Trabajo Final de Grado

Ingeniería de las Tecnologías Industriales

Predicción de cambios en el valor de mercado de futbolistas profesionales usando técnicas de Machine Learning

Autor: Ignacio Peralta Fernández-Revuelta

Tutor: Alicia Robles Velasco

Dpto. Organización Industrial y Gestión de Empresas II

Sevilla, 2022

Trabajo Final de Grado
Ingeniería de las Tecnologías Industriales

Predicción de cambios en el valor de mercado de futbolistas profesionales usando técnicas de Machine Learning

Autor:

Ignacio Peralta Fernández-Revuelta

Tutor:

Alicia Robles Velasco

Colaborador Docente Invitado

Dpto. Organización Industrial y Gestión de Empresas II

Escuela Técnica Superior de Ingeniería

Universidad de Sevilla

Sevilla, 2022

Trabajo Final de Grado: Predicción de cambios en el valor de mercado de futbolistas profesionales usando técnicas de Machine Learning

Autor: Ignacio Peralta Fernández-Revuelta

Tutor: Alicia Robles Velasco

El tribunal nombrado para juzgar el Proyecto arriba indicado, compuesto por los siguientes miembros:

Presidente:

Vocales:

Secretario:

Acuerdan otorgarle la calificación de:

Sevilla, 2022

El Secretario del Tribunal

A mi abuelo

Agradecimientos

Parece mentira, pero con en estas líneas pongo fin a 4 (+1) años de carrera, y en ellas quisiera, brevemente, acordarme de aquellos que me han acompañado en este tiempo.

Para comenzar quisiera agradecer a mi familia por su apoyo infinito e incondicional. Si tuviera que detenerme a agradeceros por todo lo que hacéis y significáis para mí no terminaría nunca, por eso solo puedo deciros GRACIAS.

A mis amigos, que también sois familia. Gracias a todos por estos años increíbles donde a parte de estudiar mucho, nos lo hemos pasado genial. Sin vosotros no entiendo mi paso por la carrera, gracias por todos los momentos vividos y por los que están por venir.

Finalmente, quisiera acordarme de todos mis profesores, pero especialmente de Alicia Robles. Sin saberlo me propusiste hacer el TFG sobre uno de los temas que más me apasionan, y eso ha hecho este camino mucho más ameno. Gracias por tu ayuda en todo momento, pero sobre todo por tu trato y cercanía.

Ignacio Peralta Fernández-Revuelta

Sevilla, 2022

Resumen

Las cifras económicas que manejan los clubes de fútbol son cada vez superiores, arrastrando consigo mucho dinero con cada pequeño movimiento. Dentro de los movimientos que más dinero generan se encuentran las transacciones de futbolistas, es decir, el mercado de fichajes. Hasta hace algunos años, el valor de un futbolista era aquel que consideraba el club poseedor de sus derechos y al que estaba dispuesto a llegar el club comprador; si bien actualemente, como en cualquier negociación, sigue dependiendo de los límites establecidos entre vendedor y comprador, el valor de mercado de los futbolistas está muy estudiado, siendo prácticamente una cifra objetiva. Es por esto, con todo el dinero que hay en juego, que se busca desarrollar en este trabajo una herramienta que sirva de apoyo a la toma de decisiones de los clubes en el mercado, ya sea la de declinar ofertas, pedir más dinero o decantarse por la compra de un futbolista sobre otro. Se pretende que la toma de decisiones se base en un histórico de datos que, aplicando técnicas de Machine Learning prediga aumentos en los valores de mercado futuros. Para comenzar el trabajo se hizo uso de una base de datos generada a mano, en la que cada muestra incluía información básica del futbolista (por simplificación se escogieron únicamente delanteros centro de LaLiga) y la temporada, además de una serie de estadísticas relacionadas con todo tipo de acciones del juego. Finalmente, el desarrollo del trabajó se llevó a cabo en *Python*, gracias a las librerías *Pandas* y *Scikit-learn.*

Abstract

Football clubs are increasingly handling larger economic figures, bringing a lot of money with every single move. Footballer transfers, or the transfer market, are among those that generate most money. Until a few years ago, a footballer's value was determined by the club that owned him and by the purchasing club's willingness to pay for that player; Although now, as in any negotiation, markets are well studied, and prices for footballers can be very objectively determined. The project aims to develop a tool to assist clubs when making transfer market decisions, whether they're declining offers, asking for more money, or picking a player over another. It is intended to make decisions based on historical data that, using Machine Learning techniques, forecast market value increases in the future. For the first step in the project, a hand-generated database was created, where each sample contained basic information about the player (for simplification, only center forwards from LaLiga were chosen) and the season, along with a series of statistics relating to everything type of game actions. The work was finally developed in Python, using the Pandas and Scikit-learn libraries.

Índice

ÍNDICE DE TABLAS

ÍNDICE DE FIGURAS

1 INTRODUCCIÓN

Ni siquiera el día que me muera seré capaz de dejar el fútbol.

- Diego Armando Maradona -

El fútbol, el deporte rey, tan presente en la cultura popular de la sociedad española. Motor de pasiones y quebraderos de cabeza. Capaz de congregar a cientos de miles de personas de todas las clases, ideologías, sexos y razas en un mismo lugar; capaz de enloquecer al más cuerdo por un par de horas y de acallar al más locuaz algunas más. El fútbol para los verdaderos apasionados va mucho más allá de los 90 minutos reglamentarios y de ver a 22 personas corriendo detrás de un balón, el fútbol es el éxtasis de un gol, el abrazo con tu compañero de asiento y las horas de nervios y previa con los amigos.

Hablando de números, el alcance de La Liga, competición liguera de España, en la última temporada prepandémica con público en las gradas (2018/2019) fue de 14.812.356 espectadores entre primera y segunda división. Según informa la propia institución en su *newsletter* del mes de julio del 2019, la cifra de espectadores en vivo se vio aumentada en un 3,8% con respecto a la temporada anterior, siguiendo una tendencia de crecimiento ascendente desde la temporada 2013/2014, desde la cual se ha aumentado la cifra de espectadores en un 11,45% [1]. En la Figura 1-1 se refleja la evolución de la asistencia a los estadios en las últimas temporadas completas a las que pudo asistir el público, contando los partidos a LaLiga Santander, LaLiga Smartbank y los partidos de playoff de ascenso a primera división.

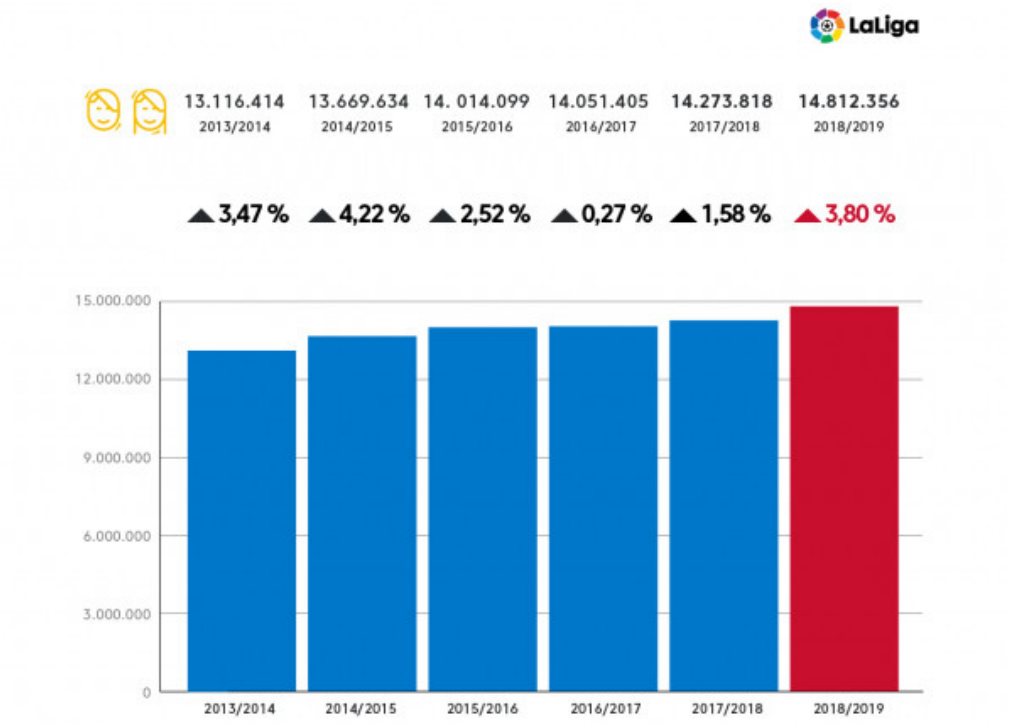


Figura 1-1. Evolución de la asistencia a los estadios (Fuente: LaLiga [1]).

En cuanto a los espectadores por televisión, según un estudio de LaLiga, la audiencia residencial aumentó un 7,85% entre la temporada 2019/2020 y la 2020/2021, aunque no supera los números anteriores al año 2019. A continuación, se presenta la evolución de las audiencias residenciales de LaLiga desde la temporada 2016/2017 hasta la 2020/2021.

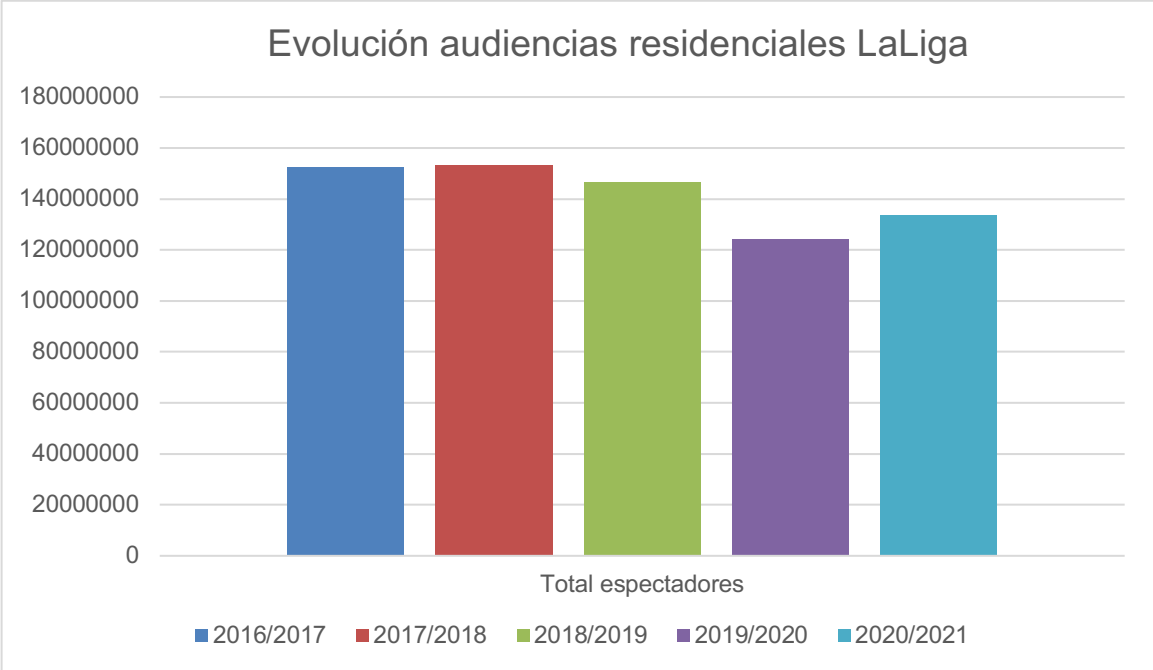


Figura 1-2. Evolución de las audiencias residenciales (Elaboración propia. Fuente ElConfidencial [2])

LaLiga, a pesar de las declaraciones de dirigentes que tratan de defender intentos de creación de competiciones paralelas, tiene un buen horizonte por delante. El mejor indicador de que le augura un futuro próspero a la competición es analizar su influencia e interés entre los más jóvenes. El mismo estudio de audiencia mentado con anterioridad segmenta por tramos de edad las audiencias residenciales, en él se puede observar como en la última temporada las audiencias entre los menores de 24 años crecieron un 9,4% frente al 7,85% de crecimiento de audiencia total. Además, el peso de este grupo de edad es cada vez más relevante, mientras que en la temporada 2016/2017 este grupo suponía el 6,4% de la audiencia residencial, en la temporada 2020/2021 alcanzó el 7,2%. Este peso es todavía mayor cuando se contabiliza la audiencia en el total de pantallas, y es que este grupo representa un 10,9% de los espectadores de LaLiga Santander (primera división del fútbol español) [2]. Se presenta a continuación, el gráfico que segmenta la evolución de las audiencias residenciales por franjas de edad.

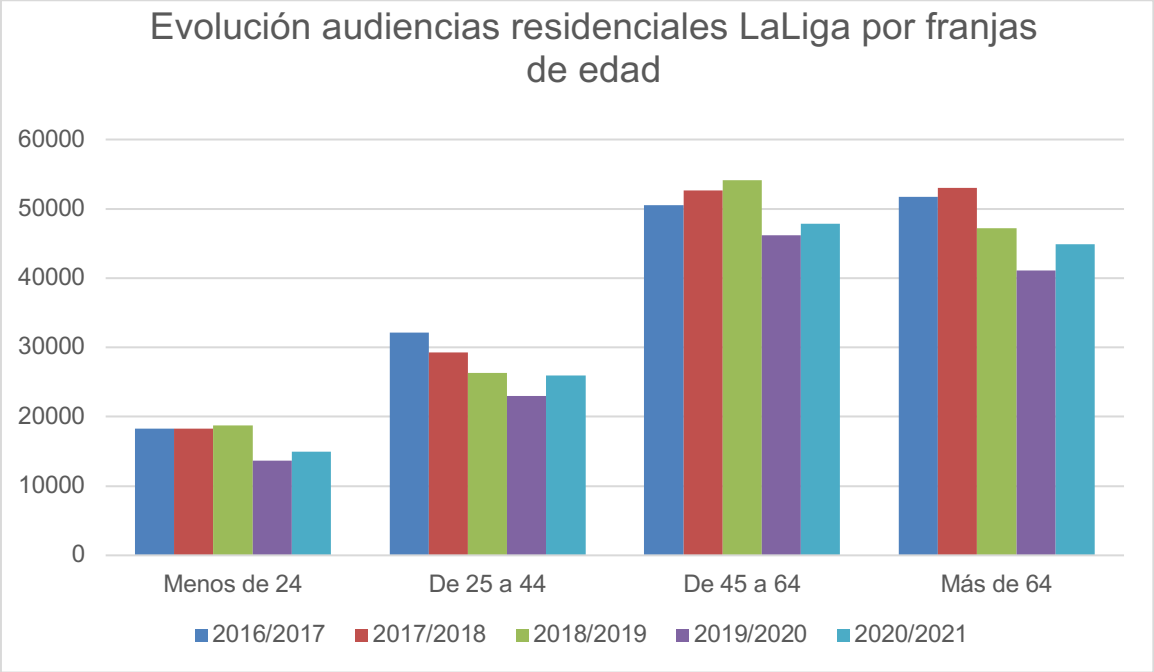


Figura 1-3.Evolución audiencias residenciales por franjas de edad (Elaboración propia. Fuente: ElConfidencial [2])

Una vez conocido y analizado el alcance del fútbol español, a nivel nacional, hay que realizarse una pregunta, ¿cuánto genera el fútbol como industria en España? Para ello, la prestigiosa consultora Price Waterhouse Cooper (pwc de ahora en adelante) realizó un informe publicado en diciembre del 2018 donde se analizaba el impacto económico, fiscal y social del fútbol profesional en España. A pesar de que los datos arrojados correspondían a la temporada 2016/2017, se presupone que el panorama de la industria futbolística no ha sufrido grandes cambios; si bien hay que tener en cuenta la crisis del coronavirus que, como a todos los sectores afectó al fútbol, perdiendo entre otros ingresos de *ticketing* por la ausencia de público, aunque esta situación se encuentra en vías de superarse tras la vuelta a la completa normalidad en la temporada 2021/2022.

A pesar de no estar muy presente en los análisis económicos de la nación, la industria futbolística en España,

como plasma el informe de pwc, supone el 1,37% del PIB del país. Dicha aportación se traduce en unos ingresos totales de 15.688 millones de euros entre impacto directo, impacto tractor, impacto indirecto e impacto inducido; por cada euro de actividad directa de LaLiga se generaron 4,2€ en el resto de la economía. Para poner en perspectiva, la aportación económica del fútbol en España equivale al 48% de la actividad económica generada por el sector de las telecomunicaciones o 1,4 veces los ingresos del transporte aéreo. Todo este volumen de capital repercute fiscalmente en 4.100 millones de euros a las arcas del estado, esta cifra equivale a 2,7 veces el gasto en política exterior estipulado en los Presupuestos Generales del Estado, o el 37% de los ingresos tributarios de la Comunidad Valenciana a lo largo del año 2017.

LaLiga genera una serie de efectos tractores sobre la economía de otros sectores que de otra forma no se generarían, estos efectos se traducen en 3.998 millones de euros en sectores tan diversos como la hostelería y restauración, juego y apuestas, transporte y alojamientos, televisión de pago, videojuegos y medios de comunicación.

En términos de empleabilidad, supone un total de 184.626 puestos de trabajo (entre todos los impactos), esta cifra es igual al 0,98% de las personas ocupadas en España o a 1,2 veces los empleos generados por la industria textil; por cada empleo directo generado por LaLiga, se crearon 4 empleos más [3].

En la época actual, marcada por el vertiginoso avance técnológico y la digitalización en todos los ámbitos, una industria de tanta importancia como es la futbolística no puede quedarse atrás. Por ello, LaLiga fundó en el 2021 su filial tecnológica *LaLiga Tech*, concebida para el desarrollo de herramientas y soluciones tecnológicas basadas en datos, para el mundo del deporte y del entretenimiento. Entre todos sus proyectos destacan *Calendar Selector*, una herramienta que haciendo uso de inteligencia artificial escoge las mejores fechas y horarios de los encuentros para maximizar así su audiencia y asistencia al estadio, o su plataforma de transmisión OTT, que como describe la propia organización a través su web "incluye herramientas de *business intelligence* para analizar patrones de visionado; creación y gestión de campañas multicanal de *engagement* con los aficionados, creando una imagen de 360° de cómo interactúan los seguidores con todas las propiedades digitales" [4]. Ambos proyectos han recibido certificaciones I+D+i de la mano de la European Quality Assuarance (EQA) y la Agencia de Certificación de Innovación Española (ACIE) por su carácter tecnológico e innovador [5]. Se muestra a continuación el esquema oficial distribuido por LaLiga donde se reflejan todos los productos desarrollados por *LaLiga Tech.*

Figura 1-4. Productos LaLiga Tech (Fuente: LaLiga).

En cuanto a la implementación de las nuevas tecnologías por parte de los equipos, principalmente aquellas relacionadas con el dato y la inteligencia artificial, de entre los 42 clubes presentes en LaLiga (tanto de primera como de segunda división), únicamente ocho de estos cuentan con un departamento específico de innovación, big data o tecnología (según reflejan los organigramas publicados por cada club en su página web oficial). Dentro de un club de fútbol, las posibilidades de implementación de herramientas que apliquen el dato son casi infinitas, no solo se utilizan para analizar métricas deportivas o tácticas, sino que están muy presentes en acciones comerciales y estratégicas, desde el control de redes sociales al modelo de *ticketing*. Viendo que la especialización en estos sectores por parte de los clubes profesionales es lenta y escasa, *LaLiga Tech* aporta soporte y herramientas para facilitar que estos puedan trabajar con la analítica de datos. LaLiga aporta a los clubes doce herramientas para que puedan medir su impacto audiovisual y en redes, tanto nacional como internacionalmente, salud de marca y patrocinios. Se ponen, además, a disposición de los clubes herramientas que ayudan a obtener información sobre data finance, mercado de fichajes, comunicaciones, perfiles de usuarios o merchandaising. El propósito de ofrecer estas herramientas es favorecer la igualdad de oportunidades entre los clubes que forman parte de LaLiga a la hora de embarcarse en la digitalización de sus instituciones; en palabras de Fernando Martín, responsable del área de *Business Intelligence* y *Analytics* de *LaLiga Tech*, en una entrevista concedida al diario El Confidencial: “Si fueran los propios clubes los que hubieran tenido que desarrollar estas herramientas, el precio sería altísimo y habrían tenido que destinar importantes recursos económicos. Es obvio que no todas las entidades podrían haberlo afrontado con la misma intensidad y en el mismo período de tiempo”.

Entre los proyectos desarrollados para el análisis deportivo destaca *Mediacoach*, una herramienta que ofrece a

entrenadores y cuerpos técnicos (junto a los espectadores), estadísticas y métricas avanzadas calculadas en vivo a través del videoanálisis con cámaras de *tracking óptico* [6]. *Mediacoach* es un claro ejemplo de que la implementación de todas estas herramientas ha favorecido positivamente a la hora de la toma de decisiones de los clubes de fútbol, gracias a los millones de datos almacenados se ofrece multitud de información a los cuerpos técnicos a partir de la cual trabajan en nuevas estrategias y variantes tácticas para tratar de batir a sus rivales [7].

1.1 Objetivos

El objetivo principal de este trabajo es el desarrollo de una herramienta que sea capaz de predecir el aumento, o no, del valor de mercado de futbolistas profesionales de una temporada a otra. Para ello, se debe hacer una extrapolación del futbolista como un activo meramente económico, y es que en las variaciones en el valor de mercado no influye únicamente el rendimiento deportivo; la edad, la reputación del futbolista o de su equipo, y el valor de mercado anterior son, entre otros, parámetros no deportivos que influyen en el valor de mercado. Es por esto por lo que con el uso de esta herramienta se busca potenciar únicamente el rendimiento económico, aunque vaya muy ligado al deportivo.

Cabe destacar que el rendimiento de un futbolista no es una cuestión matemática, está sujeto a multitud de factores externos que pueden repercutir en su rendimiento: desde lesiones, falta de entendimiento con sus compañeros y técnicos o mal rendimiento general del equipo, hasta situaciones personales como puede ser una ruptura amorosa o un problema familiar. Teniendo en cuenta esos factores externos fuera del alcance de este proyecto, y haciendo uso de técnicas machine learning, se tratará de realizar predicciones en los cambios de valor de mercado de los futbolistas.

La herramienta ha sido concebida para facilitar la toma de decisiones a ambos lados de la negociación en la compraventa de jugadores. Por un lado, el club con interés en adquirir a un futbolista puede observar si el valor de mercado del jugador en que tiene interés aumentará la temporada siguiente, razón que motivaría la compra, o no aumentará, pudiendo hacer que se rebaje el interés en dicha adquisición. Por otro lado, el club que reciba una oferta de compra por uno de sus jugadores en plantilla puede evaluar si el es momento idóneo para realizar la venta, ya que, si se predice que el valor de mercado de su futbolista no aumentará, desde un punto de vista económico será mejor desprenderse de este.

1.2 Estructura del trabajo

Se presenta a continuación la estructura del trabajo, dividido en 7 puntos, contando con este capítulo introductorio.

- **Capítulo 2**: Se introduce lo que es una secretaría técnica y su importancia para la viabilidad deportiva y económica de un club de fútbol. Una vez introducido el concepto de secretaría técnica se desarrolla el uso que estas hacen del *BigData*.
- **Capítulo 3:** En este capítulo se introduce el concepto de *Machine Learning*, desde su contexto histórico a sus aplicaciones. Se describen con detalle los diferentes tipos de aprendizajes y algoritmos utilizados en el desarrollo del proyecto.
- **Capítulo 4:** Se centra en describir el proceso de captación de los datos y sus diferentes procesamientos hasta tenerlos preparados para generar los distintos modelos de *Machine Learning*. Se realiza también un amplio análisis descriptivo de los datos obtenidos.
- **Capítulo 5:** En el se muestran los resultados obtenidos de los diferentes modelos generados. Además, se describen cambios realizados sobre los distintos conjuntos de datos para tratar de mejorar los resultados obtenidos en un principio.
- **Capítulo 6:** En el capítulo 6 se plasman las conclusiones del proyecto, realizando una reflexión entre los resultados obtenidos y futuras líneas de investigación.

2 EL DATO, NUEVO COMPAÑERO DE LA DIRECCIÓN DEPORTIVA

Los clubes de fútbol mueven una gran cantidad de dinero en el mercado de fichajes, esto motiva a buscar nuevas herramientas y argumentos que favorezcan el maximizar sus ingresos minimizando costes. En este capítulo se describe el trabajo elaborado por una dirección deportiva y la importancia de los traspasos en los balances económicos de los equipos de fútbol como instituciones. Se describe también la adopción del *Big Data* y la Inteligencia Artificial en estos ámbitos.

2.1 La dirección deportiva

Un éxito puntual, como puede ser la consecución de un título, una buena clasificación o un ascenso, puede ser fruto únicamente de las circunstancias puntuales de una temporada. Pero, cuando ese éxito es continuo en el tiempo ya no es casual, sino causal. Estos éxitos reflejan un buen modelo de gestión de la entidad que, como no, debe ser refrendado en todos los ámbitos de la misma, pero en el que tiene una carga de responsabilididad importantísima la dirección deportiva. Antes de profundizar acerca de que es una dirección deportiva y sus responsabilidades, hay que hablar acerca de los distintos modelos de gestión de un club de fútbol.

2.1.1 Modelos de gestión deportiva

La salud financiera de los clubes de fútbol, como cualquier otra institución, es un aspecto de máxima importancia a la hora de gestionarlos. Pero, no puede olvidarse que se está tratando con instituciones deportivas, cuyo objetivo principal es la consecución de objetivos en el terreno de juego. Además, a pesar de que existen multitud de fuentes de ingresos, el éxito económico siempre va acompañado del deportivo. Por ello, una buena gestión deportiva marcará diferencias remarcables en los ingresos de los clubes, desde los ingresos por clasificaciones deportivas hasta mayores ingresos en marketing por una expansión mayor de la marca.

