

# Análisis de Rendimiento con GPS



Esteban Sánchez de la Barquera Cerda.  
CODERHOUSE  
Comisión 39990.





# TABLA DE CONTENIDOS



**01**

**CONTEXTO Y  
AUDIENCIA**

**02**

**PREGUNTAS  
DE INTERÉS**

**03**

**VISIÓN GENERAL  
DEL DATASET**

**04**

**ANÁLISIS  
EXPLORATORIO**

**05**

**INSIGHTS**

**06**

**FEATURE  
ENGINEERING**

**07**

**SELECCIÓN DE  
MODELOS**

**08**

**OPTIMIZACIÓN  
DEL MODELO**

**09**

**CONCLUSIONES /  
RECOMENDACIONES**



# CONTEXTO Y AUDIENCIA

El equipo de primera división llamado WIMU TEAM ha decidido tomar en cuenta los datos GPS que sus jugadores del primer equipo tienen durante la semana de prueba de la nueva herramienta para gestionar los entrenamientos y sacar el máximo provecho a cada uno de ellos.

El análisis de datos proporcionados por un GPS es una herramienta valiosa para mejorar el rendimiento de un equipo de fútbol soccer.

La audiencia para esta presentación puede incluir entrenadores, jugadores y otros miembros del equipo, así como analistas deportivos.





# **PREGUNTAS DE INTERÉS**

Nuestra tarea es interpretar los datos que generan los GPS y generar visualizaciones que respondan las preguntas específicas que tiene el staff de entrenadores:

- 1.¿Qué tipo de entrenamiento es más popular?
- 2.¿Qué días de la semana es cuando más carga de trabajo en metros recorridos hay?
- 3.¿Cuáles son los jugadores con la mayor distancia recorrida?
- 4.¿Quiénes son los jugadores más veloces? y si existe alguna correlación respecto a su posición en el terreno de juego.

# VISTA GENERAL DATASET



## ARCHIVO



**Weeksessions .csv**

Contiene información sobre las sesiones de entrenamiento de los usuarios de una plataforma de monitoreo GPS durante una semana.

El archivo contiene 15 columnas y 525 filas de datos. Dichos datos se encuentran ya en su formato correcto para proceder con el análisis.

