***Predicción del rendimiento en Running mediante técnicas de Machine Learning.***

*Un enfoque aplicado al sector deportivo.*

*Logotipo, Icono, nombre de la empresa

El contenido generado por IA puede ser incorrecto.*

**EVOLVE ACADEMY**

**MÁSTER EN DATA SCIENCE & INTELIGENCIA ARTIFICIAL**

Realizado por: Eduardo Sánchez

Año Académico: 2025/2026

**RESUMEN**

**ABSTRACT**

**TABLA DE CONTENIDOS**

[1. INTRODUCCIÓN 4](#_Toc210752875)

[1.1. OBJETIVO DEL PROYECTO 4](#_Toc210752876)

[1.2. MOTIVACIÓN 4](#_Toc210752877)

[2. MÉTODOLOGÍA 5](#_Toc210752878)

[2.1. ENFOQUE GENERAL 5](#_Toc210752879)

[2.2. HERRAMIENTAS UTILIZADAS 6](#_Toc210752880)

[3. RECOLECCIÓN DE DATOS 6](#_Toc210752881)

[3.1. PARTICIPANTES 6](#_Toc210752882)

[3.2. FUENTES DE DATOS 7](#_Toc210752883)

[3.3. ESTRUCTURA DE DATOS RAW 8](#_Toc210752884)

[4. PROCESAMIENTO Y LIMPIEZA DE DATOS 8](#_Toc210752885)

[4.1. LIMPIEZA DEL FORMULARIO 8](#_Toc210752886)

[4.2. LIMPIEZA DE DATOS DE STRAVA 9](#_Toc210752887)

[4.2.1. Problemas identificados en los datos raw de Strava. 9](#_Toc210752888)

[4.2.2. Proceso de limpieza aplicado: 10](#_Toc210752889)

[4.2.3. Aplicación a todos los usuarios 11](#_Toc210752890)

[4.2.4. Resultado 11](#_Toc210752891)

[4.3. UNIÓN DE DATASETS 11](#_Toc210752892)

[4.4. MÉTRICAS DE CALIDAD DEL PROCESO 12](#_Toc210752893)

[5. FEATURE ENGINEERING 12](#_Toc210752894)

[5.1. OBJETIVO 13](#_Toc210752895)

[5.2. FEATURES TEMPORALES CREADAS 13](#_Toc210752896)

[5.3. MÉTRICAS DE RENDIMIENTO 13](#_Toc210752897)

[5.4. IMPLEMENTACIÓN TÉCNICA 14](#_Toc210752898)

# INTRODUCCIÓN

## OBJETIVO DEL PROYECTO

El presente Trabajo Fin de Máster tiene como objetivo diseñar y desarrollar un sistema de predicción del rendimiento en corredores mediante el uso de técnicas de Machine Learning. Para ello, se plantea la integración de dos fuentes de información complementarias:

1. Datos de actividad deportiva extraídos de Strava, que proporcionan métricas objetivas relacionadas con la práctica del running (distancia, ritmo, frecuencia cardiaca, cadencia, desnivel acumulado, entre otros).
2. Información personal y de entrenamiento recopilada mediante cuestionarios específicos, que incorpora variables contextuales y subjetivas (edad, sexo, hábitos de entrenamiento, etc.).

La combinación de ambas fuentes permite construir modelos predictivos que no solo se ajusten al rendimiento pasado, sino que también sean capaces de estimar de forma personalizada el desempeño futuro de cada corredor en distancias estándar (5 km, 10 km, media maratón).

## MOTIVACIÓN

El análisis del rendimiento deportivo mediante datos es una tendencia creciente en el sector del deporte. La capacidad de predecir el rendimiento futuro basándose en patrones de entrenamiento permite:

* Optimización del entrenamiento: Ajustar cargas, volúmenes e intensidades en función del estado real del corredor.
* Prevención de lesiones: Detectar situaciones de sobreentrenamiento y reducir el riesgo de paradas prolongadas.
* Planificación estratégica: Establecer objetivos alcanzables para competiciones y medir la progresión en ciclos de entrenamiento.
* Personalización: Diseñar planes adaptados a las características fisiológicas y al historial de cada corredor.

