



SBL Education & Movistar Estudiantes

Basket Data Analytics & Sport Management Predicción de Resultados de Partidos de Baloncesto y Análisis de Factores Determinantes Utilizando Modelos de Machine Learning y Métodos de Explicabilidad

Trabajo fin de máster presentado por:	Alberto Cebollada Solanas
Director/a:	Javier Sanguino Bautiste
Fecha:	4 de octubre de 2024

En colaboración con:



1. Introducción	5
1.1. Contexto y justificación del trabajo	5
1.2. Objetivos del trabajo	7
2. Estado del arte	8
3. Marco Teórico de los Modelos predictivos	10
3.1. Random Forest	10
3.2. Red Neuronal	13
3.3. Curvas de Expectativa Condicional Individual	18
3.3.1. Interacciones no Lineales y no Aditivas	20
3.3.2. Promedio y Curvas de Comportamiento Medio	20
3.4. Counterfactual values (Valores Contrafácticos)	21
3.5. Shapley Values	23
4. Metodología	29
4.1. Elección de la fuente de datos	29
4.2. Preprocesado de los datos	30
4.3. Modelos predictivos elegidos	34
4.3.1. Random Forest	34
4.3.2. Red Neuronal	34
4.4. Métricas de Evaluación del Modelo	35
4.5. Diferencias Entre las Métricas de Evaluación	35
4.6. Metodos de explicabilidad elegidos	36
4.6.1. Curvas de Expectativa Condicional Individual (ICE)	36
4.6.2. Counterfactuals Values	36
4.6.3. Valores de Shapley y SHAP:	37
5. Resultados	39
5.1. Modelos predictivos	42
5.1.1. Test	44
5.1.2. Importancia Variables	46
5.2. Curvas de Expectativa Condicional Individual (ICE)	47

5.3.	Conterfactual values (Valores Contrafácticos)	49
5.4.	Shapley Values	52
6.	Caso de Uso	59
7.	Conclusiones	81
8.	Bibliografía	85
9.	Anexo	88
9.1.	Ficheros de datos	88
10.	Índice de Acrónimos	92

Resumen

Este trabajo ha presentado un análisis exhaustivo sobre la predicción del número de puntos en partidos de la NBA mediante modelos de machine learning, específicamente Random Forest y redes neuronales, utilizando datos desde el año 2000 y las últimas cinco temporadas (2018-2024). Se lograron rendimientos óptimos, con RMSE entre 5.86 y 7.06, y se evaluó la precisión de las predicciones, destacando un 65.49% de aciertos dentro del 5% de margen para la red neuronal. Las variables más influyentes identificadas fueron el número de posesiones, el porcentaje de aciertos en tiros de tres y el porcentaje de pérdidas.

Se aplicaron métodos explicativos como las curvas ICE, los valores contrafácticos y los valores SHAP para entender mejor el comportamiento de los modelos. Los valores SHAP revelaron que las redes neuronales eran más sensibles a variables relacionadas con la agresividad ofensiva, mostrando un impacto superior en el número de posesiones y tiros de campo. Finalmente, los hallazgos sugieren que los métodos explicativos pueden ser útiles para equipos de baloncesto al analizar estrategias de juego, proporcionando insights valiosos para la preparación de partidos basados en datos históricos y predicciones matemáticas.

Abstract

This work has presented a comprehensive analysis of predicting the number of points in NBA games using machine learning models, specifically Random Forest and neural networks, utilizing data from the year 2000 and the last five seasons (2018-2024). Optimal performances were achieved, with RMSE between 5.86 and 7.06, and the accuracy of the predictions was evaluated, highlighting a 65.49% success rate within a 5% margin for the neural network. The most influential variables identified were the number of possessions, three-point shooting percentage, and turnover percentage.

Explanatory methods such as ICE curves, counterfactual values, and SHAP values were applied to better understand the models' behavior. SHAP values revealed that neural networks were more sensitive to variables related to offensive aggressiveness, showing a greater impact on the number of possessions and field goals. Finally, the findings suggest that explanatory methods can be useful for basketball teams when analyzing game strategies, providing valuable insights for game preparation based on historical data and mathematical predictions.

Palabras clave: Machine Learning, Shapley Values, Counterfactual, Baloncesto, Métodos de Explicabilidad

1. INTRODUCCIÓN

1.1. CONTEXTO Y JUSTIFICACIÓN DEL TRABAJO

En los últimos años, el baloncesto ha sufrido grandes cambios tácticos. La punta de lanza de estos cambios, ha estado siempre en el dato. Gracias a los avances en la recogida de los datos y a su posterior procesamiento con algoritmos estadísticos, ahora se pueden medir muchos aspectos del juego. El estudio de la eficiencia de cada acción ha supuesto una revolución en el tipo de jugadas que se ven en los partidos. Un ejemplo de este cambio ha sido el uso del triple. El acierto normal de un tiro de media distancia (vale dos puntos) es del 40%, mientras que de un triple es del 30% (vale tres puntos). Por tanto, cada 100 tiros, con un lanzamiento de media distancia se obtendrán 80 puntos y con uno de 3, 90 puntos.

La predicción de competiciones deportivas y el análisis del rendimiento son cada vez más importantes, con el aprendizaje automático demostrando gran potencial en estos campos y esto se ve reflejado en el crecimiento de publicaciones científicas relacionadas con la predicción de resultados deportivos, así como la aparición de datasets con mayor volumen datos para ser explotados.

Uno de los primeros intentos para analizar el baloncesto desde el punto de vista de los datos fue Basketball on Paper de Dean Oliver ([1](#)), que se inspiró en lo que se venía haciendo en el béisbol en las décadas anteriores. La contribución más importante del libro es los factores de Oliver, que predicen la victoria de un partido de baloncesto basado en cuatro métricas, el porcentaje de tiro efectivo es decir la eficiencia de los tiros, pero mediante un ajuste de los triples, el porcentaje de balones perdidos, la capacidad de capturar rebotes, y finalmente, la cantidad de tiros libres anotados respecto al número de tiros de campo intentados.

Aunque los métodos de aprendizaje automático pueden ser útiles para predecir quién gana un partido de baloncesto, también tienen una complejidad que los hace difíciles de interpretar, lo que se conoce como el problema de la “caja negra”. Este fenómeno ha hecho que haya entrenadores que sean reacios a adoptar soluciones basadas en modelos de Machine Learning (ML). Efectivamente, la interpretabilidad es un concepto que está intrínsecamente relacionado con el usuario que utiliza el algoritmo de ML. Una de las definiciones de interpretabilidad es el grado en el cual un humano de forma consciente puede

predecir el resultado del modelo (2), otra sería el grado en el cual un humano puede entender una decisión (3).

En el baloncesto el problema es especialmente delicado, ya que las decisiones en el campo del deporte son tomadas de forma muy intuitiva, como por ejemplo cuando un entrenador hace un cambio a lo largo de un partido o cambia el sistema de juego. Esta forma parte del pensamiento rápido como explicaría Kahneman en (4). Se toman decisiones instantáneas basadas en experiencias pasadas, instintos y patrones, siendo eficiente para decisiones rápidas pero susceptible a errores, especialmente en situaciones complejas. Mientras que entender y confiar en un modelo de ML se basaría en un sistema lento, analítico y deliberado. Este sistema se activa cuando hay que enfrentarse a problemas complejos que requieren pensamiento consciente y reflexión. Aunque este sistema es más preciso en la toma de decisiones y en la resolución de problemas complejos, en el banquillo no hay modelo de ML sino al entrenador jefe.

La manera de acercar ambos sistemas es a través de los métodos de interpretabilidad. La interpretabilidad puede permitir a las personas manipular el sistema, traducido a lenguaje baloncestístico, sería saber que factores están influyendo en un determinado outcome de un partido (victoria o derrota, número de puntos anotados, número de rebotes) y entender como modificando estos factores se puede modificar el resultado y en qué medida. Pero como se puede explicar que la predicción individual de cada una de las instancias en nuestro estudio, dicho en términos deportivos, cuales han sido los factores mas importantes en que en un partido un equipo meta un número de puntos determinado y en otro partido se hayan metido los mismos puntos, cuando los inputs del modelo son diferentes. A esta pregunta se responde a través de lo que se denominan métodos de explicabilidad (Explanation methods).

Lo que se entiende por explicación generalmente relaciona los valores de las características de una instancia con su predicción del modelo de una manera comprensible para los humanos. Una explicación sería la respuesta a la pregunta ¿Porqué mi equipo ganó el partido en la última jornada? o ¿Porqué mi equipo en el ultimo partido metió 95 puntos? (5). Quizá en el banquillo no se tiene un modelo de ML, pero sería capaz de explicarle al entrenador porque un modelo predice ese la victoria o el número de puntos, y lo más interesante, determinar cómo se puede conseguir en el siguiente partido meter 5 puntos mas, es decir que decisiones puede tomar para lograrlo, y los métodos de explicabilidad ayudan a ello.

En este trabajo se construirán dos modelos de ML que predigan el número de puntos anotados en un partido de baloncesto a partir de una serie de parámetros capturados en los box-score de la NBA entre las temporadas (2000-2001 y 2023-2024). Se planteará un modelo con los datos capturados desde el año 2000 hasta el 2024, y otro modelo con los datos capturados a partir de 2018 hasta 2024. Posteriormente se realizará un análisis de los factores clave que afectan al número de puntos anotado. En esta metodología se usarán diferentes métodos locales de explicabilidad para interpretar estos modelos, evaluar la concordancia entre métodos e intentar proporcionar una herramienta de soporte para la toma de decisiones dinámicas de los entrenadores durante el juego

1.2. OBJETIVOS DEL TRABAJO

El objetivo principal de este trabajo es identificar los factores que influyen en la capacidad ofensiva de un equipo de baloncesto, utilizando los datos obtenidos de los box-scores de la NBA desde el año 2000. Para lograrlo, se desarrollarán dos modelos de ML: Random Forest y una Red Neuronal, que permitirán predecir el número de puntos anotados en un partido. Se optimizará la precisión de las predicciones utilizando métricas como Error Medio Absoluto (MAE), el Error Cuadrático Medio (MSE), la Raíz del Error Cuadrático Medio (RMSE) o la Precisión en un intervalo. Así como se identificarán los factores clave que afectan al rendimiento ofensivo en función de los modelos. Posteriormente, se aplicarán métodos explicativos locales, como las Curvas de Expectativa Condicional Individual (ICE), los valores contra-factuales y los valores de Shapley (SHAP), para interpretar las predicciones individuales de los modelos. Además, se evaluará la variabilidad de estos métodos para cada predictor con el fin de determinar cuáles son más accionables o eficaces a la hora de modificar las predicciones.

2. ESTADO DEL ARTE

En la literatura hay numerosas publicaciones cuyo objetivo es predecir si un partido va a terminar en victoria o derrota y que factores están influenciando en que se dé un resultado. Por ejemplo en (6) a través de redes neuronales se predijo de forma precisa la victoria con un 74,3%. En (7) se utilizó el algoritmo de Naive-Bayes para predecir con un acierto del 67%. En (8) compararon cinco modelos de clasificación supervisada, encontrando que el Random Forest tuvo una precisión del 65.15%, mejorando al 68.75% al dividir la temporada en cuartiles con regresión logística. En (9) utilizando el modelo de predicción SVM logró una precisión de predicción del 85.2%. Estos son unos ejemplos de las múltiples publicaciones que tienen el objetivo de predecir derrota o victoria, pero este trabajo pretende no estudiar este outcome, sino evaluar como se comporta el número de puntos que anota un equipo en un partido, de que factores depende y se es capaz de determinar como es posible modificar estos factores para que un entrenador pueda conseguir esto de la forma mas eficaz. En (10) construyeron un árbol de regresión y regresión con SVM para predecir los puntos de cada jugador en dos equipos, con el matriz de que luego evaluaron si con esos puntos anotados por cada equipo, cual de ellos ganaba el partido.

En lo referente a métodos explicativos en el campo deportivo se pueden encontrar artículos como (11) que analizan los métodos ProtoDash y SHAP, en este caso para predecir resultados en Voleibol, en (12) se analizan los métodos LIME y SHAP, con el objetivo de proporcionar explicaciones sobre las características originales cuando se trata de predecir resultados de lanzamientos de béisbol. En (13) aplican LIME (Explicación Localmente Interpretable Independiente del Modelo) en redes neuronales y Random Forests para analizar el estilo de juego basado en los boxscores de la NBA. En (14) van mas alla y definen un sistema de soporte de decisiones completo en el cual utilizan un modelo de aprendizaje automático para predecir el resultado y valores SHAP para explicar la predicción del modelo.

Existe poca literatura que aborde los métodos explicativos en baloncesto, especialmente en lo que respecta a los modelos que predicen el número de puntos anotados. Aunque los modelos predictivos son comúnmente utilizados en el análisis deportivo, pocos estudios se centran en explicar los factores o mecanismos subyacentes que contribuyen a la producción

de puntos en los partidos de baloncesto para ello este trabajo puede contribuir a aumentar el conocimiento en el aspecto de explorar los factores clave que influyen en el número de puntos anotados en baloncesto, utilizando un enfoque explicativo que permita identificar y comprender mejor las variables determinantes en el rendimiento ofensivo de los equipos.

3. MARCO TEÓRICO DE LOS MODELOS PREDICTIVOS

El ML es la ciencia que desarrolla algoritmos y modelos estadísticos que permiten a los sistemas de computación realizar tareas sin instrucciones explícitas, basándose en el reconocimiento de patrones e inferencias. Estos algoritmos procesan grandes cantidades de datos históricos para identificar patrones, lo que les permite generar resultados con mayor precisión a partir de datos de entrada. Por ejemplo, las aplicaciones médicas pueden entrenarse para diagnosticar cáncer a partir de imágenes de rayos X, analizando millones de escaneos almacenados y sus diagnósticos asociados.

La idea central del ML es la relación matemática entre los datos de entrada y los de salida. Aunque esta relación no se conoce de antemano, los algoritmos pueden estimarla si se les proporciona suficientes datos. A medida que se incrementan los datos y la potencia de procesamiento, las predicciones se vuelven más precisas.

En este trabajo se han utilizado dos algoritmos de ML, Random Forest y Redes Neuronales. En ambos casos el objetivo ha sido predecir el número de puntos anotado a partir de una serie de predictores que se mencionan en la sección de resultados.

3.1. RANDOM FOREST

Random Forest es un algoritmo que mejora las predicciones utilizando varios árboles de decisión en lugar de uno solo. Un árbol de decisión es un modelo predictivo que divide el espacio de los predictores agrupando observaciones con valores similares para la variable respuesta o dependiente. Se construyen múltiples árboles de decisión (habitualmente cientos o miles) a partir de diferentes subconjuntos de los datos de entrenamiento.

Un **Random Forest** construye B árboles de decisión a partir de diferentes subconjuntos de los datos de entrenamiento. Cada árbol es entrenado usando un subconjunto aleatorio de las variables X y los ejemplos D .

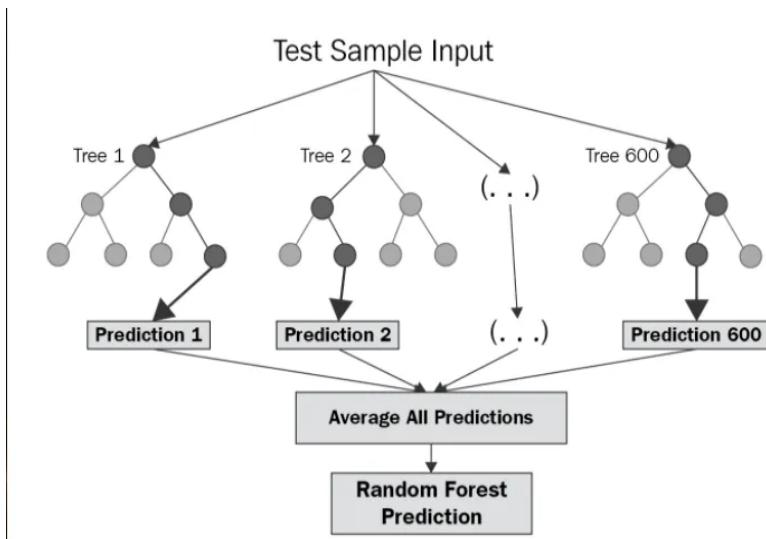


Figura 3.1: Representación gráfica de Random Forest

En la figura 3.1 ([15](#)) se muestra el funcionamiento de un Random forest, se puede ver que 600 árboles se ajustan en paralelo sin interacción entre ellos, el algoritmo funciona construyendo múltiples decisiones durante el entrenamiento y mostrando la media de dichas predicciones

Los parámetros a tener en cuenta en un Random Forest son:

- `n_estimators`: el número de árboles de decisión que se ejecutarán en el modelo.
- `criterion`: esta variable te permite seleccionar el criterio (función de pérdida) utilizado para determinar los resultados del modelo. Se puede seleccionar entre funciones de pérdida como el error cuadrático medio (MSE) y el error absoluto medio (MAE). El valor predeterminado es MSE.
- `max_depth`: establece la profundidad máxima posible de cada árbol.
- `max_features`: el número máximo de características que el modelo considerará al determinar una división.
- `bootstrap`: el valor predeterminado es True, lo que significa que el modelo sigue los principios del bootstrapping (definido anteriormente).
- `max_samples`: este parámetro supone que bootstrapping está configurado como True; si no, este parámetro no se aplica.

Si D es el conjunto de datos de tamaño N , para cada árbol se selecciona una muestra con reemplazo de N datos, lo que se conoce como **método de “bootstrap”**.

$$D_{\text{bootstrap}} = \{(X_1, y_1), (X_2, y_2), \dots, (X_N, y_N)\}$$

Donde:

- X_i son las variables predictoras,
- y_i es la variable respuesta.

Cada árbol se entrena con una **muestra aleatoria** de p variables seleccionadas aleatoriamente de las P totales.

En cuanto a la selección de características, en cada nodo de cada árbol se selecciona aleatoriamente un subconjunto de las características en lugar de considerar todas, lo que introduce variabilidad y reduce la correlación entre los árboles.

Cada árbol en el Random Forest realiza particiones sucesivas en los datos.

Para un nodo t , la mejor partición se selecciona al maximizar una métrica de importancia como la **impureza**. En el caso de problemas de clasificación, se utiliza la **Gini impurity**:

$$Gini(t) = 1 - \sum_{i=1}^C p_{ti}^2$$

Donde:

- p_i es la proporción de observaciones de la clase i en el nodo t ,
- C es el número total de clases.

Para problemas de regresión, se minimiza el **error cuadrático medio** (MSE):

$$MSE(t) = \frac{1}{N_t} \sum_{i=1}^{N_t} (y_i - \bar{y}_t)^2$$

Donde:

- N_t es el número de observaciones en el nodo t ,
- \bar{y}_t es el valor medio de y en ese nodo.

Para la clasificación, cada árbol da una predicción \hat{y}_b y el Random Forest elige la clase con más votos entre los B árboles:

$$\hat{y} = \text{mode}(\hat{y}_1, \hat{y}_2, \dots, \hat{y}_B)$$

Donde:

- \hat{y}_b es la predicción del árbol b ,
- B es el número total de árboles.

Para la regresión, el resultado final se obtiene promediando las predicciones \hat{y}_b de los B árboles:

$$\hat{y} = \frac{1}{B} \sum_{b=1}^B \hat{y}_b$$

Esto garantiza la independencia de los hiperparámetros y de los datos utilizados y se asemeja mas a modelar su distribución real.

La importancia de una variable en un **Random Forest** se evalúa a través de la **reducción de la impureza** promedio (para clasificación) o la reducción del MSE (para regresión) que dicha variable proporciona en los nodos donde es usada para la partición.

Para una variable X_j , la importancia se define como:

$$\text{Importancia}(X_j) = \frac{1}{B} \sum_{b=1}^B \sum_{t \in T_b} \Delta I_t(X_j)$$

Donde:

- T_b es el conjunto de nodos en el árbol b ,
- $\Delta I_t(X_j)$ es la reducción en la impureza o MSE debido a X_j en el nodo t .

3.2. RED NEURONAL

Las Redes Neuronales Artificiales (del inglés Artificial Neural Network ,ANN) son modelos de ML que se inspiraron en el conocimiento que había de las redes neuronales cerebrales en el momento de su creación. En la actualidad, han divergido mucho de los descubrimientos que se han hecho en el cerebro. Sin embargo, estos modelos matemáticos funcionan excepcionalmente bien en muchos conjuntos de datos

Estos modelos se consideran “cajas negras” porque el proceso exacto de transformación interna de las señales es difícil de interpretar. Una Red Neuronal Artificial consta de varios

niveles de nodos interconectados, desde el nivel de entrada hasta el de salida, pasando por uno o más niveles ocultos.

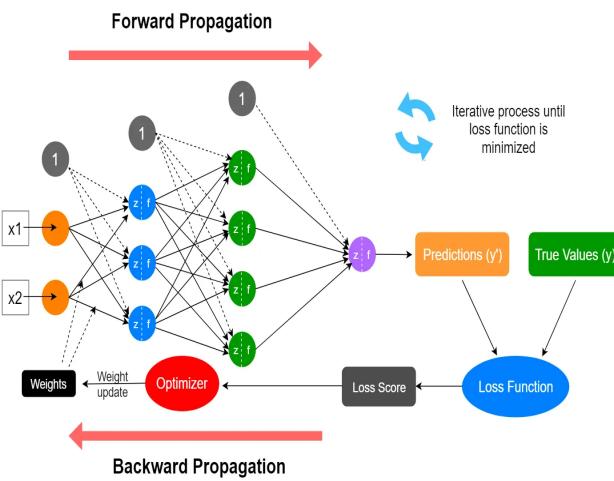


Figura 3.2: Representación gráfica de una Red Neuronal

Cada nodo puede ajustar la señal con un sesgo (bias), y la forma en que se conectan los nodos define la topología de la red.

En las redes más comunes, la información fluye de entrada a salida, y los nodos suelen conectarse con todos los nodos del siguiente nivel, aunque esto no es siempre necesario, como en las redes recurrentes.

Entrada a la capa oculta (propagación hacia adelante)

Para una neurona j en la capa oculta, el cálculo de su activación se realiza a partir de las entradas x_i y los pesos w_{ij} que conectan la entrada i con la neurona j :

$$z_j = \sum_{i=1}^n w_{ij} x_i + b_j$$

$$z_j = f \sum_{i=1}^n w_{ij} x_i + b_j$$

Donde:

- z_j es la neurona j ,
- w_{ij} es el peso entre la neurona i de la capa anterior y la neurona j ,
- b_j es el sesgo (bias) asociado a la neurona j ,

- x_j es la neurona de la capa anterior.

Función de activación

Después de calcular z_j , se aplica una **función de activación** f para generar la salida a_j de la neurona j :

La función de activación en cada nodo procesa y transmite la señal, modulada por un peso específico.

$$a_j = f(z_j)$$

Las funciones de activación más comunes son:

- **ReLU (Rectified Linear Unit):** $f(z) = \max(0, z)$

- **Sigmoide:** $f(z) = \frac{1}{1+e^{-z}}$

- **Tanh:** $f(z) = \frac{e^z - e^{-z}}{e^z + e^{-z}}$

Salida de la red

Para la capa de salida, los cálculos son similares. Si es un problema de regresión, la función de activación puede ser lineal, es decir, la salida es simplemente el valor z_j calculado. Si es un problema de clasificación, a menudo se utiliza la función **softmax** para obtener probabilidades:

$$a_j = \frac{e^{z_j}}{\sum_k e^{z_k}}$$

Donde a_j representa la probabilidad de que la entrada pertenezca a la clase j .

Cálculo de la función de pérdida

La red aprende ajustando los pesos mediante la minimización de una **función de pérdida** L .

Algunos ejemplos son:

MSE (Mean Squared Error) para regresión:

$$L = \frac{1}{N} \sum_{i=1}^N (y_i - \hat{y}_i)^2$$

Donde y_i es el valor real y \hat{y}_i es la predicción de la red.

Cross-entropy para clasificación:

$$L = - \sum_{i=1}^N y_i \log(\hat{y}_i)$$

Donde y_i es la clase verdadera y \hat{y}_i es la probabilidad predicha para esa clase.

Actualización de los pesos (backpropagation)

Resolver el problema de minimizar la función de pérdida en una NN es difícil porque la red tiene muchos parámetros. Los pesos se ajustan utilizando el algoritmo de descenso de gradiente, que calcula la derivada de la función de pérdida respecto a cada peso. La regla de actualización es:

$$w_{ij} = w_{ij} - \eta \frac{\partial L}{\partial w_{ij}}$$

Donde η es la tasa de aprendizaje y $\frac{\partial L}{\partial w_{ij}}$ es el gradiente de la pérdida con respecto al peso w_{ij} .

Donde el cálculo de la derivada se realiza de manera sencilla gracias a aplicar la regla de la cadena.

La importancia de cada característica de entrada x_i se estima combinando los valores absolutos de los pesos que conectan la capa de entrada con la capa oculta, y los pesos que conectan la capa oculta con la capa de salida. Específicamente, la importancia $I(x_i)$ de la característica x_i se define como:

$$I(x_i) = \sum_{j=1}^m |w_{ij}| \cdot |\nu_j|$$

Este método destaca las características que tienen una mayor influencia en la salida al considerar tanto los pesos entre la capa de entrada y la capa oculta, como los pesos entre la capa oculta y la capa de salida. Las características con valores de importancia más altos tienen un efecto más fuerte en las predicciones de la red neuronal ([16](#))

Métricas de Rendimiento de los modelos

Error Medio Absoluto (MAE)

El Error Medio Absoluto representa el promedio de la diferencia absoluta entre los valores reales y los valores predichos en el conjunto de datos. Mide el promedio de los residuos en el conjunto de datos y se expresa en las mismas unidades que la variable de interés.

$$\text{MAE} = \frac{1}{n} \sum_{i=1}^n |y_i - \hat{y}_i|$$

Error Cuadrático Medio (MSE)

El Error Cuadrático Medio representa el promedio de la diferencia al cuadrado entre los valores originales y los valores predichos en el conjunto de datos. Mide la varianza de los residuos y se expresa en unidades cuadradas de la variable de interés.

$$\text{MSE} = \frac{1}{n} \sum_{i=1}^n (y_i - \hat{y}_i)^2$$

Raíz del Error Cuadrático Medio (RMSE)

La Raíz del Error Cuadrático Medio es la raíz cuadrada del Error Cuadrático Medio. Mide la desviación estándar de los residuos.

$$\text{RMSE} = \sqrt{\frac{1}{n} \sum_{i=1}^n (y_i - \hat{y}_i)^2}$$

Métrica de Evaluación: Precisión dentro de un porcentaje X

Dado un conjunto de predicciones \hat{y}_i y los valores reales y_i , la métrica evalúa cuántas predicciones se encuentran dentro de un margen del $X\%$ con respecto al valor real.

Se define el margen de tolerancia como:

$$\text{Margen}_i = X\% \times y_i = \frac{X}{100} \times y_i$$

Para cada predicción \hat{y}_i , se verifica si cumple:

$$y_i - \text{Margen}_i \leq \hat{y}_i \leq y_i + \text{Margen}_i$$

Si esta condición se cumple, se considera la predicción como **correcta**. El número total de predicciones correctas se denota como $N_{\text{correctas}}$.

La precisión (Accuracy) se calcula como la proporción de predicciones correctas sobre el total de predicciones N :

$$\text{Accuracy} = \frac{N_{\text{correctas}}}{N}$$

Donde:

- N es el número total de predicciones,
- $N_{\text{correctas}}$ es el número de predicciones que caen dentro del margen tolerado.

Ejemplo con umbral $X = 10\%$

Si se define $X = 10\%$, el margen de tolerancia sería:

$$\text{Margen}_i = 0.1 \times y_i$$

Y se evalúa cuántas predicciones cumplen:

$$y_i - 0.1 \times y_i \leq \hat{y}_i \leq y_i + 0.1 \times y_i$$

3.3. CURVAS DE EXPECTATIVA CONDICIONAL INDIVIDUAL

Los gráficos de Expectativas Condicionales Individuales (ICE) muestran una línea por instancia, que indica cómo cambia la predicción de esa observación cuando varía una característica.

El gráfico de dependencia parcial (PDP) para el efecto promedio de una característica es un método global porque no se enfoca en observaciones específicas, sino en un promedio general. El equivalente a un PDP para instancias de datos individuales se llama gráfico de Expectativa Condicional Individual (ICE). Un gráfico ICE visualiza la dependencia de la predicción en una característica para cada instancia por separado, lo que da como resultado una línea por instancia, en comparación con una línea general en los gráficos de dependencia parcial.

