COMILLAS
UNIVERSIDAD PONTIFICIA

Facultad Ciencias Económicas y Empresariales

ORGANIZACIONES DEPORTIVAS Y EL ANÁLISIS DE DATOS

Autor: David Alba Burbano de Lara
Directora: María Jesús Jiménez Abad

MADRID | Junio 2020

Resumen

Este trabajo de fin de grado intenta mostrar cómo el análisis de datos, tan común hoy en día en todo tipo de negocios, puede aportar grandes beneficios a organizaciones deportivas, las cuales cuentan con la peculiar interdependencia de operaciones de negocio y rendimiento deportivo. A través del comentario de ejemplos, tanto en operaciones de negocio como deportivas en diversas organizaciones, y la introducción de conceptos básicos del análisis de datos se muestra aplicación de la ciencia de datos a este tipo de organizaciones. Además, gracias a la inspiración de artículos científicos y diversos ejemplos online, en este trabajo de fin de grado se realizan algunos experimentos replicables con datos de fútbol reales que demuestran el potencial del análisis de datos en el deporte.

Palabras clave

Análisis de datos, organizaciones deportivas, aprendizaje automático, márketing deportivo, rendimiento deportivo, puntos FIFA

Abstract

This final degree project tries to show how data analytics, common today in all types of businesses, can bring great benefits to sports organizations, which have the peculiar interdependence of business operations and sports performance. Through the discussion of examples, both in business and sports operations from various organizations, and the introduction of basic concepts of data analytics, this project shows the application of data science to this type of organizations. Furthermore, thanks to the inspiration of scientific articles and various online examples, some replicable experiments with real soccer data were carried out to show the potential of data analytics in sports.

Key Words

Data analytics, sports organizations, machine learning, sports marketing, sports performance, FIFA ratings

Índice

1. INTRODUCCIÓN

1.1 Propósito general

El presente Trabajo de Fin de Grado pretende demostrar el potencial que existe en el análisis de datos en organizaciones deportivas y hacer un caso práctico en equipos de fútbol. El análisis de datos tendría el objetivo de facilitar el proceso de toma de decisiones empresariales, tanto en el ámbito de marketing como en el ámbito deportivo. Al igual que en organizaciones empresariales más convencionales, con el uso del análisis de datos en marketing se pretende simplificar la toma de decisiones en procesos como ofertas promocionales, campañas publicitarias, eventos, etc. En el ámbito deportivo, el análisis de datos se propone como una herramienta que ayude al equipo técnico de la organización a tomar decisiones dentro y fuera del campo, desde qué jugadores fichar hasta qué jugadores poner titulares en partidos.

1.2 Justificación del tema

En los últimos años se ha puesto de moda en boca de empresarios el *business analytics*, pero realmente no es algo nuevo ni revolucionario. Existen algoritmos de análisis, predicción, clasificación e incluso de aprendizaje automático desde hace varias décadas, pero debido a la inmensa cantidad de datos que existen hoy en día, cada vez es más común hablar del análisis masivo de datos para la toma de decisiones en el entorno empresarial. Dado que las organizaciones deportivas tienen un ánimo de lucro como cualquier negocio, no hay razón para que no utilicen las mismas técnicas en su actividad empresarial.

El uso de la estadística por parte del equipo técnico de los equipos deportivos es cada vez más común, en el caso del béisbol se lleva haciendo desde hace más de 50 años, e incluso recibió una denominación propia, *sabermetrics*. El potencial del *sabermetrics* quedó claro cuando el equipo de béisbol Oakland Athletics acabó la temporada de 2002 con 103 victorias y 59 derrotas a pesar de haber perdido sus tres mejores jugadores antes de comenzar la temporada; al año siguiente todos los grandes equipos de béisbol contrataron analistas de *sabermetrics.*

1.3 Objetivos del trabajo

Durante este Trabajo de Fin de Grado se pretende revisar cuales son las prácticas de análisis de datos que se utilizan hoy en día en las organizaciones deportivas en el ámbito de márketing y en el ámbito deportivo, detallar casos concretos y discutirlos.

1.4 Metodología

La aproximación al tema la basé en tres fases consecutivas, pero no totalmente independientes: en la primera fase realizo una revisión del estado del tema recopilando bibliografía genérica sobre análisis de datos en organizaciones deportivas, en la que constato que el número de publicaciones es mucho mayor al esperado, incluyendo libros monográficos, libros con recopilación de contribuciones de expertos en temas diversos y actas de congresos. En la segunda fase realizo un cribado para seleccionar las publicaciones de mayor interés para el TFG y que además yo pudiera abarcar tanto en extensión como en comprensión de las técnicas utilizadas para el análisis de datos. En esta fase observo que la bibliografía referida al análisis de datos con objetivos de marketing es más genérica y no entra en detalles técnicos, quizás por ser demasiado específico para cada organización o por no querer revelar estrategias empresariales, sin embargo, en el análisis de datos deportivos la bibliografía es abrumadora, tanto en cantidad como en profundidad técnica. Es cuando decido enfocar la parte más técnica de este TFG sólo al caso del fútbol de la que aún hay una enorme cantidad de estudios. En esta fase ya anoto y comento artículos específicos y diseño el discurso y la sucesión de secciones de interés para construir un trabajo lo más contenido posible y de interés para futuros lectores. En la última fase abordo un caso práctico para ejemplificar cómo se trabaja con los datos hoy en día, partiendo de registros concretos de ligas de futbol a disposición pública y de técnicas estudiadas tanto durante las asignaturas de la carrera como durante la elaboración del propio TFG.

1.5 Estructura

Los siguientes capítulos de este Trabajo de Fin de Grado se estructuran de la siguiente manera. En el capítulo 2 hago un repaso del análisis de datos aplicado a la toma de decisiones en entidades deportivas en general, diferenciando entre la toma de decisiones relacionadas con el márketing de organizaciones deportivas y la toma de decisiones relacionadas con el rendimiento deportivo. En el capítulo 3 comento las técnicas de análisis de datos y modelado más generales poniendo ejemplos a partir de un trabajo con datos de jugadores de fútbol. En el capítulo 4 bajo al terreno práctico y a partir de datos disponibles entreno y pruebo mis propios modelos de predicción, inspirado por un artículo que también comento en este capítulo. En el capítulo final hago un repaso de las aportaciones de este trabajo y detallo posibles vías de continuación.

2. EL ANÁLISIS DE DATOS EN ORGANIZACIONES DEPORTIVAS

2.1. Introducción al análisis de datos

Se puede llamar análisis de datos a cualquier proceso que conlleve el examen o procesado de una serie de datos, independientemente de cuál sea la naturaleza de estos. Por ejemplo, el análisis de datos puede referirse a la inspección de ventas de una empresa a lo largo de los meses, la representación de porcentajes de votantes de cierto partido en cada provincia o los cálculos de probabilidades de precipitaciones que ofrece un pronóstico meteorológico. El análisis de datos es una ciencia presente en cualquier rama del conocimiento humano y hoy en día es común que se hable del análisis de datos en el entorno empresarial. Los avances tecnológicos están acelerando y abaratando los procesos de obtención, almacenamiento y procesamiento de ingentes cantidades de datos, lo cual lleva a las organizaciones a utilizar el análisis de datos con mayor frecuencia. Estos avances también causan que hoy en día ya no se hable de datos como antes, ahora se habla del big data, datos masivos o macrodatos. Según el glosario de términos de la consultora Gartner, el big data se puede definir como aquellos datos que tengan las tres v: volumen, velocidad y/o variedad; es decir, activos de información en gran volumen, alta velocidad de generación y/o gran variedad de fuentes y formatos que exigen métodos rentables e innovadores de procesamiento de la información que permitan mejorar la comprensión, la toma de decisiones y la automatización de los procesos.

El tipo y fuente de datos a analizar por parte de las empresas puede ser muy variado, desde datos financieros hasta datos de logística o de marketing. A pesar de que el objetivo final es claramente diferente, los análisis pueden llegar a ser muy parecidos en muchas ocasiones por mucho que la información en cuestión no tenga nada que ver. Es, pues, importante mencionar que las técnicas y métodos utilizados en esta ciencia son aplicables en muchos escenarios diferentes.

El objetivo principal del análisis de datos en una organización deportiva (como en cualquier otra) es convertir los datos sin procesar en información significativa, de valor agregado y procesable que permita tomar decisiones comerciales estratégicas, lo que luego resultará en un mejor desempeño financiero de la compañía y una ventaja competitiva medible y sostenible. Un sistema efectivo de análisis de datos debería dar como resultado mayores ingresos incrementales, costes reducidos, riesgos administrados, una utilización más efectiva de los recursos humanos (es decir, "análisis de talento"), desarrollo optimizado de productos y servicios (por ejemplo, innovación basada en datos), mejora de la

comercialización y el servicio al cliente (por ejemplo, una mayor participación de los consumidores (los fans) en la organización deportiva) y, en general, tomas de decisiones estratégicas más informadas. (Bukstein and Harrison, 2017). En un estudio empírico reciente sobre la adopción de análisis de datos por parte de organizaciones deportivas, el uso de este tipo de herramientas se vio correlacionado con un crecimiento de ingresos del 7.2% en el año posterior a la adopción de una estrategia de análisis de datos, lo cual supone un incremento en comparación con las expectativas generales de la industria de negocios deportivos que contaba con una previsión de aumentos anuales de ingresos del 3% (Trolio et al., 2016).

2.2 Análisis de datos en marketing de organizaciones deportivas

En esta sección vamos a reflejar dos áreas de actividad a nivel de márketing que pueden ayudar a incrementar los beneficios de la compañía con un uso adecuado de los datos disponibles. La primera de las áreas es la de retorno más directo: la venta de entradas y abonos. La segunda tiene un retorno más indirecto ya que tiene que ver con la fidelización de los aficionados.

2.2.1 Maximizar ingresos por venta de entradas y abonos

Es evidente que para toda organización deportiva una de las muchas preocupaciones de márketing es la venta de entradas a los eventos. Ya sea a través de entradas para partidos individuales, abonos de temporada o carnés de socio, esta es una de las áreas en las que un buen pronóstico de ventas es importante para las organizaciones deportivas a la hora de planear sus presupuestos. Dependiendo de la organización deportiva que estemos hablando, y también del deporte, los ingresos de este ámbito pueden representar un peso muy diferente en la contabilidad. Por ejemplo, en el 2018 el FC Barcelona ingresó más de 50 millones de euros por abonados y socios, frente al Sevilla FC que ingresó 14 millones o al RC Celta de Vigo que no llegó a los 5 millones (Menchén, 2018). A pesar de que el FC Barcelona sea el club que más ingrese en España por esta partida, para el ejercicio económico de 2018 esto representó solo el 5% de ingresos de la organización, en cambio para otros clubes puede que sea más importante, como en el caso del Sevilla FC, para el cual los abonados y socios representaron más del 10% de ingresos ese mismo año[1].

[1] Porcentajes de ingresos calculados con las cuentas oficiales de las organizaciones. Consultadas en https://www.fcbarcelona.es/es/club/organizacion-y-plan-estrategico/comisiones-y-organos/reportes-anuales y https://www.sevillafc.es/el-club/la-entidad/ley-de-transparencia

Toda organización deportiva desea ver que sus estadios mantienen una alta asistencia de manera frecuente, ya que esto se traduce a un alto volumen de ventas de entradas, abonos, etc. Para llegar a alcanzar esto el departamento de márketing de las organizaciones tiene que realizar una laboriosa tarea, ya que, a diferencia de otro tipo de negocios, las organizaciones deportivas ofrecen un servicio muy complejo que se basa en la actuación de los deportistas y por lo tanto la satisfacción del consumidor puede variar radicalmente de evento a evento. El análisis de datos es una herramienta que puede ayudar a los departamentos de márketing con esta tarea ya que permite a las organizaciones crear nuevas estrategias, innovar y ofrecer ofertas para atraer a consumidores a los estadios.

Un claro ejemplo de cómo el análisis de datos puede suponer una gran ayuda para los departamentos de márketing a la hora de realizar ventas es el caso de los Orlando Magic en la NBA, el cual Bukstein y Harrison explican en gran detalle en su libro *Sports Business Analytics*. En 2012 la organización encontró que una parte significativa de sus ventas eran realizadas a consumidores casuales, gente que vivía cerca pero que no atendía a partidos regularmente durante la temporada o incluso que no acudía a diferentes temporadas. Orlando Magic hizo un estudio sobre este segmento de consumidores, denominado "fans casuales", donde determinó que:

- Estos consumidores eran muy sensibles en cuanto a precio, considerando que era una barrera para atender a más partidos
- Los fans casuales acostumbraban a buscar descuentos para poder ir a partidos
- No se consideraban grandes fans del equipo, pero tenían interés por asistir a ver partidos en directo
- Estos consumidores eran mucho más propensos a incurrir en más gastos en los estadios, ya sea porque solían acudir acompañados o porque realizaban compras una vez allí
- A pesar de tener interés en el baloncesto no solían acudir a partidos hasta mediados de temporada, mostrando poco interés en los partidos de pretemporada o en los primeros meses de temporada oficial.

Una vez la organización observó más en detalle el perfil de estos consumidores, el equipo de márketing, a través del análisis de datos, encontró una solución que garantizaría un aumento de asistencia para este segmento de consumidores. A partir de 2014 Orlando Magic empezó a ofrecer un pase especial para residentes de Florida en el que los consumidores tendrían derecho a entradas para los primeros ocho u once partidos de la temporada. Lo denominaron *Fast Fall Break Pass*. El atractivo de estos pases es el precio, 49$ o 79$, el mismo precio que una entrada individual en los primeros meses de la temporada. Los pases vienen sin asignación de asientos predeterminada, de manera que en cada partido se asigna uno de los asientos que queden libres en el estadio. Este pase supuso una manera de atraer a los "fans casuales" ya que ofrecía una manera increíblemente económica para asistir a varios partidos, pero además suponía una solución para los Orlando Magic ya que aumentaba la asistencia a principios de temporada, pero no ponía en peligro las ventas de entradas individuales ya que ningún asiento estaba asignado de manera predeterminada.

El *Fast Fall Break Pass* fue todo un éxito. En la primera temporada en la que se implantó se vendieron más de 1500 pases a 700 clientes diferentes, 80% de los cuales fueron clientes completamente nuevos para la organización y 20% clientes que recientemente habían acudido a algún partido. Además, los pases fueron utilizados aproximadamente un 70% de las veces posibles. El objetivo de atraer a más "fans casuales" de manera recurrente y aumentar la asistencia durante el principio de la temporada fue completado gracias a la ayuda de Experience, una compañía de soluciones basadas en el análisis de datos especializada en eventos deportivos. Desde entonces Experience ha realizado más proyectos similares con otras organizaciones deportivas, como Sacramento Kings de la NBA, DC United de la MLS, Atlanta Braves de la MBL, etc.[2]

Esta nueva estrategia tuvo tan buenos resultados para los Orlando Magic que al año siguiente ofrecieron más opciones para el Fast Break Pass:

- Season Pass: incluye un total de 34 partidos durante toda la temporada
- Half-Season Pass: incluye partidos desde mediados de diciembre en adelante
- Week-day Pass: incluye partidos solo entre lunes y jueves
- Week-end Pass: incluye partidos solo entre viernes y sábado
- Fall Plus Pass: incluye los 14 partidos durante octubre y noviembre
- Autumn Pass: incluye 12 partidos durante octubre y noviembre
- All-Star Pass: incluye determinados partidos durante enero y principios de febrero
- Slam Dunk Pass: Incluye partidos a fines de febrero y marzo
- Buzzer Beater Pass: pases de un solo partido para partidos seleccionados

Afortunadamente la mayoría de los consumidores que adquirieron estos pases (unos 2000 clientes en la temporada 2015-2016) continuaban siendo clientes nuevos para los Orlando Magic o fans casuales cuyo compromiso con la organización deportiva había aumentado debido a la creación de estas ofertas. Además de esto, muchos fans casuales después de incrementar su compromiso con la organización a través de los Fast Break Pass durante la temporada se convirtieron en grandes seguidores de los Magic, lo cual se tradujo en oportunidades para vender más abonos de temporada (los cuales suponían ingresos mayores que los Fast Break Pass) en futuras campañas.

A raíz del gran éxito con Experience y los Fast Break Pass los Orlando Magic decidieron buscar nuevas maneras de innovar la organización. La app de los Magic es la plataforma a través de la cual se gestionan los Fast Break Pass y la organización quería que los "Loyal Blue", aquellos fans que llevaban varias temporadas siguiendo a los Magic, también tuvieran nuevos beneficios. Después de estudiar este segmento la organización llegó a la conclusión que el mayor problema era el difícil e intensivo horario que tiene la NBA, el cual suponía conflictos personales y/o profesionales para los fans y la reventa de entradas no era una solución válida en la gran mayoría de los casos. A través de un modelado predictivo los Magic encontraron que la renovación de abonos de temporada estaba influenciada sustancialmente por el porcentaje de uso del abono a través de la temporada; de manera que, si un fan se perdía varios partidos, pero además no conseguía revender sus entradas

[2] Información obtenida en la página oficial de Experience: https://www.expapp.com/

sentía una gran pérdida y sería menos propenso a renovar su abono que aquel fan que, a pesar de tener que perderse partidos, conseguía revender sus entradas. De manear que los Magic decidieron crear, a través de la plataforma app, la opción de Not Going, la cual permitía a los dueños de abonos de temporada recibir una compensación económica en una divisa virtual si avisaban a la organización de que no iban a poder asistir a determinado partido. Esta solución supuso que la organización pudiera revender esos asientos u otorgarlos a gente con Fast Break Pass, y para los fans supuso que no perderían un partido ya que la compensación que recibían, el Magic money, podían gastarlo en otros partidos para recibir asientos mejores o para realizar consumiciones durante el partido, entre otras opciones. En la primera temporada de implantación un 50% de los abonos de temporada utilizaron la opción Not Going en algún momento, y en la siguiente temporada aumentaron hasta un 80%.

Los Orlando Magic fueron capaces de ofrecer nuevos servicios para la gran mayoría de su clientela, para los fans casuales crearon los Fast Break Pass y para los Loyal Blues crearon el Not going. Ambas innovaciones se tradujeron en mayores ventas para la organización y mayor satisfacción para los fans; ambas fueron realizadas gracias a la información obtenida a través del análisis de datos, que supuso una ayuda clave para que el departamento de márketing llegara a estas ingeniosas soluciones.

2.2.2 Ingresos indirectos generados gracias a la analítica avanzada

¿Cómo puede un club aprovechar al máximo el interés de los fans para promocionarse y generar interés por consumir productos o servicios que lleven su marca, desde abonos, hasta entradas de eventos promocionados, pasando por merchandising del propio club e interacción on-line con sus apps? Una analista de la industria de entretenimiento de IBM, Stacy Nawrocki (2016), comenta cinco vías para sacar provecho comercial de la analítica avanzada de datos deportivos, siendo las cuatro primeras relacionadas directamente con la atracción de aficionados y la última relacionada con otro campo de aplicación de la analítica de datos, de máximo interés todo deporte: la forma física e integridad de los jugadores.