De este modo, este trabajo no solo busca aportar una contribución técnica en el ámbito del Machine Learning aplicado al deporte, sino también generar un MVP (Producto Mínimo Viable) que pueda tener aplicación práctica en el sector deportivo, con potencial para ser utilizado por corredores, entrenadores y empresas relacionadas con el running.

# MÉTODOLOGÍA

## ENFOQUE GENERAL

La metodología adoptada en este trabajo sigue un enfoque clásico de ciencia de datos, estructurado en distintas fases que permiten garantizar la calidad del proceso y la reproducibilidad de los resultados. Las fases principales son las siguientes:

1. Recolección de datos:

* Datos objetivos: se extrajeron registros de entrenamiento a través de la plataforma de entrenamiento Strava (exportación de archivos CSV).
* Datos contextuales: se diseñó un cuestionario en Google Forms para recopilar información personal y hábitos de entrenamiento de los participantes (edad, sexo, experiencia deportiva, historial, objetivo d entrenamiento, frecuencia de entrenamiento, etc.).

1. Procesamiento y preparación:

Se aplicaron procesos de limpieza y unificación de los diferentes orígenes de datos. Esto incluyó el tratamiento de valores ausentes, la homogeneización de unidades (p. ej., tiempo en segundos, distancia en kilómetros) y la creación de un dataset maestro con la información combinada.

1. Ingeniería de características (Feature Engineering):

Se desarrollaron variables derivadas con valor predictivo, tales como kilometraje semanal, número de sesiones por semana, ratio de carga aguda-crónica (ACR), métricas de eficiencia (ritmo ajustado a frecuencia cardiaca), y estadísticas de entrenamiento en ventanas móviles de 7 y 28 días.

1. Análisis exploratorio de datos (EDA):

Se realizaron análisis descriptivos y visualizaciones para identificar patrones de entrenamiento, distribuciones de rendimiento y posibles relaciones entre variables. Esto permitió formular hipótesis iniciales y detectar outliers.

1. Modelado predictivo (fase posterior):

Se aplicarán algoritmos de Machine Learning supervisado con el fin de predecir el rendimiento futuro de los corredores en distancias estándar (5 km, 10 km y media maratón). Se compararán modelos base (regresiones lineales, Random Forest) con modelos más avanzados (Gradient Boosting, XGBoost).

1. Validación y despliegue:

Los modelos se evaluarán mediante métricas de error (MAE, RMSE) y técnicas de validación cruzada, garantizando la robustez de los resultados. Finalmente, se implementará un MVP en Streamlit, permitiendo a los usuarios subir sus datos y obtener predicciones personalizadas.

## HERRAMIENTAS UTILIZADAS

Para el desarrollo del proyecto se emplearon herramientas ampliamente utilizadas en el ámbito de la ciencia de datos:

* Python: Lenguaje principal de análisis y modelado.
* Pandas: Manipulación y limpieza de datos tabulares.
* NumPy: Cálculo numérico eficiente.
* Matplotlib / Seaborn: Generación de gráficos y visualizaciones.
* Scikit-learn: Implementación de modelos de Machine Learning y validación.
* Google Forms: Recolección estructurada de datos personales y de entrenamiento.
* Strava / CSV Export: Obtención de registros de actividades deportivas.
* Streamlit: Creación del prototipo interactivo (MVP) para la visualización de resultados y predicciones.

# RECOLECCIÓN DE DATOS

## PARTICIPANTES

La muestra del estudio estuvo compuesta por 10 corredores voluntarios (6 hombres y 4 mujeres), seleccionados bajo los siguientes criterios de inclusión:

* Uso activo de Strava, garantizando un historial suficiente de actividades para el análisis.
* Entrenamiento regular, permitiendo observar patrones de consistencia en la práctica deportiva.
* Consentimiento informado otorgado por cada participante, autorizando el uso de sus datos de entrenamiento y de las respuestas al cuestionario para fines académicos.

Este enfoque se justifica en la necesidad de contar con datos longitudinales y reales de corredores aficionados, que sirvan como base para el desarrollo del sistema predictivo. Si bien la muestra es reducida, resulta adecuada para la construcción de un MVP (Producto Mínimo Viable) y para evaluar la viabilidad del enfoque metodológico.