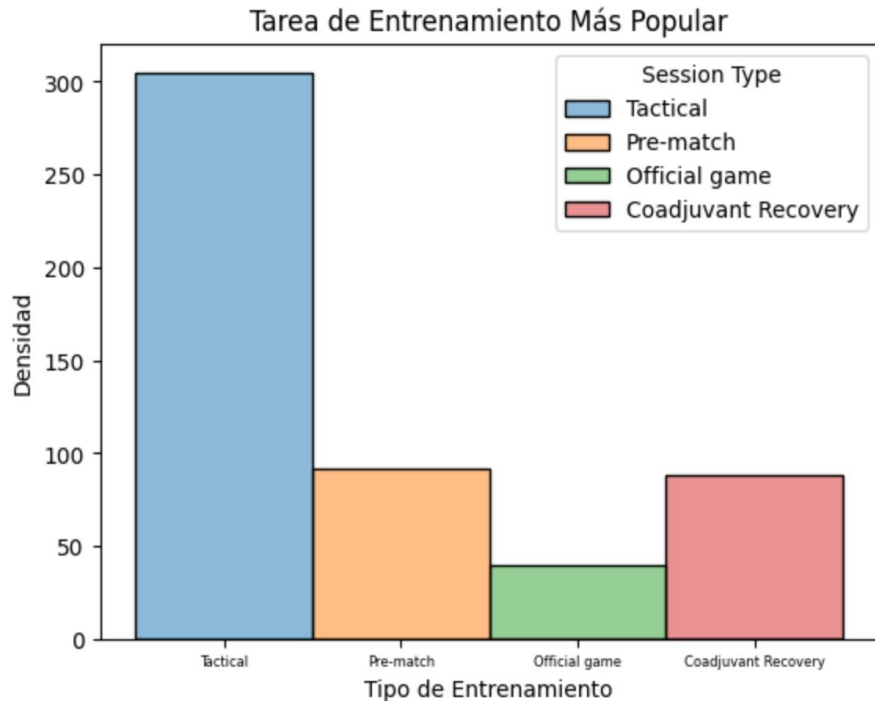


# ANÁLISIS EXPLORATORIO

## 1. ¿Qué tipo de entrenamiento es más popular?

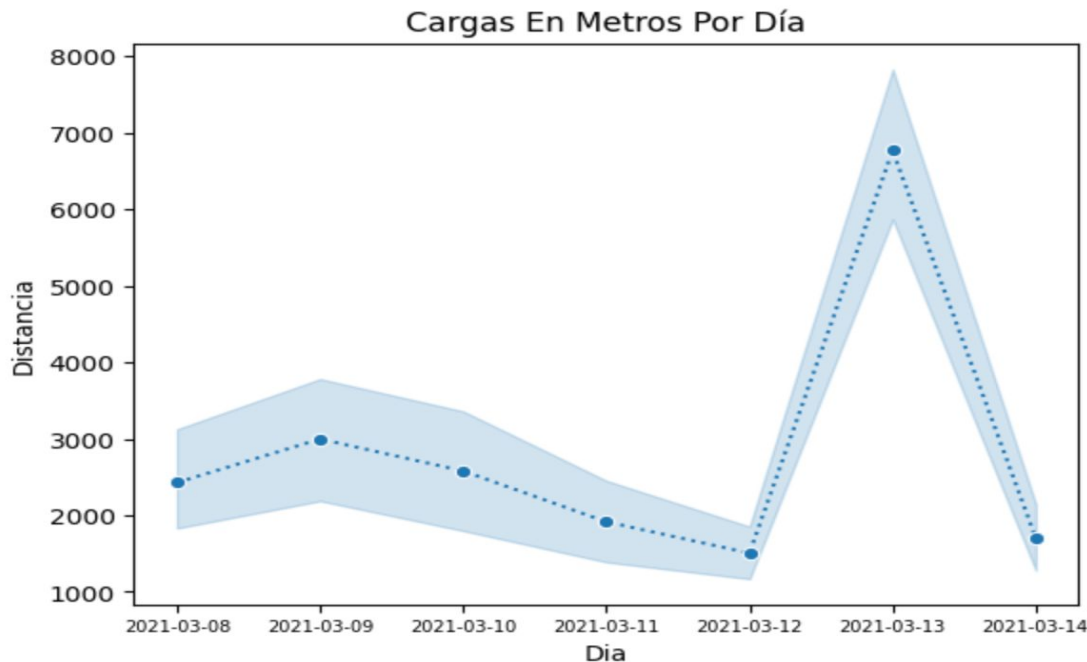
Con el gráfico de la derecha encontramos que el entrenamiento táctico colectivo tiene la mayor densidad de entrenamiento dentro de las sesiones.

Nos indica que el tipo de entrenamiento llevado a cabo forma parte de las nuevas tendencias en entrenamiento como lo es la Periodización Táctica o el Entrenamiento Estructurado, ya que la sesión se encuentra contextualizada porque no separa ninguno de los elementos del juego para entrenarlos por separado.



## ¿Qué días de la semana es cuando más carga de trabajo en metros recorridos hay?

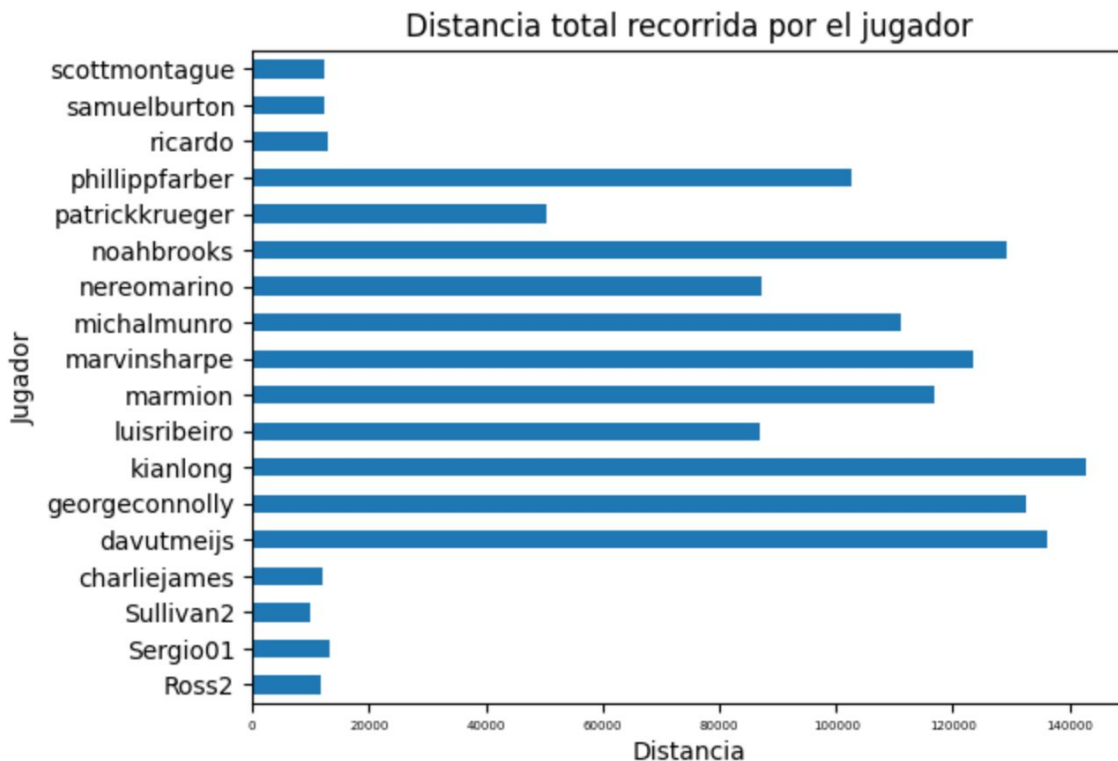
Podemos descubrir que la carga semanal de entrenamiento de la resistencia no está llegando a valores cercanos a la competencia; por lo tanto se debería ajustar la carga para que los estímulos en los días de entrenamiento de la resistencia sean lo más cercanos al día de competencia.



### 3.¿Cuáles son los jugadores con la mayor distancia recorrida?

Encontramos que hay jugadores que tienen una elevada distancia recorrida con respecto al de otros jugadores.

Podríamos interpretar que fueron los jugadores titulares el día del partido o incluso nos podría indicar a los jugadores que están apartados del entrenamiento haciendo tareas diferenciadas debido a una lesión.



