UNIVERSIDAD MIGUEL HERNÁNDEZ DE ELCHE

FACULTAD DE CIENCIAS SOCIALES Y JURÍDICAS DE ELCHE

GRADO EN ESTADÍSTICA EMPRESARIAL

TRABAJO FIN DE GRADO

Valoración de jugadores de tenis y predicción de resultados mediante el sistema Elo y modelos de clasificación supervisada

Autor: José Manuel Fernández Romero

Tutor: Juan Carlos Gonçalves Dosantos

Cotutora: Ana Meca Martínez

Curso académico 2025–2026

ÍNDICE

1. Introducción

El tenis profesional es un deporte individual en el que el rendimiento de los jugadores se evalúa habitualmente a través de rankings, siendo el ranking ATP la referencia principal para determinar la jerarquía competitiva del circuito. Este ranking se construye a partir de los resultados obtenidos en los distintos torneos y permite establecer comparaciones globales entre jugadores. No obstante, al analizar con mayor detalle enfrentamientos concretos entre jugadores, resulta evidente que la posición en el ranking no siempre coincide con la probabilidad real de victoria en un partido determinado. Al basarse en la acumulación de puntos a lo largo del tiempo, el sistema puede tardar en reflejar cambios recientes en el nivel de juego o particularidades específicas de ciertos emparejamientos.

La necesidad de comparar jugadores y predecir resultados hace del tenis un contexto especialmente adecuado para la aplicación de herramientas estadísticas. Cada partido puede entenderse como una interacción estratégica en la que intervienen tanto el nivel relativo de los jugadores como múltiples factores asociados al rendimiento en pista. Desde esta perspectiva, el análisis cuantitativo no pretende sustituir las clasificaciones tradicionales, sino complementarlas con modelos que permitan interpretar el rendimiento desde ángulos alternativos y, en algunos casos, más ajustados al enfrentamiento directo.

Desde un punto de vista teórico, la teoría de juegos proporciona un marco natural para abordar problemas de comparación y asignación de valor en situaciones competitivas (Casas Méndez, Fiestras Janeiro, García Jurado, & González Díaz, 2012). En particular, los juegos cooperativos y conceptos como el valor de Shapley (Shapley, 1953) permiten distribuir de forma razonada una medida de rendimiento entre los distintos participantes, dando lugar a rankings agregados que reflejan la contribución relativa de cada jugador dentro de un sistema competitivo. La incorporación de este enfoque al tenis resulta especialmente interesante, ya que permite reinterpretar la jerarquía competitiva no solo como acumulación de puntos, sino como resultado de interacciones entre jugadores.

Junto a estos enfoques agregados, los métodos de clasificación supervisada (James, Witten, Hastie, & Tibshirani, 2021) permiten abordar el problema desde un punto de vista complementario, centrado en la predicción del resultado de partidos individuales. En este trabajo se opta por utilizar la regresión logística debido a su equilibrio entre capacidad predictiva e interpretabilidad, lo que facilita no solo estimar probabilidades de victoria, sino también analizar la influencia de distintas variables de rendimiento. A través de este tipo de modelos es posible incorporar estadísticas detalladas y obtener una estimación cuantitativa de la fortaleza relativa de los jugadores en enfrentamientos directos.

El objetivo de este Trabajo de Fin de Grado es analizar y comparar distintos enfoques de clasificación y predicción aplicados al tenis profesional. En concreto, se estudia un sistema de tipo Elo basado en teoría de juegos cooperativos, el valor de Shapley y sus extensiones, así como un modelo de clasificación supervisada basado en regresión logística. La comparación entre ambos enfoques, junto con su contraste con el ranking ATP oficial, busca aportar una visión crítica sobre hasta qué punto estas herramientas describen adecuadamente la jerarquía competitiva del circuito y permiten estimar con mayor precisión la probabilidad de victoria en enfrentamientos individuales.

2. Preliminares

En esta sección se introducen los conceptos teóricos fundamentales que servirán de base para el desarrollo del modelo propuesto y su posterior aplicación al ranking de jugadores de tenis. En particular, se presentan herramientas procedentes del aprendizaje estadístico y de la teoría de juegos cooperativos, que permiten abordar el problema desde dos perspectivas complementarias.

Por un lado, se introduce el marco de la clasificación supervisada, centrando la atención en la regresión logística como método para modelizar la probabilidad de ocurrencia de un resultado binario a partir de un conjunto de variables explicativas. Este enfoque resulta especialmente adecuado para el análisis de resultados deportivos, donde el objetivo es pronosticar el desenlace de un enfrentamiento entre dos jugadores a partir de información observada sobre su rendimiento.

Por otro lado, se presentan los conceptos básicos de la teoría de juegos cooperativos, con especial énfasis en los juegos en forma característica y en el valor de Shapley como medida de importancia relativa de los jugadores dentro de un sistema. Este marco teórico permite cuantificar la contribución individual de cada jugador al resultado global y proporciona una interpretación axiomática de dicha contribución.

La combinación de ambos enfoques permitirá, en capítulos posteriores, construir un modelo aplicado al tenis que integre probabilidades de victoria y medidas de importancia relativa.

2.1 Clasificación supervisada

La clasificación supervisada es una rama del aprendizaje estadístico cuyo objetivo es asignar una etiqueta categórica a un individuo a partir de un conjunto de variables explicativas previamente observadas. Este tipo de problemas aparece de forma natural en numerosos ámbitos, como el diagnóstico médico, la detección de fraude o la predicción de resultados deportivos, y constituye uno de los pilares fundamentales del aprendizaje automático supervisado (James, Witten, Hastie, & Tibshirani, 2021).

Formalmente, se considera un conjunto de individuos Ω para los cuales se dispone de una variable respuesta categórica Y y un vector de predictores $X = (X_1, X_2, \dots, X_k)$. Para cada individuo $i \in \Omega$, se observa un par (x_i, y_i), donde y_i pertenece a un conjunto finito de clases. El objetivo consiste en construir un modelo que, a partir de los datos disponibles, permita predecir la clase de nuevos individuos para los que únicamente se conocen sus variables explicativas.

Desde un punto de vista teórico, el clasificador óptimo viene dado por el clasificador de Bayes, que asigna cada observación a la clase con mayor probabilidad condicionada dadas sus características. Sin embargo, en la práctica dichas probabilidades son desconocidas, lo que hace necesario recurrir a métodos estadísticos que permitan aproximar este clasificador a partir de los datos observados.

En el contexto del aprendizaje supervisado, es importante distinguir entre problemas de regresión y problemas de clasificación. Mientras que en la regresión la variable respuesta es continua, en la clasificación la variable respuesta toma valores discretos, lo que requiere el uso de técnicas específicas para la construcción de reglas de decisión.

Existen diversas técnicas de clasificación supervisada, entre las que se encuentran la regresión logística, el análisis discriminante, los métodos basados en vecinos más próximos, los árboles de clasificación y otros enfoques más complejos. En este trabajo se emplea la regresión logística, que se desarrolla en el apartado siguiente, por su interpretación probabilística, su simplicidad y su adecuación al problema considerado, siendo una de las técnicas más utilizadas en problemas de clasificación binaria (James, Witten, Hastie, & Tibshirani, 2021).

2.1.1 Regresión logística

La regresión logística es una técnica de clasificación supervisada que se utiliza cuando la variable respuesta es binaria. Su objetivo es modelizar la probabilidad de que un individuo pertenezca a una de las dos clases posibles a partir de un conjunto de variables explicativas. A diferencia de los modelos de regresión lineal, la regresión logística permite que las predicciones estén acotadas entre 0 y 1, lo que posibilita una interpretación probabilística directa de los resultados. (James, Witten, Hastie, & Tibshirani, 2021)

Aunque podría plantearse un modelo de regresión lineal para una variable respuesta binaria, este enfoque presenta importantes limitaciones, como la posibilidad de obtener predicciones fuera del intervalo [0,1] y la falta de una relación adecuada entre los predictores y la probabilidad de pertenencia a una clase. La regresión logística resuelve estos problemas mediante el uso de una función de enlace no lineal, conocida como función logística, que transforma una combinación lineal de los predictores en una probabilidad válida. (James, Witten, Hastie, & Tibshirani, 2021)

Sea Y una variable aleatoria binaria codificada como $Y \in \{0,1\}$, y sea $(X_1, X_2, \dots, X_k)$ un conjunto de variables explicativas. El modelo de regresión

logística expresa la probabilidad condicional de que $Y = 1$ mediante la función logística:

$$P(Y = 1 \mid X_1, \dots, X_k) = \frac{1}{1 + e^{-(\beta_0 + \beta_1 X_1 + \cdots + \beta_k X_k)}},$$

donde $\beta_0, \beta_1, \dots, \beta_k$ son los parámetros del modelo. La probabilidad de que la variable respuesta tome el valor cero viene dada por

$$P(Y = 0 \mid X_1, \dots, X_k) = 1 - P(Y = 1 \mid X_1, \dots, X_k).$$

La función logística presenta una forma sigmoide, lo que permite modelizar transiciones suaves entre probabilidades cercanas a 0 y a 1. Esta propiedad resulta especialmente adecuada en problemas de clasificación, ya que pequeñas variaciones en los predictores pueden producir cambios graduales en la probabilidad estimada. En la Figura 1 se muestra de forma ilustrativa el ajuste de un modelo de regresión logística a una variable respuesta binaria.

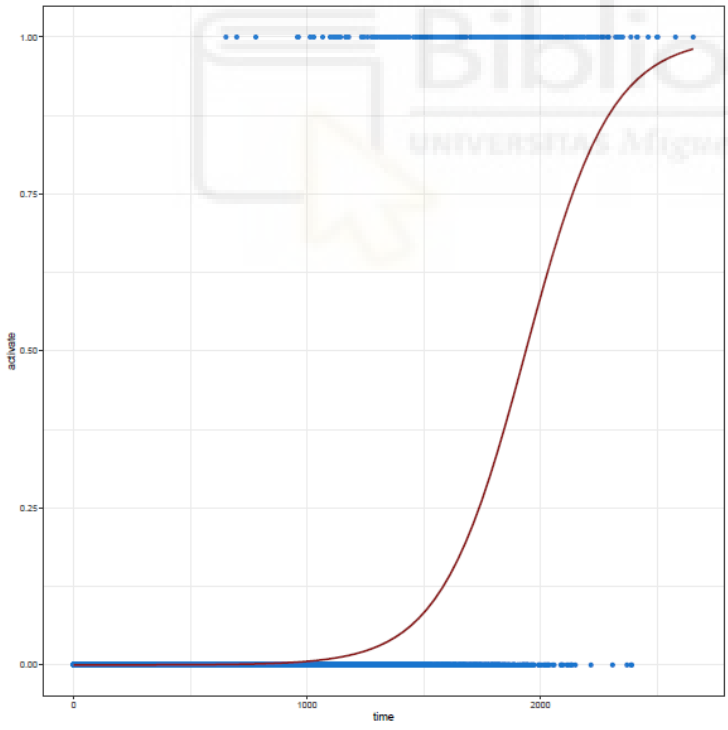

Figura 1. Ajuste de un modelo de regresión logística a una variable respuesta binaria. La curva sigmoide representa la probabilidad estimada P(Y = 1 | X) en función de una variable explicativa numérica X (eje horizontal) y una variable respuesta binaria Y (eje vertical).

Una de las principales ventajas de la regresión logística es su interpretación en términos de razones de probabilidad. Se define el *odds* o razón de probabilidades como

$$\text{odds} = \frac{P(Y = 1 \mid X_1, \dots, X_k)}{1 - P(Y = 1 \mid X_1, \dots, X_k)}.$$

El logaritmo natural del *odds*, denominado *logit*, se expresa de forma lineal en los predictores:

$$\log\left(\frac{P(Y = 1 \mid X_1, \dots, X_k)}{1 - P(Y = 1 \mid X_1, \dots, X_k)}\right) = \beta_0 + \beta_1 X_1 + \cdots + \beta_k X_k.$$

Desde el punto de vista interpretativo, el coeficiente βj representa el cambio en el logaritmo del *odds* asociado a un incremento unitario en la variable Xj, manteniendo constantes el resto de predictores.