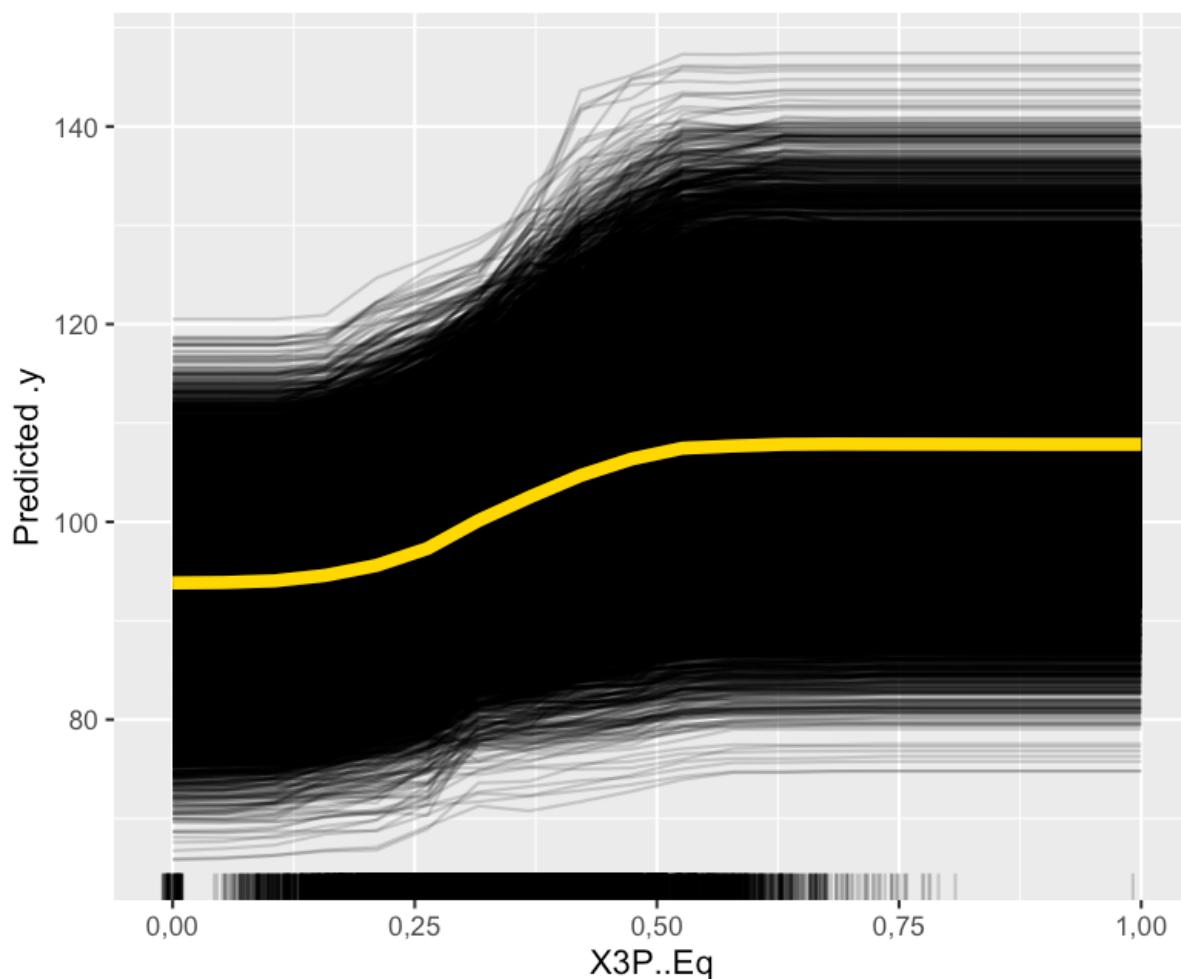


Figura 3.3: Ejemplo de Gráfica ICE y PDP

Un PDP es el promedio de las líneas de un diagrama ICE (en la figura, la línea amarilla). Los valores para una observación se pueden calcular manteniendo todas las otras características iguales, creando variantes de esta instancia al reemplazar el valor de la característica con valores de una cuadrícula y haciendo predicciones con el modelo de caja negra para estas instancias recién creadas. El resultado es un conjunto de puntos para una observación con el valor de la característica y las predicciones respectivas.

¿Cuál es el objetivo de analizar expectativas individuales en lugar de dependencias parciales? Las gráficas de dependencia parcial pueden ocultar una relación heterogénea creada por las interacciones. Los PDP pueden mostrar cómo se ve la relación promedio entre una característica y la predicción. Esto solo funciona bien si las interacciones entre las características para las cuales se calcula el PDP y las otras características son débiles. En caso de interacciones, la gráfica ICE proporcionará mucha más información.

Sin embargo, tienen limitaciones: solo pueden mostrar una característica de manera clara, y si la característica de interés está correlacionada con otras, podrían generar puntos de datos no válidos.

Además, si se incluyen demasiadas curvas, el gráfico puede volverse confuso, aunque esto puede mitigarse con transparencia o muestreo.

Para cada observación i en el conjunto de datos, la gráfica ICE representa la función de predicción condicional sobre x_j :

$$\hat{f}_{ICE}^{(i)}(x_j) = f(x_1^{(i)}, \dots, x_{j-1}^{(i)}, x_j, x_{j+1}^{(i)}, \dots, x_p^{(i)})$$

Donde:

- $\hat{f}_{ICE}^{(i)}(x_j)$ es la predicción del modelo para la observación i , modificando el valor de la variable x_j .
- x_j varía dentro de su rango o un conjunto predefinido de valores.
- Las otras variables $x_1^{(i)}, \dots, x_{j-1}^{(i)}, x_{j+1}^{(i)}, \dots, x_p^{(i)}$ se mantienen constantes en sus valores observados para la observación i .

Esta función permite visualizar cómo afecta a cada observación un cambio en la variable x_j , lo que facilita el análisis de la heterogeneidad de las respuestas del modelo.

3.3.1. Interacciones no Lineales y no Aditivas

Las gráficas ICE son particularmente útiles para detectar interacciones complejas entre las variables predictoras y no aditividad en los modelos. Si las curvas ICE de diferentes observaciones son paralelas, se puede concluir que no hay interacciones entre la variable x_j y otras variables del modelo. Sin embargo, si las curvas tienen diferentes pendientes o formas, se puede inferir la presencia de interacciones.

3.3.2. Promedio y Curvas de Comportamiento Medio

Aunque las ICE permiten el análisis individual, también es común obtener una curva promedio para observar el comportamiento medio de la variable x_j . Sin embargo, este promedio puede ocultar patrones importantes a nivel individual.

$$\hat{f}_{PDP}(x_j) = \frac{1}{n} \sum_{i=1}^n \hat{f}_{ICE}^{(i)}(x_j)$$

En este caso, $\hat{f}_{PDP}(x_j)$ es la gráfica de dependencia parcial, que es el promedio de las gráficas ICE para todas las observaciones, pero no es el foco de este análisis.

3.4. COUNTERFACTUAL VALUES (VALORES CONTRAFÁCTICOS)

Una explicación ‘contrafáctica’ describe una situación causal en la forma: “Si X no hubiera ocurrido, Y no habría ocurrido”. Por ejemplo: “Si el jugador no hubiera fallado ese tiro libre en el último minuto, el equipo habría ganado el partido”.

Pensar en contrafácticas requiere imaginar una realidad hipotética que contradiga los hechos observados (por ejemplo, un mundo en el que el jugador encestó el tiro libre), de ahí el nombre de “contrafáctico”.

Simplemente se cambian los valores de las características de una instancia antes de hacer las predicciones y se analiza cómo cambia la predicción. Lo interesante es estudiar escenarios en los que la predicción cambia de manera relevante, como un cambio en el resultado del partido (por ejemplo, victoria o derrota) o en que la predicción alcance un cierto umbral (por ejemplo, la probabilidad de ganar alcanza el 80%). Una explicación contrafáctica de una predicción describe el cambio más pequeño en los valores de la característica que altera la predicción a un resultado predefinido.

En nuestro caso, se quiere explicar un modelo que predice un resultado continuo utilizando explicaciones contrafácticas. Un entrenador quiere estimar cuántos puntos anotará su equipo en el próximo partido, por lo que entrena un modelo de aprendizaje automático para predecir la puntuación. Después de ingresar todos los detalles sobre el rendimiento pasado del equipo, la alineación, el equipo contrario, etc., el modelo le dice que su equipo anotará 85 puntos. El entrenador esperaba más de 90 puntos, pero confía en su modelo y decide ajustar los valores de las características para ver cómo puede mejorar la predicción. Descubre que el equipo podría anotar más de 90 puntos si el conjunto de los jugadores del equipo tuviera un mejor porcentaje de aciertos en triples. Este es un conocimiento interesante y fácilmente accionable ya que durante la semana se puede trabajar el tiro de 3, o en determinados momentos del partido sacar a los jugadores con mayor acierto. Finalmente, ajustando solo los factores que están bajo su control (estrategia de juego, rotación de jugadores, intensidad defensiva, etc.),

descubre que aumentando el ritmo del juego y enfocándose en tiros de tres puntos, el equipo podría superar los 90 puntos. El entrenador ha trabajado intuitivamente con los contrafácticos para cambiar el resultado.

Primero, se define una función de pérdida que toma como entrada la instancia de interés, un resultado contrafáctico y el resultado deseado. La pérdida mide qué tan lejos está la predicción contrafáctica del resultado deseado y qué tan diferente es el contrafáctico de la instancia original. Se puede optimizar esta función de pérdida directamente mediante un algoritmo de optimización o explorando el espacio alrededor de la instancia, como sugiere el método “Growing Spheres”.

Wachter et al. (2017) ([DBLP?](#)) proponen minimizar la siguiente función de pérdida:

$$L(x, x', y', \lambda) = \lambda \cdot (\hat{f}(x') - y')^2 + d(x, x')$$

El primer término representa la distancia cuadrática entre la predicción del modelo para el contrafáctico x' y el resultado deseado y' , definido por el usuario. El segundo término es la distancia $d(x, x')$ entre la instancia original x y el contrafáctico x' . El parámetro λ equilibra la distancia en la predicción contra la distancia en los valores de las características. La función de pérdida se resuelve para un λ dado y devuelve un x' contrafáctico. Un valor más alto de λ prioriza los contrafácticos que se acercan más al resultado deseado y' , mientras que un valor más bajo prioriza contrafácticos x' más similares a x . Si λ es muy grande, se seleccionará la instancia con la predicción más cercana a y' , sin importar qué tan diferente sea de x .

Los autores sugieren seleccionar una tolerancia ϵ para la distancia permitida en la predicción del contrafáctico respecto a y' . Esta restricción se expresa como:

$$|\hat{f}(x') - y'| \leq \epsilon$$

Para minimizar esta función de pérdida, se puede usar cualquier algoritmo de optimización adecuado, como Nelder-Mead. Si se tiene acceso a los gradientes del modelo, se pueden utilizar métodos basados en gradientes como ADAM. Se deben establecer previamente la instancia x que se va a explicar, el resultado deseado y' y el parámetro de tolerancia ϵ . La función de pérdida se minimiza para x' , devolviendo el contrafáctico óptimo al incrementar λ hasta que se encuentre una solución suficientemente cercana.

$$\operatorname{argmin}_{x'} \max_{\lambda} L(x, x', y', \lambda)$$

La función $d(x, x')$ para medir la distancia entre x y x' es la distancia ponderada de Manhattan con la desviación absoluta media inversa (MAD):

$$d(x, x') = \sum_{j=1}^p \frac{|x_j - x'_j|}{MAD_j}$$

La distancia total es la suma de todas las distancias p entre las características, escaladas por la inversa de la desviación absoluta media de la característica j sobre el conjunto de datos:

$$MAD_j = \text{median}_{i \in \{1, \dots, n\}}(|x_{i,j} - \text{median}_{l \in \{1, \dots, n\}}(x_{l,j})|)$$

El MAD es similar a la varianza, pero usa la mediana en lugar de la media y las distancias absolutas en lugar de las distancias cuadradas. La distancia Manhattan ponderada con MAD introduce escasez, lo que significa que dos puntos están más cerca cuando menos características difieren, y es más robusta frente a valores atípicos. Es necesario escalar con MAD para que todas las características estén en la misma escala, independientemente de sus unidades.

La receta para producir contrafácticos es simple:

Selecciona una instancia x para explicar, el resultado deseado y' , una tolerancia ϵ y un valor inicial (bajo) para λ .

Muestrea una instancia aleatoria como contrafáctico inicial.

Optimiza la pérdida con el contrafáctico muestreado inicialmente como punto de partida.

Mientras $|\hat{f}(x') - y'| > \epsilon$:

Aumenta λ .

Optimiza la pérdida con el contrafáctico actual como punto de partida.

Devuelve el contrafáctico que minimiza la pérdida.

Repite los pasos 2 a 4 y devuelve la lista de contrafácticos o el que minimiza la pérdida.

3.5. SHAPLEY VALUES

Los valores SHAP (SHapley Additive exPlanations) ofrecen una forma de interpretar los resultados de cualquier modelo de ML. Basados en la teoría de juegos, evalúan la contribución individual de cada muestra de entrenamiento (jugador) al modelo entrenado (resultado final).

En el contexto del ML, se puede interpretar como un valor que refleja la importancia de cada muestra en el modelo.

Estos valores SHAP destacan cómo cada característica influye en la predicción final, la relevancia de cada una en comparación con las demás y cómo el modelo depende de la interacción entre características ([17](#))

El valor de Shapley para una característica es lo que ha contribuido una característica a la predicción media.

El valor de Shapley mide la contribución de cada característica individual en una predicción.

Esta idea captura la importancia de una característica no solo por sí misma, sino también en combinación con otras características. Es un enfoque justo y equitativo porque considera todas las posibles interacciones entre características.

Un valor de Shapley se define mediante una función de valor val para los jugadores en S .

$$\phi_j(val) = \sum_{S \subseteq \{x_1, \dots, x_p\} \setminus \{x_j\}} \frac{|S|! \cdot (p - |S| - 1)!}{p!} \cdot (val(S \cup \{x_j\}) - val(S))$$

donde:

S es un subconjunto de las características utilizadas en el modelo, x es el vector de valores de características de la instancia a explicar y p es el número de características.

$val_x(S)$ es la predicción para los valores de características en el conjunto S que están marginados sobre las características que no están incluidos en el conjunto

$$val_x(S) = \int \hat{f}(x_1, \dots, x_p) dP_{x \notin S} - \mathbb{E}_X(\hat{f}(X))$$

Esto implica realizar múltiples integraciones para cada característica que no está contenida en S . Un ejemplo concreto: Supongamos que el modelo de aprendizaje automático utiliza 4 características x_1, x_2, x_3 y x_4 y se evalúa la predicción para la coalición S , que consta de los valores de características x_1 y x_3 :

$$val_x(S) = val_x(x_1, x_3) = \int_R \int_R \hat{f}(x_1, X_2, x_3, X_4) dP_{X_2 X_4} - \mathbb{E}_X(\hat{f}(X))$$

Esto se parece a las contribuciones de características en el modelo lineal.

No hay que confundir los muchos usos de la palabra “valor”:

El valor de la característica es el valor numérico o categórico de una característica e instancia.

El valor de Shapley es la contribución de la característica a la predicción.

La función de valor es la función de pago para coaliciones de jugadores (valores de características).

Se dice que el valor de Shapley es un método de atribución que cumple las propiedades de Eficiencia, Simetría, Dummies y Aditividad

Eficiencia Las contribuciones de características deben sumarse a la diferencia de predicción para x y el promedio de las predicciones

$$\sum_{j=1}^p \phi_j = \hat{f}(x) - \mathbb{E}_X(\hat{f}(X))$$

Simetría

Las contribuciones de dos características j y k deberían ser iguales si contribuyen de manera equivalente a todas las coaliciones posibles. Si

$$val(S \cup x_j) = val(S \cup x_k)$$

para todos

$$S \subseteq x_1, \dots, x_p \setminus x_j, x_k$$

entonces

$$\phi_j = \phi_k$$

Dummies

Una característica j que no cambia el valor predicho, independientemente de a qué coalición de valores de características se agregue, debe tener un valor Shapley de 0. Si

$$val(S \cup x_j) = val(S)$$

para todos

$$S \subseteq x_1, \dots, x_p$$

entonces

$$\phi_j = 0$$

Aditividad

Para un juego con pagos combinados $\text{val} + \text{val+}$, los valores de Shapley respectivos son los siguientes:

$$\phi_j + \phi_j^+$$

Si se entrena un modelo de random forest, lo que significa que la predicción es un promedio de muchos árboles de decisión. La propiedad de aditividad garantiza que, para una característica, puedes calcular el valor de Shapley para cada árbol individualmente, promediarlos y obtener el valor de Shapley para esa característica en el random forest.

Todas las coaliciones (conjuntos) posibles de valores de características deben evaluarse con y sin la característica j -ésima para calcular el valor exacto de Shapley. Para muchas características, la solución exacta a este problema se vuelve problemática, ya que el número de coaliciones posibles aumenta exponencialmente a medida que se agregan más características. Strumbelj et al. (2014) ([18](#)) proponen una aproximación con el muestreo de Monte-Carlo:

$$\widehat{\phi}_j = \frac{1}{M} \sum_{m=1}^M \left(\hat{f}(x_m^{+j}) - \hat{f}(x_m^{-j}) \right)$$

Donde $\hat{f}(x_m^{+j})$ es la predicción para x , pero con un número aleatorio de valores de características reemplazados por valores de características de un punto de datos aleatorio z , excepto el valor respectivo de la característica j . El vector x_m^{-j} es casi idéntico a x_m^{+j} , pero el valor x_m^j también se toma de la muestra z . Cada una de estas M nuevas instancias es una especie de “Frankenstein” ensamblado a partir de dos instancias.

Estimación aproximada de Shapley para el valor de una sola característica:

Salida: valor de Shapley para el valor de la característica j -ésima.

Requerido: Número de iteraciones M , instancia de interés x , índice de características j , matriz de datos X y modelo de aprendizaje automático f .

Para todos $m = 1, \dots, M$:

Dibuja una instancia aleatoria z de la matriz de datos X .

Elige una permutación aleatoria de los valores de la característica.

Instancia de orden x :

$$x_o = (x(1), \dots, x(j), \dots, x(p)).$$

Instancia de orden z :

$$z_o = (z(1), \dots, z(j), \dots, z(p)).$$

Construye dos nuevas instancias:

Con la función j :

$$x^{+j} = (x(1), \dots, x(j-1), x(j), z(j+1), \dots, z(p)).$$

Sin la característica j :

$$x^{-j} = (x(1), \dots, x(j-1), z(j), z(j+1), \dots, z(p)).$$

Calcular contribución marginal:

$$\phi_m^j = \hat{f}(x^{+j}) - \hat{f}(x^{-j}).$$

Calcular el valor de Shapley como el promedio:

$$\phi_j(x) = \frac{1}{M} \sum_M \widehat{\phi_m^j}.$$

Primero, selecciona una instancia de interés x , una característica j y el número de iteraciones M . Para cada iteración, se selecciona una instancia aleatoria z de los datos y se genera un orden aleatorio de las características. Se crean dos nuevas instancias combinando valores de la instancia de interés x y la muestra z . La primera instancia x^{+j} es la instancia de interés, pero todos los valores en el orden anterior e incluido el valor de la característica j se reemplazan por los valores de la característica de la muestra z . La segunda instancia x^{-j} es similar, pero tiene todos los valores en el orden anterior, pero excluye la característica j .

reemplazada por los valores de la característica j de la muestra z . Se calcula la diferencia en la predicción de caja negra:

$$\phi_m^j = \hat{f}(x_m^{+j}) - \hat{f}(x_m^{-j})$$

Todas estas diferencias se promedian y dan como resultado:

$$\phi_j(x) = \frac{1}{M} \sum_{m=1}^M \phi_m^j$$

El promedio pesa implícitamente las muestras por la distribución de probabilidad de X .

El procedimiento debe repetirse para cada una de las características para obtener todos los valores de Shapley.

4. METODOLOGÍA

Este proyecto tendrá la siguiente estructura

1. Obtención de datos y descripción detallada de ellos
2. Preprocesado de los datos
3. Generación de Modelos predictivos
4. Evaluación de los modelos a través de métricas específicas
5. Algoritmos de Explicabilidad

Pasaré a desarrollar los cinco pilares que componen este trabajo

4.1. ELECCIÓN DE LA FUENTE DE DATOS

Para la realización de este trabajo se utilizan datos de NBA debido a dos razones, Estados Unidos ha sido el país pionero en el uso de datos en el campo de los deportes, tal importancia tuvo que en 2004 se creó el primer puesto de analista estadístico en la NBA, así que el hecho de ser pioneros ha tenido como consecuencia que de todas las competiciones deportivas, se puede decir que los datos de NBA son los más curados y más fiables a la hora de hacer análisis estadísticos.

Cada equipo de la NBA tiene hasta 18 jugadores durante la temporada regular. La liga está dividida en dos conferencias (Este y Oeste), y cada equipo juega 82 partidos, enfrentando a rivales de su conferencia y de la otra. Al final de la temporada regular, los equipos compiten en un “play-in” para definir los últimos puestos de los Playoffs, que constan de cuatro rondas, todas al mejor de 7 partidos.

Los datos utilizados se han obtenido de Basketball-Reference, desde el repositorio [kaggle](#)). Este conjunto de datos ofrece una completa colección de estadísticas históricas de la NBA, obtenidas de Basketball Reference, una fuente líder en análisis de baloncesto. Incluye estadísticas de jugadores, métricas de rendimiento de equipos y registros históricos de la NBA, ABA y BAA, cubriendo más de 74,000 partidos y más de 30 equipos. También incluye estadísticas de jugadores de todos los tiempos y novatos por año. Todos los conjuntos de

datos están disponibles en un repositorio de GitHub, con un glosario detallado para la descripción de las columnas.

En este trabajo se han utilizado los datos procedentes de los boxscores desde la temporada 1983-1984 hasta la temporada 2023-2024. Los datos de cada temporada están divididos en dos ficheros csv. Uno que contiene datos basicos y otro que contiene datos avanzados. La nomenclatura utilizada en los ficheros es “NBA_AñoInicio-AñoFinal_basic.csv” y “NBA_AñoInicio-AñoFinal_advanced.csv”

La estructura de ambos ficheros está explicada en el anexo.

4.2. PREPROCESADO DE LOS DATOS

A estos dos conjuntos de datos se les hace una serie de transformaciones ya que los datos crudos tal cual aparecen en el repositorio no permiten resolver el problema planteado ni generar los modelos de forma fiable.

Se eliminan los datos de los partidos que hayan tenido prorrogas ya que se considerán posibles outliers de nuestro outcome objetivo que es el número de puntos.

Se filtran los datos de los jugadores que hayan jugado mas de 4 minutos

Se suman para cada equipo y cada partido todos los valores del boxscore como tiros de campo, rebotes y puntos.

Se calcula el porcentaje de jugadores con más de 10 puntos.

En este momento se tienen, los datos de boxscore, por equipo y partido, resumidos a través de los totales, con lo cuál en el nuevo conjunto de datos básico, habrá dos filas por partido, una con los datos de un equipo y otra con los datos del otro equipo.

En el fichero de datos avanzados se usa una estrategia similar, pero en este caso en lugar de calcular la suma de los datos, se usa la media.

Se calculan las medias de las métricas avanzadas como TS%, eFG% y ORtg.

En este caso los estadísticos avanzados vienen en formato porcentaje, así que no tiene sentido hacer la suma de todos los porcentajes de los jugadores, como si que lo tenía en los estadísticos básicos (Número de rebotes, Número de asistencias, Número de Robos...), por

ello se considerá mas conveniente reducir el dataset por equipo y partido a través de las medias.

Los datos básicos y avanzados se combinan en un solo conjunto de datos utilizando una unión por el identificador del partido y el equipo, proporcionando un conjunto de datos completo para el análisis posterior.

Se duplican y renombran las columnas del conjunto de datos completo (data_complete) para diferenciarlas entre un equipo y el rival (df_riv). Luego,

Se realiza una auto-unión (self-join) en la que se combinan los datos del equipo con los del rival basándose en el identificador del partido (Game Reference), y se filtran las filas para asegurar que los equipos comparados sean diferentes.

Con esto se tiene un conjunto de datos único, en el que los datos de cada partido están en dos filas, y cada fila contiene los datos de los dos equipos, así como del rival. En las columnas que tienen la terminación “Eq” contienen los datos de un equipo y en las que terminan en “Riv” las del rival.

Finalmente,

Se calculan las estadísticas de tiro de dos puntos ya que no aparecían en los datos de kaggle (2P, 2PA, 2P%) tanto para el equipo como para el rival, restando los tiros de tres puntos de los tiros de campo totales y luego calculando los porcentajes de tiro de dos puntos (2P%).

Para enriquecer nuestro conjunto de datos

Se calculan estadísticos avanzados ya que estos pueden ser importantes tanto para generar los modelos predictivos de este trabajo, como para en el futuro poder generar otros modelos que mejores a los realizados aquí.

Porcentaje de Tiro Efectivo (eFG%)

El **Porcentaje de Tiro Efectivo (eFG%)** ajusta el porcentaje de tiros de campo al valorar los tiros de tres puntos como 1.5 veces más valiosos que los de dos puntos.

$$\text{eFG\%} = \frac{T2C + T3C + 0.5 \times T3I}{T2I + T3I} \times 100$$

Donde:

T2C: Tiros de 2 puntos convertidos.

T2I: Tiros de 2 puntos intentados.

T3C: Tiros de 3 puntos convertidos.

T3I: Tiros de 3 puntos intentados.

Porcentaje de Pérdidas de Balón (TO%)

El **Porcentaje de Pérdidas de Balón (TO%)** mide la proporción de posesiones que terminan en una pérdida de balón.

$$TO\% = \frac{TO}{T2I + T3I + 0.44 \times TLI + TO} \times 100$$

Donde:

TO: Pérdidas de balón.

TLI: Tiros libres intentados.

Porcentaje de Rebotes Ofensivos (ORB%)

El **Porcentaje de Rebotes Ofensivos (ORB%)** indica la capacidad de un equipo para capturar rebotes ofensivos comparado con los rebotes defensivos del oponente.

$$ORB\% = \frac{ORB}{ORB + DRB_{opp}} \times 100$$

Donde:

ORB: Rebotes ofensivos.

DRB_{opp}: Rebotes defensivos del oponente.

Porcentaje de Rebotes Defensivos (DRB%)

El **Porcentaje de Rebotes Defensivos (DRB%)** mide la capacidad de un equipo para asegurar rebotes defensivos.

$$DRB\% = \frac{DRB}{DRB + ORB_{opp}} \times 100$$

Donde:

DRB: Rebotes defensivos.

ORB_{opp} : Rebotes ofensivos del oponente.

Factor Oliver de Tiro Libre (FT Rate)

El **Factor de Tiro Libre** evalúa la eficacia de un equipo en convertir tiros libres en relación con sus intentos de tiros de campo.

$$\text{FT Rate} = \frac{TLC}{T2I + T3I} \times 100$$

Donde:

TLC : Tiros libres convertidos.

Rating Ofensivo y Defensivo (Off/Def Rating)

El **Rating Ofensivo y Defensivo** mide los puntos anotados o permitidos por 100 posesiones.

$$\text{Off/Def Rating} = \frac{\text{Puntos}}{T2I + T3I + TO + 0.44 \times TLI - ORB} \times 100$$

Donde:

Puntos: Puntos anotados por el equipo.

Número de Posesiones

El número de posesiones en un partido se calcula con la siguiente fórmula:

$$\text{Posesiones} = T2I + T3I + 0.44 \times TLI - ORB + TO$$

Porcentaje de Tiros Verdaderos (TS%)

El **Porcentaje de Tiros Verdaderos (TS%)** mide la eficiencia de un equipo considerando tiros de campo, triples y tiros libres.

$$\text{TS\%} = \frac{\text{Puntos}}{2 \times (T2I + T3I + 0.44 \times TLI)} \times 100$$

Donde:

Puntos: Puntos anotados.

4.3. MODELOS PREDICTIVOS ELEGIDOS

4.3.1. Random Forest

En este trabajo para garantizar resultados generalizables y evaluar la estabilidad del modelos a lo largo de diferentes subconjuntos de datos se ha utilizado validación cruzada con 5 particiones, generadas a partir de los datos de entrenamiento. Se controlarán dos parámetros para encontrar el mejor Random Forest, por un lado el parámetro que determina el número de variables consideradas en cada división del árbol, utilizando los valores 2, 5 y 10. Además, por otro lado se fijará el número de árboles del random forest en 500.

Las principales ventajas de Random Forest incluyen su versatilidad, ya que puede utilizarse tanto para tareas de regresión como de clasificación. Es fácil interpretar la importancia que asigna a cada característica de entrada, y los hiperparámetros que utiliza son sencillos de entender y ajustar. Además, el algoritmo generalmente evita el problema del sobreajuste, siempre que se utilicen suficientes árboles en el bosque, lo que lo convierte en una herramienta eficaz y confiable en ML.

La principal limitación de Random Forest es que, al necesitar un gran número de árboles, puede volverse demasiado lento e ineficaz para hacer predicciones en tiempo real. Aunque estos algoritmos son rápidos de entrenar, pueden ser lentos al generar predicciones una vez entrenados. Cuantos más árboles se utilicen para lograr una predicción precisa, más lento será el modelo. En la mayoría de las aplicaciones, Random Forest es lo suficientemente rápido, pero en situaciones donde el rendimiento en tiempo real sea importante, habría que probar otro tipo de algoritmos. ([19](#))

4.3.2. Red Neuronal

En este trabajo se ha entrenado una red neuronal de manera que para encontrar la mejor red, se utiliza un grid de hiperparámetros, que especifica diferentes combinaciones de los parámetros *size* (el número de neuronas en la capa oculta) y *decay* (la tasa de decaimiento para la regularización).

En este caso, se prueban valores de *size* que van desde 2 hasta 10, y valores de *decay* que incluyen 0, 0.001, 0.01, 0.1, y 0.5. Además se ha establecido que el número máximo de iteraciones para el algoritmo sea de 1000.

Este grid se utiliza para encontrar la mejor combinación que minimice el error cuadrático medio RMSE, que es la métrica elegida para optimizar, como se hizo con el modelo de Random Forest

Las redes neuronales tienen tanto ventajas como desventajas. Entre las ventajas, destacan su alta capacidad de aprendizaje, flexibilidad para trabajar con diversos tipos de datos (textos, imágenes, audio, etc.), su capacidad para crear modelos no lineales, y la posibilidad de entrenarse de manera continua con nuevas muestras. Sin embargo, presentan desventajas significativas, como la necesidad de grandes cantidades de datos para su entrenamiento, así como ser un modelo complejo de caja negra, ya que su complejidad dificulta su interpretación, la tendencia al sobreajuste, y la dificultad de optimización debido al alto número de hiperparámetros a ajustar. Además, son sensibles a la escala de los datos, por lo que es necesario preprocesarlos adecuadamente ([20](#))

4.4. MÉTRICAS DE EVALUACIÓN DEL MODELO

En cualquier modelo de ML es fundamental evaluar la precisión del modelo. Las métricas utilizadas para evaluar el rendimiento del modelo en Random Forest y en Redes Neuronales cuando se trata de predecir una variable numérica y no una categoría, incluyen el Error Medio Absoluto (MAE), el Error Cuadrático Medio (MSE), la Raíz del Error Cuadrático Medio (RMSE) y en el caso de modelos de regresión el Coeficiente de Determinación (R^2). Además de estas, se ha diseñado una métrica, equivalente a un Accuracy, que calcula el porcentaje de observaciones que son predichas en unos márgenes prefijados anteriormente. En este trabajo se evaluarán los modelos con el RMSE y la Precisión dentro de un porcentaje.