## FUENTES DE DATOS

La recolección de datos se realizó a través de dos canales principales:

Google Forms - Información Personal:

Se diseñó un cuestionario digital con el fin de recopilar variables contextuales y personales de los corredores, que complementaran los registros objetivos de entrenamiento. Entre las variables recolectadas se incluyen:

* Datos demográficos: nombre, edad, sexo, peso, altura.
* Experiencia deportiva: nivel declarado por el corredor (principiante, intermedio o avanzado).
* Hábitos de entrenamiento: número de días de entrenamiento por semana y principales objetivos (rendimiento, salud, competición, etc.).

<https://docs.google.com/forms/d/12zsb6vO0CjlnBCoAv_OUS61hJFOicLKgZfXEHzDaiQI/edit?usp=forms_home&ouid=111142689510274097539&ths=true>

Esta información resulta fundamental para personalizar el análisis y enriquecer los modelos predictivos, ya que introduce factores individuales que influyen en el rendimiento.

Strava - Datos de Actividades:

Los registros de entrenamiento se obtuvieron mediante la exportación de datos en formato CSV desde la plataforma Strava, algunas de las variables recolectadas incluyen:

* Temporales: fecha y duración de la actividad.
* Distancia: kilómetros recorridos.
* Rendimiento: ritmo (pace), velocidad promedio.
* Fisiológicas: frecuencia cardíaca promedio y máxima (cuando disponible).
* Topográficas: desnivel positivo acumulado.

Estos datos proporcionan una visión cuantitativa y objetiva del rendimiento deportivo, permitiendo extraer métricas clave como la carga de entrenamiento, la eficiencia y la evolución del ritmo en distintas distancia

## ESTRUCTURA DE DATOS RAW

La organización de los datos en el proyecto siguió una estructura clara y reproducible, siguiendo buenas prácticas. La carpeta raíz data/ contiene las siguientes subcarpetas:

Texto

El contenido generado por IA puede ser incorrecto.

* forms/: incluye las respuestas recogidas mediante Google Forms.
* raw/: contiene los archivos CSV exportados de Strava para cada corredor, en su formato original sin modificaciones.
* interim/: almacena los datasets resultantes de pasos intermedios de procesamiento, previos a la creación del dataset maestro.

Este esquema asegura la trazabilidad de los datos, de modo que cualquier transformación realizada durante la limpieza o el feature engineering pueda ser reproducida.

# PROCESAMIENTO Y LIMPIEZA DE DATOS

La calidad de los datos es un aspecto crítico en cualquier proyecto de Data Science. Por ello, antes de realizar análisis exploratorio o aplicar técnicas de Machine Learning, fue necesario realizar una fase de limpieza, estandarización y unión de los distintos orígenes de información. Esta fase se documentó de manera reproducible en notebooks específicos.

## LIMPIEZA DEL FORMULARIO

Archivo: notebooks/formulario.ipynb

El cuestionario en Google Forms fue exportado a CSV (respuestas\_forms.csv). Posteriormente, se aplicaron los siguientes pasos de limpieza y estandarización:

1. Carga de datos: Importación del archivo CSV con las respuestas de los participantes.
2. Renombrado de columnas:

Texto

El contenido generado por IA puede ser incorrecto.

1. Selección de variables relevantes: Se mantuvieron únicamente 8 columnas clave, correspondientes a datos demográficos, antropométricos y de hábitos de entrenamiento.
2. Exportación: El resultado fue un dataset limpio y estructurado, guardado como formulario\_limpio.csv.

Resultado:

Un dataset tabular consistente, con una fila por corredor y sus atributos personales, que servirá como base para enriquecer los datos objetivos procedentes de Strava.

## LIMPIEZA DE DATOS DE STRAVA

Archivos individuales: usuario1.csv, usuario2.csv, etc.