1. Atraer nuevos fans y mantener a los que hay. A través del análisis de datos se puede predecir las preferencias de los fans y utilizar segmentación demográfica fina para hacer promociones casi personalizadas. Hay múltiples fuentes de datos que se pueden aprovechar para segmentar a la base de fans de un club: por la venta de entradas y abonos, por el resultado de campañas de marketing, por el análisis del CRM, por los comentarios en las redes sociales, etc. Además, se pueden usar técnicas similares a las que usan las aplicaciones de recomendación para buscar nuevos fans que comparten un perfil demográfico con fans ya existentes. Un ejemplo claro del uso de estas fuentes de datos para maximizar el ingreso por abonos es el de los Orlando Magic de la NBA, comentado en la sección previa.

2. Interactuar en las redes sociales y empoderar a los fans. Los fans que crecieron con las redes sociales, la interacción en tiempo-real y los videojuegos de deportes hiper-realistas, no se conforman con ver y disfrutar su equipo favorito. Necesitan ser parte activa mediante sus comentarios, su consumo de datos, acceso a jugadores y entrenadores y, en definitiva, sentir que forman parte del club. Las organizaciones están empezando a proporcionar apps con contenidos para "la segunda pantalla"[3] y capacidad interactiva, de tal manera que mientras el aficionado disfruta, el club recoge datos para captar las emociones (sentiment analysis[4]), preferencias y comentarios. Un buen diseño y gestión de estos canales aumenta la sensación de pertenencia y poder de los aficionados e indirectamente (a través influencers) aumentar el consumo de productos del club y atraer nuevos fans
3. Mejora de la experiencia en el estadio. Es una extensión de la medida anterior aplicada a la experiencia de asistir a un encuentro deportivo. Hoy en día todo el mundo lleva un smartphone en el bolsillo que, con la aplicación adecuada proporcionada por el club, puede facilitar todo el proceso de aparcar, buscar su asiento, pedir snacks online, etc.; e incrementar el entretenimiento mediante información deportiva en tiempo real. Una mejor experiencia influye enormemente en la asistencia y el gasto total de la misma. Un caso de uso impresionante de la tecnología de captura y análisis de datos en deportes con fines de entretenimiento es el que puso en funcionamiento IBM con su motor de Inteligencia Artificial Watson en el US Open de 2019. Con un sistema entrenado para detectar automáticamente el inicio y final de cada punto, los momentos de euforia del público y hasta la expresión facial de los jugadores, el sistema producía en tiempo real un resumen de las jugadas más interesantes (Suzor, 2019).

4. Convertir a los fans en comentaristas y ofrecerles múltiples datos del encuentro al instante. Esta vía es la combinación sinérgica de las dos anteriores. Mediante el acceso a datos estadísticos de los jugadores y datos captados y procesados in situ con las tecnologías mencionadas más adelante en la sección 2.3, el aficionado tiene en sus manos toda la información que usan los comentaristas. Sólo hay que abrirle un canal de comunicación corporativo para que se convierta en emisor y/o receptor de comentarios hacia y desde otros aficionados, con lo que el club lleva a su máximo exponente la experiencia del aficionado y la realimentación para sus propios intereses, tanto de marketing como de estrategias deportivas alternativas.

5. Mantener a los jugadores sanos...y en el campo. La medicina deportiva es todo un campo de trabajo en el que los clubes invierten bastantes recursos, tanto en la rehabilitación temprana y correcta de sus jugadores como en la prevención de las

[3] Se denomina "segunda pantalla" porque ofrece acceso a información e interactividad mientras el evento deportivo se ve en la "primera pantalla", ya sea una televisión, ordenador o incluso en directo.
[4] La interpretación y clasificación de las emociones (positivas, negativas y neutrales) dentro de los datos de texto utilizando técnicas de análisis.

lesiones. Hoy en día es posible obtener información precisa y continua de variables biométricas de la salud del jugador tanto en el gimnasio, como en la sala de rehabilitación, como en el terreno de juego mediante dispositivos wearable. Dispositivos que monitorizan variables biológicas como el ritmo cardíaco, la sudoración, el nivel de oxígeno en sangre, así como variables cinemáticas como la posición, velocidad y aceleración, permiten contrastar en cualquier momento el estado físico y de sobre esfuerzo con históricos del propio jugador y prevenir problemas musculares y de ligamentos.

2.3 Análisis de datos en la estrategia deportiva

2.3.1 Los inicios de la analítica de variables deportivas

El béisbol en Norteamérica utiliza datos estadísticos sobre jugadores para tomar decisiones sobre fichajes y estrategias de juego desde finales del siglo XX. A mediados de los 70, Bill James, un escritor especializado en beisbol bautizó al análisis empírico de las estadísticas del béisbol con el nombre de *sabermetrics.* En la temporada 2002 el equipo de béisbol norteamericano Oakland Athletics apostó fuerte por la sabermetría con un análisis de datos estadísticos de jugadores más pormenorizado y de manera científica, contratando a jugadores subvalorados por otros equipos pero que formaban un puzle perfecto para los Oakland Athletics (Cummings, 2019). Este caso fue reflejado en la novela de Michael Lewis en 2003, *Moneyball: the art of winning an unfair game*, y llevada en el 2011 a la gran pantalla con el nombre *Moneyball.* En el libro, Lewis cita al director general del equipo, Billy Beane: "*No había una forma sencilla de abordar el problema que estaba tratando de resolver: tienes 40 millones de dólares para gastar en 25 jugadores de béisbol. Tu oponente ya ha gastado 126 millones de dólares en sus propios veinticinco jugadores, y tiene quizás otros 100 millones de dólares en reserva. ¿Qué es lo que tienes que hacer con tus cuarenta millones para evitar una derrota humillante? Lo que no haces es lo mismo que los Yankees. Si lo hacemos, perdemos, porque ellos lo hacen con tres veces más dinero que nosotros. Un equipo pobre no puede permitirse el lujo de salir a comprar estrellas de las grandes ligas en la cima de sus carreras. Ni siquiera puede permitirse el lujo de salir a comprar jugadores de precio medio*" (Cortsen, 2018). La teoría de su analista de datos, que extrajo del estudio de miles de estadísticas, fue que un equipo con un alto porcentaje de alcanzar bases era un equipo más propenso a anotar carreras[5] y, como resultado, más propenso a ganar partidos. El equipo directivo reclutaba y cambiaba jugadores que se ajustaban a este sistema, y sólo esos jugadores. El resultado inmediato fue que Oakland Athletics se convirtió en un equipo

[5] En béisbol, una carrera es una anotación, análogo a un gol en el fútbol.

que sacaba más carreras que ponches[6]. Visualmente, el equipo también era diferente. El sistema de sabermétría de los Oakland no requería que los jugadores se ajustaran a los prototipos de altura, peso, velocidad o composición corporal que dictaban los movimientos de otros clubes. Con esta estrategia Oakland Athletics ganó 20 partidos consecutivos igualando el récord ostentado por Chicago White Sox desde 1906. Este sistema de alta eficiencia y bajo coste no sólo ha revolucionado la era moderna del béisbol, sino también el deporte profesional en su conjunto. A partir de aquí los clubes deportivos, especialmente de béisbol, pero progresivamente en fútbol, baloncesto y otros deportes con gran interés comercial, empezaron a tomarse en serio el potencial de los datos objetivos y el modelado matemático.

Cuando nació el sabermetrics los datos sobre el juego y jugadores que se podían recopilar se reducían a eventos observables por el ojo humano, así que la creación de modelos estadísticos con capacidades predictivas tenía una componente grande de "know-how" de los propios ojeadores, entrenadores y técnicos, y la complejidad de los modelos no podía ser muy elevada. En los inicios del siglo XXI empezaron a aparecer los primeros sistemas que utilizaban datos posicionales con información espacio-temporal de los jugadores y la pelota en el terreno de juego, como los posicionadores basados en tecnología GPS, acelerómetros acoplados como pequeños dispositivos en cinturones, camisetas y balones, y cámaras de video para observar cada pequeño detalle. El incremento en número y calidad de datos hace ahora posible entrenar modelos de aprendizaje automático, como los árboles de decisión y modelos de regresión no lineal, para así tener una descripción cualitativa y cuantitativa del rendimiento de un jugador en una posición concreta y tomar una decisión más objetiva sobre el interés de incorporarlo a un equipo concreto en un momento determinado (Liu, 2019). Con esta información el equipo de técnicos deportivos y entrenadores puede estudiar en detalle la estrategia de juego sin dejarse llevar exclusivamente por su propia experiencia y costumbres, o analizar si un jugador debe ser cedido, contratado o despedido.

2.3.2 El estado de la técnica y los servicios en la nube

Hoy en día, gracias al avance de la tecnología de sensores de posición, tecnología radar, tecnología wi-fi y las cámaras de alta velocidad y precisión, se pueden capturar una ingente cantidad de datos por minuto; y gracias al avance en las técnicas de modelado usando aprendizaje automático se puede procesar toda esa cantidad de datos en un tiempo razonable y obtener un modelo descriptivo o predictivo capaz de encontrar relaciones entre las variables medidas que un ojeador, un entrenador o un técnico deportivo no sería capaz de encontrar por sí solo (Ganguly & Pedagagi, 2019). En el reciente libro de Ben Lindbergh y Travis Sawchik, *The MVP Machine: How Baseball's New Nonconformists Are Using Data to*

[6] Se produce cuando un bateador acumula tres fallos (strikes) durante su turno de bateo, lo cual conlleva la eliminación del bateador (out). Se corresponde con el extranjerismo strike-out.

Build Better Players los autores hacen una recopilación de las diferentes estrategias que los equipos de béisbol han empleado desde el boom de *MoneyBall*, en 2002, pasando del modelo de contratar jugadores subvalorados a un modelo de desarrollo personalizado de cada jugador. En la web *The Ringer* los propios autores escribieron un extracto de un capítulo sobre lo que están haciendo *Los Houston Astros* de la MBL, la principal liga de béisbol americana, con nuevas tecnologías para maximizar el potencial de los jugadores. Hoy en día los 30 estadios de béisbol de la liga MLB utilizan el sistema Statcast, que desde el año 2015 utiliza tecnología óptica y de radar para ofrecer tal nivel de detalle del movimiento de la bola y jugadores, que permite hacer cálculos, análisis y valoraciones impensables hace solo una década (Soler, 2017). Así, el papel de los ojeadores clásicos sigue vigente, pero se ha tecnificado para extender su capacidad predictiva más allá de la subjetividad con la ayuda de miles de variables medidas de cada jugador y los modelos matemáticos que cada equipo genera. Se estima que Statcast extrae 7 TeraBytes de datos en cada partido de béisbol, es por eso que Statcast utiliza los servicios de Amazon Web Services (AWS) para digerir toda esa información y destilar lo que importa a los clientes, en este caso, los clubes, casas de apuestas, freelanders y aficionados.

Servicios como AWS, Microsoft Azure y Google Cloud son servicios de cloud computing o informática en la nube, básicamente son servicios de distribución de recursos de tecnología de la información bajo demanda a través de Internet mediante un esquema de pago por uso. En vez de comprar, poseer y mantener servidores y centros de datos físicos, los clientes de un proveedor de cloud computing obtienen acceso a servicios tecnológicos, como capacidad informática, almacenamiento y bases de datos, en función de sus necesidades (AWS a., n.d.). Los servicios de cloud computing están haciendo viable que toda empresa, no solo de cualquier sector sino también de cualquier tamaño, forme parte de esta nueva revolución industrial del análisis de datos. La computación en la nube elimina el gasto que supone comprar hardware y software, suprime tiempos de configuración, evita que la empresa busque espacio físico para los racks de servidores, incurra en el gasto que supone la electricidad las 24 horas del día para energía y refrigeración, y además no genera la contratación de expertos de TI para administrar la infraestructura (Azure, n.d.). Estos y otros beneficios, como la seguridad y la fiabilidad de estos servicios, hacen posible que el análisis de datos llegue a todos los negocios ya que no es necesario tener recursos para almacenar y procesar datos, con el cloud computing todo ocurre en servidores remotos.

De hecho, es por esto por lo que cada vez más y más organizaciones deportivas están adentrándose en el mundo del análisis de datos, lo fácil que es y los beneficios que supone hace que sea una oportunidad que nadie puede renunciar. No sólo son organizaciones como equipos o clubes, sino que las propias empresas que dirigen y administran las competiciones deportivas se quieren unir a esta revolución. La NFL, la liga de fútbol americana, también utiliza AWS para transformar una liga deportiva que cuenta con más de 100 años (AWS b., n.d.). Según Michelle McKenna, vicepresidenta y Chief Information Officer de la NFL, "*Implementar análisis de datos beneficia a toda la organización y a la liga, no solo a los*

departamentos de tecnología. Con el análisis de datos podemos aumentar nuestra efectividad, por ejemplo, trabajos tediosos como el procesado y clasificación de vídeo ahora ya no son tareas manuales, sino que se realizan a través de algoritmos de aprendizaje automático. Ahora los entrenadores pueden usar la tecnología para plantear jugadas, formaciones y dibujar automáticamente sus ideas, ahorrándoles tiempo en el banquillo. Con el poder del análisis de datos en AWS, podemos entender mejor a los fans, plantear cómo se presentan y cómo se perciben los partidos, estudiar el impacto potencial que supondría un cambio en el reglamento, calibrar cómo se arbitran los partidos, y además, medir el rendimientos de los jugadores y garantizar su seguridad."

2.4 El análisis de datos deportivos en negocios alternativos

El análisis de datos deportivos tiene algunas otras derivadas al margen del propio interés empresarial de los clubes en mejorar el rendimiento de sus equipos para maximizar la probabilidad de ganar ligas.

Una de estas derivadas es el ecosistema de apuestas deportivas. Los mismos datos que se usan para crear un modelo predictivo que ayude a decidir un fichaje, se están usando para crear otro modelo predictivo que ayude a aumentar la probabilidad de ganar una apuesta sobre cualquier evento, desde algo tan específico como el número de goles del jugador A en una temporada, hasta la quiniela semanal. Por ejemplo, la compañía Smartodds funciona como un sindicato de apuestas y una agencia de consultoría de apuestas. Está compuesta por expertos en matemáticas e informática que hacen uso de algoritmos, estadísticas e investigación de datos para determinar qué equipos tienen más posibilidades de ganar un partido. Los servicios que ofrece su empresa ayudan a los apostadores que son clientes de Smartodds a decidir a qué equipo sería más prudente apostar[7]. Otro ejemplo es el del dueño del Brighton Football Club, Tony Bloom con su empresa Starlizard, similar a la anterior pero que además utiliza herramientas de análisis de datos y rastreo de las redes sociales para capturar el sentimiento de los aficionados a las apuestas e incluso de los jugadores para obtener información "privilegiada". En palabras de Tony Bloom, "*Nuestros sistemas de mensajería distribuidos manejan miles de actividades globales cada segundo, destacando las oportunidades de apuestas en el momento en que aparecen. Continuos flujos de eventos deportivos y de apuestas convergen en nuestra plataforma de datos, proporcionando un rico análisis y una profunda comprensión a nuestros equipos de investigación y de apuestas*" (Trademade Sports, n.d.).

[7] Información del artículo en https://onlinecasinomonsters.com/matthew-benham-successful-bloke-world-online-gambling/

Otra de las derivadas tiene que ver con la explotación de los datos por parte de terceros con fines informativos o de entretenimiento. La popularidad de la toma de decisiones impulsada por los datos en los deportes ha llegado hasta los aficionados, que están consumiendo más contenido analítico que nunca. Ahora hay sitios web enteros dedicados a la investigación y análisis de las estadísticas deportivas y cómo se relacionan con una predicción en el rendimiento de los jugadores. Un ejemplo es FiveThirtyEight.com, que fue creado en marzo de 2008 por Nate Silver. Silver, que proviene del mundo del análisis de béisbol, lanzó el sitio para proporcionar más detalles que la propia cobertura de béisbol. El sitio, que cuenta con una afiliación a la cadena de deportes ESPN, tiene a más de 20 periodistas contando y haciendo números para que los aficionados puedan entender mejor un partido, una serie o una temporada.

2.5 El caso del fútbol en España: MediaCoach

La primera división de fútbol en España siempre se ha considerado una de las mejores ligas de fútbol del mundo. Es evidente que dada la historia que ha tenido esta competición, además de la actuación de los clubes españoles en otras competiciones a nivel europeo, han generado una imagen de marca espectacular; dicho esto hay que destacar que a pesar de que exista tal ventaja competitiva y que tiene cierta durabilidad, ninguna ventaja competitiva es eterna (Charles and Melissa, 2016). Desde hace años la Liga tiene fans en todo el mundo, pero durante la última década la organización ha hecho hincapié en la internacionalización de su marca y ha apostado por la innovación tecnológica. Según el director del área deportiva de La Liga, Ricardo Resta, ya en 2009 la organización se empezó a plantear el desarrollo de una herramienta de alto nivel accesible a todos los clubes con el objetivo de ofrecer una ayuda para el aumento del rendimiento deportivo, así fue como nació Mediacoach (Olmeda, 2020).

En el apartado 2.3.2 de este TFG se mencionó a Statcast, el sistema que utiliza la liga de béisbol americana para analizar partidos de manera increíblemente precisa y que permite capturar una ingente cantidad de datos para poder mejorar los análisis sobre jugadores. Mediacoach es a la La Liga lo que Statcast a la MBL. Se puede resumir como un sistema de análisis de vídeo diseñada por y para profesionales del fútbol. Mediacoach es capaz de ofrecer este servicio ya que todos los estadios de primera y segunda división, LaLiga Santander y LaLiga Smartbank, cuentan con cámaras que siguen en todo momento los partidos de fútbol. Gracias a este sistema multicámara, Mediacoach hace posible la recolección, cuantificación y gestión de multitud de parámetros del juego, y además en tiempo real; de manera que los técnicos deportivos de los clubes son capaces de estudiar antes, durante y después de los partidos a los rivales y a su propio equipo. En palabras de Francisco Joaquín Pérez, director deportivo del RDC Espanyol: "*Mediacoach te da información muy fiable del rendimiento que puede dar un jugador. Cuando el contexto es el mismo, la calidad del dato es mayor. Nosotros, además, sumamos una parte cualitativa al*

dato cuantitativo. Desde hablar con el entrenador para ver si encaja en su sistema y estilo hasta ir a ver al futbolista en persona" (Sabatés & Martín, 2020).