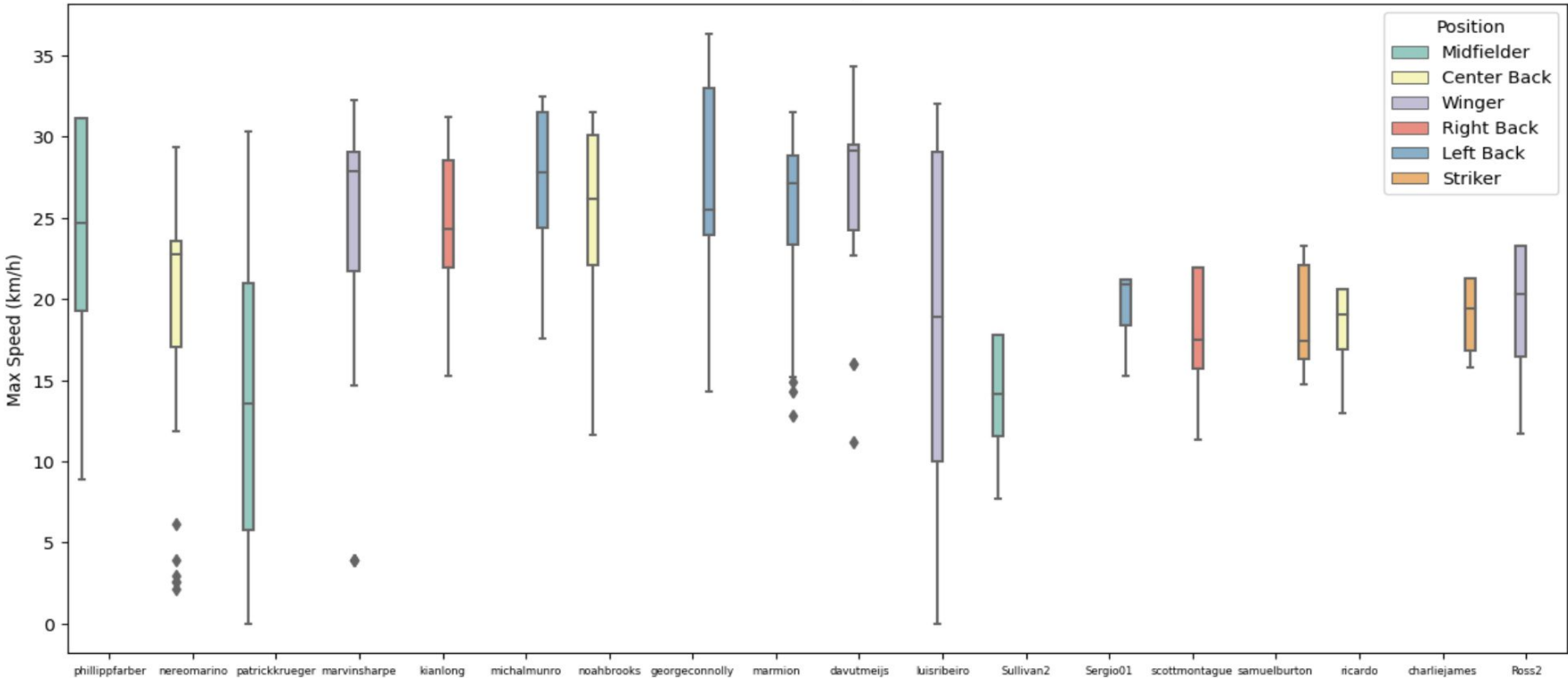


## 4 ¿Quiénes son los jugadores más veloces?

Las posiciones tienen un pico de velocidad máxima similar y, los valores por debajo nos hablan de aspectos tácticos que pudieran repercutir en el desempeño del equipo, la posición de defensor central tiene valores muy bajos lo cual nos habla de que el jugador pudo haber estado caminando y desconectado de las fases del juego.