Diferencias Entre las Métricas de Evaluación

- El Error Cuadrático Medio (MSE) y la Raíz del Error Cuadrático Medio (RMSE) penalizan más los grandes errores de predicción en comparación con el Error Medio Absoluto (MAE). Sin embargo, el RMSE se utiliza más que el MSE para evaluar el rendimiento de los modelos de regresión porque tiene las mismas unidades que la variable dependiente (eje Y).
- El RMSE es más intuitivo que el MSE, pero a nivel computacional es más costoso. En aprendizaje supervisado como arquitecturas de redes neuronales el error tiene que

ser calculado en cada iteración. Por esta razón, al ser *RMSE* más costoso es preferible usar *MSE* como función de coste.

- Un valor más bajo de *MAE*, *MSE* y *RMSE* implica una mayor precisión del modelo de regresión.

4.4.1.1. Conclusión

Tanto el *RMSE* como el *MSE* como el *MAE* nos sirven para cuantificar lo bien que un modelo se ajusta a los datos.

En nuestro caso que, nuestro objetivo es predecir el número de puntos se utilizará como indicador de calidad del modelo tanto el *RMSE* como la precisión dentro de un porcentaje.

4.5. METODOS DE EXPLICABILIDAD ELEGIDOS

4.5.1. Curvas de Expectativa Condicional Individual (ICE)

En nuestro caso se ha entrenado un modelo de Random Forest y una Red Neuronal para predecir el número de puntos anotados por los equipos en función de varios predictores.

Una vez entrenado el modelo, se utilizan las Curvas de Expectativa Condicional Individual (ICE) para analizar cómo varía la predicción de puntos para diferentes equipos en función de uno de los predictores, por ejemplo, el porcentaje de tiros de tres acertados, manteniendo los demás predictores constantes.

En este caso, cada curva ICE muestra cómo la predicción del número de puntos cambia para un equipo específico a medida que el porcentaje de triples anotados varía de su rango más bajo a su rango más alto. A diferencia de los gráficos de dependencia parcial (PDP), las curvas ICE permiten observar cómo esta relación puede ser diferente para cada equipo.

4.5.2. Counterfactuals Values

En este trabajo se han determinado los contrafácticos para que la nueva predicción del modelo esté dentro de un intervalo que sea el 5% y el 10% de puntos mas. Es decir, si la predicción del modelo ha sido de 100 puntos, como deben variar los predictores para que la predicción del modelo estuviera entre 105 y 110 puntos.

Un enfoque simple e ingenuo para generar explicaciones contrafácticas es buscar por ensayo y error. Este método consiste en cambiar aleatoriamente los valores de las características de la instancia de interés y detenerse cuando se obtiene la predicción deseada. Sin embargo, existen enfoques más eficientes que el simple ensayo y error.

Las explicaciones contrafácticas son claras y directas, ya que muestran cómo cambiarían las predicciones de un modelo al modificar ciertos valores de las características de una instancia, sin necesidad de suposiciones adicionales. El método contrafáctico no requiere acceso a los datos o al modelo. Solo requiere acceso a la función de predicción del modelo. Aunque fácil de implementar, presenta problemas como la existencia de múltiples explicaciones posibles (efecto Rashomon) y dificultades para manejar características categóricas complejas. Además, no garantiza encontrar una explicación válida para cualquier nivel de tolerancia dado, lo que puede limitar su utilidad en algunos casos.

4.5.3. Valores de Shapley y SHAP:

Para interpretar los valores de Shapley en un modelo predictivo de Red Neuronal o Random Forest que predice el número de puntos anotados en un partido de baloncesto, se descompone la predicción del modelo en contribuciones individuales de cada predictor. Por ejemplo, si el modelo predice que un equipo anotará 120 puntos, y la media general de puntos predichos es de 110, los valores de Shapley explican cómo cada predictor contribuye a esta diferencia de 10 puntos adicionales. Un valor de Shapley positivo indica que una característica específica aumentó la predicción, mientras que un valor negativo indica que la característica disminuyó la predicción. Por ejemplo, si el porcentaje de tiro de campo (FG%) tiene un alto valor de Shapley, significa que un mejor porcentaje de tiro contribuyó significativamente a la alta predicción de puntos. Por otro lado, una alta cantidad de pérdidas (TO) con un valor de Shapley negativo sugiere que las pérdidas de balón redujeron la predicción. Así, los valores de Shapley proporcionan una visión clara de cómo cada variable influyó en la predicción final del modelo.

Los valores de Shapley presentan varias ventajas. Primero, distribuyen de manera justa la diferencia entre la predicción y la predicción promedio entre todas las características, garantizando la eficiencia, lo que los distingue de otros métodos, que no siempre aseguran una distribución equitativa. Además, los valores de Shapley son el único método de explicación respaldado por una teoría sólida, fundamentada en axiomas como eficiencia,

simetría, dummies y aditividad. Esta solidez teórica los hace potencialmente compatibles con requisitos legales de explicabilidad, como el “derecho a explicaciones” de la UE. Otro beneficio es que permiten realizar explicaciones contrastantes, es decir, comparar una predicción con un subconjunto específico o incluso con un punto de datos individual, algo que modelos locales no ofrecen. Por último, los valores de Shapley tienen una base teórica clara, a diferencia de otros métodos locales, que asumen un comportamiento lineal local del modelo sin una justificación teórica sólida.

Sin embargo, los valores de Shapley también presentan inconvenientes. Uno de los principales es su alto costo computacional, ya que calcularlos requiere evaluar todas las coaliciones posibles de características, lo cual es extremadamente costoso en términos de tiempo de cómputo. En la mayoría de los casos, solo es factible una solución aproximada, lo que aumenta la varianza del resultado. Además, los valores de Shapley pueden malinterpretarse, ya que no deben ser vistos como el cambio en la predicción al eliminar una característica del modelo, sino como la contribución de una característica específica, dada la combinación actual de valores de características. Otro inconveniente es que el método de Shapley no es adecuado para explicaciones que utilicen pocas características, ya que siempre incluye todas las características. Además, para calcular los valores de Shapley de una nueva instancia, se necesita acceso a los datos originales, lo que limita su aplicabilidad cuando el acceso a los datos es restringido. Finalmente, cuando las características están correlacionadas, el método de Shapley puede generar valores poco realistas, ya que margina las características de forma independiente, lo que puede llevar a resultados que no reflejan correctamente la relación entre las características.

5. RESULTADOS

Se importaron datos de 47,741 partidos, sumando un total de 95,482 registros entre las temporadas 1983-1984 y 2023-2024. Se realizó un análisis gráfico de la distribución de puntos por temporada para identificar tendencias. La figura muestra una disminución de puntos hasta 1998-1999, seguida de una meseta de casi una década, y a partir de 2010, un aumento constante. Este cambio se relaciona con un ritmo de juego más lento y menos posesiones hasta 1999, y con el aumento progresivo del ritmo desde entonces. Además, el lanzamiento de tres puntos, introducido en 1980/81, ha ganado relevancia, con un crecimiento notable desde 2010-2011, como se muestra en las dos siguientes gráficas.

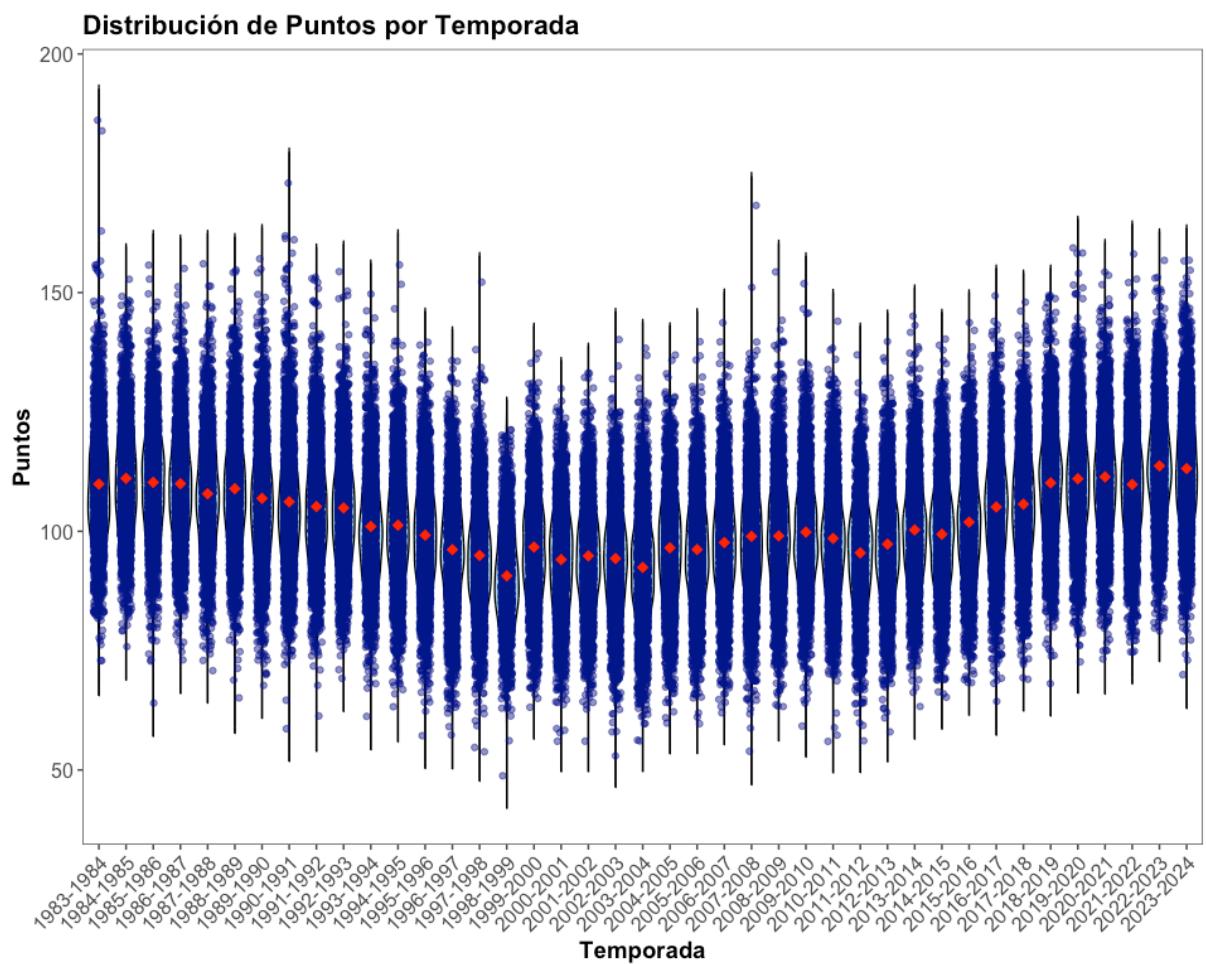


Figura 5.1: Distribución de los puntos anotados desde la temporada 1983 hasta la actualidad. El punto rojo es la media de los puntos anotados en cada temporada

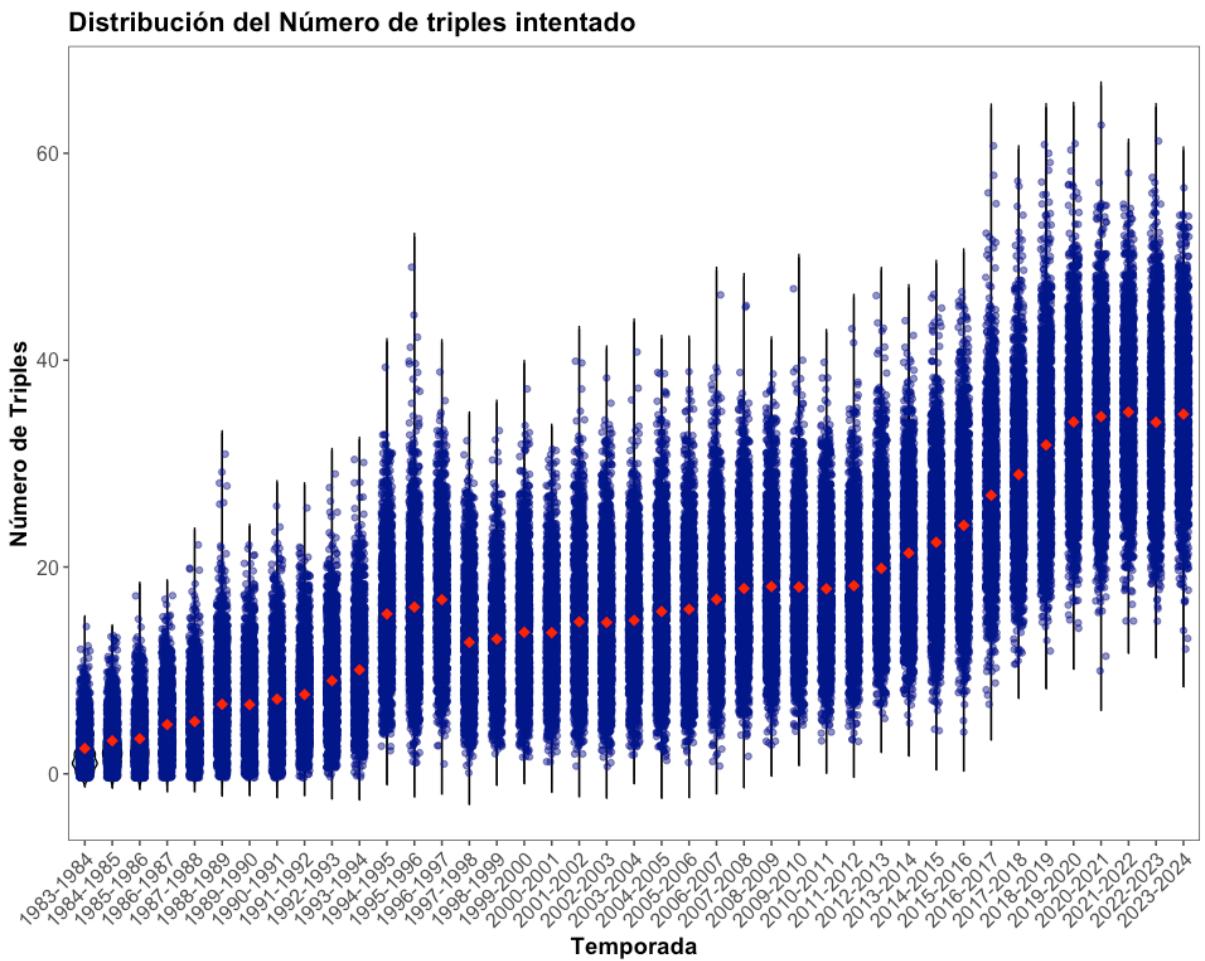


Figura 5.2: Distribución del número de triples intentados desde la temporada 1983 hasta la actualidad. El punto rojo es la media de los tiros de tres intentados en cada temporada

Tras ver estos resultados se decide abordar dos estrategias en cuanto al conjunto de datos a utilizar. Por un lado, se analizan los partidos desde la temporada 2000 en adelante, y se hace un subanálisis de estos datos utilizando únicamente los datos del año 2018 hasta la actualidad.

Desde el año 2000 se analiza la información de 27760 partidos de las NBA mientras que el número de partidos desde 2018 son 5923.

De todos los indicadores disponibles y construidos, se analiza la correlación numérica entre ellos para elegir parámetros que sean independientes entre sí y que no tengan una correlación alta.

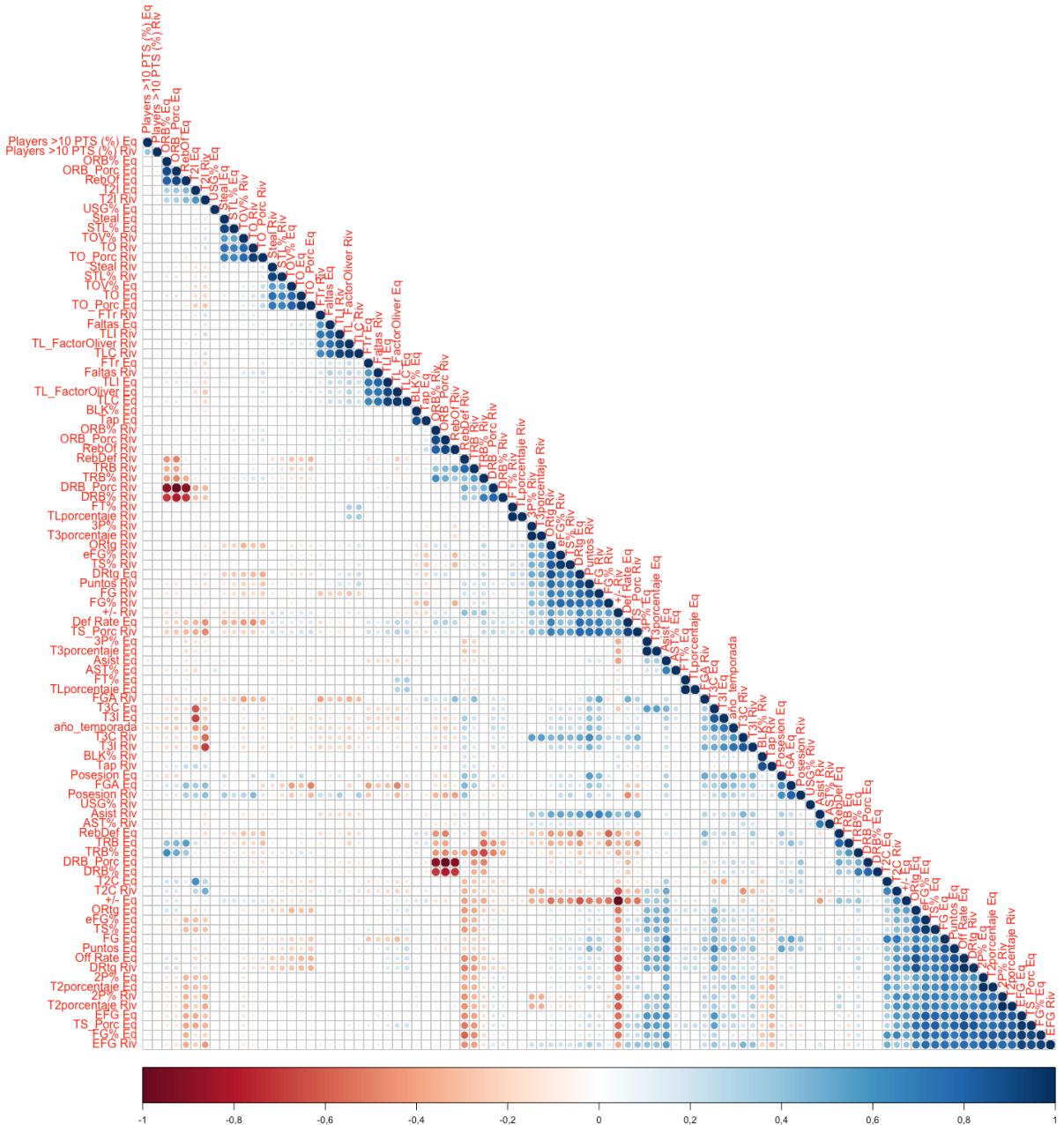


Figura 5.3: Correlaciones entre parámetros disponibles

Basado tanto en las correlaciones observadas en la figura 5.3 como en el conocimiento experto del juego, se seleccionaron las variables que presentan una relación significativa con

el rendimiento en baloncesto. Estas incluyen el porcentaje de acierto en el tiro de tres, el número de asistencias, el porcentaje de rebote ofensivo, el porcentaje de rebote defensivo, el número de tiros de campo intentados, el porcentaje de jugadores que anotan más de 10 puntos en el equipo, el número de posesiones, el número de robos y el porcentaje de pérdidas. Estas variables capturan aspectos clave sobre todo en el desempeño ofensivo y en menor medida, defensivo del equipo. Estos indicadores son seleccionados de los datos del equipo que se quiere predecir los puntos anotados. Del equipo rival se selecciona también, el porcentaje de perdidas y el número de robos

En la siguiente tabla se muestran los estadísticos descriptivos de los indicadores seleccionados en los dos datasets elegidos. Se observan ligeras diferencias en el número de asistencias, en el número de tiros de campo y en el número de posesiones. También se observa que ha habido un aumento en los puntos anotados. Si estudia desde el año 2000 la mediana son 101 puntos (RIQ 92-111) frente a los 112 puntos de mediana (RIQ 103-120) en el periodo 2018-2023.

Tabla 5.1: Descripción de los datasets elegidos para generar los modelos predictivos

Characteristic	2000-2023	2018-2023
	N = 57 870¹	N = 14 328¹
3P% Eq	0,35 (0,28, 0,43)	0,36 (0,30, 0,41)
Asist Eq	22,0 (19,0, 26,0)	25,0 (21,0, 28,0)
DRB_Porc Eq	75 (69, 80)	77 (72, 82)
ORB_Porc Eq	25 (20, 31)	23 (18, 28)
FGA Eq	83 (78, 88)	88 (83, 92)
Players >10 PTS (%) Eq	0,44 (0,36, 0,54)	0,46 (0,40, 0,56)
Posesion Eq	96 (92, 100)	100,6 (97,2, 104,1)
Steal Eq	7,0 (5,0, 9,0)	7,00 (5,00, 9,00)
Steal Riv	7,0 (5,0, 9,0)	7,00 (5,00, 9,00)
TO_Porc Eq	12,6 (10,3, 15,0)	11,7 (9,5, 14,1)
TO_Porc Riv	12,6 (10,5, 14,8)	11,8 (9,7, 13,9)
Puntos Eq	101 (92, 111)	111 (103, 120)

¹Median (Q1, Q3)

5.1. MODELOS PREDICTIVOS

Se han entrenado dos modelos predictivos, Random Forest y Red Neuronal.

En ambos casos se han generado a su vez dos modelos, uno utilizando los datos desde el año 2000 y otro desde el año 2018.

En los cuatro casos se ha entrenado el modelo con un 80% de los datos disponibles y se ha hecho el test del modelo con el otro 20%.

Tabla 5.2: Comparación entre los datos de entrenamiento de los modelos y los datos de testeо

Characteristic	2000-2023			2018-2023		
	test N = 11 572 ¹	training N = 46 298 ¹	p-value ²	test N = 2 863 ¹	training N = 11 465 ¹	p-value ²
X3P..Eq	0,35 (0,28, 0,43)	0,35 (0,29, 0,43)	0,2	0,36 (0,30, 0,41)	0,36 (0,30, 0,41)	0,9
Asist.Eq	22,0 (19,0, 26,0)	22,0 (19,0, 26,0)	0,5	25,0 (21,0, 28,0)	25,0 (21,0, 28,0)	0,8
DRB_Porc.Eq	75 (70, 80)	75 (69, 80)	0,053	77 (72, 82)	77 (72, 82)	0,6
ORB_Porc.Eq	25 (20, 31)	25 (20, 31)	0,6	23 (18, 28)	23 (18, 28)	0,072
FGA.Eq	83 (78, 88)	83 (78, 88)	0,3	88 (83, 92)	88 (83, 92)	0,4
Players..10.PTS.....Eq	0,44 (0,36, 0,54)	0,44 (0,36, 0,54)	0,6	0,45 (0,40, 0,56)	0,46 (0,40, 0,56)	0,6
Posesion.Eq	96 (92, 100)	96 (91, 100)	0,6	100,6 (97,2, 104,2)	100,6 (97,2, 104,1)	0,7
Steal.Eq	7,0 (5,0, 9,0)	7,0 (5,0, 9,0)	0,7	7,00 (5,00, 9,00)	7,00 (5,00, 9,00)	0,6
Steal.Riv	7,0 (5,0, 9,0)	7,0 (5,0, 9,0)	0,2	7,00 (5,00, 9,00)	7,00 (5,00, 9,00)	>0,9
TO_Porc.Eq	12,6 (10,2, 15,0)	12,6 (10,3, 15,0)	0,2	11,7 (9,6, 14,0)	11,7 (9,5, 14,1)	>0,9
TO_Porc.Riv	12,6 (10,4, 14,8)	12,6 (10,5, 14,8)	0,9	11,8 (9,7, 13,8)	11,8 (9,7, 13,9)	0,3
Puntos.Eq	101 (92, 111)	101 (92, 111)	>0,9	111 (103, 120)	111 (103, 120)	0,9

¹Median (Q1, Q3)²Wilcoxon rank sum test

Como se observa en la tabla 5.2, ambos conjuntos de datos, test y training, son similares y no se observan diferencias estadísticas significativas.

Se ha entrenado un modelo de Random Forest utilizando validación cruzada con 5 particiones que se generan a partir de los propios datos de entrenamiento. El objetivo de este modelo es predecir el número de puntos de un equipo en dos conjuntos de datos de baloncesto. Se ajustaron tres valores para el hiperparámetro mtry (número de variables a considerar en cada árbol, que fueron 2, 5, 10) y se optimizó el modelo según el RMSE. El preprocessamiento incluyó escalado y centrado de los datos, y se usaron 500 árboles. Además, se implementó una parallelización con 5 núcleos para acelerar el proceso y además se obtuvo la importancia de las variables.

Para la red neuronal también se utilizó validación cruzada. Se ajustaron los hiperparámetros size (2 a 6) y decay (0.5 y 0.1) mediante un grid de búsqueda, optimizando el modelo según el RMSE. El preprocessamiento, incluyó escalado y centrado, y se configuraron hasta 1000 iteraciones para el entrenamiento. La parallelización fue usada para acelerar el proceso y se extrajo la importancia de las variables.

Los resultados en el conjunto de training son los siguientes

Tabla 5.3: Rendimiento de los modelos generados con los datos de entrenamiento en Random Forest

mtry	2000-2024		2018-2024	
	RMSE	MAE	RMSE	MAE
2	7.39	5.90	7.25	5.77
5	7.16	5.72	7.04	5.61
10	7.12	5.68	6.99	5.56

En las redes neuronales

Tabla 5.4: Rendimiento de los modelos generados con los datos de entrenamiento en Red Neuronal

size	decay	2000-2024		2018-204	
		RMSE	MAE	RMSE	MAE
2	0.000	7.97	6.35	10.46	5.64
2	0.001	6.65	5.27	5.97	4.75
2	0.010	6.59	5.23	5.92	4.75
2	0.100	6.56	5.20	5.92	4.73
2	0.500	6.57	5.21	5.88	4.71
4	0.000	6.55	5.19	5.96	4.73
4	0.001	6.77	5.20	6.37	5.07
4	0.010	6.53	5.18	5.99	4.79
4	0.100	6.53	5.18	5.94	4.76
4	0.500	6.47	5.13	5.92	4.73
6	0.000	7.65	5.26	8.15	5.00
6	0.001	6.62	5.17	6.34	4.84
6	0.010	6.50	5.15	6.07	4.81
6	0.100	6.52	5.17	5.98	4.76
6	0.500	6.44	5.11	5.98	4.76
8	0.000	6.53	5.17	6.22	4.91
8	0.001	6.51	5.15	6.14	4.90
8	0.010	6.46	5.13	6.22	4.89
8	0.100	6.45	5.12	6.14	4.89
8	0.500	6.48	5.13	6.13	4.88
10	0.000	6.48	5.13	6.45	5.07
10	0.001	6.50	5.15	6.50	5.02
10	0.010	6.50	5.15	6.24	4.98
10	0.100	6.47	5.13	6.29	4.97
10	0.500	6.47	5.14	6.19	4.94

Un valor inferior de MAE y RMSE significa que el modelo tiene un error en la predicción más pequeño, bajo esta premisa, los modelos que se eligieron para testear fueron:

- Random Forest con el parámetro mtry establecido en 10, tanto para el análisis desde el año 2000 al 2024, como del 2018 al 2024
- Red Neuronal para el conjunto de datos desde el año 2000 con un tamaño de 6 y un valor de decay de 0.5, y en el caso de los datos tomados desde el 2018, el tamaño de la red fue de 2 y decay 0.5

5.1.1. Test

Se testean los modelos con el 20% de los datos disponibles. En las siguientes gráficas se puede observar una correlación lineal moderada entre puntos anotados y los puntos anotados predichos por el modelo. Una correlación lineal positiva indica que, en general, a medida que los valores reales aumentan, los valores predichos también tienden a aumentar. Esto sugiere que el modelo captura correctamente la tendencia general de los datos. Aunque no es un indicador definitivo de la calidad del modelo, un buen valor de correlación puede ser un buen indicativo de que el modelo tiene el potencial de ser útil.

