Instituto Tecnológico de Buenos Aires

82.19 - Analítica del deporte

Informe trabajo final

Grupo 1

Integrantes del grupo:

Manuel Hanono - 62265

Sofia Weintraub - 62583

Sol Winkel - 62409

Docentes titulares:

Juan Ferlaino

Diego Gaston Vilches Antao

Fecha entrega: 24/06/2024

Primer cuatrimestre 2024

Índice

1. Introducción

Se detectó que los jugadores muchas veces no regulan sus esfuerzos, lo que genera que se necesiten realizar cambios durante el partido, debido al **cansancio**, así como también se generan lesiones. Esto deriva en una baja del rendimiento individual del jugador, lo que conlleva a una baja de rendimiento colectivo, afectando a la totalidad del equipo.

Es por este motivo que se cree fundamental disponer de herramientas avanzadas de análisis en la era del deporte moderno, y la predicción del rendimiento de los jugadores se ha convertido en un factor crucial para el éxito.

En el trabajo a desarrollar se propone analizar en qué punto los jugadores llegan a su límite de rendimiento y cansancio, para así evitar cambios “obligados”, ayudar al jugador a regular mejor su performance durante el partido y a su vez colaborar con el cuerpo técnico para que puedan realizar una mejor planificación de los partidos en cuestión.

Este informe presenta un innovador proyecto de analítica deportiva desarrollado para un club de primera división, cuyo objetivo es proporcionar al cuerpo técnico una herramienta poderosa y eficiente para evaluar y optimizar el rendimiento de sus jugadores.

A través de la implementación de técnicas de machine learning, se diseñó una herramienta de predicción de rendimiento para la primera división del club de fútbol argentino. No sólo se buscará mejorar la capacidad de monitoreo y gestión del rendimiento individual, sino que también aporta una ventaja competitiva significativa, permitiendo al club mantenerse a la vanguardia en el uso de tecnologías avanzadas en el deporte.

Esta herramienta permitirá al cuerpo técnico tomar decisiones informadas de manera rápida y eficiente, basadas en datos obtenidos de los dispositivos GPS de Catapult, que ofrecen un análisis detallado y en tiempo real del desempeño de cada jugador.

El objetivo principal de este trabajo es desarrollar un sistema predictivo que permita estimar el umbral de rendimiento de cada jugador, basándose en datos obtenidos de los datos históricos de los dispositivos GPS utilizados. En otras palabras, se buscará predecir cuál será el límite físico de un jugador en un partido.

2. Base de datos

2.1 Obtención de la base de datos

El club con el que se trabajó compartió con el equipo los datos de diferentes partidos del año 2023, todos en formato PDF. Contiene los datos GPS de cada jugador, distinguiendo por partido y por tiempo jugado del partido, ya sea primer o segundo tiempo.

Los datos GPS son aquellos generados por los dispositivos Catapult, los cuales registran una variedad de métricas relacionadas con el rendimiento de los jugadores.

Se cuenta con registros desde el 4/04/2023 hasta el 11/11/2023, entre los cuales se encuentran partidos de Copa Libertadores, Copa Argentina y del torneo local.

Entre las variables originales se pueden mencionar las siguientes:

- Jugador → nombre y apellido del jugador
- Minutos → cantidad de minutos jugados por el jugador
- Avg Dist (Sess) (m) → distancia promedio recorrida, en metros
- Zona 4 19,9 A 25,1 KM/H → cantidad de piques en zona 4
- Zona 5 > 25,1 KM/H → cantidad de piques en zona 5
- Aceleraciones intensas → cantidad de aceleraciones
- Desaceleraciones intensas → cantidad de desaceleraciones
- Aceleraciones + desaceleraciones intensas → suma de las 2 variables anteriores
- Cantidad de sprints total → cantidad de ráfagas de velocidad máxima
- Promedio de esfuerzos repetidos
- Máxima velocidad (KM/H) → máxima velocidad alcanzada
- Metros/minuto → cantidad de metros recorridos por minuto
- Tiempo → primer o segundo tiempo del partido

2.2 Lectura de la base de datos

Tal como se mencionó anteriormente, se recibió, por parte del club, un archivo PDF con la información de cada partido. Estos PDF archivos se generan de manera automática por Catapult

Al necesitar estos datos en un formato en el que se pudieran manipular, se procedió al procesamiento de cada uno de los archivos a fines de obtenerlos todos en una tabla. Para esto se utilizó la librería de Python de PyPDF2.

