1. **Introducción**

Desde siempre he sido muy aficionado al fútbol, lo sigo desde que soy niño. Soy socio del Athletic desde hace casi 30 años.

Siempre he tenido esa duda dentro de mí: Qué tan fácil es pronosticar resultados? Podemos obtener grados altos de certidumbre cuándo hablamos de predecir un juego tan aparentemente complicado como el fútbol (donde no hay 2 partidos iguales)?

Es evidente que, sin comernos mucho la cabeza, todos somos conscientes de que hay hitos más probables que otros. La cuestión es… podemos identificar esos patrones? Podemos darle alguna utilidad?

1. **Dataset**

En este caso me he basado en los datos de la liga en curso (Liga Española 2022/2023).

¿Por qué únicamente la Liga española? ¿Por qué no tomar datos de ligas extranjeras: inglesa, italiana, alemana…?

Pienso que cada competición tiene su propia identidad y creo que juntar datos de diferentes competiciones iba a provocar un empobrecimiento de los resultados. Todo es fútbol, pero no todos son iguales. Por ejemplo: A día de hoy el número de expulsados en la Premier League es de 13, mientras que en España alcanza los 126. Me va ayudar en mis predicciones incluir datos tan dispares?

¿Por qué sólo la Liga en curso y no años anteriores?

Por el mismo motivo. Si bien es cierto, que quizá podríamos identificar patrones sostenidos a lo largo del tiempo, creo que la data más sustanciosa es la más próxima.

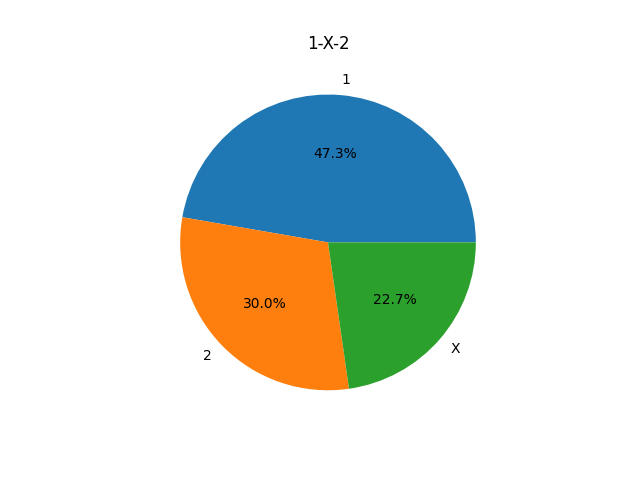
En este caso cuento con un Dataset de 330 partidos (33 jornadas \* 10 partidos) en los que he identificado hasta 113 variables. El Dataset es 100% de creación propia. Muchas de las columnas son estadísticas de los partidos obtenidos en diversas webs, pero la gran mayoría están desarrolladas a partir de ahí.

En el caso de mi estudio vamos a analizar 4 “target” de un partido de fútbol:

**Resultado (1-X-2)**

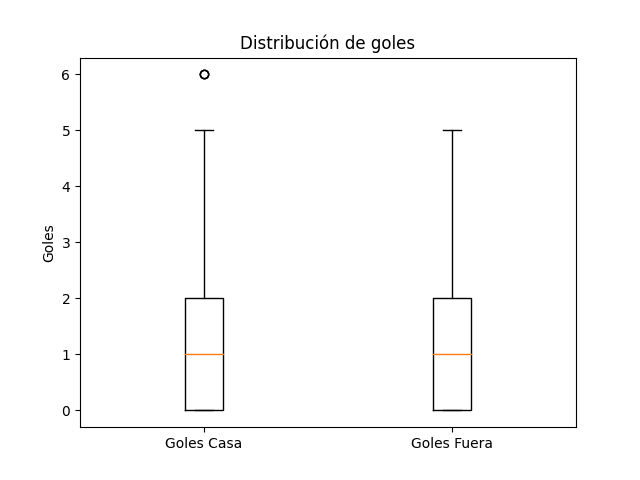
El fútbol es uno de esos deportes en los que no siempre gana uno. El empate es un resultado habitual. Lo cuál hace un poco más difícil la predicción. No se trata sólo de merecer ganar, sino de cuánto lo mereces y si es suficiente para deshacer el empate inicial.