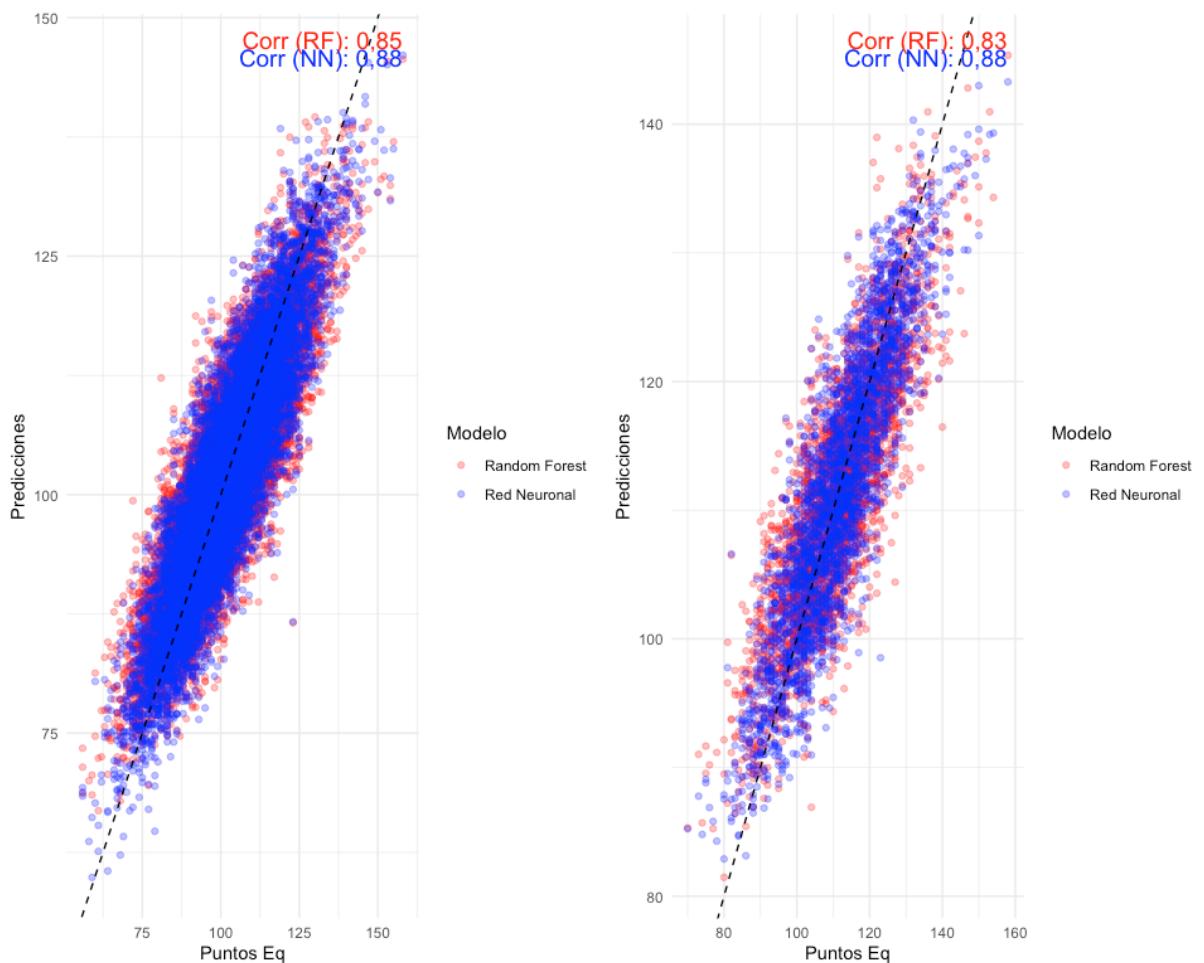


Figura 5.4: Gráfico de dispersión entre los valores reales y los valores predichos por los modelos generados. El gráfico de la izquierda representa los modelos generados con datos desde el año 2000 y el gráfico de la derecha los los modelos generados con datos desde el año 2018. En el eje horizontal se representan los valores reales y en el eje vertical las predicciones de los modelos. En rojo se ven las predicciones del Random Forest y en azul las predicciones de la Red Neuronal

Tabla 5.5: Rendimiento de los modelos con los datos de Test

Modelo	RMSE	Accuracy 5%	Accuracy 10%
RF 2000	7.06	53.53	84.95
RN 2000	6.37	57.79	88.79
RF 2018	6.82	58.92	89.28
RN 2018	5.86	65.49	94.20

Finalmente, el mejor modelo fue la red neuronal con los datos desde el 2018 al 2023, con el RMSE más bajo (5.86) y las mejores precisiones: 65.5% dentro del 5% y 94.2% dentro del 10%.

5.1.2. Importancia Variables

En el siguiente gráfico representamos la importancia de las variables en cada modelo, que ayudan a entender como las diferentes características estudiadas influyen en las predicciones generadas.

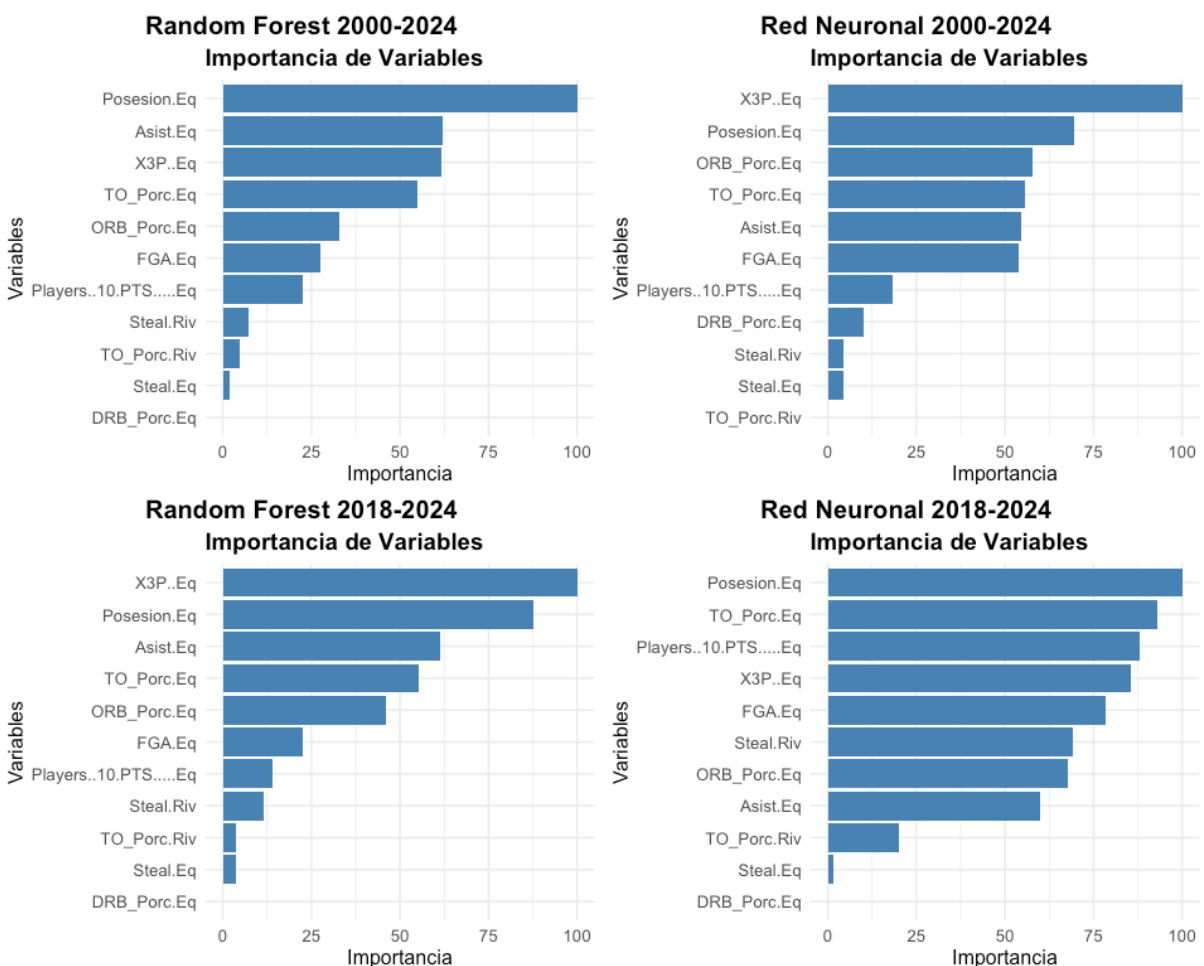


Figura 5.5: Importancia de las variables en los cuatro modelos generados

En el análisis de la importancia de las variables entre los modelos de Random Forest (rf2000, random forest entrenado con datos desde el año 2000 y rf2018, random forest entrenado con datos desde el año 2018) y Redes Neuronales (nn2000, red neuronal entrenada con datos desde el año 2000, y nn2018, red neuronal entrenada con datos desde 2018), se pueden observar tanto similitudes como diferencias significativas. Entre los puntos en común, tanto en rf2000 como en nn2000, la variable posesiones del equipo se destaca como una de las más relevantes, ocupando posiciones elevadas en ambos modelos. Asimismo, el porcentaje de triples anotados, es la variable más importante en nn2000 y también figura entre las más relevantes en rf2018, indicando su consistentemente alta influencia en las predicciones. Sin

embargo, las diferencias son notables: en rf2018, el porcentaje de triples lidera la importancia, mientras que en rf2000, es el número de posesiones es la más destacada. En el caso de nn2018, número de posesiones también ocupa el primer lugar, pero el porcentaje de perdidas le sigue de cerca, mostrando su creciente relevancia en comparación con los otros modelos. Por otro lado, variables como porcentaje de rebotes defensivos, los robos del equipo y el porcentaje de balones perdidos del rival, muestran una importancia mínima en todos los modelos, lo que sugiere que su influencia en las predicciones es escasa y constante a lo largo de los años. Estas observaciones resaltan cómo las dinámicas del juego pueden cambiar con el tiempo y cómo diferentes modelos pueden resaltar distintas variables clave.

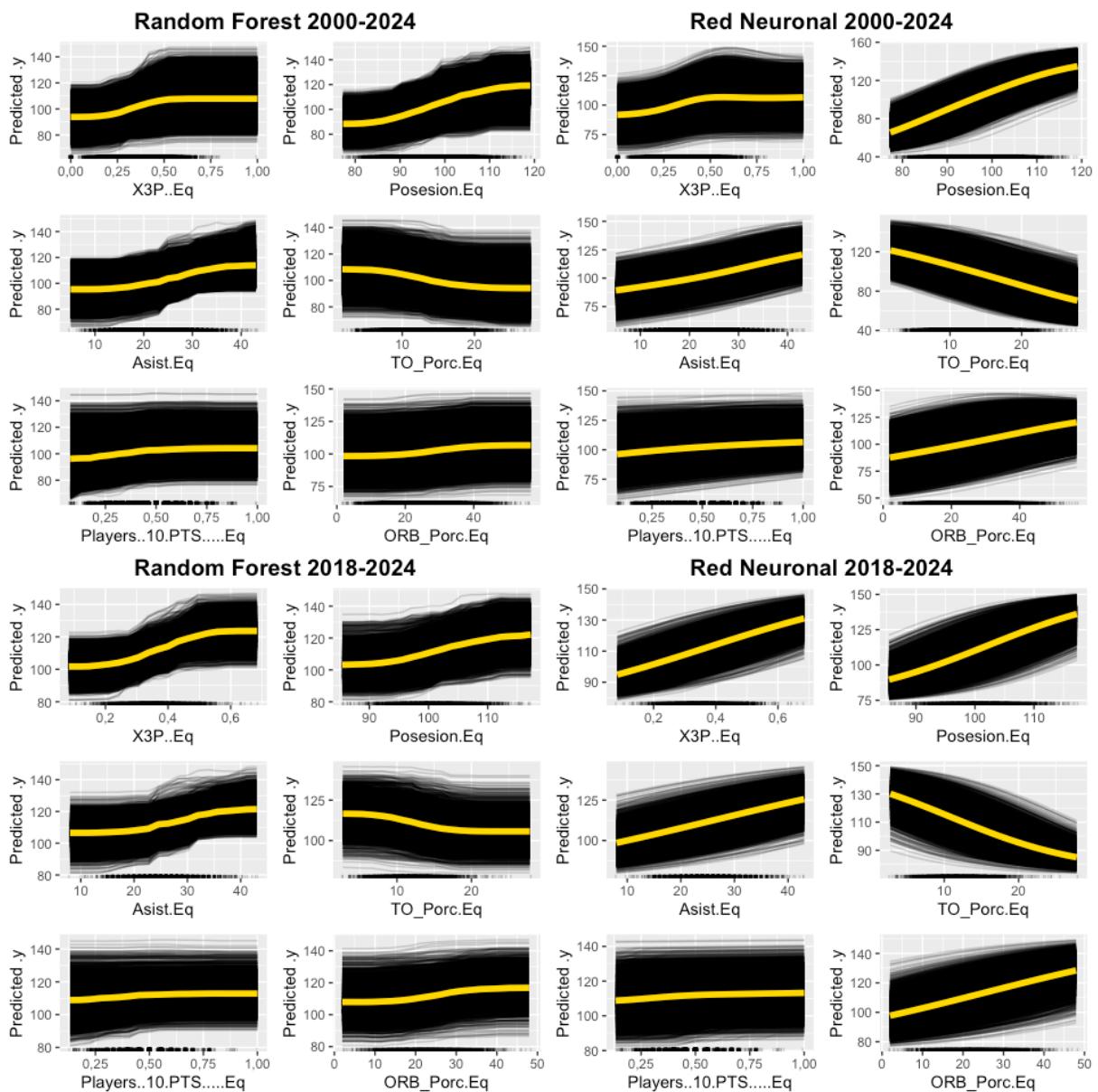
5.2. CURVAS DE EXPECTATIVA CONDICIONAL INDIVIDUAL (ICE)

Cada curva de los gráficos en la figura siguiente (curvas ICE) muestran el valor predicho de la puntos anotados para cada observación conforme se va aumentando el valor de cada uno de los predictores, mientra se mantienen constantes el resto de predictores del modelo en su valor observado. La curva resaltada en amarillo se corresponde con la curva PDP, que es la variación promedio de todas las observaciones.

En el caso del modelo Random Forest con datos entre 2000 y 2024, en el análisis de las variables porcentaje de jugadores que meten mas de diez puntos y rebotes ofensivos, las curvas son prácticamente planas, lo que apunta a que, en la mayor parte de los casos, apenas influye en el número de puntos anotados. Se observa que conforme aumenta el número de posesiones, sobre todo a partir de 90 posesiones aparece un aumento lineal en el número de puntos predichos. En el caso del porcentaje de triples anotados aumentar del 25% al 30% podemos ver un aumento de 90 a 100 puntos anotados predichos. En el caso de la Red Neuronal con datos entre 2000 y 2024, la curva asociada al porcentaje de triples es similar al del random forest. Si que se percibe, que las curvas ICE del número de posesiones es totalmente lineal, así como el porcentaje de perdidas (en este caso lineal negativa). En el modelo de Red Neuronal con datos entre 2018 y 2024, la mayoría de las variables tienen líneas paralelas con una tendencia lineal positiva o, negativa en el caso del porcentaje de perdidas, lo que significa que, para esas variables, el impacto en la predicción es bastante consistente para todos los partidos de los datos de test, independientemente del valor específico de la variable. En otras palabras, la relación entre esa variable y la predicción es uniforme: a medida

que la variable aumenta (o disminuye), la predicción cambia de manera predecible en la misma dirección para todos los individuos.

En todos los casos las líneas paralelas sugieren que el modelo no está capturando interacciones complejas entre esa variable y otras variables. Si el comportamiento de la variable es el mismo para todos los individuos, puede ser que el modelo no esté considerando adecuadamente interacciones no lineales o condicionales con otras variables que podrían ser importantes.



5.3. CONTERFACTUAL VALUES (VALORES CONTRAFÁCTICOS)

Para los cuatro modelos generados se han calculado los valores contrafácticos para todos los datos de testeo. Los contrafácticos se han calculado para todos los datos de testeo con el objetivo de obtener predicciones que generen una puntuación entre un 5% y un 10% superior al valor predicho originalmente por el modelo. En la siguiente tabla se muestran la media, desviación típica y coeficiente de variación con el objetivo de evaluar cuáles muestran más variación a lo largo de las predicciones realizadas y poder analizar en cada modelo cuáles son más estables y cuáles presentan cambios más significativos a lo largo de los datos de testeo. Valores contrafácticos con mucha dispersión indican que hay una gran variabilidad en las condiciones bajo las cuales el modelo puede cambiar su predicción hacia un valor deseado (en este caso, aumentar entre un 5% y un 10%). Esto sugiere que el modelo es sensible a múltiples combinaciones de características para lograr ese ajuste en la predicción. Un valor contrafáctico con alta dispersión puede sugerir que no es fácil identificar acciones claras respecto al parámetro que lleven a un cambio en la predicción, lo que complica la interpretación y la aplicabilidad del modelo para realizar ajustes o mejoras específicas. Un valor contrafáctico con baja dispersión indica que el modelo responde de manera consistente a cambios en las características para lograr la predicción deseada. Esto implica que hay un patrón claro y estable bajo el cual el modelo ajusta su predicción, y que las características relevantes tienen un impacto predecible y controlado.

Tabla 5.6: Descriptivos de los valores contrafácticos de las predicciones generadas por los modelos entrenados con datos desde el año 2000

Parámetro		2000-2024		
Parámetro	Variable	Random Forest		Red Neuronal
		Media	DT	CV
X3P..Eq		0.39	0.10	26.41
Asist.Eq		23.83	5.42	22.74
DRB_Porc.Eq		74.75	7.87	10.52
ORB_Porc.Eq		25.95	7.84	30.22
FGA.Eq		82.57	7.52	9.10
Players..10.PTS.....Eq		0.46	0.12	25.62
Posesion.Eq		97.66	5.95	6.09
Steal.Eq		7.51	2.89	38.52
Steal.Riv		7.48	2.87	38.32
TO_Porc.Eq		12.29	3.33	27.13
TO_Porc.Riv		12.65	3.20	25.31

DT: Desviación Típica CV: Coeficiente de Variación en porcentaje

En los modelos generados con datos desde el año 2000, las medias de las variables entre Random Forest y la red neuronal son bastante similares, lo que indica que ambos modelos generan valores contrafactuales consistentes en la mayoría de los casos. Por ejemplo,

DRB_Porc.Eq tiene una media de 74.75 en Random Forest y 74.79 en la red neuronal. Los coeficientes de variación en general son bajos, todos muestran una variabilidad inferior al 30%. En el caso los valores contrafactuals del número de posesiones (CV=0.06) y el número de tiros intentados(CV=0.09), muestran que tienen un comportamiento estable, al igual que el porcentaje de rebotes defensivos (CV=0.11) . En ambos modelos alcanzar un número de posesiones de 97, alcanzar un número de tiros de campo de 82 o llegar al 75% del rebote defensivos, podría implicar un aumento de los puntos predichos por el modelo entre el 5% y el 10%.

Tabla 5.7: Descriptivos de los valores contrafácticos de las predicciones generadas por los modelos entrenados con datos desde el año 2018

Parámetro		2018-2024		
Parámetro	Variable	Random Forest		Red Neuronal
		Media	DT	CV
X3P..Eq		0.40	0.08	20.07
Asist.Eq		26.68	5.34	20.02
DRB_Porc.Eq		76.97	7.26	9.43
ORB_Porc.Eq		23.58	7.19	30.49
FGA.Eq		87.48	6.52	7.45
Players..10.PTS.....Eq		0.48	0.12	24.86
Posesion.Eq		101.89	4.78	4.69
Steal.Eq		7.41	2.84	38.36
Steal.Riv		7.46	2.83	37.89
TO_Porc.Eq		11.39	3.11	27.32
TO_Porc.Riv		11.77	3.07	26.09

DT: Desviación Típica

CV: Coeficiente de Variación en porcentaje

Los resultados obtenidos en cuanto a la variabilidad de los valores contrafactuals, en los modelos entrenados con datos desde 2018 a 2023 son similares a los obtenidos con los datos desde el año 2000, lo cual reflejan consistencia de los modelos obtenidos. Si que se observa el cambio de tendencia en el juego, propiciado porque los datos de entrada al modelo son diferentes, y que si que en estos dos modelos para lograr aumentar el valor predicho en el modelo entre el 5% y el 10%, el valor a alcanzar en el número de asistencias es en torno a 101, el número de tiros de campo de 87. En el caso del porcentaje de rebotes defensivo se mantiene igual que en los modelos entrenados con datos desde el año 2000.

Se calcula el cambio porcentual entre los valores contrafactuals y el valor real de los datos de testeo, esto nos sirve para cuantificar que porcentaje hay que cambiar el valor real para conseguir un cambio entre el 5% y el 10% del valor predicho. Si hay un mayor cambio porcentual para conseguir un aumento entre el 5% y el 10%, conviene menos que si con un cambio porcentual pequeño, se consigue ese mismo aumento, ya que esto indica que se requiere un mayor esfuerzo para lograr el mismo cambio. Se calcula tambien el porcentaje de

ceros, esto significa que porcentaje de datos no hay cambio entre el valor contrafáctico y el valor real del parámetro. Un porcentaje alto de ceros en un parámetro significa que para mejorar la predicción de los puntos entre un 5% y un 10% no requiere modificar ese parámetro.

Los estadísticos descriptivos del cambio porcentual muestran que las variables como el porcentaje de rebotes defensivos, número de perdidas, robos propios y del rival tienen medias cercanas a cero. Esto significa es que los valores contráfacticos no se diferencian del valor real del dato de test. Los parámetros que más interes tienen son aquellos en los que el porcentaje de ceros es menor, ya que esto significa que en un mayor número de predicciones, un cambio en el parámetro suponer un incremento de la predicción de puntos anotados entre el 5% y 10%. Se observa que tanto en Random Forest como en Red Neuronal, el porcentaje de triples anotados, el número de asistencias y el número de posesiones tienen un porcentaje de ceros inferior al 70%. De todos ellos el parámetro con una media mas baja es el número de posesiones (1.87%), esto indica que en un alto porcentaje de los datos de test analizados, un ligero cambio del número de posesiones (con un aumento medio del 1.87% y el 1.51%), se hubiera logrado un incremento entre el 5% y el 10% de los puntos anotados. En el caso del porcentaje de triples lo que se observa es que para lograr un aumento entre el 5% y el 10% en los puntos anotados, la media del cambio porcentual entre el valor contrafáctico y el valor real es de un 15%

El porcentaje de perdidas propias, es un indicador que ha pasado desapercibido, pero una reducción entorno al 3% de este indicador nos supondría un aumento entre el 5% y el 10% del número de puntos anotados. Por otro lado, variables como X3P..Eq y Asist.Eq tienen medias y desviaciones típicas más altas, lo que sugiere que se requiere un cambio porcentual más grande en sus valores reales para conseguir ese mismo aumento en la predicción. Esto implica que estas últimas variables podrían ser menos eficientes en términos de accionabilidad, ya que se necesitarían ajustes más significativos en el valor real para alcanzar el objetivo de mejora en la predicción. La columna porcentaje de ceros indica, el porcentaje de observaciones en las que no ha habido un cambio entre el valor real y el conterfactual value.

Tabla 5.8: Descriptivos del cambio porcentual entre el valor contrafáctico y el valor real, generados a partir de las predicciones generadas por los modelos entrenados con datos desde el año 2000

2000-2024								
	Random Forest				Red Neuronal			
Variable	Media	DT	% de Ceros	CV	Media	DT	% de Ceros	CV
X3P..Eq	15.48	31.69	56.26	204.65	11.45	30.84	67.59	269.37
Asist.Eq	8.36	16.51	64.86	197.42	6.51	15.61	72.24	239.63
DRB_Porc.Eq	0.07	1.66	97.15	2,347.26	0.12	2.11	96.12	1,777.91
ORB_Porc.Eq	3.39	16.41	88.41	483.42	5.85	25.55	81.53	436.63
FGA.Eq	-0.29	1.87	92.40	649.04	-0.71	3.00	81.51	421.17
Players..10.PTS.....Eq	4.31	17.39	89.69	403.16	1.37	12.09	95.72	885.26
Posesion.Eq	1.87	2.81	58.44	150.46	1.51	2.57	65.11	169.56
Steal.Eq	0.06	3.51	99.15	6,157.77	0.32	8.57	97.92	2,711.85
Steal.Riv	-0.03	4.50	98.95	13,189.23	0.31	7.51	97.91	2,432.78
TO_Porc.Eq	-2.76	8.73	83.34	316.65	-3.22	9.90	79.58	307.24
TO_Porc.Riv	-0.21	3.10	97.92	1,463.43	-0.09	3.55	96.35	4,128.50

DT: Desviación Típica CV: Coeficiente de Variación en porcentaje

En el caso de los modelos entrenados con datos entre los períodos 2018 y 2024, los resultados siguen la misma tendencia que lo estudiado en el periodo desde 2000. Si que llama la atención que el porcentaje de ceros de la variable porcentaje de triples anotados, en Random Forest es del 40%, esto quiere decir que en el 60% de los valores contrafácticos obtenidos ha habido un cambio en el porcentaje de triples. También se observa que en la red neuronal entrenada con estos datos, el cambio porcentual medio del número de triples es inferior al resto de modelos. Esto significaría que en los valores contrafactuales obtenidos, se ha visto un cambio porcentual medio del 9.20%, alrededor de un 5% menos que en los otros tres modelos.

Tabla 5.9: Descriptivos del cambio porcentual entre el valor contrafáctico y el valor real, generados a partir de las predicciones generadas por los modelos entrenados con datos desde el año 2018

2018-2024								
	Random Forest				Red Neuronal			
Variable	Media	DT	% de Ceros	CV	Media	DT	% de Ceros	CV
X3P..Eq	14.52	19.06	39.96	131.30	9.20	17.99	63.96	195.62
Asist.Eq	9.31	16.62	61.75	178.50	6.66	15.98	75.86	239.98
DRB_Porc.Eq	0.02	1.72	97.55	10,487.36	0.03	1.44	97.37	4,908.24
ORB_Porc.Eq	5.59	21.03	86.92	376.42	10.14	34.81	79.46	343.21
FGA.Eq	-0.15	1.48	95.49	973.66	-0.80	3.07	81.76	384.28
Players..10.PTS.....Eq	2.41	12.74	94.53	529.50	0.62	7.19	97.27	1,160.63
Posesion.Eq	1.24	2.58	71.31	207.75	1.21	2.48	73.81	205.62
Steal.Eq	0.04	2.71	99.39	7,358.90	0.83	20.84	98.38	2,508.84
Steal.Riv	-0.06	2.88	99.18	4,909.35	0.24	10.86	98.38	4,476.32
TO_Porc.Eq	-3.49	9.98	84.11	285.98	-3.85	10.67	80.72	277.31
TO_Porc.Riv	-0.08	2.20	98.93	2,745.07	0.10	3.46	97.48	3,456.29

DT: Desviación Típica CV: Coeficiente de Variación en porcentaje

5.4. SHAPLEY VALUES

En los siguientes cuatro gráficos se muestran los valores de Shapley correspondientes a cada predicción realizada con los datos de test. Estos valores de Shapley reflejan la contribución de cada parámetro en la desviación de la predicción respecto a la media, indicando cuánto influye cada característica en que una predicción sea mayor o menor que el valor promedio de todas

las predicciones. Un valor de Shapley alto indica que una característica tiene una gran influencia en la predicción, es decir, que contribuye significativamente a que el modelo prediga un valor más alto o más bajo en comparación con otras características. Si el valor de Shapley es positivo y alto, la característica está impulsando la predicción hacia un valor mayor. Si es negativo y alto (en magnitud), significa que la característica está empujando la predicción hacia un valor menor. En resumen, cuanto mayor es el valor de Shapley, más relevante es la influencia de esa característica en la predicción del modelo.

En el modelo Random Forest entrenado con datos desde el año 2000, la predicción media fue de 101.53 puntos, se puede observar mayor variabilidad en los shapley values en el porcentaje de tiros triples, en el número de posesiones y en el número de asistencias, lo cual es consistente con lo visto en el apartado anterior de los valores contrafactuales. El resultado más destacable es el del número de posesiones, se observa que partidos en los que el número de posesiones es alto (puntos en color amarillo), ha tenido un gran impacto a que el modelo prediga un valor más alto que la media en el número de puntos. En este modelo también tiene importancia el aumento del porcentaje de perdidas de manera, el cual tiene un impacto negativo en las predicciones de los puntos anotados.

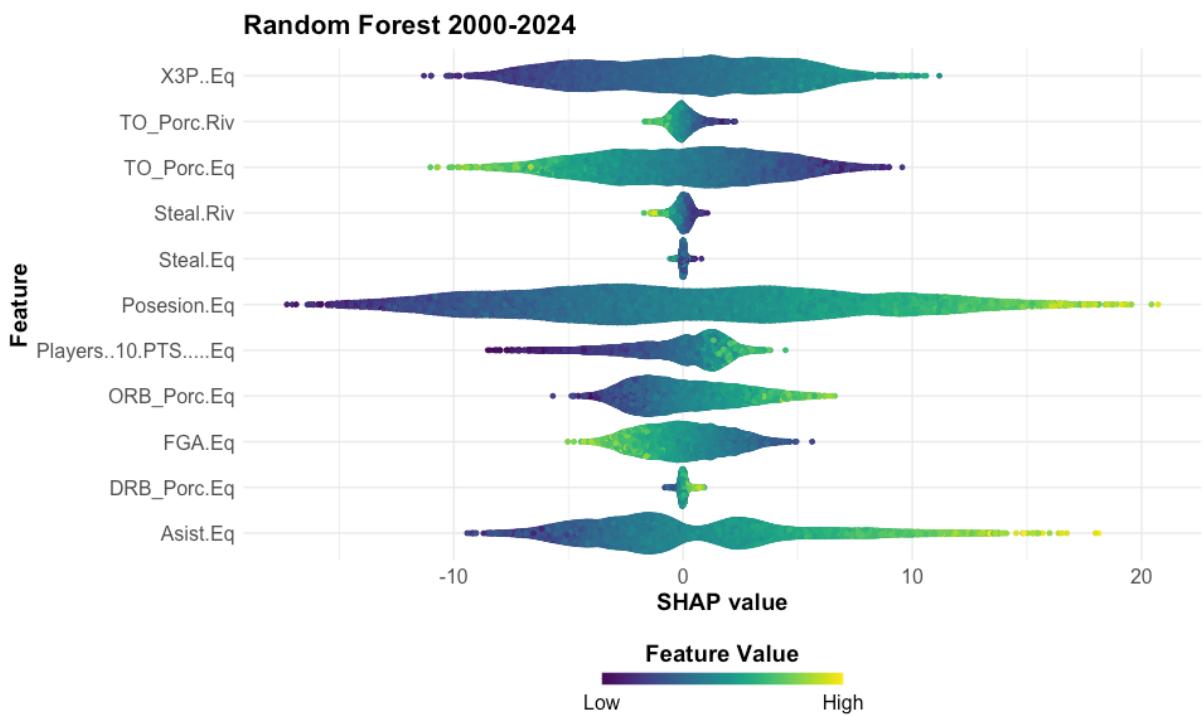
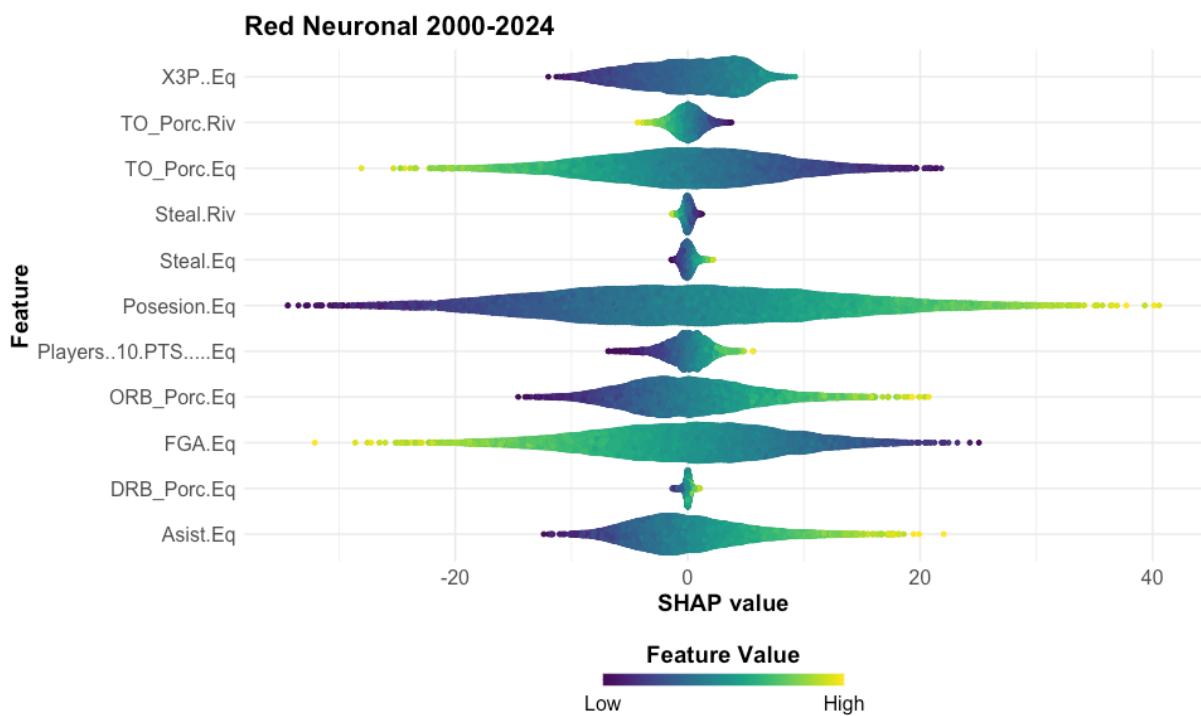


Figura 5.6: Gráfico de enjambre de los valores Shapley de Random Forest entrenado con datos desde 2000 hasta 2023

En este gráfico de valores SHAP para la Red Neuronal entrenada con datos desde el año 2000, se observa una mayor variabilidad en el impacto de ciertas características. Características como el porcentaje de tiros de tres puntos muestran un impacto muy significativo en las predicciones, con valores de SHAP que alcanzan hasta 40. Esto sugiere que estas variables pueden alterar las predicciones del modelo de manera considerable, dependiendo de sus valores originales. En el caso del número de tiros intentado, refleja un impacto negativo en el número de puntos predicho. La variabilidad en el impacto también es notable para el número de posesiones, donde sus valores de SHAP oscilan entre -30 hasta 40, reflejando una influencia tanto positiva como negativa, lo cual significa que es un factor clave a la hora de predecir el número de puntos anotado. En cambio, otras características como el número de robos o el porcentaje de rebotes defensivo tienen un impacto menos variable, con valores de SHAP que fluctúan cerca de cero, lo que sugiere que su influencia sobre las predicciones es más limitada en este modelo. El valor predicción media para este modelo fue de 101.50 puntos.



En el modelo Random Forest entrenado con datos desde 2018, la predicción media fue de 111.45 puntos, observamos cómo varias características influyen de manera significativa en las predicciones. El porcentaje de tiros de tres puntos tiene un impacto importante, con valores de SHAP positivos que alcanzan hasta 15, lo que indica que mayores valores en esta variable llevan a predicciones superiores.

De manera similar, el número de posesiones también tiene una fuerte influencia positiva en las predicciones, aunque con una variabilidad menor comparada con otras características. Sin embargo, características como el porcentaje de pérdidas del equipo muestran impactos variados, indicando que niveles bajos de estas variables pueden reducir las predicciones de puntos.

Un aspecto interesante es que la variable tiros de campo intentados tiene un comportamiento más neutro, con valores de SHAP que oscilan entre -5 y 5, lo que sugiere que su impacto es más balanceado y puede variar según el contexto del partido.

En resumen, el modelo Random Forest parece estar influenciado de manera más moderada en general, con algunas características clave como el porcentaje de tiros de tres y el número de posesiones que tienen efectos positivos claros, mientras que otras variables muestran comportamientos más distribuidos o de menor magnitud en su impacto sobre las predicciones.

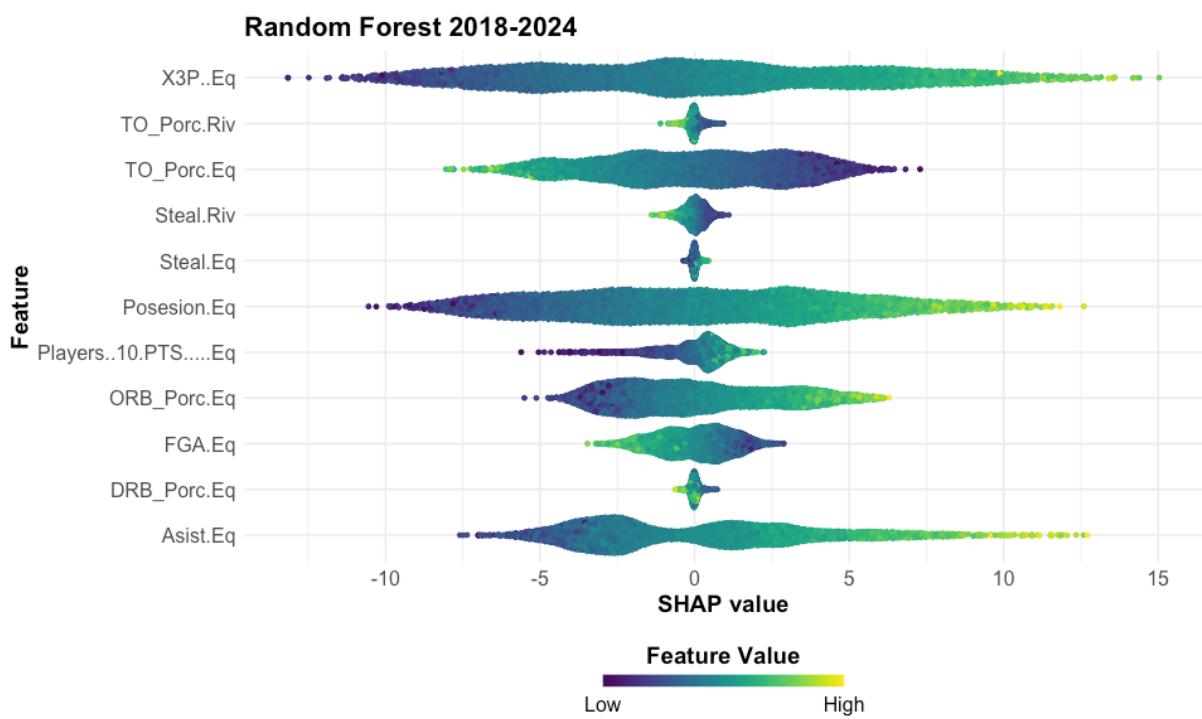


Figura 5.8: Gráfico de enjambre de los valores Shapley de Random Forest entrenado con datos desde 2018 hasta 2023

La red neuronal entrenada con los datos desde 2018 hasta 2024, tuvo una predicción media de 111.51 puntos. Se puede apreciar una gran variabilidad en los valores de SHAP asociados a distintas características. Algunas características como el porcentaje de tiros de tres puntos y

el número de posesiones tienen un impacto positivo notable, con valores de SHAP que superan los 20, lo que indica que estas características influyen en predicciones superiores a la media. Por otro lado, características como el porcentaje de pérdidas propias y el número de tiros de campo intentados presentan impactos negativos significativos en las predicciones, contribuyendo a valores inferiores en el número de puntos pronosticados. En particular, llama la atención la influencia negativa de un mayor número de tiros intentados, lo que sugiere que un mayor volumen de tiros podría estar asociado a menores puntuaciones, lo que sería relevante explorar en mayor detalle.

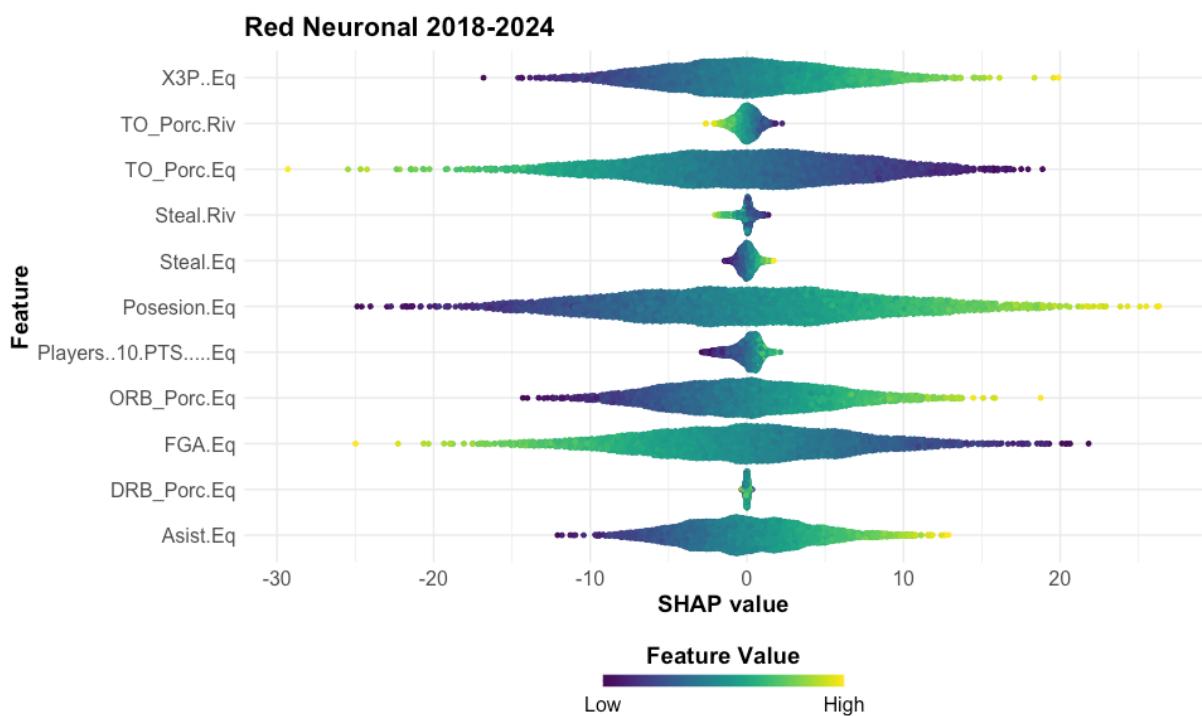


Figura 5.9: Gráfico de enjambre de los valores Shapley de Red Neuronal entrenada con datos desde 2018 hasta 2024

En la siguiente tabla se muestra la Media de los Valores Absolutos de los Shapley Values (MASV) y la desviación típica de los Shapley values en los cuatro modelos entrenados. MASV es una métrica que resume la magnitud promedio de la contribución de cada característica en las predicciones de un modelo. Se calcula tomando el valor absoluto de cada Shapley value, lo que permite medir la importancia de las características sin importar si influyen positivamente o negativamente en la predicción. Un MASV más alto para una característica indica que dicha variable tiene un mayor impacto, en promedio, sobre las predicciones realizadas en los datos

de testeo. Esta métrica es útil para identificar las variables más relevantes en el comportamiento global del modelo.

En concordancia con las observaciones anteriores, todos los modelos entrenados capturan un alto impacto en variables como el número de asistencias como el porcentaje de acierto desde la línea de tres puntos. Estas dos características parecen ser las más influyentes en las predicciones tanto para los modelos entrenados con datos desde el año 2000 como para aquellos entrenados con datos desde 2018, lo que subraya su importancia constante en el análisis.

Sin embargo, es destacable la diferencia en el comportamiento entre los modelos Random Forest y las redes neuronales en la forma en que capturan el impacto de otras variables. En particular, se observa que las redes neuronales son capaces de detectar un mayor impacto del número de posesiones, número de tiros de campo y el porcentaje de rebote ofensivo en las predicciones en comparación con los modelos de Random Forest. Por ejemplo, en el caso de los modelos entrenados con datos entre 2000 y 2024, los MASV de la red neuronal para la variable tiros de campo alcanzan un valor de 6.09, considerablemente superior al 1.44 observado en el Random Forest. De forma similar, el impacto de la variable porcentaje de rebotes ofensivos es más significativo en la red neuronal, con un de 3.92 frente al 1.77 en Random Forest. Esta tendencia persiste en los modelos entrenados con datos entre 2018 y 2024, donde se siguen observando diferencias importantes en estas variables entre ambos tipos de modelos.

Estas diferencias sugieren que las redes neuronales pueden ser más sensibles a los cambios en características relacionadas con la agresividad ofensiva del equipo, como el porcentaje de acierto del tiro de tres, los tiros de campo y los rebotes ofensivos, lo que puede hacerlas más adecuadas para capturar patrones más complejos en los datos que los modelos Random Forest no detectan con la misma precisión.

Año	Parámetro	Random Forest		Red Neuronal	
		MASV	DT	MASV	DT
2000-2024	Asist.Eq	3.43	4.22	3.55	4.51
	DRB_Porc.Eq	0.09	0.12	0.16	0.23
	FGA.Eq	1.44	1.74	6.09	7.57
	ORB_Porc.Eq	1.77	2.12	3.92	4.90
	Players..10.PTS.....Eq	1.45	1.84	1.21	1.53
	Posesion.Eq	5.93	7.12	9.28	11.47
	Steal.Eq	0.05	0.07	0.34	0.43
	Steal.Riv	0.22	0.29	0.24	0.31
	TO_Porc.Eq	2.98	3.57	5.67	7.05
	TO_Porc.Riv	0.35	0.45	0.85	1.07
2018-2024	X3P..Eq	3.48	4.10	3.41	4.04
	Asist.Eq	3.10	3.69	3.13	3.91
	DRB_Porc.Eq	0.09	0.13	0.06	0.08
	FGA.Eq	0.94	1.13	5.20	6.55
	ORB_Porc.Eq	2.10	2.50	4.02	4.98
	Players..10.PTS.....Eq	0.75	1.00	0.63	0.81
	Posesion.Eq	3.82	4.58	6.67	8.26
	Steal.Eq	0.06	0.08	0.33	0.42
	Steal.Riv	0.26	0.33	0.28	0.41
	TO_Porc.Eq	2.53	3.01	5.59	6.93
	TO_Porc.Riv	0.14	0.20	0.50	0.63
	X3P..Eq	4.56	5.55	4.21	5.22

6. CASO DE USO

El staff de data management del Equipo de los Miami Heats en la temporada 2023-2024 quiere hacer un plan de partido basándose en los modelos predictivos disponibles en este trabajo. Concretamente se usa la red neuronal obtenida con los datos desde el año 2018 al 2024, que es el modelo en el que se ha obtenido un RMSE más bajo (Ver sección Resultados). Suponiendo que se está trabajando la previa del partido Miami Heats - Chicago Bulls del día 19 de noviembre de 2023. El contexto de Miami Heat en este momento es de un balance de 8 victorias y 4 derrotas, y el de Chicago Bulls es de 4 victorias y 8 derrotas. Al disponer de los datos de boxscore de las jornadas previas a ese partido se decide obtener las medias de todos los parámetros de 5 en 5 jornadas, lo que se denomina media móvil. De la jornada 1 a la 5, de la jornada 2 a la 6, de la jornada 3 a la 7...y así sucesivamente hasta el último partido previo al encuentro de interés.

Se aplicarán los tres métodos explicabilidad a este nuevo conjunto de datos de test, tanto para los partidos de Miami Heats, como para los partidos de Chicago Bulls con el objetivo de poder estudiar que parámetros se recomienda trabajar mas para fortalecer los puntos fuertes del equipo así como evaluar que parámetros influyen en nuestro rival a la hora de anotar.

Tabla 6.1: Ejemplo del dataset que se va a validar con el modelo. Cada linea recoge las medias de los parámetros en 5 jornadas. La última columna indica que jornadas se han usado para calcular las medias

Equipo	X3P Eq	Asist Eq	DRB_Porc Eq	ORB_Porc Eq	FGA Eq	Players PTS	10 Eq	Posesion Eq	Steal Eq	Steal Riv	TO_Po rc Eq	TO_Po rc Riv	Puntos Eq	jornada usada
CHI	0.35	20.20	73.79	24.50	90.40		0.42	98.12	8.40	5.60	10.21	11.98	106.00	Jornada 1 a 5
CHI	0.35	20.60	71.46	23.15	88.80		0.41	97.56	7.80	6.40	10.16	11.41	105.40	Jornada 2 a 6
CHI	0.38	22.60	72.13	22.50	88.80		0.49	98.42	8.20	5.80	10.14	12.43	111.00	Jornada 3 a 7
CHI	0.38	23.20	69.91	23.15	89.60		0.54	97.10	9.00	5.20	8.96	12.18	112.40	Jornada 4 a 8
CHI	0.38	24.20	69.63	23.90	91.80		0.56	97.58	8.80	4.40	7.57	12.27	113.20	Jornada 5 a 9
MIA	0.38	23.80	75.54	21.62	89.00		0.38	99.69	8.40	7.60	11.40	13.60	104.60	Jornada 1 a 5
MIA	0.40	26.40	80.13	20.04	86.60		0.41	100.43	8.00	8.40	13.74	14.53	108.20	Jornada 2 a 6
MIA	0.37	27.60	83.90	20.64	86.40		0.42	100.62	8.60	8.20	14.62	14.92	107.60	Jornada 3 a 7
MIA	0.38	27.20	81.98	22.26	84.40		0.49	101.20	9.20	8.40	15.46	14.78	111.20	Jornada 4 a 8
MIA	0.37	26.40	79.88	22.35	84.00		0.52	101.44	9.80	8.00	15.61	16.03	111.80	Jornada 5 a 9

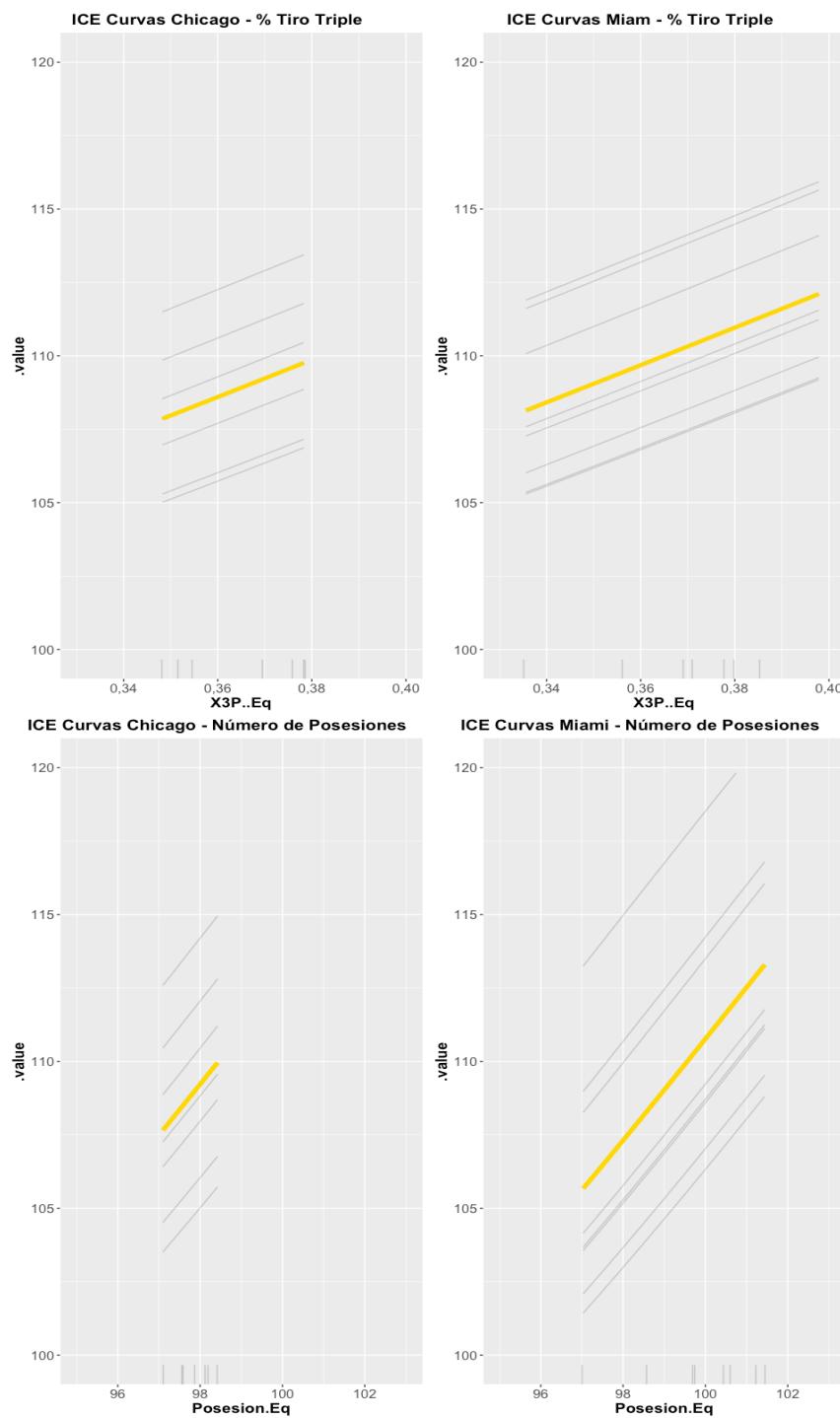


Figura 6.1: Curvas ICE de las variables Porcentaje de triples acertado y número de posesiones.

Cada línea se corresponde a una predicción de los datos que se están testeando. El eje horizontal refleja la variación del parámetro en cuestión y el eje vertical la variación del número de puntos predicho por el modelo. La línea amarilla es la media de los valores predichos

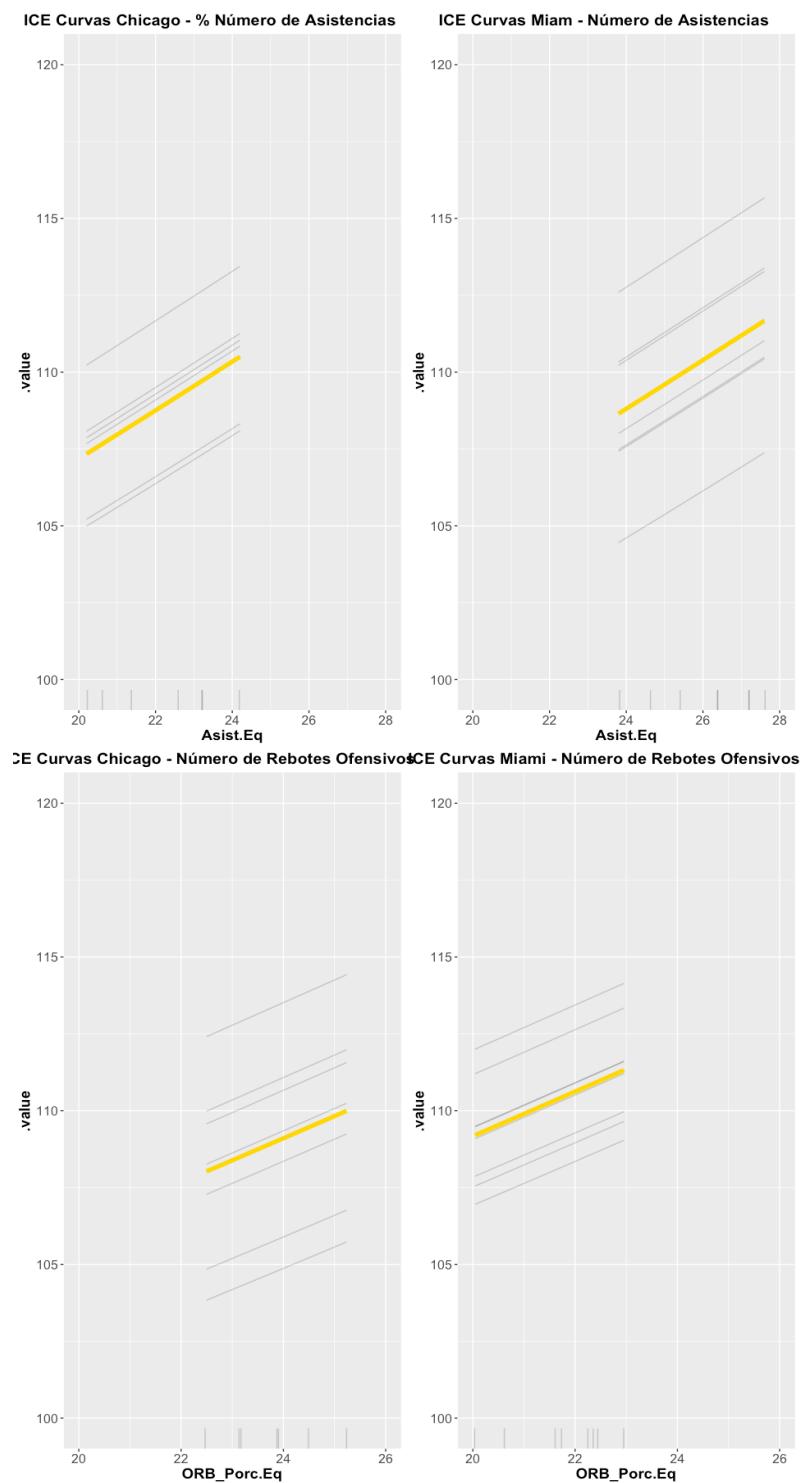


Figura 6.2: Curvas ICE de las variables Número de asistencias y número de rebotes ofensivos

Cada línea se corresponde a una predicción de los datos que se están testeando. El eje horizontal refleja la variación del parámetro en cuestión y el eje vertical la variación del número de puntos predicho por el modelo. La línea amarilla es la media de los valores predichos

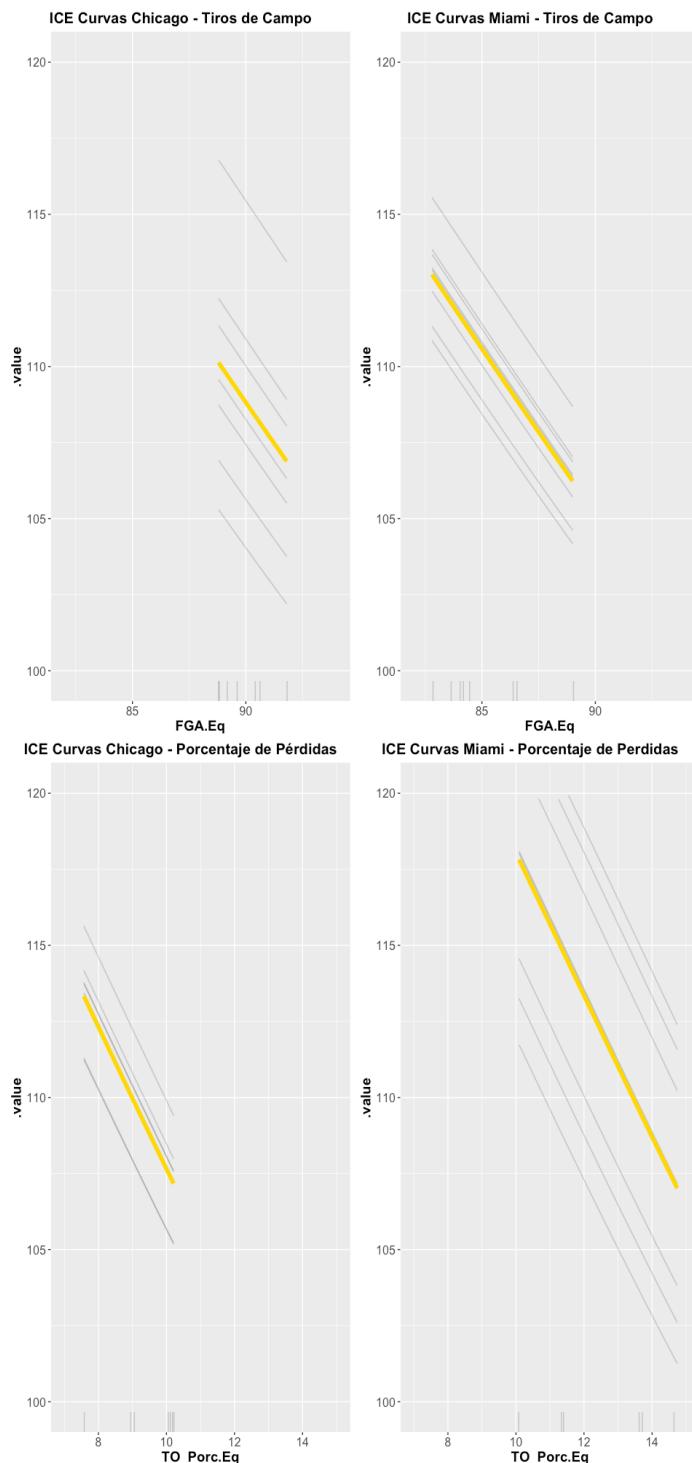


Figura 6.3: Curvas ICE de las variables Número de asistencias y número de rebotes ofensivos