Una de las grandes ventajas con las que cuenta Mediacoach y su sistema multicámaras es que está respaldado por la ciencia. Mediacoach ha sido sujeto de, al menos, tres estudios diferentes para comprobar si el sistema multicámaras utilizado para grabar partidos y recopilar datos es fiable para los clubes de fútbol. Los estudios fueron realizados por diversas entidades con la finalidad de comparar la calidad de este sistema con diferentes dispositivos GPS y con cámaras de alta precisión de captura de movimiento. El primer estudio fue realizado con datos de la temporada 2017-2018 de La Liga SmartBank por un grupo de investigadores en el cual se incluían miembros del FC Barcelona, LaLiga, la Universidad de Extremadura y la Universidad de Lisboa. Su objetivo era comparar el sistema de Mediacoach con los dispositivos GPS WIMU Pro de RealTrack, los cuales utilizaba el Barcelona B. La conclusión del estudio fue que el sistema era tan fiable como los otros dispositivos, con la ventaja de que el sistema Mediacoach es menos invasivo, lo cual supone un beneficio a la hora de realizar estudios y análisis; de manera que no sólo validaron Mediacoach, sino que lo recomendaron (Pons et al., 2019). El segundo estudio fue realizado también en la campaña 2017-2018 con datos de la segunda temporada, comparando la fiabilidad de Mediacoach contra el sistema GPS Apex de STATSports. Este estudio, llevado a cabo por académicos de la Universidad Europea de Madrid y la Universidad de Castilla-La Mancha, llegó a la misma conclusión que el organizado por el FC Barcelona, Mediacoach es tan fiable y válido como los sistemas GPS (Felipe et al., 2019). El tercer estudio fue desarrollado por investigadores de la Universidad Técnica de Múnich. En esta ocasión, el sistema multicámaras de Mediacoach fue comparado con las cámaras de alta precisión del sistema de captura de movimiento VICON, el cual es considerado el sistema de referencia en la comunidad científica para este tipo de mediciones. Al igual que en los otros dos casos, el veredicto para el sistema español fue positivo.

"*En la práctica, no solo hemos contribuido a mejorar la toma de decisiones en directo, durante la disputa de un partido. También en la gestión de las plantillas",* dice Ricardo Resta, director del Área Deportiva de LaLiga y director de Mediacoach, "*Al fomentar este conocimiento milimétrico y transparente para todos, hemos democratizado la competición y ayudado a retener talento. De que se construyen equipos cada vez más competitivos, grandes y pequeños, dan fe los resultados que obtienen cuando compiten fuera*" (Sabatés & Martín, 2020).

3. TÉCNICAS DE ANÁLISIS DE DATOS

En este capítulo vamos a hacer un repaso del proceso de análisis de datos, además de comentar las ramas principales de esta ciencia y algunas de las técnicas principales. Una buena comprensión de cuándo y cómo utilizar las técnicas y metodologías de modelado es clave para poder extraer el máximo partido a los datos disponibles y realizar las tomas de decisión más fundamentadas y con mayor garantía de éxito.

Un repaso rápido de todas las tareas que deben formar parte de una buena estrategia de análisis de datos nos deja la siguiente lista que, aunque no es exhaustiva, es común para la gran mayoría de casos.

- Preparación de los datos: la información relevante para poder predecir un resultado puede provenir de fuentes muy diversas, tanto de texto como numéricas. La información puede estar dispersa y desestructurada, como en tweets u opiniones en facebook, en paquetes de datos homogéneos como en la secuencia de movimientos captada por una pulsera con acelerómetro de un futbolista, o en estructuras de datos heterogéneos como las fichas de los jugadores a contratar. Los datos pueden estar completos, corruptos o incompletos; en rangos diversos y tipos categóricos, ordinales o numéricos. Muchos datos requerirán transformaciones a nuevas variables para poder acomodarlos a las técnicas posteriores. Un porcentaje muy significativo del tiempo dedicado al análisis se dedica a preparar bien los datos. Se habla más de este paso en la sección 3.1
- Visualización y exploración: con esta tarea tratamos de descubrir relaciones entre las variables, agrupaciones de éstas y datos erróneos (outliers) mediante la representación de gráficos, histogramas, métodos de análisis de no supervisado (explicados en la sección 3.2) y reducción de dimensionalidad[8]. Los resultados de este análisis pueden dar lugar una fase de selección y transformación de variables o simplemente permitir la elección adecuada de las técnicas de modelado posteriores. Así, supongamos un problema de seleccionar a un jugador como delantero para el equipo a partir de *N*. Si un clustering *N*-dimensional de P delanteros y defensas con k-means arroja que todos ellos se distribuyen en dos grupos compactos y bien separados (silhouette[9] $\cong$ 1), estaremos bastante seguros de que las variables elegidas para discriminar delanteros son informativas.

[8] Proceso de reducción del número de variables explicativas que se trate en un problema.
[9] Método de validación del número de clústeres, con valores de -1 a +1, donde un valor alto indica que los clústeres están bien separados.

- Selección y/o transformación de variables: en esta fase del proceso de análisis de datos el objetivo es realizar un cribado de variables redundantes (o que añaden ruido o entorpecen los cálculos de parámetros). En este paso también se pueden añadir más variables a través de transformaciones lineales o no lineales que tengan más representatividad que los datos crudos[10]. Se habla más de este paso en la sección 3.2.
- Seleccionar el modelo o modelos para el problema: una vez que tenemos más información sobre cómo se relacionan los datos, cuál es su naturaleza y cuál es el objetivo: un valor numérico (regresión) o un valor categórico (clasificación), es más fácil seleccionar entre todos los tipos de modelos disponibles en la literatura y en las librerías de programación. Esta tarea es un arte en sí misma dada la enorme variedad de modelos que se adaptan a situaciones similares.
- Partición de datos para entrenamiento y validación: uno de los principales errores a la hora de diseñar una estrategia de análisis de datos con fines predictivos es no escoger bien cómo se debe validar el sistema o modelo predictivo desarrollado. En general se suelen usar dos particiones disjuntas de los datos disponibles: una para entrenamiento de los parámetros del modelo y otra para comprobar cómo se comporta el modelo sobre datos no usados en entrenamiento. Cuando el conjunto de datos que tenemos es pequeño en comparación con el número de parámetros que se deben estimar en el modelo[11] y necesitamos aprovechar todo lo posible los que tenemos, en lugar de hacer una partición fija de entrenamiento/validación, se suelen aplicar técnicas de validación cruzada, consistentes en hacer una partición pequeña para validación, entrenar con la partición restante y anotar el resultado sobre la de validación. Repitiendo este proceso sobre un número finito de particiones diferentes y promediando los resultados se consigue estimar de manera más acertada el comportamiento real del sistema que si hubiéramos hecho solo una partición entrenamiento/validación. El caso extremo de esta técnica donde la validación se hace solo sobre una muestra y se entrena con todas las demás se conoce como leave-one-out.
- Evaluación de prestaciones: esta tarea es fundamental para poder decidir si un sistema es válido para nuestros objetivos y para informar sobre su utilidad. Si el modelo predictivo resuelve un problema de regresión sobre una variable continua, como puede ser la predicción de ventas de abonos para la siguiente temporada, entonces es habitual utilizar el coeficiente de determinación o R^2 que indica la proporción de varianza en la respuesta (ventas) que el modelo es capaz de explicar. Otra medida de evaluación de prestaciones de un regresor muy utilizada es la raíz cuadrada del error cuadrático medio (RMSE) entre el valor de respuesta estimado y el verdadero valor. Si el modelo predictivo resuelve un problema de clasificación hay

[10] Datos que no han recibido ningún tipo de tratamiento después de ser recopilados

[11] El término 'pequeño' depende del tipo de modelo, pero una regla habitual para obtener un modelo que generalice bien es tener 10 veces más datos que parámetros del modelo.

varias medidas de prestaciones. Como la gran mayoría de los problemas de clasificación son binarios (tomar una decisión entre las opciones A ó B), se han acuñado muchas medidas a partir de la matriz de confusión, que expresa en una matriz 2x2 el número de muestras clasificadas correctamente, en la diagonal, y el número de muestras clasificadas incorrectamente fuera de la diagonal (ver Figura 1). A partir de la tasa de verdaderos positivos (TPR, en inglés) y la de falsos positivos (FPR) se describe la curva ROC (del inglés, Receiver Operating Characteristic) que indica el balance entre las medidas contrapuestas TPR y FPR para cada punto de trabajo del modelo de decisión binaria, o lo que es lo mismo, el compromiso entre sensibilidad y especificidad. Por ejemplo, si el modelo de predicción debe decidir si un socio va a renovar su abono (evento positivo) o darse de baja (evento negativo) en la siguiente temporada, no es lo mismo trabajar en un punto en el que TPR y FPR son altos (estaremos prediciendo que va a haber más abonos que los que probablemente haya y tomaremos decisiones con un dinero que no tendremos) que en un punto en el que TPR y FPR son bajos (estaremos prediciendo que va a haber menos renovaciones de las que probablemente haya y tomaremos decisiones sobre campañas de retención y captación que pueden ser incorrectas). Cuanto más alejada de la diagonal sea la ROC mejor es el sistema de predicción porque tendremos más puntos de trabajo en los que mantenemos un TPR alto con un FPR bajo. Otra forma muy similar de expresar las prestaciones del modelo es la curva ROC definida como precisión-recuperación (Precission-Recall, en inglés), donde recuperación es igual a la sensibilidad (TPR), pero la precisión se define como la tasa de positivos reales con relación a todos los recuperados. En ambos tipos de definiciones de la ROC la medida que evalúa las prestaciones del modelo sin tener en cuenta su punto de trabajo es el área bajo la curva ROC (AUC en inglés). Cuanto mayor es AUC mejor separa el modelo predictivo los dos eventos de interés. Para dar un valor de prestaciones en un punto de trabajo determinado es muy conveniente utilizar lo que se denomina el valor F1, que combina precission y recall haciendo la media armónica entre la precisión y la recuperación y así permite calcular varios sistemas en su punto de trabajo óptimo. En el capítulo 4 vamos a usar el software WEKA, que permite obtener resultados de prestaciones con varias de estas medidas. Este software, por ejemplo, calcula la curva ROC como precission-recall en lugar de como sensibilidad-especificidad.

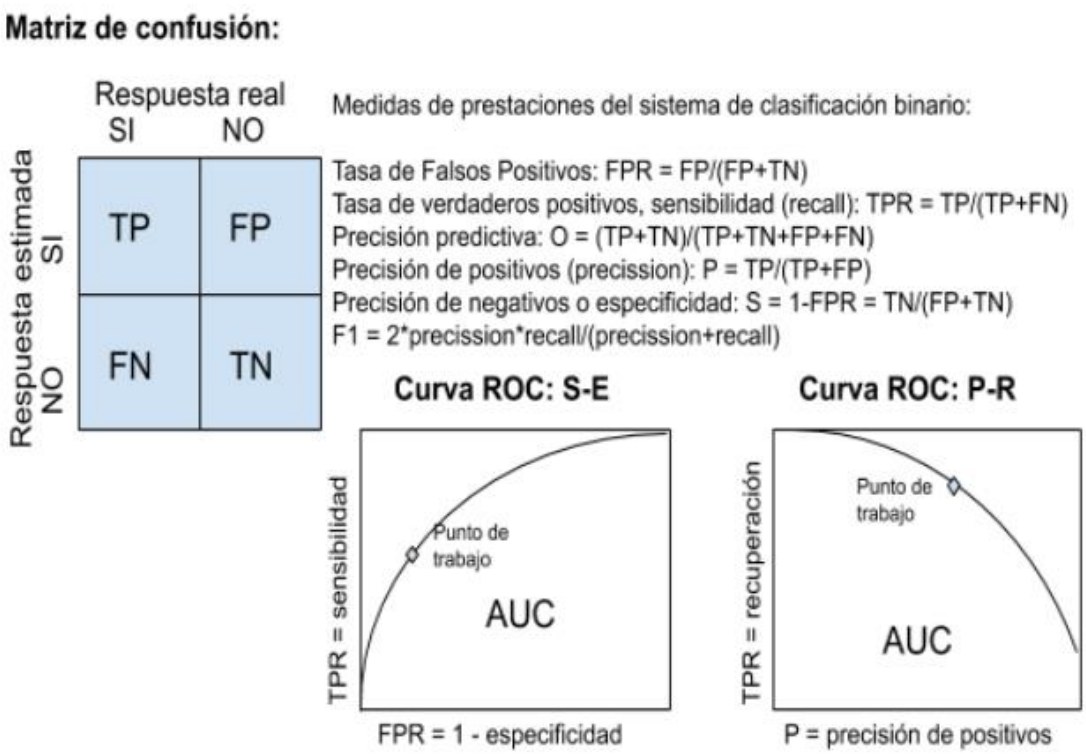


Figura 1: Medidas de prestaciones de un modelo predictivo de clasificación binaria. (Elaboración propia)

Para evaluar prestaciones en clasificadores multi-clase se suele utilizar simplemente la "accuracy" o precisión predictiva, que computa todos los resultados correctos para todas las clases respecto al número total de muestras, o lo que es lo mismo, la suma de la diagonal de la matriz de confusión multiclase dividida por la suma total de la matriz. Para utilizar las medidas definidas para clasificadores binarios es habitual tratar cada clase como la positiva y el resto como negativas y dar esas medidas para cada clase. Como veremos en el capítulo 4, WEKA calcula las prestaciones por clase y también el promedio de dichas prestaciones (TPR, FPR, precission, recall, F1, AUC) entre todas las clases para ofrecer un resultado único comparable.

Para finalizar esta sección, en la siguiente figura reflejamos todas estas tareas de análisis de datos y la relación entre ellas. Los bloques con líneas discontinuas son prescindibles, aunque aconsejables. Las flechas indican sentido del flujo de datos, y cuando regresan sobre un bloque significa que se itera sobre los datos. En las siguientes secciones vamos a desarrollar un poco las tareas más relevantes y pondremos ejemplos extraídos de la literatura de análisis de datos deportivos.

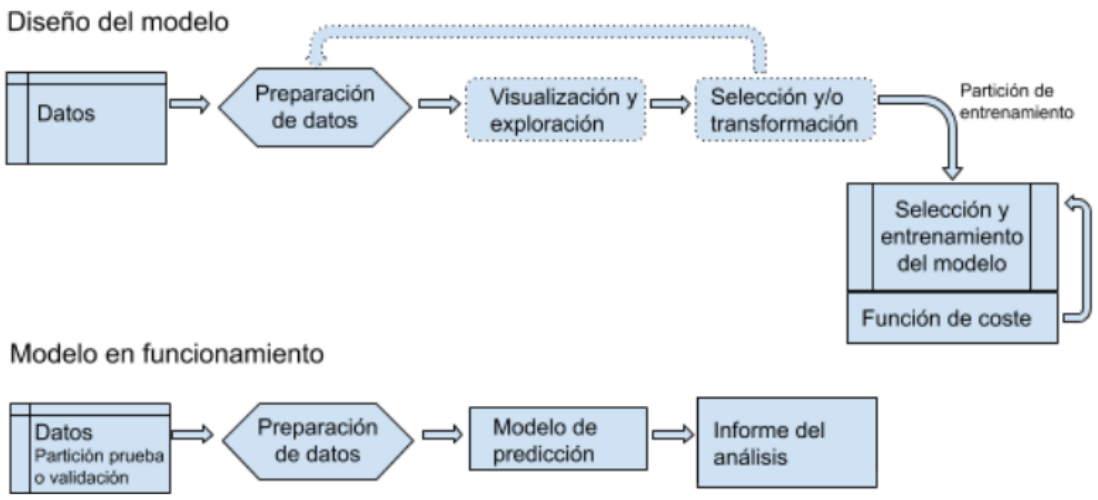


Figura 2: Tareas comunes en análisis de datos para modelado predictivo (Elaboración propia)

3.1 Preparación de datos

Dada la heterogeneidad de fuentes de datos que tenemos disponibles se hace necesario preparar los datos para que estén en el formato más adecuado para sacarles el máximo provecho. Este paso es poco dependiente de la técnica de modelado que se vaya a usar, y puede constar de varias técnicas no excluyentes

- **Tratamiento de datos perdidos.** En cualquier proceso de toma de datos es común que haya algún dato que se ha perdido por algún motivo externo. En función de qué dato se pierda, el tipo de análisis que se va a hacer y la proporción de datos perdidos, las técnicas a aplicar pueden diferir. Cuando el dato perdido es imprescindible, la metodología más habitual es la de **imputación de datos**, que consiste en "rellenar" el dato perdido con el valor más plausible. Supongamos que tenemos una serie datos de jugadores de fútbol, desde su *nacionalidad* hasta los *kilómetros corridos en* cada partido en los últimos 5 años. Si el dato que falta es la nacionalidad, es irrelevante para un estudio de su rendimiento; pero si faltase la distancia recorrida para determinado partido probablemente sería mejor utilizar una técnica de imputación de datos para estimarla.[12]
- **Tratamiento de outliers.** De la misma manera que puede haber datos perdidos, muchas veces nos encontramos con "outliers" (datos extraños o aislados), que no encajan en la lógica. Identificar outliers es complejo y además las causas de estos también pueden ser muy variadas, desde fallos de trasiego de datos entre diferentes bases de datos, hasta fallos en el sistema de adquisición del dato. Pongamos un caso

[12] La gran mayoría de estas técnicas se centran en el impacto de las imputaciones realizadas sobre la estimación de la media poblacional y su varianza. No es objeto de este TFG tratar los métodos de imputación, pero en el artículo de Muñoz Rosas y Álvarez Verdejo (2009) se describen los métodos de imputación más utilizados.

con el ejemplo anterior, si en los datos de jugadores tenemos el atributo *velocidad punta por partido,* y tenemos una observación con un valor de 60 Km/h estamos ante un outlier, ya que tal velocidad es imposible para cualquier persona. En este caso podríamos simplemente imponer unos límites naturales para detectar el outlier y luego aplicar alguna técnica de imputación de datos. Pero si el dato es de 30 Km/h podría ser perfectamente factible y pasar desapercibido... La relevancia de tratar ese dato como outlier o no depende, de nuevo, del propósito del estudio y de la técnica de modelado que se vaya a utilizar.

- **Generación de atributos a partir de datos adquiridos.** Los sistemas y dispositivos de adquisición de datos suelen proporcionar datos crudos que en muchas ocasiones aún tienen que pasar por algún tipo de transformación y/o combinación con otros para generar variables o atributos útiles para un estudio determinado. Por ejemplo, un sistema de medida utilizando las imágenes de video del campo de fútbol, con un sistema que detecta cada jugador, nos puede dar su posición en el campo 25 veces por segundo. Este dato se considera crudo porque tiene poca utilidad por sí solo. Sin embargo, combinando con el instante temporal y cogiendo una secuencia temporal de unos segundos se pueden sacar datos más perspicaces, como la velocidad punta, velocidad media, velocidad regateando.
- **Normalización numérica.** Cuando tenemos muchos atributos numéricos de diferente naturaleza que van a formar parte de un mismo modelo matemático hay que poner mucho cuidado en que las magnitudes de estos sean compatibles y no acaben dominando de manera artificial unos atributos sobre otros. Por ejemplo, el atributo de *velocidad punta* tendrá valores entre 20 y 35 Km/h, pero el atributo *número de metros recorridos por partido* podría variar desde algunos cientos a varios miles. Si no se normalizan de ninguna manera, cualquier medida usando un modelo con esos atributos podría estar dominada solo por los metros recorridos. Lo habitual es normalizar las variables para que tengan márgenes de variación similares. Existen varios tipos de normalización, siendo la min-max y la estandarización las más habituales. La normalización min-max consiste en encontrar los valores mínimo y máximo del atributo y aplicar a cada dato una transformación lineal para llevarlo al intervalo [0, 1] mediante la siguiente fórmula: $x' = (x - min)/(max - min)$[13]. Por otro lado, la estandarización a media 0 y varianza 1, calcularía la media y desviación típica de cada atributo y generaría una nueva variable $x' = (x - media)/desv$ que entonces pasa a tener media 0 y varianza 1[14].
- **Cuantificación de variables categóricas.** En el análisis deportivo, como en otros campos, podemos tener muchos datos de naturaleza categórica (p.ej. equipo o temporada) En función del tipo de modelado a utilizar puede ser conveniente convertir las variables cualitativas en numéricas. La técnica más habitual es la indexación, donde se desdobla cada variable en tantas como elementos diferentes

[13] Min-max es muy sensible a outliers, ya que establecen ficticiamente el mínimo y/o el máximo de variación y dejan reducido el margen dinámico de los nuevos valores.

