiento de los mercados deportivos y se identificaron las principales ventajas del uso de técnicas de aprendizaje automático en esta área. Cada uno de estos resultados tiene gran importancia para la comunidad científica, y en especial para los analistas deportivos que manejan gran cantidad de datos y necesitan de un adecuado basamento teórico a la hora de presentar y justificar sus resultados prácticos.

Referencias

Aguera, M. T., Blanco, A., Mendo, A. H., & Losada, J. L. L. (2015). Técnicas de análisis en estudios observacionales en ciencias del deporte. *Cuadernos de psicología del deporte, 15*(1), 13-30.

Alderson, J. (2015). A markerless motion capture technique for sport performance analysis and injury prevention: Toward a big data, machine learning future. *Journal of Science and Medicine in Sport, 19*, e79. doi: 10.1016/j.jsams.2015.12.192

Baraniuk, C. (2015). Rise of the AI sports coach. *New Scientist, 227*(3035). doi: 10.1016/S0262-4079(15)31025-3

Barshan, B., & Yüksek, M. C. (2014). Recognizing daily and sports activities in two open source machine learning environments using body-worn sensor units. *The Computer Journal, 57*(11), 1649-1667. doi: 10.1093/comjnl/bxt075

Baumer, B., & Zimbalist, A. (2014). Quantifying Market Inefficiencies in the Baseball Players' Market. *Eastern Economic Journal, 40*(4), 488-498. doi: 10.1057/eej.2013.43

Baumer, B. S., Jensen, S. T., & Matthews, G. J. (2015). openWAR: An open source system for evaluating overall player performance in major league baseball. *Journal of Quantitative Analysis in Sports, 11*(2), 69-84. doi: 10.1515/jqas-2014-0098

Bhandari, I., Colet, E., Parker, J., Pines, Z., Pratap, R., & Ramanujam, K. (1997). Advanced scout: Data mining and knowledge discovery in NBA data. *Data Mining and Knowledge Discovery, 1*(1), 121-125. doi: 10.1023/A:1009782106822

Bishop, C. (2006). *Pattern recognition and machine learning* (Springer Ed.). New York.

Bustamante, Á., & Burillo, P. (2016). Gestión y evaluación del rendimiento en baloncesto: una revisión sistemática del software. *Retos: nuevas tendencias en educación física, deporte y recreación*(29), 72-78.

Courneya, K. S., & Carron, A. V. (1992). The home advantage in sport competitions: A literature review. *Journal of Sport and Exercise Psychology, 14*(1), 13-27.

Davoodi, E., & Khanteymoori, A. (2010). Horse racing prediction using artificial neural networks. *Recent Advances in Neural Networks, Fuzzy*