Cada línea se corresponde a una predicción de los datos que se están testeando. El eje horizontal refleja la variación del parámetro en cuestión y el eje vertical la variación del número de puntos predicho por el modelo. La línea amarilla es la media de los valores predichos

Tabla 6.2: Datos de las características predictoras correspondientes a los partidos de Miami que se han usado para testear el modelo

X3P.. Eq	Asist. Eq	DRB_Porc .Eq	ORB_Porc .Eq	FGA. Eq	Players..10.PTS.. ...Eq	Posesion. Eq	Steal. Eq	Steal. Riv	TO_Porc. Eq	TO_Porc. Riv
0.38	23.80	75.54	21.62	89.00	0.38	99.69	8.40	7.60	11.40	13.60
0.40	26.40	80.13	20.04	86.60	0.41	100.43	8.00	8.40	13.74	14.53
0.37	27.60	83.90	20.64	86.40	0.42	100.62	8.60	8.20	14.62	14.92
0.38	27.20	81.98	22.26	84.40	0.49	101.20	9.20	8.40	15.46	14.78
0.37	26.40	79.88	22.35	84.00	0.52	101.44	9.80	8.00	15.61	16.03
0.38	27.20	77.85	22.44	83.60	0.61	99.70	9.20	6.40	13.58	16.10
0.34	25.40	74.43	22.96	84.20	0.58	98.53	9.80	6.20	11.34	15.63
0.36	24.60	71.34	21.74	82.80	0.58	97.03	9.20	6.00	10.09	15.19

Tabla 6.3: Valores contrafácticos que se han obtenido en el modelo para los datos de testeо correspondientes a Miami

X3P.. Eq	Asist. Eq	DRB_Porc .Eq	ORB_Porc .Eq	FGA. Eq	Players..10.PTS.. ...Eq	Posesion. Eq	Steal. Eq	Steal. Riv	TO_Porc. Eq	TO_Porc. Riv
0.42	23.80	75.54	21.62	89.00	0.38	102.56	8.40	7.60	11.40	13.60
0.40	26.40	80.13	20.04	79.00	0.41	100.43	8.00	8.40	13.74	14.53
0.37	27.60	83.90	20.64	81.00	0.42	100.62	8.60	8.20	14.62	14.92
0.38	27.20	81.98	22.26	79.00	0.49	101.20	9.20	8.40	15.46	14.78
0.37	26.40	79.88	22.35	79.00	0.52	101.44	9.80	8.00	15.61	16.03
0.38	27.20	77.85	22.44	83.60	0.61	104.88	9.20	6.40	13.58	16.10
0.34	25.40	74.43	20.00	84.20	0.58	104.88	9.80	6.20	11.34	15.63
0.36	24.60	71.34	18.92	82.80	0.58	103.80	9.20	6.00	10.09	15.19

Tabla 6.4: Datos de las características predictoras correspondientes a los partidos de Chicago que se han usado para testear el modelo

X3P.. Eq	Asist.Eq	DRB_Porc .Eq	ORB_Porc .Eq	FGA.Eq	Players..10.PTS.. ...Eq	Posesion. Eq	Steal.E q	Steal. Riv	TO_Porc. Eq	TO_Porc. Riv
0.352	20.200	73.787	24.498	90.400	0.421	98.120	8.400	5.600	10.211	11.978
0.348	20.600	71.465	23.150	88.800	0.414	97.560	7.800	6.400	10.158	11.409
0.376	22.600	72.129	22.497	88.800	0.491	98.424	8.200	5.800	10.140	12.433
0.378	23.200	69.907	23.149	89.600	0.544	97.096	9.000	5.200	8.963	12.178
0.378	24.200	69.626	23.901	91.800	0.556	97.576	8.800	4.400	7.571	12.268
0.354	23.200	68.822	25.235	90.600	0.524	97.880	9.400	6.000	9.072	13.036
0.369	21.400	68.957	23.924	89.200	0.528	98.192	10.000	5.400	10.075	14.244

Tabla 6.5: Valores contrafácticos que se han obtenido en el modelo para los datos de testeо correspondientes a Chicago

X3P.. Eq	Asist. Eq	DRB_Porc .Eq	ORB_Porc .Eq	FGA. Eq	Players..10.PTS.. ...Eq	Posesion. Eq	Steal. Eq	Steal. Riv	TO_Porc. Eq	TO_Porc. Riv
0.35	20.20	73.79	24.50	88.00	0.42	100.32	8.40	5.60	10.21	11.98
0.35	20.60	71.46	25.53	88.80	0.41	100.32	7.80	6.40	10.16	11.41
0.38	22.60	72.13	28.26	88.80	0.49	98.42	8.20	5.80	8.97	12.43
0.38	23.20	69.91	23.15	89.60	0.54	103.00	9.00	5.20	8.96	12.18
0.30	24.20	69.63	23.90	81.00	0.56	97.58	8.80	4.40	7.57	12.27
0.35	23.20	68.82	25.23	90.60	0.52	103.44	9.40	6.00	9.07	13.04
0.37	21.40	68.96	28.26	89.20	0.53	98.19	10.00	5.40	8.97	14.24

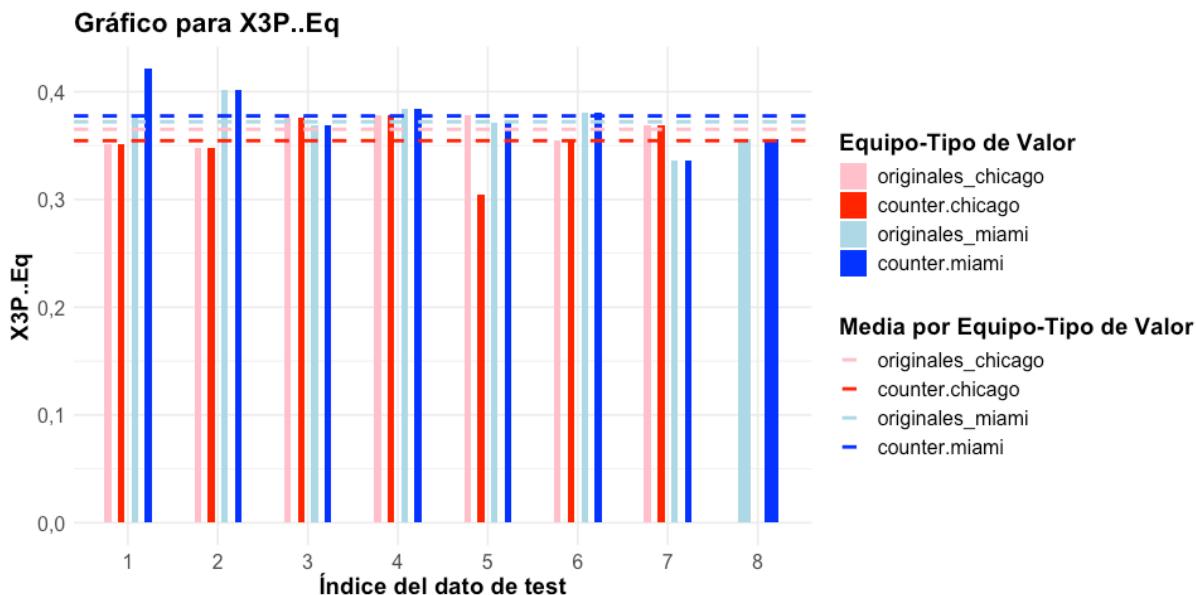


Figura 6.4: Evaluación y comparación de los valores contrafácticos con los valores originales para el porcentaje de acierto en la línea de tres. El eje horizontal representa cada uno de los datos de test. Las barras azules representan los datos analizados de Miami y las barras en rojo las de Chicago En graduación suave son los datos originales de test, y en graduación más intensa los valores contrafácticos. El eje vertical mide el valor del porcentaje de acierto en la linea de tres. La linea horizontal punteada es el valor medio

El gráfico 6.4 muestra tanto los datos de test que se han utilizado (recordar que eran las medias móviles de 5 partidos), como los valores contrafácticos que se han calculado, con el objetivo de aumentar la anotación entre un 5 y un 10%. Se muestra también la media tanto de los valores originales como la de los contrafácticos. Que las barras sean iguales entre el valor real y contrafáctico, significa que el cambio no afectaría al total de puntos predicho.

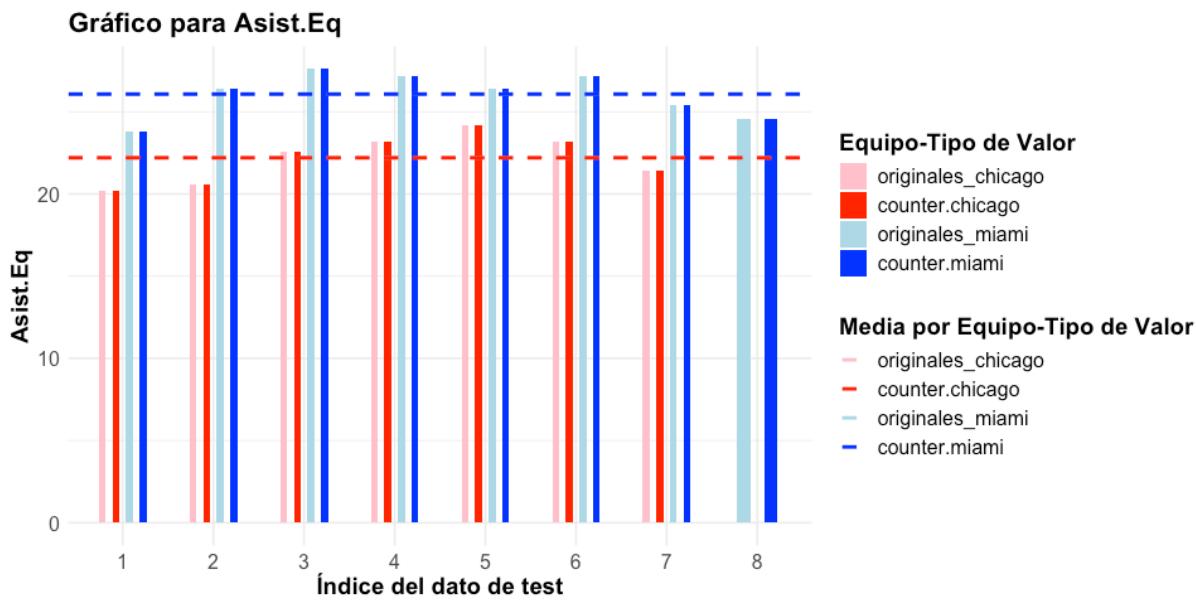


Figura 6.5: Evaluación y comparación de los valores contrafácticos con los valores originales para el número de asistencias. El eje horizontal representa cada uno de los datos de test. Las barras azules representan los datos analizados de Miami y las barras en rojo las de Chicago. En graduación suave son los datos originales de test, y en graduación más intensa los valores contrafácticos. El eje vertical representa el valor número de asistencias. La linea horizontal punteada es el valor medio

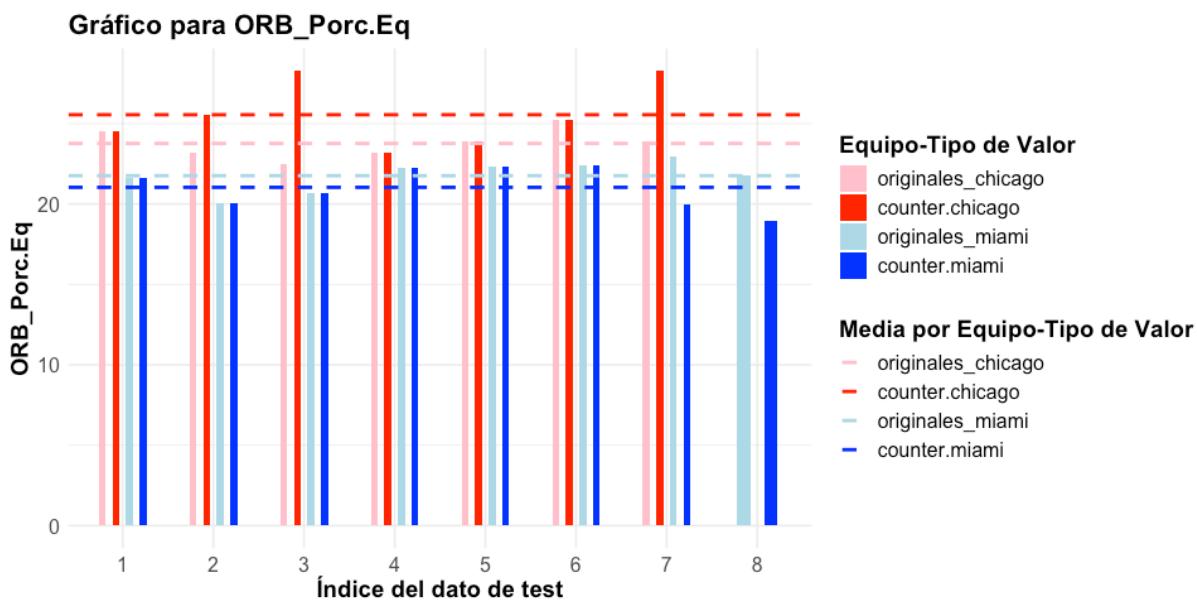


Figura 6.6: Evaluación y comparación de los valores contrafácticos con los valores originales para el porcentaje de rebotes ofensivos. El eje horizontal representa cada uno de los datos de test. Las barras azules representan los datos analizados de Miami y las barras en rojo las de Chicago. En graduación suave son los datos originales de test, y en graduación más intensa los valores contrafácticos. El eje vertical representa el valor del porcentaje de rebotes ofensivos. La linea horizontal punteada es el valor medio

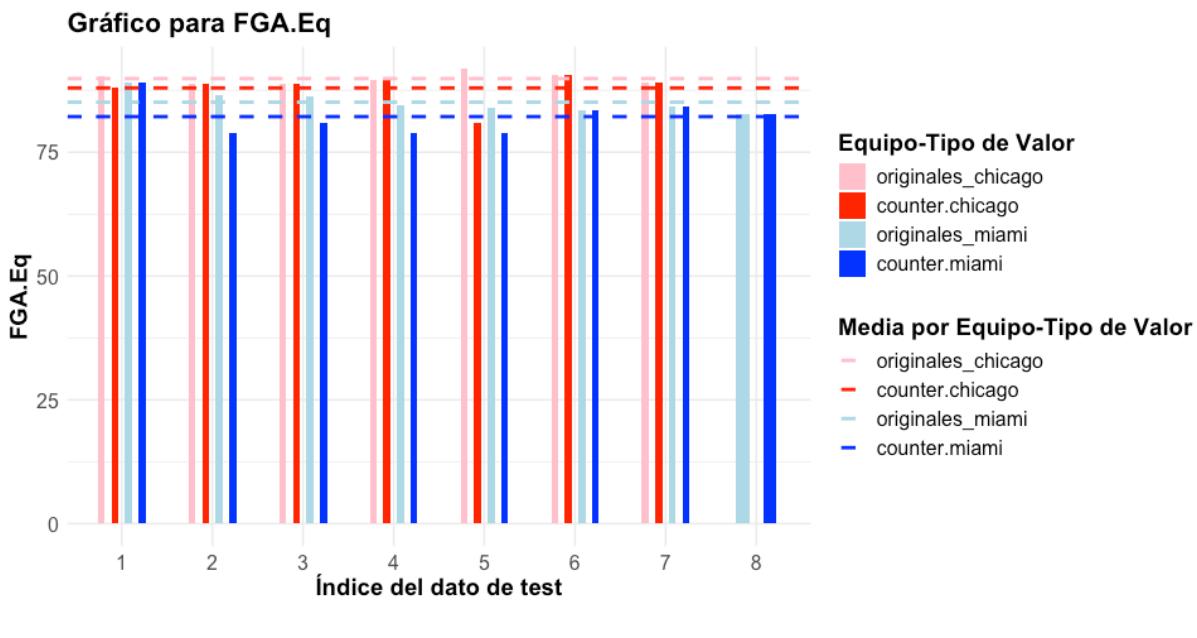


Figura 6.7: Evaluación y comparación de los valores contrafácticos con los valores originales para el número de tiros de campo. El eje horizontal representa cada uno de los datos de test.

Las barras azules representan los datos analizados de Miami y las barras en rojo las de Chicago. En graduación suave son los datos originales de test, y en graduación más intensa los valores contrafácticos. El eje vertical representa el valor del número de tiros de campo. La linea horizontal punteada es el valor medio

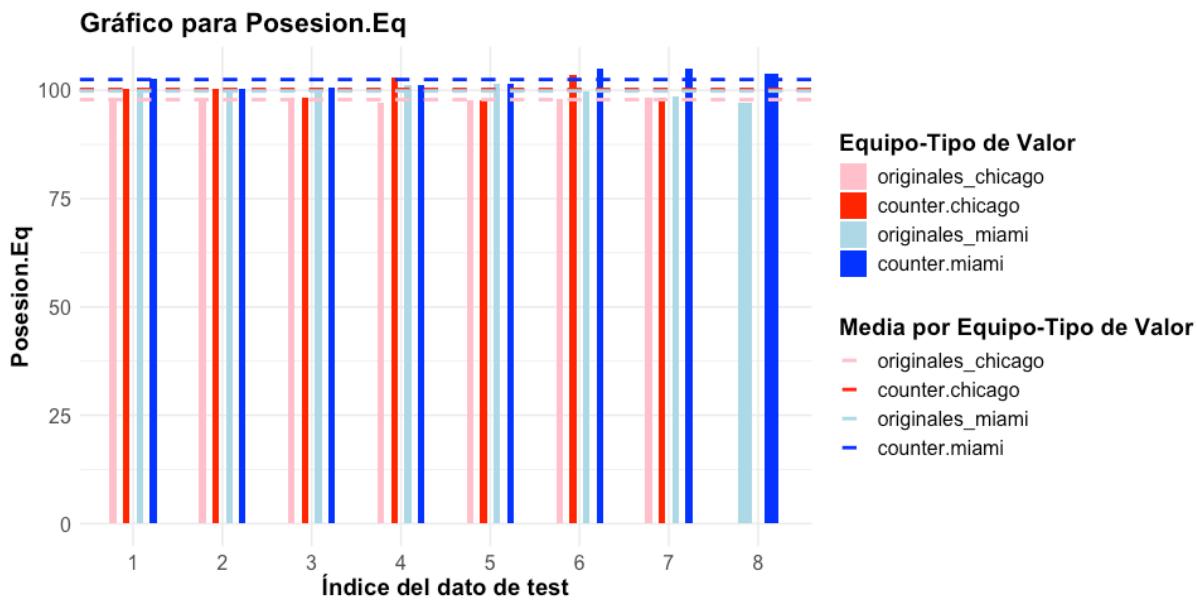


Figura 6.8: Evaluación y comparación de los valores contrafácticos con los valores originales para número de posesiones. El eje horizontal representa cada uno de los datos de test. Las barras azules representan los datos analizados de Miami y las barras en rojo las de Chicago. En graduación suave son los datos originales de test, y en graduación más intensa los valores contrafácticos. El eje vertical representa el valor del número de posesiones. La linea horizontal punteada es el valor medio

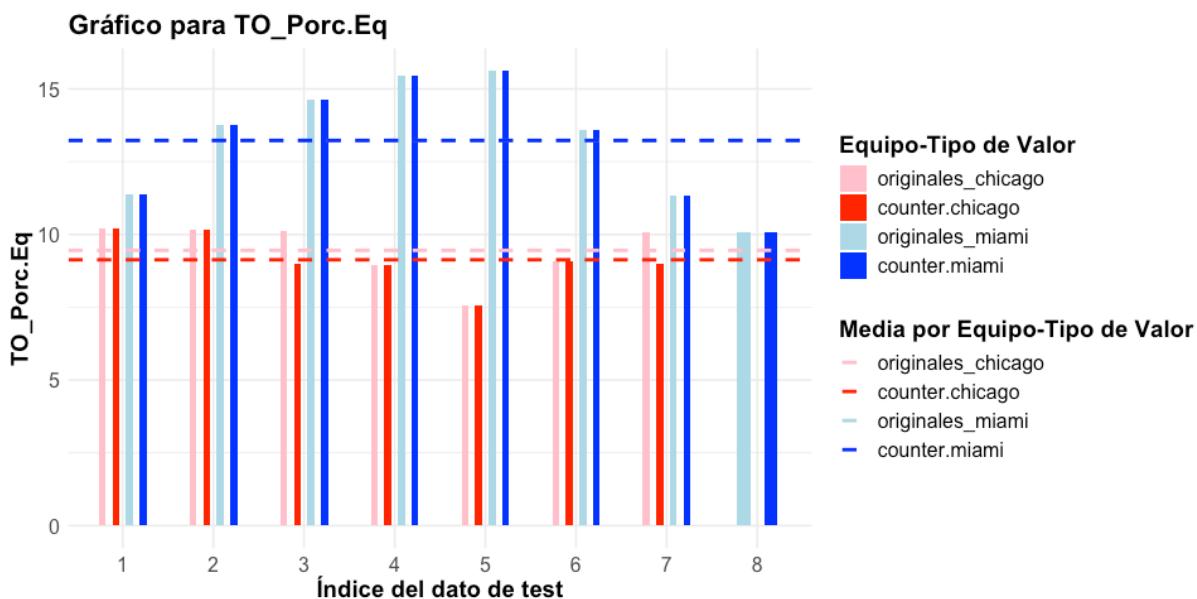


Figura 6.9: Evaluación y comparación de los valores contrafácticos con los valores originales para el porcentaje de perdidas. El eje horizontal representa cada uno de los datos de test. Las barras azules representan los datos analizados de Miami y las barras en rojo las de Chicago. En graduación suave son los datos originales de test, y en graduación más intensa los valores contrafácticos. El eje vertical representa el valor del porcentaje de perdidas. La linea horizontal punteada es el valor medio

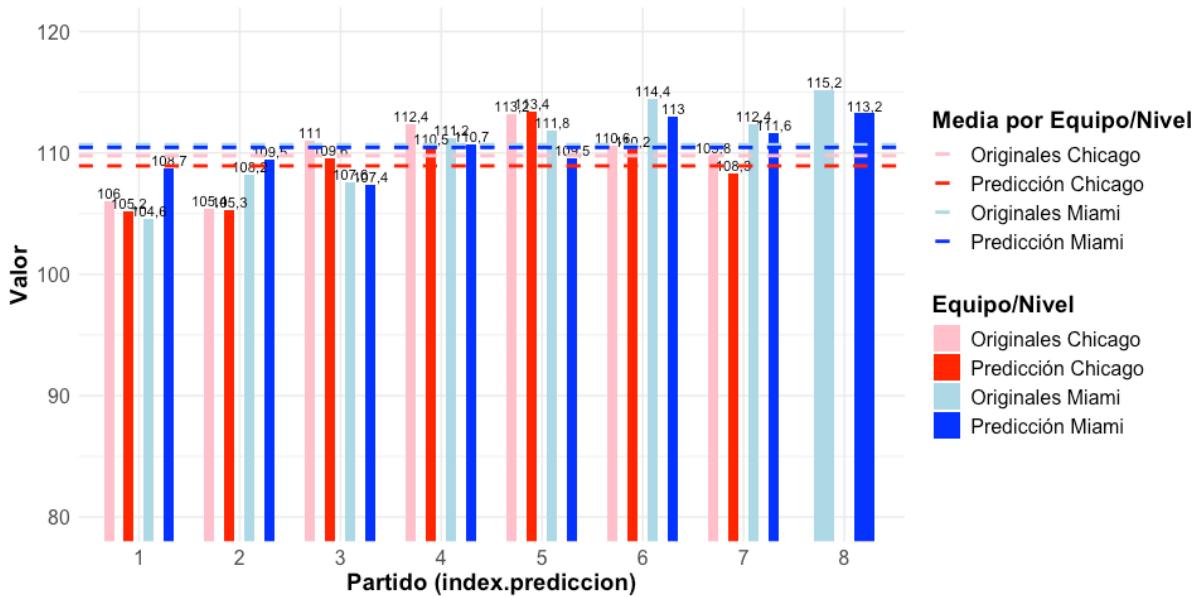


Figura 6.10: Evaluación y comparación de la predicción del número de puntos con los valores originales. El eje horizontal representa cada uno de los datos de test. Las barras azules representan los datos analizados de Miami y las barras en rojo las de Chicago. En graduación suave son los datos originales de test, y en graduación más intensa los valores predichos. El eje vertical representa el número de puntos. La linea horizontal punteada son las medias

Se calculan los Shapley Values para los datos de test que disponemos de Miami hasta la jornada previa al partido de interés. El heatmap muestra los valores de Shapley para diferentes instancias (columnas) y características del modelo (filas), donde la leyenda indica el impacto de las características en las predicciones, desde -8 (morado) hasta +4 (amarillo). Los valores positivos sugieren que la característica aumenta la predicción, mientras que los negativos indican una reducción. Por ejemplo, el número de tiros intentados, tiene un fuerte impacto positivo en las instancias 5 y 6, mientras que los porcentajes de perdidas reducen las predicciones en varias instancias. El dendrograma agrupa las instancias según similitudes en sus valores de Shapley, esto nos sugiere partidos en el que los shapley values han sido similares, lo que ayuda a identificar patrones de comportamiento compartidos entre las predicciones.

En las últimas jornadas está teniendo alto impacto en las predicciones, el número de tiros de campo intentando, aunque este resultado no se ve complementado, con el porcentaje de acierto del tiro de tres ya que no tiene demasiada importancia en los partidos analizados.

Destaca que el número de posesiones que se ha visto que en el modelo es un factor importante, en las ultimas jornadas tiene un impacto negativo sobre las predicciones y también el porcentaje de perdidas en las instancias 3, 4 y 5 ha tenido impacto fuerte en negativo.



Figura 6.11: Heatmap de los Shapley Values de Miami

Se evaluan los shapley values partido a partido para las instancias de Miami. Los valores que se obtienen a partir de los datos de las cinco primeras jornadas no son concluyentes, se obtienen valores entre -1 y 1. En la segunda instancia (datos a partir de la jornada 2 a 6), si que se observa que el porcentaje de perdidas tuvo un alto impacto en una predicción de 109.47 puntos (cuando el valor medio de las predicciones del modelo era de 111), y también un shapley value negativo del número de posesiones. En la tercera instancias (datos de la jornada 3 a 7), se sigue manteniendo ese impacto negativo del porcentaje de perdidas, aunque el número de asistencias y el número de posesiones toman valores positivos, eso si alrededor de un valor de 2. El valor predicho sigue siendo inferior a la media de las predicciones del modelo.

En las instancia 5 se observa un shapley value inferior a -5 correspondiente al porcentaje de perdidas, con gran peso en la predicción del modelo. En el caso de la instancia 8 que recoge la información de los últimos cinco partidos anteriores al que se quiere analizar, el porcentaje perdidas ha sido un factor con impacto positivo en la predicción, pero el número de posesiones, tuvo un peso negativo.

