Aplicación de Algoritmos de Aprendizaje Automático para el Análisis de Rendimiento Deportivo

Sergio Cortés Cepeda
matrícula:1731225
Universidad Autónoma de Nuevo León,
Facultad de Ciencias Físico Matemáticas

Noviembre 2025

Resumen

El presente artículo analiza un conjunto de datos deportivos obtenidos mediante un reloj Garmin, con el objetivo de predecir y comprender el comportamiento de variables fisiológicas y de rendimiento. Para ello se aplicaron algoritmos de aprendizaje supervisado **(Regresión Lineal y Random Forest)**orientados a la predicción de calorías y ritmo promedio, así como algoritmos no supervisados **(DBSCAN y Mean Shift)** destinados a identificar patrones y posibles grupos dentro de las sesiones de entrenamiento.

Este trabajo demuestra que la combinación de métodos supervisados y no supervisados proporciona una visión integral del rendimiento deportivo, facilitando la interpretación del comportamiento fisiológico y biomecánico durante sesiones de carrera.

1. Introducción

El análisis de datos deportivos ha cobrado relevancia en los últimos años debido al creciente uso de dispositivos electrónicos capaces de registrar información detallada del rendimiento humano.

El presente estudio utiliza registros reales obtenidos mediante un dispositivo Garmin, incluyendo variables como distancia recorrida, frecuencia cardiaca, pasos, cadencia y zancada. El objetivo principal es analizar la relación entre estas variables e identificar patrones mediante algoritmos de aprendizaje automático.

Para lograrlo, se aplicaron dos enfoques complementarios:

1. **Modelos supervisados** para predecir variables clave como calorías quemadas y ritmo promedio.
2. **Modelos no supervisados** para explorar la estructura interna del conjunto de datos y detectar posibles agrupamientos de sesiones de entrenamiento.

Además, se emplearon métodos de selección de características para identificar las variables con mayor influencia en cada objetivo. Este proceso permitió filtrar información relevante y optimizar el desempeño de los modelos.

El análisis realizado proporciona información valiosa sobre el comportamiento fisiológico durante las sesiones de carrera, permitiendo no solo estimar variables objetivo con precisión, sino también identificar patrones que pueden ser utilizados para mejorar la planificación del entrenamiento y comprender mejor el esfuerzo realizado.

2. Metodología

2.1. Conjunto de datos y variables

El conjunto de datos utilizado corresponde a sesiones de entrenamiento registradas con un reloj Garmin. Cada fila representa una actividad de carrera e incluye información de distancia, respuesta fisiológica y características biomecánicas.

A partir del análisis realizado en tareas previas (correlación, análisis de componentes principales y selección de variables), se consideraron como variables explicativas las siguientes:

- `distancia_km`: distancia recorrida en kilómetros.
- `pasos`: número total de pasos durante la sesión.
- `fc_media`: frecuencia cardiaca media.
- `fc_max`: frecuencia cardiaca máxima.

- `cadencia_media`: cadencia media de carrera.
- `zancada_m`: longitud media de zancada.
- `cadencia_max`: cadencia máxima (utilizada en algunas pruebas exploratorias).

Se trabajó con dos variables objetivo:

- `calorias`: calorías totales quemadas en la sesión.
- `ritmo_medio`: ritmo promedio de carrera.

Previo al modelado se realizó estandarización de las variables numéricas mediante

$$z = \frac{x - \mu}{\sigma},$$

a fin de evitar que las diferencias de escala afectaran el desempeño de los algoritmos, en especial en los modelos no supervisados.

2.2. Selección de características

Para identificar las variables más relevantes en la predicción de cada objetivo se aplicaron tres técnicas de selección de características:

1. **SelectKBest con prueba F** (*f_regression*), que evalúa la relación lineal entre cada predictor y la variable objetivo.
2. **SelectKBest con Información Mutua** (*mutual_info_regression*), que captura relaciones potencialmente no lineales.
3. **RFE (Recursive Feature Elimination)** con regresión lineal como estimador base, que elimina recursivamente las variables menos relevantes.

Los resultados fueron consistentes y permitieron definir, para cada objetivo, el subconjunto de variables más importante:

- **Para calorías**: `distancia_km`, `pasos`, `fc_media` y `cadencia_media`.
- **Para ritmo_medio**: `fc_max`, `cadencia_media`, `fc_media` y `zancada_m`.

2.3. Selección de variables

Antes de ajustar los modelos supervisados, se realizó un proceso de selección de características con el objetivo de identificar los predictores más relevantes para cada variable objetivo.

Para ello se aplicaron tres métodos complementarios: **F-score**, **Información Mutua (MI-score)** y **Eliminación Recursiva de Características (RFE)**. Cada uno de estos enfoques evalúa la relevancia de los predictores desde una perspectiva distinta, lo que permite obtener una selección más robusta y consistente.

2.4. F-score

El F-score mide la intensidad de la relación lineal entre cada variable predictora y la variable objetivo. Formalmente, evalúa cuánto varía la media del objetivo al cambiar la variable de interés. Valores altos de F-score indican que el predictor tiene una influencia significativa dentro de un modelo lineal.

2.5. Información Mutua (MI-score)

La Información Mutua cuantifica la dependