Esta tesis doctoral se centra en el estudio, implementación y aplicación de técnicas de aprendizaje automático y estadística multivariante en el emergente campo de la analítica deportiva, concretamente en el fútbol. Se aplican procedimientos comúnmente utilizados y métodos nuevos para resolver cuestiones de investigación en diferentes áreas del análisis del fútbol, tanto en el ámbito del rendimiento deportivo como en el económico. Las metodologías empleadas en esta tesis enriquece las técnicas utilizadas hasta el momento para obtener una visión global del comportamiento de los equipos de fútbol y pretenden ayudar al proceso de toma de decisiones. Además, la metodología se ha implementado utilizando el software estadístico libre R y datos abiertos, lo que permite la replicabilidad de los resultados. Esta tesis doctoral pretende contribuir a la comprensión de los modelos de aprendizaje automático y estadística multivariante para la predicción analítica deportiva, comparando su capacidad predictiva y estudiando las variables que más influyen en los resultados predictivos de estos modelos. Así, siendo el fútbol un juego de azar donde la suerte juega un papel importante, se proponen metodologías que ayuden a estudiar, comprender y modelizar la parte objetiva de este deporte. Esta tesis se estructura en cinco bloques, diferenciando cada uno en función de la base de datos utilizada para alcanzar los objetivos propuestos.

El primer bloque describe las áreas de estudio más comunes en la analítica del fútbol y las clasifica en función de los datos utilizados. Esta parte contiene un estudio exhaustivo del estado del arte de la analítica del fútbol. Así, se recopila parte de la literatura existente en función de los objetivos alcanzados, juntamente con una revisión de los métodos estadísticos aplicados. Estos modelos son los pilares sobre los que se sustentan los nuevos procedimientos aquí propuestos.

El segundo bloque consta de dos capítulos que estudian el comportamiento de los equipos que alcanzan la Liga de Campeones o la Europa League, descienden a segunda división o permanecen en mitad de la tabla. Se proponen varias técnicas de aprendizaje automático y estadística multivariante para predecir la posición de los equipos a final de temporada. Una vez realizada la predicción, se selecciona el modelo con mejor precisión predictiva para estudiar las acciones de juego que más discriminan entre posiciones. Además, se analizan las ventajas de las técnicas propuestas frente a los métodos clásicos utilizados hasta el momento. La base de datos utilizada para el análisis se compone de variables cuantitativas que almacenan información acumulada sobre las acciones de juego realizadas por los equipos a lo largo de la temporada 2018/2019.

El tercer bloque consta de un único capítulo en el que se desarrolla un código de web scraping para facilitar la recuperación de una nueva base de datos con información cuantitativa de las acciones de juego realizadas a lo largo del tiempo en los partidos de fútbol. Este bloque se centra en la predicción de los resultados de los partidos (victoria, empate o derrota) y propone la combinación de una técnica de aprendizaje automático, random forest, y la regresión Skellam, un método clásico utilizado habitualmente para predecir la diferencia de goles en el fútbol. Por último, se compara la precisión predictiva de los métodos clásicos utilizados hasta ahora con los métodos multivariantes propuestos. La base de datos contiene estadísticas partido a partido de las temporadas 2019/2020 y 2020/2021. El cuarto bloque también comprende un único capítulo y pertenece al área económica del fútbol. En este capítulo se aplica un novedoso procedimiento para desarrollar indicadores que ayuden a predecir los precios de traspaso. En concreto, se muestra la importancia de la popularidad a la hora de calcular el valor de mercado de los jugadores, por lo que este capítulo propone una nueva metodología para la recogida de información sobre la popularidad de los jugadores. La base de datos de este bloque contiene información similar a la del segundo bloque, pero relacionada con los jugadores. Además, esta base de datos se ha completado con los indicadores de popularidad propuestos.

En el quinto bloque se revelan los aspectos más relevantes de esta tesis para la investigación y la analítica en el fútbol, incluyendo futuras líneas de trabajo.