Esta formulación permite interpretar fácilmente los coeficientes del modelo. En particular, si $\beta_j > 0$, un aumento en la variable X_j incrementa la probabilidad de que $Y = 1$, mientras que si $\beta_j < 0$, dicho aumento reduce dicha probabilidad. (James, Witten, Hastie, & Tibshirani, 2021)

Los parámetros del modelo se estiman mediante el método de máxima verosimilitud, que consiste en encontrar los valores de $\beta_0, \beta_1, \dots, \beta_k$ que maximizan la función de verosimilitud asociada a las observaciones disponibles.

Sea $\{(X_i, Y_i)\}_{i=1}^{n}$ el conjunto de observaciones, con $Y_i \in \{0,1\}$. Suponiendo independencia entre las observaciones, la función de verosimilitud del modelo de regresión logística se define como

$$\ell(\beta) = \prod_{i:Y_i=1} P(Y_i = 1 \mid X_i) \prod_{i:Y_i=0} [1 - P(Y_i = 1 \mid X_i)],$$

donde la probabilidad $P(Y_i = 1 \mid X_i)$ viene determinada por el modelo logístico definido previamente.

Los estimadores de los parámetros β se obtienen maximizando esta función de verosimilitud, o de forma equivalente, maximizando su logaritmo.

Una vez estimados los coeficientes, es posible evaluar la significación estadística de cada predictor mediante contrastes de hipótesis, habitualmente contrastando la hipótesis nula $H_0: \beta_j = 0$.

Para realizar predicciones sobre nuevos individuos, se sustituyen los valores de las variables explicativas en la expresión del modelo, obteniendo así una probabilidad estimada de pertenencia a la clase $Y = 1$. A partir de estas probabilidades, se define una regla de decisión basada en un umbral $z \in [0,1]$, de modo que se asigna la clase $Y = 1$ si la probabilidad estimada es mayor que z, y la clase $Y = 0$ en caso contrario. Habitualmente se toma $z = 0.5$, aunque este valor puede ajustarse según el contexto del problema.

2.1.2 Matriz de confusión

La calidad de un modelo de regresión logística se evalúa mediante la matriz de confusión, que compara las clases observadas con las clases predichas. A partir de dicha matriz se definen distintas medidas de rendimiento, como la precisión, la sensibilidad, la especificidad y las tasas de error, que permiten cuantificar la capacidad predictiva del modelo.

En la Tabla 1 se presenta la estructura de la matriz de confusión para un problema de clasificación binaria, donde se comparan las clases observadas con las clases predichas por el modelo y se identifican los distintos tipos de aciertos y errores de clasificación.

Tabla 1. Matriz de confusión en un problema de clasificación binaria.

Clase real/Clase predicha	1	0
1	Verdadero positivo (VP)	Falso negativo (FN)
0	Falso positivo (FP)	Verdadero negativo (VN)

Tal y como se muestra en la Tabla 1, los verdaderos positivos y verdaderos negativos corresponden a clasificaciones correctas, mientras que los falsos positivos y falsos negativos representan errores de clasificación.

En el caso binario, la matriz de confusión se compone de cuatro cantidades: verdaderos positivos (VP), que corresponden a observaciones correctamente clasificadas como pertenecientes a la clase positiva; verdaderos negativos (VN), que representan observaciones correctamente clasificadas como pertenecientes a la clase negativa; falsos positivos (FP), que son observaciones clasificadas erróneamente como positivas; y falsos negativos (FN), que corresponden a observaciones clasificadas erróneamente como negativas.

A partir de la matriz de confusión se definen diversas medidas de rendimiento. La precisión global se define como la proporción de observaciones correctamente clasificadas sobre el total, y viene dada por

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

La sensibilidad mide la capacidad del modelo para identificar correctamente las observaciones de la clase positiva, y se define como

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

Por su parte, la especificidad cuantifica la capacidad del modelo para identificar correctamente las observaciones de la clase negativa, y se expresa como

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

Estas medidas permiten evaluar de forma más detallada el comportamiento del modelo de clasificación, especialmente en situaciones en las que las clases están desbalanceadas o cuando los errores de distinto tipo tienen consecuencias diferentes (James, Witten, Hastie, & Tibshirani, 2021).

2.2. Juegos cooperativos

La teoría de juegos es una disciplina matemática que estudia situaciones estratégicas en las que varios agentes, denominados jugadores, toman decisiones que afectan tanto a su propio resultado como al de los demás participantes. Dentro de esta teoría se distinguen dos grandes enfoques: los juegos no cooperativos, en los que cada jugador actúa de manera individual buscando maximizar su propio beneficio, y los juegos cooperativos, en los que los jugadores pueden formar coaliciones y actuar de manera conjunta para obtener ventajas colectivas.

En los juegos cooperativos, el objetivo principal no es analizar estrategias individuales, sino estudiar la capacidad productiva de los distintos grupos de jugadores y determinar cómo repartir el valor generado por la cooperación entre sus miembros. Para formalizar este tipo de situaciones se emplea la noción de juego en forma característica, en la que se asigna a cada coalición un valor numérico que representa la utilidad o recompensa que dicho grupo puede alcanzar si actúa de forma coordinada (Pérez Navarro, Jimeno Pastor, & Cerdá Tena, 2004).

Uno de los conceptos fundamentales en este ámbito es el valor de Shapley, introducido por Lloyd Shapley (1953), que proporciona un criterio axiomático para distribuir el valor total del juego entre los jugadores de manera coherente, en función de su contribución marginal a las coaliciones. Este valor se caracteriza por propiedades como eficiencia, simetría, jugador nulo y aditividad, que justifican su uso como regla de reparto en juegos cooperativos (Shapley, 1953) (Casas Méndez, Fiestras Janeiro, García Jurado, & González Díaz, 2012)

Finalmente, existen extensiones del valor de Shapley que permiten adaptar el marco clásico a situaciones más generales. Entre ellas destaca la generalización propuesta por Nowak y Radzik (1994), que incorpora la posibilidad de que el orden en que los jugadores se incorporan a una coalición influya en el valor generado.

2.2.1 Juegos en forma característica

En la teoría de juegos cooperativos, un juego se representa mediante un par (N, v), donde $N = \{1,2, \dots, n\}$ es el conjunto de jugadores y $v: 2^N \to \mathbb{R}$ es la función característica. Para cada coalición $S \subseteq N$, el valor $v(S)$ representa la utilidad máxima

que dicho grupo puede obtener si sus miembros cooperan entre sí. Por convenio, se suele imponer que $v(\emptyset) = 0$, de modo que la coalición vacía no genera valor alguno (Pérez Navarro, Jimeno Pastor, & Cerdá Tena, 2004).

Esta formulación constituye la base de los juegos cooperativos y permite cuantificar el potencial productivo de cada coalición dentro del sistema. El análisis de la función característica proporciona información tanto sobre la capacidad colectiva de los distintos subconjuntos de jugadores como sobre la contribución individual de cada jugador al valor total del juego (Casas Méndez, Fiestras Janeiro, García Jurado, & González Díaz, 2012).

En el modelo clásico de juegos cooperativos, el valor asignado a una coalición depende únicamente del conjunto de jugadores que la forman, sin tener en cuenta el orden en que estos se incorporan al grupo. Es decir, dos coaliciones que contienen los mismos jugadores generan el mismo valor, independientemente de cómo se haya constituido el grupo. Este enfoque resulta adecuado en contextos en los que la cooperación es simétrica y el orden de participación no influye en el resultado final.

No obstante, existen situaciones en las que el orden de incorporación de los jugadores puede afectar al valor generado por la cooperación. Para capturar este tipo de dinámicas, se consideran extensiones del modelo clásico basadas en coaliciones ordenadas, en las que la función característica se define sobre secuencias de jugadores en lugar de sobre subconjuntos. En este caso, dos secuencias que contienen los mismos jugadores, pero en distinto orden pueden generar valores diferentes.

Este tipo de formulación fue introducido formalmente por Nowak y Radzik (Nowak & Radzik, 1994), quienes generalizaron el concepto de juego cooperativo permitiendo que el valor asociado a una coalición dependa del orden de llegada de los jugadores. Esta extensión amplía el marco clásico de los juegos en forma característica y proporciona una herramienta teórica más flexible para el análisis de situaciones cooperativas con relaciones asimétricas entre los jugadores.

2.2.2. El valor de Shapley clásico

El valor de Shapley es uno de los conceptos fundamentales de la teoría de juegos cooperativos. Fue introducido por Lloyd Shapley en 1953 como un método axiomático para distribuir el valor total generado por la cooperación entre los jugadores de manera justa.

La idea principal del valor de Shapley es medir la contribución marginal de cada jugador a las distintas coaliciones que pueden formarse en el juego. Para ello se considera el siguiente proceso: los jugadores se incorporan a una coalición en un orden aleatorio, y cada vez que un jugador entra en la coalición aporta un incremento en el valor total. El valor de Shapley asigna a cada jugador la media de sus contribuciones marginales a lo largo de todas las posibles permutaciones del conjunto de jugadores.

Formalmente, el valor de Shapley de un jugador i en un juego (N, v)viene dado por:

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

En esta expresión, S recorre todas las coaliciones que no contienen al jugador i, el término $v(S \cup \{i\}) - v(S)$ representa la contribución marginal del jugador i cuando se une a la coalición S, y el coeficiente $\frac{|S|!\,(|N|-|S|-1)!}{|N|!}$ es la probabilidad de que, en un orden aleatorio de N, la coalición S aparezca exactamente antes que el jugador i. Donde |.| denota la cardinalidad.

El valor de Shapley se caracteriza de manera única por el cumplimiento de una serie de axiomas fundamentales (Shapley, 1953). En primer lugar, el axioma de eficiencia establece que la suma de los valores asignados a los jugadores coincide con el valor total del juego, es decir, el valor generado por la gran coalición N, que viene dado por $v(N)$. El axioma de simetría garantiza que dos jugadores que contribuyen de la misma manera a todas las coaliciones reciben la misma asignación. El axioma del jugador nulo establece que cualquier jugador cuya contribución marginal sea nula para todas las coaliciones recibe un valor de Shapley igual a cero. Por último, el axioma de aditividad asegura que el valor de Shapley es compatible con la suma de

juegos, de modo que el valor asignado en un juego compuesto es la suma de los valores asignados en cada juego individual.

Estas propiedades hacen del valor de Shapley una regla de reparto coherente, equitativa y ampliamente aceptada en el análisis de juegos cooperativos, lo que explica su uso extendido como medida de importancia relativa de los jugadores dentro de un sistema cooperativo.

La combinación de estos axiomas garantiza la unicidad del valor de Shapley, en el sentido de que no existe ninguna otra regla de asignación que los satisfaga simultáneamente.

Ejemplo 1. Para ilustrar el funcionamiento del valor de Shapley, consideremos un juego cooperativo sencillo con un conjunto de jugadores $N = \{1,2,3\}$. El juego viene descrito mediante una función característica v: $2^N \to \mathbb{R}$, definida de la siguiente forma:

- $v(\emptyset) = 0$,
- $v(\{1\}) = v(\{2\}) = v(\{3\}) = 0$,
- $v(\{1,2\}) = 10$,
- $v(\{1,3\}) = 6$,
- $v(\{2,3\}) = 4$,
- $v(\{1,2,3\}) = 12$.

Este juego representa una situación en la que ningún jugador es capaz de generar valor de manera individual, pero la cooperación entre jugadores sí permite obtener beneficios. Además, la contribución conjunta depende de qué jugadores formen la coalición.