El código fue buscando los archivos PDF en un directorio especificado y se extrajo el texto de la página 6 de cada archivo PDF encontrado. Se aclaró la página 6, dado que ahí estaban los datos que eran necesarios para los modelos de predicción. Una vez extraídos los datos, se concatenaron en un mismo dataframe, con la librería Pandas.

Finalmente, a través del método to_excel se guardó este data frame como archivo de excel, que fue el que luego se usó para realizar los modelos de predicción.

2.3 Tratamiento y limpieza

Una vez obtenidos los datos en formato correspondiente, en primer lugar se procedió a anonimizar la base de datos, siguiendo con el pedido realizado del club.

Es por eso que a cada jugador de la base de datos se le asignó un número, y se procedió a eliminar la variable *'jugador'*. Se cuenta con información de 29 jugadores, por lo que los ID se encuentran entre 1 y 29.

Luego, se realizó un chequeo de valores nulos, duplicados o atípicos, pero no se encontraron. Por lo que no hubo tal limpieza respecto de estos aspectos.

Al tratarse de un análisis exhaustivo que requiere precisión, se optó por incluir nuevas variables al análisis en cuestión. Dichas variables son las siguientes:

- Jugador anonimizado → se relaciona a cada jugador con un ID único para respetar la anonimidad solicitada por el club
- Rival → a qué equipo se enfrentó en cada partido
- Número de fecha → a qué fecha del cada torneo, respectivamente, corresponden los datos
- Torneo → hay tres valores posibles que toma la variable:
    - Copa Libertadores
    - Copa Argentina
    - Torneo local
- Fecha → día en que se jugó el partido
- Categoría del partido → variable que indica si el partido es “importante” o “normal”
- Posición habitual → indica la posición en la que el jugador se desempeña habitualmente
- Gol → es una variable numérica, de tipo entero, que toma:
    - Cantidad de goles realizados por el jugador en cada fecha, respectivamente
    - 0 en caso de no convertir
- Asistencia → es una variable binaria que toma los siguientes valores:
    - 1 si el jugador realizó una asistencia de gol
    - 0 en caso contrario
- Lesión → es una variable binaria que toma los siguientes valores:
    - 1 si el jugador se lesionó en dicho partido
    - 0 en caso contrario
- Edad → edad que tiene el jugador al día de la fecha (24/06/2024(
- Peso → peso del jugador, medido en kilogramos
- Altura → altura del jugador, medida en metros

2.4 Split en train y test

Para la partición en conjuntos de entrenamiento y testeo, se agruparon los datos por fecha, indicando así una agrupación por partido. Una vez agrupado, se procedió a agrupar los nuevos datos pero ahora por jugador anonimizado. Esto se hizo para asegurar de que los datos de cada jugador se mantengan juntos, pero que se dividan temporalmente.

El código garantiza que los datos de cada jugador se dividan de manera que las primeras observaciones (80%) se utilicen para el entrenamiento y las últimas observaciones (20%) para la prueba, manteniendo la coherencia temporal.

2.5 Generación del umbral

En este proceso de análisis de datos, se emplearon diversas técnicas de manipulación y agregación para preparar los datos de rendimiento de jugadores de fútbol. En primer lugar, se utilizaron bibliotecas especializadas para el escalado y la manipulación de datos en formato de tabla, lo cual facilitó las operaciones posteriores.