En función de como se organicen los órganos de responsabilidad internos, encargados de la gestión deportiva de los clubes, repartiendo responsabilidades entre una u otra figura, existen tres modelos principales, aunque en la actualidad predominan únicamente dos de ellos.

- **Modelo Anglosajón:** En este, la responsabilidad principal de la gestión deportiva recae en la figura del entrenador. El *mister* tiene poder absoluto en la toma de decisiones deportivas, únicamente debe atenerse a los límites económicos que marque el club. Un ejemplo de club que siga este modelo, al igual que en toda Inglaterra es el Southampton FC, cuyo organigrama se muestra a continuación:

Figura 2-1. Organigrama Southampton FC (Fuente: medium.com)

- **Modelo Latino:** Probablemente el más extendido en la actualidad. El peso de la gestión deportiva recae principalmente sobre tres pilares: la directiva del club, el entrenador y el director deportivo. Al igual que sucede en el modelo anglosajón, la directiva se encarga de limitar las cantidades económicas disponibles para el área deportiva. Una vez conocido el presupuesto disponible, el entrenador transmite al director deportivo las necesidades que tiene el equipo, y este, será el encargado de materializar dichas necesidades. Se incluye a continuación el organigrama del Sevilla FC, club ejemplar en la implementación del modelo latino, también conocido como modelo mixto:

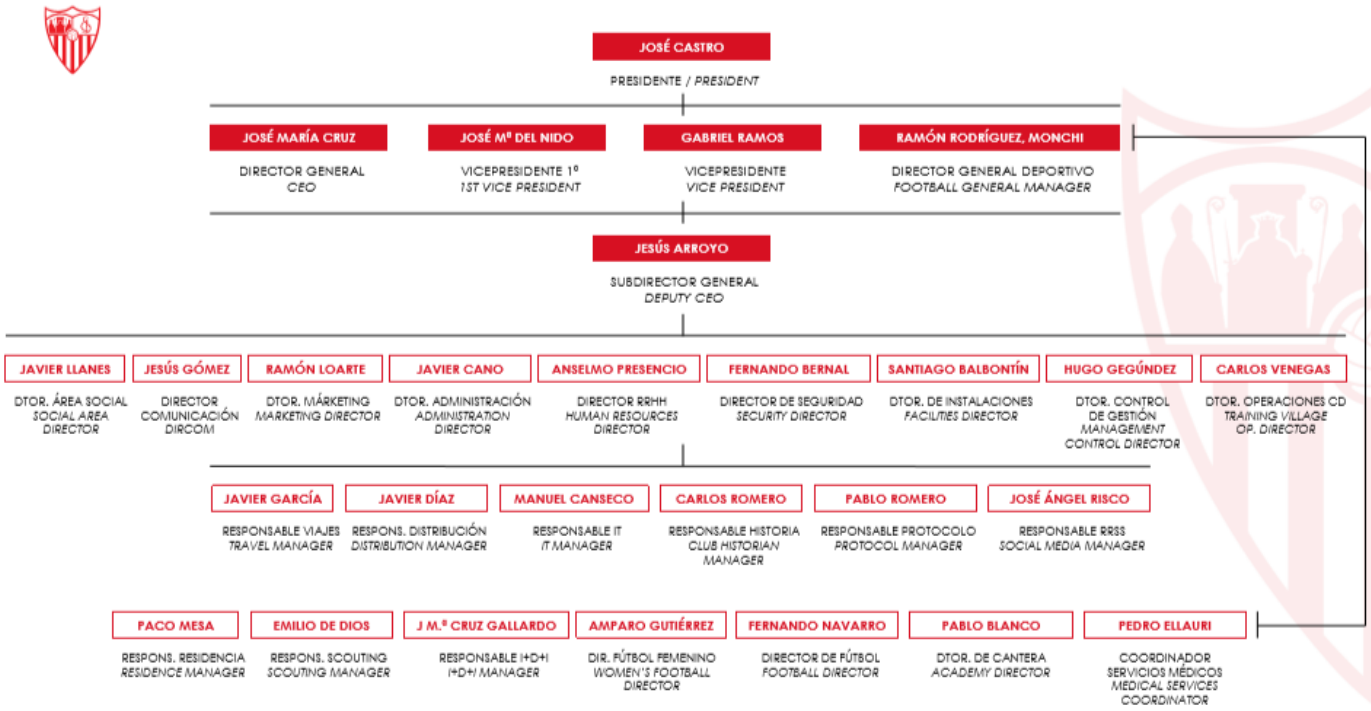


Figura 2-2. Organigrama Sevilla FC (Fuente: sevillafc.es)

- **Modelo Presidencialista:** Prácticamente extinto en la actualidad. El presidente tiene poder absoluto en el club. Generalmente cuenta con consejeros y personal especializado, pero la última palabra siempre la tiene el presidente. Un club que mantiene este tipo de modelo hoy en día es el Real Madrid CF, a continuación, se muestra su organigrama [8,9]:

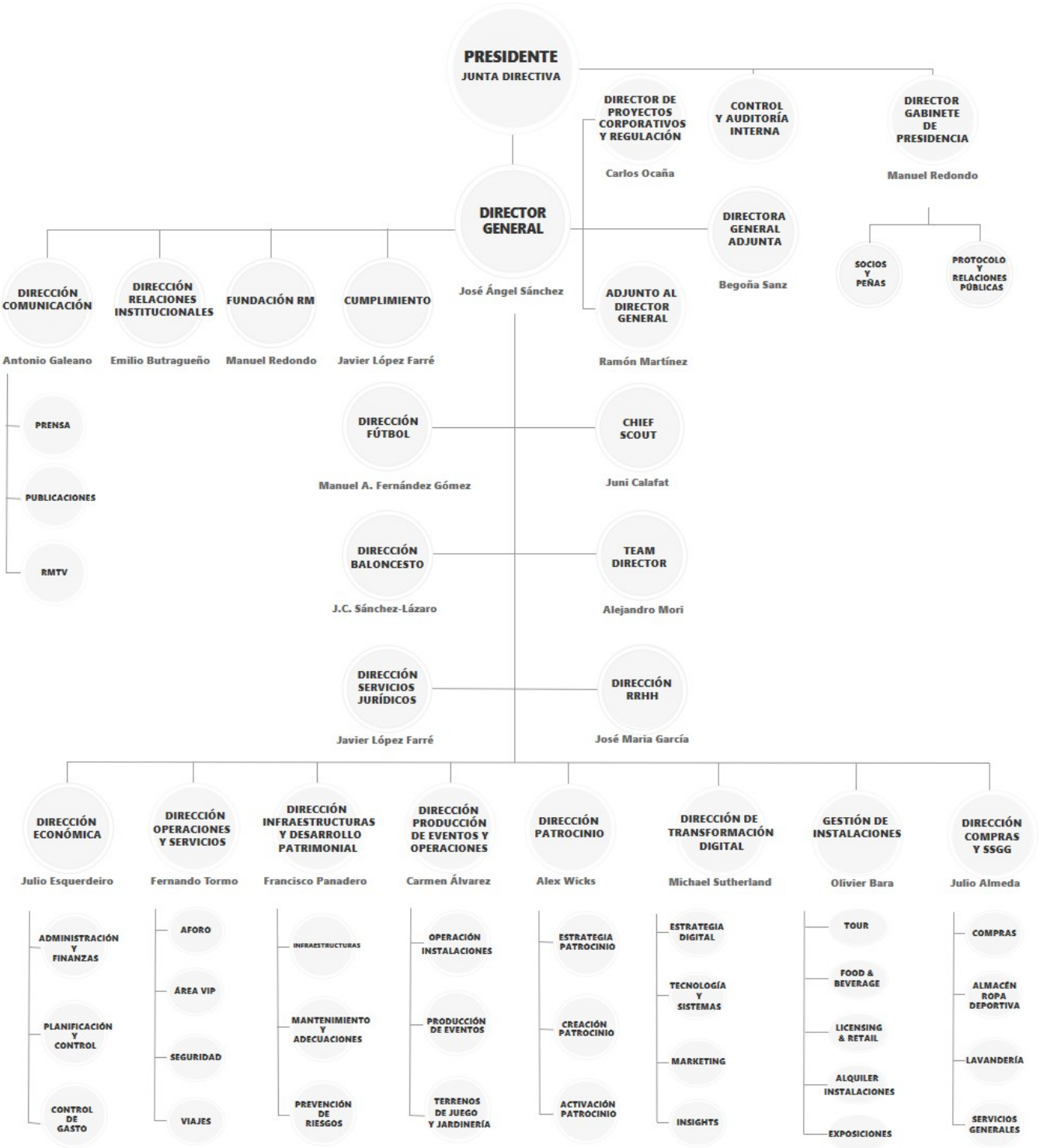


Figura 2-3. Organigrama Real Madrid CF (Fuente: realmadrid.com)

2.1.2 La figura del director deportivo

Entre todos los modelos de gestión deportiva, el más extendido en los clubes hoy en día es el modelo mixto (también conocido como modelo latino). Este modelo introduce una figura clave para la gestión deportiva de las entidades, el director deportivo. ¿Qué es un director deportivo? Respondiendo a esta pregunta, Albert Valentín, exfutbolista y exsecretario técnico del FC Barcelona entre los años 2010-2015 [10], define la figura del director deportivo en su libro titulado "Dirección deportiva en un club profesional", como el: "responsable de ejecutar la política deportiva de la entidad y de conseguir los objetivos marcados por la cúpula directiva dentro de su Plan Estratégico. Responsable máximo del Área Deportiva que dirigirá salvaguardando y defendiendo los valores de la organización".

En los últimos años, en un intento de prosperar económica y deportivamente, los clubes han optado por una mejora profunda de sus estructuras internas. Esta mejora ha traído consigo la especialización en todas las áreas. Cada vez se observan en los organigramas como van apareciendo más departamentos, subdepartamentos y áreas. Entre todas estas áreas, las áreas deportivas de los diferentes clubs cuentan con figuras cuya labor se hace difusa, no sabiendo establecer el límite claro entre una y otra, es el caso del secretario técnico, director general y director deportivo. Por ello, se describen a continuación las labores del director deportivo, las cuales se reparten en cuatro tareas principales: planificación, organización, dirección, control y evaluación.

Figura 2-4. Funciones del Director Deportivo.

- **Planificación:** Albert Martín, define el proceso de planificar como: "Partiendo de un análisis previo de la situación, definir las tareas a realizar y distribuirlas en el tiempo con el fin de alcanzar los objetivos propuestos con anterioridad". Una planificación deportiva es preparar el futuro, hacer un análisis de la situación deportiva actual de la entidad, hallar debilidades y aspectos a mejorar, para ponerles solución en forma de salidas, renovaciones y fichajes. Durante este proceso, un buen director deportivo debe abstraerse del ruido externo y de situaciones deportivas puntuales que inviten a decisiones en caliente, para así seguir su hoja de ruta y tratar de conseguir los objetivos, alcanzando cada vez cotas mayores.

- **Organización:** Definido en el libro "Dirección deportiva en un club profesional" como: "Distribución de funciones y actividades entre los miembros del Área Deportiva estableciendo los vínculos organizativos entre ellos, definiendo la estructura de poder y administrando los recursos disponibles para la consecución de los objetivos planteados". El director deportivo, como máximo responsable del área deportiva de la entidad debe organizar y definir las funciones de todos los miembros de su área, desde ojeadores a directores de cantera. Esta organización debe quedar plasmada en el organigrama de la entidad de manera que cada subárea quede claramente definida. Deberá, además, encargarse de la selección de personal dentro de la parcela deportiva.

- **Dirección:** El director deportivo debe líderar y coordinar a todos los miembros del área deportiva con el propósito de que cada departamento trabaje conjuntamente para conseguir los objetivos marcados desde el club. Será clave la capacidad de liderazgo y de motivación con la que cuente el director deportivo, así como una buena habilidad comunicativa. El director deportivo debe ser un líder, debe saber transmitir un mensaje a su equipo y que este cale en él. Además, debe ser ávido en la toma de decisiones, pues estará expuesto a multitud de decisiones en el día a día.

- **Control y evaluación:** El trabajo del director deportivo va mucho más allá de los meses de planificación. La parcela deportiva de un club de fútbol requiere trabajo y control a diario. Una vez realizada la planificación, se debe supervisar que esta se lleve a cabo según lo establecido. Además, el día a día es el que aporta información relevante para futuras planificaciones, el director deportivo debe conocer en todo momento como trabajan los jugadores y cuerpo técnico. Para valorar y controlar el trabajo del área deportiva, se evalúan, cuantitativa y cualitativamente, a todos los eslabones que componen la cadena de trabajo continuamente, tomando las medidas necesarias cuando los resultados no sean los adecuados [10].

2.1.3 El peso económico

El trabajo de la dirección deportiva no solamente se ve reflejado en el campo, con los resultados deportivos. A nivel financiero, el trabajo de la dirección deportiva es fundamental para el devenir económico de la gran mayoría de clubes. Una gran cantidad de clubes, asentados en la élite, pero sin los ingresos de los equipos más poderosos, basan su filosofía de crecimiento en un modelo de compraventa de jugadores. Con el fin de aumentar su presupuesto, buscan generar plusvalías vendiendo a sus mejores activos deportivos. La clave reside en comprar barato para vender caro, ser capaces de detectar el talento precoz, encontrar el ecosistema idóneo para que este se desarrolle y finalmente, venderlo. Se muestran en la Figura 2-5 los clubes que obtuvieron mayores beneficios por transacciones en el mercado de fichajes entre los años 2012 y 2021.

Club	Invest.	Receipts	Balance
LOSC Lille (FRA)	€321M	€663M	+€342M
Olympique Lyonnais (FRA)	€421M	€646M	+€225M
Genoa CFC (ITA)	€262M	€472M	+€210M
Udinese Calcio (ITA)	€248M	€415M	+€167M
Atalanta BC (ITA)	€368M	€532M	+€164M
Montpellier HSC (FRA)	€88M	€205M	+€117M
Athletic Club (ESP)	€109M	€224M	+€115M
TSG Hoffenheim (GER)	€229M	€340M	+€111M
AS St-Etienne (FRA)	€113M	€223M	+€110M
Empoli FC (ITA)	€73M	€164M	+€91M

Figura 2-5. Clubes con mayores beneficios por traspasos 2012-2021
(Fuente: football-observatory.com)

Se procede, a continuación, a analizar los resultados económicos directamente dependientes de la actividad de la dirección deportiva como son los traspasos y el coste de la plantilla en los clubes españoles. Según el último informe económico financiero publicado por LaLiga en el año 2021, en la temporada 2019/2010 los equipos de la competición recibieron 1130,5 M€ en concepto de traspasos de los derechos deportivos de jugadores. Esta cifra supone un 22,4% del total de los ingresos, una parte bastante importante de la economía de las entidades. A pesar de que no se aportan datos exactos de la temporada 2020/2021, comentan desde LaLiga que, entre las cinco grandes ligas estipuladas por la UEFA, se disminuyeron en un 50% los ingresos por traspasos de jugadores, en España concretamente, un 62%, la liga más mermada de las cinco. Esta disminución de los ingresos se debe a la crisis originada por el coronavirus. A continuación, se muestran los ingresos totales percibidos por los clubes de LaLiga en las últimas temporadas, desglosado por concepto. Además, se presentan en la Figura 2-6, los mismos ingresos representando el porcentaje que suponen del total.

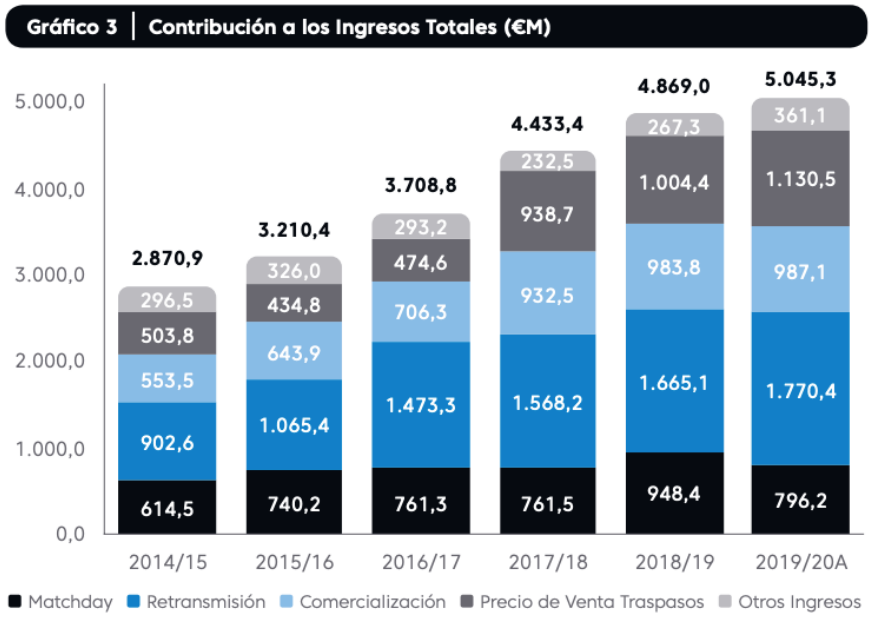


Figura 2-6. Ingresos clubes de LaLiga (Fuente: LaLiga).

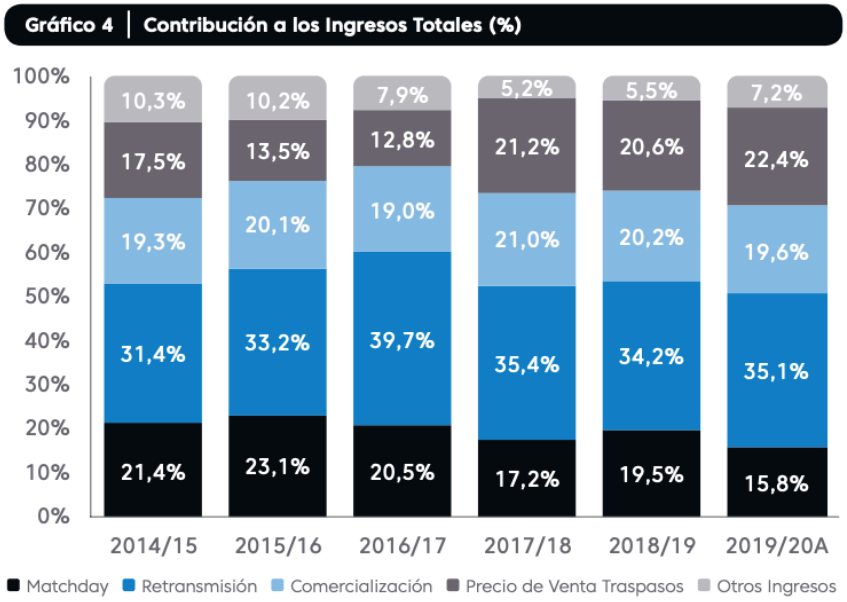


Figura 2-7. Porcentaje de ingresos sobre el total (Fuente: LaLiga)

Como se puede apreciar en las figuras anteriores, no solamente los ingresos en concepto de traspasos, al igual que los ingresos totales, aumentaron, sino que su representación en el total es cada vez mayor. A pesar de la disminución de los ingresos durante la temporada 2020/2021, se espera un repunte de las inversiones en materia de fichajes tras la vuelta a la normalidad antes del covid que supone el inicio del fin de la crisis económica.

Por otro lado se encuentran los gastos, entre los que caben destacar los gastos derivados de la adquisición de futbolistas, como son las cantidades por el traspaso así como los salarios. No solo hay que analizar el desembolso que realizan los equipos por hacerse con los derechos deportivos de los futbolistas, sino el coste salarial que suponen para la entidad. Controlar estos gastos es una de las funciones más importantes del director deportivo, pues suponen una gran parte del presupuesto disponible del club. Se plasma en la Figura 2-8, la evolución de los costes salariales desde la temporada 2014/2015 a la 2019/2020, y su relación con los ingresos totales.

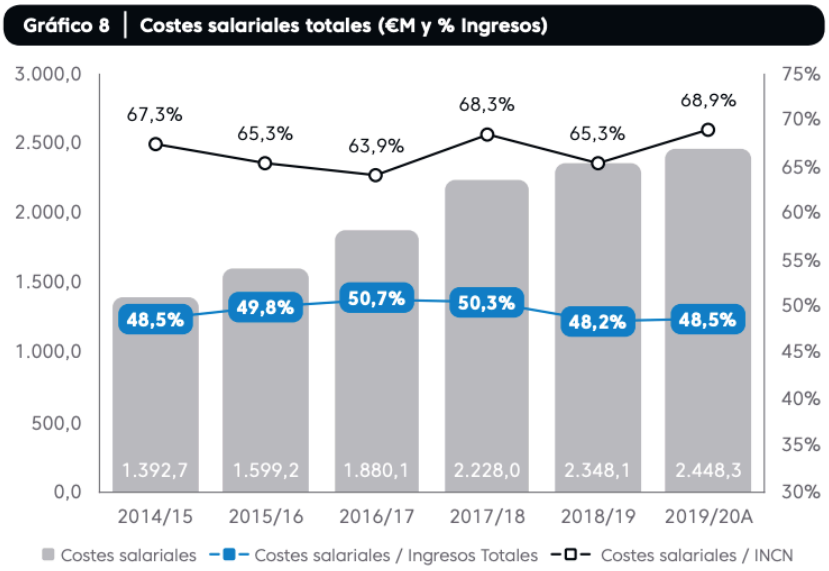


Figura 2-8. Evolución de los costes salariales (Fuente: LaLiga)

Es apreciable como los costes salariales suponen a lo largo de estas temporadas prácticamente la mitad del montante total de ingresos percibido por los clubes. Además, a pesar de mantener ese 50% prácticamente estable durante los años, se comprueba como han ido aumentando los ingresos de los clubes de LaLiga, disponiendo cada temporada de mayores presupuestos.

A continuación, en la Figura 2-9, se representa un diagrama con los gastos totales de los clubes españoles durante la temporada 2019/2020.

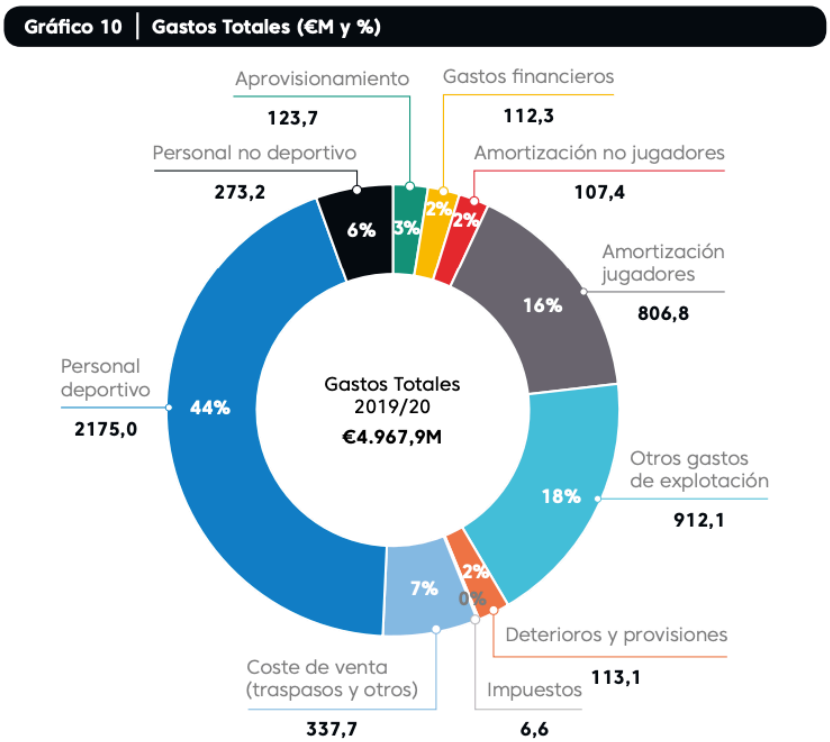


Figura 2-9. Gastos totales temporada 2019/2020 (Fuente: LaLiga)

Viendo el diagrama anterior se puede apreciar como la inversión en compra de jugadores es bastante menor en comparación con el coste salarial del personal deportivo. En la temporada 2019/2020, los clubes de LaLiga gastaron unos 337,7 millones de euros en conceptos de traspasos, un 7% de los gastos totales, frente a los 1130 millones de euros que percibieron por operaciones de venta [11].

2.1.4 Nuevas herramientas de apoyo en la toma de decisiones

Tras analizar la importancia de las direcciones deportivas en los clubes de fútbol, tanto en materia deportiva, como económica, a la hora de tomar decisiones, cualquier herramienta facilite la elección o aporte más información de la disponible, será más que bienvenida. Es ahí donde surgen las nuevas herramientas que, basadas en el dato, como el Big Data y ML, sirven a los clubes en sus labores de dirección del área deportiva.

Como se destacó en la introducción, únicamente 8 de los 42 clubes que conforman LaLiga tienen un departamento específico de I+D+i, Data o similares. En una entrevista para el diario Navarra Capital, Pablo Sanzol, '*head*' de Data Analytics en la SD Eibar comenta sobre el uso de estas herramientas en el fútbol español: "Creo que todavía no estamos ni en la rampa de salida. Hoy en día, en los clubs están empezando a crearse estas

pequeñas estructuras de analítica avanzada. Y me parece que el paradigma de los datos va a ir cada vez a más porque se están recogiendo frutos en aquellos equipos que lo utilizan" [12]. A pesar de que el fútbol en general, y el fútbol español en particular se encuentra en el arranque de una nueva etapa en la que se haga uso de todas estas nuevas tecnologías, lo cierto es que ya se le está dando un uso bastante importante e interesante a estas nuevas ayudas.

