PROPUESTA PROYECTO FINAL MACHINE LEARNING

# MACHINE LEARNING EN LA LIGA

Predicción de resultados y goles de partidos de fútbol de La Primera División de España

---

Ignacio García Santamaría

María Oliva Calero

Propuesta de Proyecto Final
Machine Learning
14/11/2021

María Oliva Calero
Ignacio García Santamaría
3ºB GITT

# Tema y justificación

El fútbol es un deporte donde siempre se ha dicho que dos más dos nunca suman cuatro, es decir, que en este deporte la ciencia no tiene cabida. "Fútbol es fútbol" y algunos dicen que es imposible de predecir. Esta aleatoriedad a la hora de predecir los resultados puede justificar el hecho de que las casas de apuestas subsistan. No obstante, nosotros queremos demostrar que el fútbol sí se puede predecir, demostrando que con el uso de datos y diferentes algoritmos podemos explicar lo que va a ocurrir en el deporte rey en España.

Por ello el objetivo principal de este proyecto es explorar el problema de predicción de resultados de fútbol mediante técnicas de Machine Learning. En nuestro proyecto utilizaremos datos de los partidos jugados en las últimas 10 temporadas de La Liga. Nuestra intención es comparar los diferentes tipos de aprendizaje y algoritmos estudiados, para demostrar que en el fútbol hay factores que pueden tener más peso que otros a la hora de determinar qué equipo se alzará con la victoria.

# Objetivo y necesidades que se cubrirían

Dos objetivos de análisis.

En primer lugar, se emplearán técnicas de clasificación para predecir el resultado final de un partido (Gana local, Empate o Gana Visitante). (María)

En segundo lugar, se emplearán técnicas de regresión para realizar dos predicciones, la primera es el número de goles que marcará el equipo local (FTHG) y la segunda es determinar el número de goles que marcará el equipo visitante (FTAG). (Ignacio)

Adicionalmente, de manera conjunta vamos a discutir los resultados obtenidos, analizando cual es la mejor manera de determinar el vencedor de un encuentro. Como a partir del estudio de regresión realizado por Ignacio se puede determinar el vencedor del encuentro, tenemos un punto común de análisis de los resultados obtenidos.

El análisis se hará con la herramienta RStudio y las técnicas aprendidas en clase.

A partir de este trabajo queremos aprender que factores son los más determinantes a la hora de averiguar quien será el ganador de un encuentro. De este modo, comprenderemos mejor este deporte, llegando incluso a mejorar la experiencia a la hora verlo porque nos fijaremos más en aquellos factores determinantes.

Además, si nos fijamos más en la utilidad de los modelos que van a ser obtenidos, y no tanto en la información obtenida a medida que los modelos son desarrollados. A toda persona relacionada con el mundo del fútbol le gustaría saber cuántos goles van a ser encajados o qué equipo va a ganar antes o

María Oliva Calero
Ignacio García Santamaría
3ºB GITT

mientras se juega un partido. Asimismo, esta predicción sería basada en datos imparciales, no en la opinión condicionada de un periodista o de un analista deportivo.

## Metodología de trabajo y técnicas previstas utilizar

Una vez encontrado un conjunto de datos y tras establecer los objetivos. Tanto María como Ignacio vamos a realizar nuestro estudio de manera independiente. Estableciendo cada uno nuestras propias conclusiones acerca del mejor algoritmo para predecir, cuáles son las variables más significativas o cualquier información adicional útil para el objetivo perseguido.

Una vez realizado el análisis individual, discutiremos los resultados a los que hemos llegado cada uno, poniendo en común la información adquirida y estableciendo que modelo emplearíamos en función de la situación analizada, determinando de este modo quien se ha acercado más al objetivo común de ambos estudios (determinar el desenlace de un partido).

Técnicas previstas a utilizar por parte de Ignacio: A la hora de hacer el análisis de regresión de las dos variables a predecir, voy a analizar de manera independiente cada una, pero empleando las mismas técnicas. Para hacer el análisis de la mejor manera, y lo más completo posible, voy a emplear técnicas de regresión lineal, técnicas de regresión lineal normalizadas, KNN en regresión y árboles de decisión en regresión. Tras estudiar cada método de manera particular, tengo la intención de aplicar métodos de ensemble y así poder obtener el mejor modelo a partir de las técnicas estudiadas.

Técnicas previstas a utilizar por parte de María: De manera similar a mi compañero, voy a emplear varias técnicas, para después compararlas y obtener el mejor modelo posible. Entre estas utilizaré técnicas de regresión logística, clustering, y árboles de decisión. Igual que Ignacio, después aplicaré métodos de ensemble para obtener el mejor resultado posible.