Para resumir los datos, se definió un diccionario que especifica cómo combinar las diferentes columnas de datos. Por ejemplo, se suman los minutos jugados y las distancias recorridas, mientras que otras columnas, como el nombre del torneo o la posición del jugador, se mantienen sin cambios. Esto asegura que los datos se agreguen de manera coherente y significativa.

Posteriormente, se agruparon los datos tanto de entrenamiento como de prueba por jugador y fecha. Durante esta agrupación, se aplica el diccionario de agregaciones para resumir la información de cada jugador en cada fecha específica. Esto significa que si un jugador tiene múltiples registros en una misma fecha, estos se combinan de acuerdo con las reglas definidas, proporcionando un resumen claro y conciso de su rendimiento en ese día.

Se identificaron ciertas columnas que representan valores acumulativos para aplicarles un procesamiento especial. Se calculó el percentil 80 de los valores de estas columnas para cada jugador. Este cálculo estableció el umbral por debajo del cual se encuentra el 80% de las observaciones, ofreciendo una métrica útil para predecir futuros rendimientos y evaluar el desempeño de los jugadores.

Luego, se creó una lista estructurada donde cada entrada corresponde a un jugador específico y sus respectivos umbrales calculados para las columnas seleccionadas. Esta lista se convierte en una nueva tabla de datos, facilitando la comparación y el análisis.

Finalmente, los datos originales de entrenamiento y prueba se combinan con la tabla de intervalos, integrando los valores observados con los umbrales calculados y los nuevos valores promedio. Esta integración permite una análisis más enriquecedor y una modelización predictiva más precisa, ya que se dispone tanto de los datos originales como de nuevas métricas derivadas.

3. Modelos de predicción

Se desarrollaron tres modelos de predicción diferentes para estimar el umbral de rendimiento de los jugadores: regresión lineal, Random Forest y XG Boost. Para realizar una comparación exhaustiva de los tres modelos se usó la métrica de evaluación del coeficiente de determinación (r^2).

3.1 Regresión Lineal

El modelo de regresión lineal se utilizó como una aproximación inicial debido a su simplicidad y facilidad de interpretación. Este modelo asume una relación lineal entre las variables independientes y la variable dependiente, permitiendo identificar tendencias generales en los datos.

3.2 Random Forest

El modelo de Random Forest se seleccionó por su capacidad para manejar relaciones no lineales y su robustez ante datos ruidosos. Random Forest crea múltiples árboles de decisión y combina sus predicciones para mejorar la precisión y reducir el riesgo de sobreajuste.

3.3 XG Boost

El modelo de XG Boost se implementó debido a su alta eficiencia y precisión en tareas de predicción. XG Boost utiliza técnicas de boosting para iterativamente mejorar las predicciones combinando varios modelos débiles en uno más fuerte. Este modelo **se destacó como el más eficaz** en la predicción del rendimiento de los jugadores.

3.4 Feature importance

La importancia de las características resultó ser un concepto fundamental en el proceso de análisis, debido a que se refiere a la medida en la que cada característica, o variable, contribuye a la predicción del modelo.

En pos de este proyecto, entender la importancia de las características permitió entender qué variables influyen más en los umbrales de rendimiento de los jugadores, facilitando la interpretación de los resultados.

Al conocer las variables que más influyen en los umbrales de rendimiento, se podrá ajustar el modelo para enfocarse en estas características, eliminando aquellas que tienen poco impacto. Esto no sólo mejora la precisión del modelo, sino que también es útil para el cuerpo técnico; entender qué variables son más importantes para alcanzar ciertos umbrales puede ayudar en la toma de decisiones estratégicas.