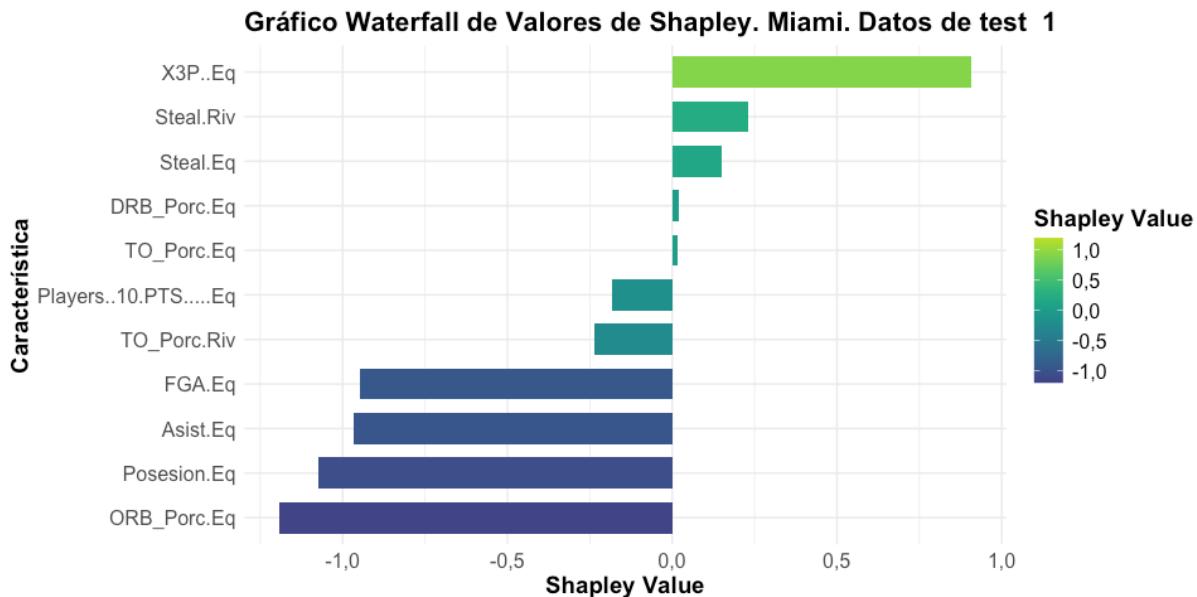


Figura 6.12: Shapley values. Datos de test de Miami. Instancia1

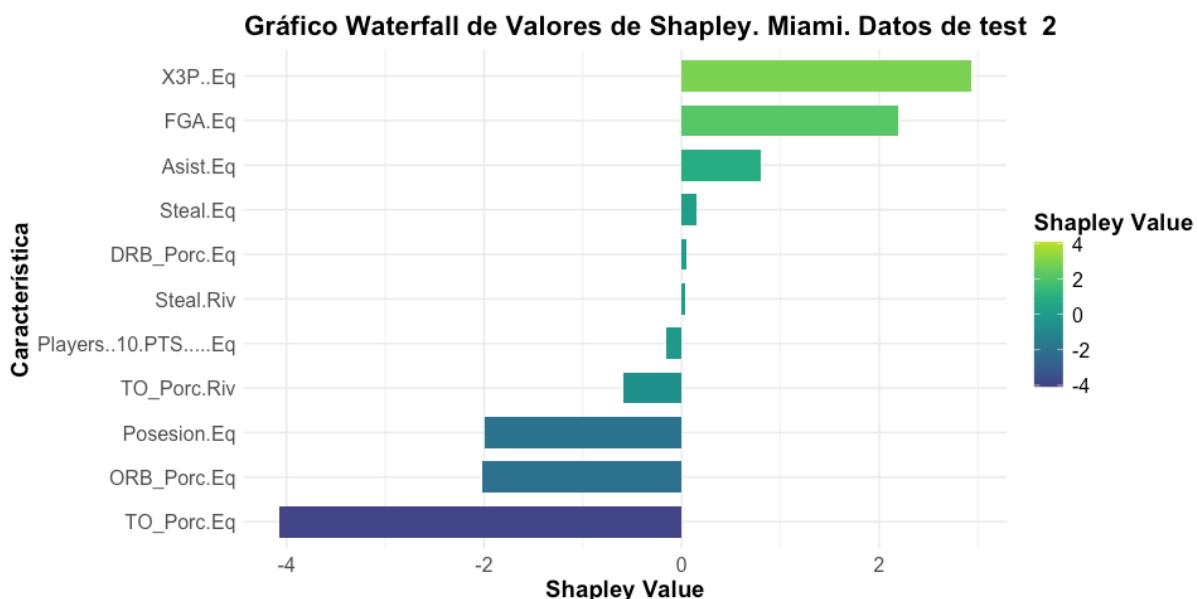


Figura 6.13: Shapley values. Datos de test de Miami. Instancia2

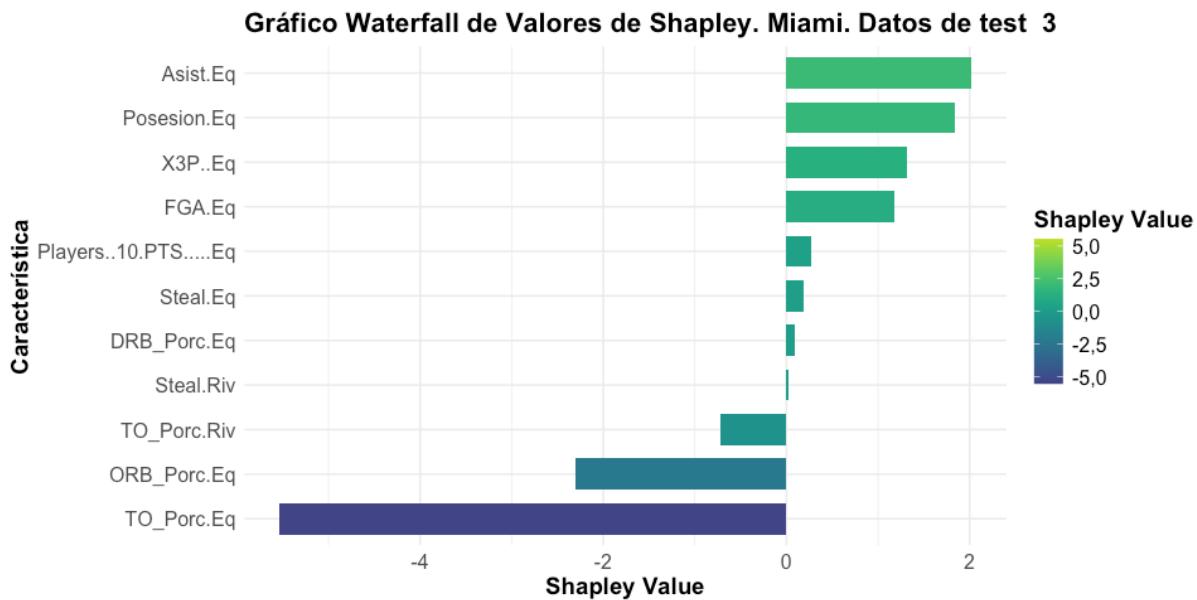


Figura 6.14: Shapley values. Datos de test de Miami. Instancia3

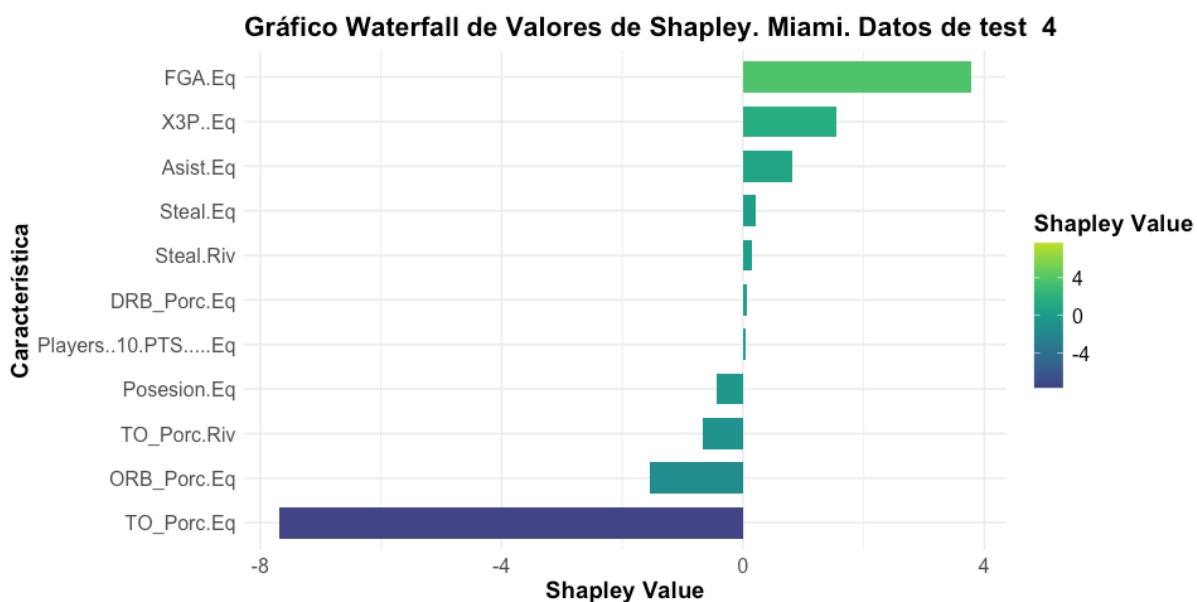


Figura 6.15: Shapley values. Datos de test de Miami. Instancia4

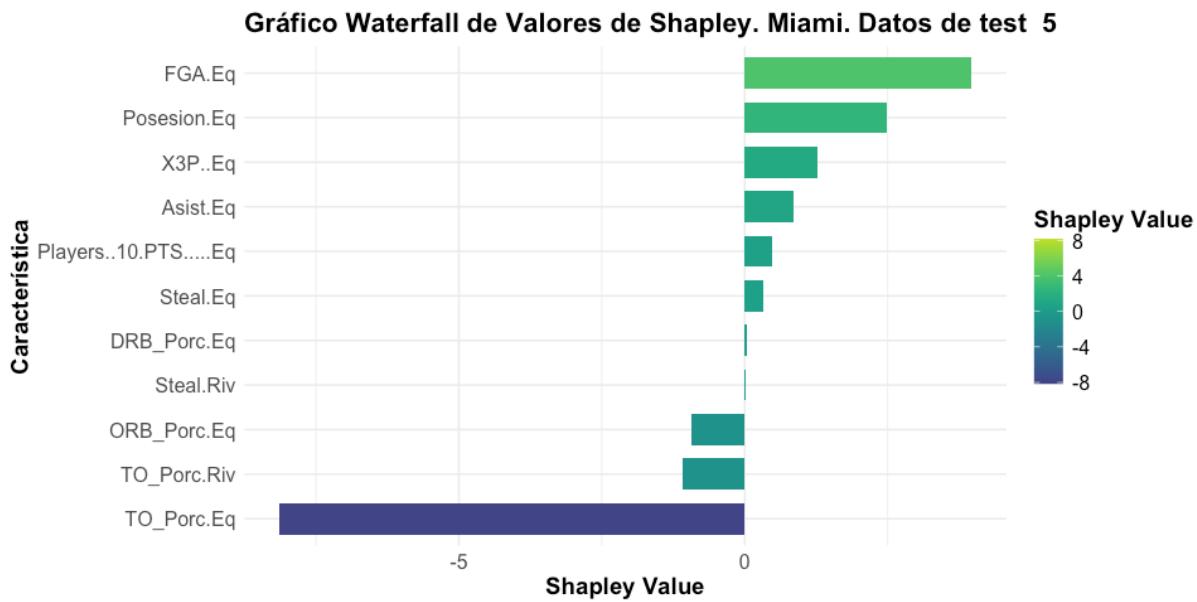


Figura 6.16: Shapley values. Datos de test de Miami. Instancia5

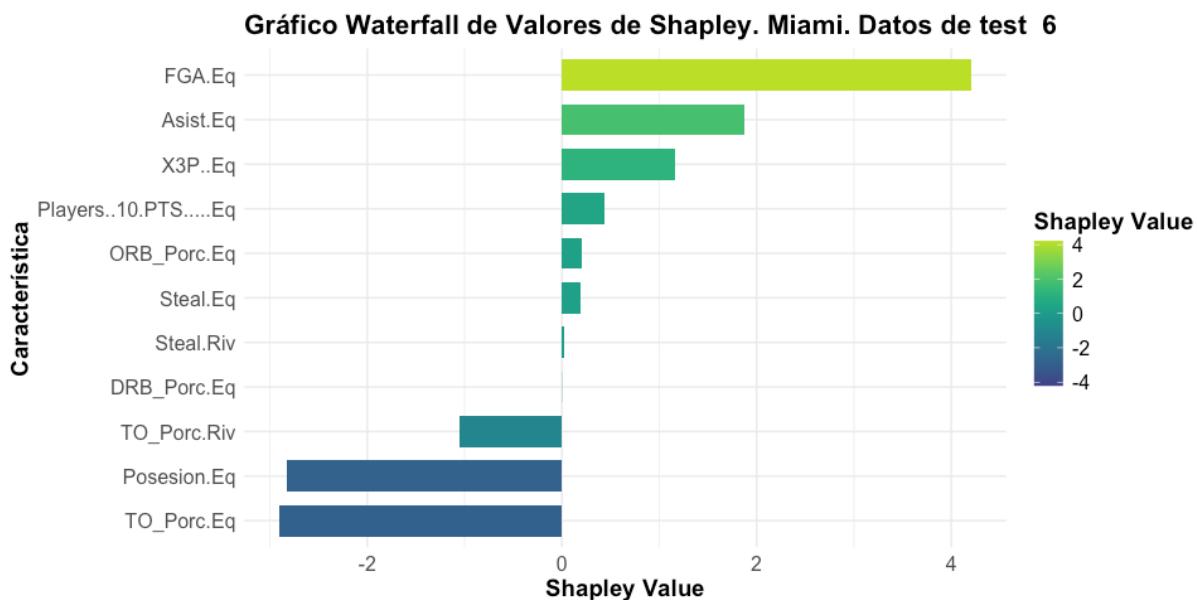


Figura 6.17: Shapley values. Datos de test de Miami. Instancia6

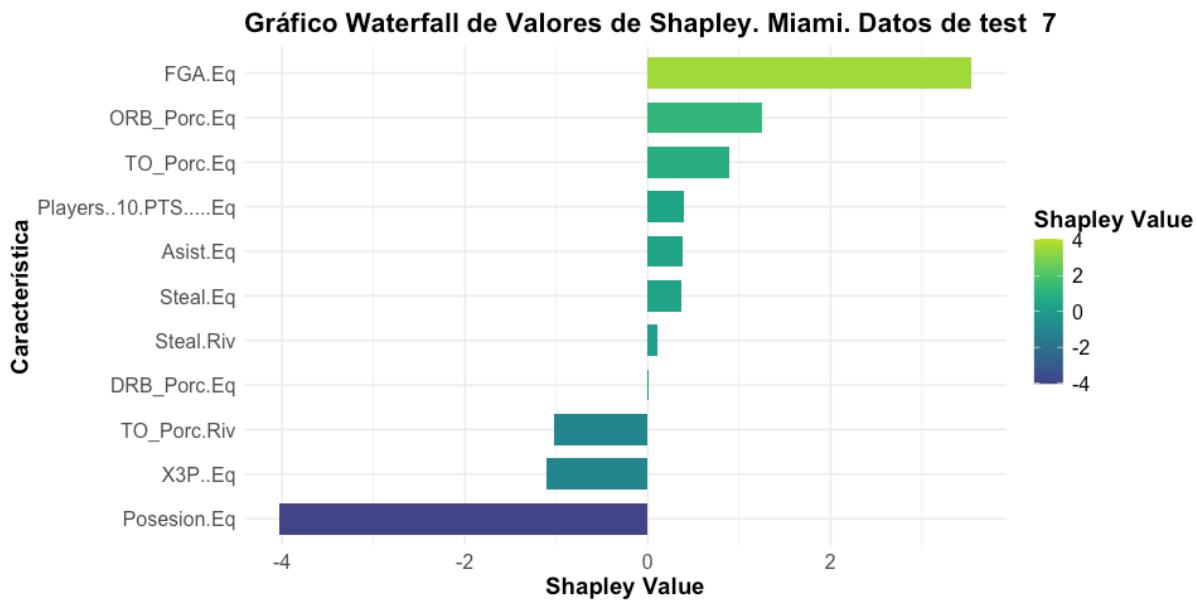


Figura 6.18: Shapley values. Datos de test de Miami. Instancia7

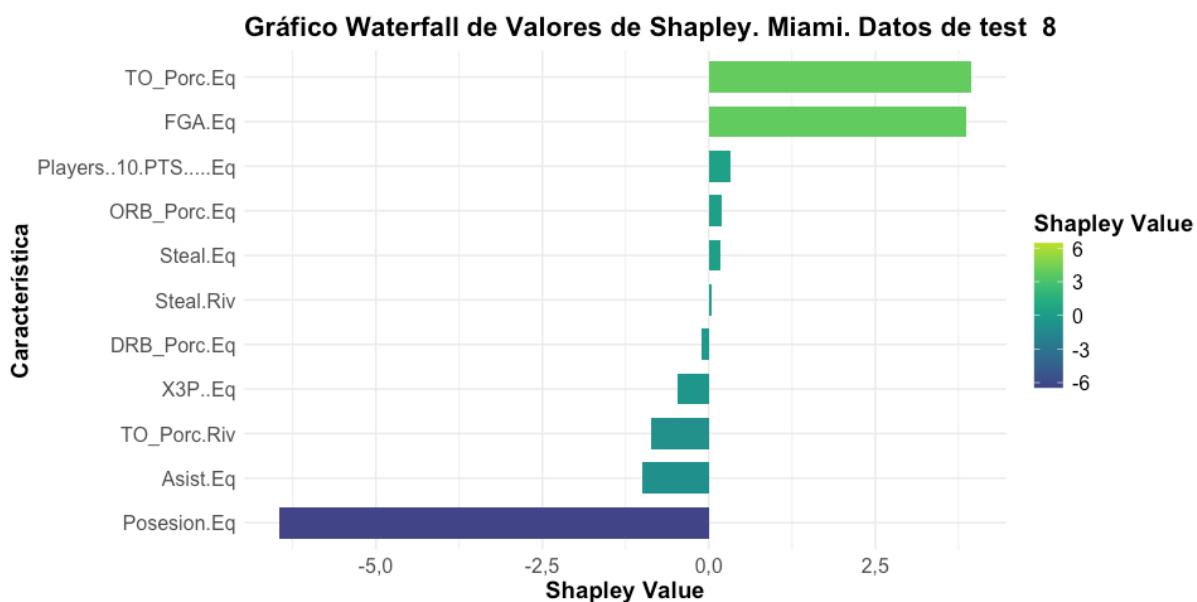


Figura 6.19: Shapley values. Datos de test de Miami. Instancia8

En el análisis de Chicago, el heatmap no proporciona información muy clara acerca del impacto que la mayoría de componentes del juego analizadas tienen en la predicción del número de puntos, si que se observa impacto negativo en el número de posesiones en varias instancias predichas, y es cierto que el porcentaje de perdidas, tiene un impacto positivo en todas las instancias predichas.

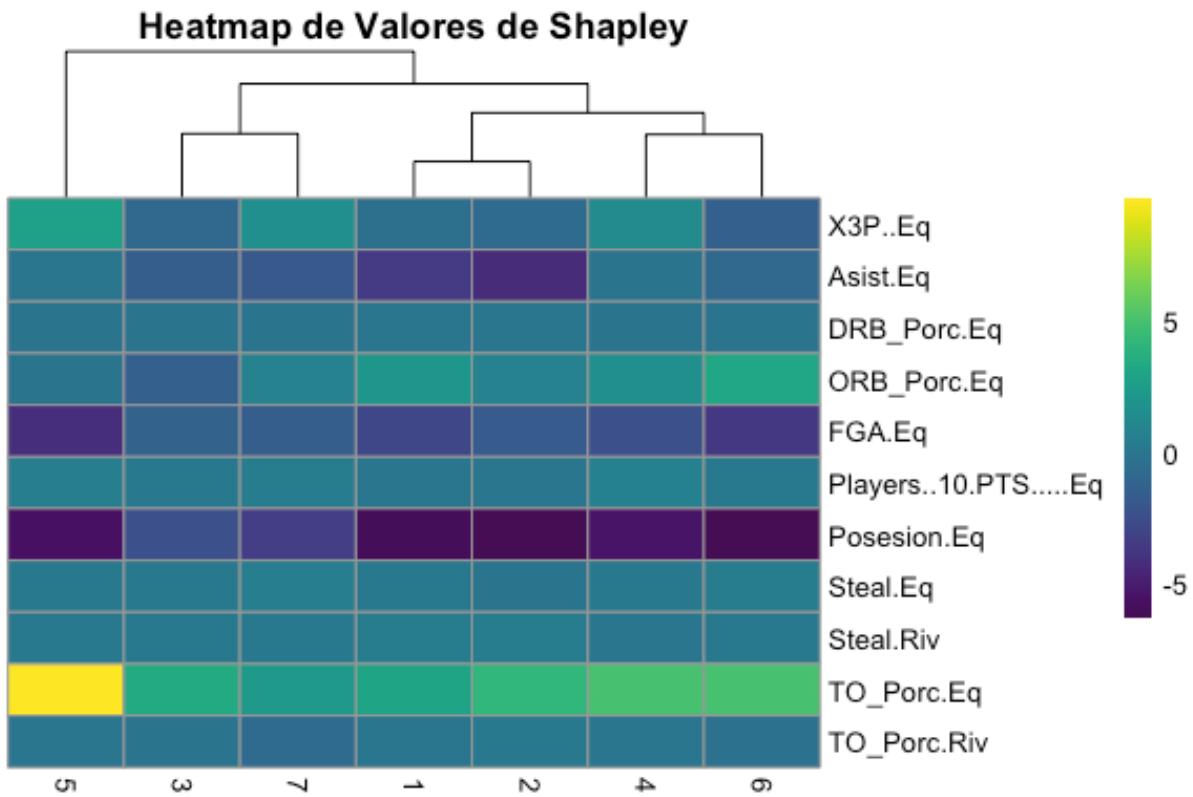


Figura 6.20: Heatmap de los Shapley values de Chicago

En los siguientes gráficos se desgranan los shapley values partido a partido, de forma visual se observa un patrón claro. Que confirma lo visto en el heatmap, un impacto muy negativo del número de posesiones y un impacto positivo del porcentaje de perdidas.