## Datos

El conjunto de datos elegido recoge la información de los partidos de la Liga en las temporadas de 2010 a 2019. Con esta información, realizaremos un análisis de qué factores incluyen más en el resultado de los partidos cuantos goles serán encajados por cada equipo.

Tuvimos que depurar los datos, pues los datos de cada temporada se encontraban en archivos CSV distintos, cada uno con un número de variables diferentes. Para ello, filtramos los datos de cada temporada en base a las variables comunes, y después los unimos en un conjunto de datos único de 3800 observaciones y 34 variables.

## Variables

María Oliva Calero
Ignacio García Santamaría
3ºB GITT

| Nombre | Descripción | Tipo |
|---|---|---|
| **Season** | Temporada en la que se jugó el partido | Categórica |
| **HomeTeam** | Equipo local | Categórica |
| **AwayTeam** | Equipo visitante | Categórica |
| **Date** | Fecha en la que se jugó el partido | Date |
| **FTHG** | Goles totales del equipo local en el partido | Int |
| **FTAG** | Goles totales del equipo visitante en el partido | Int |
| **FTR** | Resultado final | Categórica (<br>H= Home win,<br>D=Draw, A=Away Team) |
| **HTHG** | Goles del equipo local al descanso | Int |
| **HTAG** | Goles del equipo visitante al descanso | Int |
| **HTR** | Resultado al descanso | Categórica (<br>H= Home win,<br>D=Draw, A=Away Team) |
| **HS** | Tiros del equipo local | Int |
| **AW** | Tiros del equipo visitante | Int |
| **HST** | Tiros a puerta del equipo local | Int |
| **AST** | Tiros a puerta del equipo visitante | Int |
| **HF** | Faltas cometidas por el equipo local | Int |

María Oliva Calero
Ignacio García Santamaría
3ºB GITT

| | | |
|---|---|---|
| **AF** | Faltas cometidas por el equipo visitante | Int |
| **HC** | Córneres del equipo local | Int |
| **AC** | Córneres del equipo visitante | Int |
| **HY** | Tarjetas amarillas del equipo local | Int |
| **AY** | Tarjetas amarillas del equipo visitante | Int |
| **HR** | Tarjetas rojas del equipo local | Int |
| **AR** | Tarjetas rojas del equipo visitante | Int |
| **B365H** | Cuota de ganar el equipo local de la casa de apuestas Bet365 | Numeric |
| **B365A** | Cuota de ganar el equipo visitante de la casa de apuestas Bet365 | Numeric |
| **B365D** | Cuota de empate de la casa de apuestas Bet and Win | Numeric |
| **BWH** | Cuota de ganar el equipo local de la casa de apuestas Bet and Win | Numeric |
| **BWA** | Cuota de ganar el equipo visitante de la casa de apuestas Bet and Win | Numeric |
| **BWD** | Cuota de empate de la casa de apuestas Bet and Win | Numeric |
| **WWH** | Cuota de ganar el equipo local de la casa de apuestas William Hill | Numeric |
| **WHA** | Cuota de ganar el equipo visitante de la casa de apuestas William Hill | Numeric |
| **WHD** | Cuota de empate de la casa de apuestas William Hill | Numeric |

María Oliva Calero
Ignacio García Santamaría
3ºB GITT

| | | |
|---|---|---|
| **VCH** | Cuota de ganar el equipo local de la casa de apuestas BetVictor | Numeric |
| **VCA** | Cuota de ganar el equipo visitante de la casa de apuestas BetVictor | Numeric |
| **VCH** | Cuota de empate de la casa de apuestas BetVictor | Numeric |

## Resultados esperados alcanzar

Con este trabajo esperamos poder demostrar que el fútbol se puede predecir. Nuestra idea es llegar a la conclusión de que las técnicas de Machine Learning pueden ser muy útiles en este sector, no solo en la predicción de resultados y el número de goles, que es el objetivo principal de este proyecto. Sino también, en comprender qué características del juego son determinantes, de este modo, tanto los espectadores como los propios entrenadores podrán considerar determinadas variables para conseguir que su equipo gane.

Si conseguimos demostrar que mediante la aplicación de varios algoritmos podemos extraer un conocimiento que no es empleado de manera extensa por equipos de fútbol, puede que la implementación de técnicas de Machine Learning se extienda en otras áreas de este deporte, como en la elección de fichajes, por ejemplo.