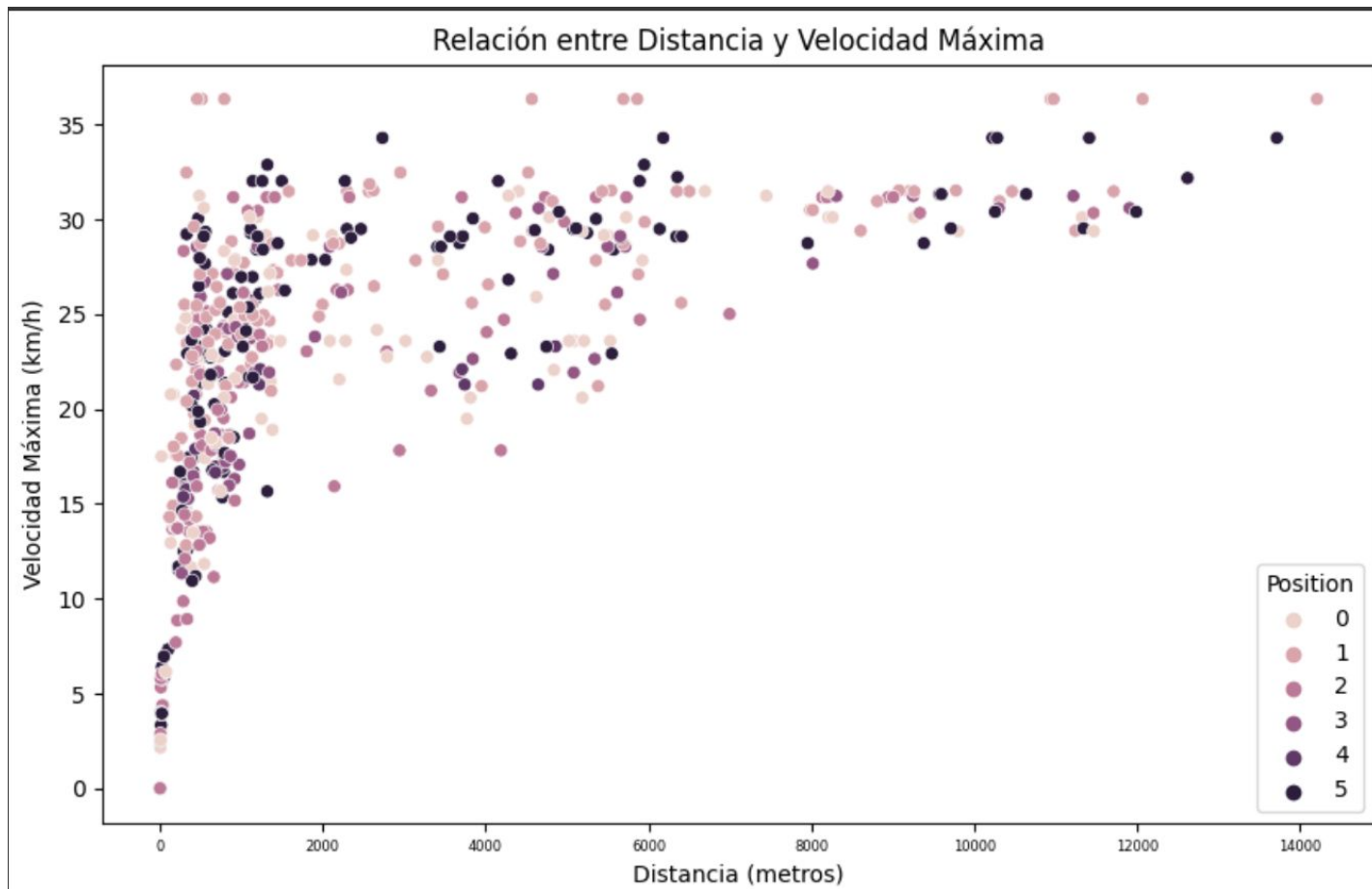
Cabe destacar el trabajo de los mediocampistas ya que pasan por la mayoría de las zonas de velocidad, por lo tanto nos indica una implicación máxima en el juego y con el equipo estando atentos a las transiciones, apoyos, desmarques, etc.

Velocidad Máxima por Posición



## 5.Relación entre velocidad y posición del jugador

Este gráfico permite identificar si existe una relación entre la distancia recorrida y la velocidad máxima, y si esta relación varía según la posición del jugador. Observamos que en TODAS las posiciones se recorren pocos metros a altas intensidades. Creemos que la posición 1 (Laterales) son los jugadores que alcanzan una mayor velocidad, ya que se asocia a que también son jugadores que tienen más espacio para lograr esas velocidades máximas.



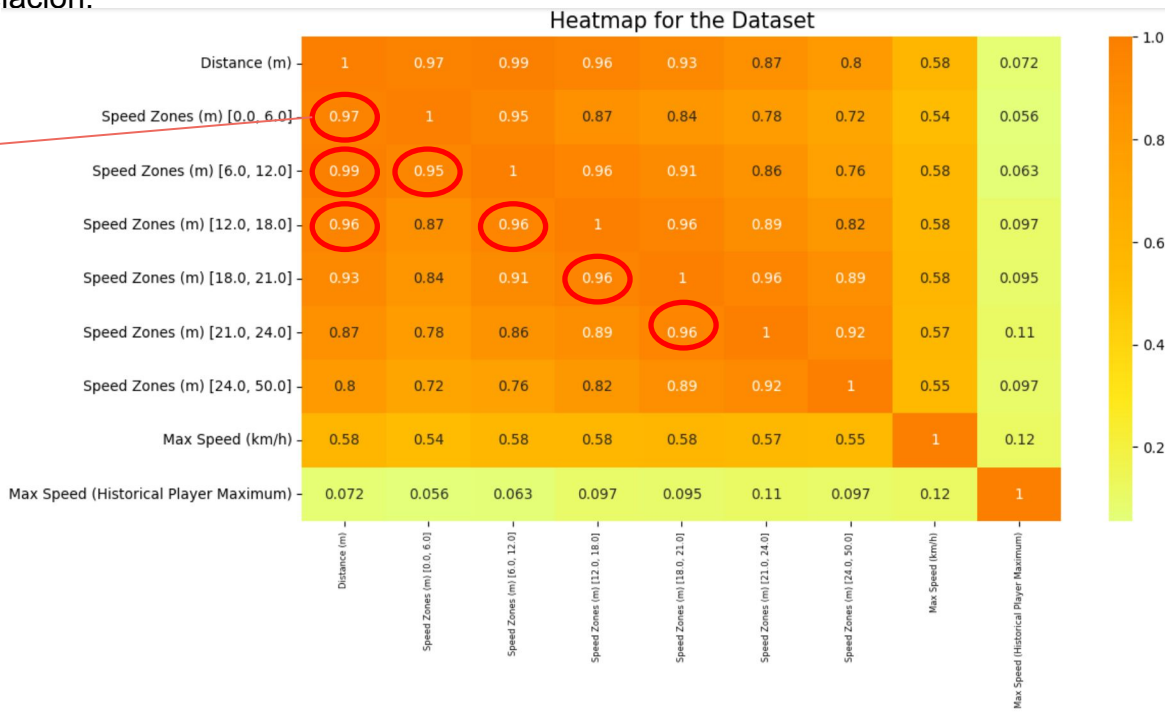


# HEATMAP



En el mapa de calor, las variables con alta correlación se representan mediante colores más fuertes, como el NARANJA. Las celdas con valores numéricos más altos indican una correlación más fuerte. Por lo tanto, para determinar las variables con alta correlación, debes buscar las celdas de color fuerte y con valores numéricos altos. Cuanto más cerca esté el valor de 1, más fuerte será la correlación.