[14] En estandarización los outliers seguirán teniendo valores alejados de la normal, pero se podrían usar reglas comunes a todos los atributos para localizarlos, como el de llevar a +/-2 todos los valores cuyo valor absoluto es mayor que 2, por ejemplo.

contiene. Por ejemplo, en una Liga con 23 equipos el nombre de equipo pasa a ser representado por 23 variables binarias, de tal modo que cada registro en lugar de tener el nombre Real Madrid CF en el campo de *nombre_de_equipo*, ahora tiene un 1 en el campo de nombre *Real_Madrid* y un 0 en el resto de los campos de los demás equipos. En general, si no hay necesidad de hacer este tipo de conversiones, es mejor dejar las variables categóricas como lo que son y utilizar técnicas específicas que respeten su propia naturaleza, como el análisis de asociaciones mediante tablas de contingencia y test Chi-cuadrado.

Para terminar esta sección sobre preparación de datos podemos comentar un ejemplo real. Aalber y Haaren (2019) plantean una solución basada en análisis de datos para ayudar en la toma de decisiones en fichajes de futbolistas. Para ello parten de un conjunto de 21 estilos de juego y proponen un método automático para identificar los estilos más adecuados para cada jugador fichable. De esta manera consiguen anticipar si cierto jugador encajaría en el rol que necesita el equipo. Es decir, no se trata de ver si un jugador es el mejor, sino predecir si va a ser el mejor fichaje para el equipo.

Los autores definieron los estilos de juego utilizando datos de la temporada 2017-2018 de 7 ligas europeas (obtenidos de Wyscout[15]) y la ayuda de expertos de fútbol. Por ejemplo, para el estilo de juego "centrocampista recuperador", cuya característica fundamental es la de recuperar la posesión, los atributos típicos son *corte de pase contrario*, *entorpecer la organización contraria*, y *hacer pases cuando se recupera la posesión*. Los jugadores prototipo para este estilo de juego son Kanté (Chelsea FC), Casemiro (Real Madrid CF) y Lo Celso (Paris Saint-Germain FC, ahora Tottenham Hotspurs).

Ahora no vamos a entrar en los detalles del modelado, simplemente vamos a ver cómo han preparado los datos a partir de los de la base de datos. Es muy habitual definir atributos como promedios de alguna combinación de datos (u otros atributos), en intervalos de tiempo que tengan sentido deportivo. Los registros de la base de datos de Wyscout ya se han obtenido mediante un proceso de filtrado de datos crudos. Así, por ejemplo, el evento "robo" proviene de detectar, a partir de la posición del balón y de los jugadores, un cambio de posesión sin que haya un evento de balón parado (saque de banda, falta, etc.).

Todas las plataformas que ofrecen datos deportivos a los clubes parten de datos crudos adquiridos en el terreno de juego y los preparan para dotarlos de un sentido deportivo fácil de interpretar. En el caso de MediaCoach, al finalizar cada partido, los 42 clubes integrados en LaLiga tienen a su disposición miles de clips de video, informes y análisis de cada jugador con decenas de estadísticas en ese partido, en todos sus partidos, las de su equipo y las de LaLiga (Figura 3).

[15] Plataforma de pago para el análisis de partidos de fútbol con videos, datos, estadísticas y herramientas varias. https://www.wyscout.com

Figura 3: Informe de jugador generado por Mediacoach al finalizar un partido (Sabatés & Martín, 2020)

En este artículo los autores construyen 242 atributos para cada jugador en dicha temporada. Algunos ejemplos son: "*número de robos de balón a corta distancia*", "*número de pases dados al área de penalti*", "*número de intentos a puerta de calidad*", etc. Cada uno de estos atributos resulta de alguna combinación de datos crudos que puede incluir reglas de construcción, como en el caso de "*número de robos de balón a corta distancia*" que implica filtrar el evento "robo" mediante un criterio de distancia entre jugadores contrarios, por lo que hace falta combinar las coordenadas en el campo de los jugadores implicados (datos crudos que están en cada registro).

Además de extraer estos atributos, los autores los normalizan en dos etapas. En una primera etapa realizan una normalización específica para el problema que quieren resolver: dividir cada valor por el número de eventos de balón que realizó cada jugador en toda la temporada. Así, para un jugador no se tiene en cuenta si su equipo ha tenido más minutos de posesión y se independiza la actividad de un jugador (que es lo que se quiere analizar) del equipo al que pertenece, ya que equipos fuertes tendrán en media más eventos de balón que equipos débiles. Además, se genera un segundo conjunto de atributos normalizando por el número de eventos, no del jugador, sino de su equipo, haciendo que resalten los jugadores en cada equipo, que dominan más el juego. Así los autores pasan a tener 484 variables normalizadas en cada registro de evento. En una segunda etapa se realiza una normalización del tipo estandarización a media 0 y varianza 1 para las 484 variables y se eliminan posibles outliers fijando a {-2,2} todos los valores con valor absoluto mayor que 2.

3.2 Selección de variables, análisis de datos y modelado

En esta sección se introducen y resumen algunas técnicas de selección de variables y modelado de datos, haciendo referencia a un estudio bastante completo que utiliza el análisis de datos y aprendizaje automático con atributos de jugadores de fútbol.

Gunjan Kumar, gracias a los datos ofrecidos por Manchester City Football Club Analytics, a través de diferentes métodos de selección de variables y de diversas técnicas de aprendizaje automático analizó los ratings[16] otorgados por Whoscored[17] (Kumar, 2012). Su objetivo era identificar los atributos más importantes a la hora de determinar los ratings oficiales, averiguar qué atributos afectan al resultado del partido y hasta qué punto el resultado del partido se puede predecir a través de estos atributos. Kumar también agregó los ratings de jugadores individuales por equipo para crear una calificación de equipo y analizó si era posible determinar el resultado del partido según estas nuevas calificaciones. Este estudio fue posible ya que durante la temporada 2011-2012 el Manchester City Football Club llegó a un acuerdo con Opta[18] para ofrecer de manera pública los datos recogidos sobre jugadores durante partidos (hallados en el repositorio de GitHub de MCFC Analytics[19]).

3.2.1 Selección de variables

Cada vez que necesitamos resolver un problema de descripción de una situación, de predicción de un suceso o de clasificación entre un conjunto de opciones, necesitamos identificar cuáles son los datos más relevantes que nos van a ayudar a generar un modelo, de manera que es muy importante saber que variables vamos a utilizar en el modelado. Por ejemplo, si queremos estimar la probabilidad de que con determinado fichaje de un delantero aumente la compra de abonos de temporada, habrá un conjunto de variables que parecen potencialmente predictivas, como su promedio de goles por partido, su progresión en las últimas temporadas, su impacto en redes sociales, , o incluso su facilidad de palabra en las ruedas de prensa tenga algo de influencia. Pero las variables que probablemente no influyan en la compra de abonos (aunque puedan influir en las decisiones estratégicas del equipo técnico) son su nivel de estudios, su velocidad de recuperación ante lesiones, el tiempo de recuperación tras un sprint, o si su familia va a mudarse con él a la nueva ciudad.

Cuando la naturaleza del proceso a modelar es extremadamente compleja como para prever qué atributos van a ser útiles o no, pero tenemos muchos registros en comparación con el número de variables, suele ser habitual acudir a técnicas de aprendizaje automático que sean capaces de atenuar implícitamente el efecto perjudicial de las variables que no tienen una relación causa-efecto con la variable o variables dependientes que queremos que nuestro modelo estime.

[16] En este contexto, los ratings son calificaciones que se otorgan a los jugadores según la calidad de su actuación en un partido
[17] Página de datos de fútbol que otorga ratings a jugadores en ligas de todo el mundo
[18] Compañía de análisis deportivos, proveedor oficial de datos para la Premier League
[19] https://github.com/iKhaled/MCFCAnalytics

A continuación, se enumeran los dos tipos de procedimientos de selección de variables más generales y una regla aplicable a todos (Ruczinski, n.d.):

- **Procedimientos paso a paso** (Stepwise Selection): Consiste en la repetición de un modelo de regresión cambiando el listado de variables utilizadas en cada iteración. Dentro de este método existen 3 grupos diferentes
    - Eliminación hacia atrás (Backward Selection): Introducción de todas las variables en él modelo y después exclusión de una (o varias) en cada iteración. En cada etapa se elimina la variable menos influyente según el criterio estadístico establecido.
    - Selección hacia adelante (Fordward Selection): Introducción de una o varias variables en cada iteración según un criterio establecido. Un ejemplo es utilizar la correlación entre variables independientes y la dependiente.
    - Regresión bidireccional (Bidirectional Regression): combinación de Backward y Fordward selection, de manera que en cada iteración se añaden y eliminan variables.
- **Procedimientos basados en criterios**: Si hay p predictores potenciales (variables), entonces hay 2^p posibles modelos. Estos métodos consisten en crear todos los posibles modelos y elegir el mejor modelo (o el más ajustado) según el criterio establecido, de manera que habremos eliminado todas las variables que no estén en ese modelo
- **Selección en jerarquía**: Algunos modelos y variables tienen una jerarquía natural. Por ejemplo, en modelos polinomiales, x^2 es un término de orden superior a x . Al seleccionar variables, es importante respetar la jerarquía y no eliminar del modelo términos de orden inferior antes de los términos de orden superior en la misma variable. Lo mismo sucede con términos de interacción[20] y las variables que interactúan entre sí, no se puede eliminar una de las variables de interacción y dejar un término de interacción en el modelo.

En el TFM *Machine Learning for Socce Analytics* (Kumar, 2012) el autor utilizó tres técnicas de selección de variables diferentes (todas de backward selection) para identificar los atributos más importantes a la hora de calcular el rating de jugador. El primer método de selección de variables, “Global Ranked Pruning”, se basó en la eliminación de atributos tras hacer una clasificación en orden decreciente del valor absoluto de los pesos asignados por los algoritmos de mejor rendimiento, y luego podando progresivamente los atributos al final de la lista de cinco en cinco (para acortar el proceso de análisis). En la segunda estrategia de selección de variables, “Iterative Local Pruning”, el algoritmo escoge atributos basándose en el modelo generado por ese algoritmo en su ejecución previa. Por lo tanto, la eliminación es "local" a la ejecución inmediata de cada algoritmo. Esta estrategia de poda es aplicable solo para algoritmos cuyo modelo puede ser interpretado para derivar una lista clasificada de atributos, ya que sigue empleando el criterio de valor absoluto de los pesos de parámetros. El tercer método de selección de variables, “Threshold Pruning”, utiliza un umbral mínimo para seleccionar los atributos elegidos por la estrategia “Iterative Local Pruning”. Por lo tanto, solo selecciona los atributos incluidos en el modelo aprendido en la

[20] Un término de interacción es una variable creada para captar un efecto (ya sea de amplificación o reducción) que ocurre cuando dos variables explicativas suceden al mismo tiempo en un modelo

ejecución anterior cuyo peso está por encima del valor establecido. El umbral se establece de acuerdo con la desviación estándar y la media de los pesos asignados a los atributos incluidos en el modelo aprendido en la ejecución previa. Los tres métodos de selección de variables se realizaron con diversos modelos predictivos para así obtener diferentes listas de atributos con el propósito de hacer una comparación más adelante.

Finalmente, para identificar los atributos más importantes a la hora de calcular el rating de jugador y determinar un solo listado, Kumar observó las listas de atributos de los mejores modelos (según r, MSE y RMSE[21]) y creó un listado de todos los atributos presentes. A continuación, estableció una estrategia de recompensa y penalización para clasificar estos atributos. Para cada atributo contó su número de apariciones en las diferentes listas de atributos, además de anotar la posición en cuanto la importancia del atributo dentro de cada una de las listas. De manera que, por cada aparición de un atributo, la recompensa se corresponde con su posición en la lista de atributos; por lo tanto, la recompensa total sería la suma de su rango en las listas de atributos de modelo en las que ese atributo estaba presente. La penalización para un atributo sería el producto del tamaño promedio de listas de atributos de modelo para esa posición y el porcentaje de listas de atributos para esa posición en el que ese atributo estaba ausente. A través de este proceso de selección de variables Kumar, consiguió reducir la cantidad de atributos de 200 a 25, de manera que el resto de su análisis sería más fácil y sus resultados más generalizables.

3.2.2 Métodos supervisados y no supervisados

Antes de comentar las tareas de regresión y clasificación que realizó Kumar con los datos de jugadores de fútbol se van a introducir los dos grupos de métodos generales de la ciencia de datos, los métodos supervisados y los métodos no supervisados, además de comentar algunas de las técnicas más típicas de cada uno de ellos.

3.2.2.1 Métodos Supervisados:

Los métodos supervisados se utilizan en los casos en los que se tiene variables de entrada (independientes) y una variable de salida (dependiente) y queremos encontrar una relación entre éstas a través de una función (p. ej. conocer la calificación de un jugador según sus acciones durante un partido). El objetivo de estos métodos es aproximar tan bien la función en cuestión que cuando se tengan nuevos datos de entrada se pueda predecir la variable de salida para dichos datos. Se denominan supervisados porque durante el proceso de entrenamiento de la función o algoritmo, conocemos las respuestas correctas para los datos. De manera que se realizan predicciones sobre los datos de entrenamiento, se supervisan estas predicciones y se van haciendo correcciones a través de la minimización de la función de coste hasta hallar la mejor solución. Los métodos supervisados se utilizan para solucionar dos tipos de problemas, de regresión y clasificación, los cuales se comentan en la sección 3.2.3 y 3.2.4 respectivamente.

A continuación, se enumeran los grupos generales de métodos supervisados más utilizados en la práctica:

[21] Coeficiente de correlación, Error Cuadrático Medio y Raíz del Error Cuadrático Medio

Regresión lineal: en este método se modela la relación lineal entre una respuesta escalar (variable dependiente) y una o más variables explicativas (variables independientes). La regresión lineal se utiliza para ajustar un modelo a un conjunto de datos observados con un objetivo y x número de características explicativas. Después se utiliza dicho modelo con datos sin un valor de respuesta objetivo para realizar una predicción de la respuesta gracias a la combinación lineal (obtenida al ajustar el modelo) de los valores de las características de los nuevos datos. Se comentan ejemplos en los capítulos 3 y 4.

Regresión Logística: es un modelo estadístico que en su forma más básica usa una función logística para modelar una variable dependiente binaria, aunque existen muchas extensiones más complejas. Como por ejemplo la regresión logística multinomial, en este caso el objetivo del modelo es solucionar una tarea de clasificación con más de dos clases, pero que se puede interpretar al considerar cada clase como un evento binario contra todos los demás. En el capítulo 4 se comenta un modelo multinomial.

Árbol de decisión: método en el que se realizan conjunciones de variables independientes que conducen a determinadas etiquetas de clase (variables dependientes). Se le llama árbol de decisión o árbol de clasificación ya que se puede visualizar el modelo, interpretando las conjunciones de variables como ramas y las etiquetas de clase como hojas. Los árboles de decisión donde la variable objetivo puede tomar valores continuos (típicamente números reales) se llaman árboles de regresión.

Random Forest: en castellano "Bosque Aleatorio", en un método en el que se combinan muchos árboles de decisión con diferentes conjunciones de variables independientes en cada uno para solucionar el mismo problema. De esta manera se obtienen muchas soluciones y se hace un recuento entre estás, como un recuento de votos, para proponer la solución del random forest. Al igual que los árboles de decisión, se pueden utilizar para regresión y para clasificación.

K-Vecinos Cercanos: popularmente conocido como knn (K-Nearest Neighbors), es un método que consiste en clasificar observaciones según la clase de los "vecinos" cercanos en el espacio de características. Una observación se clasifica por una pluralidad de votos de sus vecinos, y se asigna a la clase más común entre sus k vecinos más cercanos. Si k = 1, entonces el objeto simplemente se asigna a la clase del vecino más cercano. También se puede utilizar en casos de regresión, donde la salida es el valor de la variable dependiente de la observación, el cual será el promedio de los valores de k vecinos más cercanos. Una diferencia importante con los demás métodos es que k-NN no busca ninguna función explicativa sino que clasifica directamente las muestras nuevas al tener almacenadas las muestras con sus etiquetas (suele ser lento y ocupar mucha memoria).

Support Vector Machine: las "máquinas de vectores soporte" son un método que se basa en construir hiperplanos en el espacio de dimensiones de los datos (SVM

lineal) o en un espacio de dimensión superior (SVM RBF). El objetivo del método es separar los datos de la manera más fiable posible, lo cual se logra encontrando los vectores soporte que forman los hiperplanos que tengan la mayor distancia al punto de datos de entrenamiento más cercano de cualquier clase o valor. Se usa para problemas de clasificación y regresión respectivamente.

Redes neuronales: Este método se basa en complejos sistemas que "aprenden" a realizar tareas considerando los datos de entrada. Hay muchos tipos de redes neuronales, pero las más usadas con diferencia son los perceptrones multicapa (Multi-Layer Perceptron) porque están compuestos de capas de "neuronas" donde cada una se comporta de manera similar a una regresión logística binomial. Así cada neurona recibe información mediante una combinación lineal de los datos de su entrada, y produce una salida mediante una función no lineal (tipo logit) para otra neurona en otra capa. La forma más fácil de entender estos sistemas es pensando en el reconocimiento de imágenes. Las redes neuronales aprenden a identificar imágenes que contienen un objeto x analizando imágenes de ejemplo que se han etiquetado manualmente ("con objeto x "/ "sin objeto x ") y utilizan los resultados para identificar el objeto x en otras imágenes. Lo hacen sin ningún conocimiento previo de dicho objeto, de manera que generan automáticamente características de identificación a partir de los ejemplos que procesan. Las redes neuronales son sistemas increíblemente complejos que encuentran relaciones no lineales entre características de datos que normalmente son imposibles de describir. Cuantas más capas tenga la red, más complejo será el modelo y las relaciones que busque en los datos. En el capítulo 4 se usará el MLP tanto para clasificación como para regresión.

3.2.2.2 Métodos no supervisados

Los métodos no supervisados se utilizan en aquellos casos donde solo se tienen los datos de entrada y no hay variables de salida correspondientes. El objetivo de estos métodos es modelar la estructura o distribución subyacente en los datos para aprender más sobre estos o encontrar relaciones previamente desconocidas. A diferencia de los métodos supervisados, no hay respuestas correctas para los problemas en cuestión ya que no hay un objetivo definido, por lo tanto, no se puede realizar una "supervisión"). Las tareas de aprendizaje no supervisados pueden dividirse en problemas de agrupación y asociación.

