Universidad Central Marta Abreu de las Villas
Facultad de Matemática, Física y Computación
Maestría en Ciencia de la Computación

Modelos predictivos con aplicación en el béisbol

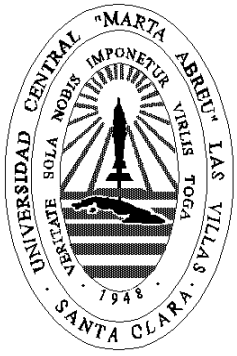
Tesis presentada en opción al título de Máster en Ciencia de la Computación

César Soto Valero

Tutor:

DrC. Irvin Pérez Morales

Santa Clara
Noviembre del año 2016



El que suscribe, César Soto Valero, hago constar que el trabajo titulado «**Modelos predictivos con aplicación en el béisbol**» fue presentado en la Universidad Central «Marta Abreu» de Las Villas como parte del programa de la Maestría en de Ciencia de la Computación, autorizando a que el mismo sea utilizado por la institución, para los fines que estime conveniente, tanto de forma parcial como total y que además no podrá ser presentado en eventos ni publicado sin la autorización de la mencionada institución universitaria.

_____

Firma del autor

Los abajo firmantes, certificamos que el presente trabajo ha sido realizado según acuerdos de la dirección de nuestro centro y el mismo cumple con los requisitos que debe tener un trabajo de esta envergadura referido a la temática señalada.

_____ _____

Firma del Tutor

Firma del Jefe del Laboratorio

Dedicatoria

A mi madre,
con el amor que su recuerdo merece.

Resumen

El aprendizaje automático de datos deportivos constituye un área de investigación reciente. Las tareas de predicción han acaparado la atención en el contexto deportivo debido al inmenso número de aplicaciones posibles, así como por la disponibilidad actual de una gran cantidad de datos y registros históricos de este tipo. El béisbol es reconocido por ser uno de los deportes que mayor cantidad de estadísticas genera durante cada partido. La sabermetría se ha consolidado como una tendencia novedosa en el estudio de este deporte. En este trabajo se abortan dos de los principales problemas predictivos del béisbol haciendo uso de métodos del aprendizaje automático. El primero de ellos es la predicción de resultados de juegos, mientras que el segundo consiste en la predicción del desempeño de los lanzadores abridores. Para su solución se proponen dos modelos pertenecientes al paradigma del aprendizaje automático supervisado, estos incluyen tanto a los métodos de aprendizaje tradicionales como a una propuesta basada en la clasificación de series de tiempo. Para evaluar los modelos propuestos se realizan experimentos empleando conjuntos de datos reales de juegos de la MLB. Los resultados obtenidos demuestran la viabilidad del uso de los métodos del aprendizaje automático para dar solución a problemas de predicción en el deporte.

Abstract

Machine learning of sport datasets is an open field of research today. In this context, predictive tasks have attracted special attention due its vast number of applications. The availability of sport data by means of historical records or play-by-play data sequences have increased significantly today. Baseball is recognized as a statistical filled sport, generating a large amount of data during each game. Sabermetrics has consolidated as a novel tendency for studying and analysing baseball. In this work, we focus on two main predictive baseball problems using machine learning methods. The first consist in predicting win–loss outcomes in baseball games, the second focus on the prediction of performance of starting pitchers. We propose two models for solving these problems using classical machine learning methods and also a time series classification model. To evaluate this approaches we performed experiments using datasets from real games of the MLB. The results obtained show the power of machine learning methods for solving predictive problems in sports.

Índice general

Listado de figuras

Listado de tablas

Listado de acrónimos

AB Acrónimo del inglés *At Bat*

ANNs Acrónimo del inglés *Artificial Neural Networks*

ARIMA Acrónimo del inglés *Autoregressive Integrated Moving-Average*

AVE Del inglés *Average*

BABIP Acrónimo del inglés *Batting Average on Balls in Play*

BP Acrónimo del inglés *backpropagation*

DTW Acrónimo del inglés *Dynamic Time Warping*

DTs Acrónimo del inglés *Decision Trees*

EDR Acrónimo del inglés *Edit Distance on Real Sequences*

ERA Acrónimo del inglés *Earned Run Average*

ERM Acrónimo del inglés *Empirical Risk Minimization*

ERP Acrónimo del inglés *Edit Distance with Real Penalty*

FFN Acrónimo del inglés *Feed-Forward Neural Networks*

FIP Acrónimo del inglés *Fielding Independent Pitching*

HA Acrónimo del inglés *Home Advantage*

MLB Acrónimo del inglés *Major League Baseball*

MLP Acrónimo del inglés *Multi Layer Perceptron*

MVP Acrónimo del inglés *Most Valuable Player*

NBA Acrónimo del inglés *National Basketball Association*

NDS Acrónimo del inglés *Negative Definite Symmetric*

OBP Acrónimo del inglés *On Base Percentage*

OPS Acrónimo del inglés *On Base Plus Slugging*

PDS Acrónimo del inglés *Positive Definite Symmetric*

PE Acrónimo del inglés *Pytagorean Expectation*

RBI Acrónimo del inglés *Run Batted In*

RC Acrónimo del inglés *Runs Created*

RNN Acrónimo del inglés *Recurrent Neural Networks*

SABR Acrónimo del inglés *Society for American Baseball Research*

SB Acrónimo del inglés *Stolen Base*

SLG Del inglés *Slugging*

SRM Acrónimo del inglés *Structural Risk Minimization*

SVMs Acrónimo del inglés *Support Vector Machines*

TWED Acrónimo del inglés *Time Warped Edit Distance*

WAR Acrónimo del inglés *Wins Above Replacement*

kNN Acrónimo del inglés *k Nearest Neighbors*

Introducción

Los logros en el deporte han estado determinados por diferentes factores, uno de ellos es la aplicación de la ciencia y la tecnología en esta área del conocimiento. Cada día se avanza más en abordar los problemas existentes con métodos de investigación científica, dejando atrás la empiria y la espontaneidad. En la actualidad existe una mayor comprensión de que no es suficiente conocer la realidad observable con vista a solucionar problemas prácticos de la actividad física y deportiva, sino que se hace necesario describir, comprender, interpretar, explicar teóricamente o predecir para transformar esa realidad, lo que requiere de la utilización de métodos y medios especiales de conocimiento. En este sentigo, es preciso elaborar sistemas teóricos confirmables en la práctica. Los conocimientos en las ciencias de deporte no pueden estar en forma de indicaciones concretas sin presentar una base teórica metodológica que los sustente.

La gestión automática de la información y la inteligencia artificial son disciplinas que están en el centro de la revolución tecnológica moderna. Una de las ramas de la inteligencia artificial que ha alcanzado un mayor auge es el aprendizaje automático. Este se ocupa de la construcción y estudio de sistemas que pueden aprender de los datos sin la necesidad de programar explícitamente el nuevo conocimiento adquirido. El auge del aprendizaje automático se debe a su gran aplicabilidad, ya que prácticamente todo conocimiento es susceptible de ser aprendido, si bien el uso de este modelo de aprendizaje puede estar limitado en determinadas aplicaciones por razones técnicas (como por ejemplo, el acceso a los datos), económicas (por ejemplo, los costos de desarrollo), legales u otras. Los métodos del aprendizaje automático son usados como una herramienta básica en ramas de la inteligencia artificial, como la minería de datos y el reconocimiento de patrones.

Han sido muchos los avances obtenidos en el desarrollo y aplicación de métodos del análisis cuantitativo de datos en el contexto deportivo, 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. Sin embargo, el aprendizaje automático

es un campo de investigación muy joven dentro de este campo, y presenta aún muchos aspectos a resolver que requieren ser estudiados en detalle. Es por ello que la presente tesis se enfoca en algunos de estos problemas, especialmente aquellos que tienen que ver con tareas de predicción, para el caso particular del juego de béisbol.

Ante la situación expuesta anteriormente, se plantea el **problema de la investigación** siguiente:

Los métodos del aprendizaje automático presentan considerables ventajas respecto a las técnicas estadísticas utilizadas tradicionalmente en el contexto deportivo. La sabermetría representa un significativo paso de avance en el análisis cuantitativo del juego de béisbol. Sin embargo, los métodos del aprendizaje automático no han sido lo suficientemente aplicados en la resolución de tareas de predicción en este deporte.

El **ojetivo general** de esta investigación consiste en:

Desarrollar modelos predictivos que permitan resolver problemas de predicción en el béisbol mediante el uso de métodos del aprendizaje automático.

Para lograr este objetivo general se plantean los siguientes **objetivos específicos**:

1. Realizar un estudio de los principales métodos de aprendizaje automático aplicables al contexto deportivo.
2. Elaborar un modelo para la predicción de resultados de juegos de béisbol usando métodos del aprendizaje automático que utilice los principales conceptos y estadísticos de la sabermetría.
3. Proponer un modelo basado en clasificación de series de tiempo que permita predecir el desempeño de los lanzadores abridores en el béisbol.

Las **preguntas de investigación** planteadas son las siguientes:

- ¿Hasta qué punto los conceptos de la sabermetría son útiles para predecir resultados de juegos particulares de béisbol?
- ¿Cuál método del aprendizaje automático ofrece los mejores resultados predictivos?
- ¿Cómo es dicho resultado en comparación con el mercado de apuestas?
- ¿Será factible transformar los datos lanzamiento-a-lanzamiento en series de tiempo para modelar el desempeño de los lanzadores abridores?

- ¿Cómo se comporta la clasificación de series de tiempo de lanzamientos en los distintos momentos del juego de béisbol?

Después de haber realizado el marco teórico se formularon las siguientes **hipótesis de investigación**:

H1: Los métodos del aprendizaje automático, en conjunción con los más novedosos estadísticos de la sabermetría, resultan ser útiles para la predicción de resultados de juegos de béisbol.

H2: La clasificación de series de tiempo constituye una solución factible para resolver el problema de la predicción del desempeño de los lanzadores abridores en el béisbol.

Para lograr los objetivos trazados y demostrar las hipótesis establecidas se acometieron las siguientes **tareas de investigación**:

1. Realizar un estudio del estado del arte acerca del aprendizaje automático supervisado y la clasificación de series de tiempo.
2. Determinar las ventajas del aprendizaje automático aplicado al análisis cuantitativo de datos deportivos.
3. Elaborar un resumen de los principales estadísticos de béisbol establecidos por la sabermetría.
4. Identificar las principales fuentes de datos de béisbol de acceso público disponibles actualmente.
5. Llevar a cabo un estudio comparativo de los resultados de los métodos de aprendizaje automático para la predicción de resultados de juegos de béisbol.
6. Elaborar un método para transformar los datos lanzamiento-a-lanzamiento de los abridores de béisbol en series de tiempo.
7. Proponer un modelo basado en clasificación de series de tiempo que permita determinar el momento indicado en el cual debe ser sustituido el lanzador abridor.

La **novedad científica** de esta investigación radica en:

- Realizar un estudio riguroso de las potencialidades de los métodos del aprendizaje automático para la predicción de resultados deportivos.
- Proponer un novedoso modelo de predicción del desempeño deportivo basado en la clasificación de series de tiempo.

El **valor práctico** de este trabajo está dado por:

- Determinación de los estadísticos que más influyen en el resultados de juegos de béisbol.
- Estimación del uso de métodos del aprendizaje automático en la predicción de resultados de juegos de béisbol.
- Implementación de un modelo para la predicción del desempeño de los lanzadores abridores de béisbol.

La **tesis** está **estructurada** en cuatro capítulos y unas conclusiones generales. Además, se incluye una sección de comentarios finales y dos apéndices. El contenido de cada uno de los capítulos se introduce a continuación:

En el Capítulo 1 se exponen los fundamentos del aprendizaje automático supervisado y la clasificación de series de tiempo. Se detalla el funcionamiento de cuatro de los métodos del aprendizaje automático más populares en la actualidad: aprendizaje basado en casos, árboles de decisión, máquinas de soporte vectorial y redes neuronales artificiales. Además, se introducen los métodos utilizados particularmente para el análisis y clasificación de series de tiempo, por ser este un tipo especial de dato.

En el Capítulo 2 se presentan los principales elementos del aprendizaje automático aplicado al análisis cuantitativo de datos deportivos. Se describen las principales ventajas del aprendizaje automático respecto al análisis estadístico empleado tradicionalmente en el deporte. Además, dado el avance que ha tenido el análisis cuantitativo en el caso particular del juego de béisbol, gracias a los aportes de la sabermetría, se presentan algunos de los principales logros en este deporte.

En el Capítulo 3 se realiza un estudio comparativo de cuatro métodos del aprendizaje automático supervisado, los cuales fueron empleados para la predicción de resultados de juegos de béisbol de la MLB. Los resultados obtenidos son comparados entre sí usando métodos estadísticos y además se realiza un estudio de estos resultados respecto al mercado de apuestas deportivas de tipo *money–line*.

En el Capítulo 4 se propone un modelo cuyo objetivo es predecir el desempeño de los lanzadores abridores de béisbol. Se describen las características particulares de este problema, enfatizando en la forma en que se transforman los datos lanzamiento-a-lanzamiento para efectuar su análisis como series de tiempo. El modelo propuesto es validado utilizando datos de juegos reales correspondientes a 20 lanzadores abridores de la MLB.

Capítulo 1

Acerca del aprendizaje automático supervisado y la clasificación de series de tiempo

En el presente capítulo se exponen los principales conceptos del aprendizaje automático supervisado y de la clasificación de series de tiempo. Primeramente, en la Sección 1.1 se abordan las características principales de algunos de los métodos del aprendizaje automático supervisado más populares de la actualidad. Seguidamente, en la Sección 1.2 se formaliza la definición de series de tiempo y se enuncian los tipos de análisis existentes para este tipo especial de dato, haciendo énfasis en el enfoque de clasificación. Por último, la Sección 1.3 concluye con un sumario donde se resaltan los principales aspectos tratados en este capítulo.

1.1. Aprendizaje automático

Desde la aparición de las computadoras estas han sido capaces de resolver problemas muy complejos para el hombre, pero aún no poseen la habilidad de aprender por sí solas. A pesar de esto, el desarrollo de la inteligencia artificial ha propuesto un gran número de algoritmos que intentan imitar esta habilidad, los cuales han demostrado ser especialmente efectivos para ciertos tipos de problemas.

El aprendizaje automático o aprendizaje de máquina (del inglés *Machine Learning*), 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 ha sido ampliamente utilizado en el descubrimiento de información valiosa a partir de datos almacenados (Witten *et˜al.*, 2011). Con frecuencia el campo de aplicación del aprendizaje automático se solapa con la estadística, ya que las dos disciplinas se basan en el análisis de datos, por lo que resulta difícil establecer una línea divisoria entre ambos. No obstante, el aprendizaje automático se centra más en el estudio de la complejidad computacional de los problemas. Muchos problemas son de clase NP-Hard, por lo que gran parte de la investigación realizada en esta rama de la ciencia se enfoca al diseño de soluciones factibles a esos problemas.

De forma más concreta, se trata de crear algoritmos capaces de generalizar comportamientos a partir de información suministrada en forma de ejemplos. Tales ejemplos sirven como entrenamiento, para que luego el algoritmo pueda enfrentarse a nuevos datos. Estos algoritmos construyen un modelo a partir de los ejemplos y lo usan para hacer predicciones, en lugar de seguir instrucciones estáticas estrictas como cualquier otro algoritmo.

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 al programa permiten que el algoritmo de aprendizaje sea capaz de extraer de ellos la información necesaria para enfrentarse a nuevos datos y realizar la función para la cual fue diseñado. Podemos definir un ejemplo de entrenamiento de la siguiente forma:

Definición 1.1. Se denomina instancia o ejemplo x a la representación de un objeto específico. Esta instancia se suele representar como un vector $\mathcal{D}$-dimensional $x = (x_1, x_2, \ldots, x_\mathcal{D}) \in \Re^D$ donde cada elemento x_a representa el valor que toma el ejemplo x en el atributo a. A la longitud $\mathcal{D}$ se le conoce como dimensionalidad del vector de atributos x (Zhu y Goldberg, 2009).

Un atributo pudiera tomar otro tipo de valores, no solamente reales sino también nominales. Esta representación de instancia es una abstracción de los objetos, pudiéndose ignorar otras características que no son representadas por los atributos.

Según el resultado que se desea obtener a partir de un sistema, existen varias categorías en las cuales se engloban las tareas de aprendizaje automático. A continuación se describen algunas de las más importantes:

- **Clasificación:** la entrada es dividida en dos o más clases, y el sistema debe producir un modelo capaz de asignarle a una nueva entrada una o más de estas clases, típicamente se lleva a cabo mediante aprendizaje supervisado.

- **Regresión:** es también una tarea supervisada, similar a la anterior pero la salida obtenida es continua en lugar de discreta.
- **Agrupamiento:** el conjunto de entrada es dividido en grupos. A diferencia de la clasificación los grupos no son conocidos de antemano, haciendo de esta una tarea no supervisada.
- **Descubrimiento de reglas:** se trata de encontrar reglas, generalmente del tipo *if-then*, que relaciones a los datos.

El aprendizaje automático tiene una amplia gama de aplicaciones, incluyendo motores de búsqueda, diagnósticos médicos, detección de fraude en el uso de tarjetas de crédito, análisis del mercado de valores, clasificación de secuencias de ADN, reconocimiento del habla y del lenguaje escrito, juegos, robótica, etc (Bishop, 2006; Mitchell, 1997). Algunos de los principales métodos del aprendizaje automático supervisado para las tareas de clasificación y regresión constituyen el objetos de análisis en esta sección.

1.1.1. Aprendizaje supervisado

La Tabla 1.1 muestra la forma en que se representa la información para llevar a cabo el aprendizaje supervisado, es importante notar que para cada ejemplo x^i existe asociada una salida y^i. El objetivo de este tipo de aprendizaje consiste en ajustar un modelo que relacione el valor de salida y^i con los valores de los atributos predictores en x^i. Formalmente podemos definir el aprendizaje supervisado de la siguiente forma:

Definición 1.2. Sea X el dominio de los ejemplos de entrenamiento y sea Y el dominio de las salidas. Dado un conjunto de l ejemplos de entrenamiento $\{(x^i, y^i)\}_{i=1}^{l}$, el aprendizaje supervisado tiene como objetivo entrenar una función $f : X \mapsto Y$ capaz de predecir el valor correcto de $y \in Y$ dado cierto valor de $x \in X$.

En dependencia del dominio que tenga y entonces se puede categorizar el problema de aprendizaje en clasificación o regresión. Cuando Y representa un dominio discreto de etiquetas o clases entonces consideramos la función f como un clasificador. Por el contrario, cuando Y representa un dominio continuo la función f se denomina función de regresión.

Ejemplos	$Atrib_1$	$Atrib_2$	$\cdots$	$Atrib_n$	**Decisión**
x_1	$x_{1,1}$	$x_{1,2}$	$\cdots$	$x_{1,n}$	y_1
x_2	$x_{2,1}$	$x_{2,2}$	$\cdots$	$x_{2,n}$	y_2
$\cdots$	$\cdots$	$\cdots$	$\cdots$	$\cdots$	$\cdots$
x_m	$x_{m,1}$	$x_{m,2}$	$\cdots$	$x_{m,n}$	y_m

Tabla 1.1: Representación de la información en el aprendizaje supervisado.

1.1.2. Métodos clásicos del aprendizaje automático supervisado

El número de algoritmos de aprendizaje supervisado que han sido desarrollados hasta este momento es impresionante, de forma que sería difísil y engorroso hacer una lista que los nombrara a todos y, menos sensato aún, describirlos. Sin embargo, es posible identificar elementos claves en el diseño de los algoritmos tradicionales, lo cual permiten agruparlos atendiendo a varios enfoques fundamentales. En esta sección se presentan y describen algunos de los algoritmos más representativos atendiendo a cuatro de los métodos más populares del aprendizaje automático supervisado: aprendizaje basado en casos, árboles de decisión, máquinas de soporte vectorial y redes neuronales artificiales.

1.1.2.1. Aprendizaje basado en casos

La esencia del aprendizaje basado en casos, también conocido como aprendizaje perezoso, es devolver como solución a un problema dado, la solución conocida de un problema similar (Mitchell, 1997). En los algoritmos de aprendizaje basados en casos cada concepto se representa mediante un conjunto de ejemplos. Cada ejemplo puede ser una abstracción del concepto o una instancia individual del concepto. El método de los k vecinos más cercanos o kNN constituye un algoritmo clásico dentro de esta corriente o forma de solucionar problemas y ha sido ampliamente empleado tanto a problemas de clasificación como de regresión (Cover y Hart, 1967).

El algoritmo kNN es un método de clasificación supervisado el cual calcula una función de densidad $F(x/C_j)$ de las instancias predictoras x para cada clase C_j. Este es un método de clasificación no paramétrico, que estima el valor de la función de densidad de probabilidad o directamente la probabilidad *a posteriori* de que un elemento x pertenezca a la clase C_j a partir de la información proporcionada por el conjunto de ejemplos. Durante el proceso de aprendizaje no se realiza ninguna suposición acerca de la distribución de las

variables predictoras. El método básicamente consiste en comparar la instancia a clasificar con los datos o casos existentes (ejemplos de entrenamiento) del problema en cuestión, recuperando los k casos más parecidos o cercanos, lo cual depende de la similitud entre los atributos del nuevo caso con los casos de la muestra de aprendizaje o entrenamiento. El resultado es la clase mayoritaria de los k casos o instancias más cercanas obtenidas. En el caso de los problemas de regresión, dicho resultado suele ser la media aritmética del valor clase de estas k instancias.

Los ejemplos de entrenamiento X son vectores en un espacio multidimensional dado. Cada ejemplo está descrito en términos de p atributos considerando q clases para la clasificación. Los valores de los atributos del i-ésimo ejemplo (donde $1 \leqslant i \leqslant n$) se representan por el vector p-dimensional:

$$x_i = (x_{i_1}, x_{i_2}, \ldots, x_{i_p}) \in X \tag{1.1}$$

Dado un ejemplo x que debe ser clasificado, sean $x_1, \ldots, x_k$ los k vecinos más cercanos a x en los ejemplos de aprendizaje, el objetivo es devolver:

$$\hat{f}(x) \leftarrow argmax_{v \in V} \sum_{i=1}^{K} \delta(v, f(x_i)) \tag{1.2}$$

donde

$$\delta(a,b) = \begin{cases} 1 & \text{si } a = b, \\ 0 & \text{en caso contrario} \end{cases}$$

El valor $\hat{f}(x)$ devuelto por el algoritmo como un estimador de $f(x)$ es solo el valor más común de f entre los k vecinos más cercanos a x. En el caso especial cuando $k = 1$, y por lo tanto es el vecino más cercano a x_i el único que determina su valor, el algoritmo es conocido como 1NN.

Para seleccionar los k vecinos el método requiere de una función que devuelva la distancia entre la instancia a clasificar y las instancias almacenadas. Luego se seleccionan las k instancias más cercanas cuyo valor de distancia es el menor. En el método se pueden utilizar tanto distancias como funciones de semejanza. Obviamente, si se calculan distancias se

seleccionaran los k ejemplos de menor distancia al problema, mientras que si se usan funciones de semejanza se seleccionaran los k ejemplos más similares.

La distancia Euclideana constituye la medida de distancia más utilizada y puede ser definida de la siguiente manera:

$$distEuclideana(x, y) = \sqrt{\sum_{a=1}^{n} (x_a - y_a)^2} \tag{1.3}$$

En este caso, se considera que x e y son dos vectores de atributos que representan ejemplos, n es la cantidad de atributos usados para describir cada ejemplo y x_a e y_a son los valores del atributo a-ésimo en los ejemplos x e y respectivamente.

Cabe señalar que este método supone que los vecinos más cercanos proveen la mejor clasificación y esto se hace utilizando todos los atributos. El problema de dicha suposición es que es posible que se tengan muchos atributos irrelevantes que dominen sobre la clasificación. Asi por ejemplo, dos atributos relevantes perderían peso ante otros veinte atributos irrelevantes. Para corregir este posible sesgo se puede asignar un peso a cada atributo, dándole así mayor importancia a los atributos más relevantes. Otra posibilidad consiste en tratar de determinar o ajustar los pesos con ejemplos conocidos de entrenamiento. Finalmente, antes de asignar pesos es recomendable identificar y eliminar los atributos que se consideran irrelevantes.

1.1.2.2. Árboles de decisión

Un árbol de decisión o DT es un modelo de predicción, perteneciente al enfoque de programación «divide y vencerás», el cual es utilizado en inteligencia artificial para la toma de decisiones (Rokach y Maimon, 2008). El aprendizaje mediante árboles de decisión constituye uno de los métodos más usados en la práctica para la realización de inferencias inductivas (Quinlan, 1986).

Un árbol de decisión tiene como entrada un conjunto de atributos en los cuales se basa para producir la decisión. A partir de una conjunto de casos se construye un modelo basado en árboles, de forma similar a los sistemas de predicción basados en reglas.

Estos métodos dividen adaptativamente el espacio de entrada en regiones disjuntas con el objetivo de crear fronteras de decisión. De esta forma en cada nodo se realiza un chequeo

sobre la región a la que pertenece un atributo y de acuerdo a esto se toma una rama para descender y continuar el proceso hasta llegar a una hoja la cual indica la salida (Kohavi y Quinlan, 2002).

La Figura 1.1 representa un ejemplo de un árbol de decisión para problemas de clasificación. El objetivo del aprendizaje en este caso es, a partir del conjunto de instancias $X = (x_1, x_2, \ldots, x_n)$ situadas en la raíz del árbol, predecir una salida Y. Los nodos X_2, X_3 y X_4 son subconjuntos disjuntos, siendo $X = X_2 \cup X_3 \cup X_4$. En este caso, los subconjuntos que no se dividieron, X_5, X_6, X_8, X_9, X_{10}, X_{11} y X_{12}, son los nodos hoja.

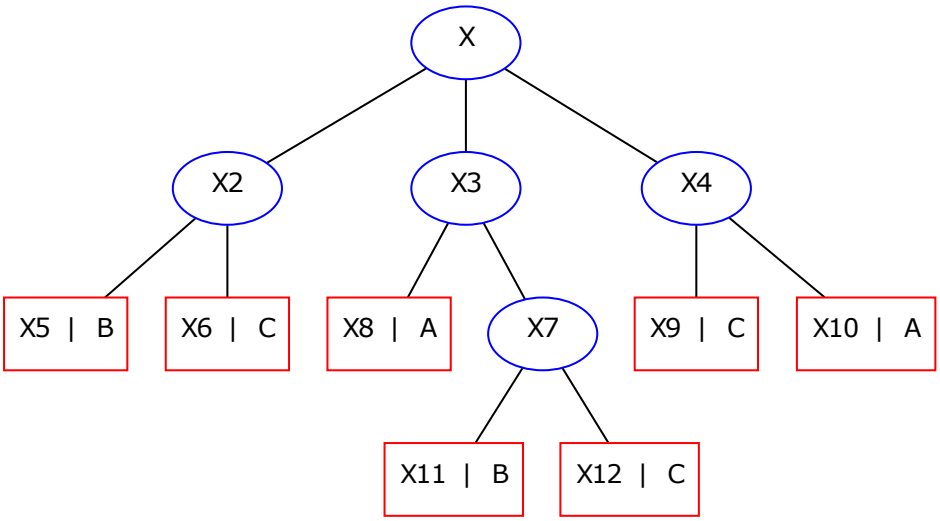


Figura 1.1: Ejemplo de árbol de decisión, los nodos internos se indican con círculos y los nodos hoja con cuadrados.

Los nodos hoja forman una partición de X, y se le asigna la etiqueta de una clase, de varias clases, o la probabilidades de pertenencia a cada una de las clases. Una partición P de X se puede obtener formando grupos con los nodos de cada clase, de ese modo, para el ejemplo anterior se obtienen las particiones siguientes:

$$P_A = X_8 \cup X_{10}$$

$$P_B = X_5 \cup X_{11} \cup X_{12}$$

$$P_C = X_8 \cup X_9$$

Estas particiones son formadas por condiciones sobre los atributos de las instancias. En este caso, una partición de X puede ser dada de la forma siguiente:

$$X_2 = \{x \in X \mid x_2 = \phi\} \textbf{ AND } X_3 = \{x \in X \mid x_2 = \varphi\} \textbf{ AND } X_4 = \{x \in X \mid x_2 = \upsilon\}$$

donde $x_2 = (\phi, \varphi, \upsilon)$ es un atributo nominal.

Por otro lado, la partición de X_2 en X_5 y X_6 puede ser dada por:

$$X_5 = \ x \in X_3 \mid x_1 >= \gamma\} \textbf{ AND } X_6 = \{x \in X_3 \mid x_1 < \gamma\}$$

donde x_1 representa un atributo continuo con $\gamma \in \Re$.

En el caso de un árbol de clasificación donde cada hoja está etiquetada con una clase para clasificar una instancia nueva x_n se procede de la manera siguiente: empezando desde el nodo raíz, se prueba la condición en cada nodo interno visitado para determinar la rama a seguir hasta alcanzar un nodo hoja, y por último, se asigna a la instancia x_i la etiqueta del nodo hoja alcanzado. Por ejemplo, para $a = (\phi, \gamma, ...)$ con $\gamma < x_1$, y las condiciones planteadas anteriormente. Al evaluar en la raíz, a_2 es igual a ϕ por lo que se pasa a X_2, mientras que X_2 se encuentra que a_1 es menor que γ y se sigue por X_6. Luego, dado que X_6 es un nodo hoja, finalmente x_n recibe la clase C.

Segun Breiman *et˜al.* (1984), la definición de un método de aprendizaje basado en árboles de decisión debería abarcar los siguientes elementos:

1. La selección de las divisiones.
2. La decisión de cuando declarar un nodo hoja o seguir dividiendo.
3. La asignación de cada nodo hoja a una clase.

Por lo tanto, la selección del criterio seguido para dividir el nodo es un aspecto fundamental en la construcción de algoritmos que implementen árboles de decisión. En tal sentido, existen un gran número de métricas disponibles las cuales por lo general miden la homogeneidad del elemento a predecir para cada subconjunto de nodos del árbol. Los valores resultantes son combinados para proveer una medida de la calidad de la división. A continuación se detallan dos de los criterios más utilizados en la literatura (Breiman *et˜al.*, 1984).

Ganancia de información

La ganancia de información mide cuán bien un atributo dado separa los ejemplos de entrenamiento acorde al atributo clase. Esta medida se estima a partir del cálculo de la entropía (Shannon, 1948). Esta última caracteriza la impureza de una colección arbitraria de instancias. Dado un atributo objetivo, que puede tomar c valores diferentes, con frecuencia de aparición p_i, como muestra la Ecuación 1.4, la entropía se calcula a partir de la Ecuación 1.5.

Esta se calcula sobre la base de la cantidad de instancias de cada clase que alcanzan un nodo determinado y tiene como significado el monto de información necesario para clasificar una instancia que alcance dicho nodo. Si todos los miembros del conjunto S pertenecen a una misma clase el valor de la entropía será cero. En otro caso tomará un valor mayor que cero pudiendo alcanzar como valor máximo $\log_2 c$.

$$p_i = \frac{S_i}{S} \tag{1.4}$$

$$Entropía(S) \equiv \sum_{i=1}^{c} -p_i \log_2 p_i \tag{1.5}$$

$$Ganancia(S, A) \equiv Entropía(S) - \sum_{v \in A} \frac{S_v}{S} Entropía(S_v) \tag{1.6}$$

La ganancia de información de un atributo A, relativa a una colección de ejemplos S se calcula a partir de la Ecuación 1.6. Esta medida es simplemente la reducción esperada de la entropía causada por la partición de los ejemplos acorde a un atributo determinado. $Valores(A)$ es el conjunto de todos los posibles valores del atributo A, y S_v es el subconjunto de S para el cual el atributo A tiene valor v. El primer término de esta ecuación es justamente la entropía presente en la colección original S y el segundo término es el valor esperado de entropía después que S es particionado usando el atributo A. La entropía descrita mediante el segundo término está compuesta por la suma de las entropías de cada subconjunto S_v, ponderadas por la fracción de los ejemplos $\frac{S_v}{S}$ que pertenecen a S_v.

Reducción de la varianza

El criterio basado en la reducción de la varianza es empleado comúnmente en problemas de regresión. Este se diferencia de otras métricas que requieren la previa discretización de las variables de entrada antes de ser utilizadas. La reducción de la varianza de un nodo N, tal como se muestra en la Ecuación **??** se define como la reducción total de la variable objetivo x debido a la partición en este nodo:

$$I_v(N) = \frac{1}{|S|}\sum_{i\in S}\sum_{j\in S}\frac{1}{2}(x_i - x_j)^2 - \{\frac{1}{|S_t|}\sum_{i\in S_t}\sum_{j\in S_j}\frac{1}{2}(x_i - x_j)^2 + \frac{1}{|S_f|}\sum_{i\in S_f}\sum_{j\in S_f}\frac{1}{2}(x_i - x_j)^2\} \quad (1.7)$$

donde S son el conjunto de índices de la muestra de prepartición, S_t es el conjunto de índices de la muestra para el que la prueba de partición es cierto y S_f es un conjunto de índices de la muestra para los cuales la prueba de partición es falsa.

1.1.2.3. Máquinas de soporte vectorial

Las máquinas de soporte vectorial o SVMs son un conjunto de algoritmos que están propiamente relacionados con las tareas de clasificación y regresión del aprendizaje automático. Este método fue creado originalmente por Cortes y Vapnik (1995) para resolver problemas de clasificación binaria, aunque ha sido ampliamente extendido al dominio de los problemas de regresión (Vapnik, 2013). Su formulación se basa en el principio de minimización del riesgo estructural o SRM, el cual ha demostrado ser superior que el principio tradicional de minimización del riesgo empírico o ERM. El principio de SRM minimiza un límite superior sobre el riesgo esperado, mientras ERM minimiza el error sobre los datos de entrenamiento. Esta diferencia provee a las SVMs de una mayor habilidad para generalizar.

Definición 1.3. En el contexto del aprendizaje automático supervisado, el funcionamiento de las SVMs se basa es encontrar un hiperplano que maximice el margen de separación entre las instancias de cada clase.

El modelo de SVMs construye un hiperplano o conjunto de hiperplanos en un espacio característico de alta dimensionalidad, obteniendo una buena separación entre las clases. En ese concepto de «separación óptima» es donde reside la característica fundamental de las SVMs. En la búsqueda del hiperplano óptimo sólo se calculan los productos escalares

de los vectores en el espacio característico. Las funciones núcleo son utilizadas con este fin, permitiendo el cálculo de los productos escalares en el espacio de entrada en lugar del espacio característico.