Cálculo de $\boldsymbol{\phi_1(v)}$

Para el jugador 1, $S \subseteq N \setminus \{1\} = \{2,3\}$, es decir $S = \emptyset, \{2\}, \{3\}, \{2,3\}$. Entonces:

$$\phi_1(v) = \frac{1}{3}(v(\{1\}) - v(\emptyset)) + \frac{1}{6}(v(\{1,2\}) - v(\{2\})) + \frac{1}{6}(v(\{1,3\}) - v(\{3\})) + \frac{1}{3}(v(\{1,2,3\}) - v(\{2,3\})).$$

Sustituyendo valores:

$$\phi_1(v) = \frac{1}{3}(0-0) + \frac{1}{6}(10-0) + \frac{1}{6}(6-0) + \frac{1}{3}(12-4) = \frac{16}{3}.$$

Cálculo de $\boldsymbol{\phi_2(v)}$

Para el jugador 2, $S \subseteq N \setminus \{2\} = \{1,3\}$, es decir $S = \emptyset, \{1\}, \{3\}, \{1,3\}$. Por tanto:

$$\phi_2(v) = \frac{1}{3}(v(\{2\}) - v(\emptyset)) + \frac{1}{6}(v(\{1,2\}) - v(\{1\})) + \frac{1}{6}(v(\{2,3\}) - v(\{3\})) + \frac{1}{3}(v(\{1,2,3\}) - v(\{1,3\})).$$

Sustituyendo:

$$\phi_2(v) = \frac{1}{3}(0-0) + \frac{1}{6}(10-0) + \frac{1}{6}(4-0) + \frac{1}{3}(12-6) = \frac{13}{3}.$$

Cálculo de $\boldsymbol{\phi_3(v)}$

Para el jugador 3, $S \subseteq N \setminus \{3\} = \{1,2\}$, es decir $S = \emptyset, \{1\}, \{2\}, \{1,2\}$. Entonces:

$$\phi_3(v) = \frac{1}{3}(v(\{3\}) - v(\emptyset)) + \frac{1}{6}(v(\{1,3\}) - v(\{1\})) + \frac{1}{6}(v(\{2,3\}) - v(\{2\})) + \frac{1}{3}(v(\{1,2,3\}) - v(\{1,2\})).$$

Sustituyendo:

$$\phi_3(v) = \frac{1}{3}(0-0) + \frac{1}{6}(6-0) + \frac{1}{6}(4-0) + \frac{1}{3}(12-10) = \frac{7}{3}.$$

Comprobación (eficiencia)

$$\phi_1(v) + \phi_2(v) + \phi_3(v) = \frac{16}{3} + \frac{13}{3} + \frac{7}{3} = \frac{36}{3} = 12 = v(N).$$

2.2.3. Generalización del valor de Shapley a coaliciones ordenadas

En algunos contextos cooperativos, no basta con conocer qué jugadores forman una coalición, sino que también resulta relevante el orden en el que estos se incorporan. En este tipo de situaciones, el valor generado por una coalición puede depender no

solo de su composición, sino también de la secuencia de participación de los jugadores.

Con el fin de extender la teoría clásica de juegos cooperativos a situaciones con coaliciones ordenadas, Nowak y Radzik propusieron una generalización del valor de Shapley en 1994 (Nowak & Radzik, 1994). En esta formulación, la función característica no se define sobre subconjuntos de jugadores, sino sobre secuencias, de modo que el valor de una coalición puede depender del orden de aparición de sus miembros.

Sea $\Pi(N)$ el conjunto de todas las secuencias finitas sin repetición de jugadores. Una secuencia típica se escribe como: $T = (i_1, i_2, \dots, i_k)$, y una función característica sobre coaliciones ordenadas es una aplicación: $v: \Pi(N) \to \mathbb{R}$ Dada una secuencia T que no contiene al jugador i, la contribución marginal de $i$en el modelo de Nowak–Radzik se define como $\Delta_i(T) = v(T \oplus \{i\}) - v(T)$, donde $T \oplus i$ representa la secuencia obtenida al añadir el jugador i al final de T.

El valor generalizado de Shapley asignado al jugador i viene dado por:

$$\phi_i^{NR}(v) = \frac{1}{n!} \sum_{\pi \in \Pi(N)} \Delta_i(Pred_i(\pi))$$

En esta expresión, para una permutación $\pi \in \Pi(N)$, se denota por $Pred_i(\pi)$ la secuencia formada por todos los jugadores que preceden al jugador i en la permutación π. Es decir, $Pred_i(\pi)$ recoge el orden de llegada de los jugadores que se incorporan a la coalición antes que el jugador i. Donde π recorre todas las permutaciones de los jugadores $N \setminus \{i\}$.

De este modo, el valor generalizado de Nowak–Radzik mide la contribución media del jugador i considerando todos los posibles órdenes de llegada dentro de la coalición.

Este valor mide la contribución media del jugador i considerando todos los posibles órdenes de incorporación dentro de la coalición. Cuando la función característica no depende del orden, el valor de Nowak–Radzik coincide con el valor de Shapley clásico, lo que muestra que esta generalización incluye al modelo tradicional como un caso particular. En cambio, cuando el orden es relevante, esta formulación

permite capturar relaciones asimétricas entre los jugadores, ampliando así el alcance teórico de los juegos cooperativos.

3. Modelo de ranking predictivo aplicado al tenis

En esta sección se desarrolla la aplicación de los modelos presentados con anterioridad al tenis profesional masculino. En primer lugar, se propone un sistema de valoración de jugadores basado en herramientas de teoría de juegos cooperativos, que permite medir su importancia relativa dentro de un conjunto determinado. Posteriormente, se construye un modelo predictivo de resultados mediante técnicas de clasificación supervisada, cuyas probabilidades estimadas se utilizan para generar un ranking alternativo al oficial. Finalmente, se comparan los resultados obtenidos con el ranking ATP con el fin de analizar las diferencias y el alcance del modelo propuesto.

3.1 Medida de importancia de los jugadores basada en juegos cooperativos

A partir del enfoque propuesto por Metulini y Gnecco (2023) sobre la importancia de los jugadores, adaptamos su formulación al caso de coaliciones ordenadas de tamaño $k = 2$.

En nuestro modelo consideramos únicamente coaliciones ordenadas de tamaño $k = 2$. En consecuencia, las coaliciones previas relevantes tienen tamaño $k - 1 = 1$, es decir, conjuntos $\{ j \}$ con $j \neq i$. Como todas las coaliciones de tamaño 0 o 1 tienen valor nulo,

$$v(T) = 0 \text{ si } | T | \leq 1,$$

la contribución marginal del jugador i al incorporarse a una coalición previa $\{j\}$ se reduce a

$$\Delta_i(\{j\}) = v(j \oplus \{i\}) - v(\{j\}) = v(\{j, i\}),$$

y, de forma análoga, cuando i aparece antes que j,

$$v(\{i,j\}).$$

Dado que en el modelo considerado únicamente las coaliciones ordenadas de tamaño dos generan valor, la contribución marginal del jugador i solo es distinta de cero cuando la coalición previa tiene tamaño uno. En consecuencia, se adopta una versión normalizada del valor de Shapley generalizado, en la que el promedio se realiza únicamente sobre las coaliciones relevantes, es decir, aquellas para las que $|T| = 1$.

Bajo esta normalización, y dado que para cada pareja (i,j) existen dos órdenes posibles, el índice de importancia del jugador i se reduce a

$$I_i = \frac{1}{n-1}\sum_{j\neq i} v(\{i,j\}),$$

que en adelante se utilizará como medida de importancia del jugador i.

Esta expresión coincide con el valor de Shapley generalizado de Nowak–Radzik aplicado a coaliciones ordenadas de tamaño dos. Por tanto, en el modelo propuesto la importancia de cada jugador puede interpretarse como su probabilidad media de victoria frente al resto de jugadores del ranking.

Una vez definida la función característica y la función de importancia adaptada al caso de coaliciones ordenadas de tamaño $k = 2$, aplicamos el valor de Shapley generalizado de Nowak–Radzik para obtener, para cada jugador i,

$$\phi_i^{NR}(v) = \frac{1}{n-1}\sum_{j\neq i} v(\{i,j\})$$

Este valor proporciona una medida agregada de la fortaleza competitiva del jugador i frente al resto del conjunto considerado. En la sección de resultados se calculan estos valores para los cien primeros jugadores del ranking ATP, lo que permite construir un ranking alternativo basado en el modelo propuesto.

3.2 Ranking de jugadores de tenis basada en Elo estático y juegos cooperativos

3.2.1 Visión general del enfoque basado en Elo

En esta sección se propone un primer enfoque para valorar la fortaleza relativa de los jugadores de tenis a partir de su rendimiento competitivo. La idea central consiste en utilizar los puntos ATP como una medida agregada del nivel de cada jugador, interpretándolos como una aproximación a una puntuación Elo.

A partir de estas puntuaciones, se calculan probabilidades de victoria en enfrentamientos uno contra uno entre jugadores, de modo que una mayor diferencia de nivel se traduce en una mayor probabilidad de victoria. Estas probabilidades permiten comparar de forma directa a cualquier pareja de jugadores del ranking.

Con el fin de obtener una medida global de rendimiento, las probabilidades de victoria se integran dentro del marco de la teoría de juegos cooperativos, interpretando cada enfrentamiento como una interacción ordenada entre dos jugadores. En este contexto, el valor de Shapley generalizado se utiliza para agregar la información procedente de todos los enfrentamientos posibles y asignar a cada jugador un único valor que resume su probabilidad media de victoria frente al resto.

El resultado es un ranking de jugadores basado en el valor de Shapley, que permite analizar la relación entre esta clasificación y el ranking ATP tradicional, así como las limitaciones de un enfoque basado exclusivamente en una medida agregada del rendimiento.

3.2.2 Modelización del tenis como un juego cooperativo

En esta sección construimos el juego cooperativo que utilizaremos para medir la importancia relativa de los jugadores de tenis. El objetivo es asignar a cada jugador un valor que refleje su capacidad competitiva frente al resto, empleando para ello el valor de Shapley generalizado introducido en la sección anterior.

En nuestro modelo, el conjunto de jugadores N está formado por los tenistas seleccionados del ranking ATP. A diferencia de un juego cooperativo tradicional, donde cualquier coalición puede tener un valor, aquí solo consideramos coaliciones

ordenadas de tamaño dos, ya que una secuencia (i, j) puede interpretarse como "el jugador i compite contra el jugador j".

De este modo, la función característica únicamente asigna valor a secuencias de la forma:

$$(i, j), \quad i \neq j.$$

Todas las secuencias de longitud distinta de dos reciben valor cero, lo cual simplifica la estructura del juego y permite centrar el análisis exclusivamente en enfrentamientos directos entre jugadores, coherentes con la naturaleza individual del tenis.

3.2.3 Definición de la función característica a partir del sistema Elo

Para cuantificar el valor de una secuencia (i, j), utilizamos la probabilidad de que el jugador i venza al jugador j. Esta probabilidad se calcula empleando el sistema de puntuación Elo (Wikipedia contributors, s.f), que relaciona las diferencias de nivel entre jugadores con probabilidades de victoria.

Si R_i y R_j son las puntuaciones Elo de los jugadores i y j, respectivamente, la probabilidad de victoria de $i$sobre $j$se define como:

$$v(\{i, j\}) = \frac{1}{1 + 10^{(R_j - R_i)/400}}.$$

Esta expresión define la función característica del juego cooperativo introducido en la sección anterior, asignando a cada coalición ordenada (i, j)un valor comprendido entre 0 y 1.

Esta formulación asigna valores entre 0 y 1 a cada enfrentamiento, de modo que:

- $v(\{i, j\})$ alto indica que es probable que i gane a j,
- $v(\{j, i\})$ puede diferir de $v(\{i, j\})$, reflejando la asimetría del enfrentamiento.

3.2.4 Resultados y limitaciones del modelo basado en Elo

En la Tabla 2 se muestran los valores obtenidos del valor de Shapley generalizado para los cien primeros jugadores del ranking ATP, junto con sus correspondientes puntos ATP. Los jugadores se ordenan de mayor a menor según el valor del índice de Shapley, lo que permite comparar directamente esta clasificación con el ranking ATP tradicional.