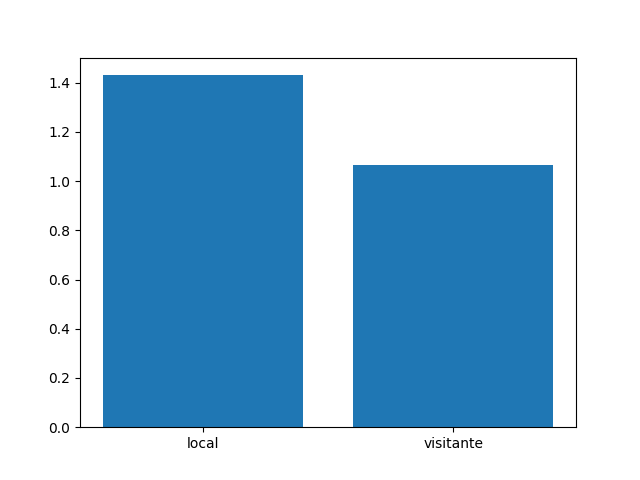
Según los datos con los que cuento, casi la mitad de los partidos se resuelven con victoria local. El resultado menos dado es el empate



**Goles**

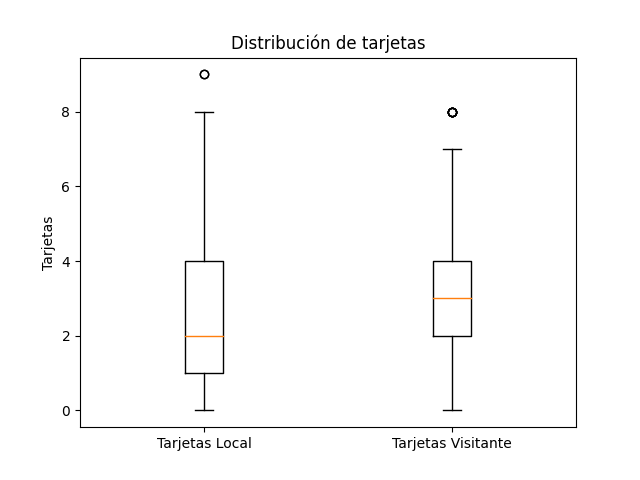
Además del signo del resultado, podemos predecir el número de goles que marcará un equipo? Vemos que la distribución de los goles en Casa y Fuera es similar. Si bien la media de goles es de 1,43 y 1,06 respectivamente.





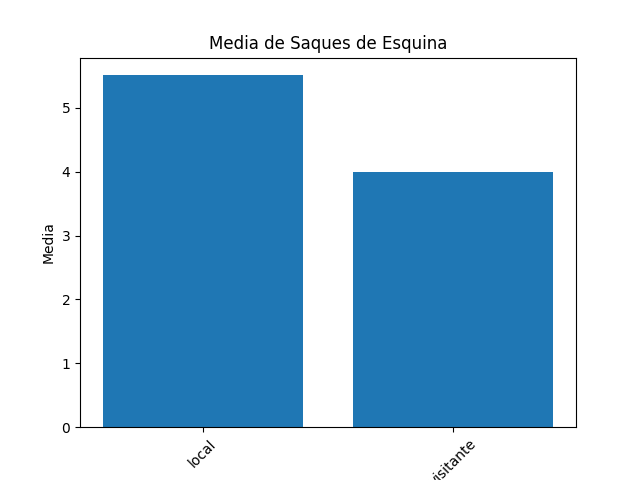
**Tarjetas (amonestaciones)**

Un clásico del fútbol y otros deportes. Podemos intuir que algunos equipos son más propensos a recibir tarjetas que otros. Analizaremos también la influencia del árbitro en cuestión. El Boxplot nos muestra un comportamiento bastante distinto entre las amonestaciones que reciben los equipos cuando actúan de local o visitante.



**Saques de Esquina**

Si bien no es una de las estadísticas más apasionante cuando analizamos un partido de fútbol, sí me parece una de las más polarizables y que son altamente significativas.