Problemas linealmente separables

La Figura 1.2 representa un problema de clasificación proyectado en un espacio de dos dimensiones $\Re^2$. Como se puede apreciar, existen infinitos hiperplanos que separan los datos, pero solo el hiperplano H_3 lo hace de manera óptima.

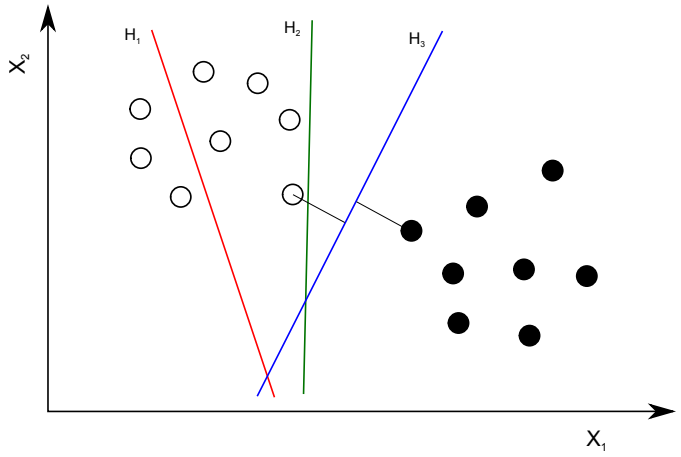


Figura 1.2: Hiperplanos de separación para un problema linealmente separable.

En este caso, el hiperplano buscado tiene la forma siguiente:

$$\langle w, x \rangle + b \tag{1.8}$$

donde w es un vector perpendicular al hiperplano y b es su distancia convenientemente normalizada desde el origen.

El vector w es una combinación lineal de vectores del conjunto de entrenamiento los cuales son cercanos al hiperplano de separación. Estos vectores se denominan vectores soporte. La etiqueta de una instancia nueva x depende de su posición respecto al hiperplano atendiendo al resultado de la Ecuación **??**, la cual arroja como resultado los valores 1 o -1. La función f se soluciona resolviendo un problema de optimización cuadrático bajo ciertas restricciones.

$$f(x) = sgn[\langle w, x\rangle + b] \tag{1.9}$$

Problema linealmente separables con datos no separables

Si los datos de entrenamiento no son linealmente separables debido a la presencia de instancias ruidosas, entonces no es posible encontrar un hiperplano de separación óptimo. En este caso, las SVMs relajan las restricciones para permitir errores en la separación de los datos de entrenamiento por medio de la disminución del peso de dichas instancias. La Figura 1.3 muestra un ejemplo donde el hiperplano separa los datos dejando algunas instancias en el lado incorrecto.

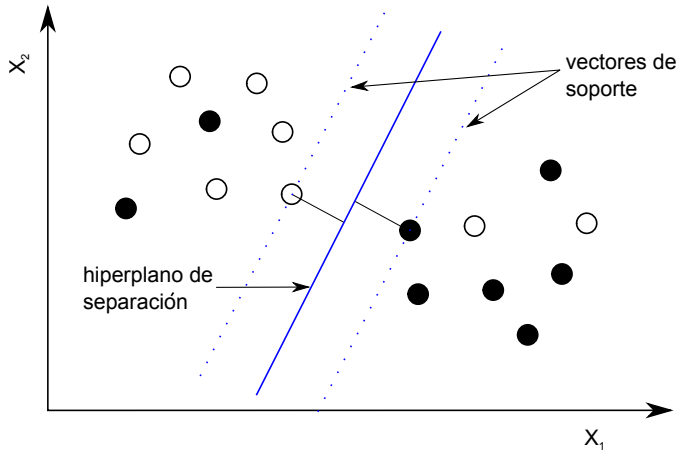


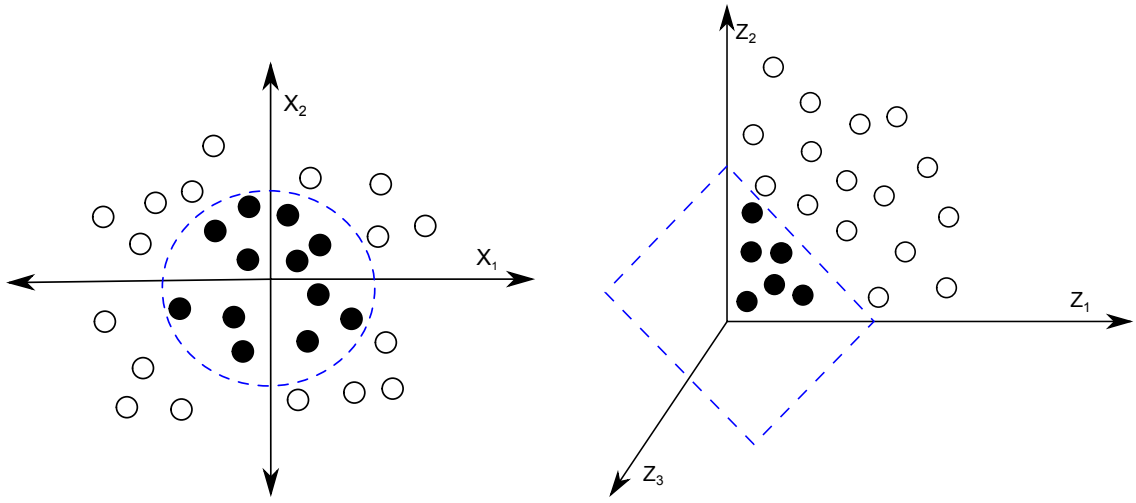
Figura 1.3: Hiperplano generalizado de separación óptimo.

Problemas linealmente no separables

En muchos problemas reales resulta imposible encontrar un hiperplano que pueda separar los datos debido a que el problema no es linealmente separable. En este caso, las SVMs realizan un mapeo no linear de los datos a un espacio característico de alta dimensión. El objetivo es transformar los datos a un espacio donde estos sean linealmente separables, permitiendo algunos errores como en el caso anterior.

La Figura 1.4 muestra a la izquierda los datos no linealmente separables en el espacio original $\Re^2$, mientras a la derecha están separados por un plano en el espacio característico $\Re^3$. La función usada para el mapeo en este caso fue $(x_1, x_2) \mapsto (z_1, z_2, z_3) := (x_1^2, \sqrt{2}x_1x_2, x_2^2)$.

Como se puede observar, se ha logrado un mapeo de los datos a otro espacio donde son linealmente separables.



(a) Espacio de entrada $\Re^2$. (b) Espacio característico $\Re^3$.

Figura 1.4: Ejemplo del mapeo de las instancias del espacio de entrada $\Re^2$ al espacio característico $\Re^3$.

En Gunn *et˜al.* (1998) se demuestra que no es necesario conocer la función de mapeo. Para la búsqueda del hiperplano óptimo solo se necesita calcular los productos escalares de los vectores, para este caso, en el espacio característico. La idea para evitar el mapeo se basa en el uso de una función núcleo. Las funciones núcleos permiten calcular los productos escalares en el espacio de entrada, en lugar del espacio característico, obteniendo los mismos resultados.

Funciones núcleo

Las funciones núcleos son aquellas que satisfacen el teorema de Mercer (1909), en el cual se establece que una función $k(x, y)$ simétrica y continua en el espacio de entrada representa un producto escalar en un espacio característico si y solo si k es semi-definida positiva.

Lei y Sun (2007) afirman que solo los núcleos simétricos definidos positivos o PDS son admisibles para la formulación estándar de las SVMs. El uso de núcleos PDS garantiza que la matriz del núcleo sea convexa y la solución sea única. También explican que los núcleos simétricos definidos negativos o NDS pueden ser empleados para construir núcleos PDS dado que existe un teorema que los relaciona.

La solución descrita para los problemas linealmente no separables es la más general, en éstas el uso de las funciones núcleos es la clave. Para los problemas lineales se puede emplear la función núcleo lineal, mientras que para el resto existen existen funciones núcleo no lineales tales como: polinomial, de base radial gaussiana o sigmoidal. La Tabla 1.1 muestra algunos de los núcleos más utilizados en la literatura.

Núcleo	**Función**	**Parámetros**
Lineal	$K(x_i, x_j) = \langle x_i, x_j \rangle$	$-$
Polinomial	$K(x_i, x_j) = (x_i \cdot x_j + 1)^d$	d
Base radial gaussiana	$K(x_i, x_j) = e^{-(\|x_i - x_j\|^2/2\sigma^2)}$	σ
Sigmoide	$K(x_i, x_j) = tanh(kx_i \cdot x_j - \delta)$	k, δ

Tabla 1.2: Ejemplos de funciones núcleo admisibles utilizadas por las SVMs.

1.1.2.4. Redes neuronales artificiales

Las Redes Neuronales Artificiales o ANNs se agrupan dentro de las técnicas conexionistas de la inteligencia artificial y constituye una de las áreas de estudio más ampliamente difundidas. Constituyen modelos matemáticos inspirados en la biología que han sido utilizados para dar solución a problemas complejos, pues son considerados excelentes aproximadores de funciones no lineales. Las ANNs son capaces de aprender las características relevantes de un conjunto de datos para luego reproducirlas en entornos ruidosos o incompletos, siendo especialmente útiles para tareas de clasificación y regresión (Hammer y Villmann, 2003).

Definición 1.4. Una red neuronal artificial es una estructura distribuida de procesamiento paralelo formada por neuronas artificiales interconectadas entre sí, en forma de grafo acíclico dirigido, las cuales son usadas para almacenar conocimiento.

Usualmente las ANNs reciben la información proveniente del exterior mediante un conjunto de neuronas de entrada y cuentan con un conjunto distinto de neuronas de salida para manejar los resultados. El resto de las neuronas se organizan en las denominadas capas ocultas. Se concibe el cálculo general de la red a partir de la información que es procesada por cada una de sus neuronas de forma independiente. Cada neurona puede recibir información de las restantes y calcular su propia salida a partir de dicha entrada y de su estado actual, transitando eventualmente hacia un nuevo estado. Por lo general, el flujo de cálculo de la red avanza progresivamente desde las neuronas de entrada hacia las

neuronas de salida, en un proceso en el que cada una de las neuronas en las capas ocultas va activándose progresivamente atendiendo al esquema de conexión particular de cada red (Hilera y Martínez, 1995).

Las ANNs pueden ser caracterizadas de acuerdo al modelo la neurona, el esquema de conexión que presentan sus neuronas, o sea su topología, y el algoritmo de aprendizaje empleado para adaptar su función de cómputo a las necesidades del problema particular. Entre los modelos de ANNs que se emplean en la solución de problemas de aprendizaje supervisado, uno de los más referenciados es el Percéptron Multicapa o MLP. El modelo MLP es reconocido actualmente como uno de los mejores para solucionar problemas de clasificación y regresión a partir de ejemplos (Han *et˜al.*, 2011), no obstante, se le señala como principal inconveniente que este se comporta como una caja negra y no facilita la interpretación de los resultados. A continuación se detallan los aspectos fundamentales necesarios para la comprensión de este modelo de aprendizaje.

El modelo de la neurona artificial

La neurona artificial está inspirada en las células neuronales del cerebro y posee un esquema simple de cálculo en una sola dirección. La neurona recoge los niveles de señal en sus entradas y los procesa atendiendo a una función de activación, la cual puede ser lineal o no lineal y es dependiente del estado actual en que se encuentre dicha neurona, para posteriormente devolver un nivel concreto de señal en la salida.

La Figuara **??** presenta un modelo de neurona artificial Grau (2011). El cálculo de la entrada neta ξ_j de la neurona es una combinación lineal de las salidas provenientes de las neuronas de entrada $(x_1, x_2, \ldots, x_n)$, cuyos coeficientes $(w_1, w_2, \ldots, w_n)$ constituyen valores asociados a cada una de las conexiones de entrada a la neurona y son denominados pesos de las conexiones, más un valor constante w_0 denominado umbral. Estos coeficientes constituyen los parámetros libres del modelo, cuyo adecuado reajuste permite una mejor aproximación a la solución de los problemas concretos. El nuevo estado de activación de la neurona $a_j(t+1)$ se determina a través de la función F teniendo en cuenta el estado de activación actual $a_j(t)$ y la entrada neta ξ_j. La salida real y_i de la neurona se calcula mediante la función f a partir de su estado de activación.

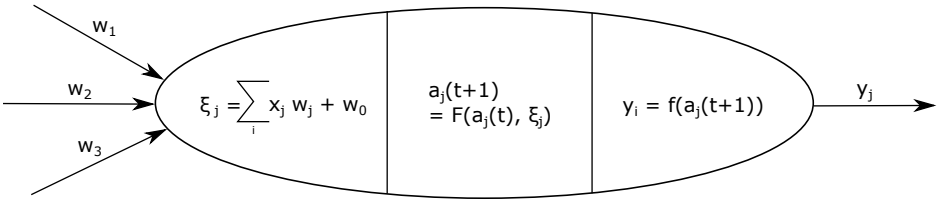


Figura 1.5: Modelo de una neurona artificial.

Existe una amplia variedad de modelos de neuronas, cada uno se corresponde con un tipo determinado de función de activación y de salida. Igualmente el cálculo de la entrada neta puede ser realizado de múltiples formas. En algunos modelos bioinspirados puede conservarse una memoria de corto período del estado de la neurona o pueden concebirse ecuaciones más complejas que ayuden a establecer tipos de procesamiento de mayor potencia. Respecto a las funciones de salida, la función sigmoidal es actualmente la más empleada en tareas de clasificación y regresión, la misma puede definirse de la forma siguiente:

$$y(\xi) = \frac{1}{1 + e^{-\xi}} \tag{1.10}$$

Esquema de conexión de una red neuronal

La topología y el número de neuronas conectadas que intervienen en una red neuronal artificial determinan directamente el poder computacional de la misma. En los últimos años se han propuesto una gran variedad de arquitecturas de ANNs (Hilera y Martínez, 1995), sin embargo, la mayoría de estas propuestas pueden ser ubicadas en los dos grupos siguientes:

- Redes multicapa hacia adelante o FFN.
- Redes recurrentes o RNN.

Una vez caracterizado un problema, la topología más adecuada para su solución debe ser seleccionada de manera empírica. Por un lado, la potencia de cálculo para resolver los problemas complejos depende directamente de la cantidad de neuronas que disponga la red, y por otro, un número muy elevado de éstas puede afectar el desempeño de los algoritmos de entrenamiento. Tal situación conlleva a adoptar soluciones de compromiso entre el poder de representación de la red y su simplicidad. Por lo general, la aplicación de heurísticas y técnicas especiales junto a la experiencia del especialista constituyen la

forma más efectiva de abordar este problema (Han *et˜al.*, 2011).

Los algoritmos de entrenamiento constituyen métodos que se aplican sobre los modelos de red para adaptar sus parámetros con el propósito de obtener un comportamiento determinado. Existe una amplia variedad de algoritmos de entrenamiento disponibles, la mayoría de ellos desarrollados para arquitecturas muy específicas. Usualmente estos algoritmos se clasifican en supervisados o no supervisados (Hilera y Martínez, 1995). Los algoritmos supervisados son los que requieren, en su funcionamiento, información sobre las respuestas esperadas de la red, y los no supervisados los que no la requieren. Los algoritmos supervisados adaptan la red mediante la optimización de una función de error que calcula las diferencias entre las respuestas deseadas y las obtenidas por el modelo, tomando como referencia un conjunto de datos fijos denominado conjunto de entrenamiento. Las técnicas supervisadas son caracterizadas por sus propiedades de convergencia global o local, rapidez con que logran dicha convergencia y los recursos en tiempo y espacio que requiere para su aplicación.

Redes neuronales con conexiones hacia adelante

La característica fundamental de las redes FFN reside en que sus neuronas están conectadas a manera de grafo acíclico dirigido, es decir, con todos sus arcos en una sola dirección. Este tipo de red define una relación de orden parcial entre sus neuronas y con frecuencia éstas pueden agruparse en forma de capas siguiendo dicha relación.

Las redes MLP constituyen un ejemplo genérico de las redes FFN. Frecuentemente se encuentran formadas por un conjunto de capas de neuronas ordenadas secuencialmente por una capa de entrada, un conjunto de capas intermedias denominadas capas ocultas y una capa de salida. En la Figura 1.6 se muestra un ejemplo típico de este tipo de red.

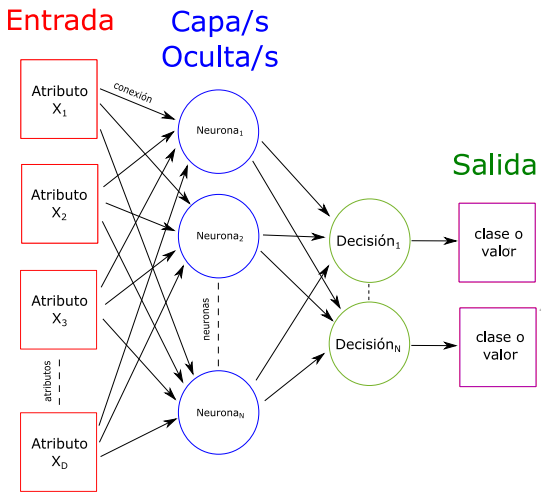


Figura 1.6: Topología de una red neuronal artificial con conexiones hacia adelante.

El modelo MLP, usando neuronas ocultas con funciones no lineales, es capaz de aproximar cualquier tipo de función continua y brindar excelentes resultados en las tareas de clasificación (Zurada, 1992; Hammer y Villmann, 2003). El poder de representación de las redes MLP está relacionado con la existencia de al menos una capa de neuronas ocultas no lineales, las cuales transforman la entrada en una representación interna. Dicha representación puede servir luego a las capas subsecuentes para resolver el problema dado.

Entre las técnicas de entrenamiento supervisado más difundidas de las que hace uso el modelo MLP se encuentran aquellas que basan su funcionamiento en el método del gradiente descendente (Thrun y Smieja, 1990). El algoritmo de propagación del error hacia atrás o BP es el de más amplio uso aplicado a redes con conexiones hacia adelante. Este algoritmo aplica la técnica del gradiente descendente para llevar a cabo la minimización del error de funcionamiento de la red. El error de la red para un patrón p dado puede calcularse como la semisuma de los errores cuadráticos de cada unidad de salida al presentarse dicho patrón a la red, tal como muestra la ecuación siguientes:

$$E_p(W) = \frac{1}{2}\sum_{i \in o}(d_i - y_i)^2 \tag{1.11}$$

donde

W es el conjunto de pesos de las conexiones de la red, y_i y d_i corresponden respectivamente a la salida actual y la salida esperada de la i-ésima neurona de la red dado un patrón de entrada determinado y O es el conjunto de neuronas de salida.

El error total de funcionamiento de la red puede evaluarse como la suma de los errores $E_p(W)$ para cada uno de los patrones presentes en el conjunto de entrenamiento. Este tipo de error es empleado frecuentemente en redes con funciones sigmoidales o lineales en la capa de salida.

La función $E_p(W)$ se representa sobre un espacio multidimensional de dimensión igual al número de pesos de la red. La búsqueda de la solución (conjunto de pesos) que minimice el error se realiza atendiendo a la siguiente ecuación:

$$\Delta w_{i,j} = -\alpha \frac{\delta E(W)}{\delta w_{i,j}} \tag{1.12}$$

El algoritmo BP determina el cambio necesario en los pesos $\Delta w_{i,j}$ mediante la medición de la magnitud de la influencia de cada peso sobre el error de respuesta de la red calculado anteriormente El reajuste de dichos pesos se realiza de manera proporcional por medio de la constante de velocidad de entrenamiento α.

1.2. Series de tiempo

En la mayoría de las ramas de la ciencia, la ingeniería o el comercio existen variables que son medidas secuencialmente a través del tiempo. Por ejemplo, los bancos registran las tasas de interés y de cambio de monedas diariamente, las oficinas de meteorología llevan el control de las precipitaciones y la temperatura en diferentes lugares y con diferente granularidad, etc. Cuando una variable es medida secuencialmente en el tiempo o en un intervalo determinado, los datos tomados forman una serie de tiempo (Cowpertwait y Metcalfe, 2009).

Los datos representados de esta manera son susceptibles a contener información valiosa para su dominio de procedencia. En la actualidad existen dos ramas fundamentales dedicadas a su estudio: el análisis de series de tiempo y la minería de datos de series de tiempo. El primero comprende tanto métodos matemáticos como estadísticos, los cuales han sido utilizados en el pronóstico de valores futuros o con la finalidad de interpretar eventos ocurridos (Chatfield, 2013). El segundo enfoque para el tratamiento de las series temporales surge con la consolidación de una rama de la minería de datos orientada específicamente al estudio de datos temporales (Fu, 2011).

Los métodos tradicionales del aprendizaje automático han sido satisfactoriamente aplica-

dos a este dominio mediante la modelación de las series como un tipo especial de dato. El uso de estos métodos, en combinación con el aumento de la potencia de cómputo, ha propiciado su aplicación en dominios diversos tales como el reconocimiento del lenguaje natural (Sakoe y Chiba, 1978; Itakura, 1975), la biométrica (Kovacs-Vajna, 2000; Niennattrakul *et˜al.*, 2007), la medicina (Bruno y Garza, 2012; Goldberger *et˜al.*, 2000) o astronomía (Java y Perlman, 2002) entre otros.

1.2.1. Definición de serie de tiempo

Según Chatfield (2013) una serie de tiempo consiste en una colección de observaciones realizadas de manera secuencial en el tiempo. Otros autores (Wang *et˜al.*, 2013; Brockwell y Davis, 2006) ofrecen una definición más rigurosa:

Definición 1.5. Una serie de tiempo l s consiste en una secuencia de n pares $((s_1, t_1), (s_2, t_2), \ldots, (s_i, t_i), \ldots, (s_n, t_n))$ $(t_1 < t_2 < \ldots < t_i < \ldots < t_n)$, donde cada elemento s_i es un punto en el espacio $\mathcal{D}$-dimensional al que pertenecen los datos, y cada t_i es el instante de tiempo al cual se corresponde la ocurrencia de s_i.

Una serie de tiempo en cada observación s_i puede contener valores de varias variables. Si la cantidad de variables medidas es igual a uno entonces se denomina univariada, de lo contrario, se llama multivariada.

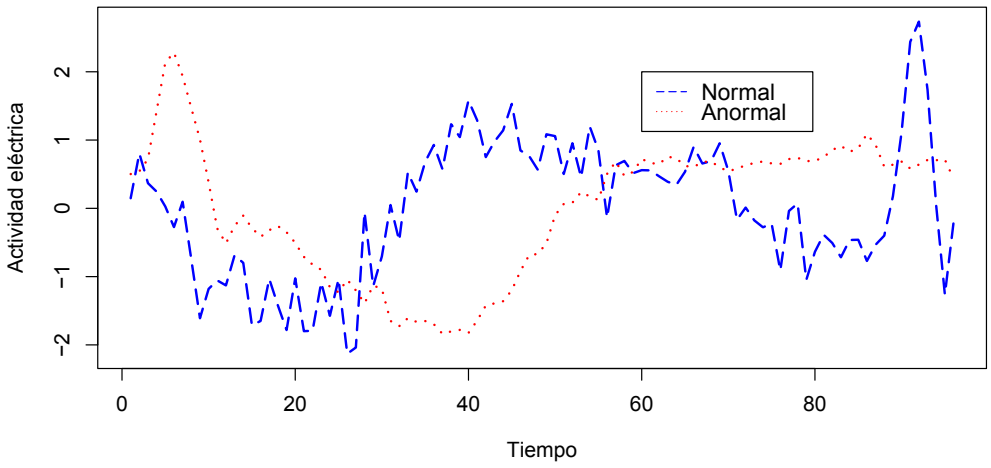


Figura 1.7: Fragmento de un electrocardiograma que describe una pulsación del corazón en una persona sana (Normal) y en una enferma (Anormal).

Una serie de tiempo es considerada como continua cuando las observaciones son hechas de manera sucesiva en el tiempo. Si las observaciones tienen lugar sólo en momentos específicos entonces es considerada como una serie de tiempo discreta. La continuidad de una serie se determina por el modo en el que se toman los valores en el tiempo y es independiente de la naturaleza continua o no de las variables medidas.

La Figura 1.7 muestra un ejemplo de dos series univariadas y discretas que describen una pulsación del corazón en dos personas diferentes (Buza *et˜al.*, 2011). En este caso, la variable fue medida utilizando un electrodo que registró el impulso eléctrico generado por el corazón a intervalos regulares de tiempo.

1.2.2. Análisis de series de tiempo

El análisis de series de tiempo comprende métodos tanto matemáticos como estadísticos, los cuales favorecen la interpretación de datos de este tipo teniendo en cuenta las correlaciones temporales existentes en los mismos. Existen múltiples objetivos que motivan el análisis de series de tiempo, éstos están enfocados principalmente tanto en la extracción de información representativa como en la posibilidad de extrapolar y predecir su comportamiento futuro. Dichos objetivos (Chatfield, 2013) pueden ser clasificados como:

- **Descripción**: permite definir las principales propiedades de la serie mediante la aplicación de técnicas descriptivas. La más simple consiste en visualizar gráficamente la serie objeto de estudio.

- **Explicación**: en ocasiones es posible usar la variación de unas series de tiempo para explicar la variación en otras. Los modelos de regresión múltiple resultan útiles en esta tarea.

- **Predicción**: es uno de los principales objetivos y consiste en predecir los valores futuros de las series analizadas. Resulta de gran importancia para la realización de análisis económicos e industriales.

- **Control**: se aplica cuando se desea controlar la calidad de determinado proceso. Existen múltiples tipos de procedimientos de control, los cuales incluyen poder tomar medidas oportunas frente al proceso que se está controlando.

El análisis clásico de series de tiempo comprende el estudio de cuatro componentes básicos, los cuales resultan ser la fuente de su variación. Varios métodos tradicionales están relacionados con la descomposición de la serie temporal en sus componentes. Algunos de dichos

componentes básicos (Chatfield, 2013) reconocidos en la literatura son los siguientes:

- **Tendencia**: es una componente de la serie que refleja la evolución a largo plazo del fenómeno observado.
- **Variación estacional**: es el movimiento periódico de corto plazo. Se trata de una componente causal producto de la influencia de ciertos fenómenos que se repiten de manera periódica en el tiempo.
- **Variación cíclica**: además de la variación estacional, y debido a una u otra causa, algunas series exhiben variaciones cada cierto período de tiempo de mayor o menor longitud.
- **Variación aleatoria**: también denominada residuo, no muestra ninguna regularidad y se obtiene una vez eliminadas la tendencia y las variaciones cíclicas de la serie.

Los métodos utilizados en el análisis de series de tiempo son típicamente divididos en dos categorías: dominio de la frecuencia (Brockwell y Davis, 2006) y dominio del tiempo (Shumway y Stoffer, 2010). El primero se basa en la función de densidad espectral y el segundo en la función de autocorrelación. Ambos enfoques resultan equivalentes aunque representan formas alternativas de analizar los procesos que originan las series.

Las técnicas de análisis de series de tiempo pueden ser divididas además según sus métodos en paramétricas y no paramétricas (Brockwell y Davis, 2006). Los enfoques paramétricos asumen que la estacionalidad fundamental del proceso estocástico tiene cierta estructura la cual puede ser descrita usando un reducido número de parámetros, por ejemplo los modelos autorregresivos de medias móviles o ARIMA (Box y Jenkins, 1976). En estos enfoques, el objetivo es estimar los parámetros del modelo que mejor describen el proceso estocástico. Por el contrario, los enfoques no paramétricos estiman explícitamente la covarianza o el espectro del proceso sin asumir que este tenga alguna estructura en particular. Adicionalmente, otras clasificaciones han sido creadas para describir los modelos, algunas de ellas son: lineales y no lineales, univariados y multivariados.

1.2.3. Minería de datos de series de tiempo

La minería de datos tiene como objetivo revelar patrones desconocidos a partir de los datos. Su singularidad radica en los tipos de problemas que es capaz de resolver, los cuales incluyen aquellos con enormes conjuntos de datos y relaciones muy complejas entre si. Su

extensión a problemas con contenido temporal explícito o implícito ha dado lugar a una rama de la minería de datos que ha experimentado un vertiginoso desarrollo.

La minería de datos temporales se encuentra en la intersección de varias disciplinas incluyendo estadística, reconocimiento de patrones temporales, bases de datos temporales y optimización entre otras. Según la revisión realizada por Lin *et˜al.* (2002), la minería de datos temporales constituye un paso en el proceso de descubrimiento del conocimiento en conjuntos de datos temporales y se relaciona con el descubrimiento de patrones temporales. En Povinelli (1999) también se hace alusión al concepto de patrones temporales como estructuras que se encuentran potencialmente ocultas en las series de tiempo. Un patrón temporal puede estar asociado a la ocurrencia de un determinado evento, por lo cual está estrechamente relacionado con la predicción del mismos.

En el contexto de la minería de datos de series de tiempo (Esling y Agon, 2012) es una práctica común representar las series como una secuencia ordenada de n observaciones o puntos $s = (s_1, s_2, \ldots, s_i, \ldots, s_n)$. En series temporales discretas, donde las observaciones son hechas en intervalos regulares de tiempo, es posible omitir la variable t_i. En este punto es posible hacer una analogía entre el valor que toma un caso x del aprendizaje automático en el i-ésimo atributo y el valor que toma la serie s en el i-ésimo instante de tiempo. La principal diferencia entre ambos radica en la relevancia del orden de los atributos. En los problemas de aprendizaje tradicionales el orden de los atributos es irrelevante y la relación entre ellos es independiente de sus posiciones. Por el contrario, para las series de tiempo este orden resulta generalmente crucial en la determinación de sus características.

Esta particularidad hace que el tratamiento de las series de tiempo constituya todo un reto para la minería de datos (Fu, 2011), puesto que su dominio específico resulta ser diferente al de los problemas tradicionales del aprendizaje automático. Otras de las características distintivas de este tipo de series son la alta numerosidad, gran número de dimensiones y una constante actualización de sus datos al transcurrir el tiempo. En el contexto del aprendizaje automático, es imprescindible considerar una serie de tiempo como un todo en lugar de tratarla como un conjunto de campos numéricos individuales.

Las tareas de minería de datos temporales que comúnmente se han enfrentado (Keogh y Kasetty, 2003; Fu, 2011; Esling y Agon, 2012; Shahnawaz *et˜al.*, 2011) pueden ser clasificadas en los grupos siguientes:

- **Indexado**: tiene como objetivo, a partir de una serie de interés s y una medida de similitud dada, determinar la serie más cercana a s en un conjunto de datos temporales.

- **Descubrimiento de patrones y conglomerados**: consiste en descubrir patrones interesantes que pueden aparecer con frecuencia o de forma repentina en las series de tiempo. En esta tarea es común emplear algoritmos de agrupamiento.
- **Clasificación**: su principal objetivo consiste en asignarle una etiqueta a una serie a partir de un conjunto de clases previamente definido.
- **Segmentación**: puede ser considera como un paso previo de preprocesamiento o como una técnica de análisis. Tiene como objetivo, a partir de una serie, obtener un conjunto reducido de segmentos que aproximen la serie original.

1.2.4. Aprendizaje automático de series de tiempo

El aprendizaje automático de series de tiempo ha seguido dos enfoques fundamentales: la transformación de las series originales a un nuevo espacio de descripción y la adaptación de los clasificadores existentes al dominio temporal. El primero elimina la relación temporal entre los atributos que describen la serie. El segundo se basa principalmente en la utilización de medidas de disimilitud adaptables a las características de las series de tiempo. Esta sección se enfoca en la descripción de este segundo enfoque.

1.2.4.1. Medidas de disimilitud

Las medidas de disimilitud constituyen el núcleo de varios métodos del aprendizaje automático. Dada la naturaleza numérica y continua de las series de tiempo, el cálculo de la similaridad entre dos series dadas se satisface de forma aproximada, a diferencia de otros tipos de datos donde el concepto de similitud se resuelve de forma exacta. Esto se debe a que resulta difícil en la práctica encontrar dos series de tiempo exactamente iguales. Otra característica que dificulta la aplicación de las medidas de disimilitud es la presencia de distorsiones en la serie tanto en el dominio temporal como en el de los valores. Una práctica extendida consiste en realizar un proceso de normalización de la serie antes de aplicar cualquier medida de disimilitud (Rakthanmanon *et˜al.*, 2012).

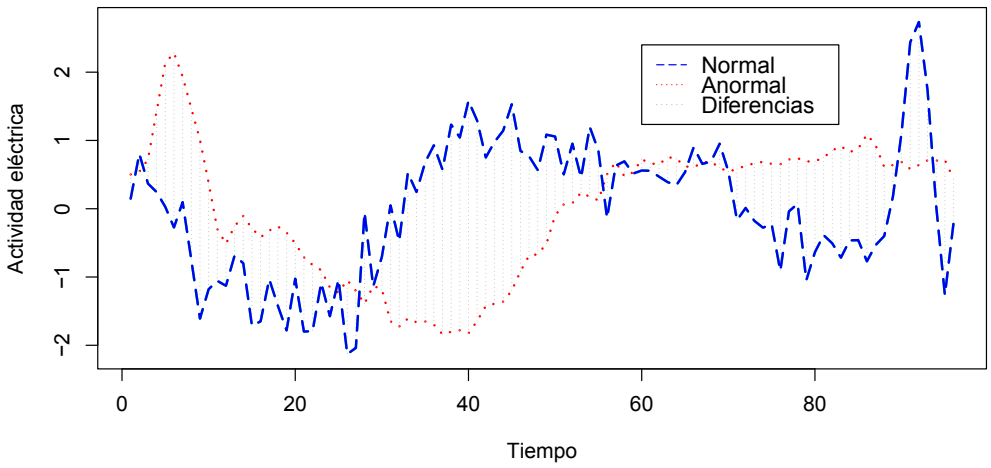


Figura 1.8: Diferencias entre cada punto de datos que se obtienen mediante el cálculo de la distancia Euclidiana entre dos series formadas por fragmentos de electrocardiogramas. El valor de distancia Euclidiana total obtenido en este caso es de 13,9.