Tabla 2. Valores del índice de Shapley generalizado y puntos ATP para los 100 primeros jugadores del ranking ATP.

JUGADOR	PUNTOS ATP	SHAPLEY	JUGADOR	PUNTOS ATP	SHAPLEY
Jannik Sinner	11830	0.999	Roberto Bautista Agut	1104	0.487
Alexander Zverev	7915	0.990	David Goffin	1037	0.448
Carlos Alcaraz	7010	0.980	Lorenzo Sonego	1026	0.441
Taylor Fritz	5100	0.966	Miomir Kecmanovic	1021	0.438
Daniil Medvedev	5030	0.963	Gael Monfils	1005	0.428
Casper Ruud	4255	0.947	Denis Shapovalov	981	0.413
Novak Djokovic	3910	0.934	Carballés Baena	981	0.413
Andrey Rublev	3760	0.926	Fabian Marozsan	935	0.383
Alex de Miñaur	3745	0.925	Arthur Rinderknech	927	0.378
Grigor Dimitrov	3350	0.904	Roman Safiullin	923	0.375
Stefanos Tsitsipas	3165	0.891	Davidovich Fokina	845	0.322
Tommy Paul	3145	0.890	Jaume Munar	832	0.313
Holger Rune	3025	0.881	Arthur Cazaux	807	0.295
Ugo Humbert	2765	0.857	Christopher O'Connell	795	0.286
Jack Draper	2685	0.849	Bu Yunchaokete	784	0.279
Hubert Hurkacz	2640	0.844	Adrian Mannarino	779	0.275
Lorenzo Musetti	2600	0.839	Alexandre Muller	778	0.274
Frances Tiafoe	2585	0.837	Aleksandar Vukic	778	0.274
Karen Khachanov	2410	0.818	Yoshihito Nishioka	776	0.273
Arthur Fils	2355	0.812	Corentin Moutet	772	0.270
Ben Shelton	2330	0.809	Zizou Bergs	768	0.267
Sebastian Korda	1985	0.770	Quentin Halys	756	0.259
Alejandro Tabilo	1943	0.764	Rinky Hijikata	746	0.252
Alexei Popyrin	1865	0.751	Thiago Seyboth Wild	732	0.242
Tomas Machac	1758	0.731	Benjamin Bonzi	730	0.241
Jordan Thompson	1745	0.728	Hugo Gaston	717	0.231
Sebastián Báez	1690	0.716	Thanasi Kokkinakis	716	0.231
Jiri Lehecka	1660	0.708	A.Shevchenko	715	0.230
Auger-Aliassime	1635	0.702	Facundo Diaz Acosta	714	0.229
F.Cerundolo	1620	0.698	Van de Zandschulp	712	0.228
Mpetshi Perricard	1561	0.681	Dusan Lajovic	710	0.227
Flavio Cobolli	1472	0.653	James Duckworth	687	0.211
Alexander Bublik	1420	0.635	Damir Dzumhur	679	0.205
Matteo Berrettini	1380	0.619	Taro Daniel	674	0.202
Nicolas Jarry	1370	0.615	Francisco Comesana	661	0.193
Nuno Borges	1355	0.609	Pavel Kotov	655	0.190
Matteo Arnaldi	1345	0.605	Gabriel Diallo	643	0.182
B.Nakashima	1335	0.601	Sebastian Ofner	643	0.182
Martin Etcheverry	1315	0.592	Daniel Altmaier	640	0.180
Tallon Griekspoor	1280	0.577	Borna Coric	639	0.179
Alex Michelsen	1245	0.560	Fabio Fognini	637	0.178
Jan-Lennard Struff	1240	0.558	Luca Nardi	637	0.178
Pedro Martínez	1205	0.541	Adam Walton	636	0.177
Luciano Darderi	1198	0.537	Otto Virtanen	634	0.176
Zhizhen Zhang	1155	0.515	Emil Ruusuvuori	628	0.172
Marcos Giron	1150	0.512	Yannick Hanfmann	627	0.172

Mariano Navone	1148	0.511	Camilo Ugo Carabelli	624	0.170
Jakub Mensik	1136	0.505	Sumit Nagal	622	0.168
Cameron Norrie	1119	0.495	Jacob Fearnley	622	0.168
Juncheng Shang	1115	0.493	Federico Coria	617	0.165

Los resultados muestran que el ranking obtenido mediante el valor de Shapley coincide prácticamente con el ranking ATP original. Este comportamiento es esperable, ya que en el modelo propuesto se emplean los puntos ATP como si fueran puntuaciones Elo. Dado que la fórmula Elo asigna una probabilidad de victoria estrictamente creciente con la diferencia de puntuaciones, un jugador situado por encima en el ranking ATP tiene siempre una probabilidad mayor de vencer a cualquier jugador situado por debajo.

Como consecuencia, el valor de Shapley generalizado, al basarse en la probabilidad media de victoria frente al resto de jugadores, preserva el orden inducido por el ranking ATP. En este sentido, el modelo no altera la clasificación existente, sino que la reafirma desde un punto de vista probabilístico.

Este resultado pone de manifiesto una limitación del enfoque basado exclusivamente en Elo: al depender únicamente de una medida agregada del rendimiento, no incorpora información adicional relevante, como el contexto de los partidos, la superficie o el historial reciente.

Esta observación motiva la introducción, en la siguiente sección, de un modelo de clasificación supervisada que permita enriquecer la predicción de resultados y evaluar en qué medida la inclusión de nuevas variables mejora el poder explicativo del modelo.

3.3 Ranking de jugadores de tenis basada en clasificación supervisada y teoría de juegos.

En esta sección se analiza en qué medida la inclusión de variables de rendimiento propias de los partidos de tenis permite mejorar la capacidad predictiva del modelo frente a un enfoque basado únicamente en el sistema Elo. Para ello, se construye un modelo de clasificación supervisada basado en regresión logística, cuyo objetivo es

estimar la probabilidad de victoria de un jugador en un enfrentamiento determinado del circuito profesional de tenis.

A diferencia del sistema Elo, que resume el nivel de los jugadores mediante una puntuación agregada, el modelo de clasificación supervisada permite incorporar información adicional procedente de las estadísticas de los partidos, proporcionando así una visión más detallada del rendimiento de los jugadores. Este enfoque permitirá, en secciones posteriores, comparar los rankings obtenidos mediante ambos métodos y analizar sus diferencias.

3.3.1 Construcción del conjunto de datos

El análisis se basa en un conjunto de datos construido a partir de enfrentamientos disputados entre jugadores del circuito ATP durante los años 2021, 2022 y 2023. Para cada jugador y cada año se dispone de estadísticas agregadas de rendimiento, obtenidas a partir de partidos oficiales disputados en torneos del ATP Tour y Grand Slams, excluyéndose exhibiciones, *walkovers* y encuentros finalizados por retirada. Estas restricciones permiten garantizar la homogeneidad y comparabilidad de los datos utilizados.

Cada observación del conjunto de datos corresponde a un enfrentamiento concreto entre dos jugadores en un año determinado. La variable respuesta recoge únicamente el resultado del partido, mientras que las variables explicativas asociadas a cada jugador corresponden a sus estadísticas agregadas a lo largo de la temporada completa en la que se disputó dicho enfrentamiento. De este modo, el modelo utiliza información representativa del nivel medio de los jugadores en ese periodo temporal, sin incorporar información específica del partido individual, evitando así posibles problemas de dependencia o filtración de información del resultado.

Con el fin de diferenciar a los dos jugadores en cada observación, se adopta una convención técnica según la cual uno de ellos se denomina *jugador local* y el otro *jugador visitante*. Esta distinción no tiene interpretación deportiva, ya que en el tenis no existe ventaja asociada a jugar como local o visitante, y se introduce únicamente

para organizar las variables explicativas del modelo y facilitar la formulación del modelo de clasificación.

La variable respuesta se define de forma binaria como

$$Resultado = \begin{cases} 0, & \text{si gana el jugador local,} \\ 1, & \text{si gana el jugador visitante.} \end{cases}$$

3.3.2 Variables explicativas

Las variables explicativas consideradas en el modelo completo se construyen a partir de estadísticas agregadas anuales de rendimiento de los jugadores. Estas variables capturan distintas dimensiones del rendimiento competitivo, incluyendo el desempeño al servicio, el desempeño al resto y el rendimiento global en el partido.

Con el fin de clarificar el significado de cada variable, en la Tabla 3 se presenta el conjunto completo de variables explicativas empleadas en el modelo inicial, junto con su interpretación y la dimensión del juego a la que pertenecen.

Tabla 3. Variables explicativas del modelo completo y su interpretación según la dimensión del juego.

Variable	Descripción	Dimensión del juego
ps1	% primer servicio	Servicio
pg1s	% puntos ganados con 1er servicio	Servicio
pg2s	% puntos ganados con 2º servicio	Servicio
prs	% puntos de rotura salvados	Servicio
jgs	% juegos ganados al saque	Servicio
pts	% puntos totales ganados al saque	Servicio
pr1	% puntos ganados restando 1er servicio	Resto
pr2	% puntos ganados restando 2º servicio	Resto
prc	% puntos de rotura convertidos	Resto
jgr	% juegos al resto ganados	Resto
ptr	% puntos ganados al resto	Resto
ptt	% total de puntos ganados	Rendimiento global

Las variables anteriores se incluyen para ambos jugadores en cada enfrentamiento, diferenciando entre jugador local y jugador visitante, y utilizando siempre la información correspondiente a la temporada en la que se disputó el partido. De este modo, el modelo captura diferencias relativas de rendimiento entre los dos jugadores enfrentados en un mismo periodo temporal, lo que resulta especialmente relevante para la comparación directa entre rivales en el contexto de un modelo de clasificación supervisada.

Con el objetivo de analizar la posible existencia de colinealidad y redundancia informativa entre las variables explicativas, se calcula la matriz de correlaciones del conjunto completo de variables. Dado que la matriz completa resulta extensa, en la Tabla 4 se presentan únicamente aquellas correlaciones con valor absoluto superior a 0.85, que evidencian relaciones lineales de elevada intensidad entre métricas que capturan dimensiones similares del juego.

Tabla 4. Correlaciones con valor absoluto superior a 0.85 entre las variables explicativas del modelo completo.

Variable1	Variable2	Correlación
jgs_c	pts_c	0.97
jgr_c	ptr_c	0.97
pr1_f	ptr_f	0.96
jgr_f	ptr_f	0.96
pr1_f	jgr_f	0.95
pr1_c	jgr_c	0.94
pr1_c	ptr_c	0.94
pr2_c	ptr_c	0.93
jgs_f	pts_f	0.92
pr2_c	jgr_c	0.91
pr2_f	ptr_f	0.90
pg1s_f	jgs_f	0.89
pg1s_c	jgs_c	0.88
pg1s_c	pts_c	0.88
pr2_f	jgr_f	0.88
pg1s_f	pts_f	0.87

Los resultados muestran la existencia de bloques de variables altamente correlacionadas, particularmente dentro de las métricas asociadas al rendimiento al saque (por ejemplo, variables relacionadas con juegos y puntos ganados al servicio) y al rendimiento al resto. Estas correlaciones elevadas indican que varias variables capturan esencialmente dimensiones similares del rendimiento, lo que puede generar inestabilidad en la estimación de los coeficientes del modelo de regresión logística.

Por este motivo, se selecciona un subconjunto representativo de variables, eliminando aquellas que aportan información redundante y conservando únicamente las que capturan de forma más estable las principales dimensiones del rendimiento. En concreto, el modelo final incorpora las variables ps1, pg2s, prs, prc, ptr y ptt (para ambos jugadores), seleccionadas por representar de manera estable el rendimiento al servicio, el rendimiento al resto y el rendimiento global, evitando la inclusión simultánea de métricas altamente correlacionadas.