### Problemas identificados en los datos raw de Strava.

Durante la exploración inicial se detectaron las siguientes incidencias:

* Nombres de columnas inconsistentes: Espacios y mayúsculas mezcladas
* Formatos de fecha diversos: Diferentes formatos de fecha entre usuarios
* Actividades no relevantes: Deportes distintos al running incluidos
* Unidades mixtas: Distancias en metros vs kilómetros según el usuario
* Datos de años anteriores: Actividades fuera del período de análisis (2024)
* Formato de pace inconsistente: Necesidad de cálculo uniforme del ritmo
* Formato de tiempo poco manejable: Tiempo transcurrido en segundos.

### Proceso de limpieza aplicado:

Para resolver los problemas detectados, se aplicaron las siguientes transformaciones:

1. Estandarización de nombres de columnas:

Texto

El contenido generado por IA puede ser incorrecto.

1. Filtrado temporal y por tipo de actividad:



1. Estandarización de unidades de distancia:

Texto

El contenido generado por IA puede ser incorrecto.

1. Cálculo uniforme del ritmo (pace):

Captura de pantalla con letras y números

El contenido generado por IA puede ser incorrecto.

1. Formato legible del tiempo transcurrido:

Texto

El contenido generado por IA puede ser incorrecto.

1. Selección de columnas relevantes:

Texto

El contenido generado por IA puede ser incorrecto.

1. Adición de identificador de usuario:



### Aplicación a todos los usuarios

* Proceso replicado: La función limpiar\_strava() se aplicó a los 10 archivos CSV individuales
* Consistencia garantizada: Mismo proceso de limpieza para todos los usuarios
* Exportación individual: Cada archivo limpio guardado en csv\_clean/

### Resultado

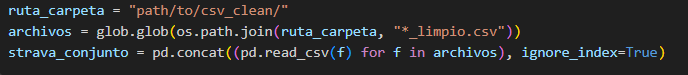
Tras el proceso de limpieza, se obtuvo un conjunto de archivos CSV individuales, limpios, homogéneos y estandarizados, listos para ser utilizados en la fase posterior de unión con la información proveniente del cuestionario de Google Forms. Estos archivos fueron almacenados en la carpeta data/interim/, garantizando la trazabilidad y reproducibilidad del proceso.

## UNIÓN DE DATASETS

Archivo: union\_csv.ipynb

En este paso se unificaron las dos fuentes de información disponibles: los archivos de actividades de Strava y el formulario de Google Forms. El procedimiento seguido fue:

1. Concatenación automática de archivos CSV limpios:



1. Preparación del formulario para el merge:



1. Unión mediante left join:



1. Exportación final: strava\_final.csv en carpeta processed

## MÉTRICAS DE CALIDAD DEL PROCESO

Transformaciones Aplicadas:

* Actividades filtradas: Solo running del año 2024
* Columnas estandarizadas: 11 variables consistentes por usuario
* Pace calculado: Formato uniforme MM.SS para análisis
* Tiempo formateado: HH:MM:SS legible
* Usuarios identificados: Columna 'user' añadida a cada registro

Validaciones Implementadas:

* Fechas válidas: Conversión con errors='coerce'
* Distancias positivas: Eliminación implícita de valores <= 0 en cálculo de pace
* Consistencia de unidades: Detección automática metros/kilómetros
* Formato de ritmo: Manejo de valores nulos y negativos

Resultado Final:

* Dataset unificado: Un solo archivo con todos los usuarios
* Consistencia total: Formato estándar para todas las variables
* Datos listos: Preparados para Feature Engineering
* Trazabilidad: Identificación clara de cada usuario y actividad

# FEATURE ENGINEERING

Archivo: feature\_engineering.ipynb

## 5.1. OBJETIVO

El objetivo de esta fase fue la creación de un conjunto de variables derivadas (features) que capturen las principales dinámicas del rendimiento de los corredores, permitiendo representar cuantitativamente los siguientes aspectos:

* Carga de entrenamiento reciente: volumen y frecuencia de las actividades en ventanas temporales.
* Consistencia en el rendimiento: estabilidad del ritmo medio y de la frecuencia cardíaca.
* Capacidad máxima del corredor: identificación de los mejores tiempos registrados en distancias estándar.
* Patrones temporales de actividad: evolución del rendimiento a lo largo del tiempo.

Estas features constituyen la base de entrada para los modelos predictivos de rendimiento.