Una de las medidas de similitud más usadas por su simplicidad y eficiencia es la distancia Euclidiana tradicional, la cual se emplea fundamentalmente en las series temporales después de alguna transformación (Keogh y Kasetty, 2003). Como se aprecia en la Figura 1.8, su cálculo se realiza a partir de la medición de la diferencia entre cada punto de datos de la serie objetivo respecto a su similar en la serie de referencia. Uno de los principales beneficios de utilizar esta medida es que tiene una complejidad computacional de orden lineal. Como consecuencia de las características particulares que poseen las series de tiempo, varios estudios revelan que esta no es la distancia indicada para dominios más específicos, pues resulta sensible a los desplazamientos y distorsiones de las series en el eje temporal (Wang *et˜al.*, 2013). Además, otra limitación de esta medida es el requerimiento de que las series tengan la misma longitud.

Existen otras múltiples medidas las cuales se han aplicado para evaluar la disimilitud entre series de tiempo, las cuales se pueden categorizar en los siguientes grupos:

- **Basadas en rasgos**: determinan la disimilitud entre dos series de tiempo utilizando rasgos discriminantes pertenecientes al dominio de la frecuencia. La trasformada discreta de Fourier es comúnmente utilizada con este fin (Oppenheim *et˜al.*, 1999).
- **Basadas en modelos**: este tipo de medidas primeramente ajustan un modelo autoregresivo, por ejemplo el ARIMA, en las series a comparar. Los parámetros obtenidos,

a partir del ajuste de los modelos son posteriormente utilizados como rasgos discriminatorios (Carden y Brownjohn, 2008; Corduas y Piccolo, 2008; Bagnall y Janacek, 2004).

- **Elásticas**: basan su funcionamiento en la determinación de los rasgos discriminantes en el dominio temporal. Específicamente, se espera que series de la misma clase compartan determinados comportamientos los cuales pueden encontrarse ocultos producto del ruido o de desplazamientos en el eje temporal. La característica distintiva de este tipo de medidas es que posibilitan el alineamiento de puntos desfasados en el tiempo durante las comparaciones (Sakoe y Chiba, 1978; Chen *et˜al.*, 2005; Marteau, 2009).

Una gran parte de las investigaciones realizadas en el campo de la clasificación de series de tiempo están basadas en la utilización de medidas elásticas. Una de las más utilizadas se denomina distorsión dinámica del tiempo o DTW (Sakoe y Chiba, 1978). Esta medida ha sido empleada satisfactoriamente en un gran número de aplicaciones para fines diversos (Rodríguez y Alonso, 2004; Bartolini *et˜al.*, 2005; Tormene *et˜al.*, 2009; Hamooni *et˜al.*, 2015). Mediante el uso de DTW no solo se consigue el valor de la disimilitud entre dos series sino que además se obtiene el alineamiento óptimo entre ellas, emparejándolas de forma no lineal mediante contracciones y dilataciones de las series en el eje temporal. Por consiguiente, este emparejamiento permite encontrar regiones equivalentes entre las series que facilitan el cálculo de su disimilitud.

Otra familia de medidas elásticas, conocidas como distancias de edición, también han sido aplicadas para calcular la disimilitud en el dominio temporal. La distancia EDR (Chen *et˜al.*, 2005) es considerada una extensión para series de tiempo a partir de la distancia original de Levenshtein (Levenshtein, 1966). La idea es calcular la disimilitud entre series como el costo mínimo de la cantidad de operaciones de edición necesarias para transformar una serie en la otra. La métrica ERP (Chen y Ng, 2004) es una variante de la distancia anterior, donde se utiliza una constante real para penalizar aquellos valores de las series donde es necesaria una transformación de inserción o eliminación. También son penalizados, de acuerdo a la distancia existente entre ellos, aquellos valores donde se aplica una transformación de reemplazo. Por otro lado, la métrica TWED (Marteau, 2009) es una extensión sumamente interesante de las distancias de edición y DTW, por que en esencia puede ser considerada como una combinación de ambos tipos de medidas.

1.2.4.2. Definición y cálculo de DTW

Suponemos que se desea comparar dos series de tiempo dadas, una serie de prueba $q = (q_1, \ldots, q_m)$ y una series de referencia $s = (s_1, \ldots, s_n)$. Además, se asume la existencia de una función f no negativa que expresa el efecto de alinear los puntos q_i y s_i de las series $d(i, j) = f(q_i, s_j) \geq 0$. La distancia Euclidiana se asume generalmente para este fin. El núcleo de esta técnica[1] consiste en encontrar un camino $\phi(t) = (\phi_q(t), \phi_q(t))$ de longitud T que defina una correspondencia entre los elementos de q y s, donde $\phi_q(t) \in \{1, \ldots, m\}$ y $\phi_s(t) \in \{1, \ldots, n\}$. Dado un camino ϕ, la distancia acumulada de los alineamientos propuestos entre las series q y s se calcula según la Ecuación 1.13.

$$d_\phi(q, s) = \sum_{t=1}^{T} d(\phi_q(t), \phi_s(t)) m_\phi(t) / M_\phi \tag{1.13}$$

donde $m_\phi(t)$ es un coeficiente de peso y M_ϕ su constante de normalización correspondiente, asegurando que las distorsiones acumuladas sean comparables entre diferentes caminos.

Con el objetivo de encontrar alineamientos razonables se imponen ciertas restricciones como las condiciones de frontera, continuidad y monotonía. A continuación se define esta última:

$$\phi_q(t + 1) \geq \phi_q(t)$$

$$\phi_s(t + 1) \geq \phi_s(t)$$

El cálculo de DTW, Ecuación 1.14, se basa en encontrar un alineamiento óptimo que garantice una distancia acumulada mínima entre las dos series. En otras palabras, la distorsión que se pretende con el alineamiento es aquella que permita acercar las series q y s tanto como sea posible. Atendiendo a su definición, DTW es considerada una pseudo-distancia debido a que no cumple la desigualdad triangular. La Figura 1.9 muestra un ejemplo del cálculo de DTW, así como los alineamientos que dan como resultado el acumulado de distancia mínima.

$$DTW(q, s) = \min_{\phi} d_\phi(q, s) \tag{1.14}$$

[1] La notación y las ecuaciones utilizadas en esta sección se tomaron del trabajo de Giorgino (2009).

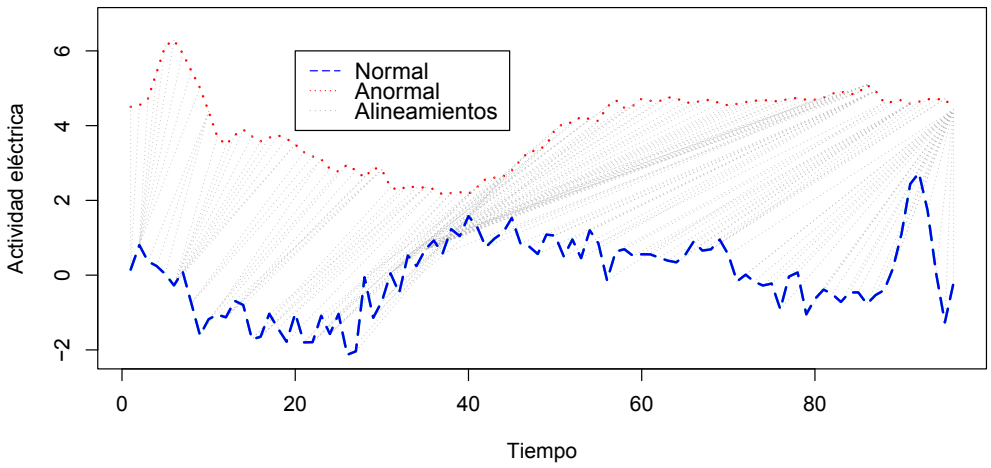


Figura 1.9: Alineamientos obtenidos al aplicar DTW a las dos series formadas por fragmentos de electrocardiogramas mostrados anteriormente en la Figura 1.7. El valor de distancia acumulado en este caso es de 106.8.

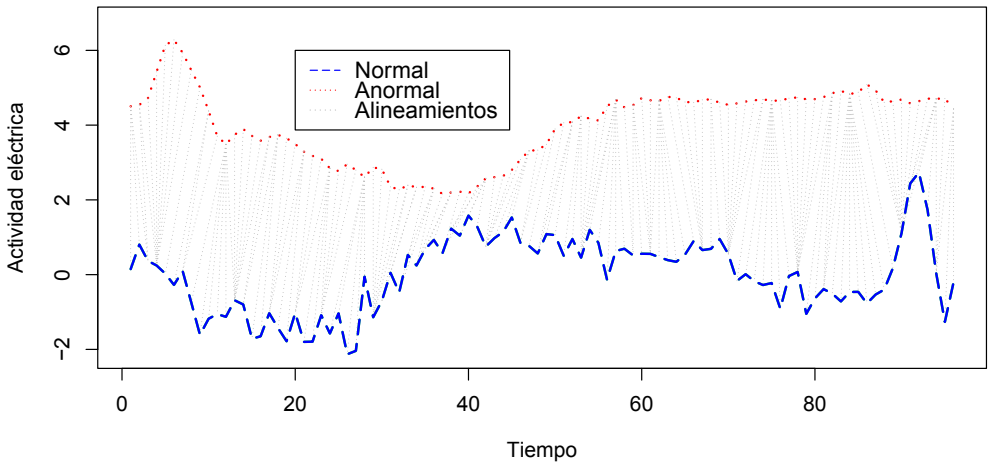


Figura 1.10: Alineamientos obtenidos al aplicar DTW combinada con la banda de Sakoe-Chiba con tamaño de ventana $w = 4$. El valor de distancia acumulado es de 154.3.

El cálculo de DTW se puede realizar utilizando programación dinámica y tiene una complejidad máxima temporal y espacial de $O(n \cdot m)$. En comparación con la distancia Euclidiana, DTW tiene un costo computacional mayor pues depende de forma cuadrática

de la longitud de las series. Dado que una de las características del dominio temporal es la alta dimensionalidad, varias investigaciones han estado encaminadas a reducir el costo computacional de esta medida. Una de estas soluciones es aplicar restricciones globales. Este tipo de restricciones controlan la elasticidad permitida en los alineamientos y por consiguiente la cantidad de celdas a visitar durante la construcción del camino mínimo. La banda de Sakoe-Chiba (Sakoe y Chiba, 1978) y el paralelogramo de Itakura (Itakura, 1975) son las restricciones globales más utilizadas.

En el caso de la banda de Sakoe-Chiba, se pasa como parámetro un tamaño de ventana w el cual limita los alineamientos, añadiendo una nueva restricción a la búsqueda del camino óptimo $|\phi_q(t) - \phi_s(t)| \leq w$. En la Figura 1.10 se muestra el efecto de aplicar esta banda con un tamaño de ventana $w = 4$. Esta restricción afecta la distancia total obtenida entre las dos series. La diferencia, lejos de ser perjudicial, aumenta la exactitud de las clasificaciones pues evita alineamientos erróneos que pueden generarse si no se limita el ancho de la ventana. El estudio desarrollado por Kurbalija *et˜al.* (2014) sobre este tema, para varias medidas elásticas incluyendo DTW, muestra que los mejores valores de exactitud se obtienen utilizando tamaños de ventana pequeños. Aplicar esta banda con un tamaño de ventana igual a cero equivaldría a calcular una distancia similar a la Euclidiana.

1.2.4.3. Métodos de clasificación

La clasificación de series de tiempo es uno de los problemas más abordados de la minería de datos temporales. Al igual que en el aprendizaje supervisado tradicional se parte de un conjunto de entrenamiento formado por ejemplos, que en este caso particular provienen del dominio temporal. A cada serie se se le asocia una clase y el objetivo del aprendizaje consiste en entrenar una función capaz de predecir la clase correcta para una serie dada.

Considerando las características que presentan las series temporales, la tarea de clasificarlas requiere un tratamiento especial. Un primer grupo de propuestas desarrolladas en este sentido lo constituye el enfoque basado en rasgos (Carden y Brownjohn, 2008; Behera *et˜al.*, 2010; Weng y Shen, 2008; Dash *et˜al.*, 2008; Fulcher y Jones, 2014), y se basan en transformar la serie original a un nuevo espacio de descripción donde los clasificadores convencionales pueden ser aplicados. Para extraer los rasgos de la serie original se utilizan comúnmente herramientas del procesamiento de señales o estadísticas. Este enfoque está estrechamente relacionado con las medidas basadas en rasgos y las medidas basadas en modelos que se describieron en la sección 1.2.4.1. Un segundo grupo de propuestas (Ro-

dríguez *et˜al.*, 2000; Povinelli *et˜al.*, 2004; Douzal-Chouakria y Amblard, 2012; Rodríguez y Alonso, 2004; Xi *et˜al.*, 2006; Kaya y GunduzOguducu, 2015) se enfoca en adaptar o desarrollar clasificadores especialmente diseñados para tratar series temporales. Esta categoría se basa principalmente en la selección de: una representación apropiada de las series y de una medida adecuada para calcular la disimilitud entre las mismas, por ejemplo las medidas elásticas. Esta categoría incluye además el enfoque basado en casos, el cual ha recibido gran atención en la literatura especializada.

Algoritmo de clasificación de series de tiempo kNN

El algoritmo de clasificación kNN es actualmente la opción estándar utilizada ampliamente para tareas de clasificación de series de tiempo. La exactitud de este algoritmo depende directamente de la efectividad de la medidas de disimilitud utilizada para comparar las series de tiempo. Específicamente, se suele utilizar 1NN dados lo buenos resultados que ha reportados en comparación con otros valores de k en la literatura (Serrà y Arcos, 2014; Wang *et˜al.*, 2013). Otra ventaja de este algoritmo es que resulta sencillo de implementar y no necesita ajustar parámetros adicionales. De manera general, en el dominio de las series de tiempo este clasificador ha obtenido muy buenos resultados (Xi *et˜al.*, 2006; Petitjean *et˜al.*, 2016).

1.3. Sumario

En este capítulo se han presentado las bases teóricas que hacen posible comprender los fundamentos del aprendizaje automático supervisado. En particular, se abordaron cuatro de los métodos más representativos de este dominio de aplicación: el aprendizaje basado en casos, los árboles de decisión, las máquinas de soporte vectorial y las redes neuronales artificiales. Además, se realizó un estudio sobre el caso particular de las series de tiempo, las cuales deben ser tratadas como un tipo especial de dato, explicando sus características principales en el contexto del aprendizaje automático.

A partir del estudio del estado del arte presentado en este capítulo se pueden resaltar los siguientes aspectos:

- El aprendizaje automático constituye actualmente una de las principales áreas de investigación de la inteligencia artificial.

- El aprendizaje automático supervisado ha sido el paradigma más estudiado. En particular, los métodos de clasificación y regresión desarrollados para este tipo de aprendizaje han resultado ser muy útiles en diversos dominios de aplicación.
- Las series de tiempo representan de manera natural fenómenos presentes en prácticamente todas las áreas del conocimiento.
- Las características especiales que presentan las series de tiempo, como la dependencia temporal entre los puntos de datos, la alta dimensionalidad y numerosidad entre otras, diferencian su tratamiento en comparación con otros problemas tradicionales del aprendizaje automático.
- El algoritmo kNN, en conjunción con la medida de disimilitud DTW, constituye la opción estándar empleada en tareas de clasificación de series de tiempo.

Capítulo 2

Sobre el aprendizaje automático aplicado al análisis cuantitativo de datos deportivos: el caso particular del béisbol

En este capítulo se presentan los elementos fundamentales del análisis cuantitativo de datos deportivos. En la Sección 2.1 se enuncian las principales tareas del aprendizaje automático para el análisis de esta clase particular de datos, señalando sus ventajas frente a los métodos estadísticos tradicionales. En especial, se fundamenta la utilidad de los métodos de clasificación y regresión para la solución de los diferentes problemas de predicción en el deporte. Por otra parte, dado el avance que ha tenido el análisis cuantitativo en el caso particular del juego de béisbol, se dedica la Sección 2.2 a la sabermetría, principal exponente del progreso que ha experimentado tanto el análisis estadístico como el aprendizaje automático en el deporte. Por último, la Sección 2.3 concluye con un sumario donde se resaltan los aspectos fundamentales tratados en este capítulo.

2.1. Aprendizaje automático de datos deportivos

En el contexto deportivo, las decisiones estratégicas son de vital importancia para la obtención de buenos resultados competitivos. La trascendencia de dichas decisiones esta directamente relacionada con su complejidad, por lo que casi siempre se requiere del conocimiento de los expertos humanos a la hora de llevar a cabo la toma de decisiones.

El análisis del rendimiento competitivo ha ganado importancia en la última década,

siendo esta la principal forma para medir y evaluar la actuación de los deportistas. La Metodología Observacional (Aguera *et˜al.*, 2015) a favorecido el estudio de las observaciones sistemáticas obtenidas en este ámbito, fundamentalmente mediante el análisis de datos jugada–a–jugada. Esto último ha sido posible fundamentalmente gracias a los avances en las tecnologías de la información y las comunicaciones (Hua *et˜al.*, 2015; Bishop, 2003).

La minería de datos y el aprendizaje automático ofrecen varias ventajas para el análisis de los datos deportivos. En particular, los métodos de aprendizaje automático han demostrado su eficacia en la medición del rendimiento competitivo en varios deportes colectivos importantes (Schumaker *et˜al.*, 2010b).

En esta sección se exponen los aspectos fundamentales del aprendizaje automático enfocado al análisis de datos deportivos. En especial, se describen los principales métodos utilizados así como los pasos necesarios para llevar a cabo la búsqueda del conocimiento necesario para sustentar la toma de decisiones en el deporte.

2.1.1. Análisis cuantitativo de datos deportivos

En la actualidad, las ciencias del deporte han despertado un gran interés en toda la comunidad científica. 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 el avance de la teoría y práctica del deporte (Link y Lames, 2009).

Definición 2.1. Se considera como deporte a toda actividad física ejercida dentro de un juego o una competición, cuya práctica está sujeta a normas específicas.

El estudio de los aspectos técnico–tácticos constituye la base del análisis del rendimiento deportivo. La Figura 2.1 representa la interrelación que se establece entre ambas disciplinas. Como se puede observar, el objetivo general es la mejora del rendimiento competitivo. En este sentido, el análisis cuantitativo de datos deportivos es un eje fundamental, ya que enlaza varios aspectos claves del análisis técnico–táctico tales como el desempeño competitivo o la estrategia competitiva.

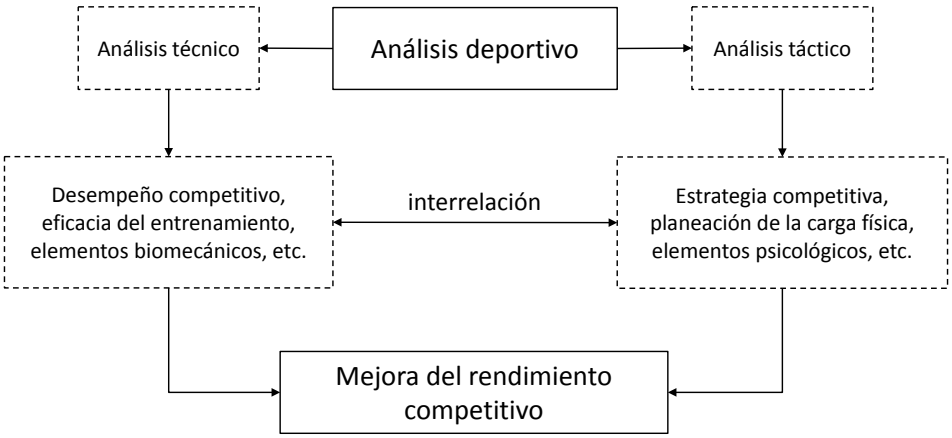


Figura 2.1: Aspectos técnico-tácticos del análisis deportivo.

La cantidad de datos disponibles en a casi todos los ámbitos deportivos ha crecido de forma significativa 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 entrenadores, o 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 determinar cuál es la información más relevante y cómo generar conocimiento útil a partir de la misma (Fister˜Jr *et˜al.*, 2015).

El objetivo final en cualquier deporte consiste alcanzar la victoria frente al contrario. El primer problema presente en la preparación para la obtención de este resultado consiste 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 contra 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.

Mediante la búsqueda del conocimiento oculto en 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, 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 y 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 relevancia en el equipo (Baumer *et˜al.*, 2015).

Antes de la utilización de técnicas estadísticas en este campo, las organizaciones deportivas dependían casi exclusivamente de la experiencia humana. Se debía asumir el supuesto de que los expertos (entrenadores, directores de equipos o cazatalentos) son realmente capaces de convertir los datos de los que disponen en conocimiento útil y veraz. Sin embargo, con el significativo incremente 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 de forma rápida y veraz, con el propósito de explicar científicamente los diversos y complejos fenómenos que ocurren en los terrenos de juego.

El uso de la computación a venido a ser una herramienta fundamental en todos los dominios de aplicación de las ciencias del deporte, y en especial en el análisis cuantitativo de datos deportivos. En general, se pueden identificar las siguientes áreas de investigación en este campo:

- Adquisición y preprocesamiento de los datos.
- Representación de la información y análisis descriptivo.
- Bases de datos y sistemas expertos.
- Simulación.

2.1.2. Aplicaciones del aprendizaje automático en el análisis cuantitativo de datos deportivos

Las técnicas estadísticas han sido usadas tradicionalmente para resolver los problemas inherentes al manejo de datos deportivos. Estas se han aplicado sobre todo para realizar comparaciones entre poblaciones las cuales han sido objeto de determinados planes de ejercicios, o para distinguir patrones significativos en determinados eventos deportivos. Estas 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 y realizar predicciones, lo cual es el propósito de la minería de datos, y en especial de los métodos del aprendizaje automático (Piatetsky, 2016).

La Figura 2.2 presenta una comparación entre las dos técnicas fundamentales que se han utilizadas para el análisis cuantitativo de datos deportivos. En este sentido, el uso de la minería de datos constituye un paso de avance en la comprensión de numerosos fenómenos, los cuales son difíciles de explicar mediante las técnicas estadísticas clásicas.

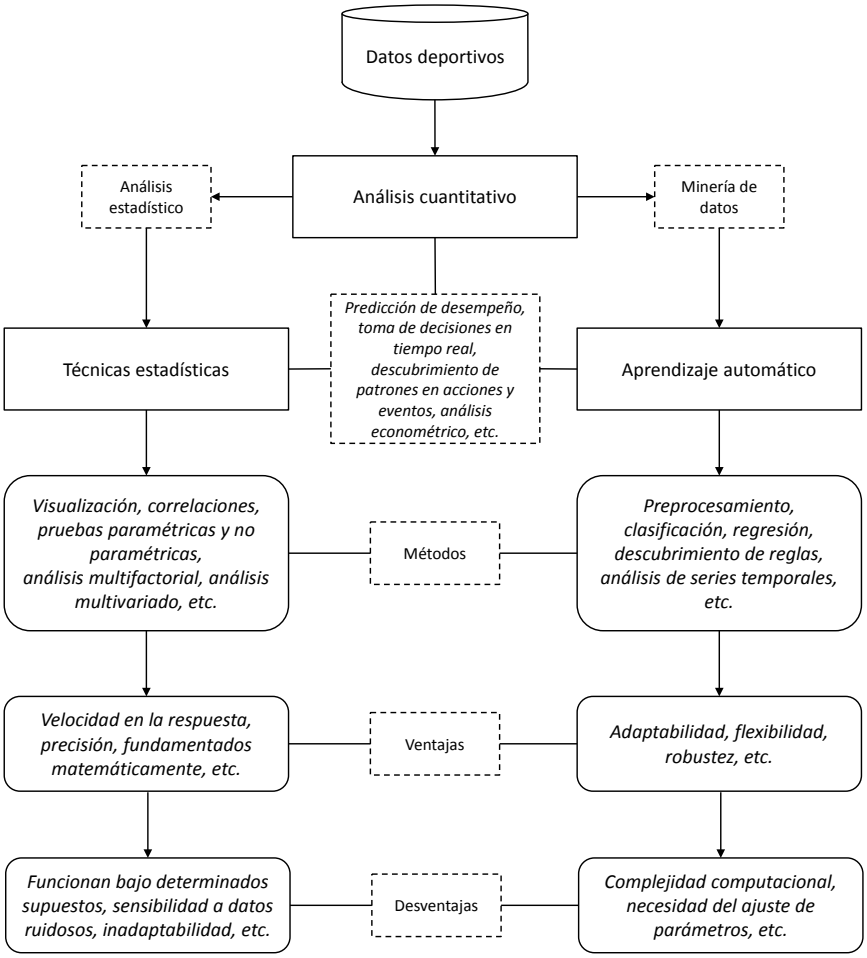


Figura 2.2: Elementos del análisis cuantitativo de datos deportivos.

La minería de datos, y en especial métodos del aprendizaje automático, difieren de las técnicas estadísticas ya que tienen el potencial de hacer generalizaciones a partir de situaciones mucho más complejas, realizando predicciones y revelando relaciones ocultas en los datos. Esta forma de obtención de conocimiento es sumamente útil, ya que puede ser usado 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 *et˜al.*, 2013).

Los métodos del aprendizaje automático son un complemento a las herramientas del análisis estadístico tradicional en el deporte posibilitando, entre otras cosas, dar respuesta a un sinnúmero 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 simulación y aprendizaje automático para planificar sus estrategias completas para la temporada (O'Reilly y 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 de esto 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 tiene la capacidad de dirigir de forma más efectiva, lo cual redunda en una mejor organización y rendimiento competitivo.

La Tabla 2.1 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 utilizadas. Podemos decir que sus principales usos incluyen el análisis del desempeño competitivo, la predicción de resultados en deportes tanto colectivos como individuales y los estudios económicos y de mercado.

2.1.2.1. Análisis del desempeño deportivo

Uno de los aspectos fundamentales en el deporte, y en especial en el de alto rendimiento, es la mejora del desempeño competitivo. Desde el punto de vista del análisis cuantitativo, el objetivo consiste en obtener mejoras significativas a partir de las mediciones realizadas a los atletas de forma periódica.

La Figura 2.3 presenta un esquema general para el monitoreo de los parámetros tecnico–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 un aspecto esencial de este proceso.

Método	Autor/es	Técnica	Deporte
Clasificación	(Hamilton *et˜al.*, 2014)	kNN	béisbol
	(Davoodi y Khanteymoori, 2010)	ANN	equitación
	(Delen *et˜al.*, 2012)	DT	bolos
	(Rama˜Iyer, 2009)	ANN	cricket
Regresión	(Shao, 2009)	ANN	gimnasia
	(Demens, 2015)	SVM	hockey
	(Lock y Nettleton, 2014)	DT	fútbol
	(Jelinek *et˜al.*, 2014)	DT	fútbol
Agrupamiento	(Ofoghi *et˜al.*, 2010a)	k-Medias	ciclismo
	(Menéndez *et˜al.*, 2016)	series de tiempo	béisbol
	(Ofoghi *et˜al.*, 2010b)	k-Medias	ciclismo
Descubrimiento de reglas	(Bhandari *et˜al.*, 1997)	Apriori	baloncesto
	(Sun *et˜al.*, 2010)	Apriori	tenis
	(Valero *et˜al.*, 2016)	Apriori	polo acuático

Tabla 2.1: Ejemplos de aplicación de los métodos del aprendizaje automático en el contexto deportivo.

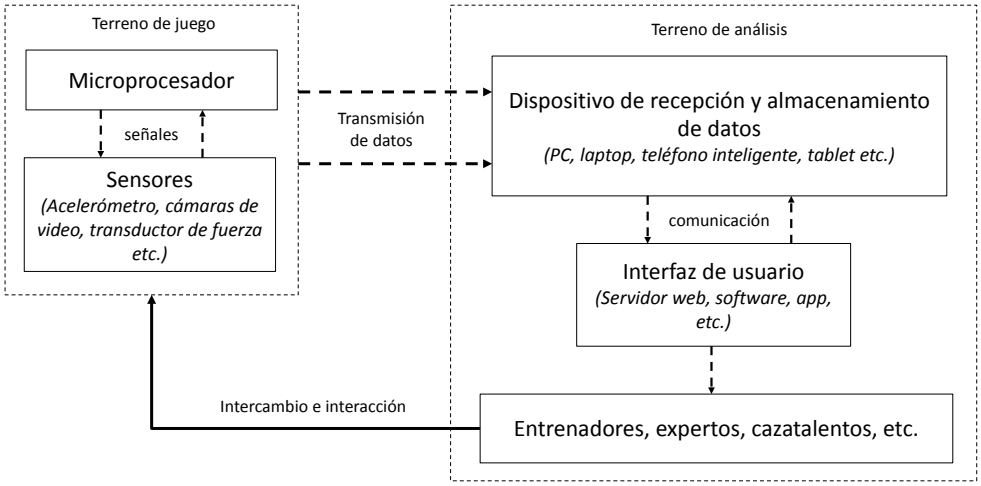


Figura 2.3: Esquema para la implementación de un sistema de monitoreo del desempeño deportivo.

Las herramientas del aprendizaje automático han resultado ser sumamente útiles para el procesamiento de los datos colectados por los diferentes dispositivos creados con este fin. 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 numerosos deportes. Por ejemplo, en Edelman-Nusser *et˜al.* (2002) proponen un modelo basado en redes neuronales artificiales para el análisis del desempeño de nadadores olímpicos, mientras que Morgan *et˜al.* (2013) utilizan árboles de decisión en la identificación de atributos importantes durante la interacción uno–contra–uno en el hockey.

2.1.2.2. Predicción de resultados competitivos

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

La predicción en el deporte tiene características especiales que lo distinguen del resto. Por ejemplo, caben señalar su carácter eminentemente competitivo así como el dinamismo presente, el cual requiere de una actualización constante del modelo de predicción. Esto hace posible en muchos casos generar resultados no solo a partir de datos propios de los deportistas objeto de estudio, sino también sobre la base de otros elementos y variables del contrario (Knottenbelt *et˜al.*, 2012). Existen marcadas diferencias entre la predicción del desempeño en deportes individuales y colectivos, siendo estos últimos los más estudiados por ser los más complejos.

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 (Rama˜Iyer, 2009) y la prevención de lesiones son algunos ejemplos.

Los deportes colectivos se caracterizan por la interacción entre los miembros de equipo. En este caso, la cuantificación de los resultados individuales debe tener en cuenta el aporte realizado por cada deportista a todo el equipo. La planificación de las estrategias tiene una base en la predicción y el análisis del contrario. Hasta el momento, de entre todos los deportes colectivos, el fútbol, el baloncesto y el béisbol han sido objeto del mayor número de investigaciones en este campo(Dubbs, 2015).

Varios han sido los sistemas computacionales desarrollados con fines predictivos, cada uno una con mayor o menor complejidad y fiabilidad. Entre ellos caben destacar los siguientes:

- **Advanced Scout:** sistema desarrollado por IBM para automatizar todo el proceso de minería de datos en juegos de baloncesto de la NBA (Bhandari *et˜al.*, 1997).
- **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 (Scout, 2016).
- **Inside Edge:** es un *software* pionero en el estudio y recopilación de datos de béisbol, actualmente es un líder mundial en exploración y análisis de datos en ese deporte (Edge, 2016).

La ventaja de jugar en casa

Desde el punto de vista predictivo, la ventaja de jugar en casa o HA ha sido objeto de estudio desde hace varias décadas como un caso particular. Esta permite conocer si existe una asociación entre el porcentaje de victorias conseguidas por el equipo local respecto a al equipo visitante (Courneya y Carron, 1992; Pollard, 1986). En deportes de equipo, se ha reportado que esta ventaja puede ser de alrededor del 60 % (Jamieson, 2010), habiéndose particularizado estudios en varios deportes colectivos específicos (Gomez *et˜al.*, 2011).

Las causas que podrían explicar el concepto del HA son inconclusas y puede tener una explicación multifactorial. Pollard y Pollard (2005) aluden a la interacción de hasta siete causas que podrían explicar el fenómeno de HA tales como: psicológicos, tácticos, territorialidad, familiaridad con el lugar, parcialidad del árbitro, apoyo del público y viajes previos al partido. Las relaciones entre los factores están desigualmente distribuidas e interaccionan entre sí. Así, factores como la parcialidad arbitral estaría condicionada por el apoyo del público. En este sentido, se conoce la relación entre el acierto del árbitro y el ruido provocado por los seguidores de los equipos locales como un producto de la densidad del público asistente, pudiendo provocar decisiones erróneas en los árbitros. Además, otro ejemplo podría ser la señalización por parte los árbitros de una menor cantidad de conductas antirreglamentarias al equipo local que al equipo visitante. En cualquier caso, son requeridas más evidencias científicas que lo corroboren. Por otro lado, otros factores como los psicológicos podrían estar relacionados con la familiaridad, viajes previos para llegar al lugar del partido y/o el apoyo del público, afectando todos ellos de forma directa o indirecta al resultado del partido (Jamieson, 2010). Cabe señalar que probablemente los jugadores asumen estrategias más ambiciosas al jugar como local, pudiendo ser este un factor táctico–estratégico que se debe tener en cuenta a nivel profesional.

En este sentido, los equipos locales suelen competir de manera más efectiva en relación a las acciones ofensivas de sus rivales visitadores, lo cual quizás esté condicionado por las expectativas de los entrenadores (Staufenbiel *et˜al.*, 2015). Se ha demostrado que la HA es inversamente proporcional a la duración de la temporada para varios deportes(Jamieson, 2010). Además, otro factor asociado a la ventaja de jugar en casa es la territorialidad, entendida como la ventaja de jugar en una zona geográfica afín al grupo y la identidad cultural de éste.

2.1.2.3. Estudios macro-económicos y de mercado

En la actualidad, el deporte ha pasado de ser una simple manifestación social, destinada a la contemplación y práctica de actividades recreativas en busca de un cierto entretenimiento o satisfacción personal, a ser considerado como un bien, cuya producción, consumo, financiación y gestión responde a criterios de racionalidad económica (Baumer y Zimbalist, 2014). Por un lado, el deporte ha abierto a la economía nuevos y rentables mercados, así como distintas oportunidades de empleo hasta hace poco tiempo desconocidos. Por otro, la economía ha dotado al deporte de una estructura de pensamiento diferente para adoptar sus decisiones, valorar sus relaciones institucionales y evaluar sus consecuencias materiales. Se ha pasado de esta manera, de una situación caracterizada por una tradicional ausencia de lo económico en el contexto deportivo, a otra en la que las relaciones ideológicas y de acuerdo con el valor, las 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 la economía del deporte que la distinguen de otras ramas de la economía es que las empresas, bien sean los clubes en los deportes de equipo o los deportistas en el caso de los deportes individuales, necesitan de competencia para maximizar sus beneficios, no pudiendo aspirar a monopolizar el mercado. Esto se debe a que el deporte es precisamente un espectáculo competitivo. Esta y otras peculiaridades son las que han propiciado un lento pero continuo avance de la disciplina en los últimos años.