Entre los clubes españoles que ya han adoptado el uso de estas nuevas herramientas, se encuentra el Sevilla FC. En una *masterclass* impartida a través de la plataforma *YouTube*, Ramón Rodríguez Verdejo (Monchi), director deportivo del Sevilla FC desgrana el uso del dato dentro de su dirección deportiva. Comenta Monchi que, durante su carrera como director deportivo, 19 temporadas a cargo de la parcela deportiva del Sevilla FC mas otra temporada en la AS Roma, siempre ha utilizado el dato. La diferencia entre el dato de antaño y el actual reside en que antes se utilizaban plantillas de informes generados con un procesador de textos, mientras que ahora se generan y conocen alrededor de ocho millones de datos por partido, acrecentando las posibilidades de sacarle más partido a dicha información. Con todos estos datos generados, la dirección deportiva del Sevilla FC enfoca el uso de herramientas de Big Data o ML en tres parcelas diferentes:

- **Scouting:** "¿Qué es scouting? Es el análisis científico que realiza un Técnico Deportivo recabando la máxima información posible sobre el elemento que se le asigne, utilizando los medios tecnológicos adecuados para transmitirla a los interesados, minimizando por tanto, los riesgos que puede causar la competencia" [13].

  Explica Monchi: "No voy a firmar a un jugador solamente por el dato, pero jamás firmaré a un jugador si antes el dato no me ha dado una señal". Los datos ayudan a reducir riesgos y ganar tiempo. Un ejemplo claro de aplicación es a la hora de filtrar jugadores. Si por ejemplo, el perfil buscado es el de un delantero con buen juego aéreo, automáticamente quedarán descartados todos aquellos jugadores que no superen un parámetro mínimo de duelos aéreos ganados, reduciendo así considerablemente la lista de jugadores a analizar.

- **Prevención de lesiones:** A lo largo de una temporada, uno de los mayores problemas a los que se enfrentan los clubes son las lesiones de sus futbolistas. Esto no influye únicamente en un detrimento del rendimiento deportivo ante la falta de uno u otro futbolista, sino que además representra un coste para la entidad en servicios médicos y salario de jugadores que no pueden ejercer su trabajo. Explica Monchi que conociendo parámetros del día a día de un futbolista, desde horas de sueño a la alimentación, se pueden detectar patrones que permitan controlar el riesgo de lesiones al que está expuesto cada futbolista, de manera que se disminuya drásticamente el número de lesiones sufridas por una plantilla a lo largo de las temporadas.

- **Mercado:** A diferencia de lo que sucedía años atrás donde el valor de mercado de los futbolistas tenía que ser marcado por los propios clubes propietarios de sus derechos deportivos, gracias al Big Data, analizando parámetros de juego de los futbolistas, se puede establecer su valor de mercado real. Además, comenta Monchi que la clave para el futuro es ser capaces de conocer cuando un jugador está en su valor de mercado máximo y mínimo para así poder venderlo o comprarlo obteniendo el mayor rendimiento económico posible [14].

Los usos descritos anteriormente, no son los únicos que se le da a este tipo de herramientas a la hora de tomar decisiones por parte de las direcciones deportivas. Uno de los usos más mediáticos del dato en el mundo del fútbol fue la renovación de Kevin De Bruyne con el Manchester City. El jugador belga firmó en el año 2020 una extensión de su contrato por cinco temporadas, percibiendo un aumento de sus emolumentos desde las 350.000 libras semanales a las 400.000. De Bruyne pertenece al reducido grupo de futbolistas de élite que no cuentan con representante, sus asuntos son gestionados por su padre y sus abogados [15]. Al inicio de los contactos entre club y jugador para abordar su renovación, el Manchester City ofrecía al centrocampista belga un contrato donde veía reducido su salario. Por su parte, De Bruyne encargó a sus abogados que recabaran informes en los que, según analítica de datos, se cuantificara el valor que aportaba al equipo. Una de las empresas encargadas en realizar dicho análisis fue *Analytics FC*, una compañía de análisis de datos de fútbol que trabaja con clubes de la *Premier League* y otras ligas europeas. Según el informe desarrollado por *Analytics FC*, Kevin De Bruyne se encontraba entre los mejores jugadores de la *Premier League*, especialmente en métricas de creación de ocasiones. Dicho informe realizó comparaciones con jugadores de toda Europa, resultando De Bruyne como el mejor jugador del viejo continente en cuanto a su contribución en ataque, según una métrica obtenida con los datos de cada toque de balón realizado por los jugadores en todos los partidos disputados los 12 meses anteriores. Analizaba *Analytics FC* también, las posibilidades de ganar la *Champions League* que tendría el Manchester City con De Bruyne y sin él, además de las posibilidades de ganarla que tendrían otros equipos en caso de ficharle, demostrando la gran aportación del centrocampista en el juego de cualquier equipo. Finalmente, comparando su salario con el de otros jugadores de nivel élite, resultó que De Bruyne estaba percibiendo un sueldo por debajo de varios futbolistas que, números en mano, aportaban menos al juego de su equipo que el belga. Presentando este informe fue como consiguió renegociar De Bruyne su contrato con el club de Manchester, aumentando su salario a pesar de la oferta inicial [16].

Otro caso muy relevante fue la salida del delantero Memphis Depay del Manchester United. Tras una muy buena campaña en la temporada 2014/2015, acompañada de un rendimiento notable en el mundial de 2014, Memphis Depay firmó en verano del 2015 por el Manchester United. Tras no conseguir hacerse con un hueco en el equipo, y habiendo disputado únicamente 20 minutos en la primera mitad de su segunda temporada, Depay buscó una transferencia de equipo en enero de 2017. Para ello el delantero neerlandés solicitó los servicios de *SciSports,* empresa dedicada a la analítica e inteligencia de datos en el deporte. El equipo especializado de *SciSports* realizó

un Informe de Éxito de Transferencia, de donde, basándose en algoritmos propios de la compañía, se obtuvo una lista de equipos que reunían los requerimientos del futbolista. Para la elaboración del informe se tuvieron muchos factores en cuenta, desde el estilo de juego del equipo, el entrenador o la competencia en su puesto. Finalmente, entre las ofertas con las que contaba el jugador, y con la información obtenida a través de *SciSports,* Memphis Depay firmó por el Olympique de Lyon. La decisión fue bastante positiva para la carrera del futbolista, en sus primeros meses en Francia anotó 5 goles y repartió 8 asistencias, mientras que en su primera temporada completa, 2017/2018, marcó 19 tantos y asistió en otros 13 [17].

3 MACHINE LEARNING

En este punto se abordará el concepto de Machine Learning, su contexto histórico, marco de aplicación y diferentes tipos de técnicas y algoritmos existentes. Se entrará más es detalle en los tipos de aprendizaje y algoritmos utilizados en el desarrollo del trabajo, estos son de tipo de aprendizaje supervisado, concretamente algoritmos de clasificación, de los cuales se aplicarán Regresión Logística y Random Forest.

3.1 Definición y usos

Machine Learning, ML a partir de ahora, o Aprendizaje Automático es una subárea de la inteligencia artificial (IA), generalmente definida como la capacidad de una máquina para imitar el comportamiento de la inteligencia humana. El término fue acuñado por Arthur Samuel, informático de IBM y pionero en el campo de la IA y de los videojuegos, como: "el área de estudio que aporta a los ordenadores la habilidad de aprender sin ser explícitamente programados" [18] .

Los orígenes del ML datan del año 1950 cuando Alan Turing creó el "Test de Turing", prueba que servía para determinar si un ordenador poseía inteligencia. Para pasar dicho examen el ordenador debía engañar a un humano para hacerle creer que estaba hablando con otra persona.

Más adelante, en el año 1952, el anteriormente mencionado Arthur Samuel desarrolló el primer programa de ordenador capaz de aprender. Se trataba de un juego de damas, donde el ordenador mejoraba con cada partida que jugaba, aprendiendo de los movimientos y estrategias de los jugadores.

En el año 1957 Frank Rosenblatt diseño el "Perceptron", la primera red neuronal. Para su desarrollo se simuló el proceso de las neuronas en el cerebro humano.

Diez años más tarde, en 1967 se escribió el algoritmo "nearest neighbor", que inició el, aunque muy básico, reconocimiento de patrones [19].

Tras muchas décadas de avances se llega hasta la actualidad, donde las posibilidades de uso de ML son prácticamente infinitas. En 2021, según el índice global de adopción de la IA de IBM, un 41% de las empresas aceleraron la adopción de técnicas de IA como consecuencia de la pandemia [20]. Al estar presente en un número tan extenso de empresas, el uso que se le da a la tecnología es bastante diverso. Entre todas las aplicaciones que se le da al ML pueden destacarse: algortimos de recomendación, frecuentemente usados en redes sociales o plataformas de vídeo bajo demanda; análisis de imágenes y reconocimiento de objetos, donde destaca el reconocimiento facial; detección de fraude bancario, analizando patrones de gasto; ciberseguridad, buscando vulnerabilidades que puedan convertirse en futuras brechas; conducción de coches autónomos, con técnicas de deep learning principalmente; diagnósticos médicos y riesgos de contracción de enfermedades, a través del reconocimiento de imágenes y el historial médico del paciente [18].

3.2 Tipos de aprendizaje de Machine Learning

Según el método de aprendizaje del algoritmo, condicionado por la estructura de los datos de entrada, y por el objetivo del modelo, existen tres tipos de aprendizajes en ML: aprendizaje supervisado, aprendizaje no supervisado y aprendizaje semisupervisado. A continuación, se realiza una descripción de todos ellos, haciendo, posteriormente, un énfasis especial en los algoritmos de aprendizaje supervisado, objeto de estudio de este trabajo.

- **Aprendizaje supervisado:** En los modelos de aprendizaje supervisado se nutre al programa con una base de datos completa. En ella, se encuentran todas las muestras con su variable de salida conocida, es decir, los datos se encuentran etiquetados. El algoritmo se encarga de buscar patrones entre los datos y, aprendiendo de todas las salidas conocidas, realiza un pronóstico que será autocorregido. De esta forma logra aprender de sus propios errores, haciéndose cada vez más preciso [21].

- **Aprendizaje no supervisado:** A diferencia de lo que sucede en el aprendizaje supervisado, en el aprendizaje no supervisado se desconoce el resultado de la variable de salida, es decir, se tienen etiquetas desconocidas. Este tipo de aprendizaje se utiliza para descubir patrones y similitudes en la propia estructura de los datos que permitan realizar una segmentación de estos, así como simplificación de la propia estructura de los datos. Destacan las técnicas de *Clustering* (agrupación) y reducción de dimensionalidad [22].

- **Aprendizaje semisupervisado:** Es una mezcla de los dos anteriores, los datos de entrada se encuentran mayoritariamente con etiquetas desconocidas, aunque existe un pequeño conjunto de estos cuyas etiquetas si se conocen. Se utilizas el conjunto etiquetado para entrenar algoritmos de aprendizaje supervisado que se encargará de etiquetar al resto de los datos disponibles [23].

3.3 Algoritmos de aprendizaje supervisado

Dentro de los algoritmos de aprendizaje supervisado destacan principalmente los de clasificación y los de regresión. Mientras que los algoritmos de regresión tienen como objetivo predecir variables continuas, los algoritmos de clasificación predicen variables categóricas. A continuación, se procede con la descripción de los algoritmos de clasificación, ya que para el desarrollo del proyecto serán las técnicas a utilizar.

3.3.1 Algoritmos de clasificación

Como se ha comentado previamente, los algoritmos de clasificación de ML tienen como objetivo predecir variables categóricas. Un ejemplo claro de modelo en el que se apliquen algortimos de clasificación es este trabajo, se busca predecir si el valor de mercado de un futbolista aumentará o no. El algoritmo debe etiquetar la muestra del tipo A o B, no detenerse en cuanto aumentará o dejará de aumentar; a pesar de que este ejemplo es una clasificación binaria, existen clasificaciones entre dos o más categorías.

Dentro de las técnicas de clasificación existen diversos algoritmos que se aplican en los modelos predictivos, cada algoritmo funciona de una manera diferente por lo que no todos obtendrán el mismo resultado ni la misma precisión. Entre todos los algoritmos se van a describir los escogidos para el desarrollo del proyecto: Regresión Logística y Random Forest, pero, para poder explicar el Random Forest es necesario introducir previamente los Árboles de Decisión,

3.3.1.1 Regresión Logística

Los modelos de regresión logística estudian la relación entre una variable dependiente cualitativa (binaria o multinomial) y una o más variables independientes llamadas covariables. A diferencia de la regresión lineal, donde se trata de predecir el valor de Y a partir de una o varias X, la regresión logística busca predecir la probabilidad de que ocurra Y conociendo los valores de X. Para ello parte de la siguiente ecuación:

$$P(Y = 1|X) = \frac{e^{(b_0 + \sum_{i=1}^{n} b_i x_i)}}{1 + e^{(b_0 + \sum_{i=1}^{n} b_i x_i)}} \quad (1)$$

donde b_0 es el término independiente del modelo, x_i el valor de cada variable independiente y b_i el coeficiente o peso de la variable independiente x_i. Gráficamente la relación entre las covariables y la probabilidad de que la salida sea 1 tiene una forma sigmoide, tal y como se muestra en la Figura 3-1 [24].

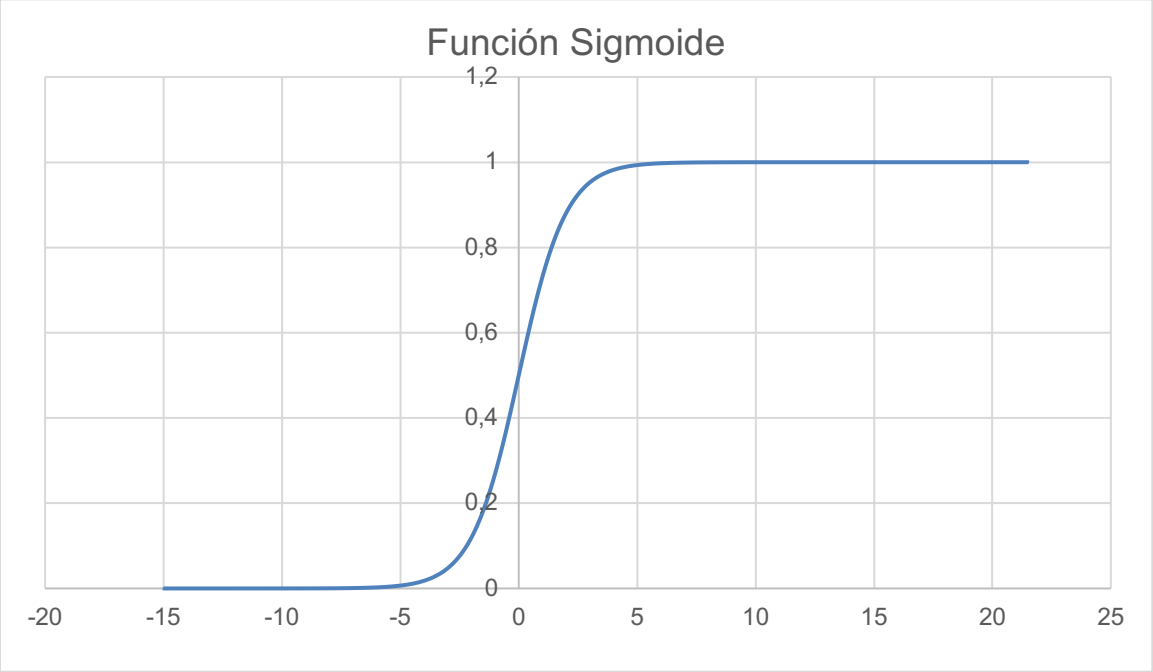


Figura 3-1. Función sigmoide

Como se puede apreciar, cuando la función tiene dos asíntotas horizontales para las cuales su valor es 0 o 1. Ambos resultados en las asíntotas significan una u otra clase, 1 podría ser el aumento del valor de mercado de un futbolista, mientra 0 significa que su valor no ha aumentado.

Dividiendo la ecuación anterior entre su complementario, es decir, probabilidad de éxito entre probabilidad de no éxito se obtiene el *odds* del problema:

$$\frac{P(Y=1|X)}{1-P(Y=1|X)} = odds = e^{(b_0+\sum_{i=1}^{n} b_i x_i)} \tag{2}$$

Finalmente, tomando logaritmo se obtiene la función *logit* (ln(odds)), que representa de forma lineal la expresión de la regresión.

$$\ln\left(\frac{P(Y=1|X)}{1-P(Y=1|X)}\right) = b_0 + \sum_{i=1}^{n} b_i\, x_i \tag{3}$$

El objetivo del algoritmo es calcular los coeficientes de las covariables y el término independiente que mejor ajusten la función sigmoide a los resultados del conjunto de datos usados para entrenar el algoritmo. Tras obtener dichos valores, todas las muestras nuevas serán evaluadas en la función y serán clasificadas como 1 o 0 según su resultado.

3.3.1.2 Árboles de decisión y Random Forest

Los árboles de decisión pueden utilizarse en la construcción de modelos tanto de regresión como de clasificación. Se denominan árboles porque su estructura es similar a los mismos. Partiendo desde la raíz se llega a los nodos, donde se divide el árbol en ramas diferentes según una determinada condición (p.e. $x_4 > 0.5$). Cada rama conduce hasta un nodo final, que supone el resultado según todas las decisiones intermedias tomadas en cada nodo que han construido esa raíz [25].

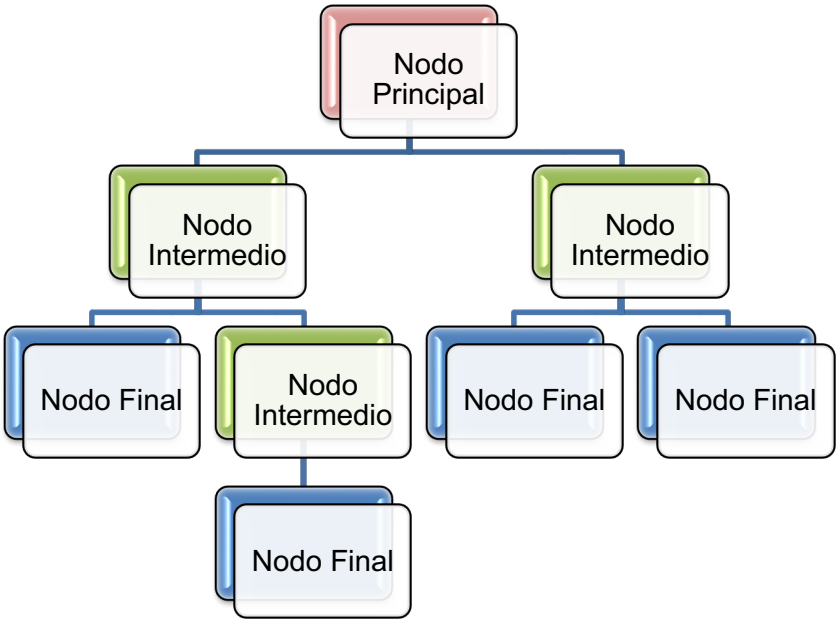


Figura 3-2. Ejemplo de árbol de decisión.

Seleccionar los atributos que componen los nodos no es una tarea que se deba dejar al azar, pues dejaría resultados con muy poca precisión. Los diferentes algoritmos que aplican árboles de decisión utilizan distintos criterios como un sistema de puntuación, aquel atributo con mayor puntación se colocará en la raíz. Entre estos criterios, para árboles categóricos, destacan:

- **Entropía:** La entropía es la medida del desorden, cuanta más entropía tengan las divisiones de los nodos, más difícil será obtener conclusiones. La fórmula de la entropía es la siguiente:

$$Entropía = -\sum_{i=1}^{c} p_i \cdot log_2(p_i) \quad (4)$$

donde c son todas las diferentes categorías del atributo, y p_i la probabilidad de formar parte de cada una de ellas. Por ejemplo, si se obtuviesen los siguientes datos para predecir según las características del conductor y del coche, la supervivencia en accidentes de tráfico:

Sexo	Coche	¿Alcohol?	¿Sobrevive?
Hombre	Deportivo	0	1
Mujer	Deportivo	1	0
Mujer	Todoterreno	1	1
Mujer	SUV	1	1
Hombre	Todoterreno	0	1
Hombre	SUV	0	1
Mujer	Deportivo	0	0
Hombre	Deportivo	1	0
Hombre	Todoterreno	0	0
Hombre	Deportivo	0	1

Figura 3-3. Ejemplo de dataset.

Las p_i de cada tipo de vehículo serían Deportivo 5/10 (0,5), Todoterreno 3/10 (0,3) y SUV 2/10 (0,2); y con esto se calcula su entropía $= -0{,}5 \cdot log_2(0{,}5) - 0{,}3 \cdot log_2(0{,}3) - 0{,}2 \cdot log_2(0{,}2) =$ 1,4854. Este resultado advierte de una alta entropía en el atributo *Coche*, si únicamente se hubiera obtenido *Deportivo* como resultado, la entropía sería igual a 0, mientras que si hubiese sido mitad y mitad *Deportivo* y *Todoterreno*, la entropía sería igual a 0,5.

- **Ganancia de información:** La ganancia de información es el estudio de como afecta a la entropía una división concreta. Para ello se utiliza la siguiente fórmula:

$$GI(T, A) = Entropía(T) - \sum_{v \in A} \frac{|T_v|}{T} \cdot Entropía(T_v) \quad (5)$$

Donde T es la columna *target* u objetivo, A el atributo a dividir, y v todas las divisiones que nacen del atributo. Continuando con el ejemplo anterior, se va a estudiar la ganancia de información que supondría crear un nodo que fuese *Sexo*. Primero se calcula la entropía inicial de la columna objetivo (*¿Sobrevive?*), que es igual a 0.97095. Seguidamente se calculan las proporciones de cada valor único de la columna A: 0,6 (Hombre) y 0,4 (Mujer). A continuación, se multiplican dichos ratios por la nueva entropía de la columna objetivo para cada división, es decir, entre los hombres hubo 4 supervivientes y 2 fallecidos, por lo que la entropía de esa división sería: $-2/3 \cdot log_2(2/3) - 1/3 \cdot log_2(1/3) =$ 0,9183 , mientras entre las mujeres hubo 2 fallecidas y 2 supervivientes, lo que hace una entropía de 1. Finalmente, $GI(T, Sexo) = 0{,}97095 - 0{,}6 \cdot 0{,}9183 - 0{,}4 \cdot 1 = 0{,}01997$. Se obtiene un resultado

positivo, lo cual significa que se ha logrado reducir la entropía con la inclusión de dicho nodo, cuanto mayor sea el resultado, mejor será la división [26].

- **Ratio de ganancia:** Surge para tratar de corregir la tendencia de la ganancia de información a dar como mejores nodos los muy fraccionados, por ello se normaliza con la entropía del atributo.

$$RG(T, A) = \frac{GI(T, A)}{Entropía(A)} \quad (6)$$

En el ejemplo anterior, se obtendría una entropía de la columna A (*Sexo)* igual a 0,97095, quedando un $RG(T, Sexo) = \frac{0{,}01997}{0{,}97095} = 0{,}0206$.

- **Índice Gini:** Se puede entender como una función de coste. Indica la probabilidad de que una variable se clasifique incorrectamente cuando esta es seleccionada aleatoriamente. Se utiliza siguiendo la siguiente fórmula:

$$Gini = 1 - \sum_{i=1}^{c} (p_i)^2 \quad (7)$$

Siendo c todas las divisiones del nodo, y p_i la proporción de cada división.

- **Chi-Cuadrado:** Se utiliza para variables categóricas binarias, cuanto mayor sea su valor, mayor diferencia estadística habrá entre los subnodos y el nodo padre. Se calcula según la siguiente fórmula:

$$\mathrm{X}^2 = \sum \frac{(O - E)^2}{E} \quad (8)$$

Siendo O el resultado obtenido y E el resultado esperado.

Tras aplicar los distintos criterios de formación de los nodos, los diferentes algoritmos de árboles de decisión alcanzan su criterio de parada y el árbol queda montado. Sucede, que los árboles muy cargados de datos obtienen una precisión muy baja. Para evitar esto se realiza la poda que consiste en eliminar, desde los nodos finales del árbol, aquellas ramas cuya eliminación no perturba a la precisión general del mismo. Este proceso se realiza utilizando los conjuntos de datos de entrenamiento, se poda a la vez que se validan los datos para así descubrir aquellos nodos que no afectan a la precisión global del modelo [27].

3.3.1.3 Random Forest

Random Forest, bosque aleatorio en español, es una técnica que utiliza diversos algortimos de árboles de decisión. Consiste en generar un bosque con la mayor cantidad de árboles de decisión diferentes. Se denomina aleatorio por dos cuestiones: para la construcción de los distintos árboles se utilizan fragmentos aleatorios de los datos de entrenamiento administrados y, además, para generar los nodos de cada árbol se seleccionan los atributos de manera aleatoria. Los modelos de clasificación que utilizan *Random Forest* funcionan de la siguiente manera: primero se dividen de manera aleatoria los datos de entrenamiento en distintas agrupaciones; seguidamente, se generan distintos árboles utilizando las agrupaciones anteriores y, dentro de estas agrupaciones atributos aleatorios para la generación de los nodos creando así el bosque aleatorio; finalmente, cuando se desea clasificar una nueva muestra se evalúa esta en cada uno de los árboles que conforman el bosque, obteniendo como resultado final el más votado entre todos los árboles. Por ello, cuantos más árboles se generen, más preciso será el algoritmo [28] .