3.3.3 Especificación del modelo

Para modelizar el resultado de un enfrentamiento entre dos jugadores se emplea un modelo de regresión logística, adecuado para problemas de clasificación binaria. El modelo relaciona la probabilidad de que gane el *jugador visitante* con un conjunto de variables explicativas que recogen las características de ambos jugadores en el año correspondiente al enfrentamiento.

Formalmente, el modelo se expresa como:

$$\mathbb{P}(\text{Resultado} = 1 \mid \text{x}) = \frac{1}{1 + \exp\left(-(\beta_0 + \beta^\top \text{x})\right)}$$

donde x representa el vector de variables explicativas asociadas a ambos jugadores en el enfrentamiento, diferenciadas según la convención de jugador local y jugador visitante y β el vector de parámetros a estimar.

El modelo se estima mediante el método de máxima verosimilitud utilizando el software R, lo que permite obtener estimaciones consistentes de los parámetros asociados a cada variable explicativa. La selección final de las variables se realiza atendiendo a criterios de estabilidad del modelo y reducción de la colinealidad,

teniendo en cuenta también su significación estadística, con el fin de evitar problemas de sobreajuste.

Los coeficientes estimados permiten interpretar el efecto relativo de cada variable sobre la probabilidad de victoria del jugador visitante, manteniendo constantes el resto de factores.

Una vez estimado, el modelo permite obtener probabilidades de victoria para enfrentamientos entre jugadores, que se utilizarán en la sección siguiente para evaluar el rendimiento predictivo del enfoque propuesto y comparar sus resultados con los obtenidos mediante el sistema Elo.

3.3.4 Predicción de enfrentamientos

Una vez estimado el modelo de regresión logística, este permite calcular probabilidades de victoria para enfrentamientos entre jugadores del conjunto considerado. Para cada partido del conjunto de datos, el modelo proporciona una estimación de la probabilidad de que el resultado tome el valor 1, es decir, que gane el *jugador visitante*, en función de las estadísticas anuales de ambos jugadores correspondientes al año en el que se disputó el encuentro.

Adicionalmente, el modelo puede aplicarse a enfrentamientos hipotéticos entre jugadores, utilizando estadísticas agregadas de una temporada concreta, con el fin de estimar probabilidades de victoria en duelos que no necesariamente han tenido lugar en el conjunto de datos original.

Estas probabilidades pueden interpretarse como una medida continua del grado de superioridad relativa de un jugador frente a su rival, proporcionando una información más rica que una predicción binaria del resultado.

Estas probabilidades estimadas se utilizan en la sección siguiente para evaluar el rendimiento predictivo del modelo y comparar sus resultados con los obtenidos mediante el enfoque basado exclusivamente en el sistema Elo, analizando en qué medida la inclusión de variables adicionales permite mejorar la calidad de las predicciones.

3.3.5 Resultados de la clasificación supervisada

En esta sección se presentan los resultados obtenidos a partir del modelo de clasificación supervisada basado en regresión logística descrito en el apartado anterior. El objetivo es evaluar su capacidad predictiva y compararla con el enfoque basado exclusivamente en el sistema Elo.

En la Tabla 5 se recogen las probabilidades estimadas de victoria para los enfrentamientos considerados entre los jugadores analizados, ordenadas de mayor a menor. Estas probabilidades constituyen el resultado principal del modelo de clasificación supervisada y permiten comparar de forma directa la fortaleza relativa de los jugadores en enfrentamientos uno contra uno. Conviene destacar que el modelo no genera un ranking global de forma directa, sino probabilidades de victoria específicas para cada enfrentamiento, a partir de las cuales puede inferirse una jerarquía relativa entre jugadores.

Tabla 5. Probabilidades estimadas de victoria en enfrentamientos uno contra uno entre los jugadores analizados.

Jugador i	Jugador j	Probabilidad estimada de victoria
Sinner	Medvedev	0.924
Zverev	Medvedev	0.901
Sinner	Tsitsipas	0.895
Sinner	Alcaraz	0.888
Sinner	Rublev	0.884
Zverev	Tsitsipas	0.864
Zverev	Alcaraz	0.856
Zverev	Rublev	0.850
Sinner	Djokovic	0.848
Alcaraz	Medvedev	0.840
Sinner	Zverev	0.835
Djokovic	Medvedev	0.828
Zverev	Djokovic	0.807
Alcaraz	Tsitsipas	0.786
Djokovic	Tsitsipas	0.770
Alcaraz	Rublev	0.766
Djokovic	Alcaraz	0.758
Sinner	Fritz	0.756
Djokovic	Rublev	0.750
Fritz	Medvedev	0.744
Alcaraz	Djokovic	0.707
Zverev	Sinner	0.699
Zverev	Fritz	0.699
Alcaraz	Zverev	0.687
Fritz	Tsitsipas	0.670
Djokovic	Zverev	0.667
Fritz	Alcaraz	0.655
Fritz	Rublev	0.645

Alcaraz	Sinner	0.573
Alcaraz	Fritz	0.573
Fritz	Djokovic	0.572
Djokovic	Sinner	0.551
Djokovic	Fritz	0.551
Fritz	Zverev	0.549
Rublev	Medvedev	0.511
Fritz	Sinner	0.427
Rublev	Tsitsipas	0.422
Rublev	Alcaraz	0.406
Tsitsipas	Medvedev	0.388
Rublev	Djokovic	0.325
Rublev	Zverev	0.304
Tsitsipas	Alcaraz	0.293
Tsitsipas	Rublev	0.283
Medvedev	Tsitsipas	0.273
Medvedev	Alcaraz	0.260
Medvedev	Rublev	0.251
Tsitsipas	Djokovic	0.225
Rublev	Sinner	0.211
Rublev	Fritz	0.211
Tsitsipas	Zverev	0.209
Medvedev	Djokovic	0.198
Medvedev	Zverev	0.184
Tsitsipas	Sinner	0.139
Tsitsipas	Fritz	0.139
Medvedev	Sinner	0.121
Medvedev	Fritz	0.121

A partir de estas probabilidades estimadas, se define un juego cooperativo que permita obtener una medida agregada de la fortaleza competitiva de los jugadores. En el apartado siguiente se construye dicho juego y se calcula el correspondiente ranking mediante el valor de Shapley generalizado.

3.3.6 Ranking obtenido a partir del modelo de clasificación supervisada

A partir de las probabilidades estimadas de victoria obtenidas mediante el modelo de regresión logística, se construye un juego cooperativo con coaliciones ordenadas de tamaño dos. En este juego, cada enfrentamiento directo entre dos jugadores (i, j) se interpreta como una coalición ordenada, a la que se asigna como valor la probabilidad estimada de que el jugador i venza al jugador j.

Formalmente, se define la función característica del juego como

$$v(\{i, j\}) = \mathbb{P}(\text{el jugador } i \text{ gana a } j),$$

donde estas probabilidades se corresponden con los valores recogidos en la Tabla 5. Por ejemplo, de acuerdo con los resultados del modelo, se tiene $v(\text{Medvedev,Sinner}) = 0.121$, mientras que $v(\text{Sinner,Medvedev}) = 0.924$, lo que refleja la naturaleza asimétrica del enfrentamiento.

El cálculo de este índice propuesto para cada jugador permite construir un ranking agregado que resume su fortaleza competitiva global a partir de los enfrentamientos directos. En la Tabla 6 se presentan los valores obtenidos del índice de Shapley generalizado para los jugadores considerados, ordenados de mayor a menor, dando lugar al ranking derivado del modelo de clasificación supervisada.

Tabla 6. Ranking derivado del índice de Shapley generalizado aplicado a las probabilidades estimadas por el modelo supervisado.

Ranking	Jugador	Shapley
1	Sinner	0.861
2	Zverev	0.811
3	Alcaraz	0.705
4	Djokovic	0.696
5	Fritz	0.609
6	Rublev	0.341
7	Tsitsipas	0.239
8	Medvedev	0.201

Este ranking proporciona una visión sintética del rendimiento relativo de los jugadores basada exclusivamente en las probabilidades estimadas por el modelo de clasificación supervisada. Los valores obtenidos reflejan la probabilidad media de victoria de cada jugador frente al resto del conjunto considerado, permitiendo identificar de forma agregada diferencias en la fortaleza competitiva que no se desprenden directamente de un ranking acumulativo. En la sección siguiente se analizan estos resultados en comparación con el sistema Elo y con el ranking ATP oficial.

3.4 Comparación de resultados

En esta sección se comparan los resultados obtenidos mediante el modelo de clasificación supervisada con distintas referencias externas, con el objetivo de analizar su capacidad para reflejar el rendimiento real de los jugadores y evaluar sus diferencias con otros enfoques de ranking. En particular, se realiza una comparación con el ranking ATP oficial correspondiente al año 2024, así como un análisis comparativo con los resultados derivados del sistema Elo.

3.4.1 Comparación con el ranking ATP en 2024

En este apartado se analiza la concordancia entre la clasificación obtenida mediante el modelo de clasificación supervisada y el ranking ATP oficial a lo largo del año 2024. Para ello, se consideran distintos momentos temporales separados por intervalos de tres meses, lo que permite estudiar la evolución del ranking y evaluar la capacidad del modelo para adaptarse a los cambios en el rendimiento de los jugadores.

Con el fin de facilitar la interpretación, el análisis se centra en los tres primeros jugadores del ranking en cada periodo considerado, comparando sus posiciones relativas con las estimaciones derivadas del modelo de clasificación supervisada.

Conviene señalar que el modelo de clasificación supervisada no proporciona un ranking global explícito, sino probabilidades de victoria en enfrentamientos directos, a partir de las cuales se infiere la fortaleza relativa de los jugadores.

1 de enero de 2024

A comienzos de enero de 2024, el ranking ATP oficial situaba a Novak Djokovic, Carlos Alcaraz y Daniil Medvedev en las tres primeras posiciones. Este ranking refleja principalmente el rendimiento acumulado por los jugadores durante la temporada anterior.

Según los resultados del modelo de clasificación supervisada, basados en las probabilidades estimadas de victoria en enfrentamientos directos, se observa que Jannik Sinner y Alexander Zverev presentan una elevada fortaleza relativa frente a Daniil Medvedev, con probabilidades de victoria claramente superiores al 0.85 en

dichos enfrentamientos. Asimismo, el modelo asigna probabilidades altas de victoria a Djokovic y Alcaraz frente a la mayoría de los jugadores considerados, en línea con su posición destacada en el ranking ATP durante este periodo.

1 de abril de 2024

A comienzos de abril de 2024, el ranking ATP oficial situaba a Novak Djokovic, Jannik Sinner y Carlos Alcaraz en las tres primeras posiciones. Este cambio respecto al inicio de la temporada refleja el impacto de los torneos disputados durante el primer trimestre del año, en los que Sinner mejora notablemente su posición en el ranking oficial.

Según los resultados del modelo de clasificación supervisada, las probabilidades estimadas de victoria indican una fortaleza relativa elevada de Jannik Sinner frente a la mayoría de los jugadores considerados, incluidos rivales situados en posiciones cercanas del ranking ATP. Asimismo, el modelo sigue asignando probabilidades altas de victoria a Djokovic y Alcaraz frente a gran parte del resto de jugadores, lo que resulta coherente con su presencia en las primeras posiciones del ranking oficial en este periodo.

En conjunto, en este momento de la temporada se observa una mayor concordancia entre el ranking ATP y las estimaciones del modelo de clasificación supervisada, especialmente en lo relativo a la consolidación de Sinner entre los jugadores con mayor fortaleza competitiva.

1 de julio de 2024

A comienzos de julio de 2024, el ranking ATP oficial situaba a Jannik Sinner, Novak Djokovic y Carlos Alcaraz en las tres primeras posiciones. Este cambio supone un punto de inflexión respecto a los meses anteriores, ya que Sinner alcanza el primer puesto del ranking tras los resultados obtenidos durante la primera mitad de la temporada, especialmente en los torneos disputados sobre hierba y pista dura.