Systems & Evolutionary Computing, 55-160.
De Marchi, L. (2011). *Data mining of sports performance data.* University of Leeds, School of Computing Studies.
Delen, D., Cogdell, D., & Kasap, N. (2012). A comparative analysis of data mining methods in predicting NCAA bowl outcomes. *International Journal of Forecasting, 28*(2), 543-552. doi: 10.1016/j.ijforecast.2011.05.002
Demens, S. (2015). Riding a probabilistic support vector machine to the Stanley Cup. *Journal of Quantitative Analysis in Sports, 11*(4), 205-218. doi: 10.1515/jqas-2014-0093
Edelmann-Nusser, J., Hohmann, A., & Henneberg, B. (2002). Modeling and prediction of competitive performance in swimming upon neural networks. *European Journal of Sport Science, 2*(2), 1-10. doi: 10.1080/17461390200072201
Fister Jr, I., Ljubiè, K., Suganthan, P. N., Perc, M., & Fister, I. (2015). Computational intelligence in sports: Challenges and opportunities within a new research domain. *Applied Mathematics and Computation, 262*, 178-186. doi: 10.1016/j.amc.2015.04.004
Gomez, M. A., Pollard, R., & Luis-Pascual, J. C. (2011). Comparison of the home advantage in nine different professional team sports in Spain 1. *Perceptual and motor skills, 113*(1), 150-156. doi: 10.2466/05.PMS.113.4.150-156
González-Ruiz, S., Gómez-Gallego, I., Pastrana-Brincones, J., & Hernández-Mendo, A. (2015). Algoritmos de clasificación y redes neuronales en la observación automatizada de registros. *Cuadernos de psicología del deporte, 15*(1), 31-40.
Hagenbuchner, M., Cliff, D. P., Trost, S. G., Van Tuc, N., & Peoples, G. E. (2015). Prediction of activity type in preschool children using machine learning techniques. *Journal of Science and Medicine in Sport, 18*(4), 426-431. doi: 10.1016/j.jsams.2014.06.003
Haghighat, M., Rastegari, H., & Nourafza, N. (2013). A review of data mining techniques for result prediction in sports. *Advances in Computer Science: an International Journal, 2*(5), 7-12.
Hamilton, M., Hoang, P., Layne, L., Murray, J., Padgett, D., Stafford, C., & Tran, H. T. (2014). *Applying machine learning techniques to baseball pitch prediction.* Paper presented at the Proceedings of the 3rd International Conference on Pattern Recognition Applications and Methods.
Hua, K.-L., Lai, C.-T., You, C.-W., & Cheng, W.-H. (2015). An efficient pitch-by-pitch extraction algorithm through multimodal information. *Information Sciences, 294*, 64-77. doi: 10.1016/j.ins.2014.09.001
Jamieson, J. P. (2010). Home field advantage in athletics: a meta-analysis. *Journal of Applied Social Psychology*, 819-1848. doi: 10.1111/j.1559-1816.2010.00641.x
Jeff, H., & John, R. (2011). Using Local Correlation to Explain Success in Baseball. *Journal of Quantitative Analysis in Sports, 7*(4), 1-29. doi: 10.2202/1559-0410.1278
Jelinek, H. F., Kelarev, A., Robinson, D. J., Stranieri, A., & Cornforth, D. J. (2014). Using meta-regression data mining to improve predictions of performance based on heart rate dynamics for Australian football. *Applied Soft Computing, 14*, 81-87. doi: 10.1016/j.asoc.2013.08.010
Knottenbelt, W. J., Spanias, D., & Madurska, A. M. (2012). A common-opponent stochastic model for predicting the outcome of professional tennis matches. *Computers and Mathematics with Applications, 64*, 3820-3827. doi: 10.1016/j.camwa.2012.03.005
Link, D., & Lames, M. (2009). Sport Informatics: Historical Roots, Interdisciplinarity and Future Developments. *International Journal of Computer Science in Sports, 8*(2), 68-87.
Lock, D., & Nettleton, D. (2014). Using random forests to estimate win probability before each play of an NFL game. *Journal of Quantitative Analysis in Sports, 10*(2), 197-205. doi: 10.1515/jqas-2013-0100
Min, B., Kim, J., Choe, C., Eom, H., & McKay, R. B. (2008). A compound framework for sports results prediction: A football case study. *Knowledge-Based Systems, 21*(7), 551-562. doi: 10.1016/j.knosys.2008.03.016
Mitchell, T. M. (1997). *Machine Learning*. New York: McGraw-Hill.
Morgan, S., Williams, M. D., & Barnes, C. (2013). Applying decision tree induction for identification of important attributes in one-versus-one player interactions: A hockey exemplar. *Journal of sports sciences, 31*(10), 1031-1037. doi: 10.1080/02640414.2013.770906
O'Reilly, N. J., & Knight, P. (2007). Knowledge Management Best Practices in National Sport Organizations. *International Journal of Sport Management and Marketing, 2*(3), 264-280. doi: 10.1504/IJSMM.2007.012405
Ofoghi, B., Zeleznikow, J., MacMahon, C., & Dwyer, D. (2010). A machine learning approach to predicting winning patterns in track cycling omnium *Artificial Intelligence in Theory and Practice III* (pp. 67-76): Springer.
Paul, R. J., & Weinbach, A. P. (2009). Sportsbook pricing and the behavioral biases of bettors in the NHL. *Journal of Economics and Finance, 36*(1), 123-135. doi: 10.1007/s12197-009-9112-4
Piatetsky, S. (2016). Difference between Data Mining and Statistics. http://www.kdnuggets.com/faq/difference-data-mining-statistics.html
Pic, M., & Castellano, J. (2016). Efecto de la localización del partido en eliminatorias de ida y vuelta de la UEFA Champions League. *RICYDE. Revista Internacional de Ciencias del Deporte, 44*(12), 149-163. doi: 10.5232/ricyde2016.04405
Pollard, R. (1986). Home advantage in soccer: A retrospective analysis. *Journal of sports sciences, 4*(3), 237-248. doi: 10.1080/02640418608732122
Pollard, R., & Pollard, G. (2005). Long-term trends in home advantage in professional team sports in North America and England (1876–2003). *Journal of sports sciences, 23*(4), 337-350. doi: 10.1080/02640410400021559
Pueo, B., & Jimenez-Olmedo, J. M. Application of motion capture technology for sport performance analysis (El uso de la tecnología de captura de movimiento para el análisis del rendimiento deportivo). *Retos*(32), 241-247.
Rama Iyer, S., & Sharda, R. (2009). Prediction of athletes performance using neural networks: An application in cricket team selection. *Expert Systems with Applications, 36*, 5510-5522. doi: 10.1016/j.eswa.2008.06.088
Sauer, R. D., Waller, J. K., & Hakes, J. K. (2010). The progress of the betting in a baseball game. *Public Choice, 142*(3-4), 297-313. doi: 10.1007/s11127-009-9544-6
Schumaker, R. P., Solieman, O. K., & Chen, H. (2010). Sports knowledge management and data mining. *Annual review of information science and technology, 44*(1), 115-157. doi: 10.1002/aris.2010.1440440110
Shalev-Shwartz, S., & Ben-David, S. (2014). *Understanding Machine Learning: From Theory to Algorithms*: Cambridge University Press.
Shao, S. (2009). Application of BP neural network model in sports aerobics performance In IEEE (Ed.), *Knowledge Engineering and Software Engineering, 2009. KESE'09. Pacific-Asia Conference*. (pp. 33-35).
Soto-Valero, C. (2017). A Gaussian mixture clustering model for characterizing football players using the EA Sports' FIFA video game system. *RICYDE. Revista Internacional de Ciencias del Deporte, 49*(13), 244-259. doi: 10.5232/ricyde2017.04904
Soto-Valero, C., & González-Castellanos, M. (2015). Sabermetría y nuevas tendencias en el análisis estadístico del juego de béisbol (Sabermetrics and new trends in statistical analysis of baseball). *Retos*(28), 122-127.
Soto-Valero, C., Pérez-Morales, I., González-Castellanos, M., & de la Celda Brovkina, A. (2016). ACI-Polo: Sistema computacional para el análisis de la actividad competitiva individual en juegos de polo acuático. *Revista Cubana de Ciencias Informáticas, 10*(1), 229-244.
Spann, M., & Skiera, B. (2009). Sports forecasting: a comparison of the forecast accuracy of prediction markets, betting odds and tipsters. *Journal of Forecasting, 28*(1), 55-72. doi: 10.1002/for.1091
Sun, J., Yu, W., & Zhao, H. (2010). *Study of association rule mining on technical action of ball* Paper presented at the 2010 International Conference on Measuring Technology and Mechatronics Automation.
Van Haaren, J., Ben Shitrit, H., Davis, J., & Fua, P. (2016). *Analyzing volleyball match data from the 2014 World Championships using machine learning techniques.* Paper presented at the Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining.
Witten, I. H., Frank, E., & Hall, M. A. (2011). *Data Mining Practical Machine Learning Tools and Techniques* (3rd ed.): Morgan Kaufmann Publishers.
Woodland, L. M., & Woodland, B. M. (1994). Market Efficiency and the Favorite-Longshot Bias: The Baseball Betting Market. *The Journal of Finance, 49*(1), 269-279. doi: 10.1111/j.1540-6261.1994.tb04429.x
Yadav, A., Sharma, A., Gautam, A., Bathla, G., & Jindal, R. (2017). Predicting English Premier League Results using Machine Learning. *Computer Engineering & Information Technology, 2017.*