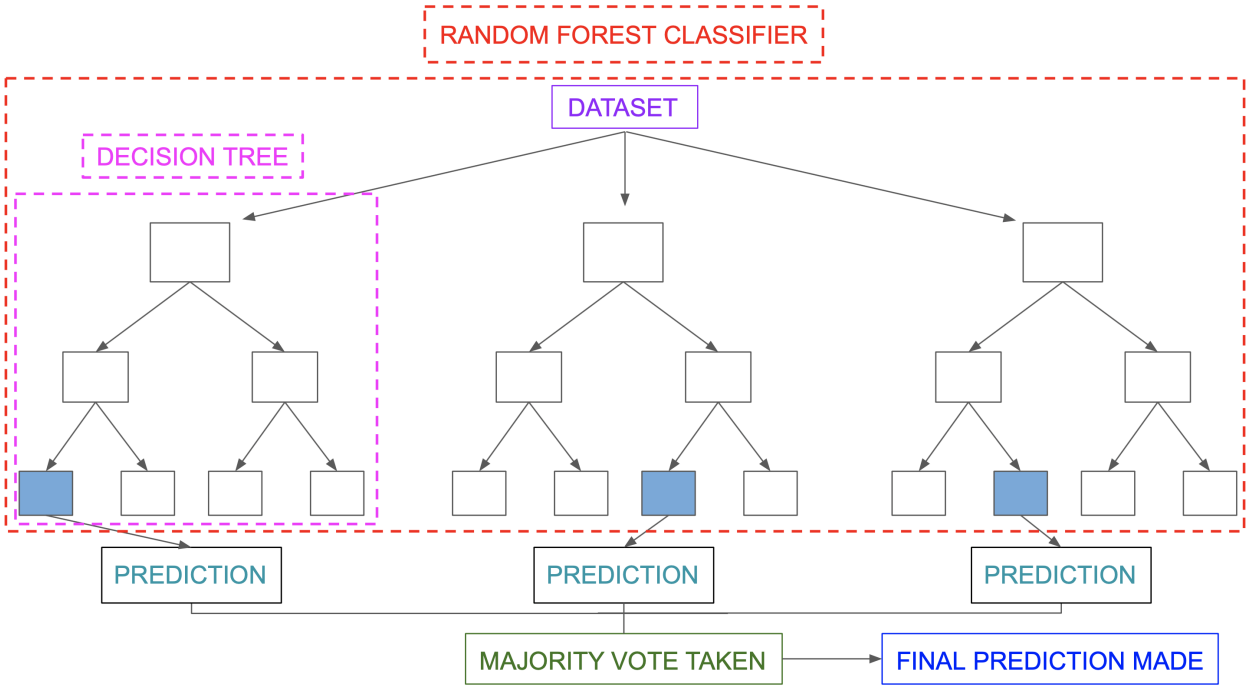


Figura 3-4. Ejemplo Random Forest (Fuente: médium.com).

En la Figura 3 se plasma de manera esquemática el funcionamiento de un bosque aleatorio tal y como se ha descrito previamente.

3.4 Métricas de evaluación

Para valorar la calidad de los modelos desarrollados se construye la matriz de confusión, a partir de la cual surgen diversas métricas que evaluarán los resultados obtenidos tras la ejecución del conjunto *test*.

La matriz de confusión está compuesta por tantas filas y columnas como resultados haya en el conjunto de datos reservado para evaluar el modelo, en ella se enfrentan los resultados reales (Eje X) frente a los resultados obtenidos (Eje Y).

Figura 3-5. Matriz de confusión binaria.

Una vez montada la matriz se clasifican los resultados como:

- **Verdadero Positivo (TP):** Predicción verdadero (1) y resultado real verdadero (1).
- **Verdadero Negativo (TN):** Predicción falso (0) y resultado real falso (0).
- **Falso Positivo (FP):** Predicción positvo (1) y resultado real falso (0).
- **Falso Negativo (FN):** Predicción falso (0) y resultado real verdadero (1).

Las métricas más comunes, y las que se utilizarán posteriormente en este trabajo son:

- **Exactitud (Accuracy):** Ratio de resultados clasificados correctamente.

$$Exactitud = \frac{(TP + TN)}{TP + TN + FP + FN} \tag{9}$$

- **Precisión (Precision):** Frecuencia de éxito en las clasificaciones como verdadero.

$$Precisión = \frac{TP}{TP + FP} \tag{10}$$

- **Exhaustividad (Recall):** También conocida como tasa de verdaderos positivos, muestra la tasa de positivos que fueron clasificados correctamente.

$$Exhaustividad = \frac{TP}{TP + FN} \tag{11}$$

- **Especificidad (Specificity):** Tasa de negativos clasificados correctamente.

$$Especificidad = \frac{TN}{TN + FP} \tag{12}$$

- **F1 Score:** La puntuación F1 congrega la precisión y la sensibilidad en una sola métrica (a través de una media armónica), sus valores oscilan entre el 0 y el 1, siendo mejor cuanto más cercano a 1 se encuentre [29].

$$F1 = 2 * \frac{precision\ * recall}{precision + recall} \tag{13}$$

3.5 Desequilibrio de clases

En ocasiones, tras analizar las métricas y resultados en los modelos de clasificación, se observa un rendimiento notablemente superior en cuanto a la predicción de una clase frente a otra. Esto puede ser debido a que el conjunto de datos de entrenamiento no se encuentra equilibrado, es decir, se encuentran más muestras de una clase que de otra, creando así una predisposición del modelo a predecir la clase mayoritaria.

Para tratar de paliar los efectos causados por el desbalanceo de clases, existen diversas posibilidades. Entre las opciones posibles destacan:

- **Modificación de los algoritmos:** Editar el funcionamiento de los algoritmos para que estos tengan en cuenta la diferencia de representación entre clases.
- **Oversampling:** Sobremuestreo en español, trata de aumentar la muestra de datos en el conjunto de entrenamiento creando artificialmente nuevas muestras de la clase minoritaria, de forma que finalmente haya el mismo número de muestras de todas las clases diferentes.
- **Undersampling:** En español submuestreo, trata de disminuir el tamaño de los datos de entrenamiento mediante la eliminación de muestras de la clase mayoritaria, hasta equilibrar las muestras de todas las clases [30].

4 CASO DE ESTUDIO

Este punto tiene como objetivo describir el desarrollo del trabajo, desde el inicio de una base datos hasta tener los datos preparados para ejecutar los modelos. Para el tratamiento de los datos, así como su posterior análisis se ha utilizado la programación en *Python*, más concretamente la librería *Pandas. Pandas* es una librería de dominio público diseñada para facilitar el tratamiento de los datos, usando sus diversas funciones se han realizado todos los cambios y transformaciones de la base de datos, además de generar las gráficas presentadas a continuación.

4.1 Origen de los datos

Debido a la gran cantidad de jugadores, posiciones y estadísticas disponibles, se estableció un criterio claro para acotar los datos que se iban a almacenar en la base de datos. Se decidió incluir en la base de datos a todos los delanteros centro presentes en La Liga Santander tras el cierre del mercado de invierno del año 2022, es decir, todos aquellos inscritos en La Liga a día 1 de febrero de 2022.
Una vez concretado el listado de jugadores, se procedió a la selección de las páginas web especializadas para realizar la extracción de datos. Finalmente, fueron escogidas dos webs de alta confianza y relevancia en el mundo del fútbol; dichas webs son: WhoScored, de donde se obtuvieron todos los datos referentes al rendimiento deportivo de cada delantero, y Transfermarkt, de donde fueron obtenidos los valores de mercado de cada futbolista durante las distintas temporadas.

Cada fila de la base de datos se compone de las variables categóricas: nombre del jugador, temporada, equipo, competición liguera en la que se encontraba su equipo; y las variables numéricas: edad en esa temporada, partidos en los que partió de titular, partidos en los que participó como suplente, minutos totales, además de las siguientes estadísticas:

Goles/90'	Promedio de goles marcados cada 90 minutos
Asistencias/90'	Promedio de goles asistidos cada 90 minutos
Amarillas/90'	Promedio de tarjetas amarillas cada 90 minutos
Rojas/90'	Promedio de tarjetas rojas cada 90 minutos
SpG	Tiros por partido
KeyP	Pases clave por partido
Drb	Regates con éxito por partido
Fouled	Faltas recibidas por partido
Off	Fueras de juego cometidos por partido
Disp	Pérdidas de posesión por partido
UnsTch	Controles mal efectuados por partido
AvgP	Promedio de pases por partido
PS%	Efectividad media de los pases
Crosses	Centros por partido
LongB	Balones en largo por partido
ThrB	Pases en profundidad por partido
Tackles	Entradas por partido
Inter	Intercepciones realizadas por partido
Fouls	Faltas cometidas por partido
Clear	Despejes por partido
DrbP	Promedio de veces que el jugador es regateado por partido
Blocks	Bloqueos por partido
AerialsWon	Duelos aéreos por partido
MotM	Número de veces nombrado jugador del partido

Tabla 4-1. Definición de las estadísticas de cada muestra de la base de datos.

4.2 Primer procesamiento

Tras el proceso de recogida de datos, se obtuvo un registro con 368 registros, correspondiente a 57 jugadores diferentes en distintas temporadas. Esta información requería un tratamiento para subsanar unas cuantas anomalías.

Primero, aunque se daba en contadas excepciones, existían futbolistas con más de una muestra por temporada. Esto se traduce en jugadores que son traspasados durante el mercado de invierno. Para ello se crearon las variables binarias 'Cambio_Equipo' y 'Cambio_Liga', que indican con un 1 si ha habido un cambio de equipo

y de liga a mitad de temporada, y con un 0 en caso contrario. Tras esto, las muestras con un 1 en 'Cambio_Equipo' fueron procesadas de manera que eran las resultantes de ponderar según los minutos jugados en cada equipo, las estadísticas de las dos muestras diferentes de la misma temporada, manteniendo en equipo y competición la información del último equipo.

Una vez obtenida un único registro por jugador y temporada se procedió a crear la variable binaria 'AValor', que reflejaba con un 1 un aumento en el valor de mercado del futbolista la temporada siguiente, y un 0 si no aumentaba. Debido a esto, a pesar de haber guardado las estadísticas de la última de las temporadas consecutivas de cada futbolista, hubo que borrar esas muestras, ya que no se conocían en el momento si su valor de mercado a final de temporada incrementó o no.

Después de este primer procesamiento del dataset, la base de datos quedó con 280 registros de 56 futbolistas diferentes.

4.3 Análisis de los datos

En la base de datos final, como se ha comentado, hay 280 muestras, distribuidas en once temporadas diferentes. A continuación, se presenta la distribución de todas las muestras por temporada en la *Tabla 4-2. Número de muestras por temporada.*

2009/2010	2010/2011	2011/2012	2012/2013	2013/2014	2014/2015	2015/2016	2016/2017	2017/2018	2018/2019	2019/2020
5	6	8	15	17	23	28	37	41	48	52

Tabla 4-2. Número de muestras por temporada.

Es destacable como las temporadas se concentran mayoritariamente en los últimos años, esta información hace intuir que la edad media de los jugadores no debe ser muy alta, como se verá a continuación.

Como ya se sabe, todas estas muestras se corresponden con las temporadas deportivas de 56 delanteros centro diferentes, con una media de cinco temporadas por futbolista. En el siguiente histograma se refleja en el eje de abscisas el número de temporadas, y en el de ordenadas los delanteros que tienen ese número de temporadas.

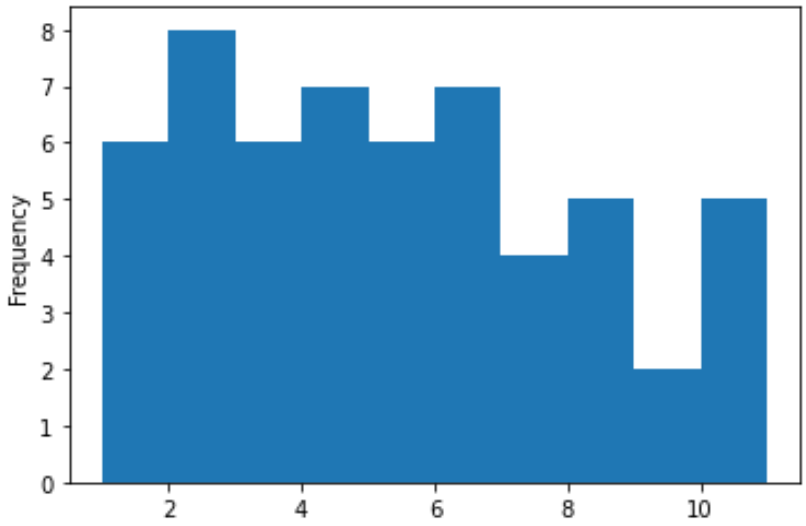


Figura 4-1. Histograma del número de temporadas por jugador.

Estas temporadas fueron disputadas en 71 equipos de 13 competiciones ligueras distintas. Entre estos equipos destacan Valencia y Real Madrid con 17 registros, seguidos de Barcelona y Celta de Vigo con 13. Con respecto a las competiciones, hay una diferencia abismal entre la primera, *LaLiga* de España, contra la segunda, la *Premier League* de Inglaterra, con un número total de registros de 178 y 29 respectivamente.

En cuanto a las edades de los jugadores, se tiene una media de 25,6 años, dato que esclarece la alta concentración de temporadas recientes frente a la escasa presencia de temporadas más antiguas. El espectro de edades oscila los 38 años, y los 17, con una mediana de 25 años. Se indexa a continuación el histograma con el reparto de edades.

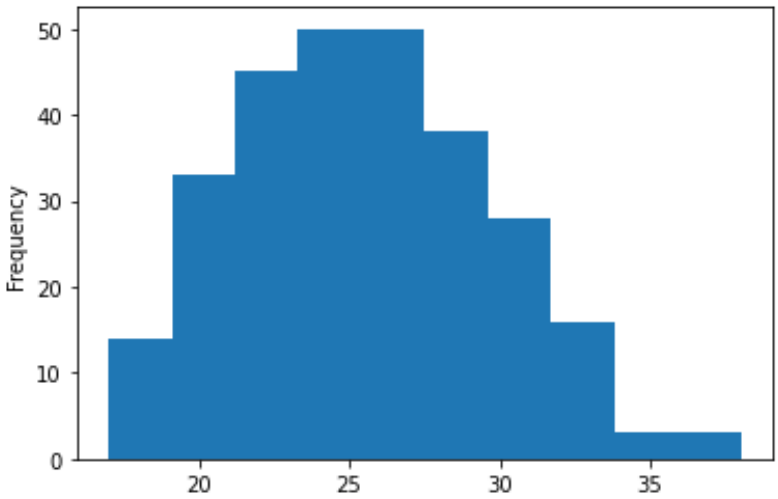


Figura 4-2. Histograma de las edades de los jugadores.

El valor de mercado de todos los jugadores, a lo largo de los 280 registros se concentra en valores que no superan la decena de millones de euros. El 80% de los jugadores tienen un valor entre los quinientos mil y los siete millones de euros, cifras que se pueden considerar relativamente bajas considerando el mercado actual y la comparativa con otras ligas extranjeras. Se presenta a continuación el histograma con la distribución de los valores de mercado en todas las muestras disponibles, donde se puede apreciar como aproximadamente 140 muestras tienen un valor de mercado inferior a los 10 millones de euros.

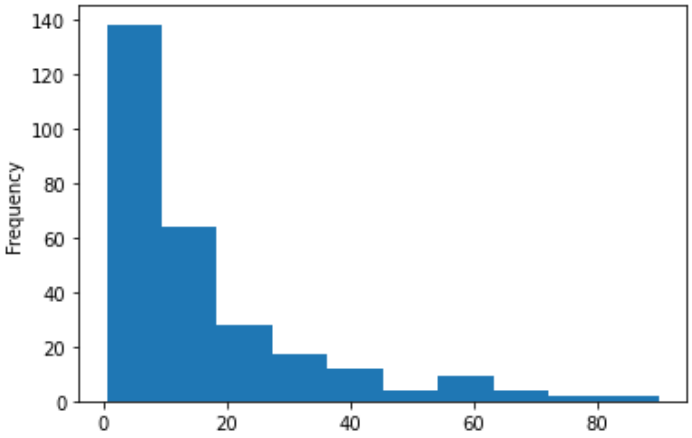


Figura 4-3. Distribución del valor de mercado de todos los registros.

Todas las estadísticas han sido obtenidas de un total de 508.853 minutos de juego, repartidos en 7681 partidos diferentes. A continuación, se incluye el histograma con el reparto de minutos por registro.

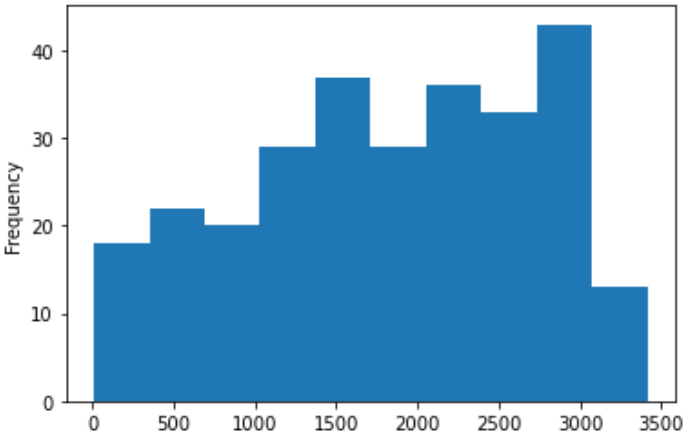


Figura 4-4. Histograma de los minutos jugados por temporada.

Es remarcable sobre el histograma anterior cómo de homogeneizado se encuentra el reparto de minutos, esto encuentra su lógica en que las temporadas representan la evolución profesional de cada delantero, en sus primeros años contará con pocos minutos hasta asentarse deportivamente en el escalafón más alto de sus equipos. Además, los futbolistas no son máquinas y sufren lesiones que impiden su participación en partidos.

Se presenta a continuación la comparativa entre los partidos en los que los delanteros partieron como titulares frente aquellos en que partieron desde el banquillo.

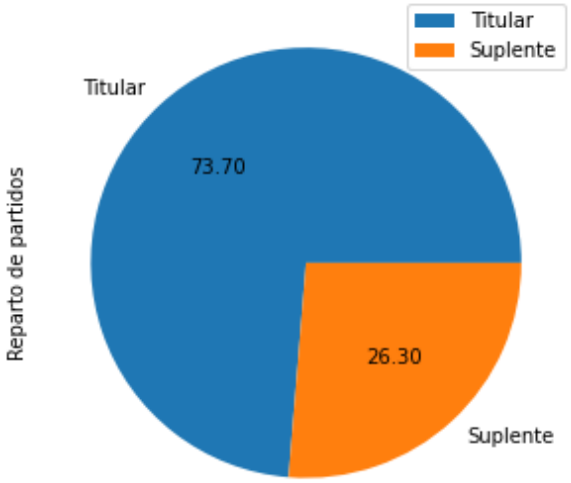


Figura 4-5. Porcentaje de los partidos en los que los jugadores comenzaron como titulares y suplentes.

Se obtienen un 73,70% de los partidos como titulares frente a 26,30% de suplentes, es un dato bastante interesante, ya que define mayoritartiamente a los futbolistas presentes en esta base de datos como actores principales en los equipos donde han estado presentes a lo largo de su carrera deportiva.

Entre todas las temporadas recogidas, se ha obtenido una mayoría de no aumentos en el valor de mercado la temporada siguiente, siendo su reparto:

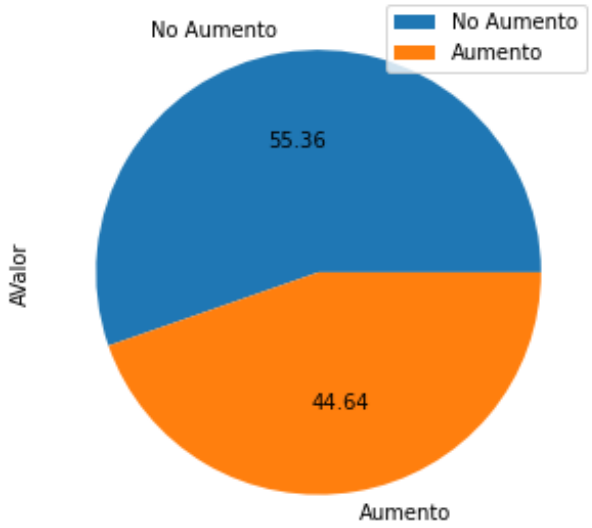


Figura 4-6. Resultados percentuales AValor.

Como se puede observar en el gráfico superior, un 55,36% de las muestras en la base de datos presentan un 0 en la variable AValor (no aumento), frente a un 44,64% de las mismas que sí aumentaron su valor de mercado en la temporada inmediatamente consecutiva. Al no ser una mayoría aplastante se considera bastante equilibrado el reparto entre los aumentos y no aumentos presentes en los datos.

Para el resto de las variables cuantitativas, correspondientes a las estadísticas por partido de cada delantero, se encuentra la siguiente ilustración que refleja los histogramas de cada una de dichas variables, además de la tabla resumen plasmando el valor máximo, mínimo (mayor que cero) y la moda de cada estadística, así como el número total de valores nulos que tiene cada una de las variables.

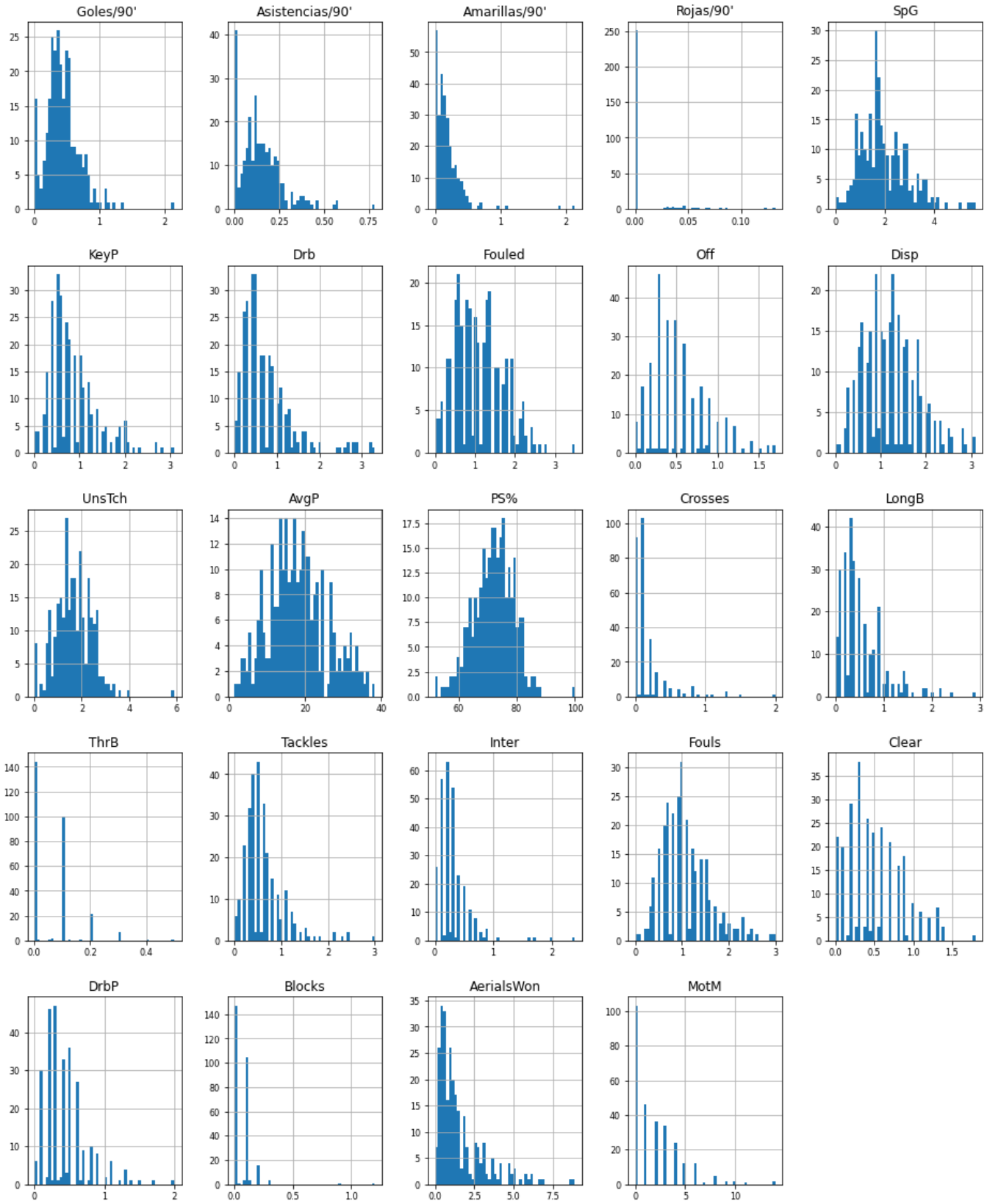


Figura 4-7. Histogramas de las estadísticas por partido.

	Mínimo	Máximo	Media	Moda	Valores Nulos
Goles/90'	0,06	2,14	0,436	0	16
Asistencias/90'	0,03	0,78	0,146	0	41
Amarillas/90'	0,03	2,14	0,185	0	48
Rojas/90'	0,03	0,13	0,005	0	251
SpG	0,10	5,70	1,984	1,8	1
KeyP	0,10	3,10	0,847	0,5	4
Drb	0,07	3,30	0,734	0,4	6
Fouled	0,10	3,50	1,096	0,6	4
Off	0,01	1,70	0,532	0,3	7
Disp	0,10	3,10	1,226	1,3	1
UnsTch	0,10	5,90	1,666	2,3	7
AvgP	1,40	38,30	18,123	20,4	0
PS%	51,90	100,00	72,045	69	0
Crosses	0,07	2,00	0,167	0,1	92
LongB	0,05	2,90	0,529	0,3	13
ThrB	0,01	0,50	0,063	0	144
Tackles	0,10	3,00	0,607	0,5	6
Inter	0,10	2,40	0,295	0,2	26
Fouls	0,10	3,00	1,053	1	1
Clear	0,07	1,80	0,502	0,3	22
DrbP	0,10	2,00	0,433	0,3	6
Blocks	0,05	1,20	0,062	0	147
AerialsWon	0,10	8,90	1,578	0,4	3
MotM	1,00	14,00	2,043	0	103

Tabla 4-3. Estadísticas de las distintas variables utilizadas en el estudio.

