Universidad de
SanAndrés

Ciencia de Datos

Propuesta de Investigación

Giliberti, Mateo; Menutti, Juan Cruz y Saravi, Valentina

Fecha de entrega: 7 de diciembre de 2024

1. Introducción

El análisis del rendimiento de los jugadores de fútbol es esencial para la toma de decisiones estratégicas en los clubes deportivos. Este análisis está basado en el desempeño de los jugadores durante los partidos, como por ejemplo cantidad de goles o pases, el cual permite asignar un puntaje a los jugadores (Lago Peñas,2022)). Comúnmente, los jugadores del primer equipo de ligas moderadamente famosas, son los que tienen un puntaje asignado en todos o gran parte de los partidos que juegan. En cambio, es poco probable que jugadores de divisiones inferiores o de clubes no tan conocidos tengan puntajes asignados en sus partidos, lo que puede generar limitaciones al interesado en el jugador. Contar con estos puntajes sería de gran utilidad, especialmente para presidentes de clubes, directores técnicos o managers, quienes obtendrían un gran beneficio al momento de elegir a quienes ascender al primer equipo, o asimismo, en la adquisición de nuevos jugadores para su club.

¿Las calificaciones de los jugadores en los partidos de fútbol, podrían ser predecidas usando sus estadísticas de partido? Esta es la pregunta de investigación que se abordará a lo largo del trabajo. Se propondrá un modelo predictivo basado en Random Forest para estimar el puntaje de rendimiento a jugadores que no lo tengan. El modelo utilizará una base de datos que contiene métricas de desempeño en distintos partidos (asistencias, minutos jugados, tiros al arco, etc) seguido del puntaje asignado a cada jugador, con el fin de poder identificar patrones que permitan predecir el puntaje a partir de estas características. Random Forest es ideal para el proyecto ya que es considerado un buen modelo predictor. Este método aleatoriza los predictores utilizados en cada nodo, lo que ayuda a no caer en un sobreajuste. Además de su capacidad de manejar una gran cantidad de datos y trabajar con variables tanto numéricas como categóricas, lo hace perfecto para la base de datos utilizada en este trabajo (Breiman, 2001).

2. Literatura Previa

En la literatura ya se han visto casos de la utilización de técnicas de Machine Learning para asignarle un puntaje a los jugadores (Harringthon y Marín, 2022; Morciano, Zingoni y Calabró, 2023).

Harringthon y Marín (2022) en su paper titulado “Football Analytics - Ranking de Jugadores” tiene como objetivo realizar un ranking de jugadores chilenos por posición, basándose en su desempeño durante el juego. Este trabajo demuestra cómo las características de un partido son capaces de demostrar el puntaje de un jugador. Para lograrlo, utiliza Machine Learning, específicamente Support Vector Machine y Regresión logística, para realizar

modelos predictivos. Los resultados obtenidos demuestran la efectividad de estos métodos en la predicción del objetivo.

Por otro lado, Morciano, Zingoni y Calabró (2023) en su estudio titulado "Prediction of Football Players' Performance Indicators via Random Forest Algorithm" también destacan la importancia de usar técnicas de Machine Learning para predecir el desempeño de los jugadores. Se recopilaron distintos datos biométricos, como la frecuencia cardiaca y la temperatura corporal durante los entrenamientos, y a partir de estos datos se implementó un modelo de Random Forest para predecir diferentes indicadores del rendimiento físico, como la distancia recorrida durante los partidos, la velocidad promedio, etc.

Estos hallazgos demuestran la importancia de implementar técnicas de Machine Learning en el ámbito deportivo, para poder predecir el rendimiento de los diferentes deportistas, y en base a esos resultados, no solo llevar a cabo sesiones de entrenamiento personalizadas, sino también obtener información valiosa para la toma de decisiones de los clubes.

3. Datos

Para realizar este proyecto, se va a utilizar la base de datos *Football Players Ratings*, que está disponible de forma gratuita en *Kaggle*. Esta base de datos fue generada para una tesis de maestría de la Universidad Tecnológica de Eindhoven y, para permitir que los usuarios puedan acceder a datos en relación a métricas de rendimiento de jugadores en partidos de fútbol, campo que aún los datos públicos son escasos (Manfredi,2020).

La base contiene datos acerca del desempeño de futbolistas en distintos partidos. Estas métricas de desempeño, fueron generadas por seis publicaciones deportivas, dentro de las que se incluyen dos calificadores algoritmos (WhoScored y SofaScore) y cuatro humanos expertos (Kicker, Bild, SkySports y The Guardian). Además de estas calificaciones, la base de datos cuenta con una amplia variedad de variables, como los goles, asistencias, pases, intercepciones, tarjetas rojas, si jugador es local o visitante, etc; y otras métricas provenientes de la teoría de redes, como la centralidad de la intermediación. Las variables clásicas ofrecen una perspectiva integral de la performance del jugador en el partido, que junto con las métricas de la teoría de redes, permiten analizar el juego de una forma más estructural. En total, el conjunto de datos cuenta con 50652 de observaciones que provienen de la temporada 2017-18 de la Premier League de Inglaterra y de la Bundesliga de Alemania, así como también partidos de la Copa Mundial de Francia 2018 y la Eurocopa 2016.