Agrupación: normalmente conocido por el término en inglés, "clustering", es un problema donde se desea descubrir las agrupaciones inherentes en los datos dadas sus características. Lo que intenta es dividir los datos en distintos grupos, de manera que los datos que pertenecen a un mismo grupo son entre ellos lo más homogéneos posible, y consecuentemente, heterogéneos a otros grupos. Existen muchos algoritmos de clustering que se basan en enfoques diferentes, como el análisis de centroides (centros de los clústeres), la agrupación de jerarquía o los basados en de distribución de los datos.

Un algoritmo de clustering muy común es el **K-Means** (K-Medias). En este algoritmo el primer paso es identificar k número de centroides, dado que es

un método no supervisado hay que designar un valor para k con el "método del codo", que implica calcular suma de los errores cuadrados dentro del clúster (Within-Cluster-Sum of Squared errors*)* para diferentes valores de k, y elegir k donde WSS multiplicado por k comienza a aumentar[22]. El nombre proviene de que en el gráfico de $WSS * k$ en función de k, la función se ve como un codo flexionado en el k óptimo. Las "medias" se refieren al promedio de los datos en cada clúster; es decir, a los centroides. En el experimento 2 de la sección 4.2.2 hay un ejemplo del uso de K-means. Otra forma de encontrar el k óptimo es mediante la medida silhouette, que no solo tiene en cuenta las distancias al centroide del cluster asignado sino también las distancias a los otros clusters.

Asociación: son problemas dónde el objetivo es descubrir las reglas que describen grandes porciones de sus datos o relaciones interesantes entre variables. Las reglas de asociación se definen como una implicación del tipo "si x entonces *Y*" ($x \Rightarrow y$). El lado izquierdo de la regla recibe el nombre de antecedente (o *Lenft-Hand-Side*) y el lado derecho el nombre de consecuente (o R*ight-Hand-Side*). Este método podría servir, por ejemplo, para descubrir que las personas que compran x también tienden a comprar y. El soporte de una regla es el número de observaciones que contienen la x de dicha regla, dividido entre el total de observaciones. Normalente una regla necesita un soporte de varios cientos de registros (observaciones) antes de que ésta pueda considerarse significativa desde un punto de vista estadístico. Otra medida de validación para una regla es la confianza que se calcula como la probabilidad de que una observación que contiene x, también contenga y (la confianza puede interpretarse como un estimador de $P(y|x)$).

El algoritmo de asociación **Apriori** fue el primer algoritmo de búsqueda de reglas de asociación y sigue siendo uno de los más empleados. Consiste en identificar todos los sucesos (itemsets) que ocurren con una frecuencia por encima de un determinado límite y después convertir esos itemsets frecuentes en reglas de asociación.

3.2.3 Problema de regresión

Los problemas de regresión comprenden tareas que utilizan el análisis de datos o algoritmos de aprendizaje automático para determinar el valor de una variable dependiente a través de otras variables independientes, ya sea con el fin de predecir el valor de la variable dependiente o para estudiar el efecto de las variables independientes. Existen muchos tipos diferentes de técnicas para modelar este tipo de problemas, pero no existe una buena teoría sobre que método de regresión usar para cada problema de esta índole. Generalmente se recomienda que se utilicen experimentos y reiteraciones para descubrir que método y configuración dan como resultado el mejor rendimiento para una tarea de

[22] Este método es uno propuesto para seleccionar k, como el de slihouette, pero también se puede determinar subjetivamente si tras explorar los datos se aprecian otro número de agrupaciones de manera muy clara

regresión dada. Los problemas de regresión requieren la predicción de una variable continua, pero las variables de entrada pueden ser contínuas o discretas. Un problema de regresión donde las variables de entrada están ordenadas en el tiempo se llama un problema de pronóstico de series de tiempo (time-series prediction).

Debido a que un modelo predictivo de regresión predice una cantidad, la habilidad del modelo debe ser reportada como un error en esas predicciones. Hay muchas maneras de estimar la habilidad de un modelo de regresión, pero quizás la más común es calcular el error cuadrático medio (Root Mean Squared Error). RMSE es la desviación estándar de los residuos (errores de predicción), los cuales son una medida de qué tan lejos están los datos reales de los valores predichos por la regresión (como en clustering para medir el error cometido al usar el centroide como representante de todo el cluster). No sólo es importante comprobar el poder de predicción del modelo, sino también la correcta estimación de los parámetros (la influencia de las variables independientes en la variable dependiente). La significación estadística de los parámetros se puede verificar mediante un test F del ajuste general, seguida de test t de parámetros individuales.

Volviendo al trabajo realizado por Kumar, es importante mencionar que la tarea de regresión tenía como objetivo principal comprender como se entendían los atributos de los jugadores de fútbol y que importancia tenían a la hora de otorgar un rating a los jugadores. La tarea de modelar los atributos para calcular el rating se realizó utilizando WEKA[23], de manera que el autor pudo utilizar muchos métodos diferentes, como regresión lineal, máquinas de vectores soporte, árboles de decisión, etc. y estudiar como cada método de regresión trataba los atributos de jugadores para proponer un rating. Esto ya se explicó en la sección 3.2.1 ya que la regresión se utilizó durante el proceso de selección de variables. Hay que destacar que los mejores modelos de regresión en el estudio de Kumar tenían aproximadamente un error absoluto medio (Mean Absolute Error) de aproximadamente 0.2. Esto quiere decir que, de media, una predicción de dichos modelos tenía 0.2 puntos de rating de fallo, y considerando que el rating tiene valores del 1 al 10, los modelos eran extremadamente precisos.

3.2.4 Problema de clasificación

Los problemas de clasificación son tareas que requieren el uso de técnicas de análisis de datos o algoritmos de aprendizaje automático para asignar una etiqueta de clase a ejemplos del dominio del problema. Un ejemplo fácil de entender es clasificar los correos electrónicos como "spam" o "no spam". Un modelo de clasificación utiliza el conjunto de datos del problema con las variables independientes seleccionadas para calcular la mejor forma de mapear ejemplos de datos de entrada a etiquetas de clase específicas. Como tal, el conjunto de datos debe representar fielmente el problema y tener muchos ejemplos de cada etiqueta de clase, ya que de no ser así la predicción de la variable dependiente (la etiqueta), en aquellos casos donde la representación de los datos no sea fiel, será mucho más complicada. Las etiquetas de clase a menudo son variables categóricas (p.ej. "Spam" y "no spam"), y deben convertirse en valores numéricos antes de utilizarse para modelar (el método de indexación para cuantificar estas variables se explicó brevemente en la sección

[23] WEKA (Waikato Environment for Knowledge Analysis) es una plataforma de software con acceso libre para el análisis de datos y aprendizaje automático https://www.cs.waikato.ac.nz/ml/weka/

3.1). Al igual que en los problemas de regresión, hay una enorme variedad de métodos para modelar los problemas de clasificación, pero cada problema es un caso diferente y no hay una teoría que dictamine que método vaya a otorgar el mejor rendimiento, de manera que experimentar y reiterar con los métodos más plausibles es la mejor forma de acercarse a la resolución de una tarea específica.

A continuación, se enumeran los tipos de problemas de clasificación más generales según el tipo de objetivo a aprender:

- **Clasificación Binaria:** Típicamente, las tareas de clasificación binaria tienen dos clases con probabilidades a priori parecidas, pero en muchos otros casos una clase es mucho más probable que la otra o tiene un coste de error mucho más alto que la otra. Por eso es importante entender bien los dos tipos de errores posibles y el punto de trabajo del sistema, como se comentó en la Figura 1 del capítulo 3. Así, por ejemplo, en diagnosis médica nos podemos encontrar muchos ejemplos en los que la probabilidad de tener una enfermedad es mucho menor que no tenerla. El mejor ejemplo ahora mismo sería la prueba de anticuerpos del COVID-19, que está arrojando en España resultados del 5% de personas que tienen los anticuerpos, han pasado la enfermedad, contra el 95% que no los tienen. El coste de cada tipo de error juega un papel importante en las decisiones médicas, así, aunque sea más probable no tener coronavirus interesa que la tasa de falsos positivos de nuestro sistema de clasificación binaria sea más alta que la de falsos negativos, para minimizar la proliferación de la pandemia. En escenarios de organizaciones deportivas las decisiones no son tan críticas, pero para el club puede tener distinto coste “renovar abono de socio” y “no renovar abono”.
- **Clasificación Multi-clase**: en estas tareas los ejemplos se clasifican como pertenecientes a uno entre un rango de clases conocidas. Por ejemplo, en el artículo de Aalbers y Van Haaren (2018) (comentado en la sección 3.1) la clasificación de jugadores según su estilo de juego es una clasificación multi-clase. Todo lo dicho en el punto anterior sobre tipos de error y el punto de trabajo es extrapolable al problema multi-clase.

Kumar también enfocó su estudio a los equipos, utilizando los atributos de los jugadores para averiguar cómo agregar las calificaciones individuales de jugadores y predecir el resultado de un partido, de tal manera que dichas calificaciones sean una buena representación del rendimiento del equipo en un partido (partiendo de la hipótesis de que sabiendo el rendimiento de los equipos se podría predecir el resultado del partido). Para resolver este problema de clasificación, el autor trató los datos para crear un nuevo conjunto de datos de partido, ya que previamente los datos contenían la actuación de un jugador en cada observación y para este problema cada observación debía de ser un partido. Este nuevo conjunto de datos contenía los siguientes atributos para cada partido: promedio de ratings de todos los jugadores (para cada uno de los equipos), promedios de ratings en cada posición de jugador (porteros, defensas, centrocampistas y delanteros), calificaciones mínimas en cada posición, clasificaciones máximas en cada posición y el

resultado del partido. Además, también se incluyó la diferencia entre los ratings ponderados de las posiciones de cada equipo y de todos los jugadores del equipo. Finalmente, Kumar seleccionó las variables más importantes de este nuevo conjunto de datos según su predicción al resultado del partido. Las variables seleccionadas fueron:

1. Rating promedio del equipo local
2. Rating promedio del equipo visitante
3. Diferencia del promedio de rating de delanteros
4. Diferencia del promedio de rating de defensas
5. Diferencia del promedio de rating de porteros
6. Diferencia del promedio de rating de centrocampistas
7. Calificaciones máximas de delanteros del equipo local
8. Clasificaciones máximas de delanteros del equipo visitante

Con este conjunto de variables seleccionadas Kumar obtuvo una precisión del 90% con dos modelos, SMO[24] y RBFNetwork[25], con un área bajo la curva de 0.946 y 0.97 respectivamente. De manera que resolviendo este problema de clasificación Kumar confirmó que al saber el rendimiento de los equipos (a través del rendimiento de los jugadores) se podría predecir de manera precisa el resultado del partido.

[24] Algoritmo de entrenamiento de máquina de vectores soporte
[25] Tipo de red neuronal

4. TRABAJANDO CON DATOS REALES

Independientemente de qué deporte de equipo estemos hablando siempre hay preguntas aplicables para todos, ¿cuánto vale cada jugador para su equipo?, ¿cuánto aportaría un determinado fichaje?,¿qué repercusión económica tendría la inversión? No cabe duda de que un técnico deportivo o un buen fanático sabe cuándo criticar y cuando aplaudir la actuación individual de un jugador, pero cuando se habla de comparaciones entre jugadores ya no es tan sencillo. ¿Quién aportó más, un delantero marcando gol o un centrocampista creando ocasiones? ¿Un portero salvando un penalti o un central metiendo gol tras un córner en el último minuto? ¿El promedio en toda la liga de un buen jugador estable o actuaciones brillantes pero esporádicas de un jugador impredecible? Cuestiones como estas son difíciles de responder porque el fútbol es un deporte con diversas posiciones en el campo y pocas anotaciones durante un partido, además, las sustituciones están muy limitadas y por lo tanto existe una correlación entre las aportaciones de los jugadores que hace difícil cuantificar la aportación individual de cada uno al resultado final del partido. En el caso del baloncesto esto no ocurre, debido a que el deporte conlleva un gran número de anotaciones y no existe un límite de rotaciones de jugadores. Gracias a estas dos condiciones en el baloncesto existe una medida que cuantifica de manera simplificada la aportación individual de cada jugador. La denominada "+/-", o Más/Menos, utiliza la anotación neta del equipo en el tiempo que un jugador está en la cancha para simplificar y cuantificar la aportación de dicho jugador al equipo. Digamos que mientras un jugador está en la cancha el equipo anota 20 puntos y el rival anota 16, su Más/Menos será de +4. Queda claro que el Más/Menos es una medida extremadamente simplificada y utilizarla por sí sola puede llevar a juicios erróneos, por eso ha de ser complementada con otra información o con variantes de esta, pero, aun así, es una herramienta muy útil a la hora de simplificar el deporte y realizar análisis.

4.1 Análisis cuantitativo de la aportación de jugadores según puntos FIFA

En la sección 3.2 se comentó brevemente un estudio que se centra en las variables deportivas para determinar los ratings de los jugadores. En esta sección vamos a analizar un estudio específico sobre la fiabilidad de ratings de esta índole, en este caso, los ratings FIFA, para poder estudiar la aportación individual de un jugador en términos de puntos de liga, además de predecir el valor de ese jugador en términos económicos.

Konstantinos Pelechrinis y Wayne Winston (Pelechrinis & Winston, 2019), realizaron un estudio para cuantificar el valor de la aportación individual de jugadores en el campo de fútbol de la misma manera que el Más/Menos captura la aportación individual en la cancha de baloncesto. En este apartado del TFG sólo se trabajará sobre este estudio, el cual utiliza una metodología de estadística clásica, pero existen muchos otros relacionados con la cuantificación de las aportaciones individuales que enfocan el tema desde otras perspectivas como pueden ser *Actions Speak Louder than Goals: Valuing Player Actions in Soccer* (Bransen, Davis, Decroos & Haaren, 2019) o *Quantifying the Performance of Individual Players in a Team Activity* (Dutch, Nunes Amaral & Waitzman, 2010).

Pelechrinis y Winston utilizaron datos de aproximadamente 20000 partidos de 11 ligas de fútbol europeas durante 8 temporadas[26], junto con las puntuaciones de los jugadores otorgadas por FIFA[27]. El estudio se planteó identificar cómo un aumento de una unidad en la puntuación de FIFA de un jugador afecta la probabilidad de que su equipo gane el partido. Por ejemplo, si sustituimos a nuestro delantero actual que tiene una puntuación FIFA de 79, con un nuevo delantero con una puntuación FIFA de 80, ¿cuánto aumentan nuestras posibilidades de ganar?

Para su modelo predictivo los autores definieron la variable dependiente como la diferencia de goles (Z = goles local - goles visitante), ya que el interés del estudio reside en el resultado global (ganar, perder o empatar) y no en el marcador específico. Inicialmente las variables independientes seleccionadas fueron las diferencias entre las puntuaciones FIFA de los jugadores de cada equipo. De manera que para la diferencia de goles (Z_i) del partido *i* se utilizarían las variables:

$$x_{i,\pi} = r_{p(h,\pi,i)} - r_{p(a,\pi,i)} , \qquad \forall \ \pi \in \Pi$$

donde $rp_{(h,\pi,i)}$ es la puntuación FIFA del jugador equipo local (*h*: *home team*) en la posición π durante el partido *i* y Π son todas las posiciones posibles de fútbol ($rp_{(a,\pi,i)}$ es el mismo caso en el equipo visitante (*a*: *away team*)).

Para evitar problemas de variables no bien definidas (p. ej. cuando los equipos no usasen las mimas 11 posiciones) los autores optaron por definir como variables independientes el promedio de las puntuaciones de FIFA por posiciones grupales es decir, en vez de hacer comparaciones entre posiciones específicas (π), el modelo compararía los porteros, los defensas, los centrocampistas y los delanteros de un equipo frente a los opuestos, sin hacer especificaciones en las posiciones exactas o en jugadores individuales (Figura 4). Con esto en mente los autores optaron por realizar una regresión basada en la distribución Skellam, modelando la diferencia de goles (Z_i) del partido *i* usando las siguientes variables:

- ❏ La diferencia entre la puntuación FIFA del portero de los dos equipos (X_{GK})
- ❏ La diferencia entre la puntuación FIFA promedio de jugadores defensa de los dos equipos (X_D)
- ❏ La diferencia entre la puntuación FIFA promedio de jugadores centrocampistas de los dos equipos (X_M)
- ❏ La diferencia entre la puntuación FIFA promedio de jugadores delanteros de los dos equipos (X_A)

[26] 2016. European Soccer Database. https://www.kaggle.com/hugomathien/soccer/
[27] https://www.sofifa.com

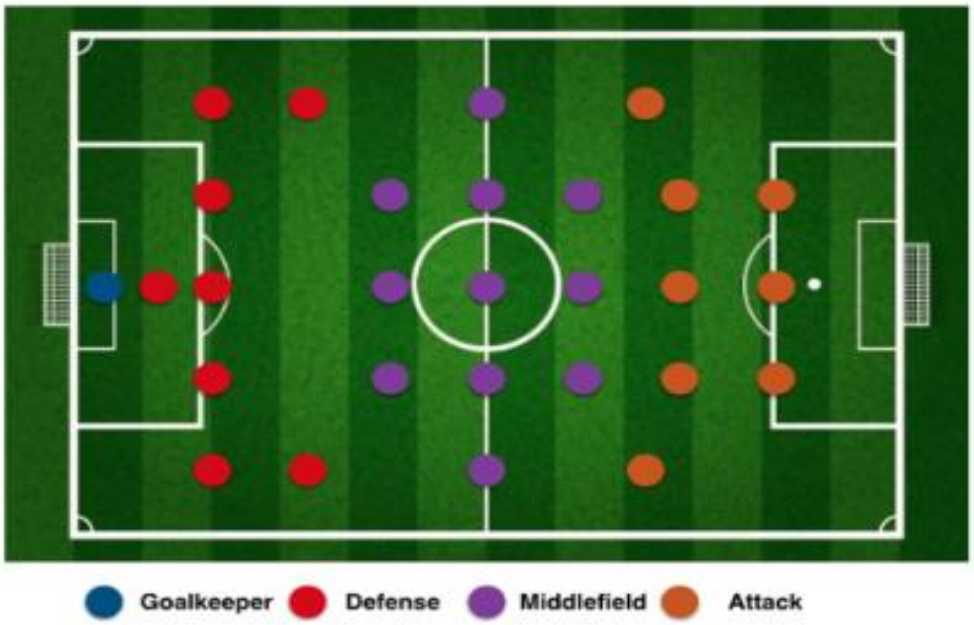


Figura 4: Grupos de posiciones utilizados en el estudio (Pelechrinis & Winston, 2019)4

De manera que la variable aleatoria *Z* sigue una distribución de Skellam, que describe la diferencia de dos distribuciones Poisson con diferente media (en este caso cada una de ellas describe la probabilidad de marcar K goles al otro equipo) y con un posible ruido aditivo común que se cancela al restarse (las condiciones del campo o del clima, por ejemplo):

$$P(z) = e^{-(\lambda_1+\lambda_2)} * \left(\frac{\lambda_1}{\lambda_2}\right)^{z/2} * I_z\left(2\sqrt{\lambda_1\lambda_2}\right)$$

donde sus parámetros dependen del modelo de variables X = (X_{GK}, X_D, X_M, X_A):

$$Z \sim Skellam(\lambda_1, \lambda_2)$$
$$log(\lambda_1) = b_1^T \cdot x$$
$$log(\lambda_2) = b_2^T \cdot x$$

Los autores ajustan el modelo mediante la maximización de la verosimilitud de que el modelo explique el 80% de los datos (alrededor de 16000 partidos), para poder probarlo con los otros 4000. Es importante mencionar que los resultados de la regresión muestran que el incremento de puntuación FIFA promedio para cualquier grupo de jugadores de un equipo conlleva a un aumento de la probabilidad de ganar de ese equipo (como era de esperar). Por ejemplo, supongamos que un central del equipo local es cambiado por otro central con una puntuación FIFA mayor. Dado que el promedio de puntuación FIFA del grupo defensa local es mayor, la variable X_D será mayor, de manera que la probabilidad del equipo local de anotar incrementará y la del equipo visitante disminuirá.

Variable	$\log(\lambda_1)$	$\log(\lambda_2)$
Intercept	0.37***	0.07***
	(0.012)	(0.015)
x_D	0.02***	-0.03***
	(0.01)	(0.002)
x_M	0.02***	-0.015***
	(0.01)	(0.002)
x_A	0.01***	-0.01***
	(0.001)	(0.001)
x_{GK}	0.001	-0.004**
	(0.001)	(0.002)
N	21,374	21,374

$^{***}p < 0.01$, $^{**}p < 0.05$, $^{*}p < 0.1$

Figura 5: Parámetros del modelo con partición aleatoria (Pelechrinis & Winston, 2019)

Con la distribución de la variable Z podemos obtener la probabilidad de ganar, perder o empatar el partido:

- ➔ Local gana/Visitante pierde: P [Z>0]
- ➔ Local pierde/Visitante gana: P [Z<0]
- ➔ Empate: P [Z=0]

Gracias a esto, Pelechrinis y Winston analizan la aportación individual de un jugador *p*. El objetivo de los autores era estimar los "puntos de liga esperados por encima del suplente" (expected league points above replacements, eLPAR). Esta idea proviene de la estadística de sabermetrics VORP (Value Over Replacement Player) de Keith Woolner (Woolner, 2007), basada en la distribución de talento de jugadores y lo que le supone a un equipo adquirir un jugador. Los jugadores estrella son escasos y difíciles de conseguir, los jugadores promedios abundan y no son complicados de conseguir, y los jugadores de calidad suplente son prácticamente ilimitados y se consiguen de manera gratuita. Partiendo de que un jugador suplente no le cuesta nada al equipo, tiene sentido comparar la aportación de un jugador *p* con la aportación de un suplente, para así poder evaluar cuánto vale realmente el jugador *p*. Pelechrinis y Winston tomaron la sugerencia de Woolner de usar los jugadores en el percentil veinteavo como los jugadores de calidad de suplente, utilizando el sistema de puntuaciones de FIFA, esto se traduce en jugadores con puntuaciones aproximadamente de 56. Para estimar eLPAR de un jugador *p* los autores plantearon utilizar el modelo de regresión de distribución Skellam previamente mencionado, partiendo de una situación en la que los dos equipos cuenten sólo con jugadores suplentes y a uno de ellos se le otorgue el jugador *p*. Aquí surge el problema que el impacto del jugador *p* depende de la formación del equipo, ya que no es lo mismo sustituir un central en equipo 5-3-2 que en un equipo 3-5-2, de manera que para calcular eLPAR es necesario contemplar la formación (algo que no

se hizo para la regresión de Skellam debido a la falta de datos). Teniendo en cuenta todo esto, y que una victoria supone 3 puntos de liga y un empate 1 punto de liga, eLPAR de un jugador *p* con una puntuación FIFA r_p se calcula de la siguiente manera:

1. Primero se calcula el aumento en la puntuación FIFA promedio de un grupo de jugadores ($g \in G$) cuando se substituye un jugador suplente por el jugador *p*, basándose en r_p, la formación (ϕ) y la puntuación del jugador suplente de ese grupo ($r_{suplente,\ \phi,\ g}$)
2. Después se calcula el cambio en la probabilidad de ganar, perder y empatar del equipo (δPw , δPd *and* δPl respectivamente) tras introducir el jugador *p*, utilizando el modelo de regresión Skellam
3. Finalmente se calculan los puntos de liga esperados por encima del suplente del jugador *p* ($eLPAR_p$ (ϕ)) como:

$$eLPARp\ (\phi)\ =\ 3\ \cdot\ \delta Pw\ +\ 1\ \cdot\ \delta Pd$$

Pelechrinis y Winston dejan claro que el valor individual que aporta un jugador depende de la formación que utilice su equipo. Esto se aprecia fácilmente pensando en cómo una formación 4-5-1 tiene una dependencia muy grande en su único delantero en comparación con una formación 4-3-3, en la que al haber otros dos jugadores en ese grupo la dependencia del equipo en los delanteros se reparte. Una de las conclusiones más notables es que el grupo de jugadores que menos valor aporta en relación con su puntuación FIFA son los porteros, independientemente de la formación que se esté utilizando.

A partir de esta información los autores plantean estudiar el mercado de jugadores, tanto el valor del jugador a la hora de negociar un fichaje como los salarios que se les paga, a través de su eLPAR. Según Pelechrinis y Winston, dividiendo el valor de mercado (v_p) de un jugador entre su $eLPAR_p$, se obtiene una estimación del coste monetario (c_p) que los equipos están dispuestos a pagar por obtener 1 "punto de liga por encima del suplente" de este jugador. En los gráficos de la Figura 5 se observa el coste de 1 punto de liga esperado para diferentes posiciones, en función del eLPAR que proporcionan al equipo a la derecha, y en función de su puntuación FIFA a la izquierda. Dado que independientemente de donde proceda 1 punto de liga (portero, defensas, centrocampistas o delanteros) los costes de los grupos deberían ser iguales o por lo menos parecidos ($c_{pi} \approx c_{pj}$, $\forall\ p_i, p_j$), pero se puede ver fácilmente cómo este supuesto no se cumple, ya que de ser así, en ambos gráficos deberíamos observar cuatro curvas con valores similares.

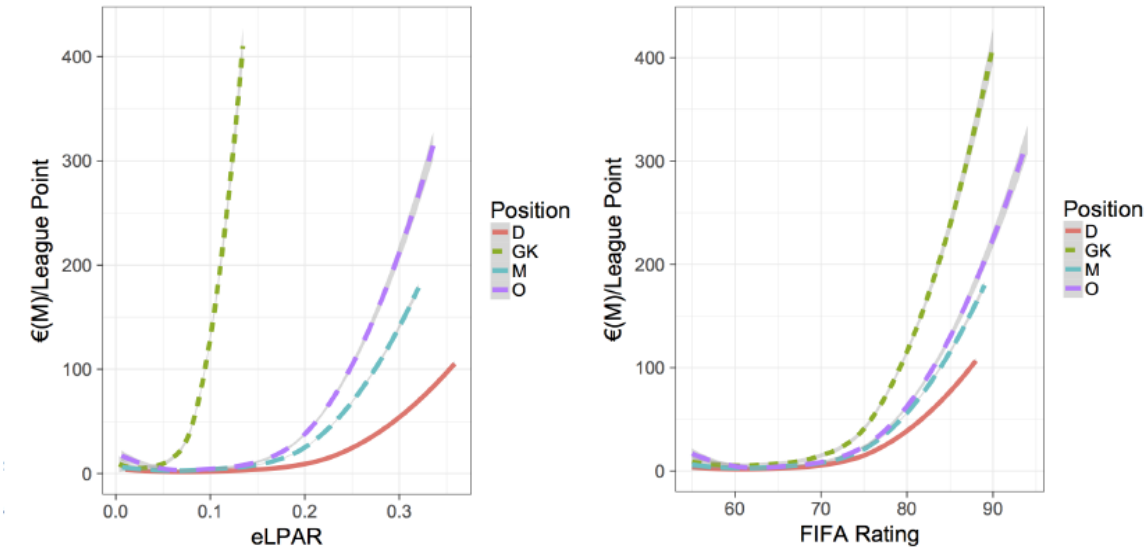


Figura 5: coste de 1 punto de liga esperado para diferentes posiciones, en función del eLPAR (derecha) y puntos FIFA (izquierda) (Pelechrinis & Winston, 2019)

Evidentemente el valor de un jugador conlleva más variables que solo su aportación en el campo, como su visibilidad o margen de mejora, sin embargo, este estudio pretende ofrecer un marco de referencia para cuantificar el valor individual de los jugadores estudiando su aportación en los partidos. Así los autores muestran el potencial que existe para las organizaciones deportivas el análisis de datos, tanto en el aspecto deportivo como en el aspecto empresarial de estas.

4.2 Predicción de los puntos de liga

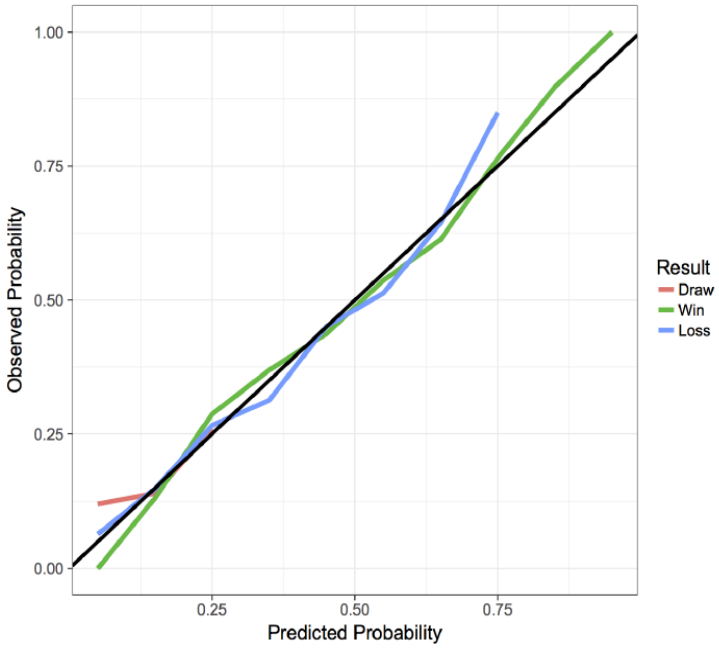
Es importante destacar que el trabajo de Pelechrinis y Winston no está enfocado a predecir el resultado de un encuentro a partir de los puntos FIFA de los jugadores. Es obvio que la cantidad de variables que afectan al resultado concreto de un partido no quedan capturadas con las puntuaciones FIFA de sus jugadores, por lo que el modelado del resultado mediante la distribución de Skellam, que a su vez depende de dos variables Poisson que modelan los goles de cada contrincante sólo teniendo en cuenta la diferencia de puntos FIFA por cada una de las 4 líneas de la configuración básica, sólo puede aspirar a estimar una probabilidad de ganar, perder o empatar. En el propio artículo los autores validan el modelo de regresión de la distribución Skellam utilizando lo que ellos denominan la curva de calibración, también denominado diagrama de confianza (Weisheimer & Palmer, 2014), que visualiza la confianza estadística de un modelo de predicción. Está definido para predicciones binarias y pone en correspondencia las probabilidades de predicción de un evento con las frecuencias de observación de dicho evento. Así, por ejemplo, para el subconjunto de partidos que tienen una probabilidad entre 0.8 y 0.9 de que gane el equipo local ($P(Z>0) \in [0.8, 0.9]$) se calcula la frecuencia del evento "gana el equipo de casa". Si esta frecuencia está entre 0.8 y 0.9 entonces tendremos un sistema de predicción fiable. Si por el contrario la frecuencia es 1, es decir, en todos los casos ganó el de casa, nuestro predictor infraestima las probabilidades en ese intervalo. Si, por el contrario, la frecuencia es inferior a 0.6, tendremos un predictor que sobreestima las probabilidades del evento de ganar en

casa. Por tanto, cuanto más pegada a la diagonal esté la curva, más confiable es el estimador. La figura 6 muestra el diagrama de confianza (extraído del artículo) para los eventos del equipo de casa: ganar, perder y empatar. Cómo se puede observar, las tres curvas están muy cerca de la diagonal por lo que se puede decir que el modelo de predicción de P(Z) es bastante confiable.

La diferencia de longitud de las curvas nos está diciendo que es más probable que el sistema asigne probabilidades altas al evento "gana equipo local" que al evento "pierde equipo local" y que el evento "empatan" tiene probabilidades estimadas siempre por debajo del 30%, lo cual es bastante ajustado a la realidad.

A la vista de estas curvas se me ocurrió probar qué potencial tiene este modelo para estimar realmente las probabilidades de estos eventos para una temporada no vista. Los autores hicieron una partición de datos estándar, un subconjunto de entrenamiento con el 80% de los registros (partidos) y un conjunto de test con el 20% restante. Pero, como también es habitual, aleatorizaron los datos antes de hacer la partición para asegurarse que los estadísticos de ambos conjuntos fueran parecidos. Esta práctica, lógica desde un punto de vista de estimación estadística, no es tan útil desde un punto de vista práctico. Si tenemos que tomar una decisión sobre el fichaje de un jugador para la siguiente temporada, a partir de su valor *eLPAR* estimado a partir de temporadas anteriores, me gustaría saber la fiabilidad de la estimación con datos que quizás no tienen exactamente la misma distribución estadística que tenían los de las temporadas anteriores.

Figura 6: Diagrama de confianza del modelo Skellam con partición aleatoria (Pelechrinis & Winston, 2019)



Experimento 1: evaluación del modelo de predicción de P(Z) en una temporada no vista durante el entrenamiento

En lugar de hacer una partición de los datos aleatoria, dividí los datos por temporadas, cogiendo las temporadas desde la 2008-2009 hasta la 2014-2015 para entrenamiento y la temporada 2015-2016 para test, que además se corresponde más o menos con el 15% de los datos. Tanto los datos como el código en R para hacer la estimación y prueba del modelo

están en el repositorio del artículo en GitHub[28]. Después de entrenar el modelo con menos temporadas, aunque un número de partidos incluso un poco superior (el porcentaje de partidos de las temporadas 2008 a 2015 es del 86,37%), se obtuvieron unos coeficientes ligeramente diferentes. Se puede ver cómo ha bajado la influencia del portero con relación a los coeficientes del artículo y además con una estimación no estadísticamente significativa.

Coefficients	**Estimate**	**Std. Error**	**z value**	**Pr(z)**
Equipo casa intercept	0.36708078	0.01170093	31.3719	< 2.2e-16 *
E.casa defensa	0.02003884	0.00302863	6.6165	3.679e-11 *
E.casa medio	0.02389181	0.00274243	8.7119	< 2.2e-16 *
E.casa delantero	0.00676768	0.00204518	3.3091	0.0009360 *
E.casa portero	-0.00081673	0.00155829	-0.5241	0.6001952
E. visitante interc.	0.05290593	0.01543904	3.4268	0.0006108 *
E. visitante defensa	-0.02838478	0.00390488	-7.2691	3.620e-13 *
E. visitante medio	-0.01859191	0.00354653	-5.2423	1.586e-07 *
E. visitante delantero	-0.01015414	0.00260212	-3.9022	9.530e-05 *
E. visitante portero	-0.00355430	0.00199231	-1.7840	0.0744214 .

Utilizando estos coeficientes para el nuevo modelo de Skellam, lo probé sobre la temporada 2015-2016. El diagrama de confianza en la predicción de este modelo se puede ver en la figura 7. Podemos decir que el modelo sigue siendo bastante fiable porque para los tres eventos binarios del equipo de casa: ganar, perder o empatar, las curvas siguen estando muy pegadas a la diagonal. Sin embargo, sorprende que el modelo no es capaz de asignar probabilidades tan altas de ganar en casa cuando se entrena sin tener en cuenta la última

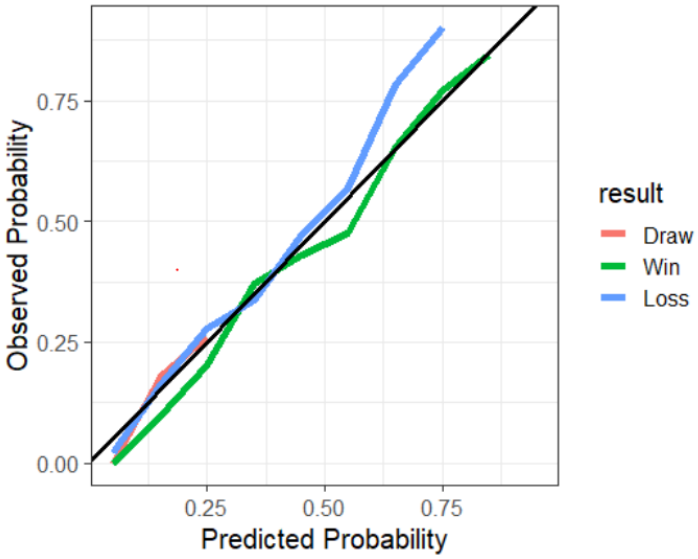


Figura 7: Diagrama de confianza del modelo Skellam con partición por temporadas (Elaboración propia)

[28] https://github.com/kpelechrinis/eLPAR-soccer

temporada (la línea roja no llega a 0.9, cuando antes pasaba holgadamente), aunque sigue siendo mayor que la probabilidad de perder en casa, y ésta mayor que la de empatar. Además, la estimación de la probabilidad de perder es inferior a la ocurrencia real en la temporada 2015-2016 (línea azul despegándose más de la diagonal hacia arriba). Con todo, se puede indicar que, aunque el modelo difiere del estimado con todas las temporadas, aún parece bastante apropiado para hacer estimaciones de probabilidades a futuro.

Pero ¿qué nos dice esto sobre la probabilidad de predecir el resultado de un partido concreto? Para comprobarlo realicé una serie de experimentos usando WEKA.

Experimento 2: predecir el resultado de un partido a partir de la diferencia de puntos FIFA.

Para un partido determinado, y conociendo las alineaciones de antemano, se puede tratar de predecir el resultado del encuentro usando los puntos FIFA promedio para cada línea (Portero, Defensa, Centrocampista, Delantero) en cada equipo. En este primer experimento no voy a usar el modelo Skellam, sino que voy a probar diferentes clasificadores.

Modelo 1 de regresión logística multinomial (partición aleatoria 85-15):

```
Correctly Classified Instances        1440               53.3729 %
Incorrectly Classified Instances      1258               46.6271 %
Kappa statistic                          0.2049
Mean absolute error                      0.3904
Root mean squared error                  0.4392
Relative absolute error                 91.363  %
Root relative squared error             95.1271 %
Total Number of Instances             2698

=== Detailed Accuracy By Class ===

                 TP Rate  FP Rate  Precision  Recall   F-Measure  MCC      ROC Area  PRC Area  Class
                 0,833    0,599    0,549      0,833    0,662      0,257    0,704     0,665     H
                 0,000    0,000    ?          0,000    ?          ?        0,563     0,283     D
                 0,508    0,206    0,497      0,508    0,503      0,301    0,727     0,525     A
Weighted Avg.    0,534    0,338    ?          0,534    ?          ?        0,676     0,530

=== Confusion Matrix ===

    a    b    c   <-- classified as
 1048    0  210 |    a = H
  483    0  186 |    b = D
  379    0  392 |    c = A
```

Como era de esperar, la capacidad de predicción del resultado de un partido es baja, aunque es mejor que una decisión aleatoria entre ganar, perder o empatar, qué sería de un 33%. Si tenemos en cuenta que la experiencia dice que el equipo que juega en casa tiene más probabilidades de ganar, y que empatar tiene una probabilidad más baja que los otros eventos, podríamos tener una probabilidad de acierto mayor que el 33%, pero vemos que usando las puntuaciones FIFA podemos llegar hasta el 53,37%. También es interesante observar que el modelo nunca decide el evento “empate”, lo que impide obtener las métricas promedio sobre los 3 eventos. En el conjunto de datos se ha respetado la distribución original de partidos con victorias en casa, derrotas y empates, igual que hicieron los autores del artículo al modelar la distribución de Skellam. Lógicamente la

distribución original (al margen de que faltan muchos registros de partidos) respeta la probabilidad a priori de los eventos ganar, perder y empatar. El evento empatar es el menos probable y el más difícil de predecir (como veremos en breve). Los modelos que no son muy precisos, como ocurre en este problema que tiene una alta componente de aleatoriedad (en caso contrario todos seríamos millonarios jugando a la quiniela), tienden a despreciar el evento menos predecible y con menor número de datos.

Hice la prueba de balancear el número de registros por clase y reentrenar, para comprobar que la clase empate era seleccionada por el clasificador.

Modelo 2 de regresión logística multinomial (partición aleatoria 85-15) pero con un balanceo previo del número de registros de cada clase (ganar, perder, empatar):

```
Correctly Classified Instances        1278.0146          47.5231 %
Incorrectly Classified Instances      1411.2328          52.4769 %
Kappa statistic                          0.2122
Mean absolute error                      0.4115
Root mean squared error                  0.4516
Relative absolute error                 92.5965 %
Root relative squared error             95.7944 %
Total Number of Instances             2689.2475

=== Detailed Accuracy By Class ===

                 TP Rate  FP Rate  Precision  Recall   F-Measure  MCC      ROC Area  PRC Area  Class
                 0,577    0,288    0,506      0,577    0,539      0,281    0,700     0,540     H
                 0,214    0,165    0,389      0,214    0,276      0,059    0,569     0,371     D
                 0,632    0,334    0,483      0,632    0,548      0,283    0,711     0,554     A
Weighted Avg.    0,475    0,263    0,460      0,475    0,455      0,208    0,660     0,489

=== Confusion Matrix ===

   a      b      c      <-- classified as
 526.04 153.61 231.86 |      a = H
 329.04 189.73 368.84 |      b = D
 183.57 144.31 562.24 |      c = A
```

Como era de esperar, ahora el clasificador asigna decisiones de empate, pero el resultado global de registros correctamente clasificados es muy inferior (47,52%) ya que se ha perdido toda la información de probabilidad a priori y, por el Teorema de Bayes, para tener una estimación de la probabilidad a posteriori de cada clase nos hace falta la probabilidad a priori. Además, podemos ver que, con el mismo número de muestras, la clase “empate” es la que tiene peor valor F1 (0.276) con gran diferencia respecto a ganar (0.539) o perder (0.548), con lo que está claro que ese evento es más difícil de predecir. Pensemos, en este sentido, que nuestras 8 variables independientes son puntos FIFA de las 4 líneas (portero, defensa, centrocampista, delantero) del equipo de casa y del equipo visitante, así que es fácilmente esperable que el clasificador tenga relativamente fácil distinguir el patrón de puntos altos de FIFA en el equipo de casa y puntos bajos en el visitante, para el evento “ganar en casa”, lo contrario para el evento “perder en casa”, pero para el evento empatar tendrá una mezcla difícil de asociar a un patrón bien determinado.

Unos resultados de clasificación de las 3 clases tan bajos hacen pensar que las muestras no están bien separadas en el espacio de los atributos (en este caso 8 dimensiones). Para cerciorarme hago un análisis de clustering con K- means y la medida silhouette, con la esperanza de que el máximo fuese para 3 clústeres, pero el máximo sale con 2 clústeres, lo que ya indica que todas las muestras están bastante mezcladas. Aun así, hago un clustering de 3 grupos y una asignación a posteriori de cada muestra de la clase "ganar en casa (H)", de "perder en casa (A)" y de "empatar (D)" al centroide más cercano (0,1,2) y se obtienen las asignaciones del cuadro siguiente. La matriz de asignación de clases a clústeres ya indica una gran mezcla en el espacio de dimensión 8. La clase H se asigna al clúster 2 por ser el que más muestras de esa clase tiene asignadas, la clase A al clúster 1, ya con una mezcla bastante mayor, y el clúster que queda se asigna por descarte a la clase D, donde ya se observa que ni siquiera es el clúster con la mayor asignación de "empates". Se confirma también así que la clase "empatar" es la que se distribuye de manera más uniforme entre las otras dos, por tanto, más difícil de construir cualquier clasificador para ella medianamente decente. Con esta decisión por "mínima distancia al clúster" tenemos un clasificador que falla el 57,33% de las muestras, o lo que es lo mismo, solo un 42,66% de acierto. Es más, si hubiéramos diseñado un clasificador de mínima distancia al clúster no podríamos haber usado todas las muestras para el cálculo de cada clúster, sino que tendríamos que haber dejado un subconjunto para test, y el resultado computarlo para ese conjunto, lo que daría resultados aún peores.

Análisis clustering con K=3 y asignación de las clases del conjunto total de muestras

```
=== Model and evaluation on training set ===

Clustered Instances

0       6733 ( 34%)
1       5328 ( 27%)
2       7733 ( 39%)


Class attribute: RESULT
Classes to Clusters:

    0    1    2  <-- assigned to cluster
 2923 1712 4471 | H
 1741 1383 1849 | D
 2069 2233 1413 | A

Cluster 0 <-- D
Cluster 1 <-- A
Cluster 2 <-- H

Incorrectly clustered instances :       11349.0  57.3356 %
```

De ahora en adelante todos los experimentos se harán con la distribución original.

Un experimento más realista y práctico debería hacerse con una división del conjunto de datos por temporadas, como se comentó al mostrar el diagrama de confianza con el modelo entrenado con todas las temporadas hasta la 2014-2015 y probado sobre la temporada 2015-2016.

Usando de nuevo el modelo de regresión logística multinomial, los resultados (ver abajo) muestran claramente que la predicción del modelo sobre partidos de una temporada no usada en entrenamiento ofrece unos resultados solo un poco inferiores (50,63%) respecto a la partición aleatoria, lo cual es lógico porque es muy difícil que las temporadas mantengan los mismos estadísticos entre sí. Con este resultado podemos concluir que, en una aplicación real de decisión a futuro, los puntos FIFA de cada línea son útiles para predecir el resultado del encuentro por encima de la decisión aleatoria.

Modelo 3 de regresión logística multinomial (partición de entrenamiento con las temporadas 2008 a 2015 y de test con la temporada 2015-2016, no vista en entrenamiento):

```
Correctly Classified Instances        1366               50.6301 %
Incorrectly Classified Instances      1332               49.3699 %
Kappa statistic                          0.1683
Mean absolute error                      0.3994
Root mean squared error                  0.4475
Relative absolute error                 92.8965 %
Root relative squared error             96.2146 %
Total Number of Instances             2698

=== Detailed Accuracy By Class ===

               TP Rate  FP Rate  Precision  Recall   F-Measure  MCC      ROC Area  PRC Area  Class
               0,865    0,656    0,509      0,865    0,641      0,239    0,682     0,628     H
               0,000    0,000    ?          0,000    ?          ?        0,549     0,283     D
               0,410    0,182    0,498      0,410    0,450      0,242    0,690     0,502     A
Weighted Avg.  0,506    0,345    ?          0,506    ?          ?        0,651     0,502

=== Confusion Matrix ===

    a    b    c   <-- classified as
 1028    0  160 |    a = H
  505    0  181 |    b = D
  486    0  338 |    c = A
```

Para salir de dudas sobre si un modelo predictivo más complejo pudiera ofrecer mejores resultados entrené una red neuronal de tipo perceptrón multicapa (MLP) con una capa oculta y un número variable de nodos en la capa oculta. Muestro a continuación los resultados con el modelo que usa 10 nodos en la capa oculta. El resultado de clasificación correcta es prácticamente el mismo (50,26%) que con el modelo 3 de regresión logística multinomial, con lo que concluimos que no compensa utilizar un modelo con muchos más parámetros.

Modelo 4 MLP con una capa oculta y 10 nodos (partición de test con la temporada 2015-2016, no vista en entrenamiento).

```
Correctly Classified Instances        1356               50.2595 %
Incorrectly Classified Instances      1342               49.7405 %
Kappa statistic                          0.1616
Mean absolute error                      0.3998
Root mean squared error                  0.4472
Relative absolute error                 92.9893 %
Root relative squared error             96.1338 %
Total Number of Instances             2698

=== Detailed Accuracy By Class ===

                 TP Rate  FP Rate  Precision  Recall   F-Measure  MCC      ROC Area  PRC Area  Class
                 0,862    0,664    0,505      0,862    0,637      0,228    0,684     0,626     H
                 0,000    0,000    ?          0,000    ?          ?        0,560     0,289     D
                 0,403    0,181    0,494      0,403    0,444      0,236    0,690     0,502     A
Weighted Avg.    0,503    0,348    ?          0,503    ?          ?        0,654     0,503

=== Confusion Matrix ===

    a    b    c   <-- classified as
 1024    0  164 |    a = H
  510    0  176 |    b = D
  492    0  332 |    c = A
```

Experimento 3: predecir el resultado de un partido a partir de la estimación de la función de densidad de probabilidad de la diferencia de goles.

En este experimento mi intención es comprobar la utilidad de la estimación de la densidad de probabilidad de la diferencia de goles que nos ofrece la distribución de Skellam que vimos en la sección anterior. ¿Habrá alguna diferencia en la estimación del evento "ganar en casa" si en lugar de usar los puntos FIFA directamente, uso el modelo de Skellam que estima la probabilidad de ganar, perder y empatar?

Para ello vuelvo a entrenar clasificadores multiclase para los 3 eventos posibles, teniendo como variables independientes simplemente las probabilidades estimadas P(Z>0), P(Z<0) y P(Z=0). Entrenando el modelo 5 de regresión logística multinomial con la partición aleatoria 85-15 sobre todas las temporadas se obtienen unos resultados prácticamente iguales a usar directamente los puntos FIFA (modelo 1), con lo que da la sensación de que no ganamos nada con el modelo Skellam si nuestro objetivo es predecir el resultado del encuentro.

Con la intención de comprobar si el modelo de Skellam tiene una capacidad predictiva mejor que el modelo directo de regresión logística sobre los puntos FIFA, sobre datos de otra temporada tengo que usar el modelo Skellam entrenado sin usar la última temporada, como hice para calcular el diagrama de confianza al principio de esta sección. De nuevo en este caso los resultados del modelo 6 son muy similares a los del modelo 3, con lo que concluimos que no compensa usar el modelo de Skellam para tomar la decisión sobre el resultado de un partido a partir de los puntos FIFA.

Modelo 5 de regresión logística multinomial con la partición aleatoria 85-15 sobre la estimación P(Z) Skelam del artículo original

```
Correctly Classified Instances        1451               53.7806 %
Incorrectly Classified Instances      1247               46.2194 %
Kappa statistic                          0.2115
Mean absolute error                      0.3901
Root mean squared error                  0.439
Relative absolute error                 91.3099 %
Root relative squared error             95.0825 %
Total Number of Instances             2698

=== Detailed Accuracy By Class ===

                 TP Rate  FP Rate  Precision  Recall   F-Measure  MCC      ROC Area  PRC Area  Class
                 0,843    0,593    0,554      0,843    0,668      0,274    0,706     0,663     H
                 0,000    0,000    ?          0,000    ?          ?        0,563     0,283     D
                 0,507    0,204    0,499      0,507    0,503      0,302    0,729     0,527     A
Weighted Avg.    0,538    0,335    ?          0,538    ?          ?        0,677     0,530

=== Confusion Matrix ===

    a    b    c   <-- classified as
 1060    0  198 |    a = H
  474    0  195 |    b = D
  380    0  391 |    c = A
```

Modelo 6 de regresión logística multinomial usando la estimación P(Z) Skellam reentrenada con una temporada menos. Estimación con la temporada 2015-2016:

```
=== Summary ===

Correctly Classified Instances        1355               50.2224 %
Incorrectly Classified Instances      1343               49.7776 %
Kappa statistic                          0.1619
Mean absolute error                      0.3992
Root mean squared error                  0.4475
Relative absolute error                 92.8579 %
Root relative squared error             96.2016 %
Total Number of Instances             2698

=== Detailed Accuracy By Class ===

                 TP Rate  FP Rate  Precision  Recall   F-Measure  MCC      ROC Area  PRC Area  Class
                 0,860    0,656    0,508      0,860    0,639      0,233    0,683     0,629     H
                 0,000    0,000    ?          0,000    ?          ?        0,541     0,271     D
                 0,404    0,188    0,486      0,404    0,441      0,229    0,691     0,504     A
Weighted Avg.    0,502    0,346    ?          0,502    ?          ?        0,649     0,500

=== Confusion Matrix ===

    a    b    c   <-- classified as
 1022    0  166 |    a = H
  500    0  186 |    b = D
  491    0  333 |    c = A
```

El artículo de Pelechrinis y Winston define eLPAR para relacionar un incremento en puntos FIFA de un jugador con un incremento en la probabilidad de que este jugador sume puntos de liga. Aunque el modelo de Skellam entrenado tiene una capacidad predictiva de las variables P(Z>0), P(Z<0) y P(Z=0) muy alta, como se vió en los diagramas de confianza, eso no se traduce en una probabilidad de acierto en el resultado del partido excepto para valores altos de dichas variables. Es decir, sólo podremos tener confianza en decidir que el equipo de casa ganará el partido si el valor de P(Z>0) es cercano a 1, y lo mismo al decidir que pierde o empata. Como se vió al discutir el diagrama, los valores de P(Z=0) nunca están por encima de 0.3, por lo que nunca podríamos usar este modelo para decidir con confianza que un partido va a acabar en empate. Como se vió en los experimentos, ninguno de los clasificadores asignó nunca la decisión de "empate", tanto si usamos el modelo de regresión logística multinomial generado con las variables independientes P(Z>0), P(Z<0) y P(Z=0), como si lo usamos directamente con los puntos FIFA.

La información contenida en los puntos FIFA promediada por cada línea de juego no parece suficiente para decidir entre los 3 posibles resultados de un partido con un porcentaje mayor del 53%. Sin embargo, está claro que contienen información predictiva por encima de la elección aleatoria, con lo que podríamos preguntarnos si la evidencia acumulada en todos los partidos de la temporada a través de la suma de diferencias de puntos FIFA podría ser un buen predictor de los puntos totales de liga conseguidos. Así podemos definir una nueva medida como la Diferencia Promedio de Puntos FIFA:

$$DPPF(e,t,l) = \frac{1}{N_{(e,t)}} * (\sum_{i=1} X_{(e,t,l,i)})$$

donde: *e* es el índice del equipo, *t* el de la temporada, *l* el de la línea (portero, defensa, centrocampista y delantero), N(*e,t*) el número de partidos por temporada para el equipo *e*, y X(*e,t,l,i*) la correspondiente diferencia de puntos FIFA en un partido concreto entre el equipo *e*, jugando en casa y cualquier otro equipo *i* de la liga.

Experimento 4: predecir los puntos de liga de equipo en una temporada a partir de la diferencia de puntos FIFA promedio contra todos sus contrincantes.

A partir de la definición de la variable DPPF(*e,t,l*) podemos calcular modelos de regresión para el número de puntos de liga que se pueden conseguir. El objetivo práctico de este experimento es determinar si con los datos temporadas anteriores, un club puede tomar decisiones sobre qué cambios debería hacer en el equipo. Obviamente no es posible prever qué alineaciones pondrá cada equipo contrincante antes de que lo haga, lo que cambiaría la variable X(*e,l,i*), pero por el histórico de cada equipo y los fichajes que puedan estar sonando, es posible diseñar varios escenarios posibles y tomar decisiones en función de la probabilidad de ocurrencia de cada uno de ellos.

Para empezar con el modelado más sencillo, primero calculo las variables DPPF(*e,t,l*) para todas las temporadas, equipos y líneas. Al hacer el promediado salta a la luz que en cada temporada no hay el mismo número de encuentros disponibles y que éste además varía con cada equipo. Es decir, faltan muchos registros. Aun así, realizo una regresión lineal simple con WEKA y obtengo los siguientes resultados:

Modelo 7 de regresión lineal para calcular los puntos de Liga a partir de las 4 variables DPPF(e,t,l):

```
Linear Regression Model

Home.team.final.points.season =

      1.5028 * Diff.D.Rating +
      1.1119 * Diff.M.Rating +
      0.463  * Diff.O.Rating +
      0.1368 * Diff.GK.Rating +
     49.0051

Time taken to build model: 0.13 seconds

=== Evaluation on test split ===

Time taken to test model on test split: 0.01 seconds

=== Summary ===

Correlation coefficient                  0.8116
Kendall's tau                            0.5418
Spearman's rho                           0.7242
Mean absolute error                      8.1239
Root mean squared error                  9.8963
Relative absolute error                 62.819  %
Root relative squared error             59.892  %
Total Number of Instances              162
```

Se puede observar como el modelo, entrenado con las variables en las temporadas 2008 a 2015, es capaz de predecir los puntos de Liga en la temporada 2015-2016, no vista en entrenamiento con un error absoluto medio de 8.12 puntos y una desviación típica de 9.89 puntos. Teniendo en cuenta que los puntos totales en la temporada 2015-2016 variaron entre [17,96], estamos hablando de un error medio de estimación sobre el margen dinámico, de un 10% aunque con una desviación también de un 11%. Tanto el coeficiente de Kendall como el de Spearman nos dicen que hay una buena correlación entre el orden de los puntos predichos y los reales, lo cual es importante si queremos usar este modelo como un predictor del incremento relativo de puntos de Liga al modificar a un jugador. Resulta interesante observar que el peso de los coeficientes de regresión está muy diferenciado para cada una de las cuatro líneas del campo, siendo la defensa el más importante, seguido del medio campo, delantera y finalmente portero. Llama la atención el poco peso de los delanteros con relación a los defensas. Según este modelo, un cambio estratégico de un defensa con más puntuación FIFA que la mayoría de los defensas de los equipos contrarios provocaría una diferencia positiva de mayor valor que el mismo cambio en un delantero. También cabe destacar que para este modelado disponemos de muchos menos datos al estar agregando todos los encuentros de cada temporada para cada equipo. Así, si antes teníamos 19794 observaciones (partidos) para las 8 temporadas, ahora tenemos 1378 observaciones (que ahora son equipos por temporada), de los que 188 se corresponden a la temporada 2015-2016 y 1190 al resto.

Como se hizo anteriormente para el clasificador por partido, voy a integrar el modelo de Skellam y comprobar si con eso mejora la estimación de los puntos de liga. Para ello uso como variables independientes, en lugar de las 4 variables DPPF(*e,t,l*), el promedio de las variables P(Z>0), P(Z<0) y P(Z=0) para cada equipo y cada temporada. Probando de nuevo con varios tipos de regresores en WEKA, el que mejor resultado ofrece sobre los datos no vistos de la temporada 2015-2016 es un MLP con 4 nodos en la capa oculta.

Modelo 8 de regresión no lineal usando un MLP con una capa oculta y 4 nodos, para calcular los puntos de Liga a partir de las 3 variables de Skellam P(Z>0), P(Z<0), P(Z=0):

Correlation coefficient	0.7769
Kendall's tau	0.4936
Spearman's rho	0.6652
Mean absolute error	7.8391
Root mean squared error	10.027
Relative absolute error	64.0267 %
Root relative squared error	63.0638 %
Total Number of Instances	188

Se puede observar que este modelo tiene prestaciones similares (menor MAE pero mayor desviación típica) al que estima directamente los puntos a partir de las variables DPPF(*e,t,l*) mediante regresión lineal, por tanto podemos concluir que el modelo de Skellam tampoco es especialmente útil para estimar los puntos de liga.

Experimento 5: predecir P(Z) a partir de la diferencia de puntos FIFA pero usando un MLP en lugar de la distribución de Skellam

Para terminar con este estudio me pareció interesante comprobar si el modelo MLP podría ser un buen predictor de P(Z), que es, al fin y al cabo, lo que se busca con el modelo de distribución de Skellam propuesto en el artículo estudiado. Dado que no tenemos el valor exacto de P(Z) para cada partido con el que entrenar un MLP para regresión, me planteé entrenar un clasificador con las clases ganar, perder o empatar P(Z>0), PZ<0) y P(Z=0), respectivamente. En este caso me interesaba comprobar si el MLP podría estimar las probabilidades P(Z), no tanto el resultado exacto del partido, que ya vimos que no supera el 53%. Es decir ¿puedo conseguir con aprendizaje automático un estimador de P(Z) similar al de Skellam usando los mismos datos de entrada (diferencia de puntos FIFA en cada línea) y la misma partición 80-20 de todas las temporadas? Para ello me propongo generar el diagrama de confianza para los eventos ganar en casa, perder en casa y empatar, como se hizo en las Figuras 6 y 7, pero no usando el código R de los autores con la distribución de Skellam sino entrenando un MLP con WEKA a partir de los mismos datos.

El primer problema que me encontré es que al testear un clasificador MLP multiclase (3 clases), WEKA solo devuelve la salida con mayor valor. Así, por ejemplo, si el resultado real de un partido fue ganar en casa, pero la salida del MLP fue 0.6 para la clase perder en casa, éste es el valor que sale, junto con la información para la matriz de confusión de que ha habido un error al predecir perder cuando debía haber predicho ganar. Pero a mí me interesa saber la probabilidad predicha para cada clase en cada partido para generar el diagrama de confianza. Por lo tanto, en lugar de entrenar un clasificador multiclase entreno 3 clasificadores binarios: ganar contra no ganar, perder contra no perder, y empatar contra no empatar. De esta forma, si el clasificador entrenado para la clase ganar saca un 0.6 en la clase perder y WEKA devuelve ese valor, ya sé que la probabilidad estimada en la clase ganar

es 0.4. Cada clasificador binario se diseña con un MLP de 5 nodos ocultos y una salida, y una partición 80-20 aleatoria como en el artículo original.

Así, con un poco de trabajo en Excel para ordenar las predicciones de los 3 MLPs y generar los diagramas de confianza de manera similar a las Figuras 6 y 7, tenemos lo siguiente:

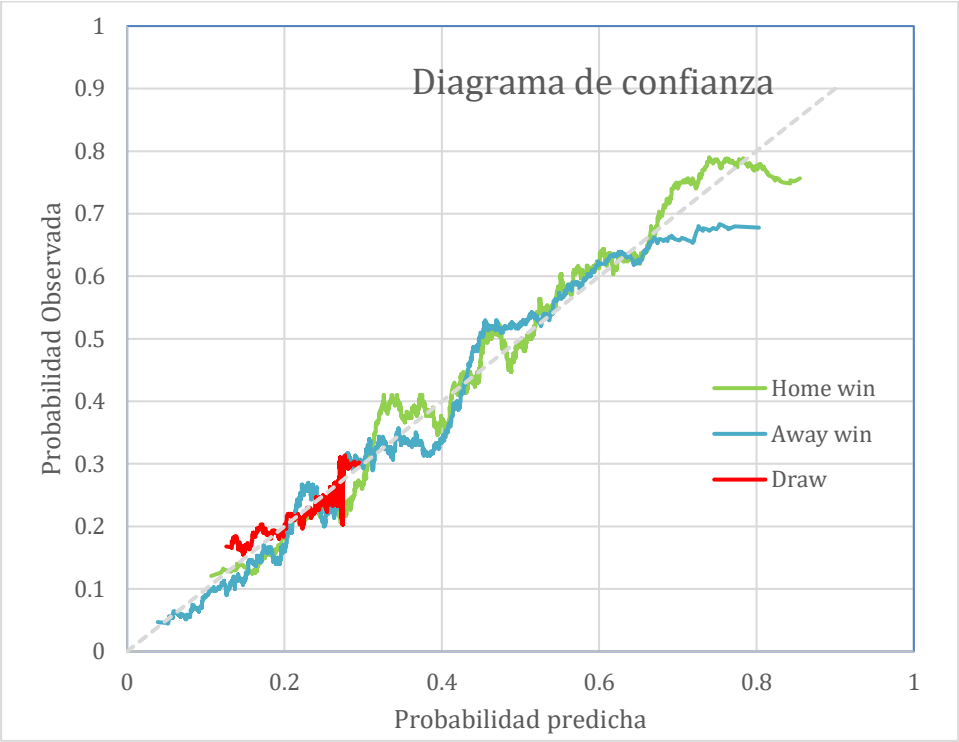


Figura 8: Diagrama de confianza del modelo MLP para predecir P(Z) con partición aleatoria 80-20 de todas las temporadas (Elaboración propia).

Nota: la gráfica se ve más "ruidosa" que las de las figuras 6 y 7 porque se han ordenado los 3959 partidos según su probabilidad predicha y se ha centrado en cada partido una "ventana" de 300 partidos de probabilidades predichas similares para calcular la probabilidad observada de ese conjunto de 300. Por tanto, la gráfica tiene muchos más puntos que las de dichas figuras calculadas con el software de los autores, que dividía en intervalos de probabilidad 0.1.

Se puede observar como la predicción de P(Z) para cada uno de los 3 eventos se comporta de manera similar a la predicción con el modelo Skellam, fallando un poco más en probabilidades altas. De nuevo se observa, como ya se comentó en el Experimento 2, que el clasificador para el evento empate nunca elige ese resultado ya que las probabilidades predichas nunca llegan a 0.5. Con este experimento comprobamos que un clasificador entrenado con técnicas de aprendizaje automático, sin hacer ninguna suposición sobre la distribución de la diferencia de goles, y basándose únicamente en los datos, es capaz de generar un predictor de la probabilidad de ganar, perder o empatar muy similar al que usa un modelo estadístico preconcebido.

5. CONCLUSIONES Y LÍNEAS FUTURAS

Con este TFG se quería mostrar el potencial del análisis de datos para las organizaciones deportivas, incluyendo casos prácticos tanto en marketing como en decisiones deportivas. Revisando la literatura se descubrió que la cantidad de estudios realizados a nivel deportivo era abrumadora, con un gran porcentaje utilizando técnicas o conceptos demasiado complejos para el alcance de este TFG. Afortunadamente se encontraron diversos artículos que no solo se ajustaban en cuanto al nivel técnico, sino que también eran reproducibles y modificables ya que aportaban datos y técnicas utilizadas. No hubo la misma suerte a nivel de márketing, ya que no se encontraron estudios con el suficiente detalle y reproducibilidad como para analizarlos en este TFG

Este TFG deja claro a través diversos ejemplos como el análisis de datos puede ofrecer no solo ayudas a la hora de tomar decisiones y optimizar recursos, sino que también puede ayudar a las organizaciones a encontrar nuevas vías de innovación. La analítica avanzada ofrece recursos de increíble potencial para cualquier departamento de márketing y no hay razón para que las organizaciones deportivas no se aprovechen de ello. Una posible línea futura para continuar este trabajo podría ser un estudio sobre que técnicas o que datos pueden utilizar las organizaciones deportivas para crear campañas de marketing más efectivas.Durante la investigación del análisis de datos en estas organizaciones y su historia, se hallaron acuerdos que están surgiendo entre compañías de servicios en la nube y ligas deportivas, las cuales quieren ser competitivas tanto a nivel nacional como internacional. Esta es prueba irrefutable de que el análisis de datos ha llegado para quedarse en el deporte y aquellas organizaciones que quieran mantenerse competitivas han de estar al día.

En cuanto al aspecto técnico de este trabajo se intentó mostrar la imagen más básica de la ciencia de datos debido a que el objetivo no era realizar ningún estudio complejo, sino mostrar un ejemplo para ayudar al lector a entender como una organización aprovecha los datos pata tomar decisiones deportivas. Esto fue posible gracias a los diversos trabajos comentados durante el TFG, además de los experimentos realizados en el capítulo 4. Estos experimentos no solo muestran cómo el análisis de datos puede ofrecer resultados de interés práctico para un equipo de fútbol, sino que sacan a la luz información relevante, como la importancia infravalorada de los defensas, el escaso papel diferencial de los porteros o la dificultad de predecir empates, cuando solo se tienen en cuenta un rating general de jugador, en este caso el de FIFA. Debido al limitado acceso a datos y la restricción de longitud del TFG, no se ha realizado ningún análisis más específico, como estudiar a quién debería fichar un equipo o qué tipo de jugadas utilizar contra un determinado contrincante, pero son ideas que surgieron al escribir este trabajo y posibles líneas futuras dignas de estudio.

No se incluye un anexo con datos ni código debido a la dimensión de estos, pero se invita al lector a solicitarlos escribiendo un email a 201601555@alu.comillas.edu

RERENCIAS Y NOTAS BIBLIOGRÁFICAS

Aalbers, B. & Haaren, J. V. (2019). Distinguishing Between Roles of Football Players in Play-by-Play Match Event Data. *Machine Learning and Data Mining for Sports Analytics Lecture Notes in Computer Science*, 31–41. doi: 10.1007/978-3-030-17274-9_3

a. AWS *What is cloud computing?* (n.d.) Acceso el 7 Marzo de 2020, desde: https://aws.amazon.com/what-is-cloud-computing/?nc1=h_ls

b. AWS *Going long on machine learning. How AWS and the NFL teamed up to transform a 100-year-old league.* (n.d.) Acceso el 7 Marzo de 2020, desde: https://pages.awscloud.com/rs/112-TZM-766/images/AWS_NFL_Interactive_eBook.pdf

Azure *What is cloud computing?* (n.d.) Acceso el 7 Marzo de 2020, desde: https://azure.microsoft.com/en-us/overview/what-is-cloud-computing/#benefits

Harrison, C. K., & Bukstein, S. (2017). *Sport business analytics: using data to increase revenue and improve operational efficiency*. Boca Raton: CRC Press, Taylor et Francis Group.

Brownlee, J. (2019, August 12) *Supervised and Unsupervised Machine Learning Algorithms*. Machine Learning Matery. Acceso el 18 Mayo de 2020, desde: https://machinelearningmastery.com/supervised-and-unsupervised-machine-learning-algorithms/

Brownlee, J. (2020, Abril 8) *4 Types of Classification Tasks in Machine Learning*. Machine Learning Matery. Acceso el 18 Mayo de 2020, desde: https://machinelearningmastery.com/types-of-classification-in-machine-learning/

Cortsen K. y Rascher D. (2018) *The Application of Sports Technology and Sports Data for Commercial Purposes*. Acceso el 7 Marzo de 2020, desde: https://www.intechopen.com/books/the-use-of-technology-in-sport-emerging-challenges/the-application-of-sports-technology-and-sports-data-for-commercial-purposes.

Cummmings, JC. (2019, Mayo 7) *Temporada 2002: el año que Billy Bean irrumpió con los analytics en el beisbol.* Supervivientes. Acceso el 7 de Marzo de 2020, desde: https://sobrevivientes.mx/deportes-y-ocio/temporada-2002-el-ano-que-billy-beane-irrumpio-con-los-analytics-en-el-beisbol/

Decroos, T., Bransen, L., Haaren, J. V., & Davis, J. (2019). Actions Speak Louder than Goals. *Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*. doi: 10.1145/3292500.3330758

Duch, J., Waitzman, J. S., & Amaral, L. A. N. (2010). Quantifying the Performance of Individual Players in a Team Activity. PLoS ONE, 5(6). doi: 10.1371/journal.pone.0010937

Felipe, J.L., García-Unanue, J., Viejo-Romero, D., Navandar, A., & Sánchez-Sánchez, J. (2019). *Validation of a video-based performance análisis system (Mediacoach®) to analyze the physical demands during matches in LaLiga.* Sensros, 19(19), 4113. https://doi.org/10.3390/s19194113

Ganguly, S. & Pedagagi, S. (2019) Tracking Data for Football Coaches and Analysts [video]. SportVU. Acceso el 7 de Marzo de 2020, desde: https://www.statsperform.com/team-performance/football/optical-tracking/

Gartner Glosario *Big Data* (n.d.) Acceso el 7 Marzo de 2020, desde https://www.gartner.com/en/information-technology/glossary/big-data.

Kumar, G. (2013). Machine Learning for Soccer Analytics. Universidad KU Leuven. doi: 10.13140/RG.2.1.4628.3761.

Liu Y., Schulte O., y Li C. (2019) Model Trees for Identifying Exceptional Players in the NHL and NBA Drafts. In: Brefeld U., Davis J., Van Haaren J., Zimmermann A. (eds) Machine Learning and Data Mining for Sports Analytics. MLSA 2018. Lecture Notes in Computer Science, vol 11330. Springer, Cham

Lindbergh, B & Sawchik, T (2019, Junio 3) How the Houston Astros disrupted player development to become the model MLB Franchise. The Ringer. Acceso el 7 de Marzo de 2020, desde: https://www.theringer.com/mlb/2019/6/3/18644512/mvp-machine-how-houston-astros-became-great-scouting

McHale, I. G., Scarf, P. A., & Folker, D. E. (2012). *On the Development of a Soccer Player Performance Rating System for the English Premier League*. Interfaces, 42(4), 339–351. doi: 10.1287/inte.1110.0589

McHale, I., & Scarf, P. (2007). *Modelling soccer matches using bivariate discrete distributions with general dependence structure*. Statistica Neerlandica, *61*(4), 432–445. doi: 10.1111/j.1467-9574.2007.00368.x

Menchén, M. (2018). Entre patrocinios y salarios: ¿Cómo ingresa, en qué gasta y cuánto gana cada club de LaLiga?. Acceso el 7 Marzo de 2020, desde https://www.palco23.com/clubes/entre-patrocinios-y-salarios-como-ingresa-en-que-gasta-y-cuanto-gana-cada-club-del-futbol-espanol.html

Muñoz Rosas, J.F. & Álvarez Verdejo, E. (2009). *Métodos de imputación para el tratamiento de datos faltantes: aplicación mediante R/Splus*. REVISTA DE MÉTODOS CUANTITATIVOS PARA LA ECONOMÍA Y LA EMPRESA, (7), 3–30. Acceso el 18 Mayo de 2020, desde http://www.upo.es/RevMetCuant/art25.pdf

Nawrocki, S. (2016, Mayo 10). *5 ways to leverage sports analytics to grow your fan base.* IBM Big Data and Analytics Hub. Acceso el 7 Marzo de 2020, desde https://www.ibmbigdatahub.com/blog/5-ways-leverage-sports-analytics-grow-your-fan-base

Olmeda, M (2020, Enero 15) *La tecnología que ayuda a los equipos a preparar los partidos.* ABC. Acceso el 7 Marzo de 2020, desde: https://www.abc.es/contentfactory/post/eslaliga/laliga-apuesta-por-la-tecnologia-para-hacerse-mas-competitiva/

Onlinecasinomonsters.com *Matthew Benham – A Successful Bloke in the World of Online Gambling.* (2017, Enero 28) Acceso el 7 Marzo de 2020, desde: https://onlinecasinomonsters.com/matthew-benham-successful-bloke-world-online-gambling/

Pons E, García-Calvo T, Resta R, Blanco H, López del Campo R, Díaz García J, et al. (2019) *A comparison of a GPS device and a multi-camera video technology during official soccer matches: Agreement between systems.* PLoS ONE 14(8): e0220729. https://doi.org/10.1371/journal.pone.0220729

Pelechrinis, K., & Winston, W. (2019). A Skellam Regression Model for Evaluating Soccer Positions. Acceso el 18 de Mayo de 2020, desde: https://arxiv.org/pdf/1807.07536.pdf

Ruczinski, I (n.d.) Chapter 10 Variable Selection. Johns Hopkins Bloomberg School of Public Health, Biostatistics. Acceso el 18 Mayo de 2020, desde: http://www.biostat.jhsph.edu/~iruczins/teaching/jf/ch10.pdf

Sabatés, R & Martín, A (2020, Febrero 4) *1.936 razones para fichar a Alcácer, En-Nesyri o R.D.T.* EL PAÍS. Acceso el 7 Marzo de 2020, desde: https://elpais.com/deportes/2020/01/31/es_laliga/1580475604_563489.html

Suzor, T. (2019, Agosto 27). *The future of the fan experiences at the US Open*. IBM Watson. Acceso el 7 Marzo de 2020, desde https://www.ibm.com/blogs/watson/2019/08/the-future-of-the-fan-experience-at-the-us-open/

Trademade Sports *How Brighton football club owner Tony Bloom got rich on sports betting* (n.d.) Acceso el 7 Marzo de 2020, desde : https://www.blog.tradematesports.com/jonas-gjelstad-professional-sports-bettor/2019/8/13/how-the-brighton-football-club-owner-tony-bloom-got-rich-on-sports-betting-10-people-who-got-rich-on-sports-betting

Troilo, M., Bouchet, B., Urban, T. L., & Sutton, W. A. (2016). Perception, reality, and the adoption of business analytics: Evidence from North American professional sport organizations. Omega, 59, 72–83.

Soler, A (2017, Septiembre 9). Statcast, una nueva herramienta para evaluar la defensa. Acceso el 7 de Marzo de 2020, desde: https://www.beisbolmlb.com/statcast-una-nueva-herramienta-para-evaluar-la-defensa/

Weisheimer, A. & Palmer, T. (2014). *On the reliability of seasonal climate forecast*s. Journal of the Royal Society Interface 11, 96 (2014), doi: 201311620

Woolner, K. (2007, Septiembre 28) *Introduction to VORP: Value Over Replacement Player*. Stathead. Acceso el 18 de Mayo de 2020, desde: https://web.archive.org/web/20070928064958/http://www.stathead.com/bbeng/woolner/vorpdescnew.htm