De la tabla anterior hay que realizar un matiz importante, cuando se habla de valores nulos hay que diferenciar claramente dos tipos: por un lado se tienen valores nulos "reales", estos son los relativos a los goles, asistencias, tarjetas y Motm, estos valores son nulos porque su valor es verdaderamente 0 y están documentados; por otro lado se encuentran el resto de estadísticas, pueden tratarse de *missing values*, es decir, que son cero porque se desconoce su valor real, o puede ser que su valor es tan pequeño que por aproximación se marque como nulo.

De los histogramas y la tabla caben destacar los siguientes resultados:

- **Rojas/90':** Es la estadística con mayor número de valores nulos, además puede observarse como su histograma, para valores distintos de cero está muy repartido en cifras muy pequeñas, con su máximo en algo más de las 0,13 unidades. La lógica de esto reside en que estadísticamente los delanteros centro reciben muy pocas expulsiones a lo largo de sus carreras, además de que al ponderar el número de expulsiones por los minutos jugados se obtienen resultados muy pequeños.
- **ThrB:** Su histograma plasma, no solo sus 144 valores nulos, sino también muy bien la poca concentración en sus resultados, eso sí, algo más variados que los de las tarjetas rojas. Todo esto se entiende con un poco de contexto futbolístico, y es que el delantero centro es el último hombre en ataque, en contadas ocasiones tendrá un compañero por delante al que darle un balón en profundidad. Esto invita a pensar que sus valores nulos mayoritariamente responderán a que su cifra real es tan pequeña que se redondea a cero.
- **Blocks:** También se mueve en valores muy bajos, y con escasa concentración, siendo el segundo parámetro con más valores nulos. Generalmente, las implicaciones defensivas de los delanteros es más una tarea posicional, ejerciendo una presión a los defensas rivales para que estos no saquen la pelota jugada y se vean obligados a sortearla, pero pocas veces bajarán a bloquear tiros de los rivales, salvo en contadas ocasiones como jugadas a balón parado.
- **Crosses:** Sigue una distribución parecida a los anteriores, algo más concentrada en determinados puntos y con más variedad de valores. Esto también obedece a una lógica futbolística, generalmente el delantero centro es el receptor de los centros, por lo que en contadas ocasiones será el quien los sirva.

A continuación, se exponen las correlaciones entre todas las variables numéricas, aunque primero hay que realizar una pequeña transformación en la base de datos: la columna *Temporada* se transforma desde categórica a numérica, asignando un valor desde el 0 (2020/2021) hasta el 10 (2009/2010) a cada temporada; además, se muestra, por simpificación la mitad superior de la matriz, pues al ser simétrica, los valores ocultos no son desconocidos.

	Temporada	Edad	Valor	mbio_Equi	Cambio_Liga	Titular	Suplente	Mins	Goles	Asistencias	Amarillas	Rojas	SpG	KeyP	Drb	Fouled	Off	Disp	UnsTch	AvgP	PS%	Crosses	LongB	ThrB	Tackles	Inter	Fouls	Clear	DrbP	Blocks	AerialsWon	MotM	AValor
Temporada	1	-0,25	0,03	-0,07	-0,03	0,02	-0,07	0,01	0,02	0,15	-0,03	0,08	0,19	0,18	0	0,07	0,17	0,17	-0,22	0,2	0,13	0,08	0,13	0,35	0,21	0,37	-0,01	-0,01	-0,2	0,08	-0,16	0,01	0,22
Edad		1	0,1	-0,04	0,03	0,34	-0,11	0,33	0,17	0,11	0,09	0	0,14	0,17	-0,11	0,1	0,2	-0,14	0,1	0,23	-0,15	-0,01	0,17	0,11	0,04	-0,01	0,13	0,2	0,16	0,06	0,16	0,22	-0,31
Valor			1	0,01	0,07	0,4	-0,32	0,39	0,46	0,37	-0,12	-0,06	0,5	0,44	0,28	0,1	0,47	0,19	0,13	0,4	0,31	0,18	0,24	0,29	-0,08	-0,08	-0,09	0,04	-0,08	0,25	-0,19	0,37	-0,19
Cambio_Equipo				1	0,74	-0,08	0,14	-0,07	-0,09	-0,01	-0,09	-0,04	-0,07	-0,08	0	-0,04	-0,07	-0,06	-0,05	-0,09	0,03	-0,02	-0,11	-0,08	-0,02	0,01	-0,04	-0,07	-0,04	0,03	-0,04	-0,09	-0,07
Cambio_Liga					1	-0,07	0,08	-0,07	-0,05	0,04	-0,1	-0,01	-0,06	-0,04	-0,07	-0,06	-0,04	-0,06	-0,06	-0,04	0,11	0,01	-0,09	-0,06	-0,08	-0,03	-0,12	-0,08	-0,07	-0,01	-0,04	-0,07	-0,08
Titular						1	-0,55	0,99	0,18	0,16	-0,13	0,06	0,65	0,54	0,33	0,5	0,49	0,48	0,5	0,66	-0,06	0,28	0,44	0,36	0,25	0,23	0,33	0,42	0,32	0,24	0,24	0,61	-0,08
Suplente							1	-0,49	-0,22	-0,19	-0,02	0,01	-0,56	-0,51	-0,38	-0,42	-0,37	-0,42	-0,41	-0,56	-0,16	-0,25	-0,36	-0,33	-0,24	-0,15	-0,25	-0,25	-0,26	-0,21	-0,1	-0,5	0,07
Mins								1	0,18	0,15	-0,14	0,05	0,65	0,53	0,33	0,49	0,49	0,48	0,51	0,64	-0,09	0,27	0,43	0,34	0,25	0,23	0,32	0,42	0,33	0,24	0,26	0,61	-0,08
Goles									1	0,2	0,15	-0,07	0,43	0,22	0,03	-0,01	0,28	-0,02	-0,01	0,11	0,17	-0,01	0,06	0,2	-0,19	-0,16	-0,11	0	-0,15	0,11	-0,1	0,39	-0,11
Asistencias										1	-0,13	0	0,26	0,49	0,18	0,02	0,27	0,09	0,05	0,34	0,23	0,27	0,2	0,33	0,05	0,05	-0,11	-0,04	-0,03	0,2	-0,16	0,23	0
Amarillas											1	-0,02	-0,14	-0,17	-0,17	0,05	-0,14	-0,2	-0,1	-0,12	-0,14	-0,07	-0,02	-0,05	0,04	0,03	0,2	-0,03	0,02	-0,01	-0,01	-0,07	-0,03
Rojas												1	0	0,01	-0,03	0,1	-0,05	0,02	-0,05	0,05	-0,11	0,01	0,13	0,05	0,08	0,13	0,06	0,08	-0,02	0	0,01	0,01	0,02
SpG													1	0,67	0,48	0,47	0,49	0,52	0,5	0,68	0,05	0,34	0,48	0,48	0,23	0,14	0,25	0,27	0,19	0,19	0,17	0,71	-0,01
KeyP														1	0,54	0,32	0,39	0,48	0,36	0,76	0,22	0,61	0,53	0,64	0,2	0,17	0,01	0,05	0,18	0,19	-0,06	0,57	0
Drb															1	0,36	0,16	0,67	0,59	0,5	0,15	0,46	0,41	0,34	0,23	0,11	0,15	0	0,35	0,1	-0,14	0,39	0,07
Fouled																1	0,24	0,53	0,53	0,57	-0,16	0,19	0,43	0,28	0,45	0,35	0,54	0,29	0,33	0,21	0,29	0,44	-0,06
Off																	1	0,25	0,25	0,28	0,06	0,12	0,16	0,29	-0,04	-0,06	0,15	0,11	-0,05	0,26	-0,06	0,27	-0,06
Disp																		1	0,68	0,56	0,11	0,35	0,36	0,33	0,27	0,16	0,34	0,23	0,27	0,14	0,07	0,35	0,03
UnsTch																			1	0,45	-0,09	0,26	0,21	0,18	0,21	-0,01	0,46	0,32	0,33	0,17	0,22	0,4	-0,09
AvgP																				1	0,06	0,41	0,71	0,52	0,49	0,44	0,27	0,36	0,43	0,22	0,24	0,63	-0,09
PS%																					1	0,14	0,03	0,1	-0,3	-0,17	-0,42	-0,39	-0,26	-0,09	-0,59	-0,04	-0,03
Crosses																						1	0,33	0,32	0,16	0,18	-0,03	-0,24	0,24	0	-0,25	0,22	0,02
LongB																							1	0,44	0,38	0,44	0,2	0,3	0,32	0,16	0,05	0,48	-0,01
ThrB																								1	0,22	0,19	0,08	0,07	0,03	0,12	-0,1	0,45	0,05
Tackles																									1	0,76	0,44	0,28	0,59	0,18	0,29	0,25	0,06
Inter																										1	0,26	0,2	0,43	0,14	0,12	0,16	0,1
Fouls																											1	0,4	0,32	0,21	0,41	0,23	-0,08
Clear																												1	0,32	0,28	0,64	0,3	-0,07
DrbP																													1	0,05	0,35	0,26	-0,05
Blocks																														1	0,11	0,22	0,02
AerialsWon																															1	0,3	-0,08
MotM																																1	-0,06
AValor																																	1

Tabla 4-4. Matriz de correlación de las variables.

En la matriz anterior, el código de colores funciona de la siguiente manera: los tonos más verdes se corresponden con valores cercanos al 1, los tonos amarillentos con valores cercanos al 0 y, los tonos rojos, con valores próximos a -1. En cuanto a los resultados reflejados, se comentan a continuación algunos de los más destacables:

- Cambio_Equipo y Cambio_Liga: es bastantante esperable ese 0,74 de correlación, ya que siempre que hay un cambio de liga implica que previamente ha habido un cambio de equipo.
- Titular y Mins: Son las dos variables más correlacionadas, con un 0,99. Se entiende que cuantos más partidos juegue un futbolista de titular, más minutos juega. En contraposición, se encuentran Suplente y Mins, con un -0,49, ya que las participaciones desde el banquillo suman pocos minutos por lo general.
- Titular y Suplente: con un -0,55, es un resultado lógico de esperar, cada partido que juegue un delantero de titular es un partido menos que juega como suplente, por eso se encuentran negativamente correlacionadas.
- Mins y SpG: se entiende esa alta correlación (0,65) puesto que cuantos más minutos dispute un futbolista, más oportunidades de chutar a puerta tendrá.
- Suplente y SpG: -0,56, de manera contaria a Mins y SpG, cuantos menos minutos esté en el campo un futbolista, menos oportunidades de tirar a puerta tendrá.
- SpG y KeyP: con un sorprendente 0,67 de correlación, a priori no se encuentran motivos claros.
- SpG y Motm: cuantas más oportunidades de gol genere un delantero, mejor valorado estará. Por ello, los tiros a puerta por partido y los galardones a hombre del encuentro tienen un índice de correlación de 0,71.
- Drb y Disp: 0,67, se entiende que los jugadores que más encaran a los rivales son más valientes con el balón y por ello pierden más la posesión.
- Disp y UnsTch: sorprende que solo sea un 0,68, ya que si un jugador efectúa un mal control con casi toda seguridad perderá la posesión del balón.
- AvgP y Titular: con un 0,66, se entiende que cuantos más minutos dispute el futbolista más pases podrá efectuar.
- AvgP y SpG: sorprende esta relación comn un índice de 0,68, no responde a priori a ninguna cuestión lógica.
- AvgP y KeyP: 0,76, podría interpretarse que el perfil de delanteros que están más en contacto con el balón, al asociarse más con sus compañeros generarán con más frecuencia grandes ocasiones fruto de dichas sociedades.

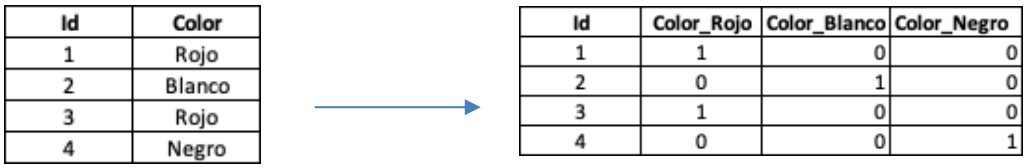
- AvgP y LongB: sucede de manera similar al anterior. Con un índice de 0,71 se supone que, al tratarse de jugadores con un perfil asociativo, estos estarán más dispuestos a surtir de balones a sus compañeros.
- Inter y Tackles: 0,76, a priori una correlación lógica. El que un futbolista realice muchas intercepciones le define como un delantero con vocación defensiva, por ello no sorprende que cuantas más intercepciones realice, más entradas efectúe.

4.4 Preparación de los distintos dataframes

Una vez organizada y analizada la base de datos, el siguiente paso a seguir es preparar los datasets para aplicar los diferentes modelos. No solo es bastante importante qué datos son los que se añaden al conjunto de entrenamiento, sino como están organizados dichos datos, las posibilidades son infinitas y pueden hacer variar sensiblemente los resultados. Por ello se proponen tres dataframes distintos, cada uno agrupando y organizando los datos de manera diferente:

- **Dataframe1:** Cada fila del dataframe arrastra información de tres temporadas consecutivas de un delantero. Se agrupa en la misma fila todas las columnas de dos temporadas consecutivas del mismo jugador, sin repetir el nombre y omitiendo *AValor*, finalmente, se añade la columna *AValor* correspondiente a la tercera temporada. Cada fila queda de la siguiente forma: nombre + datos temporada 1 + datos temporada 2 + aumento o no del valor en la temporada 3. Quedó un dataframe con 217 filas y 70 columnas.
- **Dataframe2:** Con el objetivo de comprobar si el modelo será más preciso cuando se le añade más historia a cada fila del dataframe, se crea otro análogo al anterior, pero incluyendo cuatro temporadas consecutivas de la siguiente manera: nombre + datos temporada 1 + datos temporada 2 + datos temporada 3 + aumento o no del valor en la temporada 4. Este dataframe hace que se sacrifiquen algunos datos referentes a jugadores que carecen de cuatro temporadas consecutivas, obteniendo un total de 167 filas y 104 columnas.
- **Dataframe3:** Al igual que con el primer dataframe se almacenan los datos y estadísticas de dos temporadas consecutivas y se ve si aumentó o no su valor de mercado en una tercera, pero para comprobar si la organización de los datos infiere mucho o no en los resultados, los datos numéricos de las temporadas serán una variación entre lo obtenido en la temporada 2 menos lo obtenido en la temporada 1. Por ejemplo, en el campo *Minutos*, si en la temporada 1 el jugador participó un total de 1234 minutos frente a 2341 en la siguiente temporada, se reflejarán 1107 minutos en el dataframe. El tamaño del dataframe fue de 217 filas y 42 columnas.

Tras crear los diferentes dataframes, queda únicamente preparar las variables categóricas de manera que puedan ser aceptadas por el modelo. Todas ellas, salvo la ya previamente tratada *Temporada,* se convierten en variables *dummy*, esto supone que se genera una nueva columna por categoría, los valores de cada columna son binarios: 1 si forma parte de la categoría y 0 en caso contrario. A continuación, se ejemplifica como se transforma una variable categórica en *dummy.*



Id	Color
1	Rojo
2	Blanco
3	Rojo
4	Negro

Id	Color_Rojo	Color_Blanco	Color_Negro
1	1	0	0
2	0	1	0
3	1	0	0
4	0	0	1

Figura 4-8. Ejemplo de transformación de variable categórica a dummy.

Finalmente, quedaron preparados los dataframes para ser implementados en los diferentes modelos para buscar obtener el mejor resultado posible de predicción en los cambios del valor de mercado. El tamaño de los dataframes fue finalmente:

Dataframe	Filas	Columnas
1	217	244
2	167	307
3	217	216

Tabla 4-5. Tamaño final dataframes.

4.5 División de datos en *train* y *test*

Como ya ha sido comentado, los algoritmos de ML utilizan el conjunto de datos de dos maneras diferentes. Una parte, mayoritaria, se utiliza como entrenamiento o *train* del propio algoritmo, a partir de este conjunto se identifican los patrones y variables que permiten ajustar los diferentes modelos. Una vez generados los modelos, se prueba la precisión de estos usando el conjunto *test*, ya que al conocer la salida correcta permite medir el nivel de precisión adquirido por el modelo, además de posibilitar un aprendizaje de sus propios errores.

En el caso de estudio, se decidió realizar la división de los conjuntos de datos según un criterio cronológico, estableciendo el corte *train/test* de manera que se entrenaran los modelos con los datos más antiguos, y se evaluaran con los más recientes. Siguiendo esta premisa se procedió a analizar las posibles divisiones, manteniendo una filosofía cercana al 80% *train* y 20% *test*. Finalmente, tras estudiar diversas posibilidades, se estableció el corte en la muestra más reciente, es decir, las filas de todos los dataframes que incluyesen información de la temporada 2019/2020 (temporada 0 tras el cambio de categórica a cuantitativa), fueron incluidas en el conjunto *test*. A continuación, se muestra el número de filas en cada conjunto, así como el porcentaje que suponen de cada dataframe.

Dataframe	**Filas *train***	**Porcentaje *train***	**Filas *test***	**Porcentaje test**
1	172	79,26%	45	20,74%
2	136	81,44%	31	18,56%
3	172	79,26%	45	20,74%

Tabla 4-6. Resumen conjuntos train y test.

5 RESULTADOS

Se desarrolla en este capítulo el análisis de los resultados obtenidos tras la implementación de los distintos modelos y su ejecución para los tres dataframes diferentes. Para la construcción de los modelos se hizo uso de la librería *Scikit-learn*, una librería código abierto de *Python* que contiene herramientas para el análisis predictivo de datos [31]. Haciendo uso de las métricas de evaluación desarrolladas en el capítulo 3 del trabajo, se tratará de cuantificar la calidad de los resultados obtenidos. Además, con el fin de mejorar dichos resultados se provarán diversas modificaciones de los datasets, consistentes en la omisión de algunas variables que podrían no resultar relevantes para obtener modelos más precisos.

5.1 Primeros resultados

Tras realizar la division de los datasets en los conjuntos de prueba (*test*) y entrenamiento (*train*), se procedió a construir en *Python* los modelos: Regresión Logística y Random Forest.

Los resultados obtenidos al implementar los modelos de Regresión Logística y Random Forest en los tres dataframes diferentes no fueron demasiado alentadores, principalmente en cuanto a la previsión de los aumentos de valor (clasificaciones positivas). A continuación, se muestran las matrices de confusión, así como los resultados de las métricas en los 6 resultados diferentes:

		Dataframe 1					
Modelo		**Train**			**Test**		
			Predicción			**Predicción**	
LR			**0**	**1**		**0**	**1**
	Real	**0**	82	17	**0**	31	0
		1	26	47	**1**	13	1
RF			**0**	**1**		**0**	**1**
	Real	**0**	99	0	**0**	29	2
		1	0	73	**1**	13	1

Tabla 5-1. Matrices de confusión Dataframe 1 (dos temporadas).

		Dataframe 2					
Modelo		**Train**			**Test**		
			Predicción			**Predicción**	
LR			**0**	**1**		**0**	**1**
	Real	**0**	70	11	**0**	27	0
		1	15	32	**1**	12	0
RF			**0**	**1**		**0**	**1**
	Real	**0**	81	0	**0**	27	0
		1	0	47	**1**	12	0

Tabla 5-2. Matrices de confusión Dataframe 2 (tres temporadas).

		Dataframe 3					
Modelo		**Train**			**Test**		
			Predicción			**Predicción**	
LR			**0**	**1**		**0**	**1**
	Real	**0**	82	17	**0**	29	2
		1	26	47	**1**	14	0
RF			**0**	**1**		**0**	**1**
	Real	**0**	99	0	**0**	29	2
		1	0	73	**1**	13	1

Tabla 5-3. Matrices de confusión Dataframe 3 (dos temporadas con variaciones).

		Train				
		Ac	**Prec**	**Spec**	**Rec**	**F-Score**
LR	Data1	0,75	0,73	0,83	0,64	0,69
	Data2	0,80	0,74	0,86	0,68	0,71
	Data3	0,75	0,73	0,83	0,64	0,69
RF	Data1	1	1	1	1	1
	Data2	1	1	1	1	1
	Data3	1	1	1	1	1

Tabla 5-4. Resultados iniciales conjunto train.

		Test				
		Ac	**Prec**	**Spec**	**Rec**	**F-Score**
LR	Data1	0,71	1,00	1,00	0,07	0,13
	Data2	0,69	0,00	1,00	0,00	0,00
	Data3	0,64	0,00	0,94	0,00	0,00
RF	Data1	0,67	0,33	0,94	0,07	0,12
	Data2	0,69	0,00	1,00	0,00	0,00
	Data3	0,67	0,33	0,94	0,07	0,12

Tabla 5-5. Resultados iniciales conjunto test

Antes de entrar en el análisis de los primeros resultados del conjunto de prueba, cabe destacar que, cuando se realizan predicciones sobre el conjunto de entrenamiento en un Random Forest, este siempre realizará clasificaciones perfectas, como se puede observar en los resultados. Entre los resultados de las clasificaciones del conjunto de entrenamiento correspondientes a la Regresión Logística, se obtienen buenos resultados en general, aunque notablemente mejores las clasificaciones negativas (0) frente a las positivas (1). Además, destaca que en los tres dataframes disferentes existe una mayoría de muestras negativas.

Con relación a los resultados del conjunto de prueba, se observan unos muy buenos resultados en cuanto a las clasificaciones negativas, mientras que los resultados de las clasificaciones positivas resultaron nefastos. Únicamente en dos ocasiones se consigue obtener un resultado verdadero positivo, no superando la unidad, en dos ocasiones.

5.2 Exploración de nuevos resultados

Tras los malos resultados iniciales y, con el propósito de obtener alguna mejora, se plantearon diferentes posibilidades de cambios en los dataframes:

- **Undersampling:** Tras ver la diferencia de representación de muestras positivas y negativas en los conjuntos de entrenamiento, se pensó que el problema de mala clasificación podría responder a un desbalanceo de clases. Analizando las diferentes soluciones posibles para afrontar el desbalanceo de clases, y viendo los resultados obtenidos en estudios que trataban el tema [32], se decidió apostar por el *undersampling*.
- **Codificación de las variables categóricas:** Otra posible modificación podía ser codificar las variables categóricas, en lugar de utilizar *dummies* como se hizo inicialmente, como *label*, donde cada categoría diferente se corresponde con un valor numérico entero. A continuación, se muestra un ejemplo de transformación de variable categórica a *label.*

Id	Color
1	Rojo
2	Blanco
3	Rojo
4	Negro

Id	Color
1	0
2	1
3	0
4	2

Figura 5-1. Ejemplo de transformación de variable categórica a label.

- **Reducción del número de variables:** Se propusieron tres conjuntos diferentes de variables siguiendo un criterio de eliminación hacia atrás, procedimiento de selección de varibles donde del coniunto inicial se extraen aquellas que no cumplan un criterio de correlación mínimo con la variable independiente [33]. El conjunto primero contaba con todas las variables almacenadas en los dataframes originales; el

segundo estaba formado por las diez variables, sin contar con la información básica del jugador y temporada, con más correlación con la variable de salida; y, finalmente, el tercer conjunto estaba conformado por las cinco variables más correlacionadas con *AValor*.

Con todas estas propuestas de cambios, se procedió a realizar las 24 combinaciones posibles por dataframe y ejecutar los resultados, que fueron almacenados en una hoja de cálculo *Excel*. A cada combinación se le asignó un ID, que será con el que se identifiquen posteriormente los resultados. Se muestra a continuación, la tabla con los 24 códigos y escenarios diferentes, cabe destacar que estos identificadores son los mismos para cada dataframe, variando únicamente los resultados:

ID	Modelo	UnderSampling	Cod_Cat	Vbles
0	LR	Si	Label	1
1	LR	Si	Label	2
2	LR	Si	Label	3
3	LR	Si	Dummy	1
4	LR	Si	Dummy	2
5	LR	Si	Dummy	3
6	LR	No	Label	1
7	LR	No	Label	2
8	LR	No	Label	3
9	LR	No	Dummy	1
10	LR	No	Dummy	2
11	LR	No	Dummy	3
12	RF	Si	Label	1
13	RF	Si	Label	2
14	RF	Si	Label	3
15	RF	Si	Dummy	1
16	RF	Si	Dummy	2
17	RF	Si	Dummy	3
18	RF	No	Label	1
19	RF	No	Label	2
20	RF	No	Label	3
21	RF	No	Dummy	1
22	RF	No	Dummy	2
23	RF	No	Dummy	3

Tabla 5-6. Conjunto de escenarios simulados en función del modelo, la técnica de muestreo, la codificación de las variables categóricas y el conjunto de variables de entrada.

En la tabla anterior, la columna modelo indica si se aplicó Regresión Logística (LR) o Random Forest (RF); en la columna UnderSampling, si se aplicó (Si) o no (No) *undersampling* a los datos de entrenamiento; en Cod_Cat se indica el tipo de codificación de las variables categóricas; y, finalmente Vbles indica que combinación de variables, de las detalladas anteriormente, se han utilizado.

5.2.1 Análisis de resultados

Con todo ello, se ejecutaron las 24 combinaciones en cada dataframe, obteniendo 72 resultados en total. Se presentan a continuación, las matrices de confusión por dataframe, así como los resultados de las métricas.