3.1 Estadística descriptiva

En esta sección, se presentan estadísticas descriptivas que pueden servir para conocer algunas características claves de la distribución de calificaciones asignadas a los jugadores, que provienen de evaluadores humanos y algoritmos, así como también de las métricas del desempeño por posición de los jugadores.

Para poder observar si las puntuaciones de los humanos y algoritmos difieren en promedio, se generó la Tabla 1. En esta se destaca que los puntajes asignados por humanos tienen una media mayor (6.7973) a las asignadas por los algoritmos (4.7079). También, la desviación estándar de estas calificaciones es considerablemente mayor en algoritmos (1.6682) que la de humanos (0.7084), lo que da un dato clave de la mayor variabilidad de las puntuaciones provenientes de algoritmos. No se observaron grandes diferencias de la mediana en relación a la media de las calificaciones.

	Mean	Median	Std
Humano	6.7973	6.7000	0.7084
Algoritmo	4.7079	4.5000	1.6682
Observaciones	50652	50652	50652

Tabla 1: Estadísticas descriptivas de las calificaciones asignadas por evaluadores humanos y algoritmos

La Tabla 2, fue generada con el propósito de observar los valores de métricas frecuentes por posición en la cancha y si varían por la condición de localía. Se observa que hay diferencias de las métricas entre posiciones, donde los delanteros tienen la mayor media de goles (0.318932) y asistencias (0.106061), aunque los mediocampistas tienen una media similar (0.091413). Además, los defensores tienen la mayor media de intercepciones (2.739177), seguidos por los mediocampistas (2.372288); y los suplentes tienen las medias más bajas en todas las métricas, lo que puede deberse a la menor cantidad de tiempo disponible en el partido. En relación a la condiciones de localia, se obtiene que los locales predominan en todas estas métricas. Estos resultados son consistentes con lo esperado.

	DF	FW	GK	MF	Suplentes	Visitante	Local
Goles	0.038006	0.318932	0.000000	0.105280	0.062528	0.086283	0.119050
Asistencias	0.036668	0.106061	0.000996	0.091413	0.029381	0.053785	0.073732
Pases precisos	38.781265	16.928293	16.404483	32.484801	8.760585	27.576869	28.401496
Intercepciones	2.739177	0.948845	0.880199	2.372288	0.544222	1.910857	1.971245

Tabla 2: Estadísticas de desempeño por posición en el campo y rendimiento de equipos locales y visitantes

Por último, se considera relevante evaluar las diferencias desagregadas por posición de

puntuaciones de algoritmos y humanos. Ver diferencias en este punto, sería importante para observar algunos patrones de cómo califica cada evaluador. En relación a esto, se puede ver en la Tabla 3, que hay diferencias de puntuaciones en todas las posiciones y que las diferencias son similares, exceptuando la de los suplentes (0.958263), caso en el que los algoritmos suelen puntuar mejor que los titulares.

	Algoritmo	Humano	Diferencia
DF	6.897481	4.667744	2.229737
FW	7.027897	4.616627	2.411270
GK	6.691791	4.377193	2.314598
MF	6.908831	4.689940	2.218891
Sub	6.288142	5.329879	0.958263

Tabla 3: Puntajes por posición y evaluación de humanos y algoritmos, con diferencia entre evaluadores humanos y algoritmos en terminos absolutos

4. Metodología

En esta sección se presenta el proceso que se realizará para el análisis de los datos y la estimación de los modelos. Se va a hacer foco en la limpieza de datos, el entrenamiento de modelos Random Forest y la evaluación de su desempeño.

4.1 Revisión de los datos

El primer paso será realizar una revisión de la base de datos, en la que se va a buscar detectar valores anormales. Por ejemplo, si se encuentran casos de jugadores con 10 goles metidos en el mismo partido, se procederá a realizar una verificación de la fiabilidad de los casos en cuestión. Es posible que un jugador de fútbol meta gran cantidad de goles, realice muchas asistencias o pases precisos, pero frente a valores tan grandes como el del ejemplo, se intentará corroborar la información con fuentes deportivas o sitios web especializados. En caso de encontrar inconsistencias, se corregirán esas observaciones del conjunto de datos. Esta limpieza, garantizará que los datos sean fiables para asegurar un entrenamiento confiable del modelo.

