UNIVERSITAS
Miguel Hernández

TRABAJO FINAL DE GRADO
ESTADÍSTICA EMPRESARIAL

📊 ESTUDIO SOBRE LA LIGA ESPAÑOLA DESDE LA TEMPORADA 2016-2017 A LA TEMPORADA 2024-2025

Autor: Guillermo Sánchez Belmonte

Tutores: Joaquín Sánchez Soriano y Juan Carlos Gonçalves Dosantos

Agradecimientos

En primer lugar, me gustaría agradecer a mis profesores, y en especial a mis tutores de este trabajo, su labor a lo largo de estos años, gracias a ellos, entre otras muchas cosas, he aprendido lo bonito que es este mundo de los datos.

También me gustaría agradecer a mis padres, hermano y familia en general por estar ahí siempre. Por ser mi apoyo, una gran motivación, por ser quien me habéis aguantado y por confiar siempre en mí. Lo he conseguido y esto en parte es por vosotros.

Agradecer al ayuntamiento de Beniel (un municipio de Murcia), lugar donde hice mis prácticas, por su ayuda y su compromiso conmigo. Me demostraron que lo que había estudiado en la carrera era realmente mi pasión y que esto no es solo es un trabajo, es un juego increíble.

Por último, agradecer a esas personas que no son parte de mi familia, pero que al fin y al cabo siempre me han ayudado. En especial agradecer a esa amiga que conocí en la carrera, que me ha motivado día sí y día también a querer seguir luchando por conseguir este título, que era mi mayor objetivo.

Resumen

Este informe, correspondiente al Trabajo Final de Grado en Estadística Empresarial, tiene como objetivo principal analizar los factores que influyen en el resultado final de los partidos de la liga española, desde la temporada 2016-2017 hasta la 2024-2025.

En primer lugar, se aplicaron distintas técnicas de clasificación supervisada con el fin de predecir el resultado de cada partido —victoria local, empate o victoria visitante—. Entre los métodos utilizados se incluyen el análisis discriminante, k-vecinos más cercanos (KNN) y las máquinas de vectores de soporte (SVM). La matriz de confusión se empleó como principal herramienta para evaluar la precisión y desempeño de cada modelo.

Posteriormente, se llevó a cabo un análisis de la importancia de las variables mediante tres enfoques complementarios: el ANOVA (análisis de la varianza), las medidas de importancia del modelo Random Forest y la teoría de juegos aplicada sobre las variables significativas del ANOVA. A partir de los rankings obtenidos en cada método, se construyó una clasificación general que permite identificar las variables que mejor predicen el resultado final de un encuentro.

Finalmente, se desarrolló una página web interactiva donde se presentan los resultados de ambos objetivos mediante informes generados en R Markdown y el código en R, facilitando así la interpretación y visualización de los hallazgos. Las conclusiones principales se detallan en el punto 6 de este informe.

Índice del informe

Índice de tablas

Índice de figuras

Informe

1. Introducción.

El fútbol, ese deporte que mueve masas, llena estadios y genera conversaciones en bares o en casas sin parar. Cada semana (o cada tres días hoy en día) millones de aficionados siguen a sus equipos, celebran goles, discuten decisiones arbitrales y, por supuesto, opinan como buenos entrenadores caseros que son. Pero detrás de esa magia está algo aún más interesante: los miles de datos que nos proporcionan.

Cada partido genera miles de registros: posesión, tiros a puerta, pases completados, faltas cometidas, tarjetas mostradas o goles esperados por equipo. Estamos en un momento del fútbol en el que prácticamente todo puede medirse, y aunque todavía no hemos llegado al punto de analizar si el peinado de un jugador influye en el resultado final (que igual algún día descubrimos que las coletas dan suerte), el análisis estadístico del fútbol ha llegado para quedarse.

Este trabajo tiene como objetivo analizar qué variables son realmente decisivas en el resultado de un partido de fútbol en la liga española. Conllevando a responder a algunas de las preguntas más repetidas por los aficionados y los comentaristas de los medios de comunicación: ¿tener más posesión garantiza ganar un partido? ¿Dar muchos pases, aunque sean en horizontal, sirve de algo? ¿Los árbitros influyen realmente en los resultados? ¿Realmente afectó la pandemia, cuando los estadios estaban vacíos y el fútbol se jugaba casi en silencio, a este deporte?

También se ha intentado predecir los resultados de los partidos. Porque ¿quién no ha rellenado una quiniela pensando "voy a hacer pleno al 15 seguro"? A través de métodos estadísticos se han creado modelos que intenten estimar las probabilidades de victoria, empate o derrota de un equipo, basándonos en las variables más relevantes del estudio.

Por lo tanto, el objetivo de este trabajo es descubrir esos patrones y variables significativas que nos ayudarán a predecir mejor el resultado final de un partido, teniendo en cuenta esos datos imprevisibles de cada partido, que hacen tan mágico este deporte.

2. Clasificación supervisada y no supervisada.

Con el fin de agrupar los datos en diferentes clases, para por ejemplo categorizar mejor a los clientes o los resultados de un torneo de fútbol (como en este estudio), se utiliza la clasificación supervisada y/o no supervisada.

Para ver la diferencia entre estas clasificaciones es importante fijarse en la variable respuesta y en las clases. En la clasificación supervisada estos dos conceptos son conocidos. Por el contrario, en la clasificación no supervisada las clases no son conocidas, sino que se clasifica por la homogeneidad de los datos y los criterios comunes y no se hace con una base de datos de entrenamiento, como sí sucede en la clasificación supervisada, sino con toda la base de datos completa.

Para este informe se ha estudiado la base de datos mediante clasificación supervisada (más desarrollada en el punto 2.1 de este informe) y con algunas de sus técnicas (desarrolladas en los puntos 2.2 y 2.3 de este informe).

2.1. Introducción a la clasificación supervisada.

En este tipo de técnicas la variable respuesta es conocida, y ya está organizada en diferentes clases. Esta es la variable que se utiliza para entrenar el modelo. El objetivo de estas técnicas es predecir la variable categórica de respuesta en función del resto de variables (o de las que se seleccionen).

Dado el modelo dividimos los datos según diferentes técnicas de validación (explicadas en el punto 2.5). Para comprobar la validez del modelo la herramienta que se utiliza es la matriz de confusión, que muestra el número de aciertos y errores en cada una de las clases. Si sumamos la diagonal de esta matriz y la dividimos entre el número total de datos obtenemos el porcentaje de precisión del modelo.

Algunos campos en los que se utiliza la clasificación supervisada son en medicina o en la economía de un banco, por ejemplo. También se puede utilizar (como en este estudio) en el deporte, donde la variable categórica respuesta es el resultado final del partido y se hace una predicción para ver la cantidad de precisión de cada modelo.

2.1.1. Planteamiento del problema.

En todos los problemas de clasificación supervisada existen los siguientes conceptos:

- N: es el número de individuos (observaciones) del estudio.
- K: número de variables o características del estudio. Pueden ser numéricas o categóricas. Son las variables independientes.
- $\mathbf{X_i} = (X_{i1}, X_{i2}, \dots, X_{iK})$ es el vector de características del individuo i, con $i = 1,2, \dots, N$.
- Y_i representa el valor que asume la variable respuesta correspondiente al individuo i; es decir; cada individuo tiene un valor particular de Y.

El objetivo general es buscar la relación entre las variables dependientes $\mathbf{X}_i$ y la variable de respuesta Y_i. Así, dado un nuevo individuo $\mathbf{X}_j$ será de interés obtener el posible valor $\mathbf{Y}_j$.

2.2. Técnicas de clasificación.

Existen técnicas de clasificación que nos ayudan a conseguir el objetivo de agrupar por clases a las observaciones para mejorar la eficiencia de las empresas. Como por ejemplo para ayudar a las empresas a enfocar las necesidades del cliente o para ayudar a un equipo de fútbol a enfocar sus ideas de juego contra otro rival. A continuación, se explican más detalladamente las técnicas que se han utilizado en este estudio, pero existen algunas otras.

Antes de proceder con las explicaciones teóricas de las técnicas es importante aclarar que tanto en el análisis discriminante como el KNN se usa el clasificador de Bayes, formulado de la siguiente manera:

$$P(Y = j|X_1 = x_1, \dots, X_k = x_k) = \frac{P(Y = j)P(X_1 = x_1, \dots, X_k = x_k|Y = j)}{P(X_1 = x_1, \dots, X_k = x_k)}$$

donde:

- **P($Y = j$):** Es la **probabilidad a priori** de la clase j, es decir, la probabilidad de que ocurra la clase $Y = j$ antes de observar los valores de las variables $X_1, \dots, X_k$.
- **P($X_1 = x_1, \dots, X_k = x_k|Y = j$)**: Es la **verosimilitud**, que indica la probabilidad de observar los valores $x_1, \dots, x_k$ dado que el caso pertenece a la clase $Y = j$.
- **P($Y = j|X_1 = x_1, \dots, X_k = x_k$)**: Es la **probabilidad a posteriori**, es decir, la probabilidad de que el caso pertenezca a la clase j una vez observadas las variables $X_1, \dots, X_k$. Esta es la probabilidad que se usa para **clasificar**.
- **P($X_1 = x_1, \dots, X_k = x_k$):** Es la **probabilidad total o evidencia**, que representa la probabilidad de observar los valores $x_1, \dots, x_k$ sin importar a qué clase pertenezcan.

Esta expresión permite calcular la probabilidad de que un individuo pertenezca a una determinada clase cuando se conocen sus características $X_1 \dots X_k$. Estas probabilidades se usan con el fin de clasificar a un nuevo individuo.

2.2.1. Análisis discriminante.

El análisis discriminante fue introducido originalmente por Fisher (1936) como una técnica estadística cuyo objetivo principal es clasificar observaciones dentro de las clases de la variable respuesta (variable categórica), a partir de las variables predictoras (variables cuantitativas). Tiene dos objetivos; explicar las diferencias entre los grupos a partir de las variables observadas, y predecir o clasificar nuevas observaciones dentro de cada una de las clases a las que pertenecen.

En términos generales, el análisis discriminante busca encontrar combinaciones lineales de las variables independientes que permitan maximizar la separación entre grupos. Estas combinaciones se denominan funciones discriminantes y tienen la forma:

$$Z_j = a_{j1}x_1 + a_{j2}x_2 + \cdots + a_{jk}x_k,$$

Donde

- Z_j representa la j-ésima función discriminante.
- a_{ji} representan los coeficientes de la función discriminante, j representa la importancia de cada variable en la función discriminante.
- x_i representan las variables independientes del modelo utilizadas para clasificar a los grupos.

El número máximo de funciones discriminantes posibles viene dado por:

$$\text{Número de funciones discriminantes} = \min\,(I-1, k)$$

donde I es el número de grupos y k el número de variables predictoras.

El análisis discriminante parte del supuesto de que las observaciones dentro de cada grupo siguen una distribución normal multivariante, con medias diferentes pero una misma matriz de covarianzas común a todos los grupos. Así, si $x = (x_1, x_2, \dots, x_k)$ es un vector de variables independientes, y $G_1, G_2, \dots, G_I$ representan los distintos grupos, se asume que:

$$X \mid G_i \sim N(\mu_i, \Sigma),$$

donde μ_i es el vector de medias del grupo i y Σ es la matriz de covarianzas común.

El proceso de clasificación se fundamenta en el Teorema de Bayes, que permite calcular la probabilidad posterior de una observación x que pertenece a un grupo G_i dado su conjunto de valores de entrenamiento:

$$P(G_i \mid x) = \frac{P(x \mid G_i)P(G_i)}{\sum_{j=1}^{I} P(x \mid G_j)P(G_j)}$$

donde:

- $P(G_i)$ es la probabilidad a priori del grupo i
- $P(\,x \mid G_i\,)$ es la verosimilitud de observar x dado el grupo G_i
- $P(G_i \mid x)$ es la probabilidad a posteriori, utilizada para clasificar.

Según el criterio de Bayes, una observación se asigna al grupo cuya probabilidad posterior sea mayor:

$$\text{Clasificar } x \text{ en el grupo } G_i \text{ si } P(G_i \mid x) = \max_j\, P(G_j \mid x).$$

En el Análisis Discriminante Lineal las matrices de covarianzas son iguales, por lo tanto, la función discriminante para el grupo G_i es la siguiente:

$$\delta_i(x) = x'\Sigma^{-1}\mu_i - \frac{1}{2}\mu_i'\Sigma^{-1}\mu_i + \ln\ P(G_i).$$

La observación x se asigna al grupo cuyo valor de $\delta_i(x)$ sea mayor. Esta ecuación representa una función lineal en x, lo que implica que las fronteras de decisión entre los grupos son hiperplanos lineales en el espacio de las variables predictoras.

La función discriminante puede reescribirse en términos de la distancia de Mahalanobis, que mide la proximidad de una observación al centroide de cada grupo, teniendo en cuenta la correlación entre variables:

$$D_i^2 = (x - \mu_i)'\Sigma^{-1}(x - \mu_i).$$

La observación se asigna al grupo que minimiza D_i^2. La distancia de Mahalanobis resulta más adecuada que la distancia euclídea, pues considera la estructura de covarianzas común entre las variables.

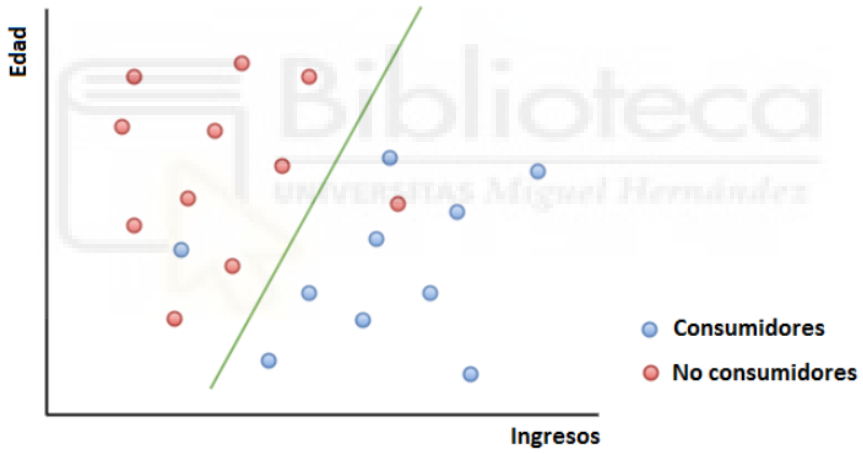

***Figura 1**: Ejemplo de análisis discriminante. Fuente: Martínez, T. L. (2000).*

En la figura 1 se muestra un ejemplo de esta técnica que nos indica dos clases (consumidores y no consumidores) y una función discriminante que es la línea verde, que divide la base de datos en dos grupos con el objetivo de encontrar la pureza en el método y clasificar a los individuos correctamente.

2.2.2. K-Nearest Neighbors.

El *K*-Nearest Neighbors también conocido como KNN (*Martínez, T. L. 2000*) es una técnica de aprendizaje supervisado sencilla y muy utilizada a día de hoy en aprendizaje automático. Consiste en clasificar a los nuevos registros según los K vecinos más cercanos, es decir establecer qué clase se repite más entre los vecinos del nuevo valor con el clasificador de Bayes.

Algunas de las cosas a tener en cuenta antes de realizar esta técnica son las siguientes:

- Si hay millones de muestras, el sistema selecciona un número de individuos, los cuales se utilizarán para realizar el modelo, los considerados como prototipos. Existen técnicas para extraer esos individuos, algunas son más complejas y otras más sencillas. Una de ellas es el método de las k-medias, que crea tantos grupos como individuos se desea seleccionar y usa los centroides de estos grupos como prototipos.
- El problema de la "Maldición de la Dimensionalidad" nos explica que si el modelo tiene muchas variables pierde precisión y hay que conseguir reducir dimensionalidad. Algunas opciones a seguir son utilizar técnicas como componentes principales o análisis factorial.

Matemáticamente este algoritmo se formula de esta manera:

$$P(Y = j \mid X = x) = \frac{1}{K} \sum_{i \in \mathrm{Nx}} I(y_i = j)$$

Donde:

- $X = (x_1, \ldots, x_k)$: Es el vector de características de la observación cuya clase queremos predecir.
- Y: Es la variable de clase o categoría que queremos estimar.
- j: Es una de las posibles clases.
- $I(y_i = j)$: es una función indicadora que vale:

$$I(y_i = j) = \begin{cases} 1 & \text{si } y_i = j \\ 0 & \text{en otro caso} \end{cases}$$

- N_x: es el conjunto de índices de los K vecinos más cercanos al punto x.
- K: es el número de vecinos más cercanos preestablecido.

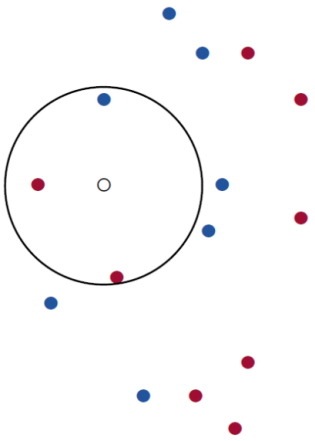

***Figura 2**: Ejemplo de KNN. Fuente: Martínez, T. L. (2000).*

En la figura 2 se muestran dos clases (diferenciadas con el color rojo y el color azul) y un nuevo individuo (en color blanco). El K seleccionado es 3 y, por lo tanto, se buscan los 3

puntos más cercanos al blanco, siendo dos rojos y uno azul, por lo que la clase seleccionada es la roja.

También hay que tener en cuenta los parámetros de este algoritmo, que se han fijado teniendo en cuenta las siguientes consideraciones:

- **Los valores K**: Este valor se puede obtener de diferentes maneras, y puede ser solo un valor K o varios valores K. Algunos de los criterios para asignar un valor K es utilizando la validación cruzada o también haciendo la raíz cuadrada del número de observaciones que tiene el estudio. En este ejemplo se ha utilizado este último criterio, seleccionando el K del redondeo de la raíz y los K+1 y K-1 como los otros K.

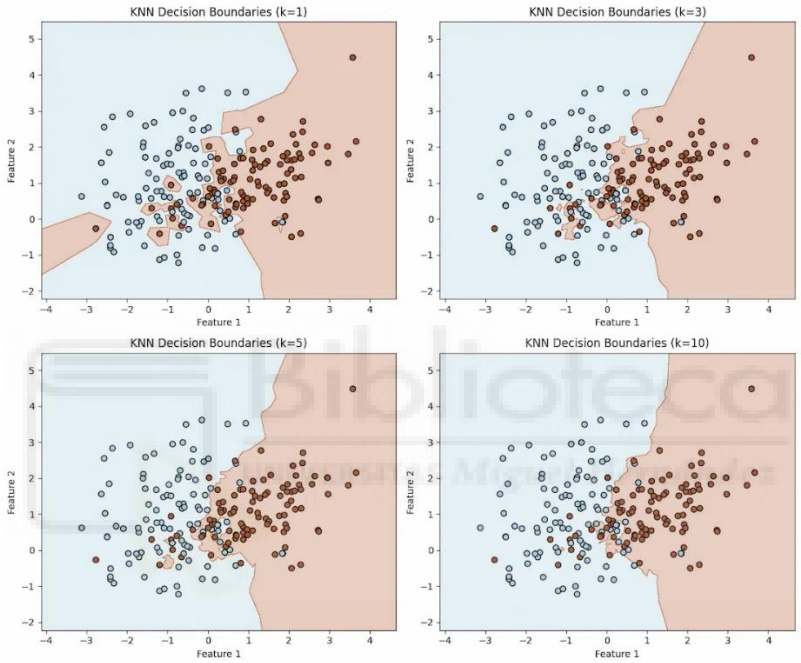

***Figura 3**: Ejemplo de diferentes K en KNN. Fuente: Sterratt, D., & Gal, K. (2024).*

En la Figura 3 se observa cómo cambia la clasificación de los puntos de datos al modificar el valor de *K* en el algoritmo KNN. Cada gráfico representa una configuración distinta de *K* (1, 3, 5 y 10). A medida que aumenta el número de vecinos, las fronteras de decisión se vuelven más suaves y generalizadas

- **Distancia**: Este concepto nos muestra la separación entre puntos en un espacio multidimensional. Hay algunas muy usadas como la distancia euclídea y la distancia de Mahalanobis, aunque existen otras que también se pueden usar. Las principales diferencias son las siguientes:
    - **Distancia Euclídea**: Se basa en la geometría clásica. Calcula la raíz cuadrada de la suma de las diferencias al cuadrado entre las coordenadas correspondientes de dos puntos. Esta distancia trata a todas las dimensiones de manera equivalente y no considera correlaciones entre variables.

- **Distancia de Mahalanobis**: Tiene en cuenta la correlación entre las variables y la dispersión de los datos. Ajusta la escala de cada dimensión según su varianza y la covarianza entre variables, de manera que las diferencias en direcciones con mayor variabilidad tienen menor peso.
- Otras distancias utilizadas son la **distancia de Manhattan**, que suma las diferencias absolutas entre coordenadas o la **distancia de Minkowski**, que generaliza la euclídea y la de Manhattan, además de las medidas basadas en similitud de coseno, que tienen como principal característica que se fijan en la orientación de los vectores más que en su magnitud.

Esta técnica presenta como principal ventaja su carácter visual e intuitivo, lo que facilita su comprensión e interpretación. No obstante, una de sus limitaciones es la elevada variabilidad y el elevado coste computacional, dado que requiere emplear la totalidad de la base de datos para clasificar a un nuevo individuo, a diferencia de otras metodologías de clasificación más eficientes.

2.2.3. Support Vector Machine.

El Support Vector Machine (SVM), o en español Máquina de Vectores de Soporte, es una técnica de clasificación supervisada ampliamente utilizada dentro del ámbito del aprendizaje automático (*Martínez, T. L. (2000)*). Su objetivo principal consiste en encontrar un hiperplano óptimo que separe las clases de datos, maximizando la distancia entre los puntos más cercanos de cada clase, denominados vectores de soporte.

El SVM busca estimar los coeficientes de un hiperplano de separación del tipo

$$\beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_k X_k = 0$$

de manera que, para un conjunto de observaciones $(\mathbf{x}_i, y_i)$ con $y_i \in \{-1, +1\}$, los datos queden correctamente clasificados según el signo de la función

$$f(\mathbf{x}_i) = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \cdots + \beta_k x_{ik}.$$

Así, los puntos para los que $f(\mathbf{x}_i) > 0$ se asignan a la clase $+1$ y aquellos con $f(\mathbf{x}_i) < 0$ a la clase -1.

El método persigue maximizar el margen entre las dos clases, es decir, la distancia entre el hiperplano y los puntos más cercanos de cada clase (denominados *vectores de soporte*). Un margen amplio se asocia con una mejor capacidad de generalización del modelo frente a nuevos datos.

Existen dos márgenes (o distancias) entre las dos clases:

- **El SVM de margen rígido** que se utiliza cuando los datos son perfectamente separables de forma lineal, es decir, cuando es posible encontrar un hiperplano que divida sin errores las clases.

El problema se formula como:

$$\min_{\boldsymbol{\beta}, \beta_0} \frac{1}{2} \| \boldsymbol{\beta} \|^2$$

sujeto a:

$$y_i(\beta_0 + \beta_1 x_{i1} + \cdots + \beta_k x_{ik}) \geq 1, i = 1, \ldots, n$$

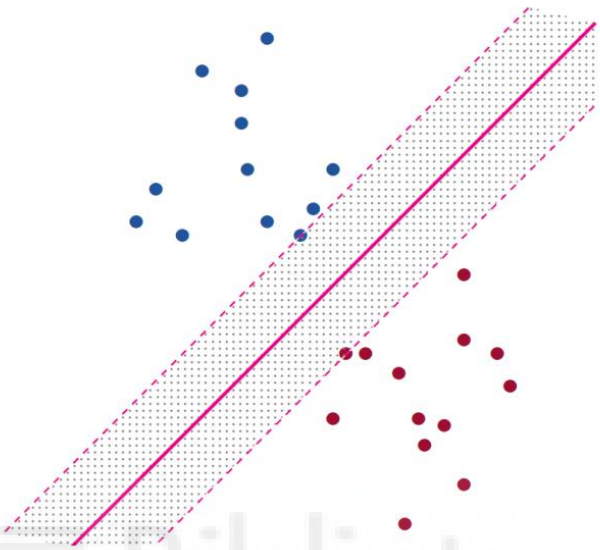

***Figura 4**: Ejemplo de SVM con margen rígido. Fuente: Martínez, T. L. (2000).*

- **El SVM de margen blando** que se utiliza cuando admitimos errores de clasificación mediante variables de holgura ξ_i.
  El problema se formula como:

$$\min_{\beta, \xi} \frac{1}{2} \sum_{j=1}^{k} \beta_j^2 + C \sum_{i=1}^{n} \xi_i$$

sujeto a:

$$y_i(\beta_0 + \beta_1 x_{i1} + \cdots + \beta_k x_{ik}) \geq 1 - \xi_i, \xi_i \geq 0, i = 1, \ldots, n$$

donde

C es un parámetro que controla el equilibrio entre la anchura del margen y la penalización de los errores.

ξ_i es la variable de holgura que mide cuanto viola el punto i la condición del margen.

  - Si $\xi_i = 0$: el punto i está correctamente clasificado y fuera o justo en el margen.
  - Si $0 < \xi_i < 1$: el punto i está dentro del margen**.**
  - Si $\xi_i > 1$: el punto i está mal clasificado.

Valores pequeños de C permiten márgenes más amplios (mayor tolerancia al error), mientras que valores grandes los reducen, haciendo el modelo más estricto.

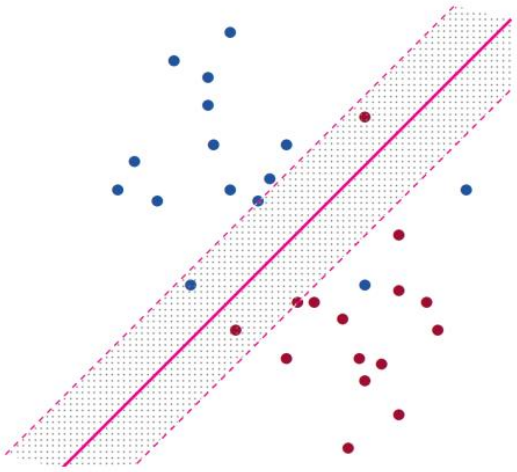

***Figura 5**: Ejemplo de SVM con margen blando. Fuente: Martínez, T. L. (2000).*

Kernels.

Cuando las fronteras de decisión no son lineales, el SVM utiliza funciones kernel, que transforman los datos originales a un espacio de características de mayor dimensión, donde es posible lograr una separación lineal. Los kernels más empleados son:

- **Kernel lineal**: Se utiliza cuando los datos son linealmente separables en el espacio original:

  $$K(x, y) = x^T y$$

  donde x, y son vectores de características.

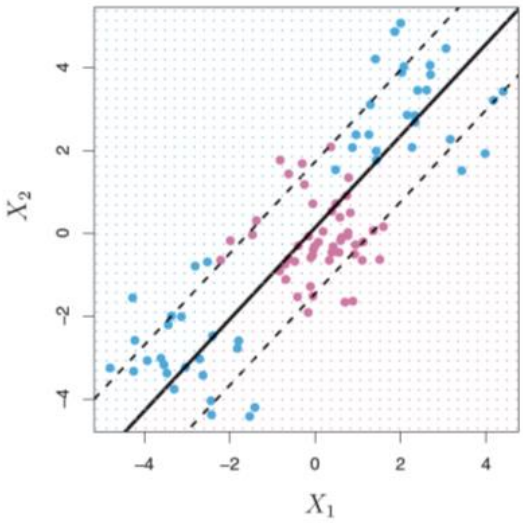

***Figura 6**: Ejemplo de un kernel lineal. Fuente: Martínez, T. L. (2000).*

- **Kernel polinómico**: Permite modelar relaciones no lineales elevando los productos internos a una potencia:

$$K(x, y) = (x^T y + c)^d$$

donde c es un parámetro de ajuste y d es el grado del polinomio.

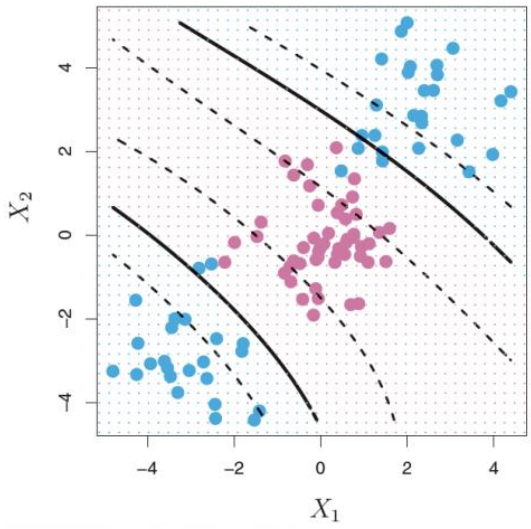

***Figura 7**: Ejemplo de un kernel polinómico. Fuente: Martínez, T. L. (2000).*

- **Kernel radial**: Evalúa la similitud en función de la distancia euclídea entre los vectores.

$$K(x, y) = \exp\left(-\gamma \parallel x - y \parallel^2\right)$$

siendo γ un parámetro que controla la influencia local de cada punto.
Este kernel es especialmente útil cuando las fronteras de decisión son altamente no lineales, ya que solo las observaciones cercanas a la observación de test influyen en su clasificación.

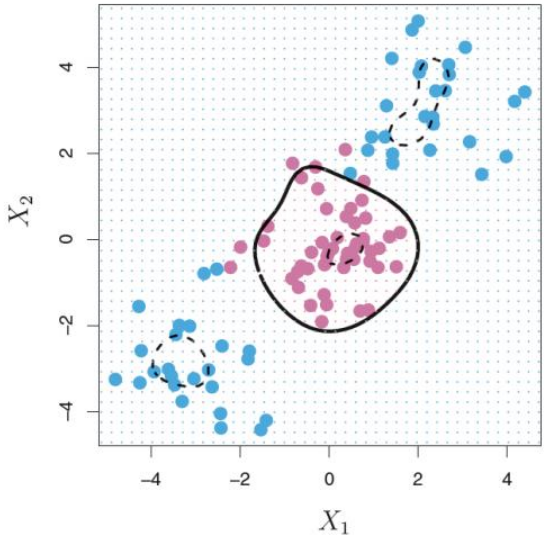

***Figura 8**: Ejemplo de un kernel radial. Fuente: Martínez, T. L. (2000).*

Aunque el SVM fue pensado para problemas de clasificación binaria, es decir 2 clases, también se puede utilizar con más de dos categorías mediante estrategias como "uno contra uno" o "uno contra todos". En la primera, se construye un clasificador SVM para cada par de clases, mientras que en la segunda se entrena un SVM para cada clase frente a las restantes).

2.3. Random forest y árboles de clasificación.

Otras técnicas también muy utilizadas en el aprendizaje supervisado son los árboles de clasificación o de decisión y los bosques aleatorios o random forest (*Martínez, T. L (2000)*).

Árbol de clasificación.

Los árboles de clasificación son una técnica visual y simple de la clasificación supervisada, que tiene como objetivo dividir la base de datos en grupos más homogéneos, con el fin de clasificar las nuevas observaciones adecuadamente, buscando el objetivo de disminuir la impureza de los nodos (decimos que un nodo es puro cuando solo hay observaciones de una misma clase).

El nodo padre (nodo raíz) se divide en dos nodos, los conocidos como nodos hijos, estos a su vez se convierten en nodos padres, que tendrán nodos hijos y así sucesivamente hasta alcanzar el criterio de parada preestablecido (estos intermedios se llaman nodos intermedios y los últimos se llaman nodos hoja). Finalmente, el nodo hoja es la clase de la variable respuesta, es decir la variable dependiente.

La regla de división se basa en dividir el nodo a partir de:

- Una cota en las variables numéricas, donde en la parte izquierda se quedan los datos inferiores o iguales al valor (por ejemplo, "edad" <= 30) y en la parte derecha los superiores a ese valor (es decir, los superiores a 30 años en el ejemplo).
- Dividir por clases en las variables categóricas, es decir NO sucede lo que indica en el lado izquierdo y SI sucede en la parte derecha (por ejemplo, "¿Va a llover?". NO llueve sería las observaciones que van en la izquierda, SI llueve los que van a la derecha).

Las medidas de pureza más comunes son el índice de Gini y la entropía cruzada, aunque existen algunas otras. La formulación de estas dos medidas de pureza es la siguiente:

$$Gini(t) = \sum_{j=1}^{I} p_j(1 - p_j)$$

$$Entropía(t) = -\sum_{j=1}^{I} p_j * \ln(p_j)$$

donde p_j representa la proporción de observaciones del conjunto de entrenamiento que pertenecen a la clase j en la región t. Siendo un valor pequeño una señal positiva de pureza del modelo.

El criterio de parada óptimo para esta técnica sería lograr la clasificación correcta de todos los individuos, es decir, que cada nodo sea completamente puro. No obstante, este escenario resulta prácticamente inalcanzable en bases de datos con un número suficiente de observaciones, ya que al hacer crecer el árbol en exceso se generan nodos con muy pocas observaciones en cada clase.

Por este motivo, el criterio de parada más habitual consiste en establecer un número mínimo de observaciones que debe contener cada nodo. Finalmente, en cada nodo hoja se asigna una clase de la variable respuesta, siendo esta la clase mayoritaria entre las observaciones que pertenecen a dicho nodo.

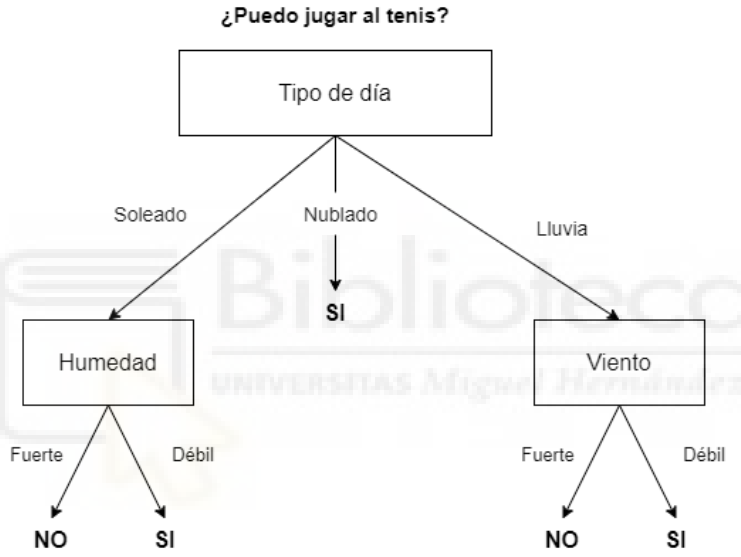

***Figura 9**: Ejemplo de un árbol de clasificación. Fuente: CDR Book.*

En la figura 9 el nodo raíz es el conjunto de datos de entrenamiento, las particiones son las preguntas sobre las variables (tipo de día, humedad, viento). Los nodos hoja son los "SI" y los "NO", que responden a la pregunta ¿Puedo jugar al tenis?

La ventaja de esta técnica es que es muy visual y sencilla de entender, además de que sirve para variables numéricas y categóricas, pero por el contrario tiene la desventaja de tener una alta variabilidad, ya que una pequeña variación en los datos puede provocar un gran desajuste en el modelo.

Bosques aleatorios (Random Forest).

El random forest es una técnica que está muy relacionada con los árboles de clasificación, ya que combina múltiples árboles de clasificación con el objetivo de mejorar la precisión y disminuir el error en el modelo. Para ello encasilla la nueva observación en cada uno de los árboles del modelo y con el grupo mayoritario entre todos los árboles, clasifica a esa nueva observación.

El criterio de parada y la ramificación de los árboles es igual que en los propios árboles de clasificación, es decir se utiliza el índice de Gini o la entropía cruzada para buscar la pureza de los nodos y se utiliza el criterio de un mínimo de observaciones para dejar de ramificar, aunque no olvidemos que el objetivo principal es un nodo puro, es decir un nodo con una sola clase (criterio óptimo).

Otra cuestión relevante al construir un bosque aleatorio es determinar cuántos árboles deben generarse para obtener un modelo con bajo error de predicción, sin que ello implique un coste computacional excesivo. En general, se recomienda que el número de árboles sea, como mínimo, diez veces superior al número de variables incluidas en el modelo.

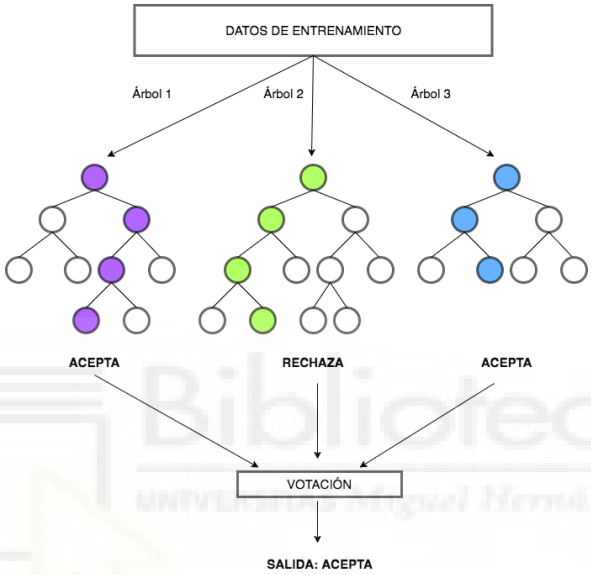

***Figura 10**: Ejemplo de un bosque aleatorio. Fuente: CDR Book.*

En la figura 10 vemos cómo realizada la clasificación en cada uno de los árboles obtenemos una respuesta final para esa nueva observación. Hecho eso, vemos entre todos los árboles cuál es la clase escogida mayoritariamente por el bosque (que es lo que se toma una especie de votación) y clasifica a esta nueva observación como “acepta”.

Esta pregunta nos lleva también a reflexionar sobre las ventajas y desventajas de esta técnica. La ventaja es la precisión (muy superior a los árboles), ya que se realizan muchos árboles conllevando a mucha fiabilidad, pero por el contrario es muy costosa computacionalmente, su gran desventaja, por la misma razón de tener que ejecutar tantos árboles.

2.4. Matriz de confusión.

La matriz de confusión (*Martínez, T. L. (2000)*) se usa en la clasificación supervisada con el fin de comprobar los aciertos y errores del modelo, es decir, la precisión en cada una de las técnicas. El otro objetivo principal es comparar entre las diferentes técnicas para visualizar cuál de ellas tiene mayor porcentaje de precisión.

El resultado de esta medida, como indica su propio el nombre, se muestra en una matriz. Tanto en las columnas como en las filas se muestra la variable categórica tenida en cuenta como variable respuesta y las clases posibles a las que pertenece cada uno de los registros. En las columnas se recogen los valores reales de la variable respuesta estudiada y en las filas se encuentran los valores predichos tras realizar cada uno de los modelos.

Por ejemplo, consideremos el caso de la pandemia que azotó a toda la sociedad, donde se predice si hay enfermedad o no y se compara con la realidad.

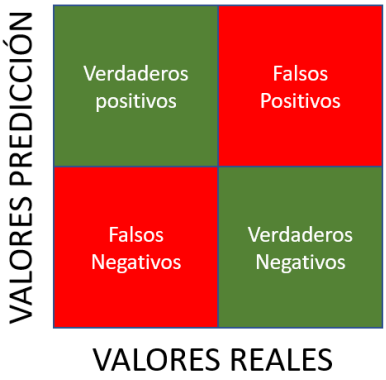

***Figura 11**: Ejemplo de la matriz de confusión. Fuente: Barrios Arce, J. I. (2019).*

Esta matriz permite calcular diversas métricas de evaluación del modelo:

- **Verdaderos Positivos (arriba izquierda)**: Son los casos correctamente clasificados como positivos.
- **Falsos Positivos (arriba derecha)**: Son los casos donde el modelo predice positivo, pero en realidad eran negativos.
- **Falsos Negativos (abajo izquierda)**: Son los casos donde el modelo predice negativo, pero realmente eran positivos.
- **Verdaderos Negativos (abajo derecha)**: Son los casos correctamente clasificados como negativos.

A partir de estos cuatro valores se calculan las siguientes métricas:

- **Precisión**: Proporción total de aciertos del modelo.

$$\text{Precisión} = \frac{VP + VN}{VP + FP + FN + VN}$$

- **Sensibilidad**: Proporción de positivos reales que el modelo identifica correctamente.

$$\text{Sensibilidad} = \frac{VP}{VP + FN}$$

- **Especificidad**: Proporción de negativos reales identificados correctamente.

$$\text{Especificidad} = \frac{VN}{VN + FP}$$

- **Tasa de falsos positivos**: % de falsos observados predichos como verdaderos.

$$\text{Falsos positivos} = \frac{FP}{VN + FP}$$

- **Tasa de falsos negativos**: % de verdaderos observados predichos como falsos.

$$\text{Falsos negativos} = \frac{FN}{FN + VP}$$

Aunque todas tienen importancia la que utilizaremos para evaluar la fiabilidad de cada una de las técnicas y mostrar los resultados será la métrica de la precisión que muestra la suma de la diagonal de la matriz entre la suma de toda la matriz.

2.5. Técnicas de validación.

En cada una de las técnicas de aprendizaje automático se utiliza un conjunto de datos de entrenamiento y otro de test. Para seleccionar esos datos se hace de forma aleatoria, pero existen técnicas de extracción de esos registros de entrenamiento cuyo objetivo es obtener el menor número de errores en las matrices de confusión de cada una de las técnicas. Esas técnicas de extracción son las siguientes:

- **El método de validación simple**: También conocido como método de retención o *holdout method* se basa en seleccionar un conjunto de los datos como base de datos de entrenamiento (normalmente el 80% de los datos; igual que en este estudio) y el resto (20%) como test. Este método tiene como ventajas que es más sencillo computacionalmente, pero como desventaja que es más variable que otros métodos.

***Figura 12**: Ejemplo del método de validación simple. Fuente: Martínez, T. L. (2000).*

- **El método de *Leave-one-out Cross Validation***: Selecciona todo el conjunto de datos exceptuando una observación como conjunto de entrenamiento y esa observación la coge como test. Este proceso lo realiza n veces (es decir, el número de observaciones que haya en el problema) y del promedio de este cálculo se obtiene el error. Este método lleva un coste computacional muy grande si la n es grande, aunque es más fiable en cuanto a variabilidad que el método anterior, ya que selecciona prácticamente todas las observaciones como conjunto de entrenamiento.

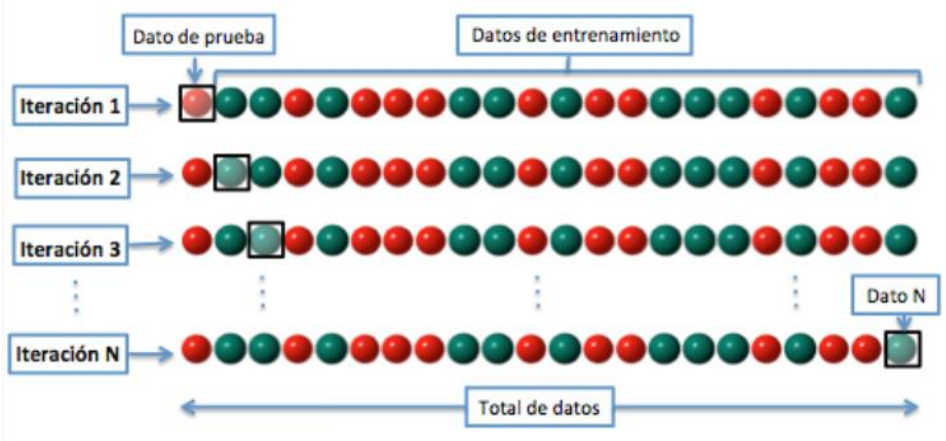

***Figura 13**: Ejemplo del método de Leave-one-out Cross Validation. Fuente: Martínez, T. L. (2000).*

- **La validación cruzada**: Este método es similar a la validación simple. En este caso, las observaciones se dividen en k grupos del mismo o similar tamaño. Uno de estos k grupos se utiliza como conjunto de test, mientras que el resto de los grupos se usan de base de datos de entrenamiento. El proceso es igual que la validación simple. Hecho esto, se hace lo mismo para cada uno de los grupos. Por ejemplo, si hay 3 grupos, se debe hacer este proceso 3 veces, siendo cada vez un grupo el conjunto de test. Realizado eso se calcula el promedio del error y ese es el error de la técnica. Aunque es menos costoso computacionalmente que las dos técnicas anteriores, si la base de datos tiene muchas observaciones es costoso igualmente, pero suele ser la técnica apropiada a utilizar.

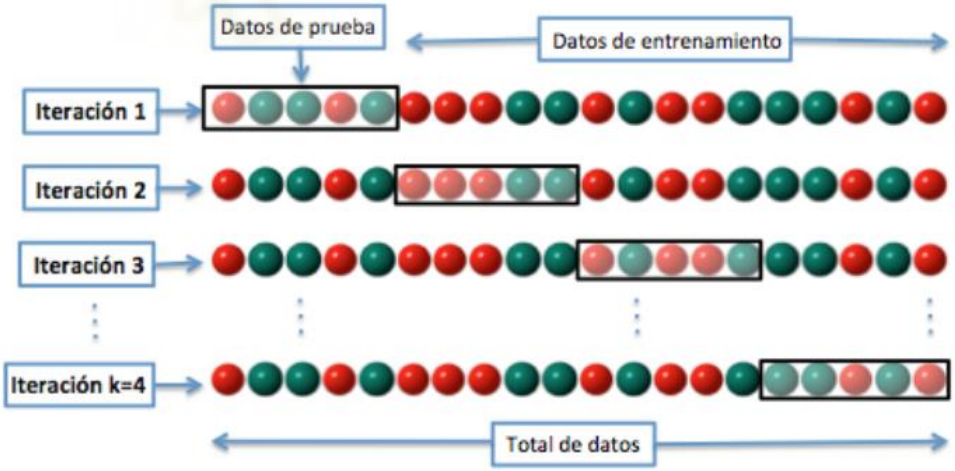

***Figura 14**: Ejemplo del método de validación cruzada. Fuente: Martínez, T. L. (2000).*

- **El método de Bootstrap**: Método Bootstrap: este método persigue el mismo objetivo que las técnicas anteriores; sin embargo, se diferencia de ellas en el procedimiento de muestreo, ya que la extracción se realiza con reemplazo. Es decir, una vez seleccionada una observación, esta puede volver a ser seleccionada en extracciones posteriores, repitiéndose el proceso múltiples.

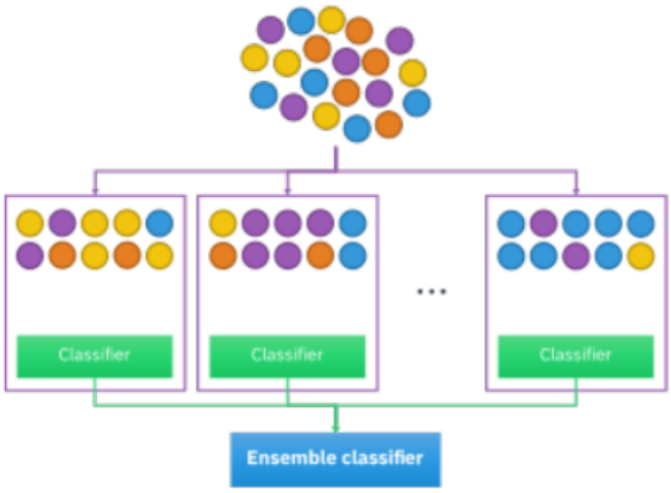

***Figura 15**: Ejemplo del método de Bootstrap. Fuente: Martínez, T. L. (2000).*

2.6. Medidas de influencia del Random Forest.

Realizada la técnica del random forest la computadora muestra algunos gráficos que sirven para responder nuestra pregunta de qué variables son más importantes para predecir la variable respuesta. Estos índices son dos:

- **Mean Decrease Accuracy (Disminución Media de la Precisión)**
  Este indicador mide cuánto disminuye la precisión del modelo al eliminar una variable del estudio. El orden en el que aparecen las variables en el grafico que nos muestra R es importante, ya que es el orden de importancia de las variables. El valor que nos muestra a la izquierda, nos indica la importancia de las variables para ver la precisión del modelo. Por lo que, si quitamos estas variables del estudio la precisión disminuirá, y, por lo tanto, el modelo a priori perderá fiabilidad.
- **Mean Decrease Gini (Disminución Media del Índice de Gini)**
  Este índice nos muestra la pureza de los nodos, es decir, los grupos que crean los árboles de clasificación. A mayor pureza, los grupos son más homogéneos, por lo tanto, un valor alto en una variable en este indicador implica una predicción mejor, y si al quitar una de esas variables implica que se reduce la homogeneidad y empeora el modelo.

3. Teoría de juegos.

La teoría de juegos (también denominada teoría de la decisión interactiva) constituye una rama de las matemáticas y de la economía que estudia los comportamientos estratégicos entre los sujetos conocidos como jugadores, los cuales persiguen maximizar sus beneficios o minimizar sus pérdidas (Von Neumann & Morgenstern, 1944).

Esta teoría se centra en la toma de decisiones en contextos donde el resultado de cada individuo no depende exclusivamente de sus propias acciones, sino también de las decisiones de los demás jugadores implicados. También es importante saber que la elección de cada uno de los jugadores se hace sin conocer con certeza la decisión final del resto de participantes.

Entre los conceptos fundamentales de esta teoría destacan los siguientes:

- **Jugadores**: Son los participantes del juego.
- **Acciones de los jugadores**: Conjunto de decisiones o estrategias posibles que cada jugador puede adoptar en función de las expectativas sobre las acciones de los demás.
- **Resultado del juego**: Es el conjunto de posibles desenlaces derivados de las combinaciones estratégicas adoptadas por los jugadores.
- **Pago**: Representa el beneficio o coste obtenido por cada jugador tras la resolución del juego.
- **Estrategia**: Es el plan o regla de decisión que cada jugador utiliza para elegir sus acciones ante las posibles alternativas del juego.

Una de las aportaciones más relevantes en esta teoría fue desarrollada por John F. Nash (1950), quien formuló el concepto de equilibrio de Nash, definido como una situación en la que ningún jugador tiene incentivos para modificar drásticamente su estrategia, dado que cualquier cambio le reportaría un perjuicio o una ganancia inferior. Este equilibrio permite analizar la estabilidad de los sistemas estratégicos no cooperativos, siendo un pilar fundamental de la teoría moderna de juegos.

La teoría de juegos actualmente se utiliza en muchos campos científicos, algunos de ellos son la biología evolutiva, sociología, psicología, política y ciencias del deporte, aunque también está en otros campos (Myerson, 1991).

Si nos fijamos en el caso de este estudio, que es el contexto deportivo, esta teoría permite modelizar el comportamiento estratégico de jugadores o variables de rendimiento que interactúan en la búsqueda de un objetivo común, como es la victoria final.

Un ejemplo puede ser tirar un penalti. El lanzador y el portero son los jugadores que deben tomar decisiones simultáneas: el primero debe elegir hacia qué lado tirar (izquierda, centro o derecha), mientras que el segundo debe decidir hacia qué lado lanzarse para intentar detener el balón. Ninguno conoce con certeza la elección del otro, por lo que ambos tratan de anticiparse estratégicamente. El equilibrio de Nash se alcanza cuando ambos adoptan estrategias mixtas (por ejemplo, variar aleatoriamente el lado del disparo

o de la parada) de modo que ninguno tiene incentivos para cambiar su patrón de decisión, ya que cualquier modificación predecible podría ser aprovechada por el rival.

3.1. Juegos cooperativos.

Dentro de la teoría de juegos, se distinguen dos grandes enfoques: los juegos no cooperativos, donde cada jugador actúa de forma independiente, y los juegos cooperativos, en los que los jugadores pueden formar coaliciones o acuerdos para maximizar beneficios conjuntos (Osborne & Rubinstein, 1994).

Los juegos cooperativos se basan en la idea de que los participantes reconocen el valor de actuar de manera conjunta. En este tipo de juegos, se busca determinar cómo debe distribuirse equitativamente el beneficio total obtenido por la coalición entre sus miembros, teniendo en cuenta la contribución de cada uno al resultado global.

En los juegos cooperativos, los jugadores pueden formar coaliciones para maximizar su ganancia conjunta. Matemáticamente, un juego cooperativo se define por:

$$G = (N, v)$$

donde:

- $N = \{1,2, \dots, n\}$ es el conjunto de jugadores.
- $v: 2^N \to \mathbb{R}$ es la función característica o valor del juego, que asigna a cada coalición $S \subseteq N$ el valor total o beneficio conjunto que pueden obtener cooperando.

Por convención, $v(\emptyset) = 0$.

Una herramienta fundamental para cuantificar estas contribuciones es el valor de Shapley, propuesto por Lloyd Shapley (1953). Este valor calcula la contribución marginal media de cada jugador (o variable) al resultado total del juego, considerando todas las posibles combinaciones de coaliciones. En otras palabras, el valor de Shapley mide la importancia relativa de cada variable dentro del sistema cooperativo, proporcionando una valoración del papel que cada factor desempeña en el éxito global.

El valor de Shapley ($\phi_i(v)$) cuantifica la contribución marginal promedio del jugador i a todas las posibles coaliciones. Se define como:

$$\phi_i(v) = \sum_{S \subseteq N \setminus \{i\}} \frac{|S|!(n - |S| - 1)!}{n!} [v(S \cup \{i\}) - v(S)]$$

donde:

- $|S|$ es el número de jugadores en la coalición S.
- $v(S \cup \{i\}) - v(S)$ es la **contribución marginal** del jugador i a la coalición S.
- El término combinatorio $\frac{|S|!(n-|S|-1)!}{n!}$ pondera cada coalición según la probabilidad de que el jugador i se una en un orden aleatorio.

El valor de Shapley cumple cuatro propiedades fundamentales:

- **Eficiencia:** $\sum_{i \in N} \phi_i(v) = v(N)$
- **Simetría:** Si dos jugadores contribuyen igual a todas las coaliciones, tienen el mismo valor.
- **Jugadores nulos:** Si un jugador no aporta nada adicional, su valor es cero.
- **Aditividad:** Si se combinan dos juegos, los valores se suman.

3.2. Selección de variables con ANOVA.

El Análisis de la Varianza (ANOVA) es una técnica estadística desarrollada por Ronald A. Fisher (1925) que permite comparar las medias de dos o más grupos con el fin de determinar si existen diferencias significativas entre ellas. Su objetivo principal es analizar la influencia que una o varias variables independientes (factores) tienen sobre una variable dependiente cuantitativa.

La hipótesis nula del ANOVA establece que todas las medias poblacionales son iguales, mientras que la hipótesis alternativa plantea que al menos una media difiere significativamente de las demás.

$$\begin{Bmatrix} H_0: \mu_1 = \mu_2 = \mu_3 = \cdots = \mu_k \\ H_1: \exists\, i, \frac{j}{\mu_i} \neq \mu_j \end{Bmatrix}$$

donde $\mu_i$representa la media del grupo i y k el número total de grupos.

En esta técnica estadística hay que tener claros dos conceptos importantes:

- La variabilidad entre grupos (SSB) que mide la diferencia entre las medias de los grupos y la media general.
- La variabilidad dentro de los grupos (SSW) que mide la variabilidad de los datos dentro de cada grupo individual.

La Suma Total de Cuadrados (SST) se expresa como:

$$SST = \sum_{i=1}^{k} \sum_{j=1}^{n_i} (x_{ij} - \bar{x})^2$$

donde:

- x_{ij} es la observación j-ésima del grupo i.
- $\bar{x}$ es la media general de todas las observaciones.
- n_i es el número de observaciones del grupo i.
- $\bar{x}_i$ representa la media muestral del grupo i
- k es el número de grupos.

Esta suma total se puede descomponer como:

$$SST = SSB + SSW$$

donde:

$$SSB = \sum_{i=1}^{k} n_i(\bar{x}_i - \bar{x})^2$$

$$SSW = \sum_{i=1}^{k} \sum_{j=1}^{n_i} (x_{ij} - \bar{\bar{x}}_i)^2$$

A partir de estas sumas de cuadrados se calculan los promedios cuadráticos (Mean Squares):

$$MSB = \frac{SSB}{k-1}, MSW = \frac{SSW}{N-k}$$

donde $N = \sum_{i=1}^{k} n_i$ es el total de observaciones.

El estadístico de contraste F se obtiene como:

$$F = \frac{MSB}{MSW}$$

Bajo la hipótesis nula H_0, el estadístico Fsigue una distribución F de Snedecor con $(k-1, N-k)$ grados de libertad.

Si el valor p (p-valor) asociado a F es menor que el nivel de significación (usualmente 0.05, aunque se pueden utilizar otros como 0.01 o 0.1), se rechaza la hipótesis nula, concluyendo que existen diferencias significativas entre las medias de los grupos.

En el contexto de selección de variables, el ANOVA se utiliza para evaluar la relevancia estadística de cada variable numérica respecto a una variable de clasificación o dependiente.

- Si una variable numérica presenta un p-valor < 0.05, se considera significativa y se mantiene en el modelo.
- Si el p-valor > 0.05, se asume que la variable no contribuye de forma significativa y se descarta para reducir la complejidad del modelo.

Para las variables categóricas, el ANOVA no es aplicable. En su lugar se utiliza la prueba Chi-cuadrado de independencia (χ^2), que permite determinar si existe una asociación significativa entre dos variables cualitativas.

El estadístico se define como:

$$\chi^2 = \sum_{i=1}^{r} \sum_{j=1}^{c} \frac{(O_{ij} - E_{ij})^2}{E_{ij}}$$

donde:

- O_{ij} es la frecuencia observada en la celda (i, j).
- E_{ij} es la frecuencia esperada bajo la hipótesis de independencia
- r y c son el número de filas y columnas de la tabla de contingencia.

El criterio de decisión es el mismo que con ANOVA: si el p-valor < 0.05, se rechaza la hipótesis nula de independencia, indicando que la variable categórica tiene un efecto significativo. Si es mayor a 0.05 no es significativa.

Seleccionadas las variables significativas, tanto cuantitativas como cualitativas, son las que se usan para definir el juego.

4. Aplicación (resultados).

En esta sección se encuentran los resultados divididos en varios bloques:

1) Lectura y visualización de los datos a través del sistema RStudio y el proceso de la obtención de los datos. Además, se explican los conjuntos de datos que se van a utilizar en los siguientes apartados.
2) Análisis descriptivo de las variables, es decir qué tipo de variable (numérica, categórica, fecha…) es cada una de ellas. También se calculan los promedios, máximos, mínimos, cuartiles de las variables numéricas y las frecuencias de las variables categóricas. Además, se han realizado algunos gráficos interesantes con los datos.
3) Categorización de las observaciones visualizadas a través de las técnicas de clasificación explicadas anteriormente y sus matrices de confusión para comprobar la precisión de las mismas. Explicando al final que modelo tiene mayor cantidad de precisión.
4) Visualización e interpretación de las tablas de las variables significativas para decidir el resultado final de los partidos en cada una de las fracciones de los datos a través del ANOVA, random forest y teoría de juegos.

4.1. Lectura y visualización de los datos.

En este estudio se ha analizado la liga española de fútbol desde la temporada 2016-2017 hasta la temporada 2024-2025 a través de diferentes análisis. Pero lo primero ha sido la obtención de los datos que se han recogido a través de varias fuentes descritas a continuación:

- **Sofascore (https://www.sofascore.com/es/)**: Página de internet y aplicación del móvil donde se han obtenido los datos referentes a la posesión, pases, duelos ganados, fueras de juego y sistemas de juego de los equipos de todo el periodo estudiado.
- **Resultados de fútbol (https://www.resultados-futbol.com/)**: Aplicación del móvil donde se han obtenido los datos de los tiros, faltas, amarillas, saques de esquina, rojas, resultado del partido, de la temporada 24-25 (hecho así, porque los he utilizado para una base de datos personal previamente).
- **Football Data (https://www.football-data.co.uk/spainm.php)**: Página de internet donde se han obtenido los datos referentes a los tiros, faltas, amarillas, saques de esquina, rojas, resultado del partido, desde la temporada 16-17 hasta la 23-24.
- **Futbolpedia**: Libro de Pedro Martín, periodista de la Cope y amante de los datos, donde se han obtenido las estadísticas anómalas o curiosas que hay en el informe.

Las herramientas utilizadas para el estudio han sido Notepad y RStudio. RStudio se ha utilizado para la parte de estadística (análisis descriptivo, técnicas de clasificación y significación de las variables sobre la variable respuesta) y Notepad con HTML para la página web que se describe en el punto 5.

Para este estudio se ha analizado la base de datos de diferentes maneras:

- Desde la temporada 2016-2017 hasta la temporada 2024-2025 al completo (3420 observaciones).
- Cada temporada (9) individualmente (es decir, 380 observaciones por temporada).
- Cada uno de los 3 equipos más laureados del fútbol español (Real Madrid Club de Fútbol, Fútbol Club Barcelona y Club Atlético de Madrid) visto tanto del punto de vista de jugar en casa como jugar fuera.
- Cada uno de los 3 sistemas de juego más repetido a lo largo del tiempo de estudio (sistema 4-3-3, sistema 4-4-2 y sistema 4-2-3-1), observando tanto los partidos que se juega de local con ese sistema como cuando se juega de visitante.
- La base de datos previamente y posteriormente al COVID.

4.2. Análisis descriptivo de las variables.

El primer paso a realizar dentro del análisis descriptivo, aparte de visualizar la base de datos, es ver si hay valores perdidos o valores atípicos. Visualizadas las bases de datos obtenemos la conclusión de que no hay NA y aunque hay valores atípicos no se eliminan en el estudio, porque los consideramos lógicos dentro de la magia que es el fútbol, donde no todo sigue un sistema fijo, sino que hay situaciones que alteran los resultados de los partidos.

Hecho esto se procede a realizar un resumen con cada conjunto de datos, donde nos muestra que tipo de variables son cada una de nuestras variables, las medias, cuartiles, medianas, máximos y mínimos.

Las variables del estudio son las siguientes:

- **Fecha**: Variable en formato fecha que muestra el día que se juega cada partido.
- **Equipo local y equipo visitante**: Variable categórica que nos indica los equipos de la liga tanto locales como visitantes.
- **% de posesión local y visitante**: Variable numérica que representa el porcentaje de tiempo que tiene la pelota cada uno de los equipos. Oscila entre 0 y 100.
- **Fueras de juego local y visitante**: Variable numérica que representa los fueras de juego de ambos equipos. Puede tomar cualquier valor entero positivo.
- **% de duelos ganados por el local y visitante**: Variable numérica que representa el porcentaje de duelos ganados por ambos equipos. Oscila entre 0 y 100.
- **Pases realizados por el equipo local y visitante**: Variable numérica que representa los pases realizados por ambos equipos. Puede tomar cualquier valor entero positivo.
- **Sistema de juego del equipo local y visitante**: Variable categórica que indica con valores del 1 al 22 de cada uno de los sistemas de juego planteados por ambos entrenadores. Para una mejor visualización se han indicado con números, pero cada sistema tiene el siguiente valor:

Tabla 1: Valores de los sistemas de juego

Sistema	Valor	Sistema	Valor
3_1_4_2	1	4_2_2_2	12
3_2_4_1	2	4_2_3_1	13
3_3_1_3	3	4_2_4	14
3_3_3_1	4	4_3_1_2	15
3_4_1_2	5	4_3_2_1	16
3_4_2_1	6	4_3_3	17
3_4_3	7	4_4_1_1	18
3_5_1_1	8	4_4_2	19
3_5_2	9	4_5_1	20
4_1_3_2	10	5_3_2	21
4_1_4_1	11	5_4_1	22

- **Tiros a puerta del equipo local y visitante**: Variable numérica que representa los tiros a puerta realizados por ambos equipos. Puede tomar cualquier valor entero positivo.
- **Faltas del equipo local y visitante**: Variable numérica que representa las faltas realizadas por ambos equipos. Puede tomar cualquier valor entero positivo.
- **Amarillas del equipo local y visitante**: Variable numérica que representa las amarillas recibidas por ambos equipos. Puede tomar cualquier valor entero positivo.
- **Rojas del equipo local y visitante**: Variable numérica que representa las rojas recibidas por ambos equipos. Puede tomar cualquier valor entero positivo.
- **Saques de esquina del equipo local y visitante**: Variable numérica que representa los saques de esquina realizados por ambos equipos. Puede tomar cualquier valor entero positivo.
- **Ganador Final**: Variable categórica, que nos muestra 3 clases. Es la variable dependiente, es decir, la variable respuesta.
    - **H**: Victoria local.
    - **D**: Empate.
    - **A**: Victoria visitante.

Obtenido el resumen se procede a realizar una tabla con las variables numéricas y alguna información de estas (medias, máximos y mínimos). Además, se ha realizado una tabla con las variables categóricas y nos muestra la frecuencia de cada una de las variables. En el informe se procede a mostrar la primera tabla y en la segunda se indica cual es la clase que mayor frecuencia tiene en cada una de las variables categóricas (para ver el resto de las frecuencias visitar la página web):

Para todos los equipos:

Tabla 2: Resumen variables numéricas para todos los equipos

Variable	$\bar{x}$	Min	Max
Posesion_L	51.4	18	82
Posesion_V	48.6	18	82
Pases_L	444.4	89	1002
Pases_V	419.4	157	993
FJ_L	2.3	0	14
FJ_V	2.1	0	14
Duelos_L	50.1	27	72
Duelos_V	49.9	28	73
Tiros_L	4.6	0	17
Tiros_V	3.6	0	15
Faltas_L	13.3	1	29
Faltas_V	13.3	0	31
Corners_L	5.3	0	26
Corners_V	4.1	0	15
Amarillas_L	2.4	0	9
Amarillas_V	2.5	0	9
Rojas_L	0.1	0	2
Rojas_V	0.1	0	2

Tabla 3: Resumen variables categóricas para todos los equipos

	Clase	Frecuencia	Porcentaje
Sistema Local	19 (4-4-2)	1005	29,4
Sistema Visitante	19 (4-4-2)	937	27,4
Ganador Final	H (Local)	1543	45,1

Para la temporada 2016-2017:

Tabla 4: Resumen variables numéricas para la temporada 16-17

Variable	$\bar{x}$	Min	Max
Posesion_L	51.6	19	81
Posesion_V	48.4	19	81
Pases_L	453.1	179	835
Pases_V	419.1	176	861
FJ_L	2.5	0	10
FJ_V	2.2	0	7
Duelos_L	50.1	27	65
Duelos_V	49.9	35	73
Tiros_L	5.0	0	14
Tiros_V	3.9	0	13
Faltas_L	14.1	1	28
Faltas_V	13.8	4	26
Corners_L	5.5	0	20
Corners_V	4.0	0	14
Amarillas_L	2.3	0	7
Amarillas_V	2.6	0	7
Rojas_L	0.1	0	2
Rojas_V	0.1	0	2

Tabla 5: Resumen variables categóricas para la temporada 16-17

	Clase	Frecuencia	Porcentaje
Sistema Local	13 (4-2-3-1)	144	37,9
Sistema Visitante	13 (4-2-3-1)	138	36,3
Ganador Final	H (Local)	181	47,6

Para la temporada 2017-2018:

Tabla 6: Resumen variables numéricas para la temporada 17-18

Variable	x̄	Min	Max
Posesion_L	51.1	24	79
Posesion_V	48.9	21	76
Pases_L	446.9	89	805
Pases_V	428.5	203	828
FJ_L	2.6	0	14
FJ_V	2.5	0	14
Duelos_L	50.2	30	67
Duelos_V	49.8	33	70
Tiros_L	4.8	0	14
Tiros_V	3.8	0	13
Faltas_L	13.7	4	29
Faltas_V	14.0	0	29
Corners_L	5.6	0	16
Corners_V	4.2	0	14
Amarillas_L	2.3	0	8
Amarillas_V	2.7	0	9
Rojas_L	0.1	0	2
Rojas_V	0.1	0	2

Tabla 7: Resumen variables categóricas para la temporada 17-18

	Clase	Frecuencia	Porcentaje
Sistema Local	19 (4-4-2)	145	38,2
Sistema Visitante	19 (4-4-2)	127	33,4
Ganador Final	H (Local)	179	47,1

Para la temporada 2018-2019:

Tabla 8: Resumen variables numéricas para la temporada 18-19

Variable	x̄	Min	Max
Posesion_L	51.2	22	82
Posesion_V	48.8	18	78
Pases_L	440.8	212	899
Pases_V	414.7	176	993
FJ_L	2.3	0	8
FJ_V	2.2	0	11
Duelos_L	50.4	36	67
Duelos_V	49.6	33	64
Tiros_L	4.8	0	15
Tiros_V	3.6	0	11
Faltas_L	13.6	1	26
Faltas_V	13.4	3	27
Corners_L	5.6	0	15
Corners_V	4.0	0	12
Amarillas_L	2.5	0	8
Amarillas_V	2.6	0	7
Rojas_L	0.1	0	1
Rojas_V	0.1	0	2

Tabla 9: Resumen variables categóricas para la temporada 18-19

	Clase	Frecuencia	Porcentaje
Sistema Local	19 (4-4-2)	127	33,4
Sistema Visitante	19 (4-4-2)	124	32,6
Ganador Final	H (Local)	168	44,2

Para la temporada 2019-2020:

Tabla 10: Resumen variables numéricas para la temporada 19-20

Variable	$\bar{x}$	Min	Max
Posesion_L	51.2	23	82
Posesion_V	48.8	18	77
Pases_L	435.0	220	1002
Pases_V	412.2	162	866
FJ_L	2.3	0	9
FJ_V	1.9	0	9
Duelos_L	50.1	35	64
Duelos_V	49.9	36	65
Tiros_L	4.3	0	17
Tiros_V	3.5	0	12
Faltas_L	13.7	4	28
Faltas_V	13.8	5	30
Corners_L	5.0	0	14
Corners_V	4.2	0	12
Amarillas_L	2.5	0	7
Amarillas_V	2.6	0	8
Rojas_L	0.1	0	2
Rojas_V	0.1	0	2

Tabla 11: Resumen variables categóricas para la temporada 19-20

	Clase	Frecuencia	Porcentaje
Sistema Local	19 (4-4-2)	156	41,0
Sistema Visitante	19 (4-4-2)	131	34,5
Ganador Final	H (Local)	174	45,8

Para la temporada 2020-2021:

Tabla 12: Resumen variables numéricas para la temporada 20-21

Variable	$\bar{x}$	Min	Max
Posesion_L	50.8	18	82
Posesion_V	49.2	18	82
Pases_L	439.9	183	938
Pases_V	425.7	167	864
FJ_L	2.1	0	9
FJ_V	1.9	0	9
Duelos_L	49.9	32	69
Duelos_V	50.1	31	68
Tiros_L	4.0	0	13
Tiros_V	3.4	0	10
Faltas_L	13.3	2	26
Faltas_V	13.2	3	30
Corners_L	4.4	0	15
Corners_V	4.3	0	14
Amarillas_L	2.2	0	8
Amarillas_V	2.3	0	7
Rojas_L	0.1	0	2
Rojas_V	0.1	0	2

Tabla 13: Resumen variables categóricas para la temporada 20-21

	Clase	Frecuencia	Porcentaje
Sistema Local	19 (4-4-2)	148	39,0
Sistema Visitante	19 (4-4-2)	142	37,4
Ganador Final	H (Local)	158	41,6

Para la temporada 2021-2022:

Tabla 14: Resumen variables numéricas para la temporada 21-22

Variable	$\bar{x}$	Min	Max
Posesion_L	51.1	22	82
Posesion_V	48.9	18	78
Pases_L	433.8	176	878
Pases_V	417.7	175	860
FJ_L	2.1	0	7
FJ_V	1.9	0	8
Duelos_L	49.4	29	64
Duelos_V	50.6	36	71
Tiros_L	4.4	0	16
Tiros_V	3.6	0	11
Faltas_L	13.6	5	25
Faltas_V	12.8	1	25
Corners_L	5.2	0	16
Corners_V	4.2	0	12
Amarillas_L	2.5	0	8
Amarillas_V	2.6	0	7
Rojas_L	0.1	0	2
Rojas_V	0.1	0	2

Tabla 15: Resumen variables categóricas para la temporada 21-22

	Clase	Frecuencia	Porcentaje
Sistema Local	19 (4-4-2)	103	27,1
Sistema Visitante	19 (4-4-2)	105	27,6
Ganador Final	H (Local)	165	43,4

Para la temporada 2022-2023:

Tabla 16: Resumen variables numéricas para la temporada 22-23

Variable	$\bar{x}$	Min	Max
Posesion_L	52.1	25	81
Posesion_V	47.9	19	75
Pases_L	444.8	186	847
Pases_V	410.8	157	740
FJ_L	2.4	0	9
FJ_V	2.0	0	9
Duelos_L	50.5	34	72
Duelos_V	49.5	28	66
Tiros_L	4.8	0	17
Tiros_V	3.6	0	15
Faltas_L	13.0	4	29
Faltas_V	12.9	3	26
Corners_L	5.5	0	19
Corners_V	4.0	0	15
Amarillas_L	2.4	0	9
Amarillas_V	2.6	0	8
Rojas_L	0.2	0	2
Rojas_V	0.2	0	2

Tabla 17: Resumen variables categóricas para la temporada 22-23

	Clase	Frecuencia	Porcentaje
Sistema Local	13 (4-2-3-1)	95	25,0
Sistema Visitante	13 (4-2-3-1)	93	24,5
Ganador Final	H (Local)	182	47,9

Para la temporada 2023-2024:

Tabla 18: Resumen variables numéricas para la temporada 23-24

Variable	x̄	Min	Max
Posesion_L	52.0	19	80
Posesion_V	48.0	20	81
Pases_L	459.4	183	811
Pases_V	425.4	185	773
FJ_L	2.4	0	10
FJ_V	2.1	0	14
Duelos_L	50.1	32	66
Duelos_V	49.9	34	68
Tiros_L	4.9	0	13
Tiros_V	3.7	0	12
Faltas_L	12.9	3	28
Faltas_V	13.0	4	31
Corners_L	5.3	0	15
Corners_V	4.0	0	15
Amarillas_L	2.3	0	8
Amarillas_V	2.4	0	8
Rojas_L	0.1	0	2
Rojas_V	0.1	0	2

Tabla 19: Resumen variables categóricas para la temporada 23-24

	Clase	Frecuencia	Porcentaje
Sistema Local	13 (4-2-3-1)	107	28,2
Sistema Visitante	13 (4-2-3-1)	110	28,9
Ganador Final	H (Local)	167	44,0

Para la temporada 2024-2025:

Tabla 20: Resumen variables numéricas para la temporada 24-25

Variable	x̄	Min	Max
Posesion_L	51.7	18	81
Posesion_V	48.3	19	82
Pases_L	446.1	156	778
Pases_V	420.1	170	793
FJ_L	2.1	0	12
FJ_V	1.7	0	11
Duelos_L	50.2	33	67
Duelos_V	49.8	33	67
Tiros_L	4.6	0	16
Tiros_V	3.6	0	11
Faltas_L	12.3	1	28
Faltas_V	12.5	3	26
Corners_L	5.4	0	26
Corners_V	4.2	0	14
Amarillas_L	2.2	0	7
Amarillas_V	2.3	0	9
Rojas_L	0.1	0	2
Rojas_V	0.1	0	2

Tabla 21: Resumen variables categóricas para la temporada 24-25

	Clase	Frecuencia	Porcentaje
Sistema Local	13 (4-2-3-1)	163	42,9
Sistema Visitante	13 (4-2-3-1)	139	36,6
Ganador Final	H (Local)	169	44,5

Para el sistema 4-2-3-1 Local:

Tabla 22: Resumen variables numéricas para el sistema 4-2-3-1 Local

Variable	$\bar{x}$	Min	Max
Posesion_L	51.6	23	81
Posesion_V	48.4	19	77
Pases_L	436.0	197	835
Pases_V	411.1	170	811
FJ_L	2.3	0	11
FJ_V	2.1	0	9
Duelos_L	50.1	32	66
Duelos_V	49.9	34	68
Tiros_L	4.7	0	17
Tiros_V	3.7	0	12
Faltas_L	13.2	1	25
Faltas_V	13.3	3	29
Corners_L	5.3	0	20
Corners_V	4.2	0	15
Amarillas_L	2.4	0	7
Amarillas_V	2.5	0	8
Rojas_L	0.1	0	2
Rojas_V	0.1	0	2

Tabla 23: Resumen variables categóricas para el sistema 4-2-3-1 Local

	Clase	Frecuencia	Porcentaje
Sistema Visitante	13 (4-2-3-1)	247	28,3
Ganador Final	H (Local)	381	43,6

Para el sistema 4-2-3-1 Visitante:

Tabla 24: Resumen variables numéricas para el sistema 4-2-3-1 Visitante

Variable	$\bar{x}$	Min	Max
Posesion_L	51.2	18	82
Posesion_V	48.8	18	82
Pases_L	436.6	89	1002
Pases_V	415.3	162	864
FJ_L	2.4	0	12
FJ_V	2.1	0	11
Duelos_L	50.3	30	66
Duelos_V	49.7	34	70
Tiros_L	4.6	0	13
Tiros_V	3.7	0	13
Faltas_L	13.4	3	28
Faltas_V	13.3	0	30
Corners_L	5.3	0	20
Corners_V	4.2	0	14
Amarillas_L	2.3	0	8
Amarillas_V	2.5	0	9
Rojas_L	0.1	0	2
Rojas_V	0.1	0	2

Tabla 25: Resumen variables categóricas para el sistema 4-2-3-1 Visitante

	Clase	Frecuencia	Porcentaje
Sistema Local	19 (4-4-2)	265	30,7
Ganador Final	H (Local)	386	44,7

Para el sistema 4-3-3 Local:

Tabla 26: Resumen variables numéricas para el sistema 4-3-3 Local

Variable	$\bar{x}$	Min	Max
Posesion_L	57.8	25	82
Posesion_V	42.2	18	75
Pases_L	526.9	217	899
Pases_V	377.5	157	794
FJ_L	2.5	0	14
FJ_V	1.9	0	14
Duelos_L	51.5	30	72
Duelos_V	48.5	28	70
Tiros_L	5.3	0	17
Tiros_V	3.4	0	11
Faltas_L	12.5	2	28
Faltas_V	13.7	1	30
Corners_L	5.8	0	17
Corners_V	3.8	0	13
Amarillas_L	2.1	0	7
Amarillas_V	2.5	0	9
Rojas_L	0.1	0	2
Rojas_V	0.1	0	2

Tabla 27: Resumen variables categóricas para el sistema 4-3-3 Local

	Clase	Frecuencia	Porcentaje
Sistema Visitante	13 (4-2-3-1)	152	25,9
Ganador Final	H (Local)	325	55,5

Para el sistema 4-3-3 Visitante:

Tabla 28: Resumen variables numéricas para el sistema 4-3-3 Visitante

Variable	$\bar{x}$	Min	Max
Posesion_L	43.7	19	81
Posesion_V	56.3	19	81
Pases_L	389.1	179	938
Pases_V	508.7	200	993
FJ_L	2.2	0	10
FJ_V	2.2	0	14
Duelos_L	48.3	27	72
Duelos_V	51.7	28	73
Tiros_L	4.3	0	13
Tiros_V	4.1	0	15
Faltas_L	14.1	1	29
Faltas_V	12.2	3	26
Corners_L	5.0	0	15
Corners_V	4.6	0	15
Amarillas_L	2.4	0	9
Amarillas_V	2.4	0	7
Rojas_L	0.1	0	2
Rojas_V	0.1	0	2

Tabla 29: Resumen variables categóricas para el sistema 4-3-3 Visitante

	Clase	Frecuencia	Porcentaje
Sistema Local	19 (4-4-2)	164	31,3
Ganador Final	H (Local)	201	38,4

Para el sistema 4-4-2 Local:

Tabla 30: Resumen variables numéricas para el sistema 4-4-2 Local

Variable	$\bar{x}$	Min	Max
Posesion_L	48.5	19	80
Posesion_V	51.5	20	81
Pases_L	408.2	89	876
Pases_V	437.5	200	993
FJ_L	2.3	0	12
FJ_V	2.0	0	11
Duelos_L	49.9	27	67
Duelos_V	50.1	33	73
Tiros_L	4.4	0	14
Tiros_V	3.6	0	12
Faltas_L	13.5	1	28
Faltas_V	13.4	3	30
Corners_L	5.2	0	26
Corners_V	4.1	0	14
Amarillas_L	2.4	0	9
Amarillas_V	2.6	0	9
Rojas_L	0.1	0	2
Rojas_V	0.1	0	2

Tabla 31: Resumen variables categóricas para el sistema 4-4-2 Local

	Clase	Frecuencia	Porcentaje
Sistema Visitante	19 (4-4-2)	284	28,3
Ganador Final	H (Local)	443	44,1

Para el sistema 4-4-2 Visitante:

Tabla 32: Resumen variables numéricas para el sistema 4-4-2 Visitante

Variable	$\bar{x}$	Min	Max
Posesion_L	54.1	26	80
Posesion_V	45.9	20	74
Pases_L	460.0	197	899
Pases_V	385.0	157	778
FJ_L	2.2	0	10
FJ_V	2.1	0	14
Duelos_L	50.5	33	68
Duelos_V	49.5	32	67
Tiros_L	4.6	0	15
Tiros_V	3.5	0	11
Faltas_L	13.1	1	29
Faltas_V	13.4	4	30
Corners_L	5.3	0	16
Corners_V	4.1	0	15
Amarillas_L	2.4	0	8
Amarillas_V	2.5	0	9
Rojas_L	0.1	0	2
Rojas_V	0.1	0	2

Tabla 33: Resumen variables categóricas para el sistema 4-4-2 Visitante

	Clase	Frecuencia	Porcentaje
Sistema Local	19 (4-4-2)	284	30,3
Ganador Final	H (Local)	412	44,0

Para el Club Atlético de Madrid Local:

Tabla 34: Resumen variables numéricas para el Club Atlético de Madrid Local

Variable	x̄	Min	Max
Posesion_L	50.6	29	73
Posesion_V	49.4	27	71
Pases_L	486.9	279	764
Pases_V	474.0	243	737
FJ_L	2.6	0	10
FJ_V	0.7	0	5
Duelos_L	52.0	37	69
Duelos_V	48.0	31	63
Tiros_L	5.1	1	12
Tiros_V	2.8	0	8
Faltas_L	12.5	1	28
Faltas_V	12.6	4	26
Corners_L	5.9	0	20
Corners_V	3.7	0	12
Amarillas_L	2.2	0	7
Amarillas_V	2.4	0	8
Rojas_L	0.1	0	1
Rojas_V	0.1	0	2

Tabla 35: Resumen variables categóricas para el Club Atlético de Madrid Local

	Clase	Frecuencia	Porcentaje
Sistema Local	19 (4-4-2)	93	54,4
Sistema Visitante	13 (4-2-3-1)	36	21,1
Ganador Final	H (Local)	115	67,2

Para el Club Atlético de Madrid Visitante:

Tabla 36: Resumen variables numéricas para el Club Atlético de Madrid Visitante

Variable	x̄	Min	Max
Posesion_L	50.6	27	73
Posesion_V	49.4	27	73
Pases_L	474.6	210	774
Pases_V	460.5	245	773
FJ_L	1.1	0	5
FJ_V	2.5	0	8
Duelos_L	48.2	37	62
Duelos_V	51.8	38	63
Tiros_L	3.8	0	13
Tiros_V	3.9	0	11
Faltas_L	13.2	6	24
Faltas_V	12.8	3	26
Corners_L	5.0	1	16
Corners_V	4.2	0	11
Amarillas_L	2.3	0	6
Amarillas_V	2.6	0	7
Rojas_L	0.1	0	1
Rojas_V	0.1	0	1

Tabla 37: Resumen variables categóricas para el Club Atlético de Madrid Visitante

	Clase	Frecuencia	Porcentaje
Sistema Local	19 (4-4-2)	95	55,6
Sistema Visitante	13 (4-2-3-1)	46	26,9
Ganador Final	A (Visitante)	78	45,6

Para el Fútbol Club Barcelona Local:

Tabla 38: Resumen variables numéricas para el Fútbol Club Barcelona Local

Variable	$\bar{x}$	Min	Max
Posesion_L	66.4	48	82
Posesion_V	33.6	18	52
Pases_L	666.3	458	1002
Pases_V	332.4	170	554
FJ_L	2.6	0	10
FJ_V	2.3	0	11
Duelos_L	52.8	38	67
Duelos_V	47.2	33	62
Tiros_L	6.7	2	17
Tiros_V	2.8	0	9
Faltas_L	11.3	3	25
Faltas_V	13.2	1	30
Corners_L	6.8	1	14
Corners_V	3.1	0	12
Amarillas_L	1.8	0	6
Amarillas_V	2.4	0	7
Rojas_L	0.1	0	2
Rojas_V	0.1	0	2

Tabla 39: Resumen variables categóricas para el Fútbol Club Barcelona Local

	Clase	Frecuencia	Porcentaje
Sistema Local	17 (4-3-3)	105	61,4
Sistema Visitante	13 (4-2-3-1)	46	26,9
Ganador Final	H (Local)	124	72,5

Para el Fútbol Club Barcelona Visitante:

Tabla 40: Resumen variables numéricas para el Fútbol Club Barcelona Visitante

Variable	$\bar{x}$	Min	Max
Posesion_L	35.3	18	56
Posesion_V	64.7	44	82
Pases_L	331.9	156	582
Pases_V	622.2	391	993
FJ_L	2.8	0	12
FJ_V	2.6	0	14
Duelos_L	45.5	27	60
Duelos_V	54.5	40	73
Tiros_L	4.0	0	14
Tiros_V	5.3	0	13
Faltas_L	14.5	4	28
Faltas_V	10.9	3	24
Corners_L	4.5	0	15
Corners_V	5.0	0	14
Amarillas_L	2.4	0	6
Amarillas_V	2.3	0	7
Rojas_L	0.2	0	1
Rojas_V	0.1	0	2

Tabla 41: Resumen variables categóricas para el Fútbol Club Barcelona Visitante

	Clase	Frecuencia	Porcentaje
Sistema Local	19 (4-4-2)	55	32,2
Sistema Visitante	17 (4-3-3)	105	61,4
Ganador Final	A (Visitante)	99	57,9

Para el Real Madrid Club de Fútbol Local:

Tabla 42: Resumen variables numéricas para el Real Madrid Club de Fútbol Local

Variable	$\bar{x}$	Min	Max
Posesion_L	60.9	31	82
Posesion_V	39.1	18	69
Pases_L	602.3	89	870
Pases_V	387.9	175	707
FJ_L	2.7	0	12
FJ_V	1.4	0	5
Duelos_L	53.9	41	72
Duelos_V	46.1	28	59
Tiros_L	6.7	2	17
Tiros_V	3.3	0	11
Faltas_L	11.0	1	23
Faltas_V	13.4	5	31
Corners_L	6.7	0	26
Corners_V	3.5	0	11
Amarillas_L	1.7	0	5
Amarillas_V	2.2	0	7
Rojas_L	0.1	0	1
Rojas_V	0.1	0	1

Tabla 43: Resumen variables categóricas para el Real Madrid Club de Fútbol Local

	Clase	Frecuencia	Porcentaje
Sistema Local	17 (4-3-3)	104	60,8
Sistema Visitante	19 (4-4-2)	43	25,1
Ganador Final	H (Local)	120	70,2

Para el Real Madrid Club de Fútbol Visitante:

Tabla 44: Resumen variables numéricas para el Real Madrid Club de Fútbol Visitante

Variable	$\bar{x}$	Min	Max
Posesion_L	41.7	23	73
Posesion_V	58.3	27	77
Pases_L	405.0	225	712
Pases_V	570.4	257	793
FJ_L	1.6	0	6
FJ_V	2.4	0	11
Duelos_L	46.0	29	60
Duelos_V	54.0	40	71
Tiros_L	3.9	0	12
Tiros_V	5.5	1	15
Faltas_L	14.2	1	29
Faltas_V	11.2	3	25
Corners_L	4.5	0	12
Corners_V	5.0	0	15
Amarillas_L	2.6	0	7
Amarillas_V	2.1	0	6
Rojas_L	0.1	0	2
Rojas_V	0.1	0	1

Tabla 45: Resumen variables categóricas para el Real Madrid Club de Fútbol Visitante

	Clase	Frecuencia	Porcentaje
Sistema Local	19 (4-4-2)	50	29,2
Sistema Visitante	17 (4-3-3)	96	56,1
Ganador Final	A (Visitante)	93	54,4

Para los partidos previos al COVID:

Tabla 46: Resumen variables numéricas para los partidos previos al COVID

Variable	$\bar{x}$	Min	Max
Posesion_L	51.2	19	82
Posesion_V	48.8	18	81
Pases_L	443.2	89	1002
Pases_V	418.7	176	993
FJ_L	2.4	0	14
FJ_V	2.2	0	14
Duelos_L	50.2	27	67
Duelos_V	49.8	33	73
Tiros_L	4.8	0	17
Tiros_V	3.7	0	13
Faltas_L	13.8	1	29
Faltas_V	13.8	0	30
Corners_L	5.5	0	20
Corners_V	4.1	0	14
Amarillas_L	2.4	0	8
Amarillas_V	2.7	0	9
Rojas_L	0.1	0	2
Rojas_V	0.1	0	2

Tabla 47: Resumen variables categóricas para los partidos previos al COVID

	Clase	Frecuencia	Porcentaje
Sistema Local	19 (4-4-2)	459	32,5
Sistema Visitante	19 (4-4-2)	419	29,7
Ganador Final	H (Local)	657	46,6

Para los partidos posteriores al COVID:

Tabla 48: Resumen variables numéricas para los partidos posteriores al COVID

Variable	$\bar{x}$	Min	Max
Posesion_L	51.6	18	82
Posesion_V	48.4	18	82
Pases_L	445.3	156	938
Pases_V	419.8	157	864
FJ_L	2.2	0	12
FJ_V	1.9	0	14
Duelos_L	50.0	29	72
Duelos_V	50.0	28	71
Tiros_L	4.5	0	17
Tiros_V	3.6	0	15
Faltas_L	13.0	1	29
Faltas_V	12.9	1	31
Corners_L	5.1	0	26
Corners_V	4.1	0	15
Amarillas_L	2.3	0	9
Amarillas_V	2.4	0	9
Rojas_L	0.1	0	2
Rojas_V	0.1	0	2

Tabla 49: Resumen variables categóricas para los partidos posteriores al COVID

	Clase	Frecuencia	Porcentaje
Sistema Local	19 (4-4-2)	546	27,2
Sistema Visitante	19 (4-4-2)	518	25,8
Ganador Final	H (Local)	886	44,1

Realizado un análisis descriptivo podemos considerar que en general los datos son similares al jugar en casa o al jugar fuera, siendo un poco superiores las medias al jugar de local, sobre todo en los apartados de ataque.

Existen datos anómalos que no se han eliminado debido a que forman parte de la magia del fútbol, ya que pueden pasar, aunque sean raros. Algunos de ellos son las 0 faltas que hizo la Real Sociedad en un partido en la Rosaleda, donde se llevó la victoria el equipo malagueño, o cuando el Fútbol Club Barcelona superó la barrera de los 1000 pases, en un partido donde el equipo de Quique Setién, que debutaba en el Camp Nou, consiguió la victoria con un gol del de siempre (Leo Messi) en los últimos minutos ante el Granada.

Como sistema más repetido podemos afirmar que el 4-4-2 ha sido clave para los entrenadores a lo largo del periodo estudiado, pero curiosamente Real Madrid y Barcelona, han utilizado mucho más el sistema 4-3-3. Sistema que al utilizarlo es el que más victorias proporciona, tanto como local como visitante.

Las victorias locales han sido lo más frecuente a lo largo del periodo estudiado, a excepción de las bases de datos de los grandes equipos cuando son visitantes, que predominan lógicamente las victorias visitantes, aunque no tanto como imagina el lector de este informe (ninguno supera el 60% de victorias, y el Atlético de Madrid no supera ni el 50%).

Por último, como análisis descriptivo se han realizado varios gráficos:

Respecto a la base de datos completa:

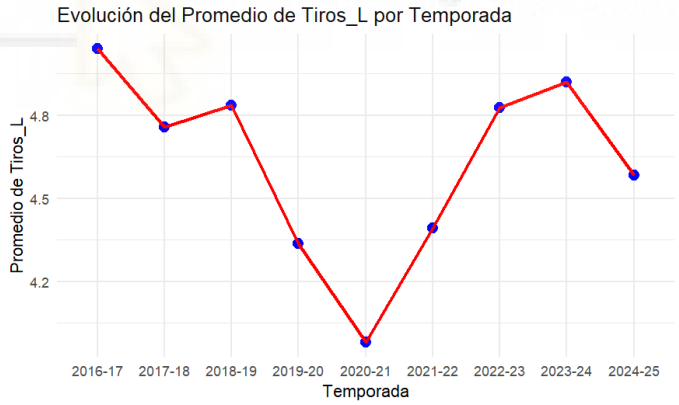

***Figura 16**: Evolución del promedio de tiros del equipo local. Fuente: Elaboración propia.*

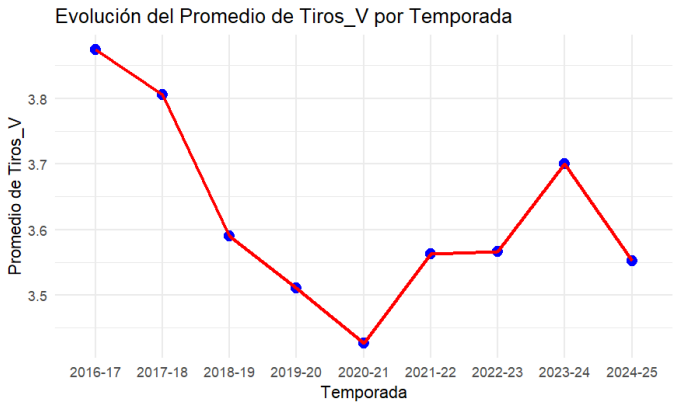

***Figura 17**: Evolución del promedio de tiros del equipo visitante. Fuente: Elaboración propia.*

En las figuras anteriores se muestra los tiros en promedio de los equipos. En la figura 16 de los equipos locales y en la figura 17 de los equipos visitantes. Estos gráficos son una evolución de cómo ha ido cambiando a lo largo de las temporadas.

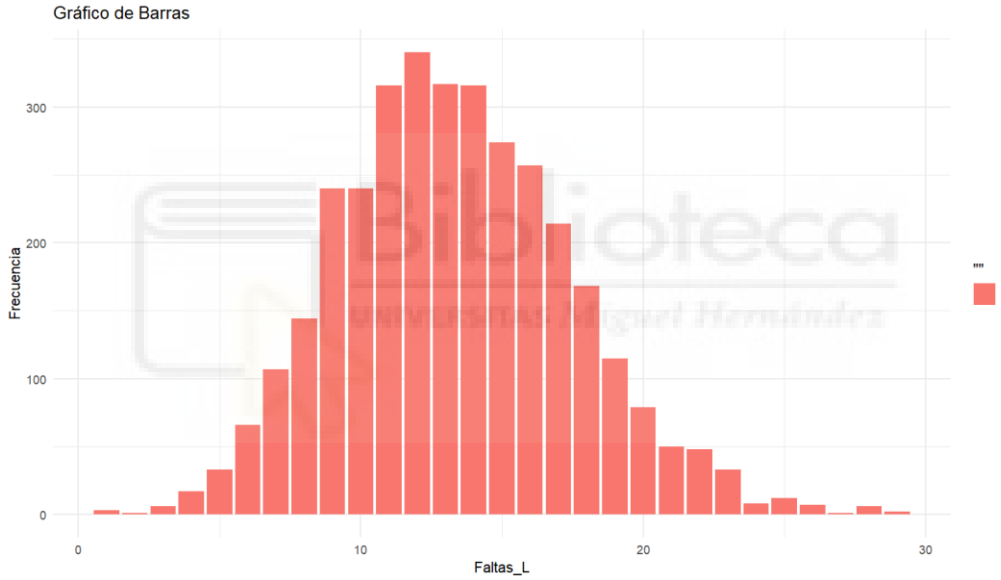

***Figura 18**: Frecuencia de faltas del equipo local. Fuente: Elaboración propia.*

En la figura 18 nos muestra en el eje X las faltas realizadas por los equipos locales y en el eje Y la frecuencia con la que ha sucedido esas faltas. Por lo tanto, la figura nos presenta un gráfico de barras donde el valor más alto es el número de faltas pitadas más recurrente a un equipo local.

Respecto al Real Madrid Local

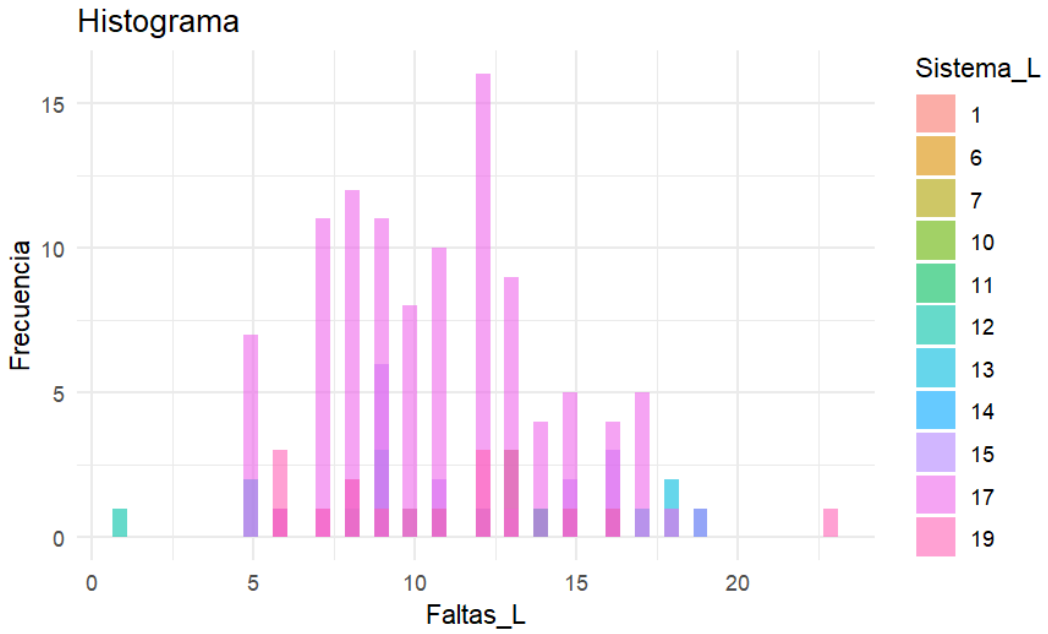

***Figura 19**: Frecuencia de faltas del Real Madrid al ser local por sistema. Fuente: Elaboración propia.*

En la figura 19 se muestra en el eje X el número de faltas del equipo local, que es el Real Madrid (ya que, se ha realizado con esa base de datos) y en el eje Y la frecuencia de cada número de faltas. En este caso cada uno de los colores nos muestra el número de veces que ha jugado con cada sistema de juego el equipo blanco, siendo un acumulado el valor. Por ejemplo, el Real Madrid el día que cometió una falta jugando de local jugó con el sistema 11, que como se ve en la tabla 1 es el sistema 4-1-4-1.

4.3. Modelización y técnicas aplicadas.

En los filtrados del estudio comentados anteriormente se ha realizado tanto la parte de clasificación de individuos (punto 4.4) como el objetivo principal de comprobar que variables son más significativas dentro de cada conjunto de datos (punto 4.5).

Además, se ha realizado un estudio comparando equipos y sistemas de juego donde se han visto las variables significativas de cada uno de los filtrados. Aunque este no es el objetivo de este estudio se explica como acceder a él en la página web (punto 5).

4.4. Técnicas de clasificación. Resultados.

En este apartado se muestran los resultados de las matrices de confusión de cada una de las técnicas en cada conjunto de la base de datos. El valor se refiere a la precisión del modelo, es decir el porcentaje de aciertos de este.

La técnica utilizada para realizar las clasificaciones en este estudio es la validación cruzada, con el objetivo de minimizar el error de estimación del modelo. Dado que la base de datos no es muy grande, se optó por un procedimiento de validación cruzada de tres

grupos, dividiendo el conjunto de datos en tres subconjuntos de igual tamaño. En cada iteración, dos tercios de los datos se emplean para el entrenamiento y el tercio restante para la validación, rotando los grupos hasta que cada uno haya sido utilizado como conjunto de validación una vez. Finalmente, se promedia el error obtenido en las tres iteraciones para obtener la probabilidad de acierto del modelo.

La extracción de las observaciones se hace de forma aleatoria, pero con el objetivo de que se obtenga la misma respuesta en cada ejecución del sistema, R tiene un comando (*set.seed*) que nos permite fijar una semilla, que hace que cada vez se ejecute tenga el mismo resultado para todos los usuarios que lo ejecuten. Esta semilla en este estudio es *set.seed(2)* pero podría ser diferente y al serlo cambiaria todo el modelo. Si no se fija una semilla cada vez que se ejecuta el problema la respuesta sería diferente.

Es importante decir que, aunque se tiene constancia del problema conocido como "maldición de la dimensionalidad", en el KNN se ha seguido con las 21 variables, ya que al intentar reducir dimensionalidad con componentes principales nos percatamos de un problema importante, no se puede hacer componentes principales en algunas fracciones de la base de datos, ya que la variabilidad por ejemplo en la variable de las "rojas" es 0, por lo tanto, se debería de eliminar la variable y es algo que no se quiere hacer, ya que una roja cambia mucho el resultado de un partido. Así que en definitiva se acepta este problema, aunque la clasificación del modelo sea peor.

Parámetros de las técnicas de clasificación.

- En el **KNN**, el parámetro k, es decir el número de vecinos se ha asignado a través de la regla de la raíz cuadrada del número de observaciones. Además, con el fin de evaluar el resultado, se probaron también los valores $k+1$ y $k-1$, comparando los porcentajes de acierto obtenidos con cada uno de ellos. Finalmente, se mantuvo aquel que daba una mayor precisión media.
- En el **SVM**, se ha utilizado un parámetro de penalización $C = 10$, se ha seleccionado debido a que este valor controla el equilibrio entre el tamaño del margen y el número de errores permitidos en el conjunto de entrenamiento.
- En el **Random Forest**, se han utilizado los parámetros por defecto. Ya que, se ha fijado $mtry = 4$, que representa el número de variables seleccionadas aleatoriamente en cada división del árbol. Valor seleccionado por ser la raíz del número de variables predictoras, que garantiza la diversidad entre los árboles y evita la correlación excesiva entre ellos. El número de árboles ($ntree$) se mantuvo en su valor por defecto de 500, suficiente para estabilizar el error sin incrementar el coste computacional. Además, se activó la opción importance = TRUE para estimar la importancia de las variables y determinar cuáles influyen más en la predicción del resultado final del partido.

También hay que destacar que como técnica con mayor precisión siempre es el análisis discriminante, con un promedio del 60% de precisión. Aquí se muestran los resultados de algunas de las tablas (para visualizar todas hay que ver la página web):

Todos los equipos

Tabla 50: Orden de matrices de confusión para todos los equipos

Modelo	Precisión
Análisis discriminante	61.28655
SVM Lineal	58.71345
Bosques Aleatorios	57.19298
Árboles de Clasificación	53.15789
SVM Polinomial	51.90058
SVM Radial	45.11696
KNN	40.97953

Temporada 2022-2023

Tabla 51: Orden de matrices de confusión para la temporada 22-23

Modelo	Precisión
Análisis discriminante	66.58230
Bosques Aleatorios	55.66075
SVM Lineal	52.35908
SVM Polinomial	49.73337
SVM Radial	47.89505
KNN	45.39641
Árboles de Clasificación	43.94659

Temporada 2023-2024

Tabla 52: Orden de matrices de confusión para la temporada 23-24

Modelo	Precisión
Análisis discriminante	56.58980
Bosques Aleatorios	45.92134
SVM Polinomial	44.74961
SVM Radial	43.95388
SVM Lineal	43.54872
Árboles de Clasificación	38.03170
KNN	38.02025

Temporada 2024-2025

Tabla 53: Orden de matrices de confusión para la temporada 24-25

Modelo	Precisión
Análisis discriminante	61.05591
Bosques Aleatorios	47.63467
SVM Polinomial	46.18381
SVM Lineal	45.65783
SVM Radial	44.47465
Árboles de Clasificación	42.89464
KNN	37.49844

Sistema 4-3-3 Local

Tabla 54: Orden de matrices de confusión para el sistema 4-3-3 Local

Modelo	Precisión
Análisis discriminante	68.94091
SVM Lineal	60.49271
Bosques Aleatorios	60.32461
SVM Polinomial	59.38794
SVM Radial	55.46178
KNN	53.84003
Árboles de Clasificación	52.47470

Sistema 4-3-3 Visitante

Tabla 55: Orden de matrices de confusión para el sistema 4-3-3 Visitante

Modelo	Precisión
Análisis discriminante	63.35906
Bosques Aleatorios	55.24740
SVM Lineal	50.28790
SVM Polinomial	49.13191
Árboles de Clasificación	46.46798
KNN	43.60755
SVM Radial	36.15982

Club Atlético de Madrid Local

Tabla 56: Orden de matrices de confusión para el Club Atlético de Madrid Local

Modelo	Precisión
Análisis discriminante	77.77778
Bosques Aleatorios	68.71345
SVM Radial	67.25146
SVM Polinomial	66.95906
KNN	66.37427
Árboles de Clasificación	59.64912
SVM Lineal	50.58480

Club Atlético de Madrid Visitante

Tabla 54: Orden de matrices de confusión para el Club Atlético de Madrid Visitante

Modelo	Precisión
Análisis discriminante	61.98830
SVM Polinomial	44.15205
Bosques Aleatorios	42.69006
SVM Lineal	42.10526
Árboles de Clasificación	38.59649
KNN	38.01170
SVM Radial	38.01170

Fútbol Club Barcelona Local

Tabla 55: Orden de matrices de confusión para el Fútbol Club Barcelona Local

Modelo	Precisión
Análisis discriminante	81.87135
Bosques Aleatorios	73.68421
SVM Radial	72.51462
KNN	72.51462
SVM Polinomial	71.63743
Árboles de Clasificación	62.86550
SVM Lineal	56.14035

Fútbol Club Barcelona Visitante

Tabla 56: Orden de matrices de confusión para el Fútbol Club Barcelona Visitante

Modelo	Precisión
Análisis discriminante	69.00585
SVM Polinomial	58.18713
SVM Radial	57.89474
Bosques Aleatorios	57.89474
SVM Lineal	56.43275
KNN	53.50877
Árboles de Clasificación	49.41520

Real Madrid Club de Fútbol Local

Tabla 60: Orden de matrices de confusión para el Real Madrid Club de Fútbol Local

Modelo	Precisión
Análisis discriminante	80.70175
SVM Radial	70.17544
KNN	69.59064
Bosques Aleatorios	69.29825
SVM Polinomial	67.54386
SVM Lineal	64.32749
Árboles de Clasificación	58.18713

Real Madrid Club de Fútbol Visitante

Tabla 61: Orden de matrices de confusión para el Real Madrid Club de Fútbol Visitante

Modelo	Precisión
Análisis discriminante	64.91228
SVM Radial	54.38596
SVM Polinomial	53.50877
KNN	51.75439
Bosques Aleatorios	51.46199
SVM Lineal	45.02924
Árboles de Clasificación	43.85965

Previo al COVID

Tabla 57: Orden de matrices de confusión para los partidos previos al COVID

Modelo	Precisión
Análisis discriminante	64.25532
Bosques Aleatorios	59.07801
SVM Lineal	58.65248
Árboles de Clasificación	53.04965
SVM Polinomial	51.17021
SVM Radial	46.59574
KNN	45.14184

Posterior al COVID

Tabla 58: Orden de matrices de confusión para los partidos posteriores al COVID

Modelo	Precisión
Análisis discriminante	59.85075
SVM Lineal	55.87065
Bosques Aleatorios	55.14925
Árboles de Clasificación	51.81592
SVM Polinomial	49.52736
SVM Radial	44.07960
KNN	43.08458

La técnica con mayor cantidad de aciertos es el análisis discriminante, que supera en casi todos los conjuntos el 60% de precisión. En segunda instancia suelen ser los bosques aleatorios y los SVM los siguientes en tener mayor precisión y casi siempre la técnica con menor acierto es el KNN que no supera muchas veces el 50% de precisión.

4.5. Variables significativas. Resultados e interpretación.

En este apartado se muestra una tabla para varios de los conjuntos descritos anteriormente (para visualizar todos hay que ver la página web) donde se muestra el orden de las variables de más significativa a menos significativa, considerando las 10 más significativas (si las hay). Para evaluar la variable como significativa se ha estudiado con un p-valor de 0.05, es decir un error de un 5% como máximo. En los filtrados realizados a los equipos (Real Madrid Club de Fútbol, Fútbol Club Barcelona y Club Atlético de Madrid) se ha seleccionado un p-valor de 0.1 debido a que si se utiliza un p-valor de 0.05 no hay variables significativas suficientes.

Otra cosa importante es la técnica de validación utilizada, en este caso es la simple, que consiste en dividir en dos subconjuntos los datos, siendo uno de ellos de entrenamiento y otro usado para validar el modelo. Se usa este tipo de validación ya que al tener que

ejecutar la teoría de juegos principalmente, es muy costosa computacionalmente, ya que, si usásemos la validación cruzada, tendríamos 3 iteraciones.

El indicador de medida precisión usado es Mean Decrease Accuracy (Disminución Media de la Precisión), ya que queremos saber qué variables no tenemos que eliminar del estudio, porque si las eliminamos perderíamos precisión en el mismo. Por lo tanto, esto nos muestra las variables que son más significativas y el orden de importancia de las mismas.

Para todo el conjunto de datos:

Tabla 59: Top 10 variables significativas para todos los equipos

ANOVA	Valor	Random forest	Valor	Teoría de juegos	Valor
Tiros Visitante	Negativo	Tiros Visitante	53.602283	Tiros Visitante	0.10811113
Tiros Local	Positivo	Tiros Local	46.480661	Tiros Local	0.10309512
Rojas Local	Negativo	Pases Local	14.995417	Rojas Local	0.06799638
Duelos Visitante	Negativo	Rojas Local	14.486502	Pases Visitante	0.04644888
Duelos Local	Positivo	Posesión Visitante	12.141449	Amarillas Local	0.04547364
Amarillas Local	Positivo	Posesión Local	11.663638	Duelos Local	0.04243131
Pases Local	Negativo	Corners Local	10.524467	Pases Local	0.04232688
Sistema Local	Indeterminado	Duelos Local	9.199626	Duelos Visitante	0.04215400
Pases Visitante	Negativo	Pases Visitante	8.238923	Sistema Visitante	0.03726968
Sistema Visitante	Indeterminado	Duelos Visitante	7.271863	Sistema Local	0.03633041

Dados los resultados podemos concluir que entre las temporadas 16-17 y la 24-25, es decir el conjunto de datos completo, las variables más significativas para decidir el resultado de un partido son los tiros a puerta, tanto visitante en primera instancia como los del local en segunda en las 3 métricas utilizadas. El tercer lugar de los rankings es diferente según la métrica utilizada, en el caso de ANOVA y teoría de juegos las rojas mostradas al equipo local son claves en el resultado final, suponiendo casi un 7% de pérdida de precisión en la teoría de juegos. En el random forest la tercera variable que más importancia le da nuestro sistema, es los pases del equipo local, con casi un 15% de pérdida de precisión en su caso, variable que tiene menos importancia en ANOVA y en la teoría de juegos, siendo en esta última métrica la décima de mayor relevancia significativa.

Para la temporada 2022-2023:

Tabla 65: Top 10 variables significativas para la temporada 22-23

ANOVA	Valor	Random forest	Valor	Teoría de juegos	Valor
Tiros Visitante	Negativo	Tiros Visitante	18.801587	Rojas Local	0.11315789
Tiros Local	Positivo	Tiros Local	10.487954	Tiros Local	0.09824561
Rojas Local	Negativo	Pases Local	9.824195	Pases Visitante	0.09473684
Rojas Visitante	Positivo	Rojas Local	7.474381	Rojas Visitante	0.09078947
Pases Visitante	Negativo	Rojas Visitante	5.246945	Tiros Visitante	0.05570175
Sistema Visitante	Indeterminado	Posesión Visitante	3.049486	Sistema Visitante	0.04736842
		Posesión Local	2.945834		
		Sistema Visitante	2.667477		
		Amarillas Visitante	2.346360		
		Duelos Visitante	1.828755		

En esta temporada tanto los tiros como las rojas, ya fuese del local o visitante, tuvieron una gran importancia en el resultado final (con una pérdida aproximadamente de un 40% en el random forest si se eliminasen esas variables y con un 35% en la teoría de juegos). Hubo también otras variables significativas, como los pases (con una pérdida del 9% de precisión) del local en el random forest o como los pases del visitante en la teoría de juegos (con un 9% también).

Para la temporada 2023-2024:

Tabla 66: Top 10 variables significativas para la temporada 23-24

ANOVA	Valor	Random forest	Valor	Teoría de juegos	Valor
Tiros Local	Positivo	Tiros Local	13.131601	Corners Local	0.11578947
Tiros Visitante	Negativo	Tiros Visitante	12.506174	Rojas Local	0.11140351
Rojas Local	Negativo	Corners Local	4.790821	Tiros Visitante	0.10153509
Corners Local	Negativo	Rojas Local	4.589814	Tiros Local	0.09495614
Corners Visitante	Positivo	Amarillas Local	4.130113	Corners Visitante	0.06315789
		Posesión Visitante	3.034015		
		Sistema Visitante	2.614165		
		Pases Visitante	2.292287		
		Corners Visitante	2.082839		
		Posesión Local	1.910380		

Durante este año podemos afirmar que sacar de esquina era importante (aunque después sea para pasarlo en corto), tanto jugando como local como jugando de visitante, ya que ambas forman parte de las 5 variables más significativas en el ANOVA y en la teoría de juegos. Como era de esperar, los tiros nos llevan a descubrir los resultados finales de los partidos siendo de nuevo las dos que copan el ranking del random forest y en el ANOVA, y aunque no en primer orden, también en la teoría de juegos es importante (con una reducción al eliminarlas de un 19%).

Para la temporada 2024-2025:

Tabla 67: Top 10 variables significativas para la temporada 24-25

ANOVA	Valor	Random forest	Valor	Teoría de juegos	Valor
Tiros Visitante	Negativo	Tiros Visitante	15.326963	Tiros Local	0.12105263
Tiros Local	Positivo	Tiros Local	10.101763	Tiros Visitante	0.11425439
Rojas Local	Negativo	Rojas Local	5.482480	Sistema Visitante	0.09627193
Sistema Visitante	Indeterminado	Sistema Visitante	5.187184	Rojas Local	0.06666667
Pases Local	Negativo	Posesión Local	3.134958	Fueras de juego Local	0.03815789
Fueras de juego Local	Negativo	Fueras de juego Local	2.329178	Pases Local	0.03728070
		Faltas Local	1.837024		
		Posesión Visitante	1.757781		
		Duelos Local	1.379770		
		Sistema Local	1.376891		

Llegamos a la última temporada del estudio y afirmamos que, si nos fijamos en temporada a temporada, todos los equipos y sistemas en conjunto, tirar es la clave del éxito. También es importante que no te expulsen a un jugador, por lo menos al equipo local, porque si lo hacen los datos demuestran que es algo significativo en el resultado final. En menor escala, pero son variables significativas, el sistema de juego del visitante y los fueras de juego local (importante para los pizarristas). Respecto a las variables locales y visitantes hay una diferencia significativa (menor que en antaño) en cuales tienen mejor ranking (sumando todas las variables de los rankings 91 sobre 54).

Para el sistema 4-3-3 como local:

Tabla 68: Top 10 variables significativas para el sistema 4-3-3 Local

ANOVA	Valor	Random forest	Valor	Teoría de juegos	Valor
Tiros Visitante	Negativo	Tiros Visitante	22.819598	Tiros Visitante	0.13561932
Tiros Local	Positivo	Tiros Local	14.724611	Tiros Local	0.10875051
Corners Local	Positivo	Corners Local	9.732132	Corners Local	0.08188509
Duelos Visitante	Negativo	Pases Local	7.061739	Rojas Local	0.07798128
Duelos Local	Positivo	Pases Visitante	3.664812	Corners Visitante	0.06005630
Pases Local	Positivo	Posesión Visitante	3.633001	Duelos Local	0.05823158
Rojas Local	Positivo	Rojas Local	3.396180	Amarillas Local	0.05724800
Amarillas Local	Negativo	Rojas Visitante	3.386117	Duelos Visitante	0.05650522
Corners Visitante	Negativo	Posesión Local	2.995953	Posesión Local	0.05367996
Posesión Local	Positivo	Faltas Local	2.900437	Pases Local	0.03653846

En el sistema 4-3-3 cuando se juega como local, existe un patrón dentro de los 3 primeros puestos de cada ranking. La primera variable son los tiros a puerta del equipo visitante (que por ejemplo supone una pérdida de precisión del 13% en la teoría de juegos). La segunda variable con mayor incidencia son los tiros a puerta de los equipos caseros (con una incidencia del 14% en la clasificación del random forest). En tercer lugar, en todos estos rankings están los saques de esquina del equipo local (teniendo una significación positiva en el ANOVA). Si nos fijamos entre la significación de las variables locales o visitantes, claramente predomina las locales con casi el doble de significación.

Para el sistema 4-3-3 como visitante:

Tabla 69: Top 10 variables significativas para el sistema 4-3-3 Visitante

ANOVA	Valor	Random forest	Valor	Teoría de juegos	Valor
Tiros Visitante	Negativo	Tiros Local	26.621003	Tiros Visitante	0.12230915
Tiros Local	Positivo	Tiros Visitante	23.459389	Rojas Local	0.07140967
Pases Visitante	Negativo	Rojas Local	7.996528	Tiros Local	0.06947468
Rojas Local	Positivo	Pases Visitante	4.727654	Pases Visitante	0.06013605
Duelos Local	Positivo	Sistema Local	3.045309	Duelos Local	0.05837491
Duelos Visitante	Negativo	Rojas Visitante	2.948551	Duelos Visitante	0.05753590
Amarillas Local	Positivo	Corners Local	2.842195	Rojas Visitante	0.05320106
Posesión Local	Positivo	Posesión Local	1.961895	Posesión Visitante	0.04162509
Posesión Visitante	Negativo	Pases Local	1.665379	Posesión Local	0.03915344
Rojas Visitante	Positivo	Posesión Visitante	1.213652	Amarillas Local	0.03630385

Cuando el equipo visitante juega con el sistema 4-3-3, podemos decir que la precisión disminuye en un 50% al eliminar las variables de los tiros a puerta (locales y visitantes) en el random forest y un 19% en la teoría de juegos. Otras de las variables importantes dentro de estos rankings tienen importancia los pases dados por el equipo visitante y las rojas mostrada a los equipos locales. Si nos centramos entre que variables tienen más significancia si las locales o las visitantes, podemos concluir que hay una igualdad entre ambas, pero hay una diferencia positiva a favor de las locales.

Para el Club Atlético de Madrid como local:

Tabla 70: Top 10 variables significativas para el Club Atlético de Madrid Local

ANOVA	Valor	Random forest	Valor	Teoría de juegos	Valor
Tiros Visitante	Negativo	Tiros Local	13.004247	Corners Local	0.16911765
Tiros Local	Positivo	Amarillas Visitante	10.625687	Tiros Visitante	0.13088235
Amarillas Visitante	Positivo	Tiros Visitante	8.387176	Faltas Visitante	0.10392157
Amarillas Local	Positivo	Faltas Visitante	6.464422	Tiros Local	0.10049020
Faltas Visitante	Negativo	Amarillas Local	5.588800	Amarillas Local	0.08529412
Corners Local	Positivo	Sistema Local	4.478577	Amarillas Visitante	0.05735294
		Pases Visitante	2.814848		
		Rojas Local	2.263530		
		Posesión Visitante	2.080920		
		Faltas Local	1.400010		

Al analizar los partidos del Atlético de Madrid como local podemos concluir con que hay una disparidad en los 3 rankings, siendo la única variable coincidente en los 3 primeros puestos los tiros a puerta por parte del visitante (con una disminución de precisión de un 8% en el random forest y un 13% en la teoría de juegos), por lo que podemos decir que los tiros que realizan al Atlético de Madrid son muy significativos en el resultado final, por lo que si al equipo rojiblanco le lanzan en más de una ocasión suele depararle un resultado adverso. Por otra lado, en la teoría de juegos podemos observar que los saques de esquina lanzados por el equipo colchonero (por eso generalmente conocido como Atlético aviación) tienen una incidencia muy notable en el resultado final, perdiendo un 16% de precisión si la eliminásemos.

Para el Club Atlético de Madrid como visitante:

Tabla 71: Top 10 variables significativas para el Club Atlético de Madrid Visitante

ANOVA	Valor	Random forest	Valor	Teoría de juegos	Valor
Tiros Visitante	Negativo	Tiros Visitante	8.1513216	Tiros Local	0.17450980
Tiros Local	Positivo	Tiros Local	4.8883046	Corners Local	0.14362745
Corners Visitante	Positivo	Sistema Local	2.7914191	Corners Visitante	0.11176471
Corners Local	Negativo	Rojas Local	2.3658434	Tiros Visitante	0.08235294
Faltas Local	Positivo	Rojas Visitante	1.1560641	Pases Visitante	0.02500000
Pases Visitante	Negativo	Posesión Visitante	0.9675016	Faltas Local	0.02156863
		Corners Visitante	0.8480462		
		Faltas Local	0.6692439		
		Posesión Local	0.5180093		
		Amarillas Local	0.4155860		

En este filtrado de datos podemos afirmar que las variables del equipo local tienen mejor ranking y por lo tanto mayor incidencia en el resultado final que las variables del equipo visitante (realizada la suma de los rankings, 76 del local sobre 69 del visitante). Al igual que pasaba en el análisis del equipo madrileño cuando jugaba en casa la única variable que forma parte en los 3 rankings es los tiros a puerta del rival, es decir los tiros a puerta del equipo local tienen una gran importancia (ya que, por ejemplo, en la teoría de juegos tiene una incidencia de un 17,45%). Otras de las variables que ocupan un buen lugar en las clasificaciones son los saques de esquina, muy incidentes en el ANOVA, o el sistema de juego del equipo local en el random forest.

Para el Fútbol Club Barcelona como local:

Tabla 72: Top 10 variables significativas para el Fútbol Club Barcelona Local

ANOVA	Valor	Random forest	Valor	Teoría de juegos	Valor
Tiros Local	Positivo	Corners Local	10.650977	Amarillas Visitante	0.14705882
Corners Local	Negativo	Tiros Local	6.946931	Tiros Local	0.14558824
Tiros Visitante	Negativo	Amarillas Visitante	6.083175	Corners Local	0.12352941
Amarillas Visitante	Positivo	Sistema Local	5.169224	Tiros Visitante	0.12352941
Amarillas Local	Positivo	Amarillas Local	4.932870	Amarillas Local	0.11029412
Faltas Visitante	Positivo	Posesión Visitante	3.655105	Faltas Visitante	0.08529412
		Rojas Local	3.250440		
		Tiros Visitante	2.530141		
		Posesión Local	2.414585		
		Fueras de juego Local	2.157192		

Para los analistas del Barça, podemos concluir que las amarillas del equipo visitante cuando ellos su equipo juega en casa tienen significación positiva en la clasificación de ANOVA en el resultado final y eliminarlas del estudio provocaría una disminución del 6% en el random forest, pero de un 14% en la teoría de juegos. Los saques de esquina y los tiros a puerta de los partidos del Fútbol Club Barcelona son muy significativos en el resultado final (con una pérdida del 16% de precisión entre ambas variables en el random forest y una de casi el 27% en la teoría de juegos).

Para el Fútbol Club Barcelona como visitante:

Tabla 73: Top 10 variables significativas para el Fútbol Club Barcelona Visitante

ANOVA	Valor	Random forest	Valor	Teoría de juegos	Valor
Tiros Visitante	Negativo	Tiros Visitante	11.9521779	Tiros Local	0.10652428
Tiros Local	Positivo	Tiros Local	11.0598999	Amarillas Local	0.09286881
Posesión Local	Negativo	Posesión Local	5.3328712	Tiros Visitante	0.07749767
Posesión Visitante	Positivo	Posesión Visitante	4.3656195	Rojas Local	0.07711251
Pases Local	Negativo	Pases Local	4.0571758	Rojas Visitante	0.07014472
Amarillas Local	Positivo	Sistema Local	2.8324938	Amarillas Visitante	0.06450747
Amarillas Visitante	Positivo	Rojas Visitante	2.0308746	Pases Local	0.04840103
Rojas Local	Positivo	Faltas Local	1.9692336	Posesión Local	0.04514472
Rojas Visitante	Positivo	Corners Visitante	0.9384208	Posesión Visitante	0.03544585
		Amarillas Visitante	0.6969546		

Si dividimos la base de datos y seleccionamos solo los partidos que juega el Fútbol Club Barcelona fuera de casa denotamos que las variables más significativas para el estudio según nuestros rankings son los tiros tanto del local como del visitante en igual medida. Curiosamente y rompiendo un mito la posesión del Barça no tiene una gran importancia en el estudio, ya que es la novena variable más significativa si sumamos todos los rankings (con un 4% y un 3% de pérdida de precisión en el random forest y teoría de juegos respectivamente). En la teoría de juegos la incidencia del árbitro es notable, pero para ambos bandos, porque si las quitásemos del estudio tanto las variables locales como visitantes de las amarillas y rojas eliminaríamos un 30% aproximadamente de precisión en el estudio, siendo mucho menos notable en el random forest.

Para el Real Madrid Club de Fútbol como local:

Tabla 74: Top 10 variables significativas para el Real Madrid Club de Fútbol Local

ANOVA	Valor	Random forest	Valor	Teoría de juegos	Valor
Tiros Visitante	Negativo	Corners Local	18.345911	Pases Local	0.10850840
Rojas Local	Negativo	Rojas Local	12.561958	Amarillas Local	0.09425770
Corners Local	Positivo	Tiros Local	7.026035	Tiros Visitante	0.08976424
Duelos Visitante	Negativo	Posesión Local	5.298029	Rojas Local	0.07930672
Duelos Local	Positivo	Posesión Visitante	5.251948	Posesión Visitante	0.07485994
Amarillas Local	Negativo	Pases Visitante	5.251494	Posesión Local	0.06954949
Posesión Visitante	Negativo	Tiros Visitante	4.899537	Duelos Local	0.06666667
Posesión Local	Positivo	Pases Local	3.494411	Duelos Visitante	0.06659664
Tiros Local	Negativo	Duelos Local	3.113918	Tiros Local	0.06602474
Pases Local	Positivo	Amarillas Visitante	2.244754	Corners Local	0.04917134

Si nos centramos en los datos del Real Madrid como local podemos obtener una conclusión muy curiosa, el Real Madrid uno de los equipos que más tiros a puerta hace en el mundo del fútbol y sus tiros son en muy baja medida significativos para ganar un partido (disminuye un 7% la precisión en el random forest y un 6% en la teoría de juegos). Incluso los tiros del visitante, que están en un ranking superior a los tiros locales, también están en un ranking inferior al normal en dos de las tres clasificaciones. Analizando estos datos podemos obtener la conclusión de la disparidad de las clasificaciones, ya que ninguna variable está en el top 3 en todas las clasificaciones, las que más incidencia tienen son las rojas al equipo blanco (con una significancia negativa en el ANOVA), los saques de esquina lanzados por el Real Madrid (con un 18% de pérdida de precisión en el random forest) y los tiros recibidos por los equipos que visitan el Santiago Bernabéu (perdiendo un 8% de precisión según la teoría de juegos).

Para el Real Madrid Club de Fútbol como visitante:

Tabla 75: Top 10 variables significativas para el Real Madrid Club de Fútbol Visitante

ANOVA	Valor	Random forest	Valor	Teoría de juegos	Valor
Tiros Visitante	Negativo	Tiros Local	6.772135	Tiros Local	0.15637255
Tiros Local	Positivo	Tiros Visitante	4.329588	Rojas Local	0.11715686
Rojas Local	Negativo	Posesión Visitante	3.988172	Amarillas Visitante	0.09509804
Amarillas Visitante	Positivo	Posesión Local	3.654985	Tiros Visitante	0.09019608
Pases Local	Positivo	Duelos Local	3.630361	Pases Local	0.07058824
		Fueras de juego Visitante	3.586377		
		Faltas Visitante	3.344641		
		Rojas Local	2.417212		
		Corners Visitante	2.118373		
		Pases Visitante	1.743427		

Realizando el estudio respecto a los datos del Real Madrid como visitante podemos afirmar que los tiros en contra del equipo madrileño tienen una incidencia significativa en el resultado final (teniendo una significancia negativa para los capitaleños), también tienen incidencia sus propios tiros, las rojas del equipo local y las amarillas del visitante, bastante significativas en el ANOVA y en la teoría de juegos. En el random forest no tiene tanta importancia la actuación del árbitro, sino que se centra más en las características tácticas del juego (por ejemplo, la posesión o los fueras de juego).

Para los partidos previos al COVID:

Tabla 76: Top 10 variables significativas para los partidos previos al COVID

ANOVA	Valor	Random forest	Valor	Teoría de juegos	Valor
Tiros Visitante	Negativo	Tiros Visitante	36.936217	Tiros Local	0.12033519
Tiros Local	Positivo	Tiros Local	34.264662	Tiros Visitante	0.11502026
Duelos Visitante	Negativo	Pases Local	10.598890	Rojas Local	0.07599206
Duelos Local	Positivo	Posesión Local	8.307134	Rojas Visitante	0.05819543
Rojas Local	Negativo	Corners Local	8.290181	Corners Local	0.04968761
Amarillas Local	Negativo	Posesión Visitante	8.079968	Amarillas Local	0.04938225
Sistema Local	Indeterminado	Rojas Local	7.940486	Sistema Local	0.04654115
Fueras de juego Local	Positivo	Duelos Local	5.889611	Fueras de juego Local	0.03970224
Rojas Visitante	Positivo	Duelos Visitante	5.659305	Duelos Local	0.03752251
Corners Local	Negativo	Pases Visitante	5.611029	Duelos Visitante	0.03528087

En el periodo previo al COVID las variables más significativas son los tiros a puerta, tanto visitantes en primera instancia como en los locales en segunda (perdiendo entre ambas aproximadamente un 71% en el random forest y un 23% en la teoría de juegos). Lo que si denotamos del periodo antes de la pandemia es que las variables locales tienen el doble de significación que las de los equipos visitantes (110 entre la suma de los rankings sobre 55). Las rojas tienen una vital importancia en la teoría de juegos, teniendo una reducción de la precisión de un 13% entre ambas. En el ANOVA curiosamente obtenemos la conclusión de que los duelos ganados por los equipos tienen una vital importancia en el resultado final, siendo las siguientes variables importantes tras los tiros a puerta.

Para los partidos posterior al COVID:

Tabla 60: Top 10 variables significativas para los partidos posteriores al COVID

ANOVA	Valor	Random forest	Valor	Teoría de juegos	Valor
Tiros Visitante	Negativo	Tiros Visitante	34.200188	Tiros Visitante	0.11977020
Tiros Local	Positivo	Tiros Local	33.989657	Tiros Local	0.10395246
Rojas Local	Negativo	Pases Local	11.325084	Rojas Local	0.06411395
Duelos Visitante	Negativo	Rojas Local	9.966613	Sistema Visitante	0.05354576
Duelos Local	Positivo	Pases Visitante	8.618229	Faltas Visitante	0.05054391
Pases Local	Negativo	Posesión Local	7.489724	Amarillas Local	0.04430427
Pases Visitante	Negativo	Posesión Visitante	7.256575	Pases Local	0.04333985
Sistema Visitante	Indeterminado	Corners Local	6.174629	Pases Visitante	0.03935284
Faltas Visitante	Positivo	Faltas Visitante	4.720671	Duelos Visitante	0.03798962
Amarillas Local	Positivo	Amarillas Visitante	4.096677	Duelos Local	0.03761451

En este periodo podemos afirmar que el equipo visitante y sus estadísticas tuvieron más incidencia en el resultado final, aunque siguieron siendo las variables locales por una mínima diferencia la más importante (88 sobre 77 en la suma de los rankings). Al igual que prácticamente en el resto de los filtrados las variables significativas son los tiros a puerta (perdiendo un 67% al eliminarlas ambas variables en el random forest). La tercera con mayor importancia, si sumamos todos los rankings, son las tarjetas rojas para el equipo local, que tiene una significación negativa en el ANOVA. Curiosamente las variables a continuación más significativas son los pases, tantos dados como local como visitante (disminuyendo la precisión aproximadamente un 8% al quitarlas). Por lo tanto, podemos afirmar que el COVID en el fútbol provocó una igualdad entre las variables significativas de los equipos que juegan en casa y los que juegan fuera, pero las variables más importantes siguieron siendo los tiros a puerta.

5. Página Web.

Esta página (https://github.com/guisabe99/TFG) ha sido realizada con la aplicación de Notepad y ejecutada con el formato HTML y CSS para los diseños gráficos. Adentrándonos en el formato de la página web se puede observar el título en la parte superior ("Estudio sobre La Liga Española desde la temporada 2016/2017 a 2024/2025") y a continuación hay una tabla con 6 apartados que muestran lo que se presenta en el resto de la página.

Al pulsar en cada uno de ellos nos lleva a la sección pertinente:

- **Todos los equipos**: En esta sección se muestran 4 páginas HTML (un documento de análisis descriptivo, uno de gráficos interactivos, uno de técnicas de clasificación y otro referente a las variables significativas). En cada uno de estos HTML se muestran los informes proporcionados por R de los scripts comentados anteriormente y los códigos. Además, en el HTML de las variables significativas se muestra las conclusiones obtenidas de estos rankings.
- **Temporadas**: En esta sección se muestran cada una de las 9 temporadas individualmente y al igual que con la sección anterior al introducirnos en cada temporada nos muestra los 4 scripts de código referentes a esa temporada (análisis descriptivo, gráficos interactivos, técnicas de clasificación y variables significativas), los archivos PDF con los informes referentes a estos códigos y en el HTML de variables significativas se muestra un apartado de conclusiones de los rankings (las temporadas mostradas en el apartado de resultados y las que no se muestran). Además, en el HTML de cada una de las temporadas existe la posibilidad de volver a la página anterior o a la de inicio, que vuelve a la sección de las temporadas o también se consigue pulsando a la flecha que indica hacia la izquierda en la parte superior de la pantalla.
- **Sistemas**: En esta sección se muestran los 3 sistemas de juego más repetidos durante el periodo estudiado. Al pulsar en cada uno de los sistemas observamos los mismos archivos descritos anteriormente, pero esta vez referentes a cada uno de los sistemas (4-3-3, 4-4-2, 4-2-3-1) cuando ocurren jugando de local y cuando ocurren jugando de visitante. De nuevo existe la página de volver al inicio o a la página anterior, o también podemos utilizar la flecha izquierda descrita en el apartado anterior.
- **Equipos**: En esta sección se presenta una tabla con 3 equipos, los más galardonados a día de hoy en el fútbol español, Real Madrid, Barcelona y Atlético (con 36, 28 y 11 títulos ligueros respectivamente). Esta sección tiene los documentos descritos en la sección del primer apartado, es decir análisis descriptivo, gráficos interactivos, técnicas de clasificación y variables significativas (código e informe R-Markdown, además del apartado de conclusiones de variables significativas) de los equipos a nivel local y a nivel visitante. Como en las secciones anteriores existe un enlace o la flecha izquierda que nos lleva a la página inicial (o a la página anterior), llevando en este caso a los equipos y a su tabla.

- **Sucesos**: En este apartado se muestra una tabla con dos campos (antes y después del COVID). Estos enlaces nos presentan los documentos descritos en secciones anteriores, es decir, las estadísticas descritas en el punto de 4 de este informe. Al pinchar en cada uno de los enlaces y sus distintos apartados, tenemos la opción de volver a las páginas anteriores, y para ello existen dos maneras: a través de "Vuelve a la página de inicio" o través de la flecha (explicada en la sección de temporadas.)
- **Otros archivos**: Este apartado nos muestra varios documentos interesantes
    - Un documento sobre las variables significativas y sus rankings (individuales y grupales con locales y visitantes.)
    - Un documento que indica el valor asociado a cada sistema de juego.
    - Un documento con todas las matrices de confusión de cada base de datos.
    - Dos documentos de diferentes filtrados (uno de código de R y otro de informe), como, por ejemplo, un duelo concreto, un equipo con un sistema de juego especifico o una comparación entre sistemas de juego. Además, se muestran los recuentos y porcentajes de cada una de las clases de la variable respuesta.

Además, como podemos observar al principio de cada sección hay una frase dicha por una persona mítica del mundo de fútbol, que nos hacen recordar lo mágico que es este deporte.

Por ejemplo, en las siguientes figuras se muestra el proceso para visualizar cada uno de los apartados en el caso del sistema 4-3-3 cuando es utilizado de local. Este proceso es igual en cada equipo, sistema, temporada o suceso.

***Figura 20**: Inicio de la web. Fuente: Elaboración propia.*

En la figura 20 se muestra el inicio de la página web. En primer lugar, tenemos el título, a continuación, el documento de informe y posteriormente una tabla con los 6 apartados

comentados anteriormente, para seguir el ejemplo del sistema 4-3-3 local hay que pulsar "Sistemas".

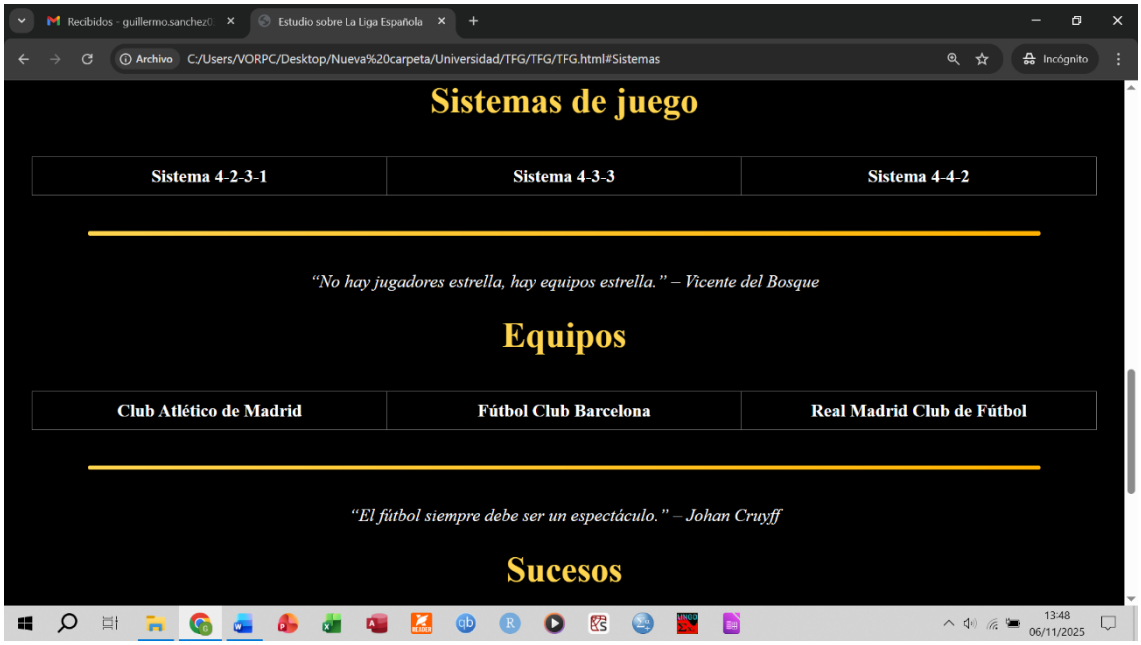

***Figura 21**: Sistemas de la web. Fuente: Elaboración propia.*

En la figura 21 nos muestra los sistemas, y por lo tanto debemos darle a "Sistema 4-3-3".

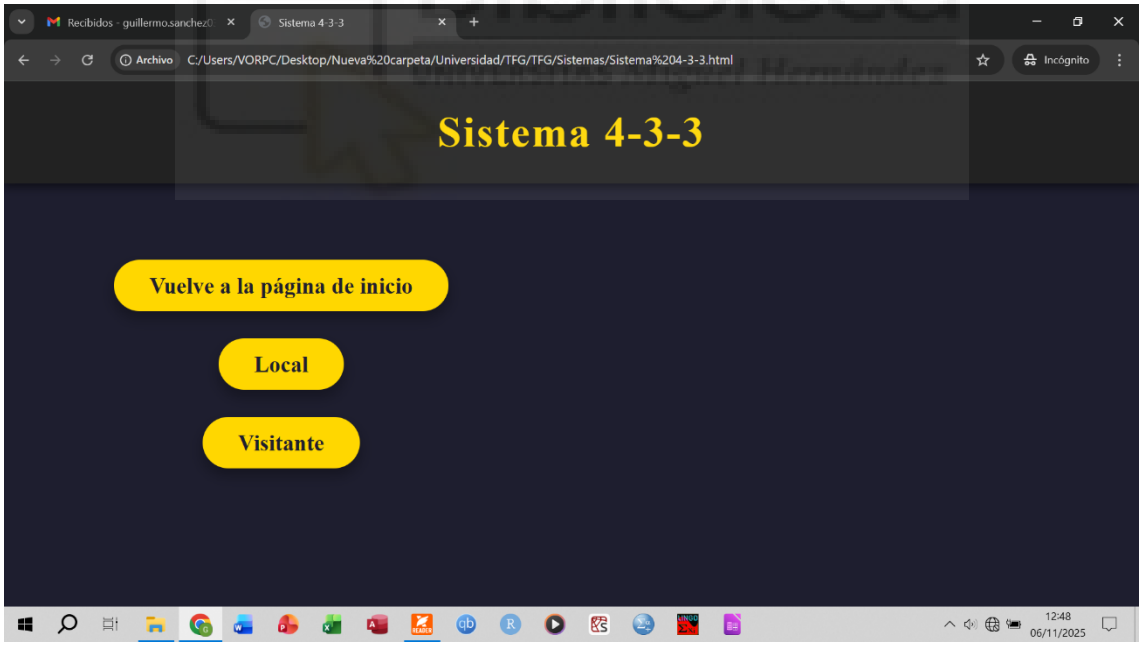

***Figura 22**: Sistema 4-3-3 en la web. Fuente: Elaboración propia.*

En la figura 22 hay que elegir entre el sistema 4-3-3 cuando es utilizado de local o cuando es utilizado de visitante, por lo tanto, en el ejemplo se pulsa en "Local".

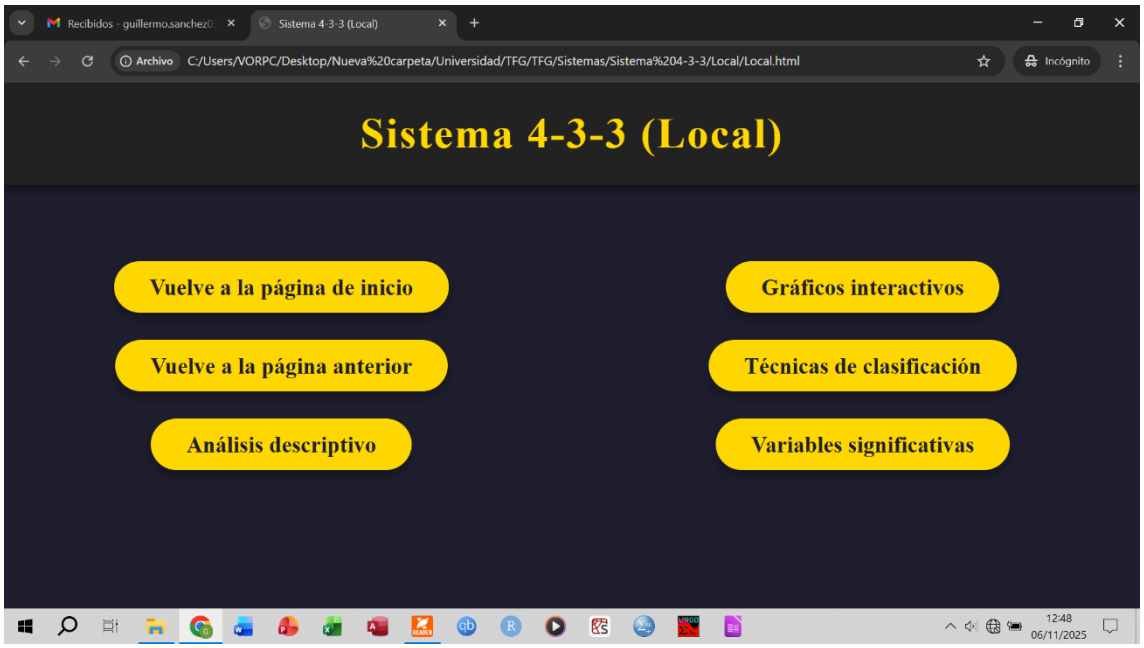

***Figura 23**: Sistema 4-3-3 como local en la web. Fuente: Elaboración propia.*

En la figura 23 se muestran cada uno de los análisis que se han realizado en el estudio. Para visualizar cada uno de estos análisis hay que pulsar en ellos, como hemos hecho en las figuras 24, 25, 26 y 27.