ID			Train			Test	
			Predicción			Predicción	
			0	1		0	1
0	Real	0	54	19	0	30	1
		1	22	51	1	11	3
			Predicción			Predicción	
			0	1		0	1
1	Real	0	56	17	0	31	0
		1	26	47	1	13	1
			Predicción			Predicción	
			0	1		0	1
2	Real	0	49	24	0	30	1
		1	21	52	1	11	3
			Predicción			Predicción	
			0	1		0	1
3	Real	0	55	18	0	30	1
		1	18	55	1	11	3
			Predicción			Predicción	
			0	1	1	0	1
4	Real	0	56	17	0	31	0
		1	26	47	1	13	1
			Predicción			Predicción	
			0	1		0	1
5	Real	0	63	10	0	31	0
		1	9	64	1	13	1
			Predicción			Predicción	
			0	1		0	1
6	Real	0	79	20	0	30	1
		1	32	41	1	12	2

			Predicción			Predicción	
7			**0**	**1**		**0**	**1**
	Real	**0**	83	16	**0**	31	0
		1	33	40	**1**	13	1
			Predicción			**Predicción**	
8			**0**	**1**		**0**	**1**
	Real	**0**	79	20	**0**	31	0
		1	32	41	**1**	14	0
			Predicción			**Predicción**	
9			**0**	**1**		**0**	**1**
	Real	**0**	82	17	**0**	31	0
		1	26	47	**1**	13	1
			Predicción			**Predicción**	
10			**0**	**1**		**0**	**1**
	Real	**0**	78	21	**0**	31	0
		1	33	40	**1**	13	1
			Predicción			**Predicción**	
11			**0**	**1**		**0**	**1**
	Real	**0**	91	8	**0**	31	0
		1	18	55	**1**	14	0
			Predicción			**Predicción**	
12			**0**	**1**	**1**	**0**	**1**
	Real	**0**	73	0	**0**	26	5
		1	0	73	**1**	10	4
			Predicción			**Predicción**	
13			**0**	**1**		**0**	**1**
	Real	**0**	73	0	**0**	28	3
		1	0	73	**1**	10	4
			Predicción			**Predicción**	
14			**0**	**1**		**0**	**1**
	Real	**0**	73	0	**0**	28	3
		1	0	73	**1**	11	3
			Predicción			**Predicción**	
15			**0**	**1**		**0**	**1**
	Real	**0**	73	0	**0**	28	3
		1	0	73	**1**	10	4
			Predicción			**Predicción**	
16			**0**	**1**		**0**	**1**
	Real	**0**	73	0	**0**	27	4
		1	0	73	**1**	10	4
			Predicción			**Predicción**	
17			**0**	**1**		**0**	**1**
	Real	**0**	73	0	**0**	24	7
		1	0	73	**1**	11	3
			Predicción			**Predicción**	

18			**0**	**1**		**0**	**1**
	Real	**0**	99	0	**0**	29	2
		1	0	73	**1**	12	2
			Predicción			**Predicción**	
19			**0**	**1**		**0**	**1**
	Real	**0**	99	0	**0**	28	3
		1	0	73	**1**	12	2
			Predicción			**Predicción**	
20			**0**	**1**	**1**	**0**	**1**
	Real	**0**	99	0	**0**	27	4
		1	0	73	**1**	11	3
			Predicción			**Predicción**	
21			**0**	**1**		**0**	**1**
	Real	**0**	99	0	**0**	29	2
		1	0	73	**1**	13	1
			Predicción			**Predicción**	
22			**0**	**1**		**0**	**1**
	Real	**0**	99	0	**0**	28	3
		1	0	73	**1**	13	1
			Predicción			**Predicción**	
23			**0**	**1**		**0**	**1**
	Real	**0**	99	0	**0**	28	3
		1	0	73	**1**	12	2

Tabla 5-7. Matrices de confusión dataframe 1 distintos escenarios.

ID			**Train**			**Test**	
			Predicción			**Predicción**	
0			**0**	**1**		**0**	**1**
	Real	**0**	35	12	**0**	24	3
		1	10	37	**1**	10	2
			Predicción			**Predicción**	
1			**0**	**1**		**0**	**1**
	Real	**0**	36	11	**0**	24	3
		1	16	31	**1**	11	1
			Predicción			**Predicción**	
2			**0**	**1**		**0**	**1**
	Real	**0**	37	10	**0**	25	2
		1	12	35	**1**	10	2
			Predicción			**Predicción**	
3			**0**	**1**		**0**	**1**
	Real	**0**	37	10	**0**	27	0
		1	9	38	**1**	12	0

			Predicción			Predicción	
4			**0**	**1**	**1**	**0**	**1**
	Real	**0**	32	15	**0**	23	4
		1	14	33	**1**	11	1
			Predicción			**Predicción**	
5			**0**	**1**		**0**	**1**
	Real	**0**	42	5	**0**	26	1
		1	2	45	**1**	10	2
			Predicción			**Predicción**	
6			**0**	**1**		**0**	**1**
	Real	**0**	67	14	**0**	27	0
		1	12	35	**1**	11	1
			Predicción			**Predicción**	
7			**0**	**1**		**0**	**1**
	Real	**0**	71	10	**0**	26	1
		1	17	30	**1**	11	1
			Predicción			**Predicción**	
8			**0**	**1**		**0**	**1**
	Real	**0**	70	11	**0**	27	0
		1	22	25	**1**	12	0
			Predicción			**Predicción**	
9			**0**	**1**		**0**	**1**
	Real	**0**	70	11	**0**	27	0
		1	15	32	**1**	12	0
			Predicción			**Predicción**	
10			**0**	**1**		**0**	**1**
	Real	**0**	69	12	**0**	27	0
		1	15	32	**1**	11	1
			Predicción			**Predicción**	
11			**0**	**1**		**0**	**1**
	Real	**0**	79	2	**0**	27	0
		1	9	38	**1**	12	0
			Predicción			**Predicción**	
12			**0**	**1**	**1**	**0**	**1**
	Real	**0**	47	0	**0**	25	2
		1	0	47	**1**	12	0
			Predicción			**Predicción**	
13			**0**	**1**		**0**	**1**
	Real	**0**	47	0	**0**	25	2
		1	0	47	**1**	11	1
			Predicción			**Predicción**	
14			**0**	**1**		**0**	**1**
	Real	**0**	47	0	**0**	26	1
		1	0	47	**1**	11	1
			Predicción			**Predicción**	

15			**0**	**1**		**0**	**1**
	Real	**0**	47	0	**0**	24	3
		1	0	47	**1**	9	3
			Predicción			**Predicción**	
16			**0**	**1**		**0**	**1**
	Real	**0**	47	0	**0**	25	2
		1	0	47	**1**	10	2
			Predicción			**Predicción**	
17			**0**	**1**		**0**	**1**
	Real	**0**	47	0	**0**	26	1
		1	0	47	**1**	10	2
			Predicción			**Predicción**	
18			**0**	**1**		**0**	**1**
	Real	**0**	81	0	**0**	26	1
		1	0	47	**1**	12	0
			Predicción			**Predicción**	
19			**0**	**1**		**0**	**1**
	Real	**0**	81	0	**0**	27	0
		1	0	47	**1**	12	0
			Predicción			**Predicción**	
20			**0**	**1**	**1**	**0**	**1**
	Real	**0**	81	0	**0**	26	1
		1	0	47	**1**	11	1
			Predicción			**Predicción**	
21			**0**	**1**		**0**	**1**
	Real	**0**	81	0	**0**	27	0
		1	0	47	**1**	12	0
			Predicción			**Predicción**	
22			**0**	**1**		**0**	**1**
	Real	**0**	81	0	**0**	27	0
		1	0	47	**1**	12	0
			Predicción			**Predicción**	
23			**0**	**1**		**0**	**1**
	Real	**0**	81	0	**0**	27	0
		1	0	47	**1**	12	0

Tabla 5-8. Matrices de confusión dataframe 2 distintos escenarios.

ID			Train			Test	
			Predicción			Predicción	
			0	1		0	1
0	Real	0	56	17	0	30	1
		1	24	49	1	13	1
			Predicción			Predicción	
			0	1		0	1
1	Real	0	53	20	0	31	0
		1	33	40	1	13	1
			Predicción			Predicción	
			0	1		0	1
2	Real	0	54	19	0	29	2
		1	21	52	1	11	3
			Predicción			Predicción	
			0	1		0	1
3	Real	0	59	14	0	29	2
		1	25	48	1	14	0
			Predicción			Predicción	
			0	1	1	0	1
4	Real	0	60	13	0	31	0
		1	17	56	1	13	1
			Predicción			Predicción	
			0	1		0	1
5	Real	0	62	11	0	28	3
		1	11	62	1	10	4
			Predicción			Predicción	
			0	1		0	1
6	Real	0	81	18	0	30	1
		1	32	41	1	13	1
			Predicción			Predicción	
			0	1		0	1
7	Real	0	83	16	0	31	0
		1	35	38	1	13	1
			Predicción			Predicción	
			0	1		0	1
8	Real	0	78	21	0	31	0
		1	32	41	1	14	0
			Predicción			Predicción	
			0	1		0	1
9	Real	0	82	17	0	29	2
		1	26	47	1	14	0
			Predicción			Predicción	

10			**0**	**1**		**0**	**1**
	Real	**0**	90	9	**0**	29	2
		1	21	52	**1**	14	0
			Predicción			**Predicción**	
11			**0**	**1**		**0**	**1**
	Real	**0**	89	10	**0**	31	0
		1	19	54	**1**	13	1
			Predicción			**Predicción**	
12			**0**	**1**	**1**	**0**	**1**
	Real	**0**	73	0	**0**	24	7
		1	0	73	**1**	10	4
			Predicción			**Predicción**	
13			**0**	**1**		**0**	**1**
	Real	**0**	73	0	**0**	28	3
		1	0	73	**1**	10	4
			Predicción			**Predicción**	
14			**0**	**1**		**0**	**1**
	Real	**0**	73	0	**0**	26	5
		1	0	73	**1**	10	4
			Predicción			**Predicción**	
15			**0**	**1**		**0**	**1**
	Real	**0**	73	0	**0**	26	5
		1	0	73	**1**	8	6
			Predicción			**Predicción**	
16			**0**	**1**		**0**	**1**
	Real	**0**	73	0	**0**	24	7
		1	0	73	**1**	11	3
			Predicción			**Predicción**	
17			**0**	**1**		**0**	**1**
	Real	**0**	73	0	**0**	25	6
		1	0	73	**1**	10	4
			Predicción			**Predicción**	
18			**0**	**1**		**0**	**1**
	Real	**0**	99	0	**0**	27	4
		1	0	73	**1**	10	4
			Predicción			**Predicción**	
19			**0**	**1**		**0**	**1**
	Real	**0**	99	0	**0**	29	2
		1	0	73	**1**	12	2
			Predicción			**Predicción**	
20		1	**0**	**1**	**1**	**0**	**1**
	Real	**0**	99	0	**0**	30	1
		1	0	73	**1**	12	2
			Predicción			**Predicción**	
21			**0**	**1**		**0**	**1**

	Real	**0**	99	0	**0**	29	2
		1	0	73	**1**	13	1
			Predicción			**Predicción**	
22			**0**	**1**		**0**	**1**
	Real	**0**	99	0	**0**	28	3
		1	0	73	**1**	10	4
			Predicción			**Predicción**	
23			**0**	**1**		**0**	**1**
	Real	**0**	99	0	**0**	27	4
		1	0	73	**1**	13	1

Tabla 5-9. Matrices de confusión dataframe 3 distintos escenarios.

					Train				
	Modelo	UnderSampling	Cod_Cat	Vbles	Ac	Prec	Spec	Rec	F-Score
0	LR	Si	Label	1	0,72	0,73	0,74	0,70	0,71
1	LR	Si	Label	2	0,71	0,73	0,77	0,64	0,69
2	LR	Si	Label	3	0,69	0,68	0,67	0,71	0,70
3	LR	Si	Dummy	1	0,75	0,75	0,75	0,75	0,75
4	LR	Si	Dummy	2	0,71	0,73	0,77	0,64	0,69
5	LR	Si	Dummy	3	0,87	0,86	0,86	0,88	0,87
6	LR	No	Label	1	0,70	0,67	0,80	0,56	0,61
7	LR	No	Label	2	0,72	0,71	0,84	0,55	0,62
8	LR	No	Label	3	0,70	0,67	0,80	0,56	0,61
9	LR	No	Dummy	1	0,75	0,73	0,83	0,64	0,69
10	LR	No	Dummy	2	0,69	0,66	0,79	0,55	0,60
11	LR	No	Dummy	3	0,85	0,87	0,92	0,75	0,81
12	RF	Si	Label	1	1	1	1	1	1
13	RF	Si	Label	2	1	1	1	1	1
14	RF	Si	Label	3	1	1	1	1	1
15	RF	Si	Dummy	1	1	1	1	1	1
16	RF	Si	Dummy	2	1	1	1	1	1
17	RF	Si	Dummy	3	1	1	1	1	1
18	RF	No	Label	1	1	1	1	1	1
19	RF	No	Label	2	1	1	1	1	1
20	RF	No	Label	3	1	1	1	1	1
21	RF	No	Dummy	1	1	1	1	1	1
22	RF	No	Dummy	2	1	1	1	1	1
23	RF	No	Dummy	3	1	1	1	1	1

Tabla 5-10. Resultados conjunto train para el dataframe 1.

					Test				
	Modelo	UnderSampling	Cod_Cat	Vbles	Ac	Prec	Spec	Rec	F-Score
0	LR	Si	Label	1	0,73	0,75	0,97	0,21	0,33
1	LR	Si	Label	2	0,71	1,00	1,00	0,07	0,13
2	LR	Si	Label	3	0,73	0,75	0,97	0,21	0,33
3	LR	Si	Dummy	1	0,73	0,75	0,97	0,21	0,33
4	LR	Si	Dummy	2	0,71	1,00	1,00	0,07	0,13
5	LR	Si	Dummy	3	0,71	1,00	1,00	0,07	0,13
6	LR	No	Label	1	0,71	0,67	0,97	0,14	0,24
7	LR	No	Label	2	0,71	1,00	1,00	0,07	0,13
8	LR	No	Label	3	0,69	0,00	1,00	0,00	0,00
9	LR	No	Dummy	1	0,71	1,00	1,00	0,07	0,13
10	LR	No	Dummy	2	0,71	1,00	1,00	0,07	0,13
11	LR	No	Dummy	3	0,69	0,00	1,00	0,00	0,00
12	RF	Si	Label	1	0,67	0,44	0,84	0,29	0,35
13	RF	Si	Label	2	0,71	0,57	0,90	0,29	0,38
14	RF	Si	Label	3	0,69	0,50	0,90	0,21	0,30
15	RF	Si	Dummy	1	0,71	0,57	0,90	0,29	0,38
16	RF	Si	Dummy	2	0,69	0,50	0,87	0,29	0,36
17	RF	Si	Dummy	3	0,60	0,30	0,77	0,21	0,25
18	RF	No	Label	1	0,69	0,50	0,94	0,14	0,22
19	RF	No	Label	2	0,67	0,40	0,90	0,14	0,21
20	RF	No	Label	3	0,67	0,43	0,87	0,21	0,29
21	RF	No	Dummy	1	0,67	0,33	0,94	0,07	0,12
22	RF	No	Dummy	2	0,64	0,25	0,90	0,07	0,11
23	RF	No	Dummy	3	0,67	0,40	0,90	0,14	0,21

Tabla 5-11. Resultados conjunto test para el dataframe 1.

					Train				
ID	Modelo	UnderSampling	Cod_Cat	Vbles	Ac	Prec	Spec	Rec	F-Score
0	LR	Si	Label	1	0,77	0,76	0,74	0,79	0,77
1	LR	Si	Label	2	0,71	0,74	0,77	0,66	0,70
2	LR	Si	Label	3	0,77	0,78	0,79	0,74	0,76
3	LR	Si	Dummy	1	0,80	0,79	0,79	0,81	0,80
4	LR	Si	Dummy	2	0,69	0,69	0,68	0,70	0,69
5	LR	Si	Dummy	3	0,93	0,90	0,89	0,96	0,93
6	LR	No	Label	1	0,80	0,71	0,83	0,74	0,73
7	LR	No	Label	2	0,79	0,75	0,88	0,64	0,69
8	LR	No	Label	3	0,74	0,69	0,86	0,53	0,60
9	LR	No	Dummy	1	0,80	0,74	0,86	0,68	0,71
10	LR	No	Dummy	2	0,79	0,73	0,85	0,68	0,70
11	LR	No	Dummy	3	0,91	0,95	0,98	0,81	0,87
12	RF	Si	Label	1	1	1	1	1	1
13	RF	Si	Label	2	1	1	1	1	1
14	RF	Si	Label	3	1	1	1	1	1
15	RF	Si	Dummy	1	1	1	1	1	1
16	RF	Si	Dummy	2	1	1	1	1	1
17	RF	Si	Dummy	3	1	1	1	1	1
18	RF	No	Label	1	1	1	1	1	1
19	RF	No	Label	2	1	1	1	1	1
20	RF	No	Label	3	1	1	1	1	1
21	RF	No	Dummy	1	1	1	1	1	1
22	RF	No	Dummy	2	1	1	1	1	1
23	RF	No	Dummy	3	1	1	1	1	1

Tabla 5-12. Resultados conjunto train para el dataframe 2.

					Test				
ID	Modelo	UnderSampling	Cod_Cat	Vbles	Ac	Prec	Spec	Rec	F-Score
0	LR	Si	Label	1	0,67	0,40	0,89	0,17	0,24
1	LR	Si	Label	2	0,64	0,25	0,89	0,08	0,13
2	LR	Si	Label	3	0,69	0,50	0,93	0,17	0,25
3	LR	Si	Dummy	1	0,69	0,00	1,00	0,00	0,00
4	LR	Si	Dummy	2	0,62	0,20	0,85	0,08	0,12
5	LR	Si	Dummy	3	0,72	0,67	0,96	0,17	0,27
6	LR	No	Label	1	0,72	1,00	1,00	0,08	0,15
7	LR	No	Label	2	0,69	0,50	0,96	0,08	0,14
8	LR	No	Label	3	0,69	0,00	1,00	0,00	0,00
9	LR	No	Dummy	1	0,69	0,00	1,00	0,00	0,00
10	LR	No	Dummy	2	0,72	1,00	1,00	0,08	0,15
11	LR	No	Dummy	3	0,69	0,00	1,00	0,00	0,00
12	RF	Si	Label	1	0,64	0,00	0,93	0,00	0,00
13	RF	Si	Label	2	0,67	0,33	0,93	0,08	0,13
14	RF	Si	Label	3	0,69	0,50	0,96	0,08	0,14
15	RF	Si	Dummy	1	0,69	0,50	0,89	0,25	0,33
16	RF	Si	Dummy	2	0,69	0,50	0,93	0,17	0,25
17	RF	Si	Dummy	3	0,72	0,67	0,96	0,17	0,27
18	RF	No	Label	1	0,67	0,00	0,96	0,00	0,00
19	RF	No	Label	2	0,69	0,00	1,00	0,00	0,00
20	RF	No	Label	3	0,69	0,50	0,96	0,08	0,14
21	RF	No	Dummy	1	0,69	0,00	1,00	0,00	0,00
22	RF	No	Dummy	2	0,69	0,00	1,00	0,00	0,00
23	RF	No	Dummy	3	0,69	0,00	1,00	0,00	0,00

Tabla 5-13. Resultados conjunto test para el dataframe 2.

					Train				
	Modelo	UnderSampling	Cod_Cat	Vbles	Ac	Prec	Spec	Rec	F-Score
0	LR	Si	Label	1	0,72	0,74	0,77	0,67	0,71
1	LR	Si	Label	2	0,64	0,67	0,73	0,55	0,60
2	LR	Si	Label	3	0,73	0,73	0,74	0,71	0,72
3	LR	Si	Dummy	1	0,73	0,77	0,81	0,66	0,71
4	LR	Si	Dummy	2	0,79	0,81	0,82	0,77	0,79
5	LR	Si	Dummy	3	0,85	0,85	0,85	0,85	0,85
6	LR	No	Label	1	0,71	0,69	0,82	0,56	0,62
7	LR	No	Label	2	0,70	0,70	0,84	0,52	0,60
8	LR	No	Label	3	0,69	0,66	0,79	0,56	0,61
9	LR	No	Dummy	1	0,75	0,73	0,83	0,64	0,69
10	LR	No	Dummy	2	0,83	0,85	0,91	0,71	0,78
11	LR	No	Dummy	3	0,83	0,84	0,90	0,74	0,79
12	RF	Si	Label	1	1	1	1	1	1
13	RF	Si	Label	2	1	1	1	1	1
14	RF	Si	Label	3	1	1	1	1	1
15	RF	Si	Dummy	1	1	1	1	1	1
16	RF	Si	Dummy	2	1	1	1	1	1
17	RF	Si	Dummy	3	1	1	1	1	1
18	RF	No	Label	1	1	1	1	1	1
19	RF	No	Label	2	1	1	1	1	1
20	RF	No	Label	3	1	1	1	1	1
21	RF	No	Dummy	1	1	1	1	1	1
22	RF	No	Dummy	2	1	1	1	1	1
23	RF	No	Dummy	3	1	1	1	1	1

Tabla 5-14. Resultados conjunto train para el dataframe 3.

					Test				
	Modelo	UnderSampling	Cod_Cat	Vbles	Ac	Prec	Spec	Rec	F-Score
0	LR	Si	Label	1	0,69	0,50	0,97	0,07	0,13
1	LR	Si	Label	2	0,71	1,00	1,00	0,07	0,13
2	LR	Si	Label	3	0,71	0,60	0,94	0,21	0,32
3	LR	Si	Dummy	1	0,64	0,00	0,94	0,00	0,00
4	LR	Si	Dummy	2	0,71	1,00	1,00	0,07	0,13
5	LR	Si	Dummy	3	0,71	0,57	0,90	0,29	0,38
6	LR	No	Label	1	0,69	0,50	0,97	0,07	0,13
7	LR	No	Label	2	0,71	1,00	1,00	0,07	0,13
8	LR	No	Label	3	0,69	0,00	1,00	0,00	0,00
9	LR	No	Dummy	1	0,64	0,00	0,94	0,00	0,00
10	LR	No	Dummy	2	0,64	0,00	0,94	0,00	0,00
11	LR	No	Dummy	3	0,71	1,00	1,00	0,07	0,13
12	RF	Si	Label	1	0,62	0,36	0,77	0,29	0,32
13	RF	Si	Label	2	0,71	0,57	0,90	0,29	0,38
14	RF	Si	Label	3	0,67	0,44	0,84	0,29	0,35
15	RF	Si	Dummy	1	0,71	0,55	0,84	0,43	0,48
16	RF	Si	Dummy	2	0,60	0,30	0,77	0,21	0,25
17	RF	Si	Dummy	3	0,64	0,40	0,81	0,29	0,33
18	RF	No	Label	1	0,69	0,50	0,87	0,29	0,36
19	RF	No	Label	2	0,69	0,50	0,94	0,14	0,22
20	RF	No	Label	3	0,71	0,67	0,97	0,14	0,24
21	RF	No	Dummy	1	0,67	0,33	0,94	0,07	0,12
22	RF	No	Dummy	2	0,71	0,57	0,90	0,29	0,38
23	RF	No	Dummy	3	0,62	0,20	0,87	0,07	0,11

Tabla 5-15. Resultados conjunto test para el dataframe 3.

A simple vista, comparando los resultados iniciales con los nuevos resultados, se puede apreciar una mejora de estos aplicando algunas de las modificaciones. Para obtener una perspectiva mejor de todos los nuevos resultados, se incluyen a continuación gráficas que plasman los resultados de cada métrica para los conjuntos de prueba y entrenamiento.

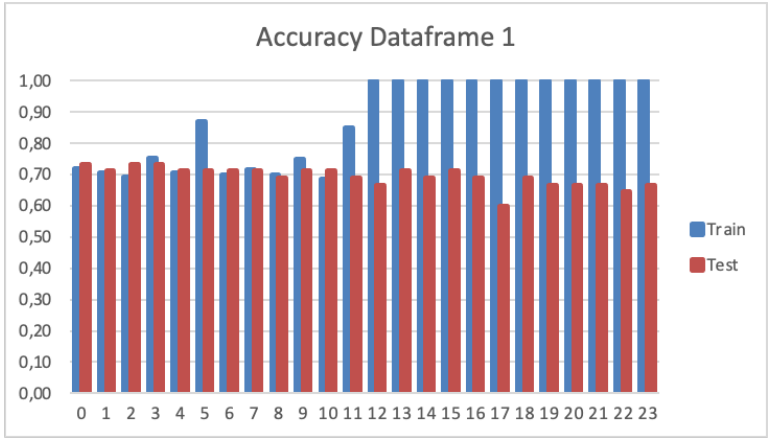


Figura 5-2. Representación accuracy en dataframe 1.

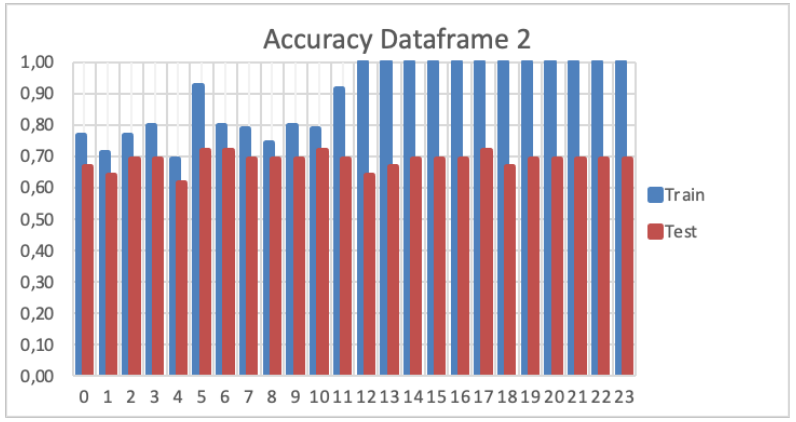


Figura 5-3. Representación accuracy en dataframe 2.