Según los resultados del modelo de clasificación supervisada, las probabilidades estimadas de victoria refuerzan la fortaleza competitiva de Jannik Sinner, que presenta valores elevados frente a la mayoría de los jugadores considerados, incluidos Djokovic y Alcaraz. Asimismo, el modelo continúa asignando

probabilidades altas de victoria a Djokovic y Alcaraz frente al resto de rivales, lo que resulta coherente con su permanencia en las primeras posiciones del ranking ATP.

En este periodo se observa una clara alineación entre el ranking ATP y las estimaciones del modelo de clasificación supervisada, ya que el liderazgo de Sinner, sugerido por el modelo en fases anteriores de la temporada, se consolida también en la clasificación oficial.

30 de septiembre de 2024

A finales de septiembre de 2024, el ranking ATP oficial situaba a Jannik Sinner, Carlos Alcaraz y Alexander Zverev en las tres primeras posiciones. Este ranking refleja los resultados acumulados durante la mayor parte de la temporada, una vez disputados los principales torneos del calendario, incluidos los Grand Slams y los Masters 1000.

Según los resultados del modelo de clasificación supervisada, las probabilidades estimadas de victoria continúan mostrando una fortaleza relativa elevada de Jannik Sinner, que mantiene valores altos frente a la mayoría de los jugadores considerados. Asimismo, el modelo asigna probabilidades competitivas a Carlos Alcaraz y Alexander Zverev frente al resto de rivales, lo que resulta coherente con su presencia en las primeras posiciones del ranking ATP en este tramo final de la temporada.

En este periodo se aprecia una elevada concordancia entre el ranking ATP y las estimaciones del modelo de clasificación supervisada, lo que sugiere que, a medida que avanza la temporada y se acumulan resultados, ambas aproximaciones tienden a alinearse en la identificación de los jugadores con mayor fortaleza competitiva.

30 de diciembre de 2024

A finales de diciembre de 2024, el ranking ATP oficial situaba a Jannik Sinner, Alexander Zverev y Carlos Alcaraz en las tres primeras posiciones. Este ranking corresponde al cierre de la temporada y refleja el rendimiento acumulado de los jugadores a lo largo de todo el año competitivo.

Según los resultados del modelo de clasificación supervisada, las probabilidades estimadas de victoria confirman una fortaleza competitiva especialmente elevada de Jannik Sinner y Alexander Zverev, ambos con valores altos frente a la mayoría de los jugadores considerados. Asimismo, el modelo asigna una elevada fortaleza relativa

a Carlos Alcaraz, lo que resulta coherente con la consolidación de estos jugadores en las primeras posiciones del ranking ATP al final de la temporada.

En este punto, las estimaciones del modelo de clasificación supervisada y el ranking ATP muestran una alta concordancia, lo que indica que, una vez finalizada la temporada y acumulados los resultados, ambas aproximaciones identifican de forma similar a los jugadores con mayor rendimiento global.

En conjunto, las comparaciones realizadas a lo largo de 2024 ponen de manifiesto que el modelo de clasificación supervisada, al basarse en estadísticas agregadas de rendimiento, puede reflejar de forma más inmediata el nivel competitivo relativo de los jugadores, mientras que el ranking ATP incorpora resultados acumulados a más largo plazo. En este sentido, el modelo muestra una mayor sensibilidad a los cambios en el rendimiento, anticipando en algunos casos modificaciones en la jerarquía competitiva que se consolidan progresivamente a medida que avanza la temporada.

3.4.2 Comparación entre el sistema Elo y la clasificación supervisada

En este apartado se comparan los resultados obtenidos mediante el sistema Elo y el modelo de clasificación supervisada, ambos integrados dentro del mismo marco de juegos cooperativos con coaliciones ordenadas de tamaño dos y utilizando el valor de Shapley generalizado como medida agregada de la fortaleza competitiva de los jugadores.

En el enfoque basado en Elo, la función característica del juego se define a partir de las puntuaciones ATP de los jugadores, lo que da lugar a probabilidades de victoria estrictamente crecientes con la diferencia de ranking. Al aplicar el valor de Shapley a estas probabilidades, se obtiene un ranking (Tabla 2) que reproduce prácticamente el orden del ranking ATP oficial a finales de 2024. Los valores del índice de Shapley se concentran en niveles elevados para los jugadores mejor posicionados, reflejando una jerarquía competitiva estable y fuertemente condicionada por la acumulación de resultados a largo plazo.

Por su parte, en el modelo de clasificación supervisada la función característica se construye a partir de probabilidades estimadas mediante regresión logística, que incorporan información adicional procedente de estadísticas de juego agregadas de

los partidos. Al aplicar el mismo juego cooperativo y el mismo valor de Shapley a estas probabilidades, se obtiene un ranking alternativo (Tabla 6) que resume la probabilidad media de victoria de cada jugador frente al resto del conjunto considerado.

La comparación entre ambos rankings pone de manifiesto diferencias relevantes. Mientras que jugadores como Jannik Sinner y Alexander Zverev mantienen posiciones destacadas en ambos enfoques, otros como Daniil Medvedev o Stefanos Tsitsipas presentan descensos más acusados en el ranking derivado de la clasificación supervisada. Estas variaciones reflejan que el modelo supervisado es más sensible a las desventajas sistemáticas observadas en los enfrentamientos directos, incluso cuando el jugador mantiene una posición elevada en el ranking ATP.

En conjunto, el sistema Elo aplicado dentro del marco cooperativo proporciona una visión más estable y agregada de la jerarquía competitiva, estrechamente ligada a la acumulación de puntos ATP. En cambio, la clasificación supervisada integrada en el mismo juego cooperativo ofrece una representación más sensible a las características reales del juego, permitiendo discriminar con mayor precisión diferencias relativas en la probabilidad de victoria entre jugadores. En este sentido, ambos enfoques resultan complementarios, si bien el modelo de clasificación supervisada aporta una mayor capacidad explicativa en el análisis de enfrentamientos individuales y en la detección de diferencias de rendimiento no capturadas por una medida puramente acumulativa.

4. Conclusiones

En este Trabajo de Fin de Grado se ha analizado el problema de la clasificación y comparación del rendimiento de jugadores profesionales de tenis desde una perspectiva cuantitativa, combinando herramientas estadísticas con conceptos procedentes de la teoría de juegos cooperativos. El objetivo principal ha sido evaluar la capacidad de distintos enfoques de ranking y predicción para describir la jerarquía competitiva del circuito y estimar la probabilidad de victoria en enfrentamientos individuales, más allá de las clasificaciones tradicionales.

En primer lugar, se ha aplicado un enfoque inspirado en el sistema Elo e integrado en el marco de los juegos cooperativos, utilizando el valor de Shapley para asignar una puntuación agregada a cada jugador a partir de sus puntos ATP. Este procedimiento permite construir un ranking global que refleja la jerarquía competitiva acumulada en el circuito profesional. En el contexto del tenis, este enfoque resulta especialmente útil para identificar de forma estable a los jugadores dominantes a largo plazo. No obstante, al basarse en una medida agregada del rendimiento, tiende a amplificar las diferencias entre jugadores y no siempre se traduce directamente en diferencias equivalentes en la probabilidad de victoria en un partido concreto.

En segundo lugar, se ha desarrollado un modelo de clasificación supervisada basado en regresión logística, estimado a partir de estadísticas agregadas de rendimiento correspondientes a los años 2021–2023 y aplicado posteriormente a datos de la temporada 2024. Este modelo permite calcular probabilidades de victoria en enfrentamientos directos, incorporando información adicional sobre el rendimiento en pista que no está presente en los rankings tradicionales. A diferencia del sistema Elo, este enfoque no produce inicialmente un ranking explícito, sino probabilidades asociadas a cada enfrentamiento, lo que lo hace especialmente adecuado para el análisis detallado de duelos individuales.

A partir de estas probabilidades estimadas, se ha construido un juego cooperativo con coaliciones ordenadas de tamaño dos y se ha aplicado el valor de Shapley generalizado de Nowak–Radzik para obtener un ranking agregado de jugadores. Este ranking resume la probabilidad media de victoria de cada jugador frente al resto del conjunto considerado y permite una comparación directa con el ranking obtenido mediante el sistema Elo. Los resultados muestran que, aunque los jugadores mejor posicionados tienden a coincidir en ambos enfoques, la clasificación supervisada introduce variaciones relevantes que reflejan de manera más directa las desventajas y ventajas observadas en los enfrentamientos reales.

La comparación entre ambos modelos pone de manifiesto que no existe un único criterio óptimo para clasificar jugadores, sino que cada enfoque responde a objetivos distintos. Mientras que el sistema Elo integrado en el marco cooperativo proporciona una visión global, estable y agregada de la jerarquía competitiva, la

clasificación supervisada ofrece una perspectiva más sensible y detallada de las probabilidades de victoria en enfrentamientos individuales, evitando sobreestimar ventajas derivadas exclusivamente de la acumulación de puntos ATP. En este sentido, ambos enfoques deben entenderse como complementarios, y su utilización conjunta permite obtener una comprensión más rica y matizada del rendimiento de los jugadores.

Desde un punto de vista aplicado, este trabajo ilustra cómo herramientas estadísticas y conceptos de la teoría de juegos pueden emplearse de forma conjunta para analizar fenómenos deportivos reales, aportando información adicional a la proporcionada por las clasificaciones tradicionales. El enfoque desarrollado demuestra que es posible combinar modelos agregados de ranking con modelos probabilísticos de predicción para estudiar el tenis profesional desde una perspectiva cuantitativa más completa.

Como limitaciones del estudio, cabe señalar que el ranking Elo/Shapley está condicionado por la disponibilidad y la fecha de los puntos ATP utilizados, y que el modelo de regresión logística se apoya en estadísticas agregadas anuales, sin incorporar factores contextuales como la superficie de juego, la forma reciente, la fatiga o las características específicas de cada torneo. Asimismo, en la construcción del conjunto de datos se han excluido los partidos finalizados por retirada de alguno de los jugadores, con el objetivo de evitar sesgos derivados de resultados no estrictamente deportivos, lo que reduce el tamaño muestral, pero mejora la coherencia del análisis.

Como líneas futuras de investigación, sería interesante extender el modelo incorporando variables contextuales adicionales, recalcular los rankings en marcos temporales homogéneos y explorar modelos dinámicos que actualicen las probabilidades de victoria a lo largo de la temporada, combinando la estabilidad de los rankings con la capacidad predictiva de los modelos supervisados.

5. Referencias

ATP Tour. (s.f.). *ATP Rankings and Statistics.* Obtenido de ATP Tour: https://www.atptour.com

Casas Méndez, B., Fiestras Janeiro, M., García Jurado, I., & González Díaz, J. (2012). *Introducción a la teoría de juegos.* Universidade de Santiago de Compostela.

James, G., Witten, D., Hastie, T., & Tibshirani, R. (2021). *An Introduction to Statistical Learning.* New York: Springer.

Metulini, R., & Gnecco, G. (2023). Measuring players' importance in basketball using the generalized Shapley value. *Annals of Operations Research, 325*(1), 441–465.

Nowak, A. S., & Radzik, T. (1994). A generalized Shapley value for cooperative games. *International Journal of Game Theory, 23*, 289–299.

Pérez Navarro, J., Jimeno Pastor, J., & Cerdá Tena, E. (2004). *Teoría de juegos.* Pearson Educación.

Shapley, L. S. (1953). A value for n-person games. En H. W. Kuhn, *Contributions to the Theory of Games II* (págs. 307–317). Princeton: Princeton University Press.