Las apuestas deportivas

Desde el propio surgimiento del deporte, las apuestas han resultado ser una parte importante la cual ha potenciado su desarrollo (Woodland y Woodland, 1994). El mercado electrónico de apuestas deportivas ha despertado un gran interés 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 y Weinbach, 2009; Spann y Skiera, 2009; Sauer *et˜al.*, 2010).

A continuación se identifican tres modelos fundamentales empleados por las agencias que manejan apuestas deportivas. En el contexto del aprendizaje automático, el primer modelo se ha abortado usando métodos de regresión, mientras que el último ha sido el más estudiado por constituir un problema clásico de clasificación binaria:

- ***Over-under:*** la apuesta tiene en cuenta la diferencia de los puntos obtenidos entre los equipos rivales.
- ***In-line:*** las apuestas se realizan durante el transcurso de la competición.
- ***Money-line:*** el resultado de la apuesta solo está determinado por conocer cuál será el equipo ganador.

La Tabla 2.2 ejemplifica el funcionamiento del modelo de apuestas *money-line*. Para cada partido, las agencias de apuestas identifican un equipo favorito (FAV) y un probable perdedor (RET), asociando los beneficios económicos en función del posible ganador.

Equipo favorito (FAV)	**Equipo retador (RET)**	**FAV-*line***	**RET-*line***
A	B	-300	+240
B	A	-100	+100

Tabla 2.2: Un ejemplo representativo del modelo de apuestas *money-line*.

La columna FAV-*line* indica cuánto dinero es necesario apostar para ganar un beneficio de \$100, mientras que la columna RET-*line* cuanto dinero se ganaría en caso de apostar \$100. En el primer caso, el equipo A fue considerado como el favorito por lo que en caso de apostarle \$100 y que ganase dicho equipo equivaldría a una remuneración de \$33.33, mientras que apostarle esa misma cantidad al equipo retador B significaría una ganancia de \$240 en caso de que éste ganase (el signo negativo se usa para indicar cual es el equipo favorito). En el segundo caso, las apuestas se encuentran perfectamente balanceadas, por lo que no es posible identificar un equipo favorito.

La Tabla 2.3 muestra las probabilidades de victoria o derrota atendiendo al modelo de apuestas *money-line*. Sea x la cantidad de dinero apostado para ganar otra cantidad y con una probabilidad de victoria p, entonces podemos esperar ganar una cantidad py o perder $x(1-p)$, a partir de lo cual se deduce la Ecuación 2.1.

$$py = (1-p)x$$

$$py = x - px$$

$$px + py = x$$

$$p(x+y) = x$$

luego

$$p = x/(x+y) \tag{2.1}$$

***Money-line* (FAV)**	***p***	***Money-line* (RET)**	***p***
100	50.00 %	100	50.00 %
-110	52.38 %	110	47.62 %
-120	54.55 %	120	45.45 %
-130	56.52 %	130	43.48 %
-140	58.33 %	140	41.66 %
-150	60.00 %	150	40.00 %
-200	66.66 %	200	33.33 %
-300	75.00 %	300	25.00 %
-400	80.00 %	400	20.00 %
-500	83.00 %	500	16.66 %
-1000	90.91 %	1000	9.09 %
-2000	95.24 %	2000	4.76 %
-3000	96.77 %	3000	3.23 %
-4000	97.56 %	4000	2.44 %
-5000	98.04 %	5000	1.96 %
-10000	99.01 %	10000	0.01 %

Tabla 2.3: Probabilidades de victoria y derrotas en apuestas *money-line*.

2.2. El juego de béisbol

El béisbol es un deporte colectivo de cooperación y oposición el cual se practica entre dos equipos de 9 jugadores cada uno. Se considera un deporte estratégico, pues la toma de decisiones es el elemento principal que guía la dinámica de las acciones y define la

victoria o derrota. El béisbol es considerado uno de los deportes más populares alrededor del mundo, siendo practicado en casi todos los continentes, especialmente en América.

En esta sección se exponen los aspectos que han hecho del béisbol uno de los deportes más estudiados desde el punto de vista estadístico. En particular, se abortan los elementos fundamentales de la sabermetría, la cual complementa y cuestiona la manera en que tradicionalmente se ha llevado a cabo su análisis estadístico.

2.2.1. Reglas del juego

El juego de béisbol (del inglés *baseball*) consiste en golpear una pelota con un bate de madera, desplazándola a través del terreno, y correr por el campo de juego buscando alcanzar la mayor cantidad de bases posibles hasta dar la vuelta a la base desde donde se bateó denominada *homeplate*. El principal objetivo es lograr anotar la carrera mientras que los jugadores defensivos buscan la pelota bateada para eliminar al jugador que bateó o a otros corredores antes de que éstos lleguen primero a alguna de las bases o consigan anotar la carrera. El equipo que anote más carreras al finalizar los nueve episodios también llamados *innings* que dura el encuentro es el que resulta ganador. Si al término de los nueve *innings* regulares persiste un marcador igualado en carreras, entonces el juego se extiende cuanto sea necesario hasta que haya un ganador. Según las reglas básicas del béisbol no existe el empate, éste solo es permitido en ligas de aficionados e infantiles para limitar el desgaste de los jugadores. La Figura 2.4 muestra las 9 posiciones de los jugadores en el terreno, la posición del lanzador es reconocida como la de mayor importancia en este deporte.

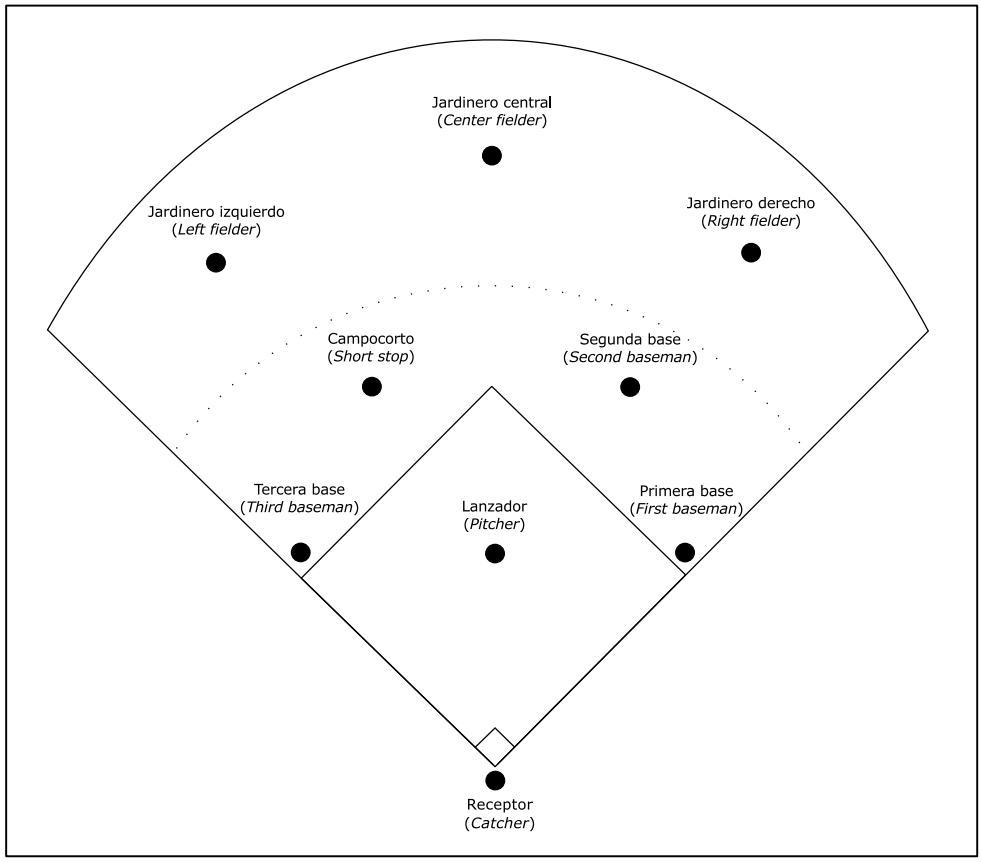


Figura 2.4: Posiciones de los jugadores de béisbol en el terreno.

El lanzador dispone de hasta tres lanzamientos erróneos o bolas, pues una cuarta bola permitiría al bateador avanzar a primera base. Por otro lado, el bateador dispone de dos fallos o *strikes* al batear, pues un tercero lo dejaría ponchado o puesto *out*. El recuento de lanzamientos y bateos malos se denomina cuenta. Si el bateador falla el tercer *strike* y el receptor pierde la bola y se le aleja, el bateador puede correr hacia la primera base, y si llega primero que el tiro del receptor, cuenta el ponche, pero no el *out* (esto es siempre y cuando la primera base esté desocupada).

El árbitro principal se coloca detrás del *homeplate* o cajón de bateo y decide si los lanzamientos son buenos (*strikes*) o malos (bolas). Este puede consultar a los otro árbitros ubicados junto a las líneas de cal de primera y tercera base si el bateador pasó el bate por la línea sin darle a la bola.

2.2.1.1. Bateo

Si el bateador consigue batear bola, hay básicamente cuatro posibilidades en el béisbol:

1. Si la bola toca el suelo antes de que ningún jugador a la defensiva la atrape, y el bateador logra alcanzar la primera base antes de que los defensores consigan tocar con la bola o pisar la primera base (*out*), se denomina sencillo. Si el bateador consigue llegar hasta la segunda base, sin que el equipo a la defensiva cometa ningún error, se denomina doble, si llega hasta la tercera base (sin error de la defensa) se denomina triple, y si llega al *homeplate* (sin error de la defensa) se denomina cuadrangular.

2. Si la bola es atrapada en el aire sin que toque el suelo por un jugador a la defensa, el bateador queda eliminado. Si la bola cae dentro del terreno de juego, el bateador tiene que correr, no puede esperar a ver si el batazo es malo.

3. Si la bola es bateada pero está fuera de los límites laterales de la zona de juego se denomina *foul*. En ese caso el bateador se suma en su cuenta un *strike*, si en su cuenta tiene menos de dos *strikes*, pero si el bateador tiene como cuenta 2 *strikes* y batea un *foul*, este seguirá bateando sin ser *out*. Ahora bien si la bola cae fuera de los límites laterales de la zona de juego, pero un jugador de la defensa la atrapase en el aire (aunque esté fuera de los límites laterales de la zona de juego), el bateador también quedaría eliminado.

4. Si la bola sale volando por encima del límite de fondo de la zona de juego, es un cuadrangular, es decir, el bateador da la vuelta al cuadro hasta llegar al home y se anota una carrera. Si además había alguno de sus compañeros en las bases, ellos también corren hasta el home y anotan carreras, una por cada jugador que hubiera en base y otra que se anota el bateador.

Si se consigue un cuadrangular con tres jugadores ocupando la primera, segunda y tercera base, es decir, con las bases llenas, se denomina *grand slam*, y se anotan 4 carreras. Para que el lanzamiento del *pitcher* sea bueno debe pasar por encima del *homeplate* a la altura determinada desde las axilas hasta las rodillas del bateador. Si el lanzamiento no cumple con estos requisitos, el árbitro la denominará bola.

Antes del inicio del partido, cada equipo debe presentar el orden de bateo de sus nueve jugadores. Este orden se respeta en cada una de las entradas. Los bateadores se pueden sustituir de manera permanente, por lo que un bateador que sale del campo de manera prematura no puede volver a jugar en el partido. Típicamente, los dos primeros bateadores

buscan obtener bases, por lo que éstos intentan poner la pelota en juego y correr velozmente hacia las bases. El tercer bateador suele ser el mejor, ya que busca remolcar a los dos bateadores anteriores, y de ser posible anotar un cuadrangular o correr para ganar bases. El cuarto bateador suele ser muy potente, buscando anotar un cuadrangular para remolcar a sus compañeros en base. El quinto y sexto bateador suelen realizar sacrificios, es decir, intentan realizar contacto para remolcar a los jugadores en base sin importar si ellos obtienen bases. El resto de los bateadores suelen ser los menos habilidosos del equipo.

2.2.1.2. Picheo

El lanzador o *pitcher* tiene que hacer contacto en la placa que está ubicada en el centro del montículo para empezar a lanzar, cuando esto sucede se pone viva la jugada y el *pitcher* puede proceder a lanzar la pelota. Cuando el pitcher hace una jugada llamada revire, esto se puede explicar con el movimiento que hace el mismo para lanzar de la placa o loma de pitcheo a alguna de las bases (primera base, segunda base, tercera base) para tratar de sacar *out* al jugador del equipo contrario que se encuentra corriendo en alguna de las bases ya mencionadas.

El pitcher, al presentar la pelota, no podrá hacer un movimiento de engaño para lanzar a las bases a menos que realmente tire, de ser lo contrario se marca un *balk*, esto quiere decir que el corredor avanza una base con la autorización del árbitro. Este engaño puede ser dado por un movimiento brusco en los hombros o literalmente por un movimiento de tirar a la base y no soltar la bola. El director del equipo puede pedir tiempo y entrar a una consulta con el lanzador, cuando esto sucede se cuenta como una entrada legal. No pueden haber dos entradas legales en el mismo *inning*, si es así, el lanzador por regla tendrá que ser cambiado del partido por un jugador de la banca.

2.2.2. Sabermetría

El béisbol es un deporte colectivo sumamente complejo. Dicha complejidad favorece la extracción de una gran cantidad de datos de diversa índole relacionados con el juego, los cuales hacen del béisbol uno de los deportes más completos en cuanto a estadísticas se refiere (Thorn *et˜al.*, 1984). Actualmente existe una tendencia en el estudio del juego de béisbol conocida como sabermetría, la cual complementa y cuestiona la manera en que tradicionalmente se ha llevado a cabo su análisis estadístico (Wolf, 2015).

Definición 2.2. La sabermetría se define como el análisis empírico del juego de béisbol mediante el estudio de la evidencia objetiva obtenida, específicamente usando técnicas cuantitativas, cuyo fin es medir de manera eficaz las actividades que se suscitan en el terreno de juego (James, 1982).

El término es derivado del acrónimo SABR, el cual hace referencia a la Sociedad para la Investigación del Béisbol Americano (Davids, 1971), y no fue acuñado sino hasta 1980, cuando Bill James hizo referencia al mismo a través de uno de sus famosos escritos sobre béisbol conocidos como «*Baseball Abstracts*».

Pero el suceso que realmente dio a conocer a escala global la filosofía propuesta por la sabermetría lo constituye sin dudas la publicación del libro «*Moneyball*» (Lewis, 2004). En éste se detallan las estrategias que utilizó la oficina de los Atléticos de Oakland, dirigidas por su gerente general Billy Beane, con el fin de hacer más productivo su reducido presupuesto. «*Moneyball*» no es un libro solo de béisbol, sino más bien de estrategia de negocios basado en la premisa de que en un mercado competitivo, el administrador audaz (de mercado pequeño) debe diferenciarse de sus competidores a través de la optimización de sus recursos mediante el uso de estrategias únicas pero eficientes. En el caso de los Atléticos, la estrategia era la de adoptar un método de selección concentrado en obtener jugadores con altos porcentajes de efectividad al embasarse, sin tomar en cuenta los métodos y estadísticos clásicos de selección usados hasta ese momento, sino con otros pertenecientes al campo de la sabermetría. Bajo este concepto y con la ayuda de genios de la matemática y la estadística, Beane logró clasificar a los Atléticos de Oakland a la postemporada durante cuatro años seguidos, sin contar con un presupuesto exorbitante ni con firmas de consideración, cambiando en gran medida la forma en que serían realizados los juicios de valor de los jugadores de béisbol a partir de esa fecha (Chang y Zenilman, 2013).

2.2.2.1. Estado actual del tema

La sabermetría se concentra fundamentalmente en evaluar cómo afectan las estadísticas individuales y colectivas de los jugadores al margen de juegos ganados y perdidos de los equipo de béisbol, tal como muestra la Figura 2.5. Siguiendo esta lógica, para que un equipo sea exitoso éste debe ganar más juegos que sus oponentes, lo cual se logra anotando carreras o por medio de la prevención de éstas. Es por ello que la sabermetría se enfoca mayormente en la medición del aporte individual de cada jugador en términos de anotación

de carreras, lo cual a la postre se traduce en victorias para todo el equipo.

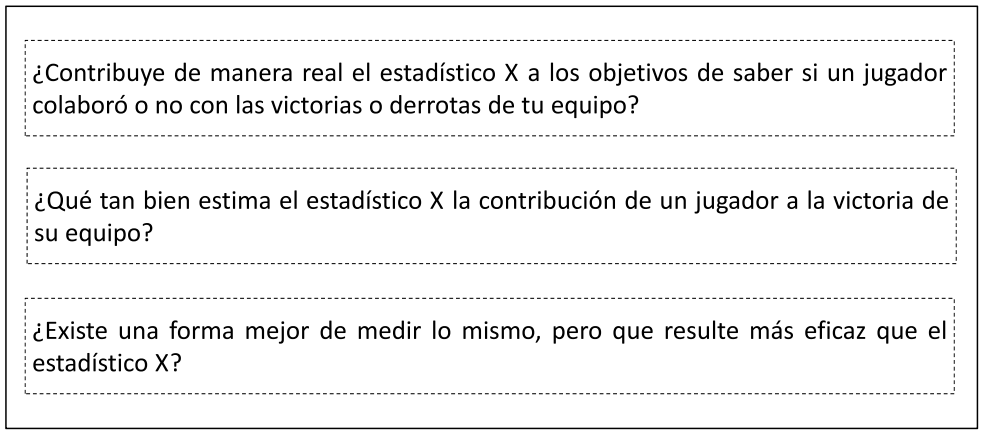


Figura 2.5: Principales interrogantes del béisbol que intenta responder la sabermetría.

Entre los usos más comunes de la sabermetría están los de evaluar el desempeño de diferentes jugadores durante una temporada con el fin de determinar quiénes son merecedores de premios como el MVP o Cy Young (Stephen˜Ockerman, 2014), o para la comparación de jugadores de diferentes épocas (Costa *et˜al.*, 2007). Además, la sabermetría busca predecir el desempeño de un jugador en el futuro y así poder estimar su competitividad con fines de uso o para su contratación en equipos a través de agencias libres. Otra función de la sabermetría es la de analizar conceptos arraigados en el béisbol que no han sido lo suficientemente estudiados, con el fin de estimar su veracidad y eficacia. Ejemplo de esto sería el estudio del efecto que tiene un estadio determinado en las estadísticas de un jugador en particular, o la medida en que contribuye la defensa a la conquista de victorias de un equipo determinado (Beneventano *et˜al.*, 2012).

2.2.2.2. Sabermetría vs estadísticos tradicionales

Los estadísticos tradicionales nacieron para tratar de soportar objetivamente con números los juicios de valor que se hacen de los jugadores. El Promedio de Bateo (AVE), las Carreras Impulsadas (RBI), las Bases Robadas (SB) o el Promedio de Carreras Limpias (ERA) de los lanzadores son estadísticos tradicionales que han brillado en todos los medios que se ocupan de cubrir los resultados del béisbol. No obstante, su valor para determinar por qué se ganan o pierden los juegos está limitado solo al contexto de lo que esas estadísticas querían medir cuando fueron creadas.

Resulta que los pocos datos que se manejaban hasta la década de los setenta del siglo veinte eran suficientes para soportar los juicios de valor que se hacían de los jugadores de béisbol. En la actualidad, los estadísticos tradicionales no son capaces por si solos de analizar las nuevas variables e ideas que han surgido de la observación del juego. El uso de las tecnologías informáticas ha extendido las capacidades de almacenamiento y procesamiento de datos deportivos a una escala insospechada (Schumaker *et˜al.*, 2010b), es por ello que se ha hecho necesario el diseño de nuevas herramientas y metodologías con el objetivo de explorar las complejas relaciones entre los miles de datos que surgen día a día en cada juego de béisbol (Adler, 2006).

La sabermetría no se trata de menospreciar estadísticos tradicionales tan importantes como el AVE o el SB, sino de integrar nuevas herramientas de análisis que dependan más del talento propio del jugador y que ayuden a determinar cuál es realmente el valor de cada uno para su equipo. La relación entre estadísticos tradicionales y sabermetría es como la relación entre contaduría y finanzas. Las primeras (estadísticos tradicionales y contaduría) describen el pasado con exactitud, mientras que las segundas (finanzas y sabermetría) usan las herramientas a su disposición para entender el pasado y predecir el futuro con la mayor precisión posible.

Por ejemplo, el AVE permite conocer la capacidad que tiene un bateador de no ser puesto *out* teniendo en cuenta solamente los batazos que conecta en sus turnos al bate, tal como se muestra en la Ecuación 2.2. El problema está en que hay otras formas de no ser puesto *out*, como recibir bases por bolas o pelotazos, que son simplemente ignorados en este cálculo, convirtiendo al AVE en un estadístico incompleto para analizar de forma correcta el rendimiento de un bateador. Además, el AVE iguala el valor de todos los batazos sin importar si son sencillos, dobles, triples o cuadrangulares. Por lo tanto, es evidente que este no constituye un estadístico del todo confiable puesto que es incapaz de proporcionar toda la información sobre la capacidad de un bateador de no ser puesto *out*, y además tampoco mide la calidad de sus conexiones.

$$AVE = \frac{H}{AB} \tag{2.2}$$

En este sentido, la sabermetría propone estadísticos como el Porcentaje de Embasado (OBP) cuyo objetivo es medir la capacidad que tiene un bateador de no ser puesto *out* teniendo en cuenta todas las posibilidades (sencillos, bases por bolas, pelotazos etc.). Además, el Slugging (SLG) permite conocer la calidad de sus conexiones asignándole

valores determinados a cada una de las bases alcanzadas en sus turnos al bate. La suma de ambas estadísticas da como resultado el estadístico conocido como OPS (OBP + SLG), una especie de calibre del bateador similar al que se obtiene con el AVE en relación a la marca de .300, pero mucho más completo que éste.

$$OPS = \overbrace{\left(\frac{H + BB + HBP}{AB + BB + HBP + SF}\right)}^{OBP} + \overbrace{\left(\frac{1B + (2 * 2B) + (3 * 3B) + (4 * 4B)}{AB + SF}\right)}^{SLG} \tag{2.3}$$

Para la sabermetría el promedio de bateo es más útil si se le calcula aislando en dicho cálculo a aquellas situaciones en las cuales la pelota es puesta en juego de aquellos turnos al bate en los cuales no hay interacción con la defensiva, como los ponches y los cuadrangulares. Por ello, el llamado Promedio de Bateo con Pelotas en Juego (BABIP) mostrado en la Ecuación 2.4, resalta la habilidad para conectar un batazo solo cuando se pone la pelota en juego y la defensiva tiene oportunidad de intervenir. O sea, el BABIP excluye los ponches y cuadrangulares del cálculo tradicional del promedio de bateo, constituyendo una métrica muy efectiva cuando se le relaciona con los porcentajes de *rollings* (batazos en los que la trayectoria de la pelota es rastrera y no se eleva en ningún momento del terreno de juego) y *flies* (batazos que elevan la pelota por los aires sin tocar antes el terreno de juego), pudiéndose inferir de este modo las causas de ciertos periodos de declive ofensivos en los bateadores.

$$BABIP = \frac{H + HR}{AB – K – HR + SF} \tag{2.4}$$

Otro ejemplo es el caso del estadístico tradicional RBI, el cual es un indicador muy usado para medir la habilidad individual de un bateador para conectar un batazo que permita al mismo bateador (con un cuadrangular), o a otro corredor en base anotar una carrera. Esto tiene mucho que ver en el sentido de buscar héroes y villanos en el béisbol. Pero por ser una estadística acumulativa que no depende por entero del bateador, el estadístico RBI no permite conocer su efectividad real para impulsar carreras. Por ejemplo, se podría dar el caso de tener bajos resultados en este indicador simplemente porque las conexiones realizadas no encuentran suficientes corredores en base.

En este sentido, desde la perspectiva de la sabermetría, el incremento de las carreras anotadas de un equipo se logra más a través del esfuerzo colectivo al mezclar sabiamente los

talentos individuales en el orden ofensivo que por actos de heroicidad individual. Mientras más veces consigan una base los jugadores y mientras más bases alcancen con sus conexiones encontrando corredores en posiciones anotadoras, más carreras anotará el equipo. De este modo, quien las anota y quien las impulsa es una resultante de la posición en el orden al bate de cada miembro del equipo. Lo cual depende fundamentalmente de que aquellos jugadores con mejores habilidades para alcanzar una base se encuentren en base mientras los bateadores con mayor fuerza para conectar extrabases estén al bate. Por eso en las historia del béisbol es muy difícil hallar bateadores que hayan liderado el renglón de carreras impulsadas bateando toda la temporada en puestos de la alineación en los cuales encontraron pocos corredores en base, como sucede tradicionalmente con los octavos y novenos bates.

Teniendo todo esto en cuenta, la sabermetría hace una estimación muy real de la productividad del equipo para anotar carreras al comparar las carreras anotadas como equipo con el producto del OBP y el SLG de todos sus bateadores en sus turnos al bate. A ese producto se le suele llamar Carreras Creadas (RC) y constituye uno de los aportes teóricos más importantes de la sabermetría para cálculo del potencial ofensivo real de los equipos. La Ecuación 2.5 muestra la fórmula inicial propuesta por Bill James, con el paso del tiempo ha tenido varias modificaciones donde se le da una mayor o menor importancia al OBP, así como se añaden otros estadísticos en su cálculo tales como el promedio con hombres en posición anotadora, el éxito en el robo de bases etc.

$$RC = \frac{TB(H + BB)}{AB + BB} \tag{2.5}$$

La sabermetría no se ha limitado solamente al aspecto ofensivo en el béisbol. Numerosos estadísticos de la sabermetría han surgido para evaluar los aspectos defensivos, y particularmente el desempeño de los lanzadores. La Ecuación 2.6 muestra el cálculo del popular Pitcheo Independiente de Fildeo (FIP), reconocido como un excelente medidor de la efectividad del lanzador calculado únicamente sobre la base de los ponches, boletos y cuadrangulares recibidos, los cuales son estadísticos que no dependen de la defensa.

$$FIP = \frac{13HR + 3BB – 2K}{IP + FIP_{Const}} \tag{2.6}$$

donde FIP_{Const} es un escalar utilizado para ajustar el valor del FIP al promedio de carreras limpias de la liga y suele calcularse del modo siguiente:

$$FIP_{Const} = ERA_{lg}\left(\frac{13HR_{lg} + 3(BB_{lg} + HBP_{lg}) – 2K_{lg}}{IP_{lg}}\right)$$

Cada uno de los análisis anteriores fundamentan el por qué surge la sabermetría como una necesidad de perfeccionar los estadísticos usados tradicionalmente en el béisbol. Cabe señalar que la sabermetría se ha encargado además de estudiar cómo afecta al resultado del juego el contexto alrededor de los jugadores, tales como las dimensiones de los terrenos de juego, el arbitraje, o la mano de lanzar y de batear tanto del lanzador como del bateador etc.

2.2.2.3. Algo más que estadísticos individuales

En la búsqueda de la mejor evaluación integral de un jugador, y para lidiar con las relaciones económicas entre jugadores y dueños de equipos, la sabermetría ha desarrollado conceptos tales como la Expectativa Pitagórica (PE), la Probabilidad de Victoria (Log5) o las Victorias Sobre el Reemplazo (WAR). Este último es uno de los estadísticos más polémicos que existe en el béisbol actualmente.

La Ecuación 2.7 muestra la forma tradicional del cálculo de PE, la cual estima cuántos juegos debería ganar un equipo atendiendo al número de carreras anotadas y permitidas (James, 1982). Puede ser usado como un evaluador de la «suerte» que ha tenido un equipo en su liga. El número esperado de victorias sería el resultado de multiplicar PE por el número de juegos jugados por el equipo.

$$PE = \frac{(\text{Carreras anotadas})^2}{(\text{Carreras anotadas})^2 + (\text{Carreras permitidas})^2} \tag{2.7}$$

El estadístico Log5 fue también creado por por Bill James (James, 1982) y tiene el propósito de estimar la probabilidad de que el equipo A derrote al equipo B, dados los promedios de victorias p_A de A y p_B de B. La Ecuación 2.8 muestra la forma en que se realiza su cálculo. Cabe señalar que dicho estadístico ha sido aplicado satisfactoriamente, además del béisbol, en varios otros deportes.

$$Log5 = \frac{p - pq}{p + q - 2pq} \tag{2.8}$$

Por otro lado, el WAR tiene el ambicioso propósito de aglutinar dentro de un marco integral

los valores ofensivos y defensivos de un jugador en una posición específica, con el fin de calcular el costo para el equipo de un hipotético reemplazo de jugador y cuyo resultado son (en teoría) las victorias adicionales que aporta tener a ese jugador en lugar de tener a un hipotético jugador de ligas menores en su misma posición (que sería la opción más rápida y menos costosa a la que recurriría el equipo para sustituirlo). Otra lectura más pragmática del estadístico WAR pudiera ser la siguiente: «Si un jugador se lesiona y tiene que ser reemplazado por otro de una liga menor ¿cuánto perdería o ganaría el equipo con la sustitución?». Un jugador de posición que tenga solo perfil defensivo será más rápido y económico de reemplazar con un jugador de liga menor que cumpla solo con el trabajo defensivo y más difícil de reemplazar si se busca un jugador también con aporte ofensivo.

Aún no existe un criterio unificado respecto a las variables a tener en cuenta a la hora de medir el WAR, por lo la mayoría de sus ecuaciones son propietarias. Esto ha dado como resultado la formulación de diversos estadísticos relacionados (WARP, rWAR, fWAR etc.) y un gran debate en torno al tema en sitios como Fangraphs[1] y TangoTiger[2]. Recientemente fue liberado un paquete de software, en el lenguaje de programación `R`, el cual hace posible el cálculo del llamado OPENWAR, cuyas potencialidades han sido teóricamente bien fundamentadas en Baumer *et˜al.* (2015).

A manera de resumen

La sabermetría surgió dada la necesidad de perfeccionar los estadísticos usados tradicionalmente en el juego de béisbol. La misma se sustenta gracias al aporte en ideas de un gran número de personas, desde jugadores, entrenadores y cazatalentos pasando por periodistas y analistas hasta los más diversos profesionales que de alguna u otra forma se involucran en el estudio de este deporte. La sabermetría ha cambiado significativamente la concepción del juego que se tenía décadas atrás. En la actualidad, los conceptos y principios de la sabermetría constituyen referencia obligada para todos aquellos que deseen realizar análisis estadísticos precisos a partir de la gran cantidad de datos que se generan continuamente. Desde matemáticos, físicos y médicos hasta abogados y agentes de jugadores utilizan los conceptos de la sabermetría, perfeccionándolos y enriqueciéndolos día a día, lo que hace de la esta una perspectiva de juego sumamente útil y dinámica.

[1] http://www.fangraphs.com
[2] http://www.tangotiger.com

2.3. Sumario

Este capítulo estuvo dedicado al análisis cuantitativo de datos deportivos, así como a las aplicaciones del aprendizaje automático en este dominio. Específicamente, fueron planteadas las principales características de este tipo de datos y los retos particulares que enfrenta el deporte en el área de la predicción. Se identificaron tres objetivos fundamentales en el análisis cuantitativo de datos deportivos: el análisis del desempeño deportivo, la predicción de resultados competitivos y la realización de estudios macro–económicos y de mercado.

Se dedicó por entero una sección al estudio del juego de béisbol, reconocido actualmente como uno de los deportes más completos en cuanto a estadísticas se refiere, por ser éste uno de los deportes más estudiados desde el punto de vista cuantitativo. En particular, se detallaron los aspectos esenciales de la sabermetría, la cual representa una forma novedosa de analizar todo lo acontecido en el juego de béisbol.

A partir de los argumentos presentados en este capítulo, cabe resaltar los siguientes aspectos:

- 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 del aprendizaje automático en el contexto deportivo brinda importantes ventajas para su análisis cuantitativo, en comparación con las técnicas estadísticas tradicionales.
- El béisbol es uno de los deportes más complejos, y a su ves más completos, en cuanto a estadísticas se refiere, de ahí que ha sido uno de los más estudiados.
- La sabermetría es la base del análisis cuantitativo en el béisbol, sus descubrimientos y conceptos representan un paso de avance para todos los teóricos de ese deporte.

Capítulo 3

Predicción de resultados de juegos de béisbol usando métodos del aprendizaje automático

En este capítulo se realiza un estudio comparativo de cuatro métodos del aprendizaje automático supervisado, los cuales fueron empleados en la predicción de resultados de juegos de béisbol de la MLB. En la Sección 3.1 se reflejan algunas de las características particulares de este problema de predicción. La Sección 3.2 está dedicada a la descripción de los experimentos a partir del modelo de predicción propuesto, desde el proceso de obtención y manejo de los datos hasta la ejecución de los algoritmos de aprendizaje. En la Sección 3.3 se muestran los resultados obtenidos con los diferentes métodos y esquemas de aprendizaje propuestos, los cuales son comparados entre sí usando métodos estadísticos y además se realiza un estudio respecto a los resultados del mercado de apuestas. Finalmente, la Sección 3.4 presenta algunas conclusiones que se desprenden a partir de los resultados obtenidos.

3.1. El problema de la predicción de juegos de béisbol

La MLB constituye en la actualidad un negocio multimillonario. Esto ha motivado la realización de grandes esfuerzos con el objetivo de desarrollar sistemas que arrojen buenos resultados en la predicción de juegos particulares. En este sentido se han dado grandes avances, sobre todo gracias a los aportes del análisis estadístico y la sabermetría.

Sin embargo, la mayoría de los sistemas predictivos creados a partir de los conceptos sabermétricos están demasiado influenciados por quien los utiliza. Esto ocurre debido a

que las estadísticas de desempeño son en su mayoría valores numéricos que pueden ser tenidos o no en cuenta por los expertos a la hora de realizar algún juicio valorativo o de predicción. Por tal motivo se hace necesario desarrollar sistemas expertos que no estén influenciados por las emociones humanas. Dichos sistemas deberán ser capaces, no solo de realizar buenas predicciones de resultados competitivos en este deporte, sino también de evaluar el alcance y las potencialidades de la sabermetría para la predicción de juegos de béisbol.

Dada la naturaleza cuantitativa del béisbol, una gran cantidad de datos de este deporte se encuentran disponibles actualmente de manera pública, ya sea en forma de variables numéricas o como estadísticos simbólicos. Sin embargo, la mayoría de los estudios realizados a partir de estos datos se originan desde el punto de vista del análisis macro-económicos o como respuesta a intereses del mercado deportivo profesional(Baumer y Zimbalist, 2014; Chang y Zenilman, 2013; Sauer *et˜al.*, 2010; Witnauer *et˜al.*, 2007). Por consiguiente, y a pesar de los avances de la sabermetría, la predicción de resultados de juegos de béisbol no ha recibido demasiada atención en los últimos tiempos (Sykora *et˜al.*, 2015).

La mayoría de los estudios de la sabermetría relacionados con la predicción se concentran en analizar y ordenar jugadores de acuerdo a sus cualidades con el propósito de identificar los más capaces (Stephen˜Ockerman, 2014; Robinson, 2014; Lyle, 2007). Dado que le objetivo final del béisbol profesional es que un equipo pueda ganar más juegos, se hace necesario prestar una atención particular a este aspecto. Un aspecto a tener en cuenta es que se ha reportado que la predicción de resultados en deportes colectivos usando técnicas del aprendizaje automático es uno de los problemas más complejos en este dominio de aplicación (Leung y Joseph, 2014; Gumm *et˜al.*, 2015).

3.1.1. Contaminación de los datos

Durante la predicción de resultados deportivos existe un fenómeno conocido popularmente como «contaminación» de los datos. Este ocurre cuando los datos utilizados por el modelo predictivo contienen «conocimiento sobre el futuro», y constituye un problema a tener en cuenta durante el preprocesamiento y manejo de los datos (Yuan *et˜al.*, 2015). Si el modelo realiza su predicción tomando en consideración datos obtenidos durante o posteriores al encuentro deportivo, entonces sus resultados no tendrán un verdadero valor desde el punto de vista predictivo.

Es un problema grave para el aprendizaje automático si los datos con los cuales se entrena

un modelo determinado contienen información implícita a los datos futuros. Por ejemplo, si ejecutásemos un algoritmo de selección de atributos para un conjunto de datos de béisbol este revelaría que un atributo potencialemente relevante para la predicción del resultado del juego es la longitud del partido en *outs*. Esto se debe a que un juego de béisbol normal consta de 9 entradas para un total de 27 *outs*, pero cuando los equipos llegan al final empatados los juegos extienden más allá de estas 9 entradas. En esta situación se ha demostrado estadísticamente que el equipo local tienen una importante ventaja. Por tanto, el atributo que representa la duración del juego constituye un excelente predictor del resutado final del encuentro, pero contamina el modelo y por lo tanto no debe ser considerado durante el aprendizaje.

Dado que la utilización de datos contaminados puede conducir a resultados erróneos en la predicción, el modelos propuesto evita ante todo incurrir en este problema. En este sentido, solo se usan datos históricos acumulados para entrenar los diferentes algoritmos de predicción utilizados. Además, las instancias se construyen a partir de las estadísticas previas al desarrollo de cada encuentro, tal como muestra la Figura 3.3.

3.2. Diseño de los experimentos

La minería de datos es una ciencia experimental, por lo que se conoce que no puede existir un algoritmo «universal» que resuelva cualquier tipo de problemas (Wolpert y Macready, 1997). En concordancia con dicho axioma, este estudio sigue la metodología de minería de datos CRISP-DM (Shearer, 2000). Dicha metodología proporciona una manera estructurada de llevar a cabo todo el manejo y análisis de los datos, con la consiguiente mejora en la obtención de resultados más precisos y confiables. El desarrollo con CRISP-DM tiene seis pasos principales, los cuales se enumeran a continuación:

1. Comprensión del dominio del problema y definición de los objetivos del estudio.
2. Identificación, acceso y manejo de las fuentes de datos.
3. Preprocesamiento de los datos.
4. Desarrollo del modelo usando técnicas del aprendizaje automático.
5. Evaluación y validación de la utilidad del modelo de acuerdo a los objetivos del estudio.
6. Despliegue del modelo para su uso en el proceso de toma de decisiones.

En primer lugar, los datos fueron obtenidos de dos de las fuentes de datos gratuitas más populares de la MLB: la organización de béisbol sin fines de lucro denominada Retrosheet, y la base de datos Lahman (ver Apéndice B). Durante el proceso de preparación y preprocesamiento de los datos, los estadísticos más populares de la sabermetría fueron calculados y añadidos de forma acumulativa. Seguidamente, se emplearon varios algoritmos de selección de atributos con el objetivo de reducir y refinar el conjunto original de atributos. Este procedimiento aumenta la precisión y la velocidad de aprendizaje de los algoritmos empleados además de mejorar la comprensibilidad de los resultados (Han *et˜al.*, 2011).

Para la predicción se utilizaron y compararon cuatro métodos del aprendizaje automático diferentes en cuanto a la concepción de su funcionamiento. Dichos algoritmos fueron seleccionados dada su capacidad para modelar tanto problemas de clasificación como de regresión, además debido a su popularidad en la literatura Liao *et˜al.* (2012). La validación de los resultados predictivos se llevó a cabo utilizando validación cruzada en 10 particiones, esto se hizo con el fin de evaluar de manera objetiva la capacidad predictiva de los algoritmos utilizados.

La disposición del modelo de predicción general propuesto, desde la obtención y preprocesamiento de los datos hasta la fase del análisis de los resultados del modelo, se ilustra en la Figura 3.1. Cada una de las partes de su estructura se describen en las siguientes subsecciones.

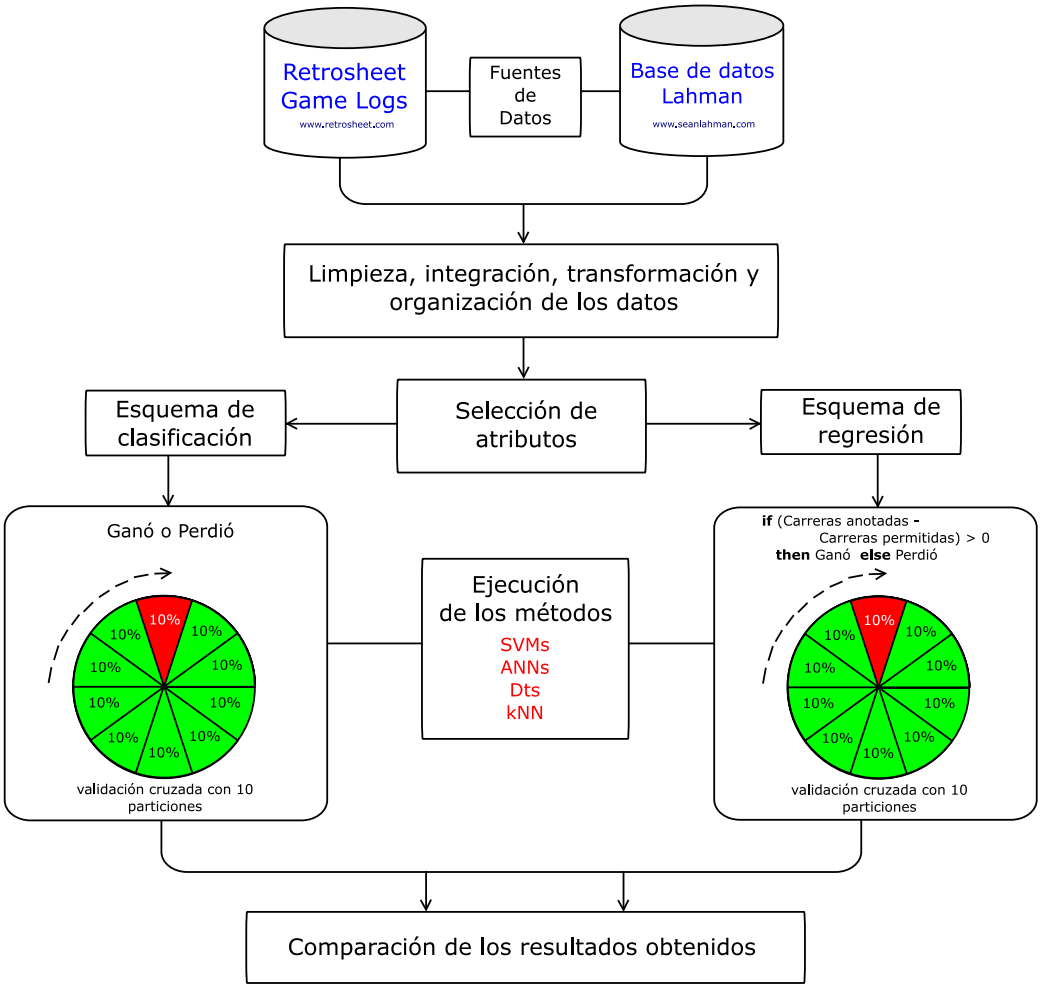


Figura 3.1: Esquema general del modelo de predicción propuesto.

3.2.1. Pre-procesamiento de los datos

La MLB incluye 30 equipos divididos en dos ligas: la Liga Americana (AL) y la Nacional (NL). Cada equipo juega un total de 162 partidos durante la temporada regular desde abril hasta octubre, la cual no comprende los partidos de pre-temporada o los de post-temporada.

Siguiendo los preceptos de la minería de datos y del aprendizaje automático, se decidió incluir la mayor cantidad de información posible en el conjunto de datos inicial. Dichos datos fueron obtenidos a partir de los *game logs* de Retrosheet y la base de datos de Lahman, desde 2005 hasta 2014, representando un total de 1620 partidos para cada equipo. Con el objetivo de llevar a cabo la extracción y el preprocesamiento de los datos fue creada una

biblioteca en el lenguaje de programación Java.

Por medio de este preprocesamiento, se redujo cada juego de cada equipo a un conjunto de estadísticos de bateo, picheo y defensa. Dichos atributos fueron organizados y agregados por cada temporada, obteniéndose un conjunto de atributos estadísticos acumulativos *day-by-day* para las 10 temporadas regulares objeto de estudio.

La Figura 3.2 muestra la representación de una instancia de aprendizaje correspondiente a un partido entre dos equipos A y B. Como se puede observar, cada instancia está conformada por atributos identificadores, atributos sabermétricos, un atributo clase (para el esquema de clasificación) y un atributo objetivo (para el esquema de regresión).

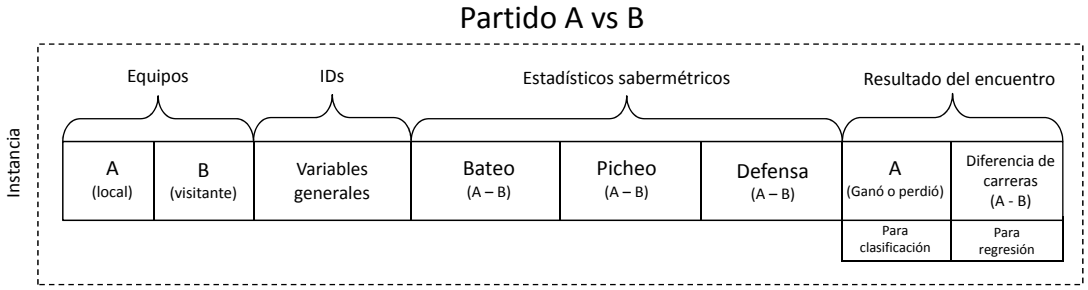


Figura 3.2: Esquema de organización de los datos de un partido entre los equipos A y B.

Con el objetivo de representar los atributos estadísticos previos a la predicción del juego, la información de cada partido fue calculada atendiendo a las diferencias entre los estadísticos acumulativos del equipo local y el visitante. Se determinó representar cada uno de los atributos desde la perspectiva del equipo local. Por ejemplo, el atributo WPDiff (Diferencia entre los porcientos de juegos ganados) representa la resta del porciento de juegos ganados del equipo local y el visitante. De la misma forma, la variable de salida representa el resultado del juego para el equipo local. Si la variable RunDiff (esquema de regresión) toma un valor entero positivo entonces esto significa que el equipo local ganó el partido por esa diferencia de carreras, por otro lado, si RunDiff toma un valor entero negativo entonces significa que el equipo local perdió el partido por esa diferencia de carreras. Así mismo, en el caso del esquema de clasificación la variable de salida es de tipo nominal e indica el resultado del juego desde la perspectiva del equipo local (Ganó o Perdió).

Durante la etapa de preprocesamiento fueron reformulados de forma acumulativa los estadísticos para tener en cuenta lo acontecido antes de la celebración de cada uno de los 162 juegos de la MLB. La Figura 3.3 muestra los valores de cuatro estadísticos acumulativos calculados de acuerdo a los resultados obtenidos antes del inicio de cada juego.

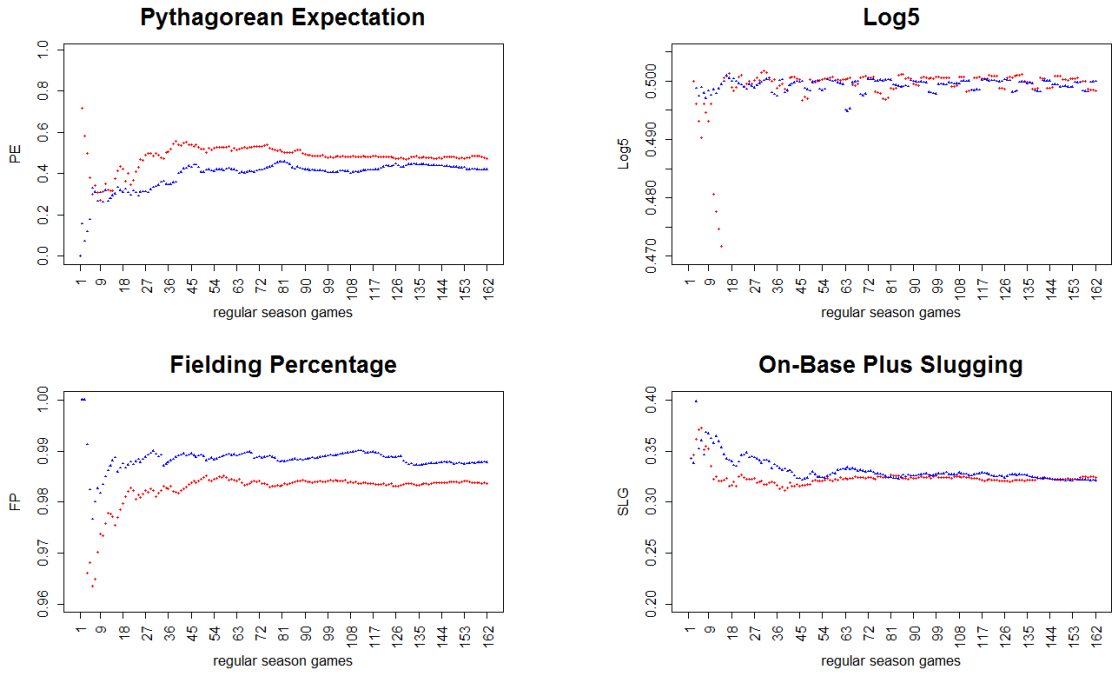


Figura 3.3: Ejemplo de cuatro estadísticos acumulativos correspondientes a los Gigantes de San Francisco (triángulos azules) y los Atléticos de Oakland (círculos rojos) durante la temporada regular de 2014.

El conjunto de datos final contiene los estadísticos acumulativos, sin valores perdidos, para cada uno de los equipos de la MLB durante un período de 10 años. El total de atributos fue de 144, de los cuales 12 representan detalles generales del encuentro tales como el nombre de los equipos, su liga, la fecha del juego, el estadio, la asistencia del público, el clima ese día etc., seguidos por 130 estadísticos sabermétricos de bateo, picheo y defensa, y finalmente las dos variables de salida.

3.2.2. Selección de atributos

Las técnicas de selección de atributos posibilitan la disminución de la dimensionalidad de los conjuntos de datos mediante la eliminación de características irrelevantes o redundantes en los datos. A partir de un conjunto de atributos, esta técnica permite a los algoritmos del aprendizaje automático ejecutarse con una mayor rapidez y de forma más eficaz, al tiempo que mejora los resultados y la comprensibilidad del modelo (Han *et˜al.*, 2011).

Dado que el conjunto de atributos creado en este estudio se considera como bastante grande, esto provoca que no sea fácil distinguir qué atributos son los más importantes

para la tarea de predicción. Además, es posible que muchos de estos atributos sean irrelevante o redundante en la práctica. Atendiendo a esto, y como un paso previo de análisis, fueron utilizados varios métodos de selección de atributos. Dicho métodos cuales están incorporados en el ambiente de aprendizaje automático WEKA. Este es un software de código abierto escrito en Java, creado en la Universidad de Waikato en Nueva Zelanda, y publicado bajo la Licencia Pública General de GNU. Cuenta con una colección de algoritmos de aprendizaje automático para tareas de análisis de datos y modelado predictivo, que se aplican a la minería de datos. Además contiene una interfaz gráfica de usuario que facilita el acceso a sus funcionalidades (Witten *et˜al.*, 2011). WEKA que ha sido empleado en diversos dominios de aplicación, incluyendo el deportivo, con resultados satisfactorios (Lyle, 2007; Trawinski, 2010; De˜Marchi, 2011; Odachowski y Grekow, 2013).

La Figura 3.4 representa de forma esquemática el proceso de selección de atributos propuesto en este estudio. Este consta de tres etapas básicas: generación, evaluación y criterio de parada. En primer lugar, se toma como entrada el conjunto original de atributos, el cual incluye en este caso los 130 atributos obtenidos. A continuación, se inicia la primera etapa de selección de atributos, denominada generación de subconjuntos, en donde se utiliza una estrategia de búsqueda para la producción de los posibles subconjuntos de atributos para su evaluación. Existen varios procedimientos de búsqueda para encontrar el mejor subconjunto del conjunto de atributos original (Dash y Liu, 2003). En este estudio la técnica *attribute ranking* de WEKA ha sido la seleccionada para esta tarea.

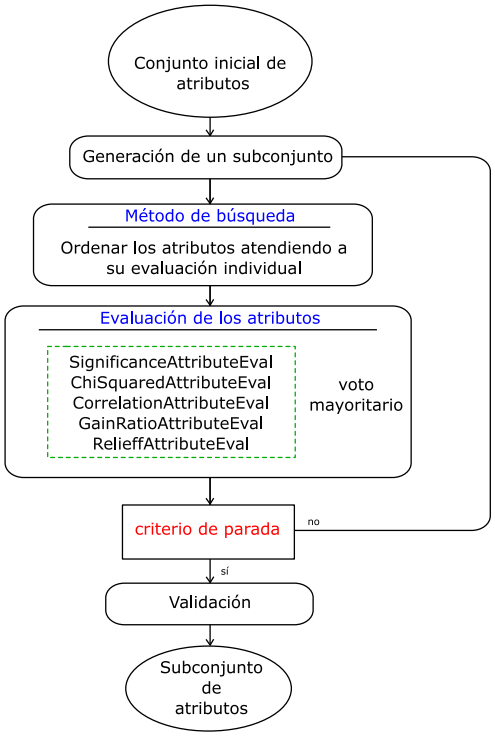


Figura 3.4: Proceso de selección de atributos utilizado.

Una vez que los subconjuntos de atributos candidatos son generados, el algoritmo de evaluación determina el mejor de estos conjuntos atendiendo a un criterio de selección determinado. En este caso, se utiliza un voto mayoritario entre los resultados de varios métodos de selección diferentes. La Tabla 3.1 describe los 5 algoritmos de evaluación empleados durante el proceso de ordenamiento que realiza WEKA. El criterio de parada es un elemento necesario para detener la búsqueda durante la selección de atributos. Eso se hace para evitar una búsqueda exhaustiva y disminuir la complejidad computacional en tiempo de ejecución de los algoritmos (Witten *et˜al.*, 2011).

El conjunto final de datos reducido contiene 60 atributos. La Tabla 3.2 muestra los primeros 15 atributos seleccionados ordenados de acuerdo a tu importancia. Como se puede observar, este método sugiere la ventaja de jugar en casa y los estadísticos Log5 y PE como los más importantes.

Evaluadores	Descripción
SignificanceAttributeEval	Evalúa la importancia del atributo mediante el cálculo de su significación probabilística por medio de una función en dos direcciones (Ahmad y Dey, 2005).
ChiSquaredAttributeEval	Evalúa la importancia del atributo calculando el valor del estadístico chi-cuadrado respecto a la clase.
CorrelationAttributeEval	Evalúa la importancia del atributo midiendo la correlación de Pearson entre dicho atributo y la clase.
GainRatioAttributeEval	Evalúa la importancia del atributo midiendo su Gain-Ratio con respecto a la clase.
ReliefAttributeEval	Evalúa la importancia del atributo tomando muestras repetidamente de una instancia y considerando el valor de ese atributo para la instancia más cercana a la clase (Robnik-Šikonja y Kononenko, 1997).

Tabla 3.1: Descripción de los métodos utilizados para la evaluación de atributos.

Orden	Atributo	Tipo	Descripción
1	isHomeClub	Nominal	Si el equipo visitante es el anfitrion o no
2	Log5	Numérico	Log5
3	PE	Numérico	Expectativa Pitagórica
4	WP	Numérico	Porciento de juegos ganados en la temporada
5	RC	Numérico	Carreras Creadas
6	HomeWonPrev	Nominal	Si el equipo anfitrión ganó el juego anterior o no
7	VisitorWonPrev	Nominal	Si el equipo visitante ganó el juego anterior o no
8	BABIP	Numérico	BABIP
9	FP	Numérico	Porcento de fildeo
10	PitchERA	Numérico	Promedio de Carreras Limpias de los lanzadores abridores
11	OBP	Numérico	OBP
12	Slugging	Numérico	Slugging
13	HomeVersusVisitor	Nominal	Resultados particulares entre el equipo visitante y el anfitrión
14	Stolen	Numérico	Total de bases robadas
15	VisitorLeague	Nominal	Liga del equipo visitante

Tabla 3.2: Lista ordenada de los primeros 15 atributos seleccionados.

3.2.3. Parámetros de los métodos

La Tabla 3.3 muestra los parámetros principales de los algoritmos de aprendizaje utilizados en los experimentos. En este caso, fueron empleados los parámetros por defecto WEKA para cada uno de los algoritmos de predicción seleccionados.

Método	Algoritmo	Parámetros principales
Aprendizaje basado en casos	k-NN	$k = 1$ Distancia Euclideana
Árboles de decisión	REPTree	proporción máxima de varianza= 0,001 peso mínimos de la instancia en las hojas = 2 profundidad máxima = ∞
Máquinas de soporte vectorial	SMO	kernel polinomial con $d = 1$ tolerancia = 0,001 épsilon = $1{,}0E - 12$
Redes neuronales artificiales	MLP	capas ocultas = $(atributos + clases)/2$ *learning rate* = 0,3 *momentum*= 0,2

Tabla 3.3: Parámetros principales de los algoritmos empleados.

3.2.4. Medida de evaluación

La decisión de qué medida emplear para evaluar la calidad de la clasificación es un asunto importante ya que determinará la validez así como la comparabilidad de los resultados. Esta no es una decisión trivial dada la gran cantidad de opciones y de factores a tener en cuenta. Cada medida de permite evaluar un aspecto diferente de la clasificación, es por ello que en distintos campos de aplicación se prefieren medidas de evaluación específicas.

En un problema de clasificación de dos clases existen cuatro posibles salidas, las cuales estan representadas en la matriz de confusión que se muestra en la Tabla 3.4. Esta puede considerarse como la información más básica sobre el desempeño de la clasificación, a partir de la cual se definen medidas más avanzadas.

Dado que en este caso el problema representa clases perfectamente balanceadas se seleccionó la medida de exactitud general o *accuracy*. Dicha medida se define como la efectividad predictiva general del algoritmo y mide el número de aciertos respecto al número total de

	Clasificado como Positivo	**Clasificado como Negativo**
Realmente Positivo (AP)	Verdaderos Positivos (TP)	Falsos Negativos (FN)
Realmente Negativo (AN)	Falsos Positivos (FP)	Verdaderos Negativos (TN)

Tabla 3.4: Matriz de confusión resultante de un problema de clasificación de dos clases.

clasificaciones, estimando la probabilidad de clasificar correctamente una instancia dada. La Ecuación 3.1 muestra cómo se realiza el cálculo de *accuracy*.

$$\text{Accuracy} = \frac{TP + TN}{TP + TN + FP + FN} \tag{3.1}$$

3.2.5. Esquema de validación

La técnica tradicional para evaluar y comparar la capacidad predictiva de los métodos de aprendizaje automático, conocida como *holdout*, consiste dividir los datos en dos subconjuntos, uno de entrenamiento y otro de prueba o control (Han *et˜al.*, 2011). A menudo dos tercios de las instancias son usados para construir el modelo y el resto es usado para probar su aprendizaje. Sin embargo, esta simple división de los datos es a menudo susceptible a errores, dado que ambos conjuntos son excluyentes entre sí, sin importar qué tipo de particionamiento de los datos sea empleado.

Por tal motivo, y con el objetivo de probar de una mejor manera la capacidad predictiva de los algoritmos, se empleó la técnica de validación cruzada en 10 particiones. Como muestra la Figura 3.1, esta técnica consiste en dividir los datos en 10 subconjuntos de forma aleatoria, de modo que el aprendizaje se realice con una parte de ellos, mientras el resto se utiliza para validar los resultados. A partir de cada conjunto de datos se crearon 10 particiones mutuamente excluyentes de forma aleatorea, manteniendo la misma proporción de instancias por clases entre las capas.

Esto es, en la iteración i, la partición P_i se reserva como conjunto de prueba y el resto es usada para el entrenamiento, iterándose en cada una de las 10 particiones. Como se muestra en la Ecuación 3.2, el resultado final es el valor promediado de los resultados obtenidos con la medida de evaluación utilizada.

$$\text{CV} = \frac{\sum_{i=1}^{10} Pi}{10} \tag{3.2}$$

Dado que durante la validación cruzada los resultados de dicha medida de evaluación depende de la asignación de instancias a las diferentes capas, a menudo se emplea una técnica utilizada para asegurar el equilibrio entre las clases conocida como «estratificación». Resultados de estudios empíricos han demostrado que la técnica de validación cruzada estratificada en 10 particiones es uno de los mejores métodos para estimar la efectividad de un modelo de predicción (Zeng y Martinez, 2000).

3.3. Resultados experimentales

En esta sección se detallan los resultados obtenidos en los experimentos. Para llevar a cabo todo el proceso de ejecución y comparación de los resultados de los algoritmos se utilizó el Ambiente Experimental de WEKA o *WEKA Experiment Environment Interface* (Hall *et˜al.*). La Tabla 3.5 muestra los resultados obtenidos con los cuatro métodos de aprendizaje para cada equipo de la MLB atendiendo a los dos esquemas de predicción empleados: clasificación y regresión.

De los cuatro métodos del aprendizaje automático, las SVMs produce los mejores resultados tanto para la predicción basada en clasificación como en regresión, con un valor medio de *accuracy* de 59 % y 58 % respectivamente y una desviación estándar de 1.64 y 1.82. El segundo mejor método resultó ser ANN con valor de *accuracy* para clasificación cercano al 58 %. La Figura 3.5 muestra de forma gráfica estos resultados.

Equipo	1-NN		MLP		REPTree		SMO	
	Clasificación	Regresión	Clasificación	Regresión	Clasificación	Regresión	Clasificación	Regresión
ANA	56.16	56.55	**58.07**	55.62	57.18	55.18	56.57	57.16
ARI	55.34	55.49	56.22	53.63	58.67	56.67	**59.12**	58.53
ATL	56.22	56.05	57.19	53.01	56.96	**58.29**	57.43	56.17
BAL	55.55	54.53	56.33	56.58	57.79	58.51	**59.25**	54.21
BOS	56.77	57.04	58.82	58.72	57.58	58.6	**59.58**	56.91
CHA	57.12	57.07	56.76	56.64	56.81	54.78	57.88	**58**
CHN	54.97	54.68	56.1	54.25	56.65	55.12	**57.04**	56.42
CIN	56.78	56.83	58.41	58.63	59.72	59.68	**60.13**	59.61
CLE	56.3	55.86	58.37	56.92	58.07	59.03	**60.75**	58.96
COL	58.92	58.38	61.41	55.9	58.76	**62.29**	61.5	59.99
DET	55.64	56.52	57.35	57.82	58.59	58.32	**58.64**	57.13
FLO	52.35	52.11	55.78	53.16	57.01	52.85	**56.41**	56.14
HOU	57.59	59.47	60.59	56.86	59.92	60.27	61.06	**61.82**
KCA	53.48	54.06	59.35	59.34	58.75	55.74	60.08	**60.27**
LAN	55.53	56.27	**57.46**	56.64	55.75	55.34	55.65	54.03
MIL	52.81	52.76	**58.82**	57.48	58.48	58.54	58.65	56.85
MIN	54.96	55.06	58.05	56.11	**60.78**	58.78	60.43	58.65
NYA	55.63	55.74	**60.35**	57.85	59.63	60.27	60.26	58.72
NYN	56.38	56.24	56.57	58.04	58.57	56.73	**59.65**	58.22
OAK	54.07	53.04	57.61	56.02	56.37	58.57	**58.88**	58.31
PHI	55.48	55.18	56.83	55.43	57.23	57.54	**59.18**	58.28
PIT	57.68	57.01	58.74	58.82	60.29	60.56	**62.27**	59.99
SDN	56.09	56.64	55.57	55.4	57.68	57.57	**59.21**	57.13
SEA	58.13	58.41	**58.99**	54.62	57.04	57.17	57.87	54.77
SFN	54.08	54.28	56.49	53.78	55.94	54.78	56.47	**57.72**
SLN	54.37	53.53	58.24	57.33	59.1	59.25	59.03	**59.5**
TBA	55.2	55.34	58	56.33	55.95	**60.05**	57.45	56.38
TEX	57.25	56.39	58.28	54.96	57.04	**58.94**	58.36	55.52
TOR	57.43	56.73	58.3	53.94	56.3	**58.91**	57.83	57.1
WAS	61.26	**62.08**	57.72	55.49	57.24	58.73	61.16	57.35
Media	55.98	55.97	57.89	56.17	57.86	57.9	**58.92**	57.66

Tabla 3.5: Valores de *accuracy* obtenidos para todos los algoritmos y esquemas de predicción, los valores máximos estan señalados en negrita.

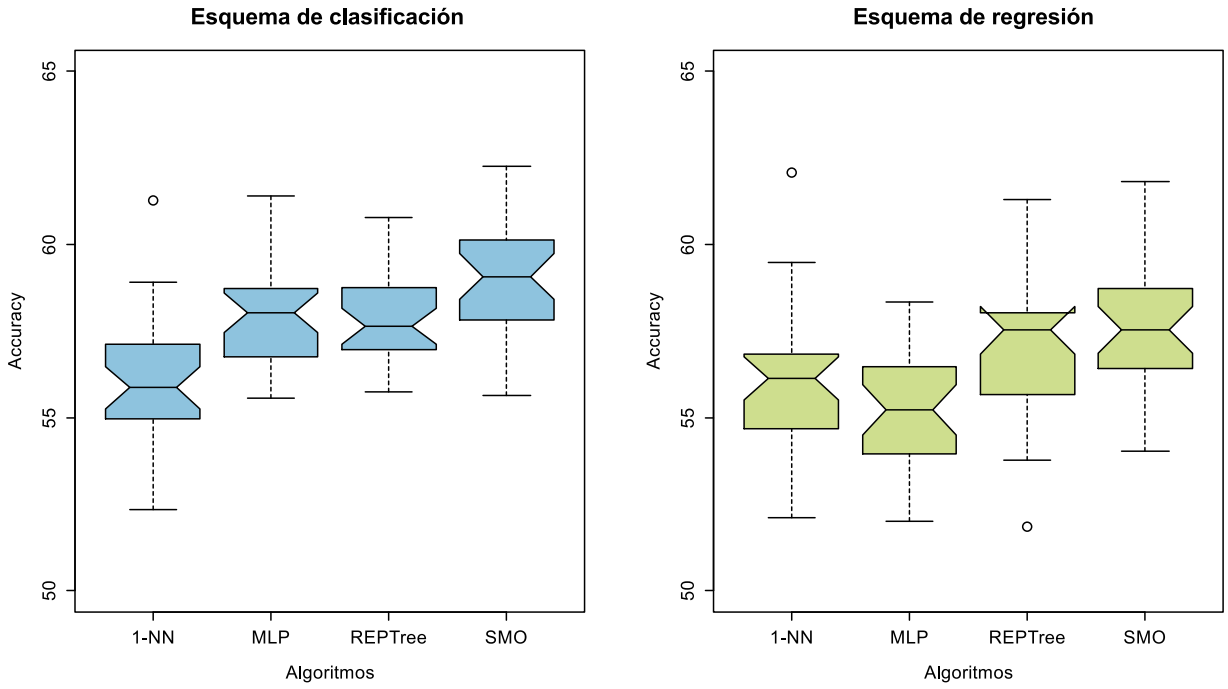


Figura 3.5: Comparación gráfica de los resultados obtenidos con ambos esquemas de predicción.

3.3.1. Determinación del mejor algoritmo de predicción

A partir de los resultados obtenidos, y con el propósito de seleccionar el mejor algoritmo de predicción, se llevaron a cabo varias pruebas estadísticas. Para ello se realizó la prueba de signos alineados de Friedman, la cual detecta diferencias significativas para un $\alpha = 0{,}05$ (para el esquema de clasificación el ρ-valor= $3{,}06 \cdot 10^{-5}$ y para el de regresión es de $2{,}18 \cdot 10^{-5}$).

La Tabla 3.6 muestra los resultados obtenidos, como se observa, el mejor algoritmo es seleccionado como método de control para la aplicación del procedimiento *post-hoc* de Hochberg. En ambos casos, algoritmo seleccionado es SMO, mientras que 1NN muestra los valores más bajos de *accuracy*. El procedimiento *post-hoc* de Hochberg detecta diferencias significativas en favor del algoritmo SMO respecto a 1NN, MLP y REPTree, con la excepción del algoritmo de regresión REPTree.

Esquema	Algoritmos	Rangos promedio	$\rho_{Horcherber}$	Hipótesis
Clasificación	1-NN	97.4	0	Denegada
	MLP	57.13	0.001638	Denegada
	REPTree	58.62	0.001638	Denegada
	SMO	28.85	-	-
Regresión	1-NN	70.63	0	Denegada
	MLP	86.4	0.000134	Denegada
	REPTree	50.13	0.88473	Aceptada
	SMO	34.83	-	-

Tabla 3.6: Prueba de rangos alineada de Friedman con los resultados de ambos esquemas predictivos, p-valores ajustados con el procedimiento *post-hoc* de Hochberg.