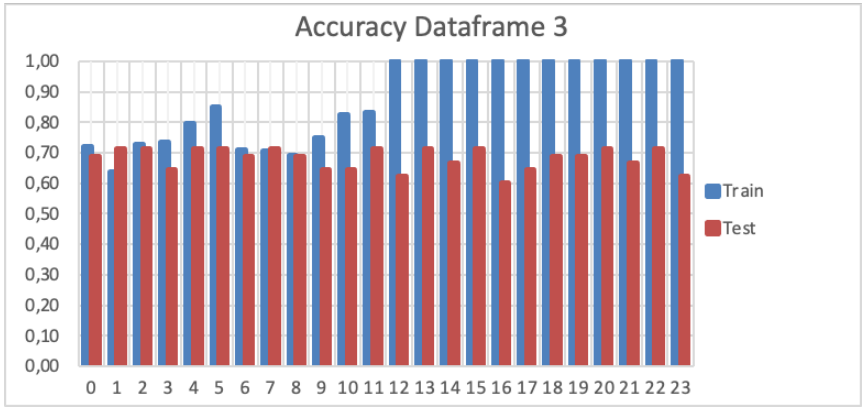


Figura 5-4. Representación accuracy en dataframe 3.

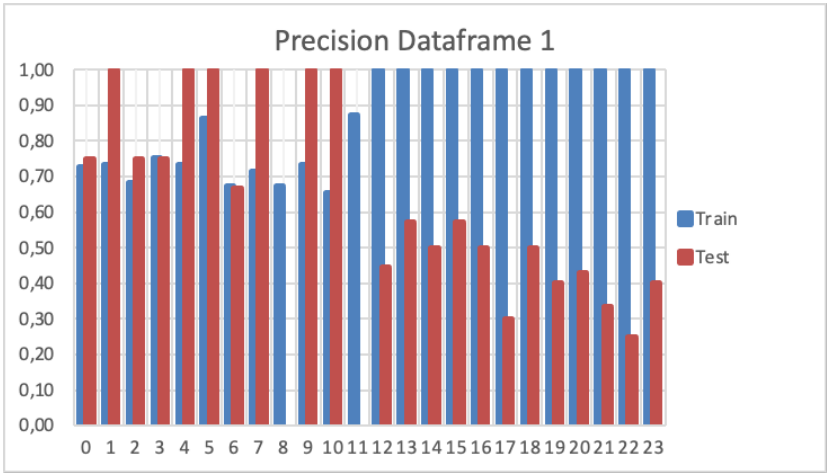


Figura 5-5. Representación precision en dataframe 1.

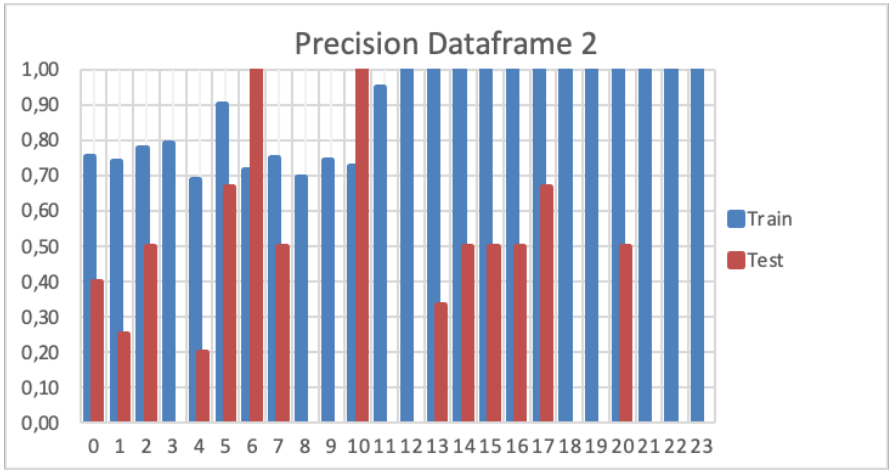


Figura 5-6. Representación precision en dataframe 2.

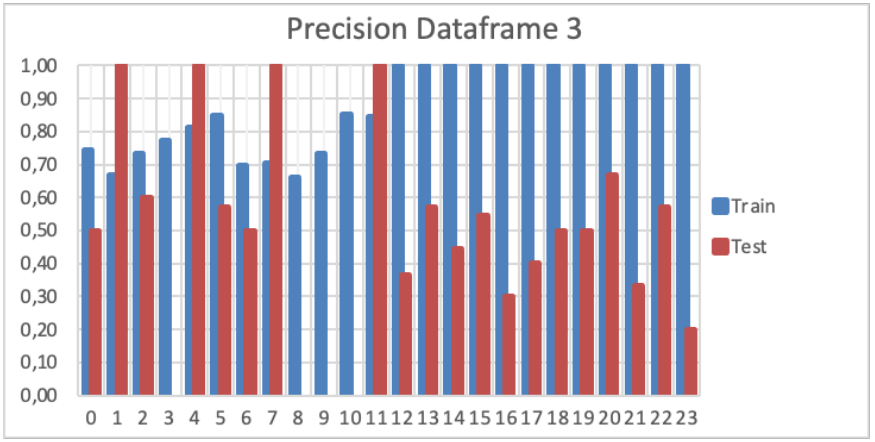


Figura 5-7. Representación precision en dataframe 3.

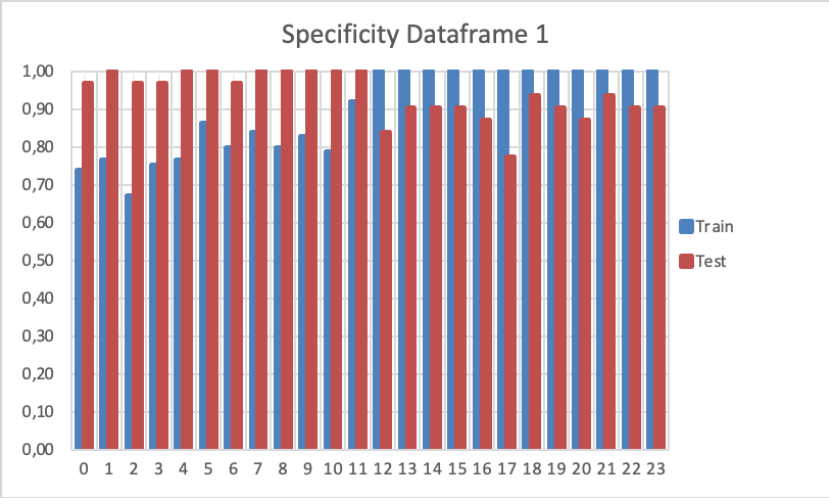


Figura 5-8. Representación specificity en dataframe 1.

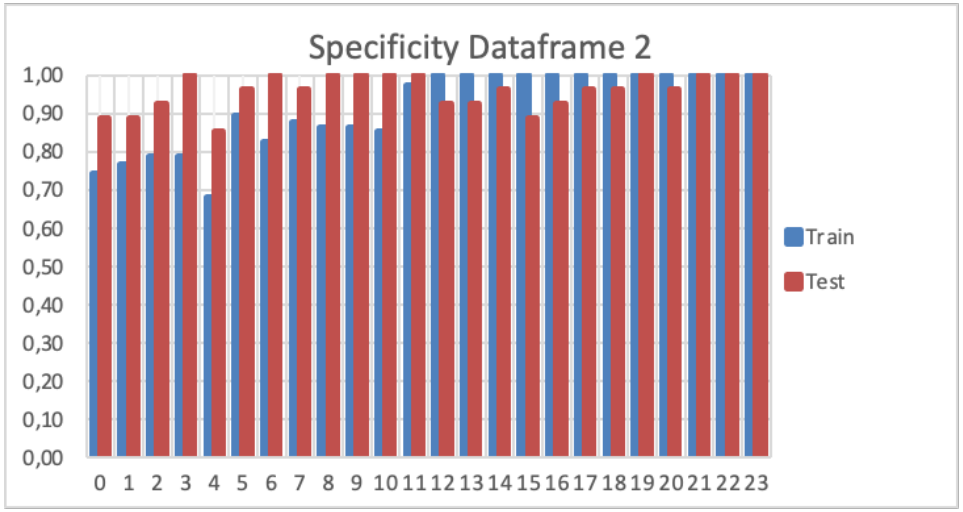


Figura 5-9. Representación specificity en dataframe 2.

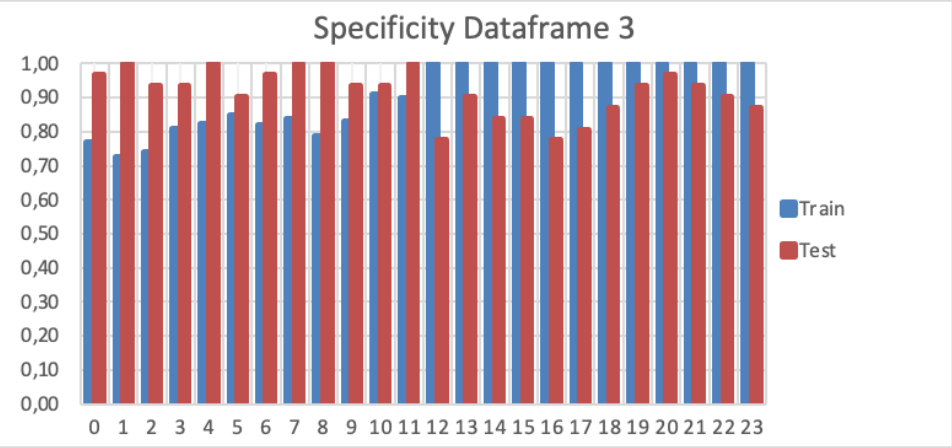


Figura 5-10. Representación specificity en dataframe 3.

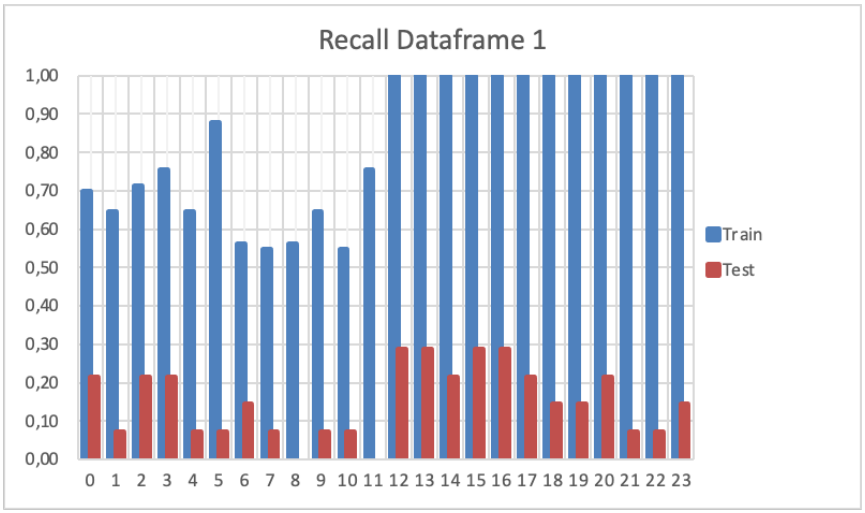


Figura 5-11. Representación recall en dataframe 1.

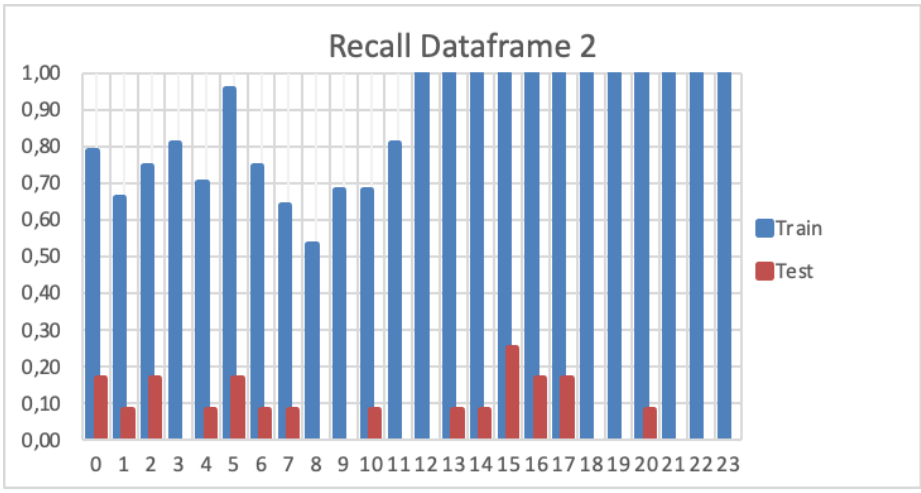


Figura 5-12. Representación recall en dataframe 2.

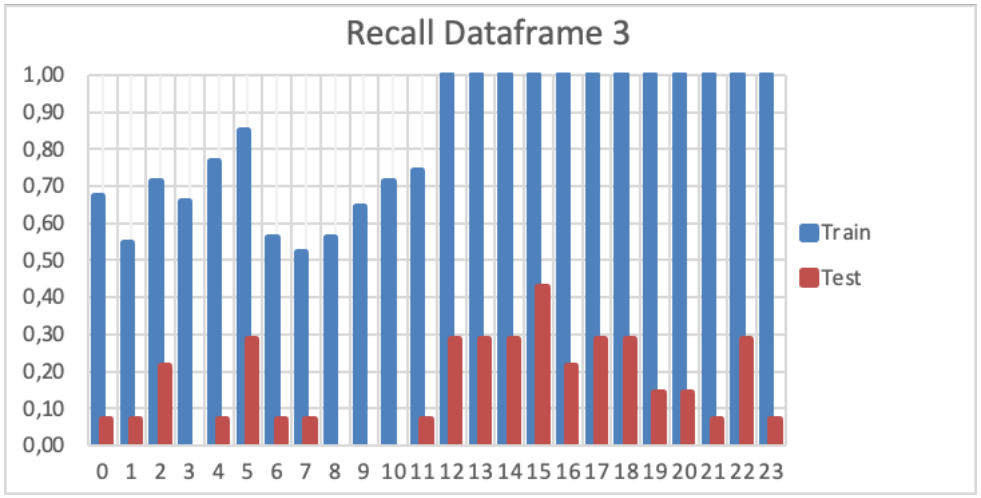


Figura 5-13. Representación recall en dataframe 3.

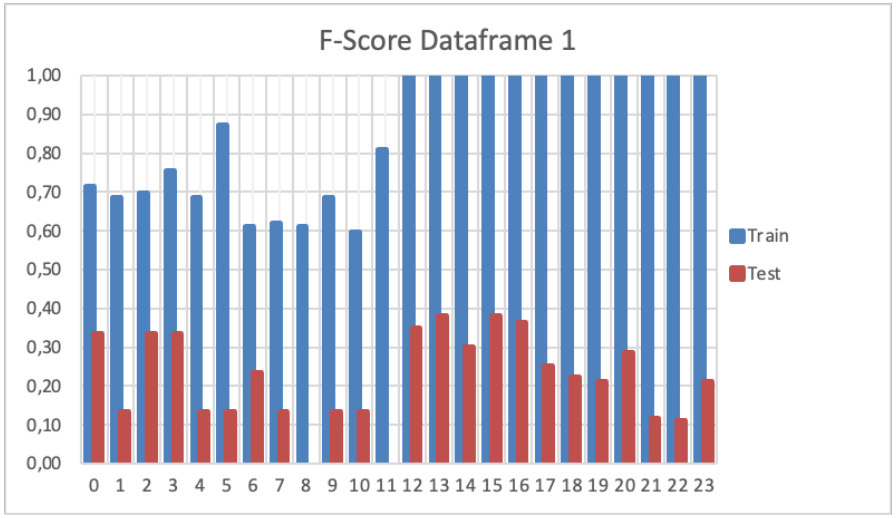


Figura 5-7. Representación f-score en dataframe 1.

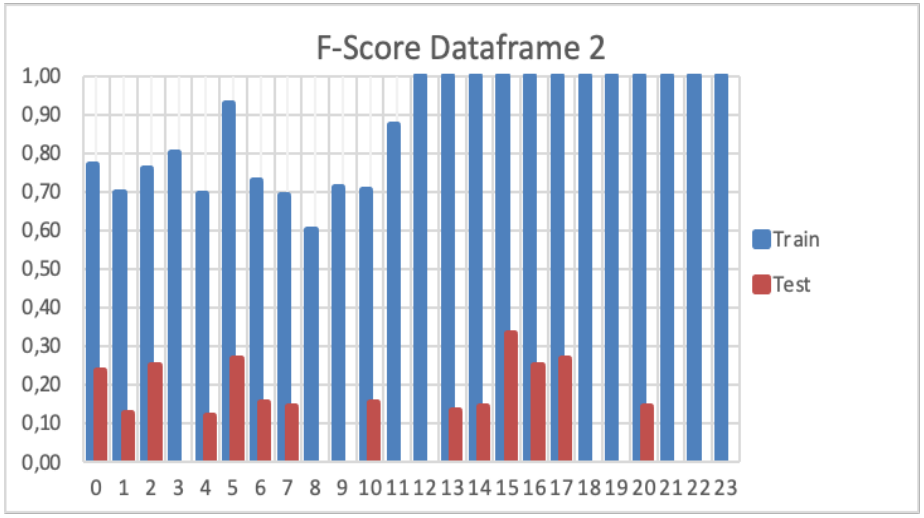


Figura 5-8. Representación f-score en dataframe 2.

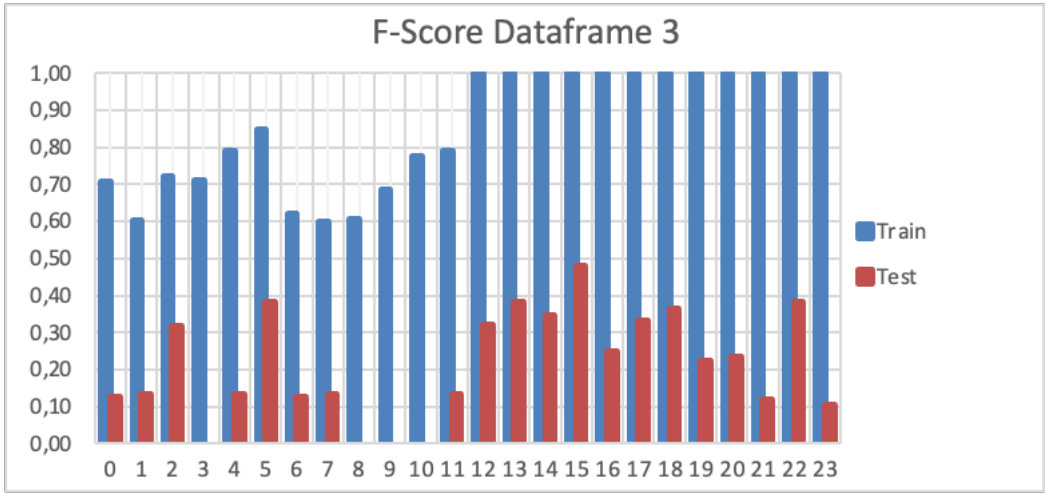


Figura 5-6. Representación f-score en dataframe 3.

Se observan resultados, aunque distintos, similares en cuanto a rango de valores obtenidos para los diferentes dataframes. Por ello, y para poder analizar los resultados con mayor precisión, se plasman en las siguientes tablas resumen, las métricas de evaluación de los conjuntos de prueba de cada dataframe.

	Test			
	Mínimo	**Máximo**	**Media**	**Moda**
Accuracy	0,60	0,73	0,69	0,71
Precision	0,00	1,00	0,59	1,00
Specificity	0,77	1,00	0,94	1,00
Recall	0,00	0,29	0,15	0,07
F-Score	0,00	0,38	0,22	0,13

Tabla 5-16. Resumen métricas de evaluación test, dataframe 1.

	Test			
	Mínimo	**Máximo**	**Media**	**Moda**
Accuracy	0,62	0,72	0,69	0,69
Precision	0,00	1,00	0,31	0,00
Specificity	0,85	1,00	0,96	1,00
Recall	0,00	0,25	0,07	0,00
F-Score	0,00	0,33	0,11	0,00

Tabla 5-17. Resumen métricas de evaluación test, dataframe 2.

	Test			
	Mínimo	**Máximo**	**Media**	**Moda**
Accuracy	0,60	0,71	0,68	0,71
Precision	0,00	1,00	0,48	0,50
Specificity	0,77	1,00	0,92	0,94
Recall	0,00	0,43	0,15	0,07
F-Score	0,00	0,48	0,21	0,13

Tabla 5-18. Resumen métricas de evaluación test, dataframe 3.

Analizando los resultados de las tablas resumen anteriores, se aprecia la similitud de resultados percibida en las gráficas. Entre los dataframes, destacan los resultados de accuracy, precisión y specificity obtenidos por el primero, teniendo mejor valor medio y máximo que los otros dos; pero en cuanto al recall y f-score, destacan los valores máximos del tercero, a pesar de manejar medias similares, es en este dataframe donde se consiguen valores sensiblemente mejores que en los otros dos para estas métricas. A continuación, se muestran las tablas resumen de los conjuntos de entrenamiento. Cabe destacar de estas tablas que han sido omitidos los resultados de clasificación aplicando Random Forest, puesto que como ya se ha comentado, estos resultados son perfectos.

	Train			
	Mínimo	**Máximo**	**Media**	**Moda**
Accuracy	0,69	0,87	0,74	0,71
Precision	0,66	0,87	0,74	0,73
Specificity	0,67	0,92	0,79	0,77
Recall	0,55	0,88	0,66	0,64
F-Score	0,60	0,87	0,70	0,69

Tabla 5-19. Resumen métricas de evaluación train, dataframe 1.

	Train			
	Mínimo	**Máximo**	**Media**	**Moda**
Accuracy	0,69	0,93	0,79	0,77
Precision	0,69	0,95	0,77	-
Specificity	0,68	0,98	0,83	0,79
Recall	0,53	0,96	0,73	0,74
F-Score	0,60	0,93	0,75	-

Tabla 5-20. Resumen métricas de evaluación train, dataframe 2.

	Train			
	Mínimo	**Máximo**	**Media**	**Moda**
Accuracy	0,64	0,85	0,75	-
Precision	0,66	0,85	0,76	-
Specificity	0,73	0,91	0,82	-
Recall	0,52	0,85	0,66	0,71
F-Score	0,60	0,85	0,70	-

Tabla 5-21. Resumen métricas de evaluación train, dataframe 3.

Destacan de los resultados de las clasificaciones del conjunto de entrenamiento, como de superior son para todas las métricas los resultados del dataframe 2, principalmente cuando se ha visto que para las clasificaciones del conjunto de prueba era el menos destacable.

Dado que los resultados en el conjunto de prueba son los mas representativos, y que además estos están desequilibrados, se descarta el accuracy como metrica para detectar el mejor escenario. Concretamente se decidide que el recall es la metrica mas importante, no obstante, también lo es el specifity ya que una tasa de verdaderos positivos (recall) elevada junto a una tasa de verdaderos negativos (specificity) baja no puede considerarse buena. Lo ideal es obtenener valores equilibrados para estos parámetros, por ello se representan las 24 simulaciones por dataframe, enfrentando los resultados de estas métricas.

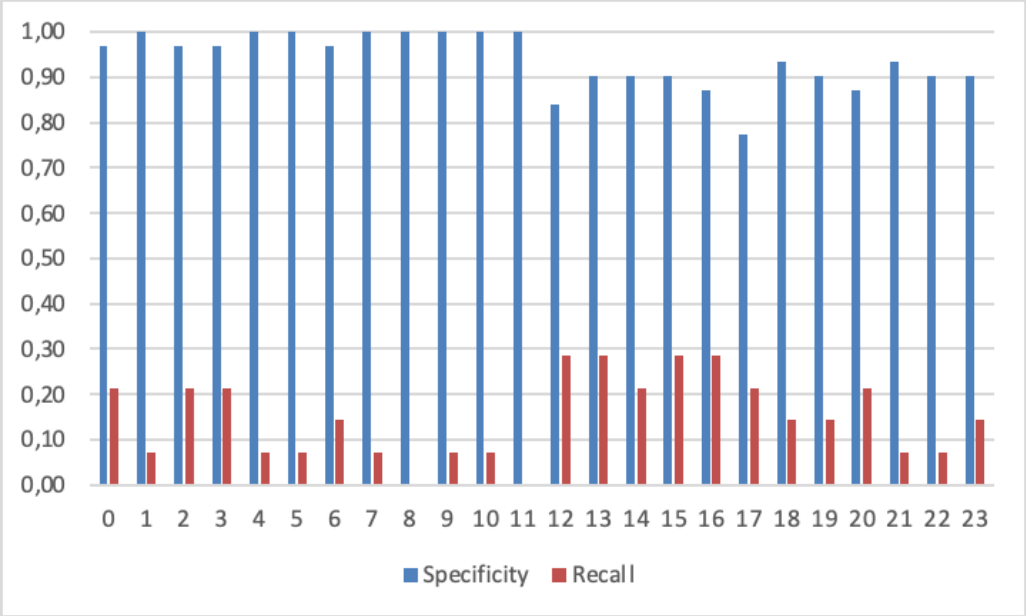


Figura 5-9. Specificity y Recall para todos los escenarios, dataframe 1.