Al tener el modelo XG Boost como el más eficaz y robusto, se procedió a ver qué variables son las más importantes. El feature importance del modelo es el siguiente:

1. Número de aceleraciones intensas (num_aceleraciones_intensas_predictor)
2. Número de aceleraciones desintensas (num_acel_desintensas_predictor)
3. Zona 5, más de 25,1 kilómetros por hora (zona_5_mas_25.1_kmh_predictor)
4. Distancia promedio, medida en metros (avg_dist_sess_m_predictor)
5. Número total de aceleraciones y desaceleraciones intensas (num_desaceleraciones_intensas_predictor)
6. Zona 4, entre 19,9 y 25,1 kilómetros por hora (zona_4_19.9_25.1_kmh_predictor)
7. Cantidad de minutos jugados (minutos_predictor)
8. Cantidad de sprints totales (num_sprints_total_predictor)
9. El peso del jugador (peso)
10. Posición habitual del jugador
11. Altura del jugador (altura)

4. Interfaz de usuario

La interfaz del usuario se desarrolló en Streamlit, un framework open source para la creación de aplicaciones web interactivas y basadas en datos. Se realizó la carga de la base de datos dividida en train y test en el framework. Luego se corrieron los modelos dentro de ella y se guardó un Json del modelo dentro del mismo. Se le agregó al Json la línea correspondiente y se realizó la predicción con el modelo. Debido a que el framework está diseñado para facilitar la creación de aplicaciones de

machine learning y visualización de datos, es simple la actualización de los umbrales en caso de obtener más datos, partido a partido. Únicamente se debe correr nuevamente el código, recargar la base y volver a correr el modelo para realizar la actualización, lo que es clave para la implementación de la herramienta.

Para el uso de la herramienta, el usuario selecciona un jugador (es un número ya que está anonimizado) y la dificultad del partido. Se visualiza información como la edad, información corporal del jugador, su posición y la cantidad de partidos jugados como titular y totales. Al realizar las predicciones se visualizan los umbrales (límites inferiores y superiores) para cada una de las variables a predecir.

5. Implementación a futuro

Para mejorar y automatizar el proceso de obtención de datos, se busca establecer un convenio o acuerdo de colaboración con el club para facilitar el acceso a estos datos. Se propone la integración directa con la API proporcionada por Catapult. Este avance tecnológico permitiría una ingesta de datos en tiempo real, eliminando la necesidad de extracción manual desde archivos PDF y asegurando una actualización constante y precisa de la información.

Esta integración no sólo optimizaría el flujo de trabajo, sino que también permitiría al cuerpo técnico disponer de datos actualizados y fiables de manera continua, mejorando así la capacidad de toma de decisiones y la eficiencia en la gestión del rendimiento de los jugadores.

6. Conclusiones

En conclusión, la herramienta planteada presenta nuevas posibilidades en el uso de los datos en el fútbol, específicamente, respecto al cansancio de los jugadores. Los beneficios de la misma implican beneficios para los jugadores, para el cuerpo técnico y para el equipo en general.

Teniendo el umbral de cansancio para constatar en vivo con la información del partido, le da al cuerpo técnico la posibilidad de ir modificando la estrategia del partido en función del cansancio y el objetivo del partido, siempre tomando decisiones informadas basadas en los datos crudos.

Para los jugadores, indirectamente los beneficia en que no bajan su rendimiento en cancha, sino que salen a tiempo, e incluso, podría avanzarse en un análisis futuro, se eviten ciertas lesiones, producto de la sobreexigencia durante un partido.

Para el club es claramente positivo ya que abre un nuevo paradigma para pensar los partidos, encarar los campeonatos y así, pensar en ganar y mejorar continuamente utilizando la herramienta tanto durante los partidos, como en la planificación y en la previa a los mismos.

7. Anexo

Anexo 1: Repositorio de datos y archivos

En el siguiente link podrán encontrar tanto la base de datos anonimizada, el código que se armó para el desarrollo del trabajo y las presentaciones

Link: repositorio de GitHub

Anexo 2: Interfaz de usuario

A continuación se adjunta el link a la interfaz de usuario desarrollada para que los clientes del club puedan hacer uso.

Link: aplicación para cliente