Wikipedia contributors. (s.f). *Sistema de puntuación Elo*. Obtenido de Wikipedia: https://es.wikipedia.org/wiki/Sistema_de_puntuaci%C3%B3n_Elo

Apéndice A. Scripts en R

En este apéndice se recogen los scripts completos en lenguaje R utilizados para la implementación de los modelos y cálculos desarrollados en el trabajo. En particular, se incluyen los códigos empleados para el cálculo del ranking basado en el sistema Elo integrado en un marco de juegos cooperativos mediante el valor de Shapley, así como el script correspondiente al modelo de clasificación supervisada y a la obtención del ranking derivado de las probabilidades estimadas. Estos scripts permiten reproducir íntegramente los resultados presentados en el cuerpo del trabajo.

A.1 Cálculo del ranking Elo–Shapley

En esta sección se presenta el script en R utilizado para el cálculo del ranking de jugadores basado en el sistema Elo y el valor de Shapley generalizado.

```
############################################################

# TFG – Valoración de jugadores mediante Elo y valor de Shapley

#

# En este script se emplean los puntos ATP como proxy de

# una puntuación Elo (rating) para obtener probabilidades

# de victoria mediante la fórmula estándar de Elo.

############################################################


#------------------------------------------------------------

# 1) Datos: lista de jugadores y puntos ATP

#------------------------------------------------------------

jugadores <- c(

 "Jannik Sinner","Alexander Zverev","Carlos Alcaraz","Taylor Fritz","Daniil Medvedev",
```

"Casper Ruud","Novak Djokovic","Andrey Rublev","Alex de Miñaur","Grigor Dimitrov",

"Stefanos Tsitsipas","Tommy Paul","Holger Rune","Ugo Humbert","Jack Draper",

"Hubert Hurkacz","Lorenzo Musetti","Frances Tiafoe","Karen Khachanov","Arthur Fils",

"Ben Shelton","Sebastian Korda","Alejandro Tabilo","Alexei Popyrin","Tomas Machac",

"Jordan Thompson","Sebastián Báez","Jiri Lehecka","Felix Auger-Aliassime","Francisco Cerundolo",

"Giovanni Mpetshi Perricard","Flavio Cobolli","Alexander Bublik","Matteo Berrettini","Nicolas Jarry",

"Nuno Borges","Matteo Arnaldi","Brandon Nakashima","Tomas Martin Etcheverry","Tallon Griekspoor",

"Alex Michelsen","Jan-Lennard Struff","Pedro Martínez","Luciano Darderi","Zhizhen Zhang",

"Marcos Giron","Mariano Navone","Jakub Mensik","Cameron Norrie","Juncheng Shang",

"Roberto Bautista Agut","David Goffin","Lorenzo Sonego","Miomir Kecmanovic","Gael Monfils",

"Denis Shapovalov","Roberto Carballés Baena","Fabian Marozsan","Arthur Rinderknech","Roman Safiullin",

"Alejandro Davidovich Fokina","Jaume Munar","Arthur Cazaux","Christopher O'Connell","Bu Yunchaokete",

"Adrian Mannarino","Alexandre Muller","Aleksandar Vukic","Yoshihito Nishioka","Corentin Moutet",

"Zizou Bergs","Quentin Halys","Rinky Hijikata","Thiago Seyboth Wild","Benjamin Bonzi",

```
 "Hugo Gaston","Thanasi Kokkinakis","Alexander Shevchenko","Facundo Diaz Acosta","Botic Van de Zandschulp",

 "Dusan Lajovic","James Duckworth","Damir Dzumhur","Taro Daniel","Francisco Comesana",

 "Pavel Kotov","Gabriel Diallo","Sebastian Ofner","Daniel Altmaier","Borna Coric",

 "Fabio Fognini","Luca Nardi","Adam Walton","Otto Virtanen","Emil Ruusuvuori",

 "Yannick Hanfmann","Camilo Ugo Carabelli","Sumit Nagal","Jacob Fearnley","Federico Coria"

)

puntos <- c(

 11830,7915,7010,5100,5030,

 4255,3910,3760,3745,3350,

 3165,3145,3025,2765,2685,

 2640,2600,2585,2410,2355,

 2330,1985,1943,1865,1758,

 1745,1690,1660,1635,1620,

 1561,1472,1420,1380,1370,

 1355,1345,1335,1315,1280,

 1245,1240,1205,1198,1155,

 1150,1148,1136,1119,1115,

 1104,1037,1026,1021,1005,

 981,981,935,927,923,

 845,832,807,795,784,

 779,778,778,776,772,

 768,756,746,732,730,
```

```
 717,716,715,714,712,
 710,687,679,674,661,
 655,643,643,640,639,
 637,637,636,634,628,
 627,624,622,622,617
)
# Construimos un data.frame con jugadores y puntos
atp <- data.frame(Jugadores = jugadores, Puntos = puntos, stringsAsFactors = FALSE)
# Comprobaciones básicas de consistencia (mismo número de jugadores y puntos)
length(jugadores)
length(puntos)
nrow(atp)

#------------------------------------------------------------
# 2) Matriz de diferencias de rating: D[i,j] = R_j - R_i
#------------------------------------------------------------
# outer crea una matriz combinando todos los pares (Ri, Rj).
# Aquí se define D[i,j] = Rj - Ri, que es lo que aparece en Elo.
D <- outer(atp$Puntos, atp$Puntos, FUN = function(Ri, Rj) Rj - Ri)

#------------------------------------------------------------
# 3) Probabilidades tipo Elo
#------------------------------------------------------------
```

```
# Fórmula estándar de Elo:
#   v(i,j) = 1 / (1 + 10^((R_j - R_i)/400))
# Devuelve una matriz P donde P[i,j] es la probabilidad de que
# el jugador i gane al jugador j.
P <- 1 / (1 + 10^(D / 400))
# La diagonal no tiene sentido (jugador contra sí mismo)
diag(P) <- NA
# Etiquetamos filas y columnas con nombres de jugadores
rownames(P) <- atp$Jugadores
colnames(P) <- atp$Jugadores

#------------------------------------------------------------
# 4) Valor de Shapley (generalizado) en el caso k = 2
#------------------------------------------------------------
# En el modelo del TFG (coaliciones ordenadas de tamaño 2),
# el valor asignado a cada jugador i coincide con la media de:
#   (1/(n-1)) * sum_{j != i} v(i,j)
# Es decir, la probabilidad media de victoria de i frente al resto.
shapley <- apply(P, 1, function(row) mean(row, na.rm = TRUE))
# Guardamos los valores en el data.frame
atp$ShapleyPuntos <- shapley

#------------------------------------------------------------
# 5) Ordenación y salida de resultados
```

```
#------------------------------------------------------------

# Redondear Shapley a 3 decimales

atp$ShapleyPuntos <- round(atp$ShapleyPuntos, 3)

# Ordenamos de mayor a menor según el índice de Shapley calculado

atp_shapley <- atp[order(-atp$ShapleyPuntos), ]

# Mostramos las primeras filas (top 100)

head(atp_shapley, 100)
```

A.2 Clasificación supervisada y ranking Shapley.

A.2.1 Script en R

```
#############################################################
###

# TFG – Clasificación supervisada aplicada al tenis

# Modelo de regresión logística para la predicción de

# probabilidades de victoria entre jugadores ATP

#############################################################
###


#------------------------------------------------------------

# 1) Lectura y preparación de los datos

#------------------------------------------------------------

library(readxl)

# El archivo contiene enfrentamientos entre jugadores ATP.

# Cada fila corresponde a un partido concreto.

# Resultado = 0 si gana el jugador "local", 1 si gana el "visitante".
```

```
partidos <- read_excel("enfrentamientosjugadoresATP.xlsx")
# Conversión de porcentajes a proporciones (0–1)
# Todas las columnas excepto 'Resultado' están en porcentaje.
decimal <- setdiff(names(partidos), "Resultado")
partidos[decimal] <- partidos[decimal] / 100
# Comprobación básica de la variable respuesta
table(partidos$Resultado)


#------------------------------------------------------------
# 2.A) Modelo completo
#------------------------------------------------------------
# Se ajusta inicialmente un modelo con un gran número de variables
# para detectar posibles problemas de colinealidad y ruido.
log.fit <- glm(
  Resultado ~
    ps1_c + ps1_f +
    pg1s_c + pg1s_f +
    pg2s_c + pg2s_f +
    prs_c + prs_f +
    jgs_c + jgs_f +
    pts_c + pts_f +
    pr1_c + pr1_f +
    pr2_c + pr2_f +
    prc_c + prc_f +
```

```
    jgr_c + jgr_f +
    ptr_c + ptr_f +
    ptt_c + ptt_f,
  data = partidos,
  family = binomial
)
summary(log.fit)
# OBSERVACIÓN:
# El modelo completo presenta coeficientes grandes y errores estándar elevados,
# lo que indica la presencia de colinealidad entre variables explicativas.
# Muchas variables miden aspectos similares del rendimiento (servicio, resto
# y puntos totales), generando inestabilidad en la estimación.

#------------------------------------------------------------
# 2.B) Diagnóstico de colinealidad – Modelo completo
# Se calcula la matriz completa de correlaciones y se
# extraen aquellas con |r| ≥ 0.85
#------------------------------------------------------------
# Variables explicativas (todas excepto la variable respuesta)
vars_completo <- setdiff(names(partidos), "Resultado")
X_completo <- partidos[, vars_completo]
# Matriz de correlaciones
cor_completo <- cor(X_completo, use = "complete.obs")
# Matriz redondeada a 2 decimales
```

```
cor_completo_round <- round(cor_completo, 2)

cor_completo_round

#------------------------------------------------------------

# Extracción de correlaciones elevadas (|r| ≥ 0.85)

#------------------------------------------------------------

thr <- 0.85

cm <- cor_completo

cm[lower.tri(cm, diag = TRUE)] <- NA

altas <- which(abs(cm) >= thr, arr.ind = TRUE)

tabla_cor_altas <- data.frame(

  Variable1 = rownames(cm)[altas[,1]],

  Variable2 = colnames(cm)[altas[,2]],

  Correlacion = round(cm[altas], 2)

)

tabla_cor_altas[order(-abs(tabla_cor_altas$Correlacion)), ]


#------------------------------------------------------------

# 3) Modelo reducido (modelo final)

#------------------------------------------------------------

# Las variables se eliminan no por falta de relevancia deportiva,

# sino por redundancia estadística (alta colinealidad),

# con el objetivo de reducir ruido y mejorar la estabilidad del modelo.

#

# El modelo final conserva variables representativas de distintas
```

```
# dimensiones del rendimiento: servicio, resto y rendimiento global.
log.fit <- glm(
  Resultado ~
    ps1_c + ps1_f +
    pg2s_c + pg2s_f +
    prs_c + prs_f +
    prc_c + prc_f +
    ptt_c + ptt_f +
    ptr_c + ptr_f,
  data = partidos,
  family = binomial
)
summary(log.fit)


#------------------------------------------------------------
# 4) Estadísticas anuales agregadas de los jugadores (temporada 2024)
#------------------------------------------------------------
# Cada vector contiene estadísticas agregadas del jugador en la temporada.
# Aunque el modelo final solo utiliza un subconjunto de variables,
# se mantienen todas las métricas por coherencia y posible extensión.
alcaraz_2024 <- c(ps1=0.65, pg1s=0.73, pg2s=0.57, prs=0.62, jgs=0.85,
                  pts=0.68, pr1=0.34, pr2=0.54, prc=0.43, jgr=0.31,
                  ptr=0.42, ptt=0.54)
djokovic_2024 <- c(ps1=0.64, pg1s=0.75, pg2s=0.55, prs=0.65, jgs=0.86,
```

```
            pts=0.68, pr1=0.33, pr2=0.55, prc=0.41, jgr=0.29,
            ptr=0.41, ptt=0.54)
medvedev_2024 <- c(ps1=0.61, pg1s=0.73, pg2s=0.49, prs=0.61, jgs=0.80,
            pts=0.64, pr1=0.32, pr2=0.53, prc=0.44, jgr=0.27,
            ptr=0.40, ptt=0.52)
zverev_2024 <- c(ps1=0.71, pg1s=0.77, pg2s=0.55, prs=0.70, jgs=0.90,
           pts=0.70, pr1=0.30, pr2=0.51, prc=0.37, jgr=0.22,
           ptr=0.38, ptt=0.53)
tsitsipas_2024 <- c(ps1=0.61, pg1s=0.75, pg2s=0.54, prs=0.68, jgs=0.86,
            pts=0.67, pr1=0.28, pr2=0.48, prc=0.41, jgr=0.21,
            ptr=0.35, ptt=0.51)
sinner_2024 <- c(ps1=0.61, pg1s=0.79, pg2s=0.58, prs=0.74, jgs=0.91,
           pts=0.71, pr1=0.32, pr2=0.56, prc=0.42, jgr=0.28,
           ptr=0.41, ptt=0.55)
rublev_2024 <- c(ps1=0.62, pg1s=0.77, pg2s=0.54, prs=0.69, jgs=0.87,
           pts=0.68, pr1=0.28, pr2=0.52, prc=0.36, jgr=0.20,
           ptr=0.37, ptt=0.52)
fritz_2024 <- c(ps1=0.62, pg1s=0.78, pg2s=0.55, prs=0.69, jgs=0.88,
          pts=0.70, pr1=0.27, pr2=0.50, prc=0.38, jgr=0.20,
          ptr=0.36, ptt=0.53)

#------------------------------------------------------------
# 5) Predicción de enfrentamientos entre jugadores
#------------------------------------------------------------
```

```
# Lista de jugadores y sus estadísticas
players_stats <- list(
  Alcaraz   = alcaraz_2024,
  Djokovic  = djokovic_2024,
  Medvedev  = medvedev_2024,
  Zverev    = zverev_2024,
  Tsitsipas = tsitsipas_2024,
  Sinner    = sinner_2024,
  Rublev    = rublev_2024,
  Fritz     = fritz_2024
)
# Variables necesarias para el modelo final
needed <- c("ps1", "pg2s", "prs", "prc", "ptt", "ptr")
# Comprobación de consistencia
miss <- lapply(players_stats, function(v) setdiff(needed, names(v)))
miss <- miss[sapply(miss, length) > 0]
if (length(miss) > 0) {
  print(miss)
  stop("Faltan variables necesarias en uno o más jugadores.")
}
# Columnas explicativas usadas por el modelo (con sufijos _c y _f)
pred_cols <- all.vars(delete.response(terms(log.fit)))
# Construcción de una observación (jugador local y visitante)
make_match_row <- function(player_c, player_f, pred_cols, players_stats) {
```

```
  row <- as.data.frame(setNames(replicate(length(pred_cols), NA, simplify=FALSE),
pred_cols))
  for (col in pred_cols) {
    base <- sub("_(c|f)$", "", col)
    side <- sub("^.*_([cf])$", "\\1", col)
    if (side == "c") row[[col]] <- unname(players_stats[[player_c]][base])
    if (side == "f") row[[col]] <- unname(players_stats[[player_f]][base])
  }
  row
}
# Cálculo de probabilidades P(i gana a j)
jugadores <- names(players_stats)
resultados <- data.frame()
for (i in jugadores) {
  for (j in jugadores) {
    if (i != j) {
      fila <- make_match_row(i, j, pred_cols, players_stats)
      # Resultado = 1 implica victoria del visitante (j)
      p_visitante_gana <- predict(log.fit, newdata=fila, type="response")
      # Probabilidad de que gane el jugador i
      p_i_gana <- 1 - p_visitante_gana
      resultados <- rbind(
        resultados,
        data.frame(
```

```
        jugador_i = i,
        jugador_j = j,
        prob_i_gana_a_j = round(p_i_gana, 6)
      )
    )
  }
 }
}
# Resultados finales
resultados
# Ordenados por mayor probabilidad de victoria
resultados <- resultados[order(-resultados$prob_i_gana_a_j), ]
resultados
resultados$prob_i_gana_a_j <- round(resultados$prob_i_gana_a_j, 3)


#------------------------------------------------------------
# 6) Ranking Shapley (Nowak–Radzik) a partir de las probabilidades v(i,j)
#------------------------------------------------------------
# n = número total de jugadores considerados
n <- length(jugadores)
# Índice de Shapley generalizado (Nowak–Radzik) para coaliciones ordenadas de tamaño 2:
# φ_i^{NR}(v) = (1/(n-1)) * sum_{j ≠ i} v(i,j)
# En este caso, equivale a la probabilidad media de victoria del jugador i
```

```
# frente al resto de jugadores considerados.
shapley_NR <- aggregate(
  prob_i_gana_a_j ~ jugador_i,
  data = resultados,
  FUN = mean
)
# Renombrar la columna con el índice de Shapley
names(shapley_NR)[2] <- "shapley_NR"
# Ordenar de mayor a menor y asignar ranking
shapley_NR <- shapley_NR[order(-shapley_NR$shapley_NR), ]
shapley_NR$ranking <- seq_len(nrow(shapley_NR))
# Reordenar columnas
shapley_NR <- shapley_NR[, c("ranking", "jugador_i", "shapley_NR")]
# Mostrar tabla final del ranking
shapley_NR
```