Variables  
con fuerte  
relación.

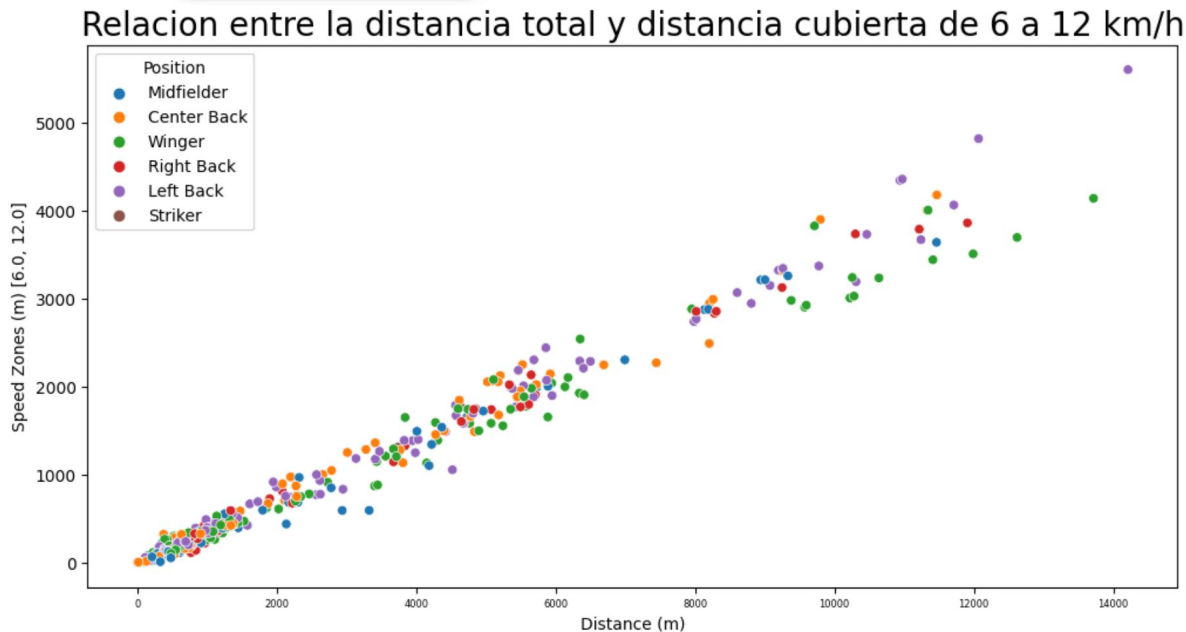




# SCATTERPLOT



Al observar el gráfico, se indica una relación entre la distancia total y la distancia cubierta a esta velocidad específica. Significa que los jugadores casi siempre están activos y en movimiento, lo cual nos podría decir que regularmente están atentos en el modelo de juego en la dimensión táctica y cubren de buena manera sus posiciones, recorridos, etc.





**INSIGHTS**



# INSIGHTS



- A través del histograma, se observa que la tarea de entrenamiento más popular es "Competitive Game". Esto podría sugerir que el equipo dedica una parte significativa de su tiempo de entrenamiento a juegos competitivos.
- El gráfico de líneas muestra la distancia recorrida por día ("Distance (m)") con un marcador distintivo para cada día. Puede observarse que en ciertos días se alcanzan distancias mayores, lo que podría relacionarse con entrenamientos más intensivos o partidos.
- El gráfico de barras horizontales muestra la distancia total recorrida por cada jugador. Esto proporciona información sobre qué jugadores han recorrido distancias más largas en general durante la semana.
- El gráfico de cajas (boxplot) muestra la distribución de las velocidades máximas por posición. Se observan diferencias en las velocidades máximas entre diferentes posiciones de jugadores. Por ejemplo, los delanteros pueden tener velocidades máximas más altas en comparación con los defensores.
- El heatmap muestra las correlaciones entre las variables en el conjunto de datos. Puedes observar qué variables están fuertemente correlacionadas positiva o negativamente. Esto puede ayudar a comprender las relaciones entre las métricas de rendimiento.
- El gráfico de dispersión muestra la relación entre la distancia total y la distancia cubierta en la velocidad de 6 a 12 km/h, con distinción por posición de los jugadores. Puede ayudar a identificar patrones en el rendimiento de los jugadores en función de la distancia y la velocidad.



# FEATURE ENGINEERING

- Creación de nuevas variables
- Transformación de variables existentes (normalización de variables, encoding)

## **LABEL ENCODING**

El Label Encoding es una técnica que asigna un valor numérico único a cada categoría en una variable categórica. Dicha técnica fue utilizada en las variables de “Posición” y “Tipo de Sesión”.



## **NORMALIZACIÓN DE DATOS**

La normalización de los datos es útil cuando se tienen variables con diferentes escalas y queremos que todas estén en la misma escala para que los modelos de ML funcionen de manera más efectiva. Esto nos ayuda a evitar que las variables con escalas más grandes dominen la influencia en el modelo en comparación con las variables con escalas bajas.