4.2 Entrenamiento de los modelos

Luego de la revisión, se entrenarán dos modelos de Random Forest con árboles de regresión (CART), debido a que se utilizaran dos variables continuas como outcomes (puntaje de humanos y algoritmos). Se tomó la decisión de entrenar dos modelos distintos, debido

a que en el análisis exploratorio se encontró que las puntuaciones de humanos y algoritmos difieren en promedio. Entrenar dos modelos, podría proporcionar estimaciones más específicas y lograr una visión más completa de los puntajes. De esta manera, se podría evitar distorsiones, porque los puntajes estimados del modelo, no se podrían comparar directamente con el puntaje proveniente de un tipo de evaluador específico (humano o algorítmico).

4.3 Evaluación del modelo y ajuste de hiperparámetros

Por último, se evaluarán los modelos con distintas complejidades utilizando k-fold cross validation con un valor de k = 5 para elegir el modelo con menor error cuadrático medio (MSE). Para esto, se va a generar un proceso de bootstraping con 1000 muestras para obtener estimaciones robustas de los parámetros del modelo. A continuación, se ajustarán los hiper parámetros de los modelos de Random Forest para optimizar su rendimiento y garantizar que los resultados sean lo más precisos posible. Los siguientes hiper parámetros serán considerados durante el proceso de ajuste:

- max_depth: Que controla la profundidad máxima de los árboles para evitar árboles excesivamente complejos que puedan producir overfitting, se probarán los valores: *None*, 5, 10, 15 y 20.
- $min\_samples\_split$: Define el número mínimo de muestras necesarias para dividir un nodo, y se evaluarán valores de 2, 5, 10 y 20.
- $min\_samples\_leaf$: Este parámetro indica el número mínimo de observaciones finales por nodo. Se probarán los valores 1, 2, 5 y 10.
- $max\_features$: Determina cuántas variables (m) se utilizaran en cada nodo para realizar una división. Se probarán las configuraciones, *auto*, *sqrt* y el total de variables (p).

5 Conclusiones

Hoy en día el uso de la tecnología en el mundo deportivo se ha vuelto una herramienta crucial para potenciar el rendimiento de los deportistas, diseñar entrenamientos especializados y adoptar nuevas estrategias de juego. Por lo tanto, en este contexto, la implementación de un modelo de Random Forest para predecir los puntaje/calificaciones promedio de los jugadores en los partidos representa una oportunidad significativa y revolucionaria para los presidentes de los clubes, en especial, aquellos clubes que están en busca de nuevos

talentos, ya sea dentro de su propia institución o en clubes de ligas inferiores.

Sin embargo, este análisis afronta un par de desafíos que debemos tener en cuenta. Una de las limitaciones principales es la representatividad de los datos. Si bien la base utilizada presenta gran cantidad de estadísticas significativas, provienen principalmente de ligas "top". Esto puede significar un sesgo a la hora de generalizar los datos a jugadores de ligas inferiores o de equipos de reserva, ya que el nivel de competencia es diferente. Por otro lado, si no se realiza un buen ajuste de los hiperparametros, existe el riesgo de sobreajustar el modelo, reduciendo su efectividad. Y por último, factores como el estado emocional de los jugadores o decisiones tácticas de los entrenadores (como por ejemplo utilizar a un jugador en una posición que no es la suya) no se incluyen en la base de datos, lo que podría limitar la precisión de las predicciones.

En conclusión, aunque la metodología propuesta puede ayudar a cubrir la falta de información que se tiene sobre el rendimiento de los jugadores de la reserva y/o ligas inferiores, y así proporcionarle a los distintos presidentes de los clubes una visión más clara sobre qué decisiones tomar a la hora de comprar o ascender jugadores al primer equipo, es necesario tener en cuenta las limitaciones mencionadas y realizar ajustes constantes a los datos para así afrontar los cambios en la competencia y maximizar el éxito de la propuesta.

Referencias

Breiman, L. (2001). Random Forests. *Machine Learning, 45*, 5–32. https://doi.org/10.1023/A:1010933404324

Lago Peñas, C. (2022). EL ANALISIS DEL RENDIMIENTO EN LOS DEPORTES DE EQUIPO. ALGUNAS CONSIDERACIONES METODOLÓGICAS. *Acción Motriz, 1*(1), 41–58. Recuperado de https://www.accionmotriz.com/index.php/accionmotriz/article/view/5

María, H. A. A., Ricardo, M. F. W. (2022). Football Analytics - Ranking de jugadores. Recuperado de https://repositorio.udd.cl/items/bcce9e2a-fc8b-499e-9764-1c9cb113b4d4

Morciano, G., Zingoni, A., Calabrò, G. (2023). Prediction of football players' performance indicators via random forest algorithm. 2023 IEEE International Conference on Metrology for eXtended Reality, Artificial Intelligence and Neural Engineering (MetroXRAINE), 201–205. https://doi.org/10.1109/MetroXRAINE58569.2023.10405671

Manfredi, G. (2020). Football players ratings. Kaggle. Recuperado de https://www.kaggle.com/datasets/gabrielmanfredi/football-players-ratings