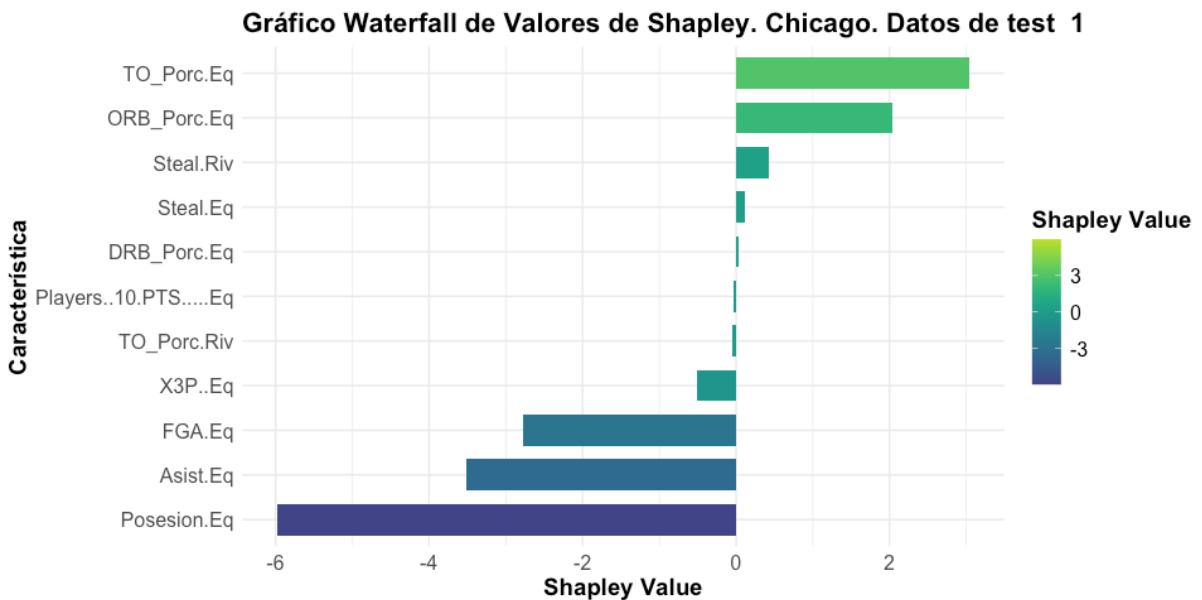


Figura 6.21: Shapley values. Datos de test de Chicago Instancia1

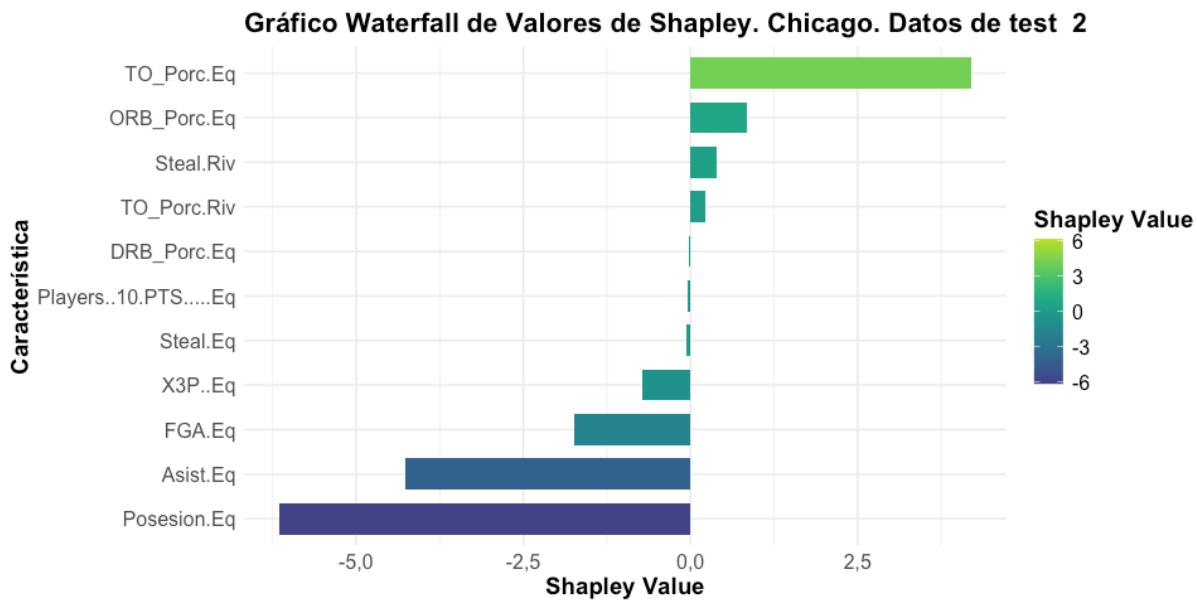


Figura 6.22: Shapley values. Datos de test de Chicago Instancia2

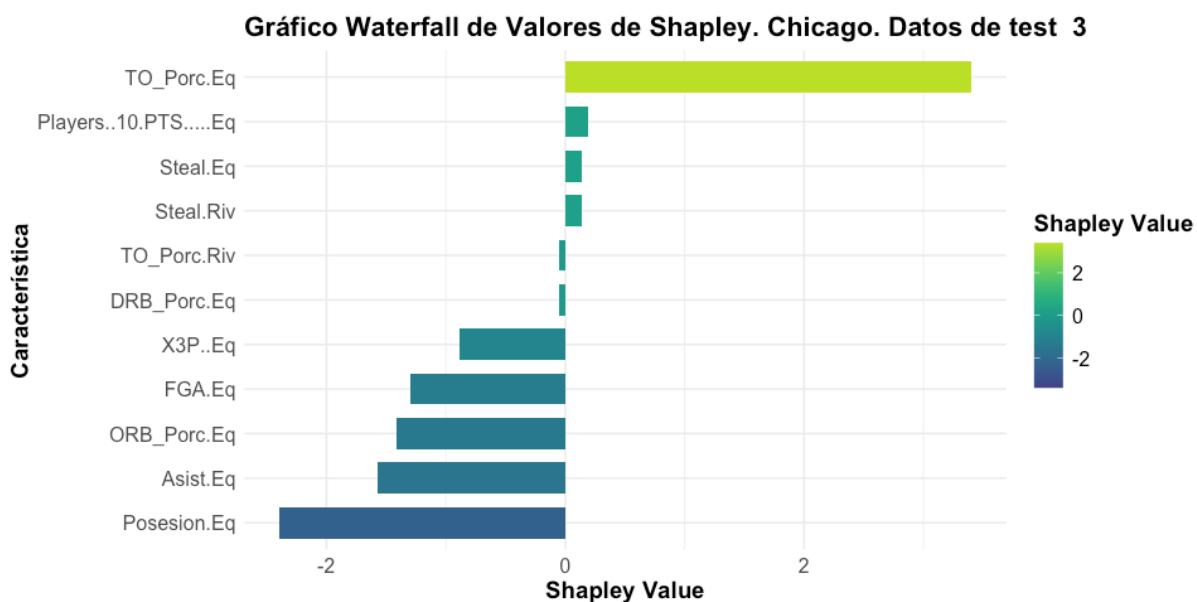


Figura 6.23: Shapley values. Datos de test de Chicago Instancia3

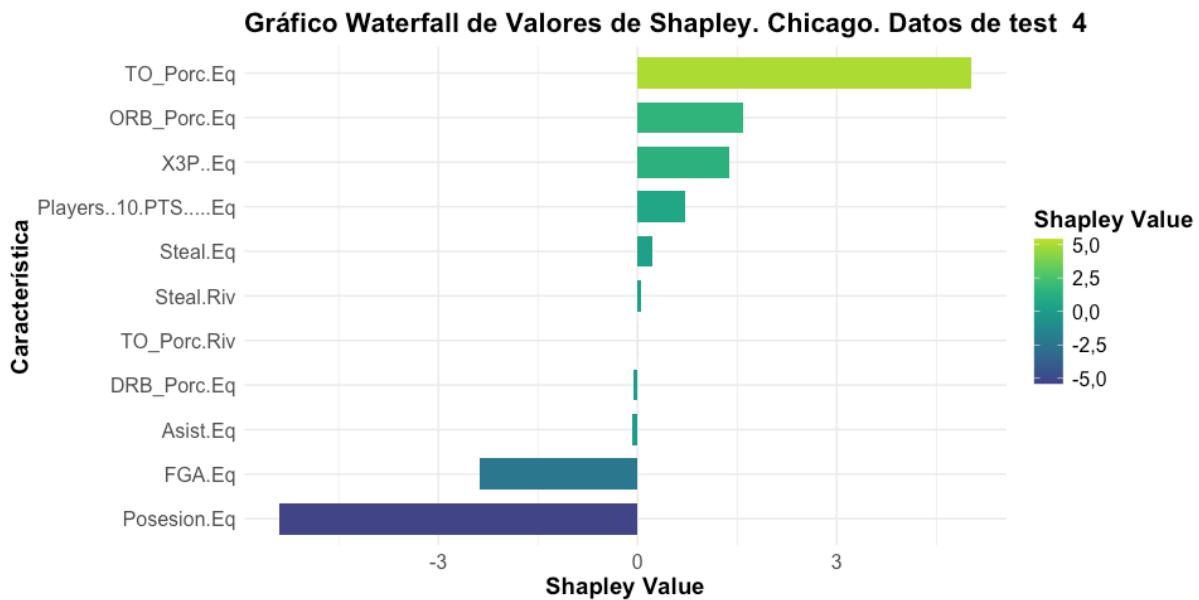


Figura 6.24: Shapley values. Datos de test de Chicago Instancia4

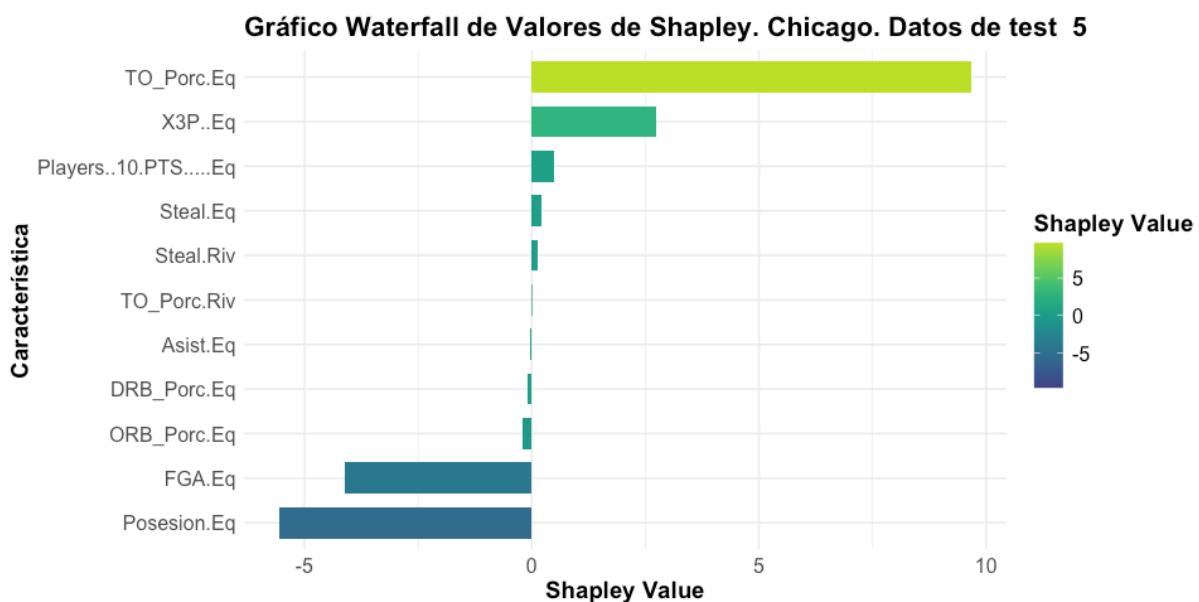


Figura 6.25: Shapley values. Datos de test de Chicago Instancia5

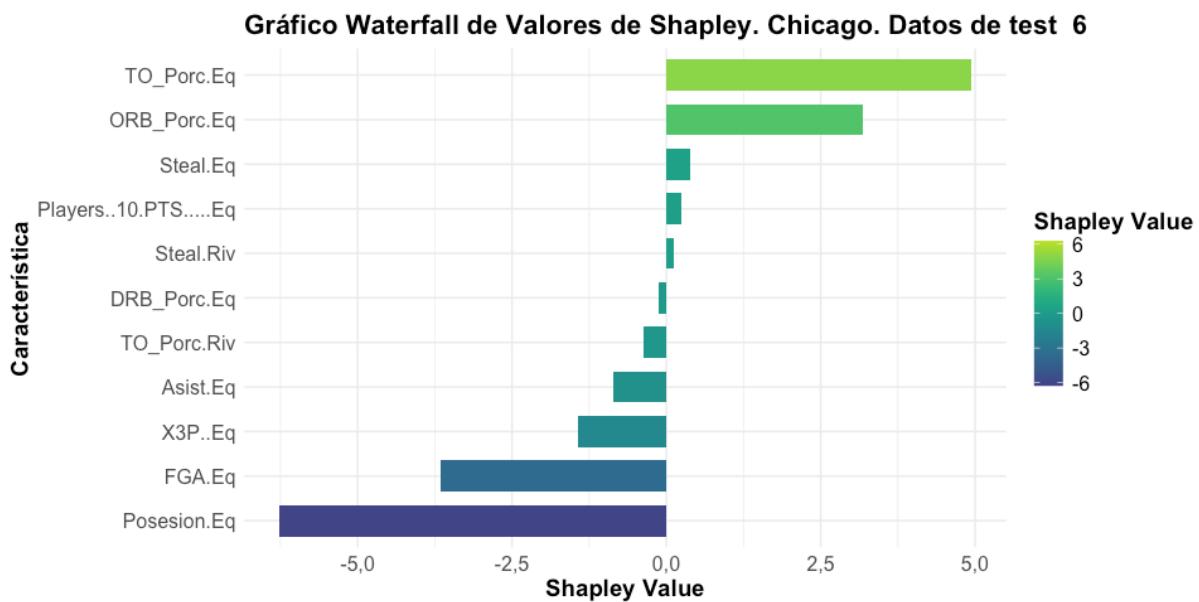


Figura 6.26: Shapley values. Datos de test de Chicago Instancia6

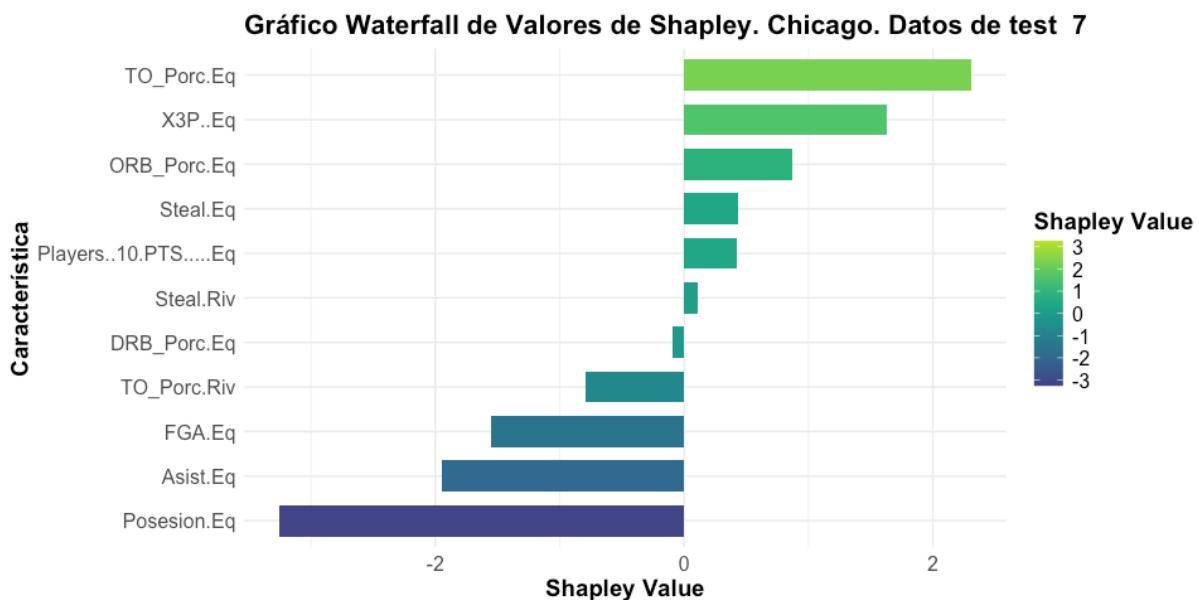


Figura 6.27: Shapley values. Datos de test de Chicago Instancia7

Con la información obtenida de los métodos explicativos aplicados al modelo predictivo, se sugiere que, para que Miami maximice su puntuación, debería aumentar el ritmo de juego. Esto permitiría obtener más posesiones y, por ende, más tiros de campo, además de reducir el porcentaje de pérdidas de balón que le penaliza. En cuanto al plan de juego del rival, es crucial ser más agresivos en defensa para provocar más pérdidas del equipo contrario y aumentar la presión en los primeros segundos para ralentizar su ritmo, ya que hasta ahora el

número de posesiones ha afectado negativamente la cantidad de puntos predicha por el modelo.

El modelo, a partir de los datos del boxscore del partido, predecía que Miami alcanzaría 98.08 puntos, mientras que Chicago anotaría 104, que es una precisión adecuada al resultado real del partido, que fue de 102-97. Se puede decir que si Miami, hubiera reducido el número de pérdidas en dicho partido y aumentado la velocidad del juego, como nos sugerían los métodos explicativos del modelo, se podría haber conseguido un resultado distinto y haber aumentado el número de puntos a favor.

Tabla 6.6: Resultado real del partido de interés

Equipo	X3P Eq	Asist Eq	DRB_Porc Eq	ORB_Porc Eq	FGA Eq	Players	10 PTS Eq	Posesion Eq	Steal Eq	Steal Riv	TO_Porc Eq	TO_Porc Riv	Puntos Eq	Rival
CHI	0.356	24	87.179	25.641	81		0.500	91.36	9	9	11.839	14.135	102	MIA
MIA	0.333	23	74.359	12.821	77		0.333	97.12	9	9	14.689	12.325	97	CHI

7. CONCLUSIONES

Se han entrenado un modelo de ML de Random Forest y una Red Neuronal con el objetivo de predecir el número de puntos. Ambos modelos se entrenaron con dos conjuntos de datos obtenidos de los Boxscore de partidos de la NBA. Por un lado, datos desde el año 2000 hasta el año 2024, y debido a que en los últimos años ha habido cambios sustanciales en el tipo de juego que se hace, ya que se ha visto que por ejemplo el número de triples y el número de puntos anotados, ha aumentado sustancialmente, se decidió también entrenar los modelos con datos únicamente de las últimas cinco temporadas, desde 2018.

La parametrización elegida para los modelos fue la siguiente,

Un modelo de Random Forest con el parámetro mtry establecido en 10, tanto para el análisis desde el año 2000 al 2024, como del 2018 al 2024

Una red neuronal para el conjunto de datos desde el año 2000 con un tamaño de 6 y un valor de decay de 0.5, y en el caso de los datos tomados desde el 2018, el tamaño de la red fue de 2 y decay 0.5.

Los cuatro modelos mostraron un rendimiento óptimo con un RMSE entre 5.86 y 7.06. Se utilizó también una media de rendimiento que tiene en cuenta el porcentaje de predicciones que están entre un +-5% del valor real y un +-10%.

La red neuronal tuvo un porcentaje de acierto del 65.49% cuando la precisión era del 5% y del 94.20% cuando la precisión era del 10%.

En cuanto a la importancia de las variables que se extrae de la información de los modelos, todos ellos muestran similitudes, siendo las variables más importantes aquellas que tienen en cuenta aspectos ofensivos del juego, como era de esperar. En general el número de posesiones, el porcentaje de tiros de tres y el porcentaje de perdidas, son factores comunes a los que todos los modelos les dan importancia. Se esperaba que un indicador como el porcentaje de jugadores que anotarán más de 10 puntos fuera un factor que tuviera peso específico en los modelos, pero en los resultados obtenidos, solo en la red neuronal entrenada con datos desde el año 2018 aparece entre los 3 primeros factores.

Posteriormente se evaluaron tres métodos explicativos de dichos modelos con el objetivo de ir más allá, del modelo propiamente dicho, y superar el problema de caja negra que tienen los modelos de ML.

Se decidió el analizar un método visual como son las curvas de expectativa condicional individual (ICE), los valores contrafácticos y los shapley values.

De los tres métodos los que más utilidad han tenido a la hora de proporcionar capacidad de entender las predicciones han sido los dos segundos.

En el caso de las curvas ICE, debido al gran numero de datos con los que se testeó el modelo, mostraron complejidad visual que dificultó la interpretación, volviéndose abarrotadas y poco claras. Además, el ruido en los datos llevó a variaciones, posiblemente, engañosas, que no reflejan la relación real entre las características y el número de puntos predichos. Al centrarse en una característica a la vez, también ocultaron interacciones importantes entre múltiples variables.

En el caso de los valores contrafácticos se analizó la variabilidad de estos con el objetivo de poder concluir que parámetros del juego son menos variables ya que esto significaría que son mas accionables a corto plazo y por otro lado aquellos que tienen un menor cambio respecto a los valores reales de los datos de prueba.

Los cuatro modelos entrenados con datos, muestran medias similares en las variables entre Random Forest y la red neuronal, indicando que ambos producen valores contrafactuales consistentes. Además, los bajos coeficientes de variación asociados al porcentaje de rebotes defensivos, el número de tiros de campo y sobre todo el numero de posesiones, sugieren un comportamiento estable dichas métricas, lo cual supone que incrementar los valores de dichos parámetros, podría incrementar los puntos predichos entre el 5% y el 10%. Se calculó el cambio porcentual entre los valores contrafactuales y los valores reales de los datos de prueba para determinar qué porcentaje se necesita modificar para lograr un aumento del 5% al 10% en las predicciones. Además, se analiza el porcentaje de ceros, que muestra cuántos datos no requieren ajustes en el parámetro para mejorar las predicciones. Un alto porcentaje de ceros sugiere que no es necesario modificar ese parámetro para alcanzar el incremento deseado.

En el análisis de los datos desde el año 2000, se observó que variables como el porcentaje de rebotes defensivos y el número de pérdidas mostraron cambios porcentuales cercanos a cero, indicando que los valores contrafactuales y reales eran similares. Se identificaron parámetros de interés, como el número de posesiones y triples anotados, que presentaron porcentajes de ceros inferiores al 70%. El número de posesiones, con un cambio medio del 1.87%, demostró que pequeños ajustes podrían generar incrementos significativos en los puntos anotados. En el análisis de los datos entre 2018 y 2024, se mantuvo la misma tendencia, destacando que el porcentaje de ceros para triples en Random Forest fue del 40%, lo que sugiere que en el 60% de los casos hubo cambios. La red neuronal mostró un cambio porcentual medio en triples del 9.20%, un 5% menor que en otros modelos. En cuanto al análisis de los shapley values se hizo un análisis visual a través de los gráficos de enjambre y numérico obteniendo las medias de los valores absolutos y la desviación típica de los valores.

Los análisis de los valores SHAP para los cuatro modelos (dos Random Forest y dos redes neuronales) muestran algunas similitudes y diferencias en cuanto a la influencia de las características sobre las predicciones de puntos. En general, el porcentaje de tiros de tres puntos y el número de posesiones, el número de asistencias se destacan como factores clave, con un impacto positivo notable en las predicciones, especialmente en los modelos entrenados con datos más recientes (2018-2024). Además, en todos los modelos, se observa una variabilidad en el impacto de las características, lo que sugiere que su influencia puede variar significativamente según los contextos de los partidos. Además, el porcentaje de pérdidas tiene un impacto negativo en todos los modelos, pero su variabilidad es más marcada en las redes neuronales, donde también se destaca la influencia negativa de un alto número de tiros intentados. Por otro lado, características como el número de robos o el porcentaje de rebotes defensivos tienen un impacto limitado en las predicciones en los modelos de red neuronal, lo que sugiere que no son factores determinantes en este contexto.

El análisis numérico de los shapley values, nos permite observar un alto impacto de variables como el número de asistencias y el porcentaje de acierto en tiros de tres puntos, lo que resalta su relevancia constante tanto en los modelos entrenados con datos desde el año 2000 como en los de 2018. Sin embargo, se identifican diferencias significativas entre los modelos Random Forest y las redes neuronales en la captura del impacto de otras variables. Las redes neuronales muestran un mayor MASV para características como el número de posesiones,

tiros de campo y el porcentaje de rebotes ofensivos, lo que sugiere una mayor sensibilidad a cambios en estas variables. Por ejemplo, el MASV para tiros de campo en la red neuronal fue de 6.09, en comparación con 1.44 en Random Forest, evidenciando que las redes neuronales pueden ser más efectivas para detectar patrones complejos relacionados con la agresividad ofensiva del equipo, que los modelos Random Forest no capturan con la misma precisión.

Finalmente, este trabajo ha permitido mostrar una aplicación de los métodos explicativos a un contexto mas cercano a la realidad, demostrando su posible utilidad en el marco de un staff de un equipo de baloncesto. Se ha visto que, a partir del modelo de red neuronal entrenado con datos desde el 2018 y 2024, y aplicando los métodos explicativos, a las predicciones generadas con datos de las primeras jornadas de la NBA de la temporada pasada, se pueden obtener conclusiones válidas a la hora de preparar un partido, desde un punto de vista puramente matemático.

A modo de resumen las conclusiones de este trabajo son:

- La Red Neuronal mostró un mayor porcentaje de aciertos con precisiones del 5% (65.49%) y 10% (94.20%) en comparación con el Random Forest, y ambos modelos destacaron variables ofensivas como claves para la predicción.
- Se evaluaron tres métodos explicativos: curvas ICE, valores contrafácticos y Shapley values. Los contrafácticos y los SHAP resultaron más útiles que las ICE, que fueron difíciles de interpretar por su complejidad visual.
- Los valores contrafácticos y SHAP mostraron que ajustar variables como las posesiones o los triples podría aumentar los puntos predichos en un 5%-10%, y que los modelos basados en redes neuronales capturaron mejor la variabilidad en variables ofensivas.
- Los métodos explicativos (contrafácticos y SHAP) mostraron su aplicabilidad en un contexto real, permitiendo que el staff de un equipo de baloncesto use información matemática para preparar estrategias antes de los partidos.

•

8. BIBLIOGRAFÍA

1. Oliver D. Basketball on paper: Rules and tools for performance analysis [Internet]. Brassey's, Incorporated; 2004. Available from: <https://books.google.es/books?id=hDUK-rAVwbQC>
2. Kim B, Khanna R, Koyejo OO. Examples are not enough, learn to criticize! Criticism for interpretability. Lee D, Sugiyama M, Luxburg U, Guyon I, Garnett R, editors. 2016;29. Available from:
https://proceedings.neurips.cc/paper_files/paper/2016/file/5680522b8e2bb01943234bce7bf84534-Paper.pdf
3. Miller T. Explanation in artificial intelligence: Insights from the social sciences. 2018; Available from: <https://arxiv.org/abs/1706.07269>
4. Kahneman D, Mielke JC. Pensar rápido, pensar despacio [Internet]. DEBATE; 2012. (DEBATE). Available from: <https://books.google.es/books?id=Ypj75lf86zsC>
5. Molnar C. Interpretable machine learning: A guide for making black box models explainable [Internet]. 2nd ed. 2022. Available from:
<https://christophm.github.io/interpretable-ml-book>
6. Bernard L, Earl B, W BK. Predicting NBA Games Using Neural Networks. Journal of Quantitative Analysis in Sports [Internet]. 2009;5(1):1–17. Available from:
<https://ideas.repec.org/a/bpj/jqsprt/v5y2009i1n7.html>
7. Miljković D, Gajić L, Kovacevic A, Konjovic Z. [The use of data mining for basketball matches outcomes prediction](#). In 2010. p. 309–12.
8. Lin J, Short L, Sundaresan V. Predicting national basketball association winners. 2014.
9. Pai PF, ChangLiao LH, Lin KP. Analyzing basketball games by a support vector machines with decision tree model. Neural Computing and Applications [Internet]. 2016 Apr;28(12):4159–67. Available from: <http://dx.doi.org/10.1007/s00521-016-2321-9>

10. Huang ML, Lin YJ. Regression tree model for predicting game scores for the golden state warriors in the national basketball association. *Symmetry* [Internet]. 2020 May;12(5):835. Available from: <http://dx.doi.org/10.3390/sym12050835>
11. Lalwani A, Saraiya A, Singh A, Jain A, Dash T. Machine learning in sports: A case study on using explainable models for predicting outcomes of volleyball matches [Internet]. 2022. Available from: <https://arxiv.org/abs/2206.09258>
12. Hickey K, Zhou L, Tao J. Dissecting moneyball: Improving classification model interpretability in baseball pitch prediction. In: Proceedings of the 53rd hawaii international conference on system sciences [Internet]. Hawaii International Conference on System Sciences; 2020. (HICSS). Available from: <http://dx.doi.org/10.24251/HICSS.2020.031>
13. Wang Y, Liu W, Liu X. Explainable AI techniques with application to NBA gameplay prediction. *Neurocomputing* [Internet]. 2022 Apr;483:59–71. Available from: <http://dx.doi.org/10.1016/j.neucom.2022.01.098>
14. Bautiste FJS, Brunner D, Koch J, Laborie T, Yang L, El-Assady M. [The Big Three: A Practical Framework for Designing Decision Support Systems in Sports and an Application for Basketball](#). 2024;103–16.
15. Chaya. Random Forest Regression — levelup.gitconnected.com. <https://levelup.gitconnected.com/random-forest-regression-209c0f354c84>;
16. Olden JD, Joy MK, Death RG. An accurate comparison of methods for quantifying variable importance in artificial neural networks using simulated data. *Ecological Modelling* [Internet]. 2004;178(3):389–97. Available from: <https://www.sciencedirect.com/science/article/pii/S0304380004001565>
17. Una introducción a los valores SHAP y a la interpretabilidad del machine learning — datacamp.com. <https://www.datacamp.com/es/tutorial/introduction-to-shap-values-machine-learning-interpretability>;
18. trumbelj E, Kononenko I. Explaining prediction models and individual predictions with feature contributions. *Knowledge and Information Systems* [Internet]. 2014;41:647–65. Available from: <https://api.semanticscholar.org/CorpusID:2449098>

19. Random Forest: A Complete Guide for Machine Learning | Built In — [builtin.com](https://builtin.com/data-science/random-forest-algorithm).

<https://builtin.com/data-science/random-forest-algorithm>;

20. Ventajas y desventajas | Interactive Chaos — [interactivechaos.com](https://interactivechaos.com/es/manual/tutorial-de-machine-learning/ventajas-y-desventajas-0).

<https://interactivechaos.com/es/manual/tutorial-de-machine-learning/ventajas-y-desventajas-0>;

9. ANEXO

9.1. FICHEROS DE DATOS

El fichero de datos básicos contiene las columnas

- **Game Reference:** Un identificador único para cada partido en la base de datos.
- **Team:** Nombre del equipo al que pertenece el jugador.
- **Period:** Periodo del partido en el que se registraron las estadísticas (Partido Completo, Número de Cuarto, Número de Parte).
- **Starter:** Indica si el jugador fue parte del quinteto titular (sí/no).
- **Player Name:** Nombre completo del jugador.
- **Player Reference:** Un identificador único para cada jugador en la base de datos.
- **MP:** Minutos jugados por el jugador en el partido.
- **FG:** Canastas de campo anotadas por el jugador.
- **FGA:** Intentos de canastas de campo realizados por el jugador.
- **FG%:** Porcentaje de aciertos en canastas de campo. La fórmula es FG / FGA .
- **3P:** Triples anotados por el jugador.
- **3PA:** Intentos de triples realizados por el jugador.
- **3P%:** Porcentaje de aciertos en triples. La fórmula es $3P / 3PA$.
- **FT:** Tiros libres anotados por el jugador.
- **FTA:** Intentos de tiros libres realizados por el jugador.
- **FT%:** Porcentaje de aciertos en tiros libres. La fórmula es FT / FTA .
- **ORB:** Rebotes ofensivos capturados por el jugador.
- **DRB:** Rebotes defensivos capturados por el jugador.
- **TRB:** Rebotes totales (ofensivos + defensivos) capturados por el jugador.
- **AST:** Asistencias realizadas por el jugador.

- **STL:** Robos de balón realizados por el jugador.
- **BLK:** Tapones realizados por el jugador.
- **TOV:** Pérdidas de balón cometidas por el jugador.
- **PF:** Faltas personales cometidas por el jugador.
- **PTS:** Puntos anotados por el jugador.
- **+/-:** Diferencia de puntos del equipo mientras el jugador estuvo en cancha (puntos a favor menos puntos en contra).
- **Reason:** Motivo por el cual un jugador no participó o tuvo limitaciones en el partido (por ejemplo, lesión, descanso, etc.).

La estructura del fichero de datos avanzados es la siguiente:

- **Game Reference:** Un identificador único para cada partido en la base de datos.
- **Team:** Nombre del equipo al que pertenece el jugador.
- **Period:** Periodo del partido en el que se registraron las estadísticas.
- **Starter:** Indica si el jugador fue parte del quinteto titular (sí/no).
- **Player Name:** Nombre completo del jugador.
- **Player Reference:** Un identificador único para cada jugador en la base de datos.
- **MP:** Minutos jugados por el jugador en el partido.
- **TS%:** Porcentaje de Tiros Verdaderos; una medida de eficiencia de tiro que toma en cuenta tiros de campo, triples y tiros libres. La fórmula es $PTS / (2 * TSA)$, donde TSA son los Intentos de Tiro Verdadero.
- **eFG%:** Porcentaje de Tiros de Campo Efectivos; ajusta el porcentaje de tiros de campo para considerar los triples. La fórmula es $(FG + 0.5 * 3P) / FGA$.
- **3PAr%:** Proporción de intentos de triples en relación con los intentos totales de tiros de campo. La fórmula es $3PA / FGA$.
- **FTr:** Tasa de Tiros Libres; representa el número de tiros libres intentados por cada intento de tiro de campo. La fórmula es FTA / FGA .

- **ORB%:** Porcentaje de Rebotes Ofensivos; el porcentaje de rebotes ofensivos capturados por el jugador en relación con los rebotes ofensivos disponibles. La fórmula es $100 * (\text{ORB} * (\text{Tm MP} / 5)) / (\text{MP} * (\text{Tm ORB} + \text{Opp DRB}))$.
- **DRB%:** Porcentaje de Rebotes Defensivos; el porcentaje de rebotes defensivos capturados por el jugador en relación con los rebotes defensivos disponibles. La fórmula es $100 * (\text{DRB} * (\text{Tm MP} / 5)) / (\text{MP} * (\text{Tm DRB} + \text{Opp ORB}))$.
- **TRB%:** Porcentaje de Rebotes Totales; el porcentaje de rebotes totales (ofensivos y defensivos) capturados por el jugador en relación con los rebotes totales disponibles. La fórmula es $100 * (\text{TRB} * (\text{Tm MP} / 5)) / (\text{MP} * (\text{Tm TRB} + \text{Opp TRB}))$.
- **AST%:** Porcentaje de Asistencias; la proporción de canastas de campo de un equipo que un jugador asistió mientras estuvo en cancha. La fórmula es $100 * \text{AST} / (((\text{MP} / (\text{Tm MP} / 5)) * \text{Tm FG}) - \text{FG})$.
- **STL%:** Porcentaje de Robos de Balón; el porcentaje de posesiones defensivas del equipo en las que el jugador realizó un robo. La fórmula es $100 * (\text{STL} * (\text{Tm MP} / 5)) / (\text{MP} * \text{Opp Poss})$.
- **BLK%:** Porcentaje de Bloqueos; el porcentaje de intentos de tiro de los oponentes bloqueados por el jugador mientras estuvo en cancha. La fórmula es $100 * (\text{BLK} * (\text{Tm MP} / 5)) / (\text{MP} * (\text{Opp FGA} - \text{Opp 3PA}))$.
- **TOV%:** Porcentaje de Pérdidas de Balón; el porcentaje de pérdidas de balón por posesión. La fórmula es $100 * \text{TOV} / (\text{FGA} + 0.44 * \text{FTA} + \text{TOV})$.
- **USG%:** Porcentaje de Uso; mide la proporción de jugadas de equipo terminadas por un jugador mientras estuvo en cancha. La fórmula es $100 * ((\text{FGA} + 0.44 * \text{FTA} + \text{TOV}) * (\text{Tm MP} / 5)) / (\text{MP} * (\text{Tm FGA} + 0.44 * \text{Tm FTA} + \text{Tm TOV}))$.
- **ORtg:** Rating Ofensivo; estima los puntos producidos por un jugador por cada 100 posesiones individuales.
- **DRtg:** Rating Defensivo; estima los puntos permitidos por un jugador por cada 100 posesiones individuales.
- **BPM:** Box Plus-Minus; una estadística avanzada que estima el impacto de un jugador sobre el equipo por cada 100 posesiones en comparación con un jugador promedio.

- **Reason:** Motivo por el cual un jugador no participó o tuvo limitaciones en el partido (por ejemplo, lesión, descanso, etc.).

-

10. ÍNDICE DE ACRÓNIMOS

- **X3P..Eq:** Proporción de triples anotados por el equipo (Tiros de tres puntos exitosos).
- **Asist.Eq:** Número de asistencias realizadas por el equipo, es decir, pases que resultan en canastas.
- **DRB_Porc.Eq:** Porcentaje de rebotes defensivos obtenidos por el equipo, lo que mide su capacidad para capturar el balón después de un intento de tiro fallido del oponente.
- **ORB_Porc.Eq:** Porcentaje de rebotes ofensivos obtenidos por el equipo, reflejando la capacidad de recuperar el balón tras sus propios tiros fallidos.
- **FGA.Eq:** Número de intentos de tiro de campo del equipo (Field Goal Attempts), que incluye tanto tiros de dos como de tres puntos.
- **Players..10.PTS.....Eq:** Porcentaje de número de jugadores del equipo que anotaron al menos 10 puntos en el partido.
- **Posesion.Eq:** Número total de posesiones que tuvo el equipo durante el partido.
- **Steal.Eq:** Número de robos de balón realizados por el equipo, que ocurre cuando un jugador defensivo le quita el balón a un oponente.
- **Steal.Riv:** Número de robos de balón realizados por el equipo rival.
- **TO_Porc.Eq:** Porcentaje de pérdidas de balón cometidas por el equipo (Turnovers).
- **TO_Porc.Riv:** Porcentaje de pérdidas de balón cometidas por el equipo rival (Turnovers).