## **NUEVAS VARIABLES**

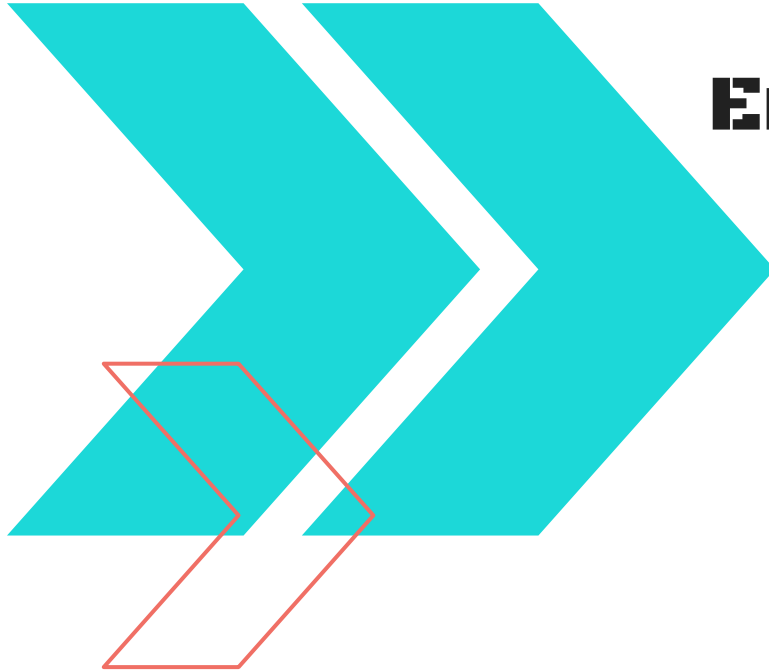
Crearemos nuevas características a partir de las características existentes para mejorar la capacidad predictiva del modelo.

1. Calcular la suma de las duraciones de las sesiones para cada usuario y agregarla como una nueva característica a nuestro Dataframe.
2. Promedio de la duración de las sesiones por jugador y agregarlo como una nueva característica.





# SELECCIÓN DE MODELOS



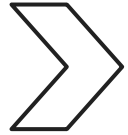
## Entrenamiento y Testeo

- Primer Ronda de Selección.
- Segunda Ronda de Selección.

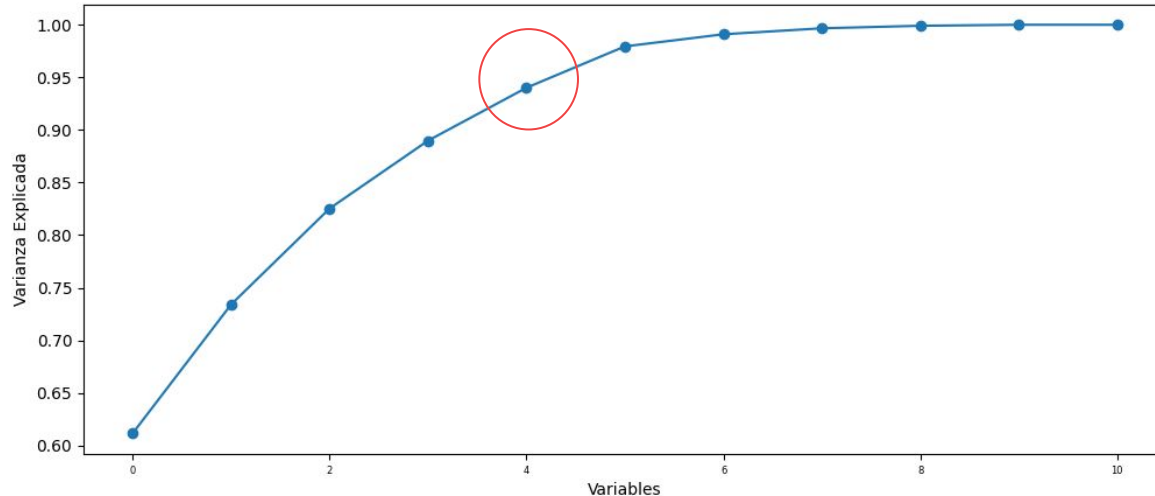




# Primer Ronda de Selección



Como primera instancia de selección se decidió realizar un modelo de regresión simple utilizando las variables que el componente PCA nos arrojó y creemos que seleccionar 4 variables nos era útil para trabajar con este modelo de regresión, dichas variables fueron: 'Speed Zones (m) [6.0, 12.0]', 'Speed Zones (m) [18.0, 21.0]', 'Speed Zones (m) [24.0, 50.0]' y 'Max Speed (Historical Player Maximum)' y nos damos cuenta que con esta selección la varianza puede ser explicada con un 94% de fiabilidad, lo cual siento que es muy alto y tenemos el riesgo de tener overfitting en nuestra predicción.



El rendimiento del modelo fue de:

**R2: 0.35737789254724206**

**MSE: 38.40476438433717**

**RMSE: 6.197157766616659**

**MAE: 4.688017874011857**

En conclusión podemos decir que nuestro modelo no ajusta a predecir muy bien nuestra proyección de velocidades máximas.



# Segunda Ronda de Selección



Con el trabajo realizado de Encoding e Ingeniería de Atributos, se pudieron crear nuevas características a partir de las ya existentes para mejorar la capacidad predictiva del modelo.