Con el objetivo de determinar el mejor esquema de predicción se aplicó la prueba de signos de Wilcoxon a los valores de *accuracy* obtenidos con el algoritmo SMO (el mejor algoritmo). La Tabla 3.7 muestra que el esquema de clasificación es mejor que el de regresión. La diferencia obtenida resultó significativa para un $\alpha = 0{,}05$.

Esquemas SMO	Rank. negativo	Rank. Positivo	p-valor
Clasificación – Regresión	6	24	$7{,}2 \cdot 10^{-5}$

Tabla 3.7: Prueba alineada de signos de Wilcoxon para los valores de *accuracy* los esquemas predictivos del algoritmo SMO.

3.3.2. Comparación con el mercado de apuestas

Una de las principales razones que ha motivado el desarrollo de modelos de predicción en el contexto deportivo es el notable auge de los mercados de apuestas deportivas. Por tal motivo, se decidió incluir en este estudio una comparación con uno de los mayores mercados de apuestas del béisbol de la MLB denominado Covers[1].

Se colectaron los registros en Covers de todas las apuestas de tipo *money-line* correspondientes a la temporada regular de 2014 usando un *script* implementado en el lenguaje de programación Phyton. A continuación se calcularon los valores de *accuracy* de las predicciones realizadas por Covers para esa temporada. El algoritmo de clasificación SMO fue entrenado con los datos de 9 años, desde 2005 hasta 2013, y probado con los resultados de los partidos de la temporada 2014 para su comparación con los resultados *money-line*.

La Tabla 3.8 muestra los valores de *accuracy* obtenidos. La prueba de signos de Wilcoxon determinó que no existen diferencias significativas entre las predicciones *money-line* de Covers y las realizadas por el algoritmo SMO para un $\alpha = 0{,}05$ ($\rho-$valor $= 0{,}066$). De acuerdo con este resultado, el modelo de predicción SMO puede ser considerado competitivo en comparación con el mercado de apuestas.

3.4. Sumario

En este capítulo se realizó un estudio comparativo del desempeño de cuatro métodos del aprendizaje automático, los cuales fueron empleados en la predicción de resultados de juegos de béisbol. El modelo de predicción propuesto llevó a cabo todo el proceso de minería de datos siguiendo la metodología CRISP-DM. Los datos fueron obtenidos a partir del cálculo acumulativo de estadísticos de la sabermetría utilizando los *game logs* de Retrosheet y la base de datos Lahman. Los conjuntos de datos creados corresponden a 10 temporadas regulares de la MLB.

Después de analizar los resultados de la experimentación, podemos extraer las siguientes conclusiones:

- Dada la gran cantidad de elementos que inciden en el marcador final, la predicción de resultados de juegos de béisbol es un problema complejo.

[1] http://covers.com

Equipo	Accuracy (%)	
	Money-line	SMO
Baltimore	50.89	**54.49**
Boston	50.00	**54.00**
NY Yankees	55.56	**59.34**
Tampa Bay	51.23	**52.08**
Toronto	**50.62**	50.22
Atlanta	**52.47**	51.52
Miami	**53.09**	47.00
NY Mets	59.26	**60.00**
Philadelphia	51.85	**58.46**
Washington	**62.65**	55.69
Chi. White Sox	**56.79**	52.06
Cleveland	56.17	**56.88**
Detroit	**54.55**	53.04
Kansas City	**58.19**	47.40
Minnesota	**59.88**	55.38
Chi. Cubs	**56.79**	53.55
Cincinnati	57.41	**64.00**
Milwaukee	52.47	**53.49**
Pittsburgh	**61.35**	56.00
St. Louis	**57.31**	56.43
Houston	54.94	**57.47**
LA Angels	**60.00**	56.43
Oakland	**63.19**	49.28
Seattle	**53.70**	47.00
Texas	**59.26**	57.88
Arizona	**58.64**	55.49
Colorado	**64.20**	58.46
LA Dodgers	**60.24**	50.67
San Diego	54.94	**60.00**
San Francisco	**60.89**	57.38
Media	**56.62**	54.70

Tabla 3.8: Comparación entre los valores predictivos de *accuracy* obtenidos con el algoritmo de clasificación SMO y el mercado de apuestas *money-line* de Covers.

- Existen fuentes de datos históricos, disponibles públicamente, las cuales posibilitan el análisis detallado de todo lo acontecido en juegos de la MLB.
- Las técnicas de selección de atributos son una buena opción para reducir la alta dimensionalidad de los datos disponibles en este deporte.

- Las SVM constituye un algoritmo del aprendizaje automático con un buen desempeño para la predicción de resultados de juegos de béisbol.
- El esquema predictivo de clasificación es mejor en comparación con el de regresión en el caso particular del béisbol.
- Los resultados obtenidos con los métodos del aprendizaje automático resultan ser competitivos con relación a las estimaciones realizadas en el mercado de apuestas.

Capítulo 4

Modelo basado en clasificación de series de tiempo para la predicción del desempeño de lanzadores abridores de béisbol

En el presente capítulo se propone un modelo para predecir el desempeño de los lanzadores abridores en el béisbol. El modelo construye una serie de tiempo a partir de los datos del lanzador colectados durante del desarrollo del encuentro y luego clasifica la serie teniendo en cuenta sus registros históricos. En la Sección 4.1 se describen las características del problema, así como los principales aspectos a tener en cuenta en la predicción del desempeño de los lanzadores abridores de béisbol. La Sección 4.2 detalla la forma en que se manejaron los datos lanzamiento-a-lanzamiento para efectuar su análisis como datos de tipo series de tiempo. Los detalles sobre el diseño de los experimentos realizados para probar la utilidad del modelo propuesto se presenta en la Sección 4.3. La Sección 4.4 muestra los resultados experimentales obtenidos para los 20 lanzadores abridores de la MLB que fueron objeto de estudio. Por último, la Sección 4.5 presenta un sumario con las principales conclusiones a las que se arriba en este capítulo.

4.1. Sobre el análisis del desempeño de los lanzadores abridores de béisbol

En el juego de béisbol, el picheo es considerado como la habilidad más difícil de aprender y dominar. La mayoría de los expertos coinciden en que este es el componente más de mayor

importancia para obtener la victoria (Pavitt, 2011). Recientemente, se han propuesto algunos modelos para el análisis y predicción del desempeño de los lanzadores. Por ejemplo, Piette *et˜al.* (2010) estudian la factibilidad y consistencia de varios estadísticos para evaluar la efectividad de los lanzadores de la MLB usando un modelo aleatorio bayesiano, Sidhu y Caffo (2014) exploran la toma de decisiones de los lanzadores mediante la modelación de información desde la perspectiva del bateador (selección del lanzamiento y conteo) como un Modelo de Decisión de Markov (MDP), mientras que Hoang *et˜al.* (2015) introduce una novedosa estrategia adaptativa usando Análisis Lineal Discriminante (LDA) para predecir los tipos de lanzamientos de forma binaria (rectas o curvas).

Sin embargo, uno de los problemas más complejos que deben ser enfrentados por los directores de equipos de béisbol durante los juegos consiste en decidir en qué momento el lanzador abridor debe ser sacado del juego y sustituido por un lanzador de relevo. Este es un proceso de toma de decisiones, el cual se hace más difícil aun por el hecho de que el lanzador sustituto necesita alrededor de 10 minutos de «calentamiento» antes de entrar a jugar. Además, dicho relevista no debe «calentar» por un período de tiempo excesivo ya que esto podría agotarlo demasiado.

No existe una fórmula definitiva para decidir de forma correcta el momento en el que el lanzador abridor debe ser relevado. Por tal motivo, los directores de equipo usualmente utilizan ciertas heurísticas (Zimniuch, 2010), como por ejemplo: el conteo del lanzador, el resultado del juego, su propia experiencia e intuición etc.

La sabermetría ha demostrado las ventajas de utilizar estadísticos históricos para el análisis del juego y la toma de decisiones. En este sentido, los registros de resultados de los lanzamientos son una gran fuente de datos, la cual podría revelar importantes patrones de comportamiento de los lanzadores. En consecuencia, a continuación se presentan los detalles de un modelo para la predicción del desempeño de los lanzadores el cual toma ventaja de estos registros históricos. El modelo propone transformar y analizar los datos lanzamiento-a-lanzamiento como una serie de tiempo.

4.2. Manejo de los datos lanzamiento-a-lanzamiento como series de tiempo

Partiendo del trabajo desarrollado por Sidran (2005), en este trabajo se propone manejar las secuencias de lanzamientos como series de tiempo. Con este objetivo, se introduce una

nueva métrica para evaluar el desempeño del lanzador durante el encuentro. Dicha métrica, denominada RP (*Running Performance*), extrae la información sobre el resultado de cada uno de los lanzamientos previos realizados por el picher.

La métrica RP consiste en una puntuación acumulativa de cada lanzamiento realizado. Su salida consiste en un valor numérico, en un formato que permite realizar una representación gráfica del desempeño del lanzador en cualquier momento del juego.

Denotamos como U al conjunto de valores correspondiente a cada posible resultado p_i de un lanzamiento. Los resultados positivos y negativos son punteados de acuerdo a su impacto en el resultado del juego. Dicho de otro modo, las puntuaciones se definen de un modo que hagan posible un balance entre todas los posibles resultados de cada aparición al plato del bateador. Por ejemplo, el valor absoluto de un out es el mismo que el de un sencillo puesto que se considera que ambos eventos tienen la misma relevancia en el juego (sus probabilidades de ocurrencia están equilibradas).

Lanzamiento con resultado Positivo	Puntuación	Lanzamiento con resultado Negativo	Puntuación
Strike	+1/3	Bola	-1/4
Foul	+1/3	Bola intencional	-1/4
Roletazo de out	+1	Base robada	-1
Ponche	+2	Pelotazo	-2
Out	+1	Sencillo	-1
Elevado de out	+1	Base por bolas	-1
Out dentro del terreno	+1	Doble	-2
Línea de out	+1	Triple	-3
Out en la inicial	+1	Cuadrangular	-4

Tabla 4.1: Definición de los valores positivos y negativos de U basados en cada uno de los posibles resultados de los lanzamientos.

Las puntuaciones RP pueden ser vistas intuitivamente como datos de serie de tiempo. La idea es la siguiente, al inicio del juego (instante $i = 0$) el lanzador inicia con una puntuación de desempeño $p_0 = 0$. Más adelante, la puntuación asociada p_i en U correspondiente a cada lanzamiento i se adiciona al valor acumulado de $\mathrm{RP}(i-1)$. La Ecuación 4.1 resume este procedimiento:

$$\mathrm{RP}(i) = \sum_{n=1}^{i} p_i \tag{4.1}$$

A partir de cada lanzamiento del picher, su puntuación total es actualizada y almacenada con el propósito de construir una serie de tiempo que represente su desempeño durante el juego. Dicho desempeño es etiquetado, en el momento i, de acuerdo a la siguiente función:

$$\text{Desempeño}(i) \quad = \quad \begin{cases} \textit{Positivo} & \text{si RP}(i) \geqslant 0, \\ \textit{Negativo} & \text{en otro caso.} \end{cases}$$

A modo de ejemplo, la Figura 4.1 representa dos series de tiempo del del desempeño de Justin Verlander, lanzador estrella del equipo de los Tigres de Detroit, en dos juegos celebrados durante la temporada regular de 2009. Como se puede observar, en (a) su puntuación RP se incrementa de manera estable hasta su lanzamiento número 60 y luego decrece algunos puntos, para culminar finalmente con una puntuación positiva de valor 7. Por otra parte, en el caso del juego (b) su desempeño se considera bueno hasta su lanzamiento número 57, entonces Verlander comienza a tener imprecisiones y a descontrolarse hasta ser reemplazado por un lanzador sustituto en la sexta entrada. En ese momento la puntuación RP fue de $-4{,}2$ puntos.

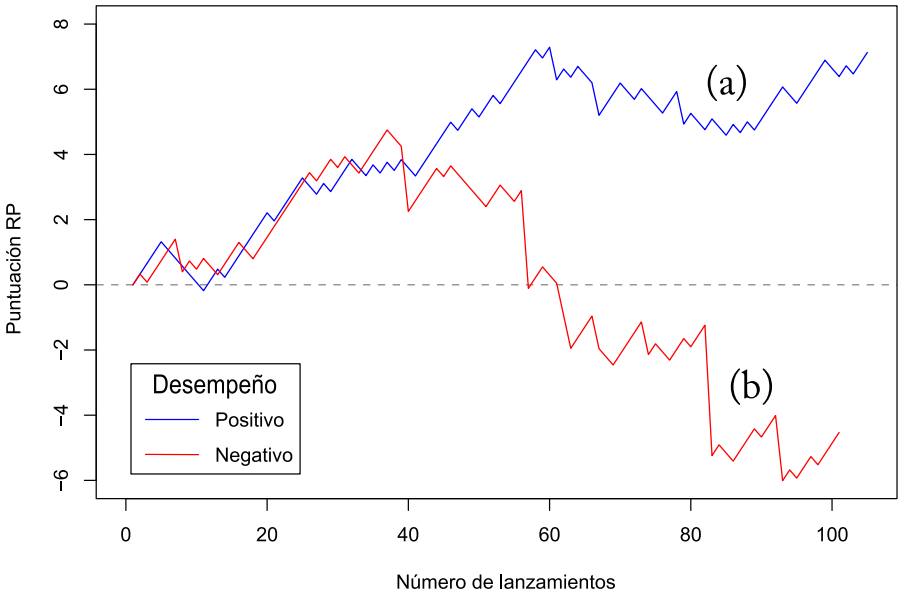


Figura 4.1: Representación gráfica del desempeño de un lanzador abridor visto como una serie de tiempo.

4.2.1. Clasificación de series de tiempo de lanzamientos

Geurts (2001) señala los pasos necesarios para llevar a cabo el planteamiento de la solución a un problema de clasificación de series de tiempo. El primer paso es esencial y consiste en encontrar propiedades locales o patrones en las series. En el segundo paso dichos patrones se combinan para construir reglas de clasificación usando métodos de minería de datos.

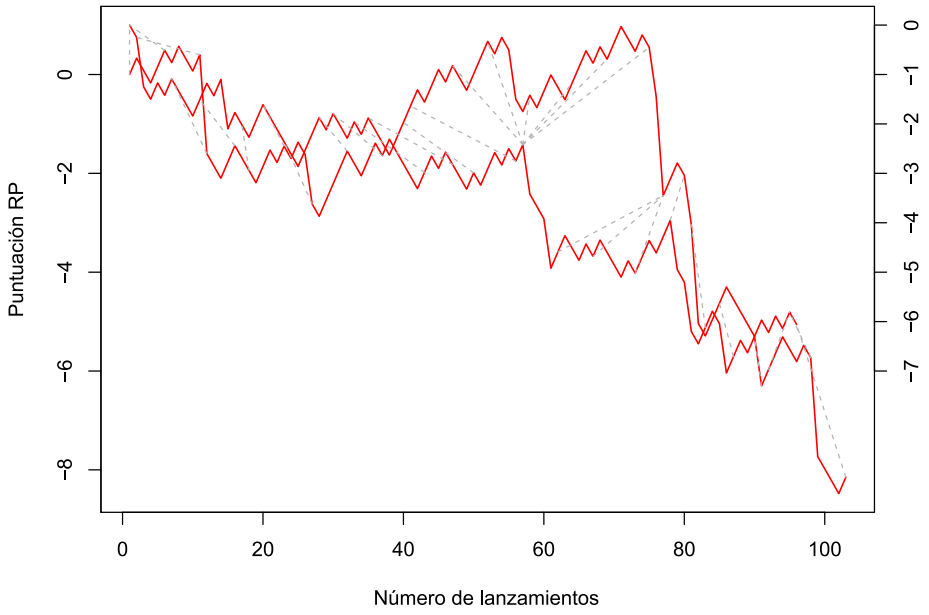
Como se planteó en el Capítulo 1, entre los diversos métodos que pueden ser utilizados, el grupo de clasificadores basados en vecinos más cercanos son de los que mejores resultados han reportado en la literatura. Por esta razón, el modelo de predicción propuesto implementa el clasificador 1NN como algoritmo de aprendizaje. La idea es muy simple, y consiste en comparar las series de tiempo de lanzamientos entre sí con el objetivo de encontrar aquella que se la más similar a la serie objetivo.

A modo de ejemplo, dada una serie de lanzamientos P y un conjunto de entrenamiento de series de lanzamientos $S = (s_1, \ldots, s_n)$, el algoritmo busca la serie en S que es más similar a P siguiendo un criterio de comparación determinado. La serie encontrada es considerada como el «vecino más cercano» de P en S.

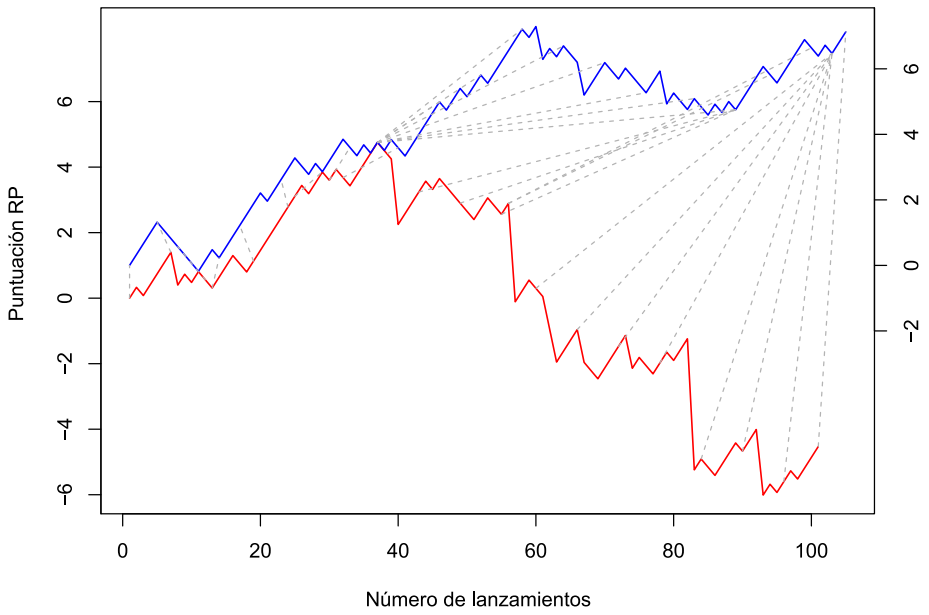
4.2.2. Comparación de series de tiempo de lanzamientos

En el Capítulo 1 se planteó cómo para los problemas de clasificación el algoritmo 1NN usando DTW como función de similaridad ha probado ser superior respecto a otras métricas de distancia Wang *et˜al.* (2013). Esto se debe a que DTW se adapta a los cambios y desfasajes de las series que compara. El modelo de predicción de desempeño propuesto implementa DTW como criterio de comparación para medir la similitud entre las series de tiempo de lanzamientos.

La Figura 4.2 muestra el alineamiento de dos de estas series usando DTW. Como se ilustra, DTW compara eficientemente ambas series incluso cuando tienen diferente longitud o estas desfasadas. En este caso, el costo del alineamiento en (a) es 63.64, mientras que en (b) es 573.95.



(a)



(b)

Figura 4.2: Comparación entre dos series de tiempo de lanzamientos usando la medida de similitud DTW.

Aunque DTW sobrepasa a muchas otras medidas de distancia, es sabe que ésta tiene un alto costo computacional con una complejidad temporal de orden $O(n^2)$. Dada esta situación, se cota inferior propuesta por Keogh y Ratanamahatana (2005) es una buena opción para reducir dicha complejidad y mejorar el rendimiento de la clasificación. La cota inferior de Keogh, $LB\_Keogh(P, P')$, entre dos series de tiempo P y $P' = (p'_1, \ldots, p'_i, \ldots, p'_n)$ se calcula segun la siguiente función:

$$LB\_Keogh(P, P') = \sum_{i=1}^{n} \begin{cases} |p'_i - u_i| & \text{si } p'_i > u_i, \\ |l_i - p'_i| & \text{si } p'_i < l_i, \\ 0 & \text{en otro caso.} \end{cases}$$

donde $u_i = \text{máx}\{p_{i-r_i}, \ldots, p_{i+r_i}\}$ y $l_i = \text{mín}\{p_{i-r_i}, \ldots, p_{i+r_i}\}$ son elementos de envoltorio calculados a partir de una constante global$R = (r_1, \ldots, r_i, \ldots, r_n)$. Ratanamahatana y Keogh (2005) mostraron cómo con el uso de LB_Keogh se hace posible podar alrededor del 90 % de todos los cálculos de DTW en varios conjuntos de datos experimentales.

4.2.3. Algoritmo de predicción

De acuerdo a los datos obtenidos lanzamiento-a-lazamiento, la longitud de las series de tiempo construidas a partir de estos datos se incrementa de una forma dinámica con cada nuevo lanzamiento. Por tanto, para predecir se hace necesario determinar el desempeño futuro del lanzador mientras el encuentro esta en progreso. Está claro que mientras mayor sea la longitud de la serie de lanzamientos a analizar entonces mayor será la información disponible acerca del comportamiento del lanzador, y por lo tanto mejor deberá ser el desempeño del algoritmo de predicción implementado.

En nuestro caso, la predicción está basada en los resultados obtenidos por el lanzadador bajo esas mismas condiciones en situaciones similares en el pasado. El modelo de predicción propuesto implementa el algoritmo de aprendizaje 1NN, el cual usa DTW como medida de similitud en conjunción con la cota inferior de Keogh, para encontrar la secuencia de lanzamientos más similar a la actual. El Algoritmo 4.1 muestra los detalles del funcionamiento de dicho modelo.

La cota inferior entre la serie de entrada P y la serie candidata de p de S es calculada en la línea 7. Si la el valor de dicha cota es suficientemente grande, entonces se evita el cálculo de DTW tal como muestra la línea 8. La función dtw, en la lína 9, calcula el valor del

Algoritmo 4.1: Predice el desempeño del lanzador usando 1NN y DTW como medida de similitud. Usa la cosa inferios de Keogh para disminuir la complejidad computacional.

```
1  Función PredecirDesempeño(P, S)
       Entrada: La serie de tiempo P del lanzador.
       Salida: Conjunto de entrenamiento S de series de tiempo de lanzamientos.
2      minDist ← ∞
3      minLB ← ∞
4      class ← ∅
5      mientras (S ≠ ∅) hacer
6          p ← una serie en S
7          distLB ← lbKeogh(P, p)
8          si (distLB < minLB) entonces
9              dist ← dtw(P, p)
10             si (dist < minDist) entonces
11                 class ← getClass(p)
12                 minDist ← dist
13             fin
14             bestLB ← minLB
15         fin
16         S ← S \ p
17     fin
18     retornar class
19 fin
```

alineamiento óptimo entre P y p. Si dicho valor es menor que $minDist$ entonces la serie de tiempo P se etiqueta como Positiva o Negativa, de acuerdo a la etiqueta correspondiente a la serie p más similar a P en S.

4.3. Diseño de los experimentos

A continuación se presentan la metodología utilizada para validar el modelo de predicción propuesto. Los datos empleados en el análisis proceden de resultados de juegos reales de la MLB. Debido a que es imposible conocer por adelantado la cantidad de lanzamientos que un lanzador realizará antes de ser sustituido, se hace necesario evaluar la capacidad predictiva del modelo dada una cantidad diferente de lanzamientos. El marco experimental presentado evalúa el modelo en diferentes momentos del juego.

4.3.1. Caracterización del conjunto de datos utilizado

Desde la temporada 2007, el sistema PITCHf/x de Sportvision ha grabado, en tiempo real, detalles acerca de cada lanzamiento durante cada juego de la MLB (Fast, 2010). El sistema PITCHf/x incluye además detalles sobre el resultado de las apariciones al plato resultado de cada lanzamiento. Una gran ventaja es que estos datos estan disponibles gratuitamente a través del sitio web GameDay[1] de la MLB. La información recopilada por este sistema ha resultado ser especialmente valiosa para entrenadores, analistas deportivos y fanáticos del béisbol.

Albert (2010) elaboró una colección de datos PITCHf/x de 20 lanzadores los cuales participaron en la temporada 2009 de la MLB. Los datos fueron obtenidos a partir de cada juego de la temporada regular en la dichos lanzadores participaron como abridores. La Tabla 4.2 muestra información acerca del número de partidos jugados y la clasificación de su desempeño en el momento en que abandonaron el juego, de acuerdo al criterio de evaluación del desempeño propuesto.

El promedio de lanzamientos por juego fue de 101. Nuevo de los lanzadores, marcados con una *e*, son considerados élite debido a que éstos recibieron (o fueron nominados) al prestigioso premio al mejor lanzador de la MLB denominado Cy Young.

En resumen, de un total de 649 juegos, cerca de un cuarto fueron clasificados con un desempeño Positivo y el resto se clasificó como Negativo. Una explicación a esto es que los directores de equipo por lo general esperan demasiado para reemplazar a sus lanzadores abridores. Esta es una decisión que a menudo es tomada demasiado tarde, cuando ya las consecuencias negativas de la tardanza son imposibles de revertir.

4.3.2. Marco experimental

A continuación se presenta al marco experimental mediante el cual será validado el modelo predictivo propuesto. Con el objetivo de obtener resultados lo más cercanos posible a la realidad se han concebido dos esquemas de validación. El primero consiste en evaluar el resultado de la predicción para cada lanzador de forma individual mediante un esquema *holdout*, mientras que el segundo ofrece una estimación más general sobre la capacidad de predicción del modelo a través de un esquema de validación cruzada para todo el cujunto de datos.

[1] http://gd2.mlb.com/components/game/mlb/

Nombre del lanzador	Juegos clasificados como Positivos	Juegos clasificados como Negativos	Juegos totales
Zack Greinke[e]	11	22	33
Roy Halladay[e]	14	18	32
Danny Haren[e]	15	18	33
Felix Hernandez[e]	12	22	34
Cliff Lee[e]	13	21	34
Tim Lincecum[e]	15	17	32
C C Sabathia[e]	10	22	32
Justin Verlander[e]	18	17	35
Adam Wainwright[e]	9	25	34
Roy Oswalt	10	20	30
Brett Anderson	7	23	30
Bronson Arroyo	6	27	33
Scott Baker	13	20	33
Joe Blanton	4	27	31
Scott Feldman	4	30	34
Gavin Floyd	4	26	30
Cole Hamels	7	25	32
Derek Lowe	1	33	34
Ricky Nolasco	11	20	31
Andy Pettitte	4	28	32
Total	188	461	649

[e] Nominados al prestigioso premio Cy Young.

Tabla 4.2: Sumario de la clasificación del desempeño de los 20 lanzadores considerados en este estudio en el momento en que abandonaron el juego.

La Figura 4.3 representa gráficamente la metodología empleada para la evaluación del modelo. Primeramente, los datos lanzamiento-a-lanzamiento son transformados en series de tiempo usando el método RP propuesto en la Sección 4.2. A continuación, el conjunto de datos resultante es particionado en dos conjuntos independientes: una de entrenamiento y otro de prueba. El conjunto de entrenamiento es usado por el modelo en el proceso de aprendizaje, mientras que el conjunto de prueba evaluará la calidad de dicho aprendizaje.

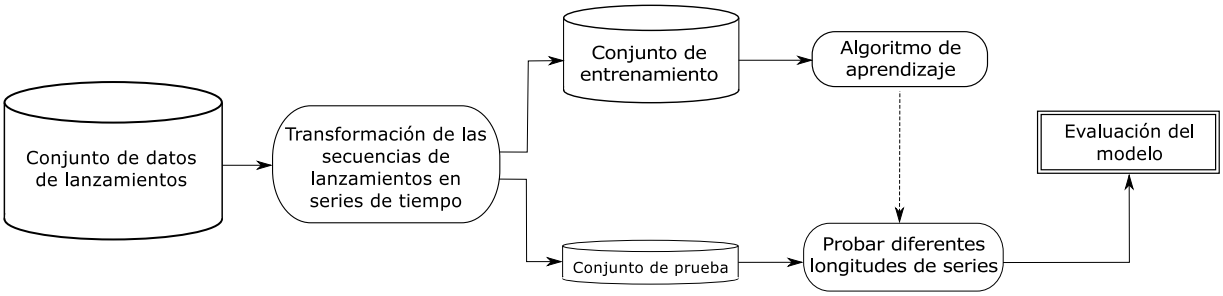


Figura 4.3: Representación gráfica de la metodología empleada para validar el modelo de predicción propuesto.

Una vez definidas las particiones de entrenamiento y de prueba, se procede a reducir la longitud de las series de tiempo de control con el objetivo de evaluar la capacidad predictiva del modelo en diferentes momentos del juego de béisbol. La Figura 4.4 muestra los tres porcientos de longitud de la series utilizados en este estudio. En correspondencia con los propósitos de esta evaluación, el valor de la clase de las series de prueba reducidas permanece igual al de la serie con tu longitud completa.

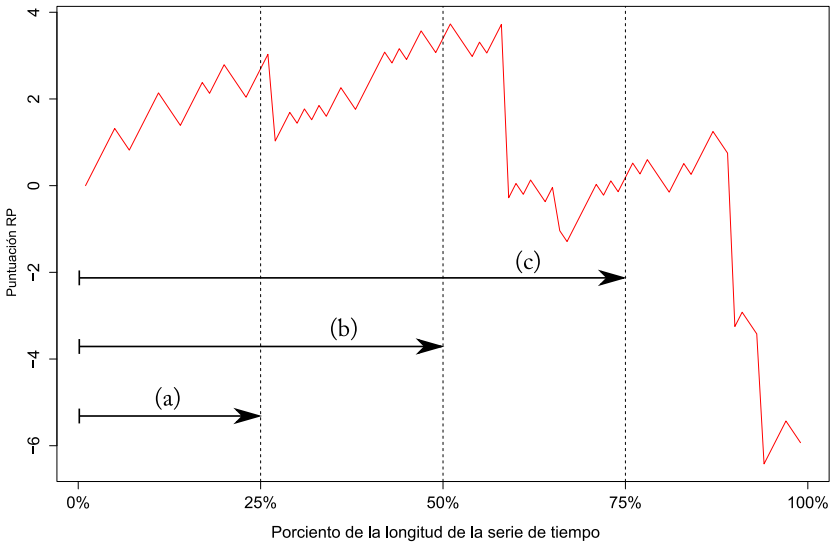


Figura 4.4: Porciento de la longitud de las series usadas para la predicción, (a) el 25 %, (b) el 50 % y (c) el 75 %.

4.3.3. Medidas de evaluación

A continuación seleccionaremos las medidas de evaluación empleadas para validar el modelo de predicción propuesto. Debido a que hay una mayor cantidad de series de lanzamientos etiquetadas como Positivas que Negativas, entonces se hace necesario considerar aquellas medidas de evaluación que permitan manejar problemas de clasificación con desbalance entre la distribución de las clases.

Las medidas seleccionadas en este caso son las denominadas *precision* y *recall*. La primera mide el porciento de las series predichas como Positivas que realmente lo son, mientras que la segunda mide el porciento de las series predichas como Negativas que realmente lo son. Las Ecuaciones 4.2 y 4.3 presentan ambas medidas.

$$Precision = \frac{TP}{TP + FP} \tag{4.2}$$

$$Recall = \frac{TN}{TN + FN} \tag{4.3}$$

Las medidas *precision* y *recall* ofrecen informacion valiosa sobre proporción de series clasificadas correcta e incorrectamente. Sin embargo, se hace necesario tener una sola medida de evaluación que caracterice con un solo valor el rendimiento general del modelo propuesto. La medida conocida como F1 o F-Measure este objetivo se considera un criterio de evaluación más general, dado que combina ambos valores de *precision* y *recall*. La Equación 4.4 muestra la forma en que se calcula dicha medida.

$$F\text{-}Measure = \frac{2 \cdot TP}{2 \cdot TP + FP + FN} \tag{4.4}$$

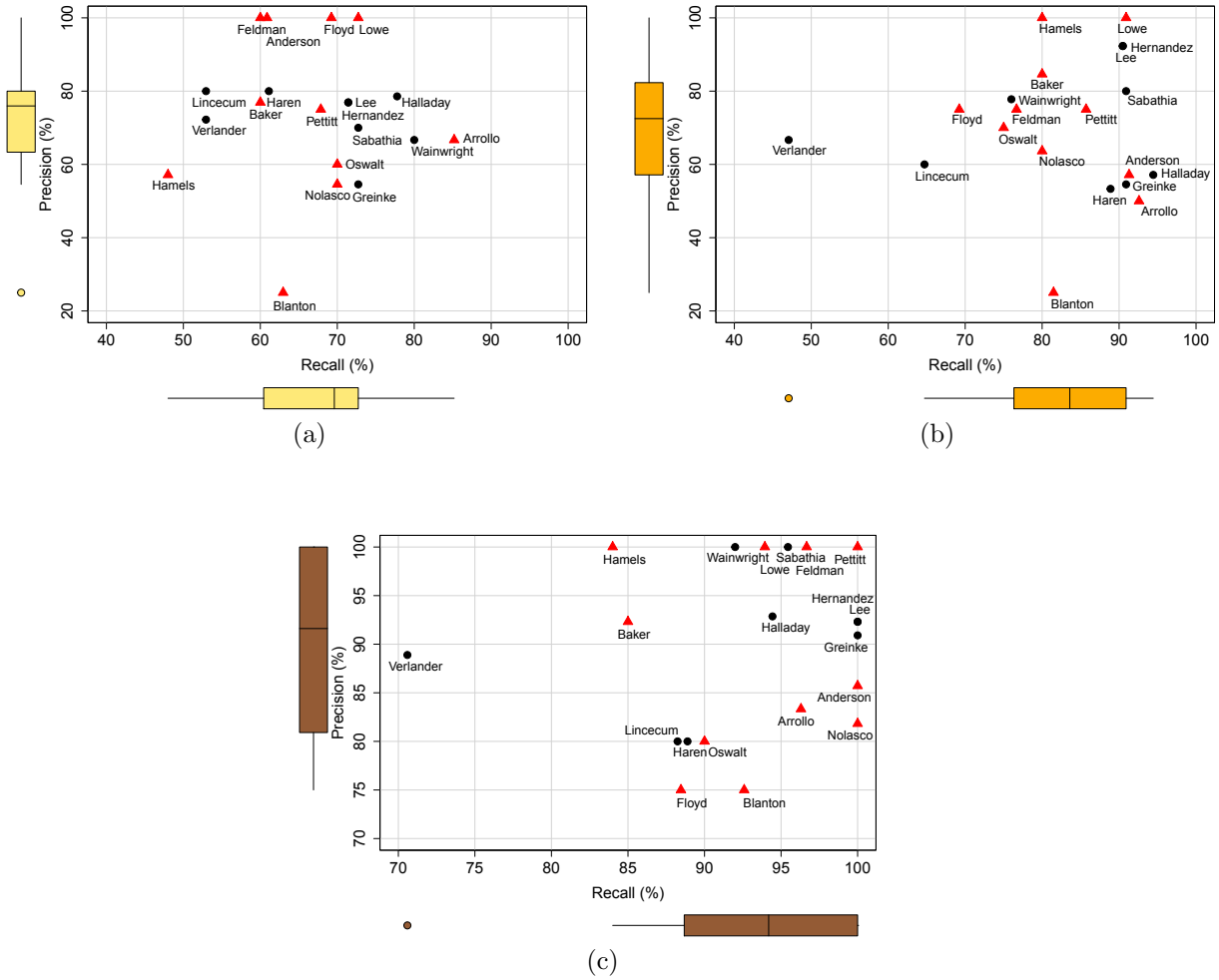
4.4. Resultados experimentales

En esta sección se presentan los resultados obtenidos con el modelo de predicción para el conjunto de datos conformado por los 20 lanzadores de la MLB descritos anteriormente. Los experimentos se conformaron siguiendo el marco experimental indicado en la Sección 4.3. La capacidad de predicción del modelo se evalúa de forma tanto individual para cada lanzador como colectiva en general. Adicionalmente, se realiza una comparación entre

lanzadores promedio y aquellos considerados élite.