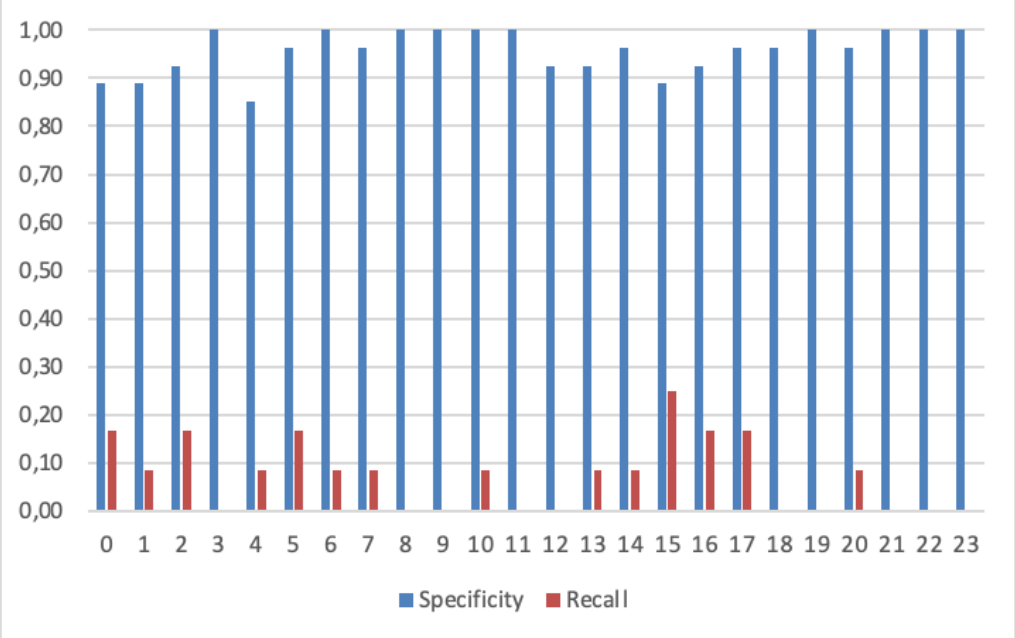


Figura 5-10. Specificity y Recall para todos los escenarios, dataframe 2.

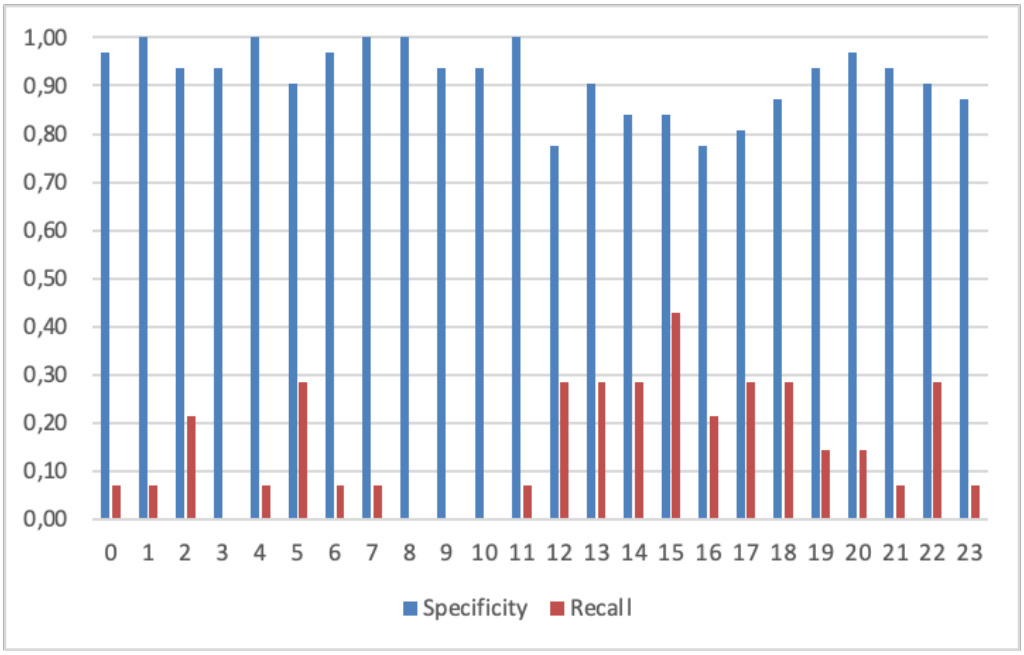


Figura 5-11. Specificity y Recall para todos los escenarios, dataframe 3.

Se puede observar, como ya era conocido, la disparidad de resultados entre el specificity y el recall, denotando la mala clasificación de positivos que realizan todas las simulaciones. Destacan los resultados del tercer dataframe, en el que se almacenan los datos de dos temporadas consecutivas según las variaciones de sus parámetros, estos resultados mejoran sensiblemente a los de los otros dos dataframes.

5.2.2 Análisis de la mejor solución

Finalmente, para cerrar este análisis de resultados, se procede a analizar con mayor detalle la mejor solución obtenida. Siguiendo el criterio descrito con anterioridad, se estableció como mejor solución aquella que poseyera un mejor equilibrio entre el specificity (tasa de negativos bien predichos) y el recall (tasa de positivos bien predichos); concretamente, para determinar de manera rigurosa la mejor solución, se realizó la media aritmética entre estas dos métricas, estableciendo como mejor solución aquella cuya media fuera superior a las demás. El escenario que resultó con mayor media entre specificity y recall y, por tanto, la mejor solución, fue la solución 15 del dataframe 3, resaltada en amarillo en la Tabla 5-15.

	Modelo	UnderSampling	Cod_Cat	Vbles	TP	TN	FP	FN	Ac	Prec	Spec	Rec	F-Score
15	RF	Si	Dummy	1	6	26	5	8	0,71	0,55	0,84	0,43	0,48

Tabla 5-22. Resultados mejor solución.

El escenario en el que se han conseguido los mejores resultados es en el que se aplicó *Random Forest* partiendo del dataframe 3 (donde en cada fila se almacena la información de dos temporadas consecutivas de cada delantero, a través de la variación de sus parámetros), realizando *UnderSampling* al conjunto de entrenamiento, codificando las variables categóricas de tipo *Dummy*, y haciendo uso del conjunto total de las variables.

A continuación, se muestran en la siguiente tabla todas las clasificaciones del conjunto de prueba realizadas por la mejor solución. Cada fila de la tabla incluye el nombre del futbolista y su edad previa a la previsión, se incluyen además los porcentajes de clasificación devueltos por el modelo junto a la predicción y el valor real. Cabe destacar también que se ha utilizado un código de colores, donde las filas con clasificaciones correctas quedan señaladas en verde, mientras que las incorrectas en rojo.

		Probabilidad			
Nombre	**Edad**	**0**	**1**	**Clasificación**	**Real**
Iago Aspas	32	0.68	0.32	0	0
Memphis Depay	26	0.57	0.43	0	0
Youssef En-Nesyri	23	0.4	0.6	1	1
Alexander Isak	20	0.65	0.35	0	1
Gerard Moreno	28	0.72	0.28	0	1
Anthony Martial	24	0.29	0.71	1	0
Luka Jovic	22	0.71	0.29	0	0
Aubameyang	30	0.68	0.32	0	0
Paco Alcacer	26	0.46	0.54	1	0
Matheus Cunha	21	0.41	0.59	1	1
Karim Benzema	32	0.9	0.1	0	0
Inaki Williams	25	0.53	0.47	0	0
Maxi Gomez	23	0.51	0.49	0	0
Willian Jose	28	0.65	0.35	0	0
Santi Mina	24	0.52	0.48	0	1
Boulaye Dia	23	0.38	0.62	1	1
Alexander Sorloth	24	0.4	0.6	1	0
Luis Suarez	33	0.56	0.44	0	0
Borja Iglesias	27	0.57	0.43	0	0
Borja Mayoral	23	0.38	0.62	1	1
Enes Unal	23	0.62	0.38	0	0
Chimy Avila	26	0.48	0.52	1	0
Munir El-Haddadi	24	0.45	0.55	1	1
Martin Braithwaite	28	0.82	0.18	0	0
Mariano Diaz	26	0.5	0.5	0	0
Joselu Mato	30	0.76	0.24	0	1
Roger Marti	29	0.77	0.23	0	1
Vedat Muriqi	26	0.52	0.48	0	0
Jose Luis Morales	32	0.66	0.34	0	0
Loren Moron	26	0.34	0.66	1	0
Luuk De Jong	29	0.57	0.43	0	1

Sandro Ramirez	24	0.45	0.55	1	1
Sergi Guardiola	29	0.57	0.43	0	0
Jaime Mata	31	0.77	0.23	0	0
Kike Garcia	30	0.83	0.17	0	1
Raul Garcia	33	0.76	0.24	0	0
Lucas Perez	31	0.7	0.3	0	0
Angel Rodroguez	33	0.79	0.21	0	0
Radamel Falcao	34	0.73	0.27	0	0
Guido Carrillo	29	0.77	0.23	0	0
Carlos Bacca	33	0.79	0.21	0	0
Ruben Sobrino	28	0.6	0.4	0	0
Roberto Soldado	35	0.78	0.22	0	1
Jorge Molina	38	0.68	0.32	0	0
John Guidetti	28	0.8	0.2	0	0

Tabla 5-23. Resumen de las clasificaciones de la mejor solución.

Con los datos de la tabla anterior, se obtiene que cuando el modelo se equivoca en sus clasificaciones lo hace con una media de error del 16,38%. Se conoce, además, que la edad media de los futbolistas cuya clasificación fue incorrecta fue de 27 años. En cuanto a los falsos positivos, se obtiene una media de 10,6% de error en la clasificación de futbolistas con un promedio de edad de 25,2 años; por otro lado, los falsos negativos obtuvieron una media de 20% de error, en delanteros con un promedio de edad de 28,13 años. Estos datos nos muestran la tendencia que tiene el modelo hacia las clasificaciones negativas que, si bien, se ha mejorado con respecto a los resultados iniciales, sigue sin ser todo lo bueno que gustaría.

Plasmando los resultados en un contexto real, si el director deportivo de cualquier equipo recibiera este informe, siendo conocedor de la precisión, exhaustividad, exactitud, especificidad y f-score del modelo, podría apoyar su toma de decisiónes en estos datos. Por ejemplo, si cualquier equipo deseara fichar a Matheus Cunha, tendría la información de que esta temporada va a tener un precio menor al que tendría la próxima, haciendo así que se refuerce su interés por el jugador, sabiendo además que podría pujar algo más alto por el futbolista, ya que este se revalorizaría. Sin embargo, si fuese el Atlético de Madrid, poseedor de los derechos de Matheus Cunha, quien contase con esta información, debería negarse a la venta o pedir un precio bastante superior a su valor de mercado actual, ya que sabe que aguantando al futbolista una temporada más le podría sacar mayor rédito económico.

5.3 Análisis de influencia de variables a través del modelo de Regresión Logística

Complementando el análisis de los resultados, gracias a una herramienta disponible en la librería *Scikit-learn* para las clasificaciones usando Regresión Logística, se procede a analizar la influencia de cada variable gracias al valor absoluto de los coeficientes de las covariables. Como lo interesante de este análisis es contar con todas las variables, se decidió almacenar la importancia de las variables para los modelos con *Vbles* = 1 que, aplicasen Regresión Logística. Además, se utilizó los modelos con codificación de variables categóricas de tipo label, de modo que se pudiese analizar la importancia de las variables categóricas. Esto se traduce en dos resultados por dataframe (IDs 0 y 6), escenarios donde el modelo aplica Regresión Logística (LR), la codificación de las variables categóricas es de tipo *label*, y el conjunto de variables es el total (*Vbles* = 1); únicamente se diferencian entre sí en que en uno se aplica *underampling* y en otro no. Es decir, seis resultados en total que serán analizados a continuación.

Escenario	AvgP1	Edad1	Edad2	Temporada1	Temporada2
Undersampling	0,100	0,097	0,097	0,073	0,073
No Undersampling	0,128	0,127	0,118	0,118	0,117

Tabla 5-22. Coeficientes más altos dataframe 1.

Escenario	AvgP2	Temporada1	Temporada2	Temporada3	Edad1	Edad2	Edad3
Undersampling	0,142	0,134	0,133	0,132	0,115	0,114	0,113
No Undersampling	0,143	0,113	0,112	0,112	0,099	0,098	0,097

Tabla 5-23. Coeficientes más altos dataframe 2.

Escenario	Temporada1	Competicion1	Temporada2	AvgP
Undersampling	0,212	0,192	0,163	0,150
No Undersampling	0,195	0,174	0,141	0,079

Tabla 5-24. Coeficientes más altos dataframe 3 (incrementos).

Como se puede observar en las tablas anteriores, las variables que más influyen en los modelos analizados son:

- **Edad:** Para los dataframes 1 y 2, tienen una alta importancia todas las edades presentes en cada fila del dataframe (dos en el primero y tres en el segundo). Esto encuentra un sentido lógico ya que los futbolistas suelen alcanzar su valor de mercado máximo entre los 25 y 29 años. A partir de una cierta edad, sin importar apenas su rendimiento deportivo, ven como se disminuye su valor ya que hacer una gran venta con un futbolista veterano es bastante complicado.
- **Temporada:** Todas las temporadas presentes en cada muestra del dataframe cuentan con una importancia mayor que el resto de las variables. Se puede interpretar por ello que existen temporadas en las que los valores de los jugadores, por muy diversas razones como puede ser una crisis económica o la inflación del mismo mercado, sufren subidas o bajadas para ajustarse a los valores reales de su mercado.
- **AvgP:** Con una alta importancia en los tres dataframes. El modelo ha detectado que los futbolistas que promedian más pases por encuentro suelen aumentar su valor de mercado. Futbolísticamente se puede traducir como que los delanteros más cotizados son aquellos que participan más del juego de sus equipos, cuantos más pases dé el futbolista, más involucrado estará en el juego y más importante será para su equipo.
- **Competición:** Con una importancia notable únicamente en el tercer dataframe. La primera competición de las dos presentes en cada muestra adquiere un peso relevante en la clasificación. Esto se puede entender por el aspecto mediático que tienen las competiciones, si un futbolista promedia buenas cifras, pero viene de una liga menor, su valor de mercado será inferior al que venga de alguna de las grandes ligas, por repercusión mediática y nivel deportivo de estas.

6 CONCLUSIONES

Con el claro objetivo de desarrollar una herramienta que fuera capaz de predecir el aumento o no, del valor de mercado de futbolistas profesionales, de una temporada a otra, se comenzó en el desarrollo de la base de datos. Para montar la base de datos, hubo que establecer un criterio de que jugadores iban a formar parte de la misma, decidiendo que la compondrían todos los delanteros centro inscritos en LaLiga Santander a día 1 de febrero de 2022. Tras tener la lista cerrada de jugadores, se procedió a guardar todas las estadísticas de estos desde la web WhoScored, así como su valor de mercado obtenido a través de Transfermarkt. Una vez almacenados los datos, se procedió a eliminar todas aquellas temporadas que no tuvieran registro de la anterior ni de la siguiente, además de solucionar el problema de los traspasos en el mercado invernal, unificando las estadísticas de los distintos equipos para la misma temporada. Seguidamente se formaron tres dataframes diferentes, según la manera de concatenar las temporadas, en el primero se guardaron datos de una temporada, seguidos de datos de la siguiente para, finalmente, indicar si aumentó o no su valor de mercado en una tercera; el segundo seguía la misma filosofía que el primero, pero aumentando los datos con una temporada más, de modo que permitiera analizar algo más la trayectoria del futbolista; el tercero contenía la misma información que el primero, pero en vez de plasmar los datos de las temporadas individualmente, mostraba la variación de los parámetros de una temporada a otra. Una vez preparados y filtrados los datos, se procedió a la división de los mismos en los conjuntos de prueba (test) y entrenamiento (train). Generalmente, este procedimiento suele efectuarse de manera aleatoria, seleccionando un porcentaje menor del total que será destinado al conjunto de prueba, mientras que el resto formará parte del entrenamiento del modelo. Al querer mantener el realismo en el desarrollo del trabajo, la división entre *train* y *test* se hizo siguiendo un criterio temporal, dejando para entrenamiento todas las temporadas excepto la más reciente, que se usaría para evaluar al modelo.

Una vez preparados todos los datos, se procedió a la implementación de los modelos: Regresión Logística y Random Forest. Al obtener los resultados, se detectó un comportamiento deficiente de los modelos en cuanto a las clasificaciones positivas de las muestras, es decir, los aumentos del valor de mercado (*AValor* = 1). Dicha deficiencia podía venir de un desbalanceo de clases en los conjuntos de entrenamiento, que en efecto existía, habiendo una mayor presencia de clasificaciones negativas que positivas. Con el objetivo de arreglar este problema de desbalanceo de clases, se propuso realizar *undersampling* a la muestra de entrenamiento, pero, como se buscaba obtener los mejores resultados posibles, se propusieron otras modificaciones que pudieran dar lugar a un mejor funcionamiento. Estas otras modificaciones consistían en cambiar la codificación de las variables categóricas y modificar las variables presentes en los dataframes, para lo que se hicieron tres conjuntos de varibles diferentes siguiendo un criterio de correlación con la variable de salida. Todas estas modificaciones

dieron lugar a 24 escenarios diferentes por dataframe, o lo que es lo mismo, en vez de obtener 6 resultados se obtuvieron 72.

Tras obtener los 72 resultados, se pudo apreciar como en bastantes de estos mejoraban a los primeros. A pesar de esta mejora, los resultados finales tampoco fueron excepcionales, llegando a clasificar correctamente 6 positivos en el mejor de los casos. Se estableció la mejor solución siguiendo un criterio que buscara equilibrio entre los resultados de las métricas *specificity* y *recall*, para las clasificaciones del conjunto de prueba. La mejor solución, a pesar de destacar notablemente sobre el resto, únicamente obtuvo un *recall* del 43%, cifra que se hubiera deseado mejorar.

En cuanto a las futuras líneas de investigación derivadas de este proyecto, se abren muchas puertas. Una de ellas sería tener acceso a una base de datos de mayor tamaño. Al haber construido a mano la base de datos, el tamaño de esta no pudo ser muy extenso, pudiendo haber ocasionado esto el no tan buen resultado de las clasificaciones positivas. Es por esto por lo que se alberga la esperanza de que teniendo acceso a una gran base de datos de alguna de las muchas empresas que se dedican a la analítica de datos en el fútbol profesional, se mejorarían considerablemente los resultados. Otra futura línea de investigación sería la de realizar un análisis exhaustivo de las variables, para reunir así el conunto de variables que optimice los resultados.

Las posibilidades de desarrollo de herramientas similares en el ámbito futbolístico son prácticamente infinitas; a pesar de ello, se comentan a continuación proyectos que, derivados de este trabajo, podrían desarrollarse en el futuro:

- Una herramienta que prediga la variación del valor de mercado y/o del rendimiento de un futbolista al pasar de un equipo a otro. Se seleccionaría un jugador en activo, al cual se le asigna un destino para la próxima temporada y el modelo debe predecir si mejoraría o no su rendimiento y/o valor con esa transferencia.
- Una herramienta como la aquí desarrollada pero que en vez de predecir el aumento o no del valor de mercado, prediga el valor exacto que alcanzaría.
- Una herramienta que, nutrida con el histórico de minutos de juego, picos de esfuerzo, minutos perdiendo, minutos ganando, y otros patrones similares, sea capaz de predecir cuando un jugador está al borde de sufrir una lesión muscular.

REFERENCIAS

[1] LaLiga finaliza la 2018/2019 batiendo un nuevo record de asistencia en los estadios por sexta temporada consecutiva | Colores que laten n.d. https://newsletter.laliga.es/colores-que-laten/laliga-finaliza-la-20182019-batiendo-un-nuevo-record-de-asistencia-en-los-estadios-por-sexta-temporada-consecutiva (accessed May 9, 2022).

[2] LaLiga sí interesa a los jóvenes: los menores de 24 años crecen un 10% desde 2017 n.d. https://www.elconfidencial.com/deportes/futbol/liga/2021-07-01/laliga-tebas-jovenes-audiencia_3160368/ (accessed May 9, 2022).

[3] PwC. Impacto socioeconómico, fiscal y social del fútbol profesional en España 2016.

[4] LaLiga Tech presenta su paquete de soluciones tecnológicas, diseñadas para la era digital del deporte y entretenimiento | Colores que laten n.d. https://newsletter.laliga.es/colores-que-laten/laliga-tech-presenta-su-paquete-de-soluciones-tecnologicas-disenadas-para-la-era-digital-del-deporte-y-entretenimiento (accessed May 12, 2022).

[5] LaLiga recibe certificaciones I+D+i que refrendan su labor en materia tecnológica y de innovación | Fútbol Global n.d. https://newsletter.laliga.es/futbol-global/laliga-recibe-certificaciones-i-d-i-que-refrendan-su-labor-en-materia-tecnologica-y-de-innovacion (accessed May 12, 2022).

[6] Big Data: La 'data science' al servicio de la competición, el contenido y el fan: así se adaptará LaLiga a ti n.d. https://www.elconfidencial.com/tecnologia/2022-02-25/laliga-data-science-datos-fan-bra_3380604/ (accessed May 16, 2022).

[7] Big Data: La puesta a punto de los clubes de LaLiga: de la pizarra y el excel a la analítica de datos avanzada n.d. https://www.elconfidencial.com/empresas/2022-04-26/analitica-datos-rendimiento-negocio-laliga-bra_3411834/ (accessed May 12, 2022).

[8] MONCHI 13 - Masterclass: el factor Suerte (I) - YouTube n.d. https://www.youtube.com/watch?v=y9pmDtZY3PY&t=200s (accessed May 17, 2022).

[9] ALBERT VALENTIN. DIRECCIÓN DEPORTIVA EN UN CLUB DE FÚTBOL PROFESIONAL. FDL; 2017.

[10] Albert Valentín – Fútbol de Libro n.d. https://futboldelibro.com/autor/albert-valentin/ (accessed May 18, 2022).

[11] LaLiga. Informe Económico Financiero LaLiga 2020 n.d.

[12] "El fútbol español todavía no está ni en la rampa de salida con el 'Big Data'"""" n.d.

https://navarracapital.es/el-futbol-espanol-todavia-no-esta-ni-en-la-rampa-de-salida-con-el-big-data/ (accessed May 24, 2022).

[13] Jesús M. Botello Hermosa. ¿Qué es scouting? ABFútbol 2012:45–57.

[14] MONCHI 13 Masterclass: Big Data (XI) - YouTube n.d. https://www.youtube.com/watch?v=vOEBCSWWnvY (accessed May 22, 2022).

[15] Kevin De Bruyne uses data analysts to broker £83m Man City contract without agent - Mirror Online n.d. https://www.mirror.co.uk/sport/football/news/kevin-de-bruyne-uses-data-23870686 (accessed May 24, 2022).

[16] Inside De Bruyne's data report: Sancho comparison and impact of playmaker's possible City exit crucial to new deal - The Athletic n.d. https://theathletic.com/2509349/2021/04/12/inside-de-bruynes-data-report-sancho-comparison-and-impact-of-playmakers-possible-city-exit-crucial-to-new-deal/ (accessed May 24, 2022).

[17] Nuestro rol en su proceso de toma de decisiones - SciSports n.d. https://www.scisports.com/es/trayectoria-de-exito/nuestro-rol-en-su-proceso-toma-de-decisiones/ (accessed May 24, 2022).

[18] Machine learning, explained | MIT Sloan n.d. https://mitsloan.mit.edu/ideas-made-to-matter/machine-learning-explained (accessed April 6, 2022).

[19] A Short History of Machine Learning -- Every Manager Should Read n.d. https://www.forbes.com/sites/bernardmarr/2016/02/19/a-short-history-of-machine-learning-every-manager-should-read/?sh=6663b9c715e7 (accessed April 16, 2022).

[20] 2022 AI, Machine Learning and Enterprise Data Trends to Watch n.d. https://www.itprotoday.com/data-analytics-and-data-management/2022-ai-machine-learning-and-data-trends-watch (accessed April 7, 2022).

[21] A Tour of Machine Learning Algorithms n.d. https://machinelearningmastery.com/a-tour-of-machine-learning-algorithms/ (accessed April 17, 2022).

[22] Aprendizaje No Supervisado en Machine Learning: Agrupación | by Victor Roman | Ciencia y Datos | Medium n.d. https://medium.com/datos-y-ciencia/aprendizaje-no-supervisado-en-machine-learning-agrupaci%C3%B3n-bb8f25813edc (accessed May 6, 2022).

[23] ¿Cómo aprenden las máquinas? Machine Learning y sus diferentes tipos | datos.gob.es n.d. https://datos.gob.es/es/blog/como-aprenden-las-maquinas-machine-learning-y-sus-diferentes-tipos (accessed May 6, 2022).

[24] FEIR 45: Regresión logística n.d. https://gauss.inf.um.es/feir/45/ (accessed April 20, 2022).

[25] Árboles de decisión: qué son y cuál es su uso en Big Data n.d.

https://www.unir.net/ingenieria/revista/arboles-de-decision/ (accessed April 19, 2022).

[26] Entropy and Information Gain in Decision Trees | by Jeremiah Lutes | Towards Data Science n.d. https://towardsdatascience.com/entropy-and-information-gain-in-decision-trees-c7db67a3a293 (accessed April 28, 2022).

[27] Decision Tree Algorithm, Explained - KDnuggets n.d. https://www.kdnuggets.com/2020/01/decision-tree-algorithm-explained.html (accessed April 28, 2022).

[28] Introduction to Random Forest in Machine Learning | Engineering Education (EngEd) Program | Section n.d. https://www.section.io/engineering-education/introduction-to-random-forest-in-machine-learning/ (accessed April 19, 2022).

[29] Precision, Recall, F1, Accuracy en clasificación - IArtificial.net n.d. https://www.iartificial.net/precision-recall-f1-accuracy-en-clasificacion/ (accessed May 9, 2022).

[30] Krawczyk B. Learning from imbalanced data: open challenges and future directions. Progress in Artificial Intelligence 2016;5:221–32. https://doi.org/10.1007/S13748-016-0094-0.

[31] scikit-learn: machine learning in Python — scikit-learn 1.1.1 documentation n.d. https://scikit-learn.org/stable/ (accessed May 30, 2022).

[32] Superior EP. Mejora de las predicciones en muestras desbalanceadas 2021.

[33] Métodos de selección de variables en el análisis de regresión lineal - Documentación de IBM n.d. https://www.ibm.com/docs/es/spss-statistics/SaaS?topic=regression-linear-variable-selection-methods (accessed May 26, 2022).