Justificación del Clasificador

Bustamante Callisaya Hansel Alain

Para este proyecto, se seleccionó el clasificador Random Forest (RF) debido a sus múltiples ventajas y adecuación al problema de clasificación planteado. Este modelo es ampliamente reconocido en la literatura por su capacidad para manejar datos heterogéneos y ofrecer resultados precisos y robustos en escenarios complejos.

Una de las principales razones para elegir Random Forest es su capacidad de manejar características de naturaleza mixta. En este caso, el dataset incluye variables continuas como la posición en el campo (*posición_x*, *posición_y*), la velocidad y la distancia al balón, así como variables categóricas como la formación del equipo y la zona controlada. Breiman (2001) destaca que este clasificador, al combinar múltiples árboles de decisión entrenados sobre subconjuntos diferentes del conjunto de datos, genera predicciones más precisas y menos propensas al sobreajuste.

Otro punto clave es que el Random Forest es una herramienta robusta frente a datos desbalanceados, lo cual es relevante en este problema. Por ejemplo, las categorías de jugadores (*Titular*, *Suplente* y *Reservista*) podrían no estar distribuidas uniformemente. En este sentido, el modelo permite ajustar pesos para garantizar que las clases menos representadas no queden ignoradas. Esto ha sido confirmado por investigaciones recientes como las de Chen et al. (2016), quienes resaltan su efectividad en problemas multiclase con desbalances.

Además, el modelo es útil para identificar las variables más influyentes en la clasificación. A través de la métrica de *importancia de las características*, es posible entender cómo factores como la posición en el campo, la velocidad o incluso la formación del equipo contribuyen al resultado. Esta capacidad explicativa es esencial para proporcionar interpretaciones útiles que apoyen la toma de decisiones, como señalan Kuhn y Johnson (2013).

Otra ventaja destacada es su capacidad para manejar relaciones no lineales entre las variables. En el contexto deportivo, factores como la presión rival y la posición relativa al balón podrían tener interacciones complejas que afectan la

clasificación del jugador. Random Forest es especialmente adecuado para capturar estas relaciones, incluso cuando no son explícitas en los datos, como se menciona en el libro de Hastie, Tibshirani y Friedman (2009).

Por último, el Random Forest combina precisión con un costo computacional razonable, lo cual resulta práctico para datasets de tamaño mediano como el presente, con 1500 registros. Esto lo convierte en una opción eficiente para entrenar modelos sin comprometer la calidad de las predicciones.

En conclusión, el clasificador Random Forest es una elección adecuada para este problema por su capacidad de manejar datos variados, capturar relaciones complejas, y ofrecer interpretaciones claras. Estos atributos garantizan que el modelo sea efectivo y práctico en este contexto.

Fuentes utilizadas

[1] Breiman, L. (2001). Random Forests. *Machine Learning*, 45(1), 5-32. DOI: 10.1023/A:1010933404324.

[2] Chen, T., Guestrin, C. (2016). XGBoost: A Scalable Tree Boosting System. *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining.* DOI: 10.1145/2939672.2939785.

[3] Kuhn, M., & Johnson, K. (2013). *Applied Predictive Modeling*. Springer. ISBN: 978-1461468486.

[4] Hastie, T., Tibshirani, R., & Friedman, J. (2009). *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*. Springer. ISBN: 978-0387848570.