A.2.2 Matriz completa de correlaciones

Tabla 7. Matriz completa de correlaciones entre las variables explicativas del modelo completo

	ps1_c	ps1_f	pg1s_c	pg1s_f	pg2s_c	pg2s_f	prs_c	prs_f	jgs_c	jgs_f	pts_c	pts_f	pr1_c	pr1_f	pr2_c	pr2_f	prc_c	prc_f	jgr_c	jgr_f	ptr_c	ptr_f	ptt_c	ptt_f
ps1_c	1	0.05	0.16	0.04	-0.19	0	0.16	-0.12	0.32	-0.01	0.36	0.07	0.02	-0.24	-0.03	-0.19	-0.1	-0.05	0	-0.26	-0.08	-0.23	0.2	-0.16
ps1_f	0.05	1	0.1	-0.14	0.05	-0.47	0.13	-0.4	0.15	-0.13	0.12	-0.09	0.16	-0.07	0.14	-0.02	0.13	-0.31	0.18	-0.03	0.14	-0.04	0.21	-0.07
pg1s_c	0.16	0.1	1	-0.06	-0.07	-0.09	0.41	-0.09	0.88	-0.08	0.88	-0.06	-0.1	-0.06	-0.05	-0.05	-0.17	-0.07	-0.14	-0.05	-0.13	-0.08	0.52	-0.11
pg1s_f	0.04	-0.14	-0.06	1	0.19	0.21	-0.06	0.28	0.01	0.89	0.04	0.87	0.21	-0.29	0.28	-0.04	0.26	0.08	0.3	-0.18	0.26	-0.23	0.19	0.43
pg2s_c	-0.19	0.05	-0.07	0.19	1	-0.02	0.36	0.19	0.27	0.21	0.28	0.15	0.51	0.15	0.63	0.18	0.28	0.11	0.6	0.19	0.63	0.15	0.55	0.27
pg2s_f	0	-0.47	-0.09	0.21	-0.02	1	-0.16	0.33	-0.13	0.48	-0.11	0.58	-0.16	-0.09	-0.09	-0.09	0	-0.17	-0.11	-0.1	-0.12	-0.05	-0.16	0.25
prs_c	0.16	0.13	0.41	-0.06	0.36	-0.16	1	-0.04	0.67	-0.09	0.54	-0.07	0.25	-0.01	0.25	0.01	0.03	-0.02	0.22	-0.01	0.26	-0.02	0.54	-0.1
prs_f	-0.12	-0.4	-0.09	0.28	0.19	0.33	-0.04	1	-0.03	0.52	-0.06	0.26	0	0.06	0.09	0.03	-0.05	-0.1	0.01	-0.02	0.03	0.06	0	0.27
jgs_c	0.32	0.15	0.88	0.01	0.27	-0.13	0.67	-0.03	1	-0.01	0.97	0	0.2	-0.08	0.25	-0.03	-0.05	-0.06	0.17	-0.06	0.19	-0.08	0.77	-0.07
jgs_f	-0.01	-0.13	-0.08	0.89	0.21	0.48	-0.09	0.52	-0.01	1	0	0.92	0.14	-0.21	0.23	-0.02	0.21	-0.09	0.24	-0.14	0.2	-0.12	0.13	0.53
pts_c	0.36	0.12	0.88	0.04	0.28	-0.11	0.54	-0.06	0.97	0	1	0.02	0.16	-0.05	0.21	-0.01	-0.06	-0.01	0.14	-0.03	0.15	-0.07	0.75	-0.03
pts_f	0.07	-0.09	-0.06	0.87	0.15	0.58	-0.07	0.26	0	0.92	0.02	1	0.13	-0.34	0.23	-0.12	0.26	-0.14	0.24	-0.26	0.19	-0.27	0.14	0.42
pr1_c	0.02	0.16	-0.1	0.21	0.51	-0.16	0.25	0	0.2	0.14	0.16	0.13	1	-0.06	0.79	0.05	0.53	-0.01	0.94	0	0.94	-0.02	0.69	0.09
pr1_f	-0.24	-0.07	-0.06	-0.29	0.15	-0.09	-0.01	0.06	-0.08	-0.21	-0.05	-0.34	-0.06	1	-0.11	0.8	-0.1	0.57	-0.07	0.95	-0.05	0.96	-0.12	0.65
pr2_c	-0.03	0.14	-0.05	0.28	0.63	-0.09	0.25	0.09	0.25	0.23	0.21	0.23	0.79	-0.11	1	0	0.57	-0.05	0.91	-0.05	0.93	-0.07	0.74	0.12
pr2_f	-0.19	-0.02	-0.05	-0.04	0.18	-0.09	0.01	0.03	-0.03	-0.02	-0.01	-0.12	0.05	0.8	0	1	0.04	0.36	0.06	0.88	0.06	0.9	-0.01	0.68
prc_c	-0.1	0.13	-0.17	0.26	0.28	0	0.03	-0.05	-0.05	0.21	-0.06	0.26	0.53	-0.1	0.57	0.04	1	-0.09	0.66	-0.03	0.56	-0.06	0.34	0.1
prc_f	-0.05	-0.31	-0.07	0.08	0.11	-0.17	-0.02	-0.1	-0.06	-0.09	-0.01	-0.14	-0.01	0.57	-0.05	0.36	-0.09	1	-0.02	0.55	0.01	0.47	-0.06	0.43
jgr_c	0	0.18	-0.14	0.3	0.6	-0.11	0.22	0.01	0.17	0.24	0.14	0.24	0.94	-0.07	0.91	0.06	0.66	-0.02	1	0	0.97	-0.02	0.71	0.16
jgr_f	-0.26	-0.03	-0.05	-0.18	0.19	-0.1	-0.01	-0.02	-0.06	-0.14	-0.03	-0.26	0	0.95	-0.05	0.88	-0.03	0.55	0	1	0.02	0.96	-0.06	0.69
ptr_c	-0.08	0.14	-0.13	0.26	0.63	-0.12	0.26	0.03	0.19	0.2	0.15	0.19	0.94	-0.05	0.93	0.06	0.56	0.01	0.97	0.02	1	-0.01	0.73	0.14
ptr_f	-0.23	-0.04	-0.08	-0.23	0.15	-0.05	-0.02	0.06	-0.08	-0.12	-0.07	-0.27	-0.02	0.96	-0.07	0.9	-0.06	0.47	-0.02	0.96	-0.01	1	-0.09	0.71
ptt_c	0.2	0.21	0.52	0.19	0.55	-0.16	0.54	0	0.77	0.13	0.75	0.14	0.69	-0.12	0.74	-0.01	0.34	-0.06	0.71	-0.06	0.73	-0.09	1	0.03
ptt_f	-0.16	-0.07	-0.11	0.43	0.27	0.25	-0.1	0.27	-0.07	0.53	-0.03	0.42	0.09	0.65	0.12	0.68	0.1	0.43	0.16	0.69	0.14	0.71	0.03	1