Decidimos probar un Modelo de Clasificación y 2 Modelos de Regresión, para decidir cuál es el que mejor se adapta a nuestras necesidades que son predecir la velocidad máxima que un jugador podría alcanzar.

## **Modelo de Clasificación (Random Forest Classifier):**

Con este modelo nos dimos cuenta que no puede ser tomado en cuenta para nuestro objetivo ya que no queremos clasificar las velocidades actuales de nuestro equipo, sino proyectar quiénes podrían darnos mayor velocidad con el tiempo; este modelo nos dio un Accuracy de: **0.3619047619047619**, quiere decir que es muy bajo y su varianza no explica en casi nada a nuestro dataset.


**Modelo de Regresión (Regresión Linear):** Nos da un resultado de:  
**MSE: 359.27068281548077**

**Modelo de Regresión (Random Forest Regressor):** Nos da un resultado de:  
**MSE: 224.02365999999998**

De acuerdo a los resultados mostrados por estos modelos creemos pertinente que el mejor modelo para trabajar será el **Random Forest Regressor** ya que comparando los 2 modelos nos da un resultado más bajo, por lo tanto más cercano a los resultados que tenemos en nuestra data original.



# OPTIMIZACIÓN DEL MODELO

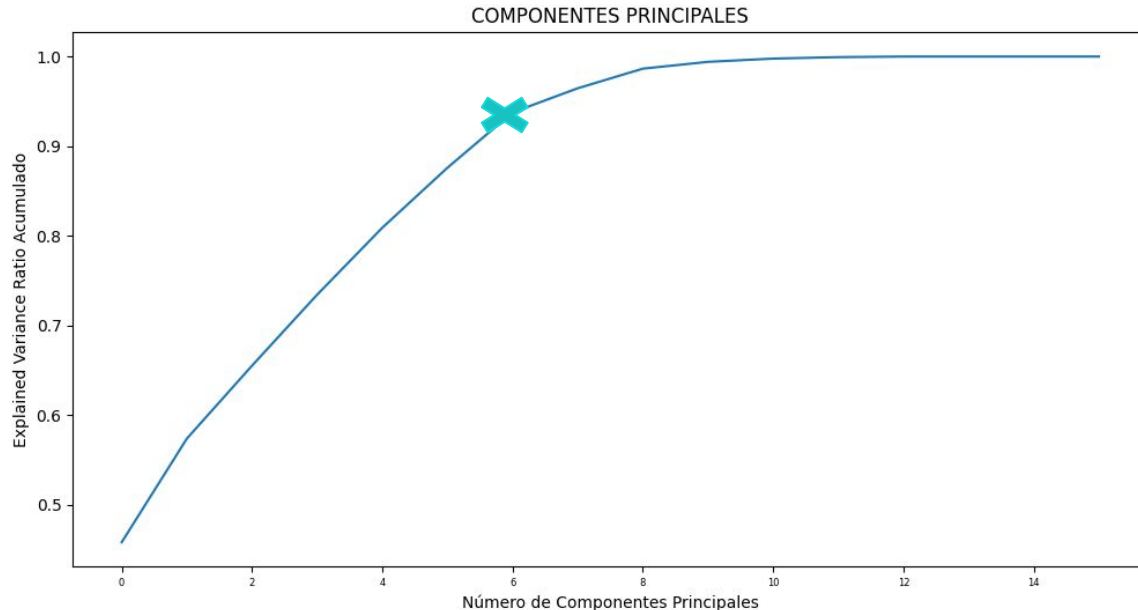
- PCA
  - CROSS-VALIDATION
  - XGBoost
- 



# Análisis de Componentes Principales (PCA)



Con este trabajo buscamos con cuántas variables se puede explicar la varianza de nuestros datos, como hemos agregado nuevas variables como la suma de las distancias y el promedio de dichas distancias veremos con cuantas variables podríamos trabajar de acuerdo al componente PCA. A diferencia de nuestro primer modelo podemos observar que nuestro quiebre se encuentra en 6 variables de acuerdo al gráfico.



Al considerar este componente con un resultado de 6 variables y considerar Bias-Variance Tradeoff para equilibrar la capacidad de ajuste de datos con la creación de nuevos datos los resultados en MSE que nos da este ajuste en nuestro modelo es de:

**MSE: 222.75205428571434**

Observamos que el resultado nos disminuye en 2 puntos con respecto al modelo; sin empezar a ajustar parámetros, por lo cual podemos decir que seguimos mejorando nuestro modelo.



# Cross-Validation



La validación cruzada es una herramienta esencial para evaluar, ajustar y mejorar modelos de aprendizaje automático. Ayuda a identificar problemas de sobreajuste, seleccionar los mejores hiper parámetros y proporcionar una estimación más precisa del rendimiento del modelo en datos reales. Por lo tanto, es una práctica recomendada que eleva la confianza de modelos, haciéndolos más robustos y confiables para realizar predicciones más confiables.

**Estimación más precisa del rendimiento:** Con esto nos aseguramos de no depender de una sola división de los datos en entrenamiento y prueba, la validación cruzada realiza múltiples divisiones y promedia los resultados.

Esta vez decidimos dividir nuestro conjunto de datos en **8 folds**, con esto nos aseguramos que tengamos una estimación más robusta del rendimiento del modelo ya que evalúa su rendimiento en los 8 múltiples subconjuntos de datos.

**Promedio de MSE utilizando validación cruzada:**  
**20.79007592525397**

Este resultado sugiere que las predicciones del modelo RandomForestRegressor tiene una tendencia a predecir las velocidades máximas con un error relativamente bajo en comparación con los valores reales, en función del contexto y escala de datos. Y al disminuir otra vez en **2** puntos el valor de MSE podemos decir que nuestro modelo sigue siendo optimizado, respecto al MSE con el análisis de componentes PCA.



# XGBoost



XGBoost nos puede ayudar a mejorar nuestro modelo de Random Forest Regressor al proporcionar un rendimiento superior, técnicas de regularización avanzadas, manejo de datos faltantes y una selección de características más eficiente. Además de que está diseñado para tener un mejor desempeño computacional, reduciendo la dimensionalidad de los datos y mejorar nuestras predicciones.

**Estimación más precisa del rendimiento:** Con esto nos aseguramos de no depender de una sola división de los datos en entrenamiento y prueba, la validación cruzada realiza múltiples divisiones y promedia los resultados.

Esta vez decidimos dividir nuestro conjunto de datos en **8 folds**, con esto nos aseguramos que tengamos una estimación más robusta del rendimiento del modelo ya que evalúa su rendimiento en los 8 múltiples subconjuntos de datos.

El rendimiento que nos da el modelo con XGBoost es de:  
**MSE: 171.54**

Comparado con el modelo RandomForestRegressor, obtenemos una señal positiva y sugiere que **XGBoost** está produciendo predicciones más precisas y un mejor ajuste a los datos.

En conclusión podemos decir que; un MSE más bajo y un mejor rendimiento en datos de prueba son indicativos de que **XGBoost** es un modelo más confiable y preciso ya que es una herramienta confiable para hacer predicciones en nuestro conjunto de datos y puede ser una opción confiable para nuestras predicciones futuras.



**CONCLUSIONES**

**Y**

**RECOMENDACIONES**







# CONCLUSIONES



El análisis de datos y la aplicación de modelos de machine learning pueden proporcionar información valiosa para mejorar el rendimiento de un equipo de fútbol soccer. El equipo puede utilizar estos hallazgos para ajustar su enfoque de entrenamiento y estrategia, así como para tomar decisiones informadas sobre la gestión de su equipo a continuación se enlistan las conclusiones más valiosas que identificamos en el presente proyecto.

- **Tipo de Entrenamiento Más Popular:** Táctico Colectivo es el entrenamiento más popular en el equipo. Esto sugiere que el equipo se enfoca en la integración de diferentes aspectos del juego en lugar de entrenar componentes individuales por separado.
- **Días de Mayor Carga de Trabajo en Metros Recorridos:** Sábados son los días de la semana en los que se registra la mayor carga de trabajo en términos de metros recorridos. Indica que la carga de entrenamiento semanal debería ajustarse a la alza para que los jugadores estén más cerca de las demandas de la competencia.
- **Jugadores con la Mayor Distancia Recorrida:** Esto puede ser útil para evaluar el rendimiento de los jugadores y detectar posibles diferencias en la carga de trabajo.



# CONCLUSIONES



- **Jugadores Más Veloces y su Posición:** Se analizó la velocidad máxima alcanzada por los jugadores y se observó que la posición en el terreno de juego no muestra una correlación clara con la velocidad máxima. Sin embargo, se identificaron diferencias en las velocidades máximas entre los jugadores, lo que podría ser relevante para la estrategia del equipo.
- **Modelos de Machine Learning:** Se entrenaron varios modelos de machine learning para predecir la velocidad máxima de los jugadores. El modelo **RandomForestRegressor** obtuvo un **MSE** de **220.79** en la predicción de velocidad máxima, lo que sugiere un buen rendimiento del modelo. Además, se utilizó la validación cruzada para evaluar los modelos y mejorar su rendimiento, terminando con un **MSE** de **171.54**, lo cual mejoró en gran medida nuestro modelo.
- **Feature Selection y PCA:** Se realizó una selección de características utilizando el modelo Random Forest para identificar las variables más importantes en la predicción de la velocidad máxima. También se aplicó PCA para reducir la dimensionalidad de los datos y explorar la importancia de los componentes principales.



# RECOMENDACIONES



**Ajustar la Carga de Entrenamiento:** Dado que los sábados son los días de mayor carga de trabajo en términos de metros recorridos, se recomienda ajustar la carga de entrenamiento en las sesiones de resistencia para que los jugadores estén más preparados en la competencia.

**Seguir Monitorizando el Rendimiento de los Jugadores:** Continuar monitorizando el rendimiento de los jugadores, especialmente en términos de distancia recorrida y velocidad máxima, para identificar tendencias a lo largo de la temporada y realizar ajustes en el entrenamiento.

**Considerar Análisis Más Detallados de Posiciones:** Aunque no se encontró una correlación clara entre la posición en el terreno de juego y la velocidad máxima, se recomienda realizar análisis más detallados por posición para entender mejor el rendimiento individual de los jugadores.

**Explorar Modelos Avanzados:** Además de los modelos utilizados en este análisis, se podría explorar el uso de modelos más avanzados de machine learning, como redes neuronales, para mejorar aún más la precisión en la predicción de la velocidad máxima.

**Seguir Refinando el Análisis de Datos:** El análisis de datos es una herramienta poderosa para la toma de decisiones en el deporte. Continuar refinando el análisis y explorando nuevas preguntas puede proporcionar insights adicionales para optimizar el rendimiento del equipo.