## 5.2. FEATURES TEMPORALES CREADAS

Para capturar la dinámica temporal del entrenamiento, se aplicaron ventanas deslizantes (rolling windows) de 7 y 28 días sobre cada usuario, obteniendo métricas agregadas de distancia, ritmo y esfuerzo.

Ventana de 7 días

* weekly\_km: distancia total recorrida en los últimos 7 días (km).
* sessions\_last\_7d: número de sesiones realizadas durante la última semana.
* elevation\_weekly: desnivel positivo acumulado en los últimos 7 días (m).

Estas variables describen la carga de entrenamiento semanal, reflejando el volumen y la intensidad reciente de la actividad.

5.2.2 Ventana de 28 días (4 semanas)

* mean\_pace\_4w: ritmo promedio (min/km) durante las últimas 4 semanas.
* std\_pace\_4w: desviación estándar del ritmo, indicador de consistencia del rendimiento.
* mean\_hr\_4w: frecuencia cardíaca promedio en el último mes.

Estas variables permiten representar tendencias de medio plazo y la estabilidad del entrenamiento, así como el esfuerzo fisiológico sostenido del corredor.

## 5.3. MÉTRICAS DE RENDIMIENTO

Además de las variables temporales, se crearon features de rendimiento absoluto y eficiencia.

Mejores Ritmos (Best Paces)

* best\_5k\_pace: mejor ritmo registrado en distancias comprendidas entre 4.8 y 5.2 km.
* best\_10k\_pace: mejor ritmo registrado en distancias entre 9.5 y 10.5 km.

Estas variables representan el nivel máximo de rendimiento alcanzado por cada corredor en distancias estándar de competición, sirviendo como indicadores de potencial atlético.

Eficiencia y Dinámica Competitiva

* efficiency\_kmh: relación entre la distancia total recorrida y el tiempo total de entrenamiento, expresada en km/h. Esta métrica sintetiza la eficiencia global del corredor.
* days\_since\_last\_race: número de días transcurridos desde la última carrera o evento registrado. Permite modelar el efecto del descanso y la recuperación sobre el rendimiento futuro.

## IMPLEMENTACIÓN TÉCNICA

Pantalla de computadora con letras

El contenido generado por IA puede ser incorrecto.

# ANALISIS EXPLORATORIO DE DATOS (EDA)

Se implementó una función general denominada EDA\_visual(), diseñada para realizar un análisis exploratorio completo e interactivo del dataset. Esta función incluía los siguientes componentes:

Análisis de Distribuciones

* Histogramas: para observar la distribución de cada variable y su dispersión.
* Boxplots: detección visual de posibles valores atípicos (outliers).
* KDE plots: estimación de densidad para variables continuas.

Análisis de Correlaciones

* Matriz de correlación (Pearson): análisis de relaciones lineales entre variables cuantitativas.
* Heatmap: visualización de las correlaciones más significativas.

Análisis por Usuario

* Boxplots por usuario: identificación de variabilidad individual en métricas de rendimiento.
* Comparación de rendimientos: análisis de heterogeneidad dentro de la muestra.

## DETECCIÓN DE OUTLIERS

Durante el EDA se identificaron posibles valores atípicos en varias variables, utilizando boxplots y reglas estadísticas (rango intercuartílico).

|  |  |
| --- | --- |
| **Variable** | **Nº de Outliers Detectados** |
| weekly\_km | 5 |
| sessions\_last\_7d | 2 |
| elevation\_weekly | 35 |
| mean\_pace\_4w | 41 |
| std\_pace\_4w | 31 |
| best\_5k\_pace | 122 |
| mean\_hr\_4w | 0 |
| best\_10k\_pace | 0 |
| efficiency\_kmh | 0 |

### TRATAMIENTO DE OUTLIERS

Criterios aplicados:

Texto

El contenido generado por IA puede ser incorrecto.

Justificación:

mean\_pace\_4w > 8.0 min/km: Ritmos poco realistas para corredores regulares

sessions\_last\_7d > 7: Más de una sesión diaria (posibles errores)

best\_5k\_pace > 7.0 min/km: Tiempos inconsistentes con perfil de corredores