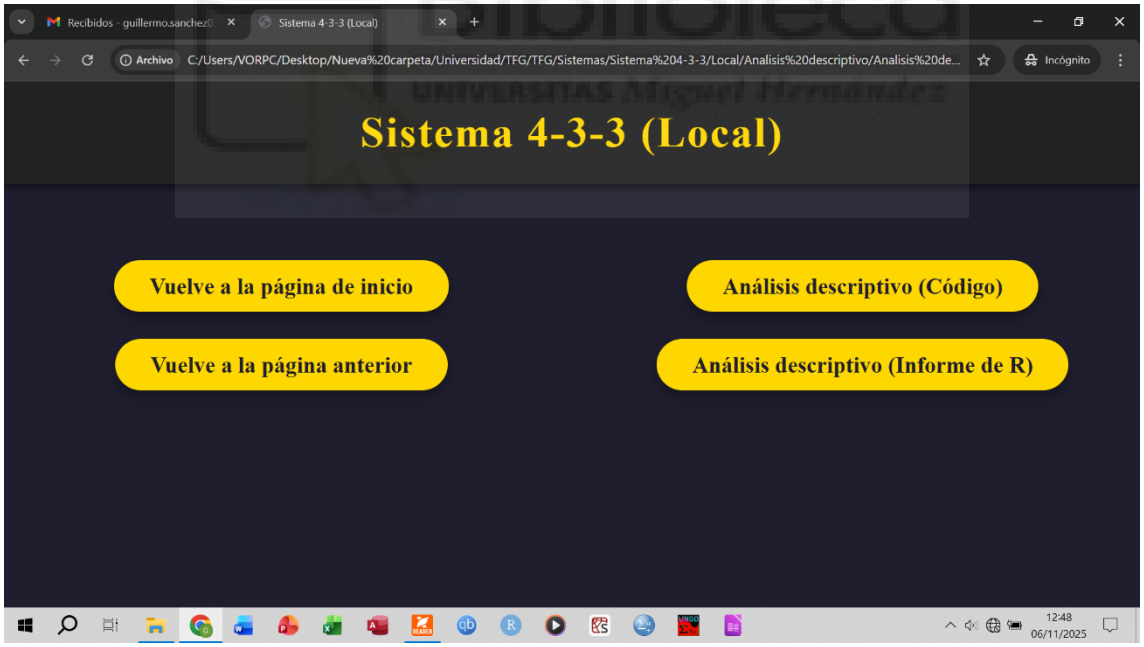

***Figura 24**: Análisis descriptivo del sistema 4-3-3 local en la web. Fuente: Elaboración propia.*

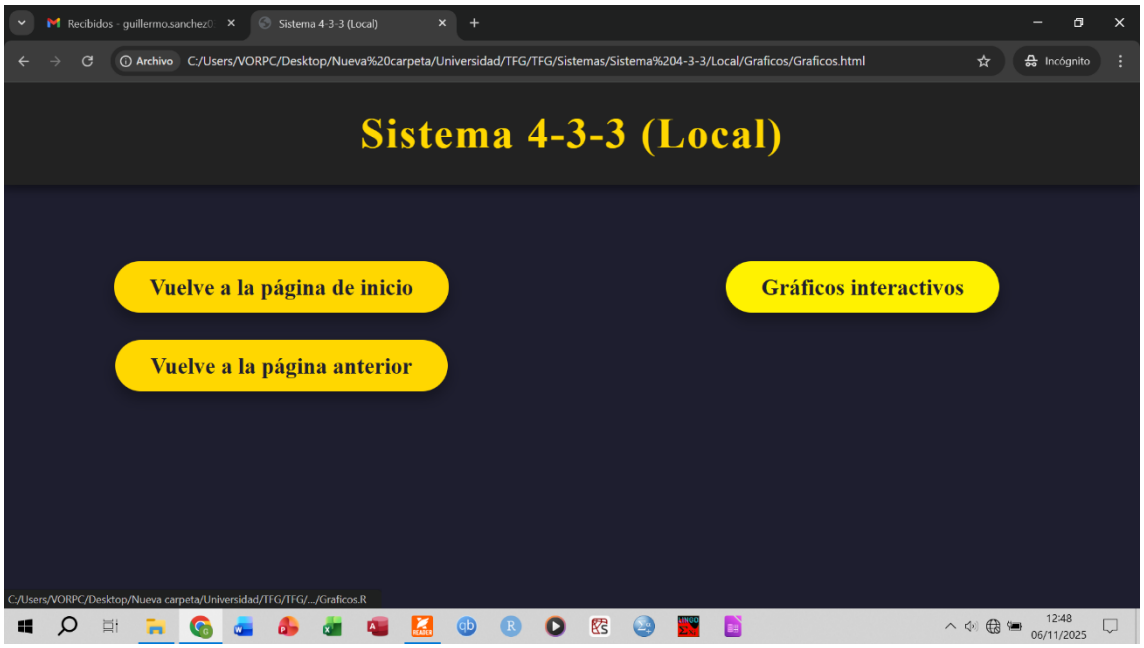

***Figura 25**: Gráficos interactivos del sistema 4-3-3 local en la web. Fuente: Elaboración propia.*

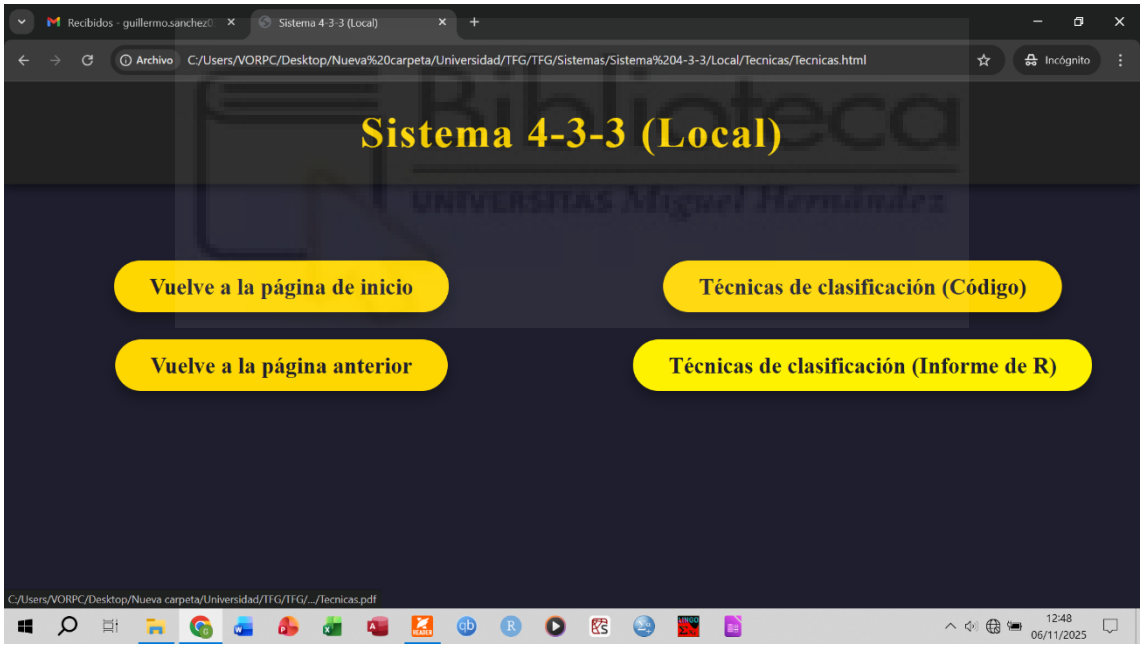

***Figura 26**: Técnicas de clasificación del sistema 4-3-3 local en la web. Fuente: Elaboración propia.*

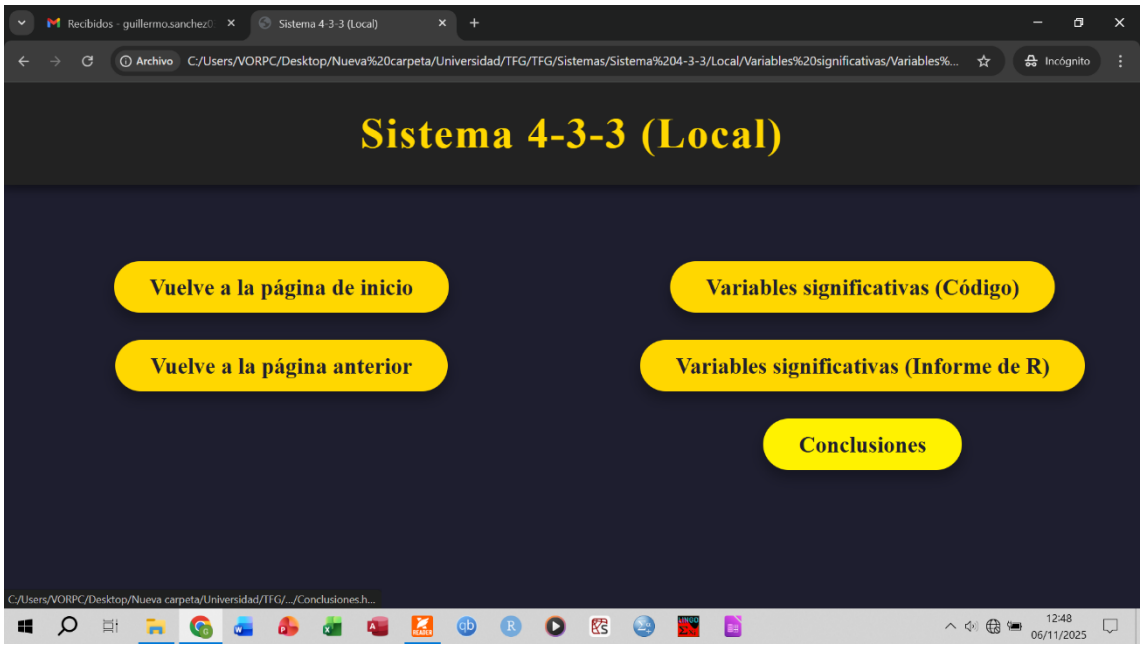

***Figura 27**: Variables significativas del sistema 4-3-3 local en la web. Fuente: Elaboración propia.*

6. Conclusiones.

Basándonos en el estudio de las técnicas de clasificación, podemos afirmar que la técnica con mayor cantidad de aciertos es el análisis discriminante, siendo la cantidad de aciertos un valor dentro del intervalo del 55% y 65%, siendo superior en los análisis de los grandes equipos en el apartado local, ya que supera el 80% en Real Madrid y Barcelona. Con estos datos podemos concluir que las variables estudiadas presentan relaciones predominantemente lineales, lo que se ajusta adecuadamente a los supuestos del modelo discriminante (normalidad y homogeneidad de covarianzas).

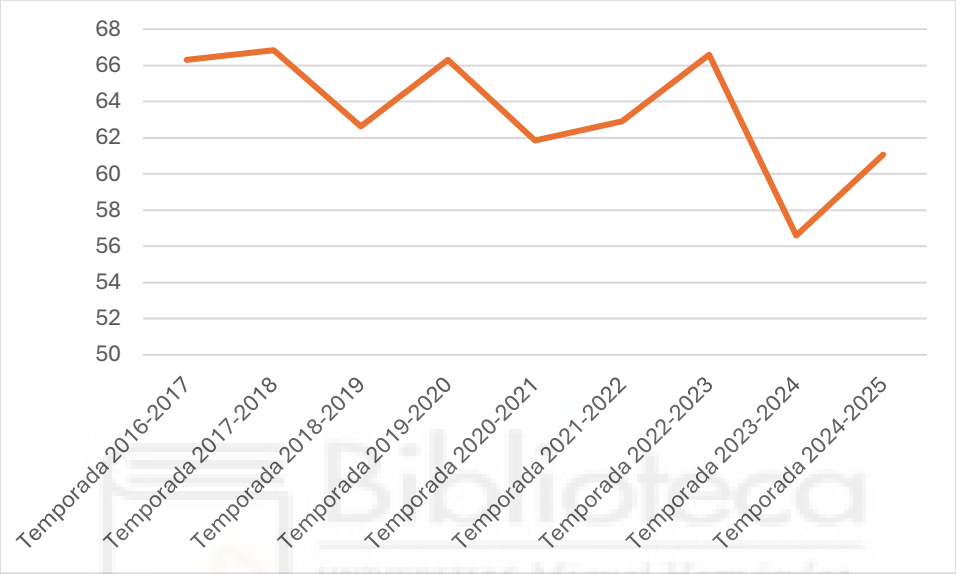

***Figura 28**: % de aciertos del análisis discriminante por temporada. Fuente: Elaboración propia.*

En la figura 28 podemos ver el porcentaje de aciertos del análisis discriminante por temporada. En él podemos comprobar que donde menos acierto hay es en la temporada 2023-2024, la única que reduce del 60% de precisión, pero también se acerca a ese valor.

Al ver la influencia del COVID podemos afirmar que, aunque haya una menor precisión (59,85 sobre 64,26) en el caso de los partidos posteriores a la pandemia, la diferencia no es tan abismal como se preveía. Eso sí, habiendo un descenso a partir del confinamiento, con un solo repunte significativo en la temporada 2022-2023.

Por otro lado, podemos afirmar que la técnica del KNN presenta los peores resultados, con precisiones frecuentemente por debajo del 50%. Este bajo rendimiento puede explicarse por la presencia de ruido en los datos, diferencias de escala entre variables o una dimensionalidad elevada, factores que tienden a afectar negativamente al método.

Si nos centramos en los equipos estudiados cuando juegan en casa tienen una gran cantidad de acierto, ya que es más probable ganar en casa, como hemos demostrado con el análisis descriptivo realizado anteriormente, y más siendo un gran equipo. La cantidad de precisión de estos equipos es superior al 75% en los 3 clubes. Por el contrario, estos equipos cuando juegan de visitantes, aunque tienen una precisión alta es aproximadamente similar a la media de todos los equipos (alrededor del 60%).

Respecto a los sistemas con el que más precisión se obtiene es el sistema 4-3-3, un sistema muy recurrente en los entrenadores de Real Madrid y Barcelona, aunque no tanto en el resto de los equipos. Con el sistema 4-3-3 podemos decir que la precisión es similar a la de los otros sistemas, pero este tiene una precisión un poco superior.

Finalmente, si estudiamos las variables significativas (objetivo principal del estudio) podemos afirmar que los tiros a puerta son muy importantes para decidir el resultado final en todos los rankings utilizados. Afirmamos esto debido a que si las suprimimos perderíamos entre un 40% y un 70% de precisión en el modelo. El ANOVA identifica una relación positiva entre los tiros locales y el resultado final, mientras que los tiros visitantes presentan una relación negativa, lo que confirma que la efectividad ofensiva es el factor más determinante para resultado final.

Tras los tiros a puerta, las tarjetas rojas, especialmente las mostradas al equipo local, son valores influyentes en el resultado final, ya que si a un equipo le expulsan un jugador probablemente pierda el partido. Otras variables, como los pases, los saques de esquina o la posesión, tienen relevancia en algunos conjuntos de datos, pero su incidencia en el resultado final es menor.

Tras analizar el conjunto de datos y sus diferentes fracciones podemos concluir en que la influencia de las variables locales es más significativa que las variables visitantes, pero que a raíz de la pandemia esa diferencia disminuye (especialmente en la temporada 2020-2021 que no había público), aunque sigue siendo a favor de los locales. A raíz de la vuelta de los aficionados, aunque no volvió a ser como en las temporadas anteriores la diferencia entre significancia de las variables locales y visitantes volvió a aumentar.

Estudiando los 3 equipos más laureados sacamos conclusiones curiosas como las siguientes:

- En el **Real Madrid** cuando juega de local sus tiros a puerta no tienen una relevancia significativa tan importante como la esperada, como por ejemplo si las tienen las rojas que son muy determinantes en el resultado final. Como visitante, los tiros en contra condicionan claramente el resultado.
- En el **FC Barcelona** podemos afirmar que jugar en casa conlleva a que los saques de esquina y los tiros a puerta sean variables muy significativas, con una pérdida de precisión superior al 25% al ser eliminadas. La posesión, tradicionalmente asociada al estilo de juego del club catalán, no muestra una relación significativa con el resultado final. Cuando juega de visitante, aunque sus tiros tienen una gran relevancia, los tiros recibidos en contra también son significativos en el resultado.
- En el caso del **Atlético de Madrid**, podemos afirmar que cuando juega tanto de local como de visitante, los tiros lanzados por el rival son la variable más influyente, lo que nos dice que la defensa del equipo colchonero, comandado por Simeone, es muy importante en el objetivo de evitar ocasiones del rival, ya que son determinantes para el resultado final si los hubiese.

Si estudiamos la base de datos por cada uno de los sistemas tácticos, podemos decir que los tiros a puerta son lo más relevante para predecir el resultado final en cada uno de ellos.

Teniendo otras variables influyentes en el tercer puesto de significancia (saques de esquina en el 4-3-3, o las rojas en el 4-4-2 o en el 4-2-3-1).

En definitiva, los resultados de este estudio nos reafirman que en todos los conjuntos de datos el ataque (los tiros a puerta y en menor instancia los saques de esquina) son las variables más significativas para el resultado final. También es clave en el resultado la disciplina, ya que las expulsiones afectan directamente a las probabilidades de victoria y, por último, podemos concluir que el impacto del COVID ha tenido una relevancia menos significativa de la que podríamos pensar, pero que también tuvo relevancia, ya que las variables de los equipos visitantes tuvieron más importancia durante el tiempo de ausencia del público.

7. Bibliografía.

AdvancedTech. (14 de abril de 2008). Clasificación supervisada y no supervisada. WordPress. https://advancedtech.wordpress.com/2008/04/14/clasificacion-supervisada-y-no-supervisada/

Barrios, J. (s. f.). La matriz de confusión y sus métricas. JuanBarrios.com. https://www.juanbarrios.com/la-matriz-de-confusion-y-sus-metricas/

Capítulo 5: Árboles de decisión. (s. f.). Computación de Datos y R: Libro abierto (CDR Book). https://cdr-book.github.io/cap-arboles.html

Capítulo 7: Análisis discriminante. (s. f.). Computación de Datos y R: Libro abierto (CDR Book). https://cdr-book.github.io/cap-discriminante.html

Capítulo 8: Bagging y Random Forest. (s. f.). Computación de Datos y R: Libro abierto (CDR Book). https://cdr-book.github.io/cap-bagg-rf.html

Fisher, R. A. (1925). Statistical methods for research workers. Oliver and Boyd.

Fisher, R. A. (1936). The use of multiple measurements in taxonomic problems. Annals of Eugenics, 7(2), 179–188.

Football Data. (s. f.). Spanish football data. Football-Data.co.uk. https://www.football-data.co.uk/spainm.php

Gil, C. (s. f.). Validación cruzada y bootstrap. RPubs. https://rpubs.com/Cristina_Gil/CV_Bootstrap

Herrera Briones, J. (2020). Análisis de datos mediante aprendizaje automático aplicado a fútbol profesional [Trabajo de Fin de Grado, Universidad Politécnica de Madrid]. Repositorio UPM. https://oa.upm.es/67548/1/TFG_JUAN_HERRERA_BRIONES.pdf

IBM. (s. f.). Aprendizaje no supervisado (unsupervised learning). IBM Think. https://www.ibm.com/es-es/think/topics/unsupervised-learning

IBM. (s. f.). K-nearest neighbors (KNN). IBM Think. https://www.ibm.com/es-es/think/topics/knn

IBM. (s. f.). Support Vector Machine (SVM). IBM Think. https://www.ibm.com/es-es/think/topics/support-vector-machine

Martínez, T. L. (Ed.). (2000). Técnicas de análisis de datos en investigación de mercados. Editorial Pirámide.

Myerson, R. B. (1991). Game theory: Analysis of conflict. Harvard University Press.

Nash, J. F. (1950). Equilibrium points in n-person games. Proceedings of the National Academy of Sciences, 36(1), 48–49.

OpenAI. (s. f.). ChatGPT. OpenAI. https://chatgpt.com/

Osborne, M. J., & Rubinstein, A. (1994). A course in game theory. MIT Press.

Resultados Fútbol. (s. f.). Resultados de fútbol en directo. Resultados-Fútbol.com. https://www.resultados-futbol.com/

SofaScore. (s. f.). Resultados y estadísticas deportivas en tiempo real. SofaScore. https://www.sofascore.com/es/

Universidad de Alicante. (s. f.). Apuntes de análisis discriminante. Repositorio Institucional RUA. https://rua.ua.es/server/api/core/bitstreams/2476203a-5908-474c-8a58-524f26c4c2b2/content

Universidad Internacional de La Rioja (UNIR). (s. f.). Teoría de juegos. UNIR Revista Empresa. https://www.unir.net/revista/empresa/teoria-de-juegos/

Von Neumann, J., & Morgenstern, O. (1944). Theory of games and economic behavior. Princeton University Press.

Zhang, H. (2005). Exploring conditions for the optimality of naïve Bayes. Pattern Recognition Letters, 26(15), 2101–2112. https://doi.org/10.1016/j.patrec.2005.03.007

Casas, B., Janeiro, F., Jurado, I. G., & Díaz, J. G. (2012). Introducción a la Teoría de Juegos. USC editora. https://d1wqtxts1xzle7.cloudfront.net/115292161/teoria_de_juegos.pdf