Por otro lado, buscando hacer un estudio más completo. He creado 2 metodologías de predicción:

1. Predicción en base a las estadísticas del partido. Es decir, le muestro las estadísticas de un partido acabado y quiero que me “acierte” las variables en estudio. Ejemplo: viendo el número de centros al área, despejes, paradas del portero, chuts a gol… me adivine el número de corners que ha habido en el partido.
2. Predicción antes de disputarse el partido. Tomando en consideración variables conocidas días antes de la disputa del partido, que me prediga el resultado, los goles, las tarjetas y los corners.
3. **Preprocesamiento de los Datos**

Como comentaba anteriormente la base da datos usada en este estudio ha sido creada desde 0. Las estadísticas de base las he tomado de la web <https://fbref.com/es/>.

A partir de los datos de cada uno de los partidos he configurado variables nuevas que me resultaban interesantes de cara a nuestro análisis, como pueden ser: “Media de goles marcados en casa”, “Media de goles recibidos en casa”, “Media de goles marcados fuera”, Media de goles recibidos fuera”… esto es un ejemplo de cómo subdividir una variable en varias en función de lo que queramos estudiar. Para “predecir” los goles que hará un equipo en el próximo partido es tan importante la capacidad goleadora de ese equipo, como la habilidad defensiva del contrario…

Además creé una variable totalmente subjetiva en la que se daba un peso a cada uno de los equipos en función de su clasificación (5 grupos: de más flojos a más fuertes)

1. **Modelado**

Al ser un proyecto de estudio y práctica personal, he decidido probar todos los modelos aprendidos durante las lecciones: Regresión Lineal, Regresión Polinómica, Regularización, Regresión Logística, Árboles de Decisión, Ensembles, KNN y SVM.

Como se podía intuir desde un principio, el que mejores resultados me ha dado ha sido el Random Forest.

1. **Predicción y Resultados Finales**
2. Predicción en base a las estadísticas del partido:
   1. Resultado (1-X-2): Acierta el signo del partido un 66% de las veces
   2. Goles: El MAE de local y visitante es de 0,43 y 0,37 respectivamente. En una liga con una media de 2,5 goles por partido es un número con el que se puede trabajar aunque es alto.
   3. Tarjetas: Se equivoca en 1,36 y 1,58 (local y visitante). Con una media de 5.5 tarjetas por partido... puede servir para orientarnos. Pero no es demasiado fiable
   4. Saques de esquina: He usado un clasificador agrupando el número de corners en relaciones cortas (0-2,3-5,6-7…). acierta en un 51% y 57% respectivamente. Teniendo en cuenta que son grupos pequeños no es mala cifra
3. Predicción Pre-partido:
   1. Resultado (1-X-2): En la j.32 acertó 5 de los 10 resultados. De los 5 que falló, 3 se quedó cerca (viendo la probabilidad) y 2 más lejos

En la j.33 acertó 6 de 10. De los 4 que falló, 3 se quedó cerca y 1 más lejos

* 1. Goles: Decido hacer un clasificador para que me de probabilidades, no "goles". En 3 partidos de la jornada 33 acertó el resultado exacto. En la mitad de los equipos (10/20) acertó el número de goles exactos.
  2. Tarjetas: Las únicas conclusiones positivas que se pueden sacar es que acierta en predecir cuándo un equipo va a tener menos amarillas que su rival (siempre que la diferencia sea grande). En caso contrario, veo que no hay conclusiones a destacar.
  3. Saques de esquina: Acierto de 10/20. En los partidos con diferencia amplia, los ha acertado.

1. **Conclusiones**

Tal y cómo podía imaginar, no hay una bola de cristal ni un Almanaque del futuro que nos vaya a dar la clave para hacernos millonarios con apuestas pre-claras. Pero sí es cierto que es una herramienta más para identificar patrones y ver, con datos, si son ciertas algunas intuiciones que podemos tener.

Obviamente no buscamos un modelo que nos prediga la gran sorpresa de la jornada, sino que nos diga qué es más probable que pase y en qué medida.