4.4.1. Resultados individuales de los lanzadores

Primeramente, los experimentos se han llevado a cabo con el objetivo de evaluar los resultados de *precision* y *recall* de forma individual para cada lanzador. De esta forma, se probó el modelo con las series de lanzamientos de cada picher y se entrenó con las del resto. La Figura 4.5 muestra los resultados obtenidos.



(a) (b) (c)

Figura 4.5: Valores de *precision* y *recall* de los 20 lanzadores abridores incluidos en este estudio usando (a) el 25 % de la longitud de la serie, (b) el 50 % y (c) el 75 %. Los lanzadores élite se indican con círculos.

Los valores de *precision* y *recall* mejoran de forma proporcional a la longitud de la serie de prueba utilizada. Como se muestra, el 75 % de la longitud de la serie ofrece valores para estas medidas los cuales pueden ser considerados como buenos para este problema, con medias de *precision* y *recall* de 89.5 y 92.8 respectivamente.

Para una evaluación más general, el gráfico de barras de la Figure 4.6 compara los valores de F1 para cada lanzador de forma individual. Tal como se muestra, el rendimiento del modelo en general es favorable, con un valor medio de F1 de 73.45, incluso cuando solo se conocen la mitad de los lanzamientos del picher. Las predicciones pueden considerarse como muy buenas, con valores de F1 superiores al 90 %, cuando se conoce el 75 % de la longitud de la serie. Resulta notable el 100 % de F1 obtenido en el caso del lanzador Andy Pettitte.

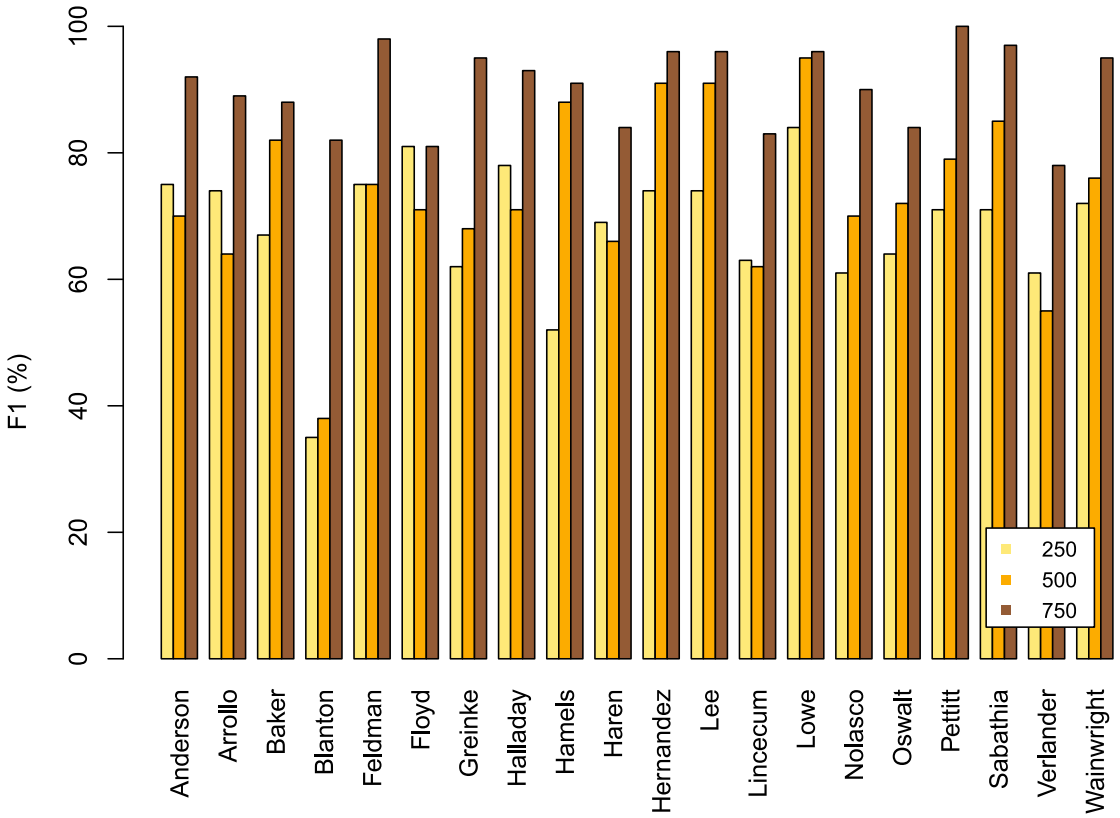


Figura 4.6: Valores de F1 obtenidos para los 20 lanzadores abridores utilizando el 25 %, 50 % y 75 % de la longitud de la serie de lanzamientos.

4.4.2. Resultados generales de los lanzadores

La Tabla 4.3 presenta los resultados obtenidos usando validación cruzada estratificada con 10 particiones en todo el conjunto de lanzadores. Además de las medidas *precision*, *recall* y F1 se incluye la medida *accuracy*. De acuerdo con los resultados obtenidos, el modelo se desempeña favorablemente a medida que aumenta la longitud de las series de prueba. Este resultado se corresponde con lo visto hasta el momento. Otra vez, resultan significativos los valores de *accuracy*, *precision* y *recall* superiores al 90 % cuando el 75 % de los lanzamientos del picher abridor son conocidos.

Longitud de la serie	Matriz de confusión[a]			Accuracy (%)	Precision (%)	Recall (%)	F1 (%)
		Positiva	Negativa				
25 %	Positiva	130	58	70,57	69,15	71,15	57,65
	Negativa	133	328				
50 %	Positiva	118	70	77,97	62,77	84,16	62,27
	Negativa	73	388				
75 %	Positiva	170	18	91,06	90,43	91,32	85,43
	Negativa	40	421				

[a] Las filas representan la clase actual, las columnas representan las predicciones.

Tabla 4.3: Resultados generales de predicción para todo el conjunto de datos utilizando validación cruzada estratificada con 10 particiones.

4.4.3. Comparación entre clases de lanzadores

Como una comparación adicional, se aplicó la prueba de rangos con signos de Wilcoxon para los resultados de lanzadores normales y élite obtenidos(Wilcoxon, 1945). La Tabla 4.4 los ρ-valores obtenidos con esta prueba para las tres longitudes de series estudiadas. Como se puede apreciar, los valores de F1 estan distribuidos de forma similar, esto es, el número de rangos postivos y negativos no difieren significativamente (para un valor de $\alpha = 0{,}05$). Por lo tanto, se concluye que el rendimiento del modelo de predicción propuesto es igual tanto para los lanzadores élite como para los otros lanzadores.

Longitud de la serie	Lanzadores	Rank. Positivos	Rank Negativos	ρ-valor
25 %	élite - normal	4	4	0.726
50 %	élite - normal	4	5	0.906
75 %	élite - normal	4	5	0.512

Tabla 4.4: Prueba signada de rangos de Wilcoxon para los valores de F1 entre lanzadore élite y normales. Los rangos positivos y negativos se muestran en conjunción con el ρ-valor.

4.5. Sumario

La predicción del desempeño de los lanzadores abridores de béisbol constituye uno de los problemas más complejos de ese deporte. En este capítulo se ha presentado un modelo cuyo principal objetivo es predecir el desempeño de los lanzadores abridores, y por ende, ayudar a decidir en qué momento el lanzador debe ser sacado del partido y sustituido por un relevista. El modelo propuesto es novedoso debido a que utiliza los datos lanzamiento-a-lanzamiento del sistema PITCHF/X para construir series de tiempo, para ello se definió una medida acumulativa que representa el desempeño del lanzador en cada momento del partido. Una vez obtenidas las series de lanzamientos, el aprendizaje y la clasificación de dichas series se realiza utilizando el algoritmo 1NN en conjunción con la medida de similitud DTW. Con el objetivo de validar el modelo propuesto se usaron datos de 20 lanzadores de la MLB, algunos de los cuales son considerados de la élite en este deporte.

Después de analizar los resultados obtenidos en la experimentación, se puede arribar a las siguientes conclusiones:

- La fiabilidad de la predicción del modelo se incrementa en la medida en que aumenta la longitud de serie de lanzamientos.
- El modelo ofrece buenos resultados (con valores de *precision*, *recall* y F1 cercanos al 90 %) cuando el 75 % de los resultados de los lanzamientos son conocidos.
- No se observan diferencias en cuanto a la calidad de la predicción entre los lanzadores promedio y aquellos que son considerados de élite.

En términos generales, los resultados mostraron que la clasificación de series de tiempo constituyen una solución factible para enfrentar problemas de decisión en el contexto

deportivo.

Conclusiones generales

En esta tesis se realizó un estudio del estado del arte sobre aspectos importantes del aprendizaje automático supervisado y de la clasificación de series de tiempo. En particular, se profundizó en el funcionamiento de cuatro métodos clásicos del aprendizaje automático (aprendizaje basado en casos, árboles de decisión, máquinas de soporte vectorial y redes neuronales artificiales). Las características especiales que presentan las series de tiempo, como la dependencia temporal entre los puntos de datos, la alta dimensionalidad entre otras, diferencian su tratamiento en comparación con el de los problemas tradicionales del aprendizaje automático. En este sentido, se plantearon las ventajas que tiene el uso del algoritmo kNN en conjunción con la medida de similitud DTW para la clasificación de series de tiempo.

Por otro lado, se abordaron los elementos fundamentales del análisis cuantitativo de datos deportivos. En particular, se realizó una comparación entre las técnicas estadísticas tradicionales y los métodos del aprendizaje automático, resaltando las principales ventajas y aplicaciones que han tenido estos últimos en este dominio de aplicación. Específicamente, se estudió el juego de béisbol, por ser reconocido como uno de los deportes más complejos y completos en cuanto a estadísticas se refiere. Se profundizó en el estudio de la sabermetría, la cual representa una forma novedosa de analizar todo lo acontecido en el juego de béisbol, detallando sus aportes en el campo del análisis cuantitativo de ese deporte.

Como principal resultado de esta investigación se propusieron dos modelos predictivos, ambos basados en la utilización de métodos del aprendizaje automático, para su aplicación en el béisbol. El primero se diseñó con el propósito de ser empleado para la predicción de resultados de juegos de béisbol, mientras que el segundo modelo tiene como objetivo dar solución a uno de los problemas de decisión más acuciantes de este deporte: la predicción del desempeño de los lanzadores abridores.

Para la evaluación de los modelos se utilizaron fuentes de datos históricos, disponibles públicamente, las cuales posibilitan el análisis detallado de todo lo acontecido en juegos

de la MLB. Se identificó al método de SVM como un algoritmo del aprendizaje automático con un buen desempeño para la predicción de resultados de juegos de béisbol, probándose que sus resultados son competitivos en relación a las estimaciones realizadas por el mercado de apuestas. Respecto a la predicción del desempeño de los lanzadores abridores en el béisbol, los resultados del modelo propuesto demuestran que la clasificación de series de tiempo constituye una solución viable para enfrentar este tipo de problemas de decisión en el contexto deportivo.

Comentarios finales y trabajo futuro

El estudio experimental demuestra la complejidad inherente a la predicción de resultados de juego de béisbol de la MLB. Los cuatro métodos de aprendizaje aplicados muestran resultados de *accuracy* inferiores al 60 %, incluso cuando los más novedosos estadísticos de la sabermetría son usados como predictores base, lo cual representa una mejora respecto a una selección aleatoria pero no es del todo significativo en el contexto de las apuestas deportivas.

El modelo de predicción de juegos propuesto tiene la ventaja de que puede ser expandido con más métodos de aprendizaje automático y también replicado a otros deportes, en los que existan suficientes datos disponibles. La implementación de otros métodos de aprendizaje (basados en reglas, algoritmos genéticos, multi-clasificadores, etc.), así como la adición de un mayor número de atributos a partir de otras fuentes de datos como Baseball Reference[2] podría arrojar resultados diferenes. Sin embargo, dada la complejidad y extensión del conjunto de datos analizado en esta tesis, resulta difícil pensar en la posibilidad de una mejora significativa de estos resultados. No obstante, sería interesante experimentar con datos de ligas de aficionados u otras ligas profesionales tales como la Liga Coreana de Béisbol o la Liga Japonesa, con el objetivo de afianzar este criterio.

En opinión de este autor, en el contexto deportivo las predicciones llevadas a cabo durante durante la celebración de los encuentros debieran arrojar mejores resultados (Percy, 2015). En este sentido, los métodos de aprendizaje automático han resultado ser efectivos en la elaboración de estrategias para algunas situaciones específicas del béisbol. Por ejemplo, en la evaluación de la influencia del lanzador y receptor en el robo de bases (Loughin y Bargen, 2008) o para la predicción de las probabilidades de ponches para diferentes marcadores (Healey, 2015). Sin embargo, como se ha visto en esta tesis, la creación de un modelo capaz de realizar predicciones acertadas de resultados de juegos de béisbol de la MLB resulta ser un campo de investigación abierto actualmente en el dominio del análisis

[2] http://baseball-reference.com

cuantitativo de datos deportivos.

En el futuro, se espera evaluar el modelo de predicción propuesto en otros deportes de equipo como son el baloncesto, voleibol o fútbol. Creemos que el modelo propuesto podría resultar útil en ciertos momentos de la temporada o contra oponentes específicos en estos deportes, siempre con la premisa de que el objetivo final del análisis es transformar el conocimiento en victorias para el equipo.

Respecto al modelo propuesto para la predicción del desempeño de lanzadores abridores, el análisis de los resultados experimentales demostró que su desempeño durante el juego no es un proceso homogéneo. No obstante, el análisis de los datos colectados y su transformación a series de tiempo puede mejorar significativamente nuestra comprensión de este importante renglón del juego. En este sentido, los métodos de clasificación de series de tiempo resultan en una herramienta útil para la predicción del desempeño de lanzadores abridores.

Se debe tener en cuenta que, aunque el modelo de predicción de desempeño propuesto resulta ser un buen predictor cuando se conoce el resultado del 75 % de los lanzamientos, es evidente que los directores de equipo no pueden conocer por adelantado el número total de lanzamientos que el picher abridor realizará en el partido. Por tal motivo, podría no quedar claro cuál sería el momento indicado para aplicar el modelo. No obstante, dado que la cantidad de lanzamientos promedio realizados por los lanzadores analizados en este estudio es de 101, se espera que el resultado obtenido sea confiable a partir del lanzamiento número 50.

El modelo de predicción propuesto puede ser fácilmente extendido a otros deportes en los cuales se disponga de datos jugada-a-jugada, tales como baloncesto (Vračar *et˜al.*, 2016) o cricket (Iyer y Sharda, 2009), siendo posible su inclusión en cualquier sistema experto para la toma de decisiones que requiera de ordenamiento o evaluación de jugadores. En la opinión de este autor, trabajos futuros en este modelo podrían incluir las líneas siguientes:

- Comparar el desempeño del algoritmo 1NN con otros métodos de clasificación de series de tiempo del estado del arte para el aprendizaje de las series de lanzamientos.
- Realizar un estudio más extenso con otros datos disponibles a través del sistema PITCHf/x con el objetivo de llevar a cabo una validación más general de las potencialidades predictivas del modelo propuesto.
- Estimar con mayor exactitud el número de lanzamientos necesarios que deben ser

conocidos para conseguir una predicción aceptable.

- Evaluar la factibilidad de usar el modelo propuesto en otros deporte, especialmente en aquellos que generan una gran cantidad de datos jugada-a-jugada.

Producción científica

A continuación se muestra un listado con las producciones científicas asociadas a la presente tesis:

Publicaciones en revistas arbitradas

- C. Soto y M. González, Sabermetría y nuevas tendencias en el análisis estadístico del juego de béisbol, *Retos. Nuevas tendencias en deportes, educación física y recreación*, Vol. 28, Núm. 2, pág. 122–127, 2015.
- C. Soto, I. Pérez, M. González y A. Brovkina, ACI-Polo: Sistema computacional para el análisis de la actividad competitiva individual en juegos de polo acuático, *Revista Cubana de Ciencias Informáticas*, Vol. 10, Núm. 1, pág. 229–244, 2016.
- C. Soto, Predicting Win-Loss outcomes in MLB regular season games: A comparative study using data mining methods, *International Journal of Computer Science in Sports*, artículo aceptado.
- C. Soto, M. González e I. Pérez, A time series classification model for predicting in-play pitcher's performance in baseball, *Journal of Quantitative Analysis in Sports*, artículo enviado.
- C. Soto, Sobre el análisis cuantitativo de datos deportivos usando métodos de aprendizaje automático, *Revista Internacional de Ciencias del Deporte*, artículo enviado.

Publicaciones en eventos

- C. Soto y M. González, Paquete para la clasificación de series temporales en Weka, *III Conferencia Internacional en Ciencias Computacionales e Informáticas (CICCI'2016)*, ISBN: 978-959-289-122-7, del 14 al 18 de marzo de 2016, La Habana, Cuba.

Monografías

- M. González y C. Soto, Minería de datos para series temporales, Editorial Samuel Feijóo. ISBN: 978-959-250-924-5.

Registro de softwares nacionales

- C. Soto, M. González y Y. Sarabia, timeSeriesClassification: Paquete para la clasificación de series temporales en Weka, CENDA 3833-2014, 2014.
- C. Soto, P. Stanislas y M. González, ACI-Polo: Sistema para el análisis de juegos de polo acuático, CENDA 3265-2013, 2015.

Bibliografía

Adler, J. (2006). *Baseball hacks*. O'Reilly Media, Inc.

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

Ahmad, A. y Dey, L. (2005). A feature selection technique for classificatory analysis. *Pattern Recognition Letters* **26**(1), 43–56.

Albert, J. (2010). Baseball data at season, play-by-play, and pitch-by-pitch levels. *Journal of Statistics Education* **18**(3).

Bagnall, A. J. y Janacek, G. J. (2004). Clustering time series from arma models with clipped data. En: *Proceedings of the Tenth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '04. New York, NY, USA: ACM.

Bartolini, I., Ciaccia, P. y Patella, M. (2005). Warp: accurate retrieval of shapes using phase of fourier descriptors and time warping distance. *IEEE Transactions on Pattern Analysis and Machine Intelligence* **27**(1), 142–147.

Baumer, B. y Zimbalist, A. (2014). Quantifying market inefficiencies in the baseball players' market. *Eastern Economic Journal* **40**(4), 488–498.

Baumer, B. S., Jensen, S. T. y 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.

Behera, H., Dash, P. y Biswal, B. (2010). Power quality time series data mining using S-transform and fuzzy expert system. *Applied Soft Computing* **10**(3), 945–955.

Beneventano, P., Berger, P. D. y Weinberg, B. D. (2012). Predicting run pro-

duction and run prevention in baseball: the impact of sabermetrics. *Int J Bus Humanit Technol* **2**(4), 67–75.

Bhandari, I., Colet, E., Parker, J., Pines, Z., Pratap, R. y Ramanujam, K. (1997). Advanced scout: Data mining and knowledge discovery in nba data. *Data Mining and Knowledge Discovery* **1**(1), 121–125.

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

Bishop, D. (2003). Performance analysis: What is performance analysis, and how can it be integrated within the coaching process to benefit performance? URL http://www.pponline.co.uk/encyc/sports-performance-analysis. Peak Performance.

Box, G. E. y Jenkins, G. M. (1976). *Time series analysis: forecasting and control*. Holden-Day, San Francisco.

Breiman, L., Friedman, J. H., Olshen, R. A. y Stone, C. J. (1984). *Classification and regression trees*. Wadsworth & Brooks/Cole Advanced Books & Software.

Brockwell, P. J. y Davis, R. A. (2006). *Introduction to time series and forecasting*. Springer Science & Business Media, second ed.

Bruno, G. y Garza, P. (2012). Temporal pattern mining for medical applications. *Intelligent Systems Reference Library* **25**, 9–18.

Buza, K., Nanopoulos, A., Schmidt-Thieme, L. y Koller, J. (2011). Fast classification of electrocardiograph signals via instance selection. En: *First IEEE International Conference on Healthcare Informatics, Imaging and Systems Biology (HISB), 2011*.

Carden, E. P. y Brownjohn, J. M. (2008). ARMA modelled time-series classification for structural health monitoring of civil infrastructure. *Mechanical Systems and Signal Processing* **22**(2), 295–314.

Chang, J. y Zenilman, J. (2013). A study of sabermetrics in major league baseball: The impact of moneyball on free agent salaries.

Chatfield, C. (2013). *The analysis of time series: an introduction*. CHAPMAN & HALL/CRC Texts in Statistical Science. CRC press, sixth edition ed.

Chen, L. y Ng, R. (2004). On the marriage of lp-norms and edit distance. En: *Proceedings of the Thirtieth International Conference on Very Large Data Bases - Volume 30*, VLDB '04. VLDB Endowment.

Chen, L., Özsu, M. T. y Oria, V. (2005). Robust and fast similarity search for moving object trajectories. En: *Proceedings of the 2005 ACM SIGMOD International Conference on Management of Data*, SIGMOD '05. New York, NY, USA: ACM.

Corduas, M. y Piccolo, D. (2008). Time series clustering and classification by the autoregressive metric. *Computational Statistics & Data Analysis* **52**(4), 1860–1872.

Cortes, C. y Vapnik, V. (1995). Support-vector network. *Machine Learning* **20**, 1–25.

Costa, G. B., Huber, M. R. y Saccoman, J. T. (2007). *Understanding sabermetrics: An introduction to the science of baseball statistics*.

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

Cover, T. y Hart, P. (1967). Nearest neighbor pattern classification. *IEEE Transactions on Information Theory* **13**(1), 21–27.

Cowpertwait, P. S. y Metcalfe, A. V. (2009). *Introductory time series with R*. Springer Science & Business Media.

Dash, M. y Liu, H. (2003). Consistency-based search in feature selection. *Artificial Intelligence* **151**(1), 155–176.

Dash, P., Behera, H. y Lee, I. (2008). Time sequence data mining using time-frequency analysis and soft computing techniques. *Applied Soft Computing* **8**(1), 202–215.

Davids, B. (1971). Society for american baseball research. URL http://sabr.org/.

Davoodi, E. y 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. Tech. rep., University of Leeds, School of Computing Studies.

Delen, D., Cogdell, D. y Kasap, N. (2012). A comparative analysis of data mining methods in predicting ncaa bowl outcomes. *International Journal of Forecasting* **28**(2), 543–552.

Demens, S. (2015). Riding a probabilistic support vector machine to the stanley cup. *Journal of Quantitative Analysis in Sports* **11**(4), 205–218.

Douzal-Chouakria, A. y Amblard, C. (2012). Classification trees for time series. *Pattern Recognition* **45**(3), 1076–1091.

Dubbs, A. (2015). Statistics-free sports prediction. *arXiv preprint arXiv:1512.07208* .

Edelman-Nusser, J., Hohmann, A. y Henneberg, B. (2002). Modeling and prediction of competitive performance in swimming upon neural networks. *European Journal of Sport Science* **2**(2), 1–10.

Edge, I. (2016). URL http://www.inside-edge.com/.

Esling, P. y Agon, C. (2012). Time-series data mining. *ACM Comput. Surv.* **45**(1), 12:1–12:34.

Fangraphs (2016). URL http://www.fangraphs.com/.

Fast, M. (2010). What the heck is pitchf/x. *The Hardball Times Annual* **2010**, 153–158.

Fister Jr, I., Ljubič, K., Suganthan, P. N., Perc, M. y Fister, I. (2015). Computational intelligence in sports: Challenges and opportunities within a new research domain. *Applied Mathematics and Computation* **262**, 178–186.

Fu, T.-c. (2011). A review on time series data mining. *Engineering Applications of Artificial Intelligence* **24**, 164–181.

Fulcher, B. D. y Jones, N. S. (2014). Highly comparative feature-based time-series classification. *IEEE Transactions on Knowledge and Data Engineering* **26**(12), 3026–3037.

Geurts, P. (2001). *Pattern Extraction for Time Series Classification*. Berlin, Heidelberg: Springer, pp. 115–127.

Giorgino, T. (2009). Computing and visualizing dynamic time warping alignments in r: the dtw package. *Journal of statistical Software* **31**(7), 1–24.

Goldberger, A. L., Amaral, L. A., Glass, L., Hausdorff, J. M., Ivanov, P. C., Mark, R. G., Mietus, J. E., Moody, G. B., Peng, C. y Stanley, H. E. (2000). Physiobank, physiotoolkit, and physionet components of a new research resource for complex physiologic signals. *Circulation* **101**(23), e215–e220.

Gomez, M. A., Pollard, R. y Luis-Pascual, J.-C. (2011). Comparison of the home advantage in nine different professional team sports in spain. *Perceptual and motor skills* **113**(1), 150–156.

Grau, I. (2011). Aprendizaje de redes neuronales recurrentes con instancias de longitud variable. aplicaciones a la resistencia antiviral del vih. Universidad Central de Las Villas.

Gumm, J., Barrett, A. y Hu, G. (2015). A machine learning strategy for predicting march madness winners. En: *16th International Conference on Software Engineering, Artificial Intelligence, Networking and Parallel/Distributed Computing (SNPD)* (IEEE, ed.).

Gunn, S. R. *et al.* (1998). Support vector machines for classification and regression. *ISIS technical report* **14**.

Haghighat, M., Rastegari, H. y 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.

Hall, M., Frank, E., Holmes, G., Pfahringer, B., Reutemann, P. y Witten, I. H. (). The weka data mining software: an update. *SIGKDD Explor. Newsl.* **11**(1), 10–18.

Hamilton, M., Hoang, P., Layne, L., Murray, J., Padget, D., Stafford, C. y Tran, H. (2014). Applying machine learning techniques to baseball pitch prediction. En: *ICPRAM*.

Hammer, B. y Villmann, T. (2003). Mathematical aspects of neural networks. En: *ESANN*.

Hamooni, H., Mueen, A. y Neel, A. (2015). Phoneme sequence recognition via dtw-based classification. *Knowledge and Information Systems* , 1–23.

Han, J., Pei, J. y Kamber, M. (2011). *Data mining: concepts and techniques*. Elsevier.

Healey, G. (2015). Modeling the probability of a strikeout for a batter/pitcher matchup. *Knowledge and Data Engineering, IEEE Transactions* **27**(9), 2415–2423.

Hilera, J. R. y Martínez, V. J. (1995). Redes neuronales artificiales. *Fundamentos, modelos y aplicaciones. Ed. Ra-ma (l 995)* .

Hoang, P., Hamilton, M., Murray, J., Stafford, C. y Tran, H. (2015). A dynamic feature selection based lda approach to baseball pitch prediction. En: *Trends and Applications in Knowledge Discovery and Data Mining* (Springer, ed.).

Hua, K.-L., Lai, C.-T., You, C.-W. y Cheng, W.-H. a. (2015). An efficient pitch-

by-pitch extraction algorithm through multimodal information. *Information Sciences* **294**, 64–77.

Itakura, F. (1975). Minimum prediction residual principle applied to speech recognition. *IEEE Transactions on Acoustics, Speech, and Signal Processing* **23**(1), 67–72.

Iyer, S. R. y Sharda, R. (2009). Prediction of athletes performance using neural networks: An application in cricket team selection. *Expert Systems with Applications* **36**(3, Part 1), 5510 – 5522.

James, B. (1982). *The Bill James historical baseball abstracts*. Ballantine Books: New York.

Jamieson, J. P. (2010). Home field advantage in athletics: a meta-analysis. *Journal of Applied Social Psychology* , 819–1848.

Java, A. y Perlman, E. S. (2002). Predictive mining of time series data. En: *American Astronomical Society Meeting Abstracts #200*, vol. 34 of *Bulletin of the American Astronomical Society*.

Jeff, H. y John, R. (2011). Using local correlation to explain success in baseball. *Journal of Quantitative Analysis in Sports* **7**(2), 1–29.

Jelinek, H. F., Kelarev, A., Robinson, D. J., Stranieri, A. y 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.

Kaya, H. y GunduzOguducu, S. (2015). A distance based time series classification framework. *Information Systems* **51**, 27–42.

Keogh, E. y Kasetty, S. (2003). On the need for time series data mining benchmarks: a survey and empirical demonstration. *Data Mining and knowledge discovery* **7**(4), 349–371.

Keogh, E. y Ratanamahatana, C. (2005). Exact indexing of dynamic time warping. *Knowledge and Information Systems* **7**(3), 358–386.

Knottenbelt, W. J., Spanias, D. y 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.

Kohavi, R. y Quinlan, J. R. (2002). Data mining tasks and methods: Classification:

decision-tree discovery. En: *Handbook of data mining and knowledge discovery.* Oxford University Press, Inc.

Kovacs-Vajna, Z. M. (2000). A fingerprint verification system based on triangular matching and dynamic time warping. *IEEE Transactions on Pattern Analysis and Machine Intelligence* **22**(11), 1266–1276.

Kurbalija, V., Radovanović, M., Geler, Z. y Ivanović, M. (2014). The influence of global constraints on similarity measures for time-series databases. *Knowledge-Based Systems* **56**, 49–67.

Lei, H. y Sun, B. (2007). A study on the dynamic time warping in kernel machines. En: *Signal-Image Technologies and Internet-Based System,* (IEEE, ed.).

Leung, C. K. y Joseph, K. W. (2014). Sports data mining: predicting results for the college football games. *Procedia Computer Science* **35**, 710–719.

Levenshtein, V. I. (1966). Binary codes capable of correcting deletions, insertions, and reversals. En: *Soviet physics doklady*, vol. 10.

Lewis, M. (2004). *Moneyball: The art of winning an unfair game.* WW Norton & Company.

Liao, S.-H., Chu, P.-H. y Hsiao, P.-Y. (2012). Data mining techniques and applications – a decade review from 2000 to 2011. *Expert Systems with Applications* **39**(12), 11303–11311.

Lin, W., Orgun, M. A. y Williams, G. J. (2002). An overview of temporal data mining. En: *Proceedings of the 1st Australian data mining workshop.*

Link, D. y Lames, M. (2009). Sport informatics: Historical roots, interdisciplinarity and future developments. *International Journal of Computer Science in Sports* **8**(2), 68–87.

Lock, D. y 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.

Loughin, T. M. y Bargen, J. L. (2008). Assessing pitcher and catcher influences on base stealing in major league baseball. *Journal of sports sciences* **26**(11), 15–20.

Lyle, A. (2007). *Baseball Prediction Using Ensemble Learning*. Tesis de Maestría, The University of Tulsa.

Marteau, P. F. (2009). Time warp edit distance with stiffness adjustment for time series matching. *IEEE Transactions on Pattern Analysis and Machine Intelligence* **31**(2), 306–318.

Menéndez, H. D., Vázquez, M. y Camacho, D. (2016). Mixed clustering methods to forecast baseball trends. En: *Intelligent Distributed Computing VIII*. Springer.

Mercer, J. (1909). Functions of positive and negative type and their connection with the theory of integral equations. *Philosophical Transactions of the Royal Society* , 415–446.

Mitchell, T. M. (1997). *Machine Learning*. McGraw-Hill.

Morgan, S., Williams, M. D. y 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.

Niennattrakul, V., Wanichsan, D. y Ratanamahatana, C. A. (2007). *Knowledge-Based Intelligent Information and Engineering Systems: 11th International Conference, KES 2007, XVII Italian Workshop on Neural Networks, Vietri sul Mare, Italy, September 12-14, 2007. Proceedings, Part II*, chap. Hand Geometry Verification Using Time Series Representation. Berlin, Heidelberg: Springer Berlin Heidelberg, pp. 824–831.

Odachowski, K. y Grekow, J. (2013). Using bookmaker odds to predict the final result of football matches. En: *Knowledge Engineering, Machine Learning and Lattice Computing with Applications: 16th International Conference* (Springer, ed.).

Ofoghi, B., Zeleznikow, J., MacMahon, C. y Dwyer, D. (2010a). A machine learning approach to predicting winning patterns in track cycling omnium. En: *IFIP International Conference on Artificial Intelligence in Theory and Practice*. Springer.

Ofoghi, B., Zeleznikow, J., MacMahon, C. y Dwyer, D. (2010b). A machine learning approach to predicting winning patterns in track cycling omnium. En: *IFIP International Conference on Artificial Intelligence in Theory and Practice*. Springer.

Oppenheim, A. V., Schafer, R. W. y Buck, J. R. (1999). *Discrete-time Signal Processing*. Upper Saddle River, NJ, USA: Prentice-Hall, Inc., second ed.

O'Reilly y Knight, N. . P. (2007). Knowledge management best practices in national sport organizations. *International Journal of Sport Management and Marketing* .

Paul, R. J. y Weinbach, A. P. (2009). Sportsbook pricing and the behavioral biases of bettors in the nhl. *Journal of Economics and Finance* **36**(1), 123–135.

Pavitt, C. (2011). An estimate of how hitting, pitching, fielding, and basestealing impact team winning percentages in baseball. *Journal of Quantitative Analysis in Sports* **7**(4), 1–18.

Percy, D. F. (2015). Strategy selection and outcome prediction in sport using dynamic learning for stochastic processes. *Journal of the Operational Research Society* **66**(11), 1840–1849.

Petitjean, F., Forestier, G., Webb, G. I., Nicholson, A. E., Chen, Y. y Keogh, E. (2016). Faster and more accurate classification of time series by exploiting a novel dynamic time warping averaging algorithm. *Knowledge and Information Systems* **47**(1), 1–26.

Piatetsky, S. (2016). Difference between data mining and statistics. URL http://www.kdnuggets.com/faq/difference-data-mining-statistics.html.

Piette, J., Braunstein, A., McShane, B. B. y Jensen, S. T. (2010). A point-mass mixture random effects model for pitching metrics. *A Point-Mass Mixture Random Effects Model for Pitching Metrics* **10**(2), 1–15.

Pollard, R. (1986). Home advantage in soccer: A retrospective analysis. *Journal of Sports Sciences* **4**(3), 237–248.

Pollard, R. y 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**, 337–350.

Povinelli, R., Johnson, M., Lindgren, A. y Ye, J. (2004). Time series classification using gaussian mixture models of reconstructed phase spaces. *IEEE Transactions on Knowledge and Data Engineering* **16**(6), 779–783.

Povinelli, R. J. (1999). *Time series data mining: identifying temporal patterns for characterization and prediction of time series events*. Tesis de Doctorado, Faculty of the Graduate School, Marquette University.

Quinlan, J. R. (1986). Induction of decision trees. *Machine learning* **1**(1), 81–106.

Rakthanmanon, T., Campana, B., Mueen, A., Batista, G., Westover, B., Zhu, Q., Zakaria, J. y Keogh, E. (2012). Searching and mining trillions of time series subsequences under dynamic time warping. En: *Proceedings of the 18th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '12. New York, NY, USA: ACM.

Rama Iyer, S. R., Subramanian (2009). Prediction of athletes performance using neural networks: An application in cricket team selection. *Expert Systems with Applications* **36**.

Ratanamahatana, C. y Keogh, E. (2005). Three myths about dynamic time warping data mining. En: *SDM'05*.

Robinson, S. J. (2014). Extracting individual offensive production from baseball run distributions. *International Journal of Computer Science in Sport* **13**(2).

Robnik-Šikonja, M. y Kononenko, I. (1997). An adaptation of relief for attribute estimation in regression. En: *Machine Learning: Proceedings of the Fourteenth International Conference (ICML'97)*.

Rodríguez, J. J. y Alonso, C. J. (2004). Interval and Dynamic Time Warping-based decision trees. En: *Proceedings of the 2004 ACM Symposium on Applied Computing*, SAC '04. ACM.

Rodríguez, J. J., Alonso, C. J. y Boström, H. (2000). Learning first order logic time series classifiers: Rules and boosting. En: *Principles of Data Mining and Knowledge Discovery*, vol. 1910 of *Lecture Notes in Computer Science*. Springer Berlin Heidelberg, pp. 299–308.

Rokach, L. y Maimon, O. (2008). *Data mining with decision trees: theory and applications*. World Scientific.

Sakoe, H. y Chiba, S. (1978). Dynamic programming algorithm optimization for spoken word recognition. *IEEE Transactions on Acoustics, Speech and Signal Processing* **26**(1), 43–49.

Sauer, R. D., Waller, J. K. y Hakes, J. K. (2010). The progress of the betting in a baseball game. *Public Choice. Springer* **142**, 297–313.

Schumaker, R. P., Solieman, O. K. y Chen, H. (2010a). Predictive modeling for

sports and gaming. En: *Predictive modeling for sSports and gaming* (Springer, ed.), vol. 26.

Schumaker, R. P., Solieman, O. K. y Chen, H. (2010b). Sports data mining. *Integrated series in information systems. New York, NY: Springer.* **26**.

Scout, D. (2016). URL http://www.digitalscout.com/.

Serrà, J. y Arcos, J. L. (2014). An empirical evaluation of similarity measures for time series classification. *Knowledge-Based Systems* **67**(0), 305–314.

Shahnawaz, M., Ranjan, A. y Danish, M. (2011). Temporal data mining: an overview. *International Journal of Engineering and Advanced Technology* **1**(1).

Shannon, C. (1948). Mathematical theory of communication. *Bell System Technical Journal* **27**, 379–423.

Shao, S. (2009). Application of bp neural network model in sports aerobics performance evaluation. En: *Knowledge Engineering and Software Engineering, 2009. KESE'09. Pacific-Asia Conference.* IEEE.

Shearer, C. (2000). The crisp-dm model: the new blueprint for data mining. *Journal of Data Warehousing* **5**, 13–22.

Shumway, R. H. y Stoffer, D. S. (2010). *Time series analysis and its applications: with R examples.* Springer Science & Business Media, third ed.

Sidhu, G. y Caffo, B. (2014). Moneybarl: Exploiting pitcher decision-making using reinforcement learning. *The Annals of Applied Statistics* **8**(2), 926–955.

Sidran, D. E. (2005). A method of analyzing a baseball pitcher's performance based on statistical data mining.

Spann, M. y 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.

Staufenbiel, K., Lobinger, B. y Strauss, B. (2015). Home advantage in soccer: A matter of expectations, goal setting and tactical decisions of coaches? *Journal of Sports Sciences* **33**(18), 1932–1941.

Stephen Ockerman, M. N. (2014). Predicting the cy young award winner. *Pure Insights* **3**(1).

Sun, J., Yu, W. y Zhao, H. (2010). Study of association rule mining on technical action of ball games. En: *2010 International Conference on Measuring Technology and Mechatronics Automation*, vol. 3. IEEE.

Sykora, M., Chung, P. W. H., Folland, J. P., Halkon, B. J. y Edirisinghe, E. A. (2015). Advances in sports informatics research. En: *Computational Intelligence in Information Systems* (Springer, ed.).

Tango, T. (2016). Tangotiger. URL http://www.tangotiger.com.

Thorn, J., Palmer, P. y & Reuther, D. (1984). *The hidden game of baseball: A revolutionary approach to baseball and its statistics*. Doubleday Garden City, New York.

Thrun, S. y Smieja, F. (1990). A general feed-forward algorithm for gradient descent in connectionist networks .

Tormene, P., Giorgino, T., Quaglini, S. y Stefanelli, M. (2009). Matching incomplete time series with dynamic time warping: an algorithm and an application to post-stroke rehabilitation. *Artificial Intelligence in Medicine* **45**(1), 11–34.

Trawinski, K. (2010). A fuzzy classification system for prediction of the results of the basketball games. En: *Fuzzy Systems (FUZZ), 2010 IEEE International Conference* (IEE, ed.).

Valero, C. S., Morales, I. P., Castellanos, M. G. y 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.

Vapnik, V. (2013). *The nature of statistical learning theory*. Springer Science & Business Media, second ed.

Vračar, P., Štrumbelj, E. y Kononenko, I. (2016). Modeling basketball play-by-play data. *Expert Systems with Applications* **44**, 58–66.

Wang, X., Mueen, A., Ding, H., Trajcevski, G., Scheuermann, P. y Keogh, E. (2013). Experimental comparison of representation methods and distance measures for time series data. *Data Mining and Knowledge Discovery* **26**(2), 275–309.

Weng, X. y Shen, J. (2008). Classification of multivariate time series using two-dimensionalsingular value decomposition. *Knowledge-Based Systems* **21**(7), 535–539.

Wilcoxon, F. (1945). Individual comparisons by ranking methods. *Biometrics* **1**(80-83).

Witnauer, W. D., Rogers, R. G. y Saint Onge, J. M. (2007). Major league baseball career length in the 20th century. En: *Population research and policy review* (Springer, ed.), vol. 26.

Witten, I. H., Frank, E. y Hall, M. A. (2011). *Data Mining: Practical machine learning tools and techniques*. Boston: Morgan Kaufmann, third edition ed.

Wolf, G. H. (2015). The sabermetric revolution: Assessing the growth of analytics in baseball by benjamin baumer and andrew zimbalist (review). *Journal of Sport History* **2**, 239–241.

Wolpert, D. H. y Macready, W. G. (1997). No free lunch theorems for optimization. *IEEE Transactions on Evolutionary Computation* **1**(1), 67–82.

Woodland, L. M. y Woodland, B. M. (1994). Market efficiency and the favorite-longshot bias: The baseball betting market. *The Journal of Finance* **49**(1), 269–279.

Xi, X., Keogh, E., Shelton, C., Wei, L. y Ratanamahatana, C. A. (2006). Fast time series classification using numerosity reduction. En: *Proceedings of the 23rd International Conference on Machine Learning*, ICML '06. New York, NY, USA: ACM.

Yuan, L.-H., Liu, A., Yeh, A., Kaufman, A., Reece, A., Bull, P., Franks, A., Wang, S., Illushin, D. y Bornn, L. (2015). A mixture-of-modelers approach to forecasting ncaa tournament outcomes. *Journal of Quantitative Analysis in Sports* **11**(1), 13–27.

Zeng, X. y Martinez, T. R. (2000). Distribution-balanced stratified cross-validation for accuracy estimation. *Journal of Experimental & Theoretical Artificial Intelligence* **12**(1), 1–12.

Zhu, X. y Goldberg, A. B. (2009). Introduction to semi-supervised learning. *Synthesis lectures on artificial intelligence and machine learning* **3**(1), 1–130.

Zimniuch, F. (2010). *Fireman: The Evolution of the Closer in Baseball*. Triumph Books.

Zurada, J. M. (1992). *Introduction to artificial neural systems*, vol. 8. West St. Paul.

Anexos

Anexo A

Descripción de algunos de los principales estadísticos propuestos por la sabermetría

Han sido muchos los aportes de la sabermetría en el béisbol, pero tal vez los más evidentes resultan ser los novedosos estadísticos creados con el objetivo de medir con mayor precisión lo acontecido en el juego. En este apéndice se ofrece una descripción de algunos de los estadísticos de bateo, picheo y defensa que, por su importancia, se han consolidado como herramientas indispensables para una mejor comprensión y análisis de este deporte.

A.1. Estadísticos de bateo

A continuación se presentan algunos de los principales estadísticos de bateo propuestos por la sabermetría. La idea fundamental en este caso consiste en neutralizar los factores que no dependan del bateador, tales como el desempeño de sus contrarios, las características del terreno donde juega, etc.

OBP

El OBP es considerado el estadístico que comenzó la revolución de la sabermetría en el béisbol, demostrando que es más importante evitar que al bateador lo pongan *out* (lo que mide el OBP) a dar un sencillo (lo que mide el AVG). Los jugadores con buen OBP (más de .350) suelen estar en los primeros lugares en la alineación (1-4). Para tener un buen

OBP ayuda que el bateador sea paciente y se tome un buen número de boletos.

$$OBP = \frac{H + BB + HBP}{AB + BB + HBP + SF}$$

SLG

Relaciona los turnos al bate con las bases alcanzadas. Cada base equivale a mil puntos, por ejemplo HR=4.000, 3B=3.000, 2B=2.000, etc. Un turno fallido tiene valor cero. Un buen *slugging* depende del poder del bateador y de su capacidad para mantener un buen promedio de bateo. Regularmente un bateador con un *slugging* elevado supera la marca de .600.

$$SLG = \frac{1B + (2 \cdot 2B) + (3 \cdot 3B) + (4 \cdot 4B)}{AB + SF}$$

OPS

Es la suma del OBP y el SLG para unir la utilidad de ambos en un solo estadístico más completo del bateador bateadores. Como el valor máximo de OBP es 1 y el máximo de Slugging es 4, el valor máximo de OPS posible es 5. Se estima que un bateador es realmente completo una vez que su OPS supera el valor de 1.000.

$$OPS = OBP + SLG$$

OPS+

Constituye un ajuste al OPS. Si se compara el OPS jugador por jugador, se puede notar que regularmente beneficia a los bateadores que tienen un *slugging* alto sobre los que tienen un OBP alto. El *slugging* promedio es casi dos veces el OBP promedio. Por ello se ha creado este estadístico que los normaliza y además ajusta la ventaja o desventaja de jugar en una liga determinada. Un jugador promedio debe tener un OPS+ de 100, así que todo número por encima de 100 será positivo y por debajo será negativo.

$$OPS+ = 100 \cdot \left(\frac{OBP}{lgOBP} + \frac{SLG}{lgSLG} - 1 \right)$$

BABIP

Se trata de la medición del promedio de bateo tomando en consideración únicamente las pelotas puestas en juego que no resultan en un error o en un cuadrangular, descartando igualmente las bases por bolas y los ponches. En otras palabras, determina cuántas pelotas que fueron bateadas cayeron terminaron como sencillos. En el caso de los bateadores, ellos sí tienen control sobre su BABIP. Por ejemplo, bateadores de líneas que suelen hacer buen contacto con la pelota o bateadores rápidos tienden a superar los .300 puntos de BABIP (y viceversa). Un BABIP en períodos cortos (hasta una temporada) se puede comparar con el BABIP vitalicio del bateador y ver qué tanto influyó la suerte en su desempeño. Un BABIP muy alto puede indicar que más pelotas de lo normal han resultado en sencillos, por lo que se puede esperar que tanto el BABIP, como el promedio y el OBP disminuyan en el futuro.

$$BABIP = \frac{H + HR}{AB - K - HR - SF}$$

BB %

Estima con qué regularidad un jugador recibe una base por bolas cuando va a batear, lo que en cierto modo indica lo paciente que puede ser un bateador en el plato. Considera que más que dar un hit, lo más importante es evitar que pongan *out* al bateador, lo cual se consigue obteniendo bases por bolas. Comparar los boletos con las apariciones legales al plato permite determinar qué tanto ayuda un bateador con su paciencia a su OBP. Favorece a los bateadores de poder y aquellos que reciben muchos lanzamientos por turno.

$$BB\,\% = \frac{BB}{AB}$$

K %

Estima con qué regularidad un jugador se poncha cuando va a batear. De este modo, es posible comparar lo factible que puede ser que un bateador ponga la pelota en juego en una situación determinada con respecto cualquier otro, incluso para distintas épocas y equipos. Favorece a los bateadores de contacto y hace lo contrario con los de poder. Un

nivel bueno se encuentra por debajo de 10 % y uno malo se considera por encima del 27 %.

$$K\,\% = \frac{K}{AB}$$

ISO

Resta los sencillos a las bases conseguidas con dobles, triples y cuadrangulares, con lo cual se mide la capacidad de dar extrabases, calculándose el poder bruto del bateador. Aunque lo ideal sería que el jugador lograse un equilibrio entre los aportes en promedios de bases alcanzadas y de poder, la mayoría de las veces no sucede así, por lo que el ISO constituye una medida importante para determinar los méritos ofensivos del bateador. En este sentido, se puede decir que bateadores que cuentan con un ISO por debajo del promedio son jugadores de velocidad que dependen mucho de los sencillos para mantener sus promedios de bateo. Igualmente es importante señalar que el ISO es un estadístico que alcanza una significación predictiva confiable a los 550 turnos al bate o más, por lo que un ISO de 0.350 durante unos 50 turnos es una muestra insuficiente para predecir si este valor corresponde al talento real del bateador o no.

$$ISO = SLG - AVE$$

RC

Estima cuántas carreras ha aportado un bateador a su equipo, independientemente del rendimiento de dicho equipo. Esta es la ecuación inicial propuesta por Bill James, la cual ha tenido varias modificaciones en las que se le da una mayor o menor importancia al OBP, así como se le añade el promedio con hombres en posición anotadora o el éxito en el robo de bases etc.

$$RC = \frac{(H + BB) \cdot TB}{PA}$$

wOBA

El wOBA tiene como fin equilibrar los valores del OPS. La premisa de este estadístico es que el OPS tiende a favorecer porcentualmente a jugadores con *sluggings* más altos, ya que este último sobrestima el valor de los extra bases (por ejemplo, estadísticamente un doble

tiene una expectativa de carreras de 0.77, mientras que un sencillo tiene una expectativa de 0.47, lo cual no se ajusta con la relación 2-1 que se aplica para el cálculo del *slugging*). Es importante señalar que los eventos ofensivos son previamente multiplicados por 0.15 (15 %) con la finalidad de ajustarse a una escala similar a la del OBP.

$$wOBA = \frac{wBB + wHBP + w1B + wRBOE + w2B + w3B + wHR}{PA}$$

donde

$wBB = 0{,}72 * BB$

$wHBP = 0{,}90 * 1B$

$w1B = 0{,}90 * 1B$

$wRBOE = 0{,}92 * RBOE$

$w2B = 1{,}24 * 2B$

$w3B = 1{,}56 * 3B$

$wHR = 1{,}95 * HR$

wRAA

Tiene como propósito medir cuantas carreras aporta un jugador por encima del promedio de la liga en término de carreras. En este sentido, un jugador que cuente con un wRAA positivo es un jugador que está aportando ofensivamente a su equipo en términos de carreras por encima del jugador promedio de la liga. Dado que el wRAA se mide en carreras, es fácil convertir su valor en victorias para determinar cuánto aporta un jugador en este renglón por encima del promedio. Si se tienen dos jugadores con dos promedios de wOBA idénticos, el que tenga mayor cantidad de turnos al bate tendrá el mayor wRAA.

$$wRAA = \left(\frac{wOBA - lgWOBA}{1{,}15}\right) \cdot PA$$

wRC

El wRC una versión mejorada, desarrollada por Tango (2016), de las carreras creadas que había introducido al mundo de la sabermetría Bill James. Año tras año se ha ido mejorando

la ecuación para conocer el total de carreras que aporta ofensivamente cada bateador a su equipo. Una vez creado el wOBA, Tango incluyó una breve ecuación que demuestra que con los valores predeterminados de cada evento del juego se puede precisar cuántas carreras agrega o sustrae cada jugador al equipo. El wOBAScale es una constante que varía según la temporada y cuyo valor ronda los 1.15. Valores de wRC iguales o superiores a 105 son considerados muy buenos.

$$wRC = \left(\frac{wOBA - lgWOBA}{wOBAScale} + \frac{lgR}{PA}\right) \cdot PA$$

A.2. Estadísticos de picheo

A continuación se presentan algunos de los principales estadísticos de picheo propuestos por la sabermetría. Su principio básico consiste en separar la actuación de los lanzadores de la de su equipo tanto a la ofensiva como a la defensiva partiendo de la premisa de que cuando los estadísticos no dependen únicamente del lanzador entonces no son los adecuados para evaluar su actuación o predecir sus futuras temporadas

K/9

Promedio de ponches de un lanzador por cada nueve entradas lanzadas.

$$K/9 = \left(\frac{K}{IP}\right) \cdot 9$$

BB/9

Promedio de bases por bola de un lanzador por cada nueve entradas lanzadas.

$$BB/9 = \left(\frac{BB}{IP}\right) \cdot 9$$

HR/9

Promedio de jonrones permitidos por cada nueve entradas lanzadas.

$$HR/9 = \left(\frac{HR}{IP}\right) \cdot 9$$

K/BB

Cantidad de ponches por cada boleto.

$$K/BB = \frac{K}{BB}$$

WHIP

Mide el número de desplazamientos entre bases que un lanzador permite por entradas lanzadas. Constituye uno de los estadísticos más usados actualmente para evaluar la efectividad de un lanzador. Valores cercanos a 1.00 o inferiores son considerados muy buenos.

$$WHIP = \frac{BB + H}{IP}$$

BABIP

Es el promedio de bateo de los oponentes sin contar los ponches ni los jonrones. En otras palabras, determina cuántas pelotas de las que le batearon al lanzador constituyeron sencillos. Los cuadrangulares no los cuenta porque ellos no dependen de los defensores. Si un lanzador tiene un BABIP mucho menor a .300 esto significa que ha tenido suerte y podemos esperar cierta regresión a la norma en otros de sus estadísticos, como por ejemplo los de efectividad. Y viceversa, un lanzador con un BABIP mucho mayor a .300 ha tenido mala suerte, y lo más seguro es que en el futuro si sigue con la misma relación de K/BB/IP su BABIP y su efectividad bajen.

$$BABIP = \frac{H + HR}{AB - K - HR - SF}$$

LOB %

Calcula de todos los corredores que se le embasaron a un lanzador cuántos de ellos quedaron en circulación cuando se terminó la entrada. Es un indicador parecido al BABIP, en el sentido que ayudan a predecir si un lanzador ha tenido suerte o no. Eso sí, los lanzadores buenos tienden a tener un LOB % mayor a los lanzadores malos, simplemente porque permiten menos sencillos con hombres en base y les anotan menos carreras. No obstante, si esa cifra se aleja demasiado del 71 %, es muy improbable que se pueda mantener.

$$LOB\,\% = \frac{H + BB + HBP - R}{H + BB + HBP\text{–}(1{,}4 \cdot HR)}$$

PR

Ajusta el promedio de carreras limpias de un lanzador atendiendo al rendimiento general de picheo en la liga. Un valor de cero representa un rendimiento promedio, uno por encima de cero indica que el lanzador tiene un rendimiento superior a la media de la liga, y si es menor de cero significa que su rendimiento es peor que el de la liga.

$$PR = \left(\text{Entradas lanzadas} \cdot \frac{lgERA}{9}\right) - \text{Carreras permitidas}$$

GB %

Es una relación entre los batazos de *rolling* que recibe un lanzador en comparación con el total de batazos recibidos. Aparte de la relación de ponches y boletos por cada nueve entradas, este es otro estadístico importante a la hora de evaluar a un lanzador, ya que sí depende por entero de su desempeño. Dependiendo de su repertorio y su forma de lanzar, un pitcher permitirá más o menos *rollings* que otro (batazo que va por el piso, que no puede ser cuadrangular, y que difícilmente sea un extrabase, además de producir más jugadas que permiten hacer dos *outs*, conocidas como *dobleplays*).

$$GB\,\% = \frac{\text{Batazos de } \textit{rolling}}{Batazostotales}$$

FB %

Se usa para medir, de todos los batazos elevados recibidos por un lanzador, cuántos de ellos fueron cuadrangulares. Al igual que con el BABIP, y el LOB % (a menor escala), el pitcher no controla su HR/FB %. Se ha demostrado que mientras más crece la muestra, más tiende la cifra a acercarse a un 11 %. Sabiendo esto, para predecir cuadrangulares recibidos resulta mejor tomar en cuenta el FB % en vez de otras de relaciones (como la comúnmente usada HR/9).

$$FB\,\% = \frac{\text{Batazos de elevados}}{Batazostotales}$$

FIP

Es un medidor de la efectividad del lanzador calculado únicamente sobre la base de los ponches, boletos y cuadrangulares recibidos, los cuales son las únicos estadísticos que no dependen de la defensa. En otras palabras, mide cuántas carreras por cada nueve entradas ha debido recibir un lanzador sobre la base de sus ponches, bases por bolas y cuadrangulares.

$$FIP = \frac{13 \cdot HR + 3 \cdot BB – 2 \cdot K}{IP + ConstanteFIP}$$

donde la $ConstanteFIP$ es un escalar utilizado para ajustar el valor del FIP al promedio de carreras limpias, varía según la temporada pero ronda el valor de 3.10, puede calcularse de la siguiente manera:

$$ConstanteFIP = lgERA \cdot \frac{(13 \cdot lgHR) + (3 \cdot (lgBB + lgHBP)) – (2 \cdot lgK)}{lgIP}$$

xFIP

Simplemente se utilizan los «cuadrangulaes normalizados» en vez de los cuadrangulares recibidos. Los cuadrangulares normalizados se obtienen multiplicando los batazos elevados recibidos por 0.11. El xFIP ayuda en la determinación del efecto de cada estadio sobre la actuación del lanzador. Que a un lanzador la bateen más o menos *rollings* no depende del estadio donde lance, pero los cuadrangulares que le conecten sí.

$$xFIP = \frac{13 \cdot HR_{Normalizados} + 3 \cdot BB – 2 \cdot K}{IP + ConstanteFIP}$$

A.3. Estadísticos de defensa

A continuación se presentan algunos de los principales estadísticos de defensa propuestos por la sabermetría. La idea fundamental es neutralizar los factores que no dependan del bateador, como el desempeño de sus contrarios, las características del estadio donde juega, etc.

F %

Muy utilizado para medir la efectividad de un jugador a la defensiva, este porcentaje orienta en qué grado el jugador no cometió error en las jugadas a la defensiva en las cuales intervino.

$$F\,\% = \frac{Outs \text{ realizados} + Asistencias}{Outs \text{ realizados} + Asistencias + \text{Errores cometidos}}$$

FR

Propuesto por Bill James, el estadístico FR parte de la premisa de que el número total de *outs* en los que participa un jugador en una posición determinada es un indicador defensivo más efectivo que el porcentaje de fildeo (F %). Sin embargo, cabe señalar que algunas posiciones (especialmente la primera base) pueden acumular una mayor cantidad de *outs* realizados y de asistencias (sobre todo debido a jugadas de *doblepay*) lo que le permite conseguir valores mayores de FR.

$$FR = \frac{Outs \text{ realizados} + Asistencias}{\text{Entradas jugadas en la posición}}$$

UZR

El UZR es una estadístico defensivo avanzado mediante el cual se mide la contribución en de un jugador sobre la base de las carreras evitadas, en una determinada posición, por encima o por debajo de otro jugador promedio en su misma posición. En el caso del UZR, los eventos que se toman en cuenta son los siguientes:

- Convertir la jugada en *out*.
- Permitir que un batazo se convierta en sencillo.

- Realizar un error que permita que un jugador alcance una base.

Por ejemplo, si un jardinero central tiene un UZR de igual a cero, su contribución es neutra en comparación con los demás jugadores en el jardín central, pero si este jugador cuenta con un UZR positivo esto implica que el jugador ha contribuido en salvar más carreras que el jugador promedio en su posición. Lo contrario sucederá en el caso de que el UZR sea negativo.

En Fangraphs (2016) se define la estructura principal para el cálculo del UZR, la cual se basa en la suma de la totalidad de eventos en que participa un defensor, multiplicada por el valor positivo (en caso de que realice un *out*) o negativo (en caso de que permita un sencillo o permita que un jugador consiga una base por error), en comparación con la cantidad de veces que una jugada similar (en términos de locación, velocidad y tipo de pelota bateada) es hecha por un jugador promedio en determinada posición del campo durante varios años.

TZL

Igual a como se lee el UZR, el resultado se obtiene a partir de las carreras salvadas por encima de un jugador promedio. De igual modo, el TZL de un jugador promedio será igual a cero, uno por encima del promedio tendrá un TZL positivo, y uno que le cueste carreras a su equipo tendrá un valor negativo. La comparación es realizada por posiciones por lo que un campo corto y un jugador de tercera base no son comparables con esta métrica, sino que lo son en comparación a jugadores de su misma posición.

Su cálculo varía dependiendo de las especificidades de los datos brindados. En vista de que la fuente del Total Zone son los datos obtenidos de Retrosheet (**?**), este estadístico permite analizar la defensa de cualquier jugador en la historia del béisbol.

Anexo B

Descripción de las fuentes de datos históricos de la MLB utilizadas en esta tesis

En el presente apéndice se ofrece información detallada acerca de dos de las fuentes de datos históricos de béisbol más completas disponibles en la actualidad: los *game logs* de Retrosheet y la base de datos Lahman. Los datos de estas fuentes contienen un gran nivel de detalle, incluyendo estadísticas desde la temporada de 1871 de la MLB, pasando por los *box scores* de cada juego en particular, hasta los refinados *play-by-play*, de la mayoría de los juegos celebrados desde 1945. Es importante comprender la estructura particular de cada una de estas fuentes de datos, para poder entender y utilizar de forma correcta la información de ofrecen.

B.1. Los *game logs* de Retrosheet

La organización Retrosheet fue fundada en 1989 con el propósito de colectar la información detallada *play-by-play* cada juego celebrado en la historia de la MLB. El sitio web[1] de Retrosheet ofrece *game logs* desde 1871. Un *game log* contiende detalles sobre varios aspectos del juego, esto incluye detalles sobre dónde se celebró el juego, cuántos espectadores acudieron, los equipos que se enfrentaron, el resultado final de juego etc. Además, los ficheros *game logs* incluyen estadísticos de los equipos tanto ofensivos como defensivos, así como resultados de jugadores, directores de equipo y árbitros. La Tabla B.1 ofrece una descripción de los 161 campos que se registran para cada juego.

[1] http://retrosheet.org/gamelogs/index.html

Campo(s)	Descripción
1	Feha de la forma «yyyymmdd».
2	Número del juego.
3	Día de la semana.
4 a 5	Liga del equipo visitador.
6	Número del juego del equipo visitador.
7 a 8	Liga del equipo local.
9	Número del juego del equipo local.
10 a 11	Carreras del equipo local y el visitador.
12	Duración del juego en outs. Un juego completo de nueve entradas deberá tener el número 54 en este campo. Si el equipo local gana sin batear en la parte baja de la novena entrada este campo será 51.
13	Día o noche de realización del partido («D» o «N»).
14	Información sobre la culminación del juego. Indica si el juego fue completado en la fecha prevista o en una fecha posterior (debido a una suspesión o por protesta).
15	Información sobre penalización.
16	Información sobre protesta.
17	Identificador del parque.
18	Asistencia del público al encuentro.
19	Tiempo de juego en minutos.
20-21	Carreras del equipo local del visitador en cada una de las entradas. Por ejemplo: «010000(10)0x» indica un juego donde el equipo local anotó una carrera en la segunda entrada, 10 carreras en la séptima y no bateó en el final de la novena entrada.
22-38	Estadísticos ofensivos del equipo visitador (en este orden): turnos al bate, sencillos, dobles, triples, cuadrangulares, RBI, sencillos de sacrificios, elevados de sacrificio, pelotazos, bases por bolas, bases por bolas intencionales, ponches, bases robadas, cojido robando, doble plays, interferencia del receptor, hombres dejados en base.
39-43	Estadísticos de picheo del equipo visitador (en este orden): lanzadores utilizados (1 significa que fue un juego completo), carreras limpias permitidas, carreras limpias del equipo, wild pitch, balks.
44-49	Estadísticos defensivos del equipo visitador (en este orden): puestos out, asistencias, errores, passed balls, doble plays, triple plays.
50-66	Estadísticos ofensivos del equipo local (igual a los del visitador).
67-71	Estadísticos de picheo del equipo local (igual a los del visitador).
72-77	Estadísticos ofensivos del equipo local (igual a los del visitador).
78-79	Identificador y nombre del árbitro principal del encuentro.
80-81	Identificador y nombre del árbitro de primera base.
82-83	Identificador y nombre del árbitro de segunda base.
84-85	Identificador y nombre del árbitro de tercera base.
86-87	Identificador y nombre del árbitro del jardín izquierdo.
88-89	Identificador y nombre del árbitro del jardín derecho.
90-91	Identificador y nombre del director del equipo visitador.
92-93	Identificador y nombre del director del equipo local
94-95	Identificador y nombre del lanzador ganador del encuentro.
96-97	Identificador y nombre del lanzador perdedor del encuentro.
98-99	Identificador y nombre del lanzador salvador del encuentro. Si no hubo juego salvado este campo recibe «none».
100-101	Identificador y nombre del bateador con mayor número de carreras impulsadas. Si hubo empate este campo recibe «none».
102-103	Identificador y nombre del lanzador abridor del equipo visitador.
104-105	Identificador y nombre del lanzador abridor del equipo local.
106-132	Identificadores de los jugadores abridores del equipo visitador, nombre y posición defensiva, listados del 1 al 9 tal como aparecen en el orden al bate del equipo.
133-159	Identificadores de los jugadores abridores del equipo local nombre y posición defensiva, listados del 1 al 9 tal como aparecen en el orden al bate del equipo.
160	Información adicional del encuentro.
161	Información sobre adquisiciones en el encuentro.

Tabla B.1: Sumario de los campos *game logs* de Retrosheet.

B.2. La base de datos de Lahman

Sean Lahman, quien es un activo periodista de béisbol y además autor de varios libros sobre este deporte, ha compilado y hecho accesible públicamente a través de su sitio web[2] una de las más completas bases de datos de estadísticas de este deporte disponibles actualmente. La base de datos Lahman, como se conoce, ofrece información para cada temporada de una gran cantidad de estadísticos. Estos incluyen el picheo, defensa, bateo de todos los jugadores de la MLB desde la primera liga profesional en 1891 hasta nuestro días. Además, esta base de datos almacena varias tablas con información suplementaria tales como detalles sobre resultados de juegos Todos-Estrellas, votos para el Salón de la Fama, estadísticos de los directores de equipo y de los jugadores en el post-temporada etc. Estos datos se encuentran disponibles en varios formatos: base de datos SQL, ficheros CSV y recientemente también como un paquete de `R`. La Tabla B.2 muestra una descripción de cada tabla correspondiente a la versión disponible en formato SQL.

Tabla	Descripción
AllStarFull	Apariciones de jugadores en juegos de Todos-Estrellas.
Appearances	Apariciones de jugadores por sus posiciones.
AwardsManagers	Directores galardonados con la condición de Mejor Director del Año.
AwardsPlayers	Jugadores galardonados con varias distinciones.
AwardsShareManagers	Resultados de los votos para la selección del Mejor Director del Año.
AwardsSharePlayers	Resultados de los votos de los jugadores galardonados con varias distinciones.
Batting	Estadísticos de bateo de la temporada regular.
BattingPost	Estadísticos de bateo de la post-temporada.
Fielding	Estadísticos de fildeo de la temporada regular.
FieldingOF	Apariciones por temporada de las tres posiciones de outfield.
FieldingPost	Estadísticos de fildeo de la post-temporada.
HallOfFame	Resultados de los votos para los premios del Salón de la Fama.
Managers	Datos de los directores de equipo por temporada.
ManagersHalf	Datos de los directores divididos por temporada.
Master	Información biográfica adicional de los individuos de la base de datos.
Pitching	Estadísticos de picheo de la temporada regular.
PitchingPost	Estadísticos de picheo de la post-temporada.
Salaries	Salarios de los jugadores por temporada.
Schools	Listado de equipos escolares.
SchoolsPlayers	Listado de las escuelas atendidas por jugadores.
SeriesPost	Resultados de las series de juegos de post-temporada.
Teams	Estadísticas de los equipos por temporada.
TeamsFranchises	Franquicias de los equipos.
TeamsHalf	Estadísticas de los equipos divididos por temporada.

Tabla B.2: Descripción de las tablas de la base de datos Lahman.

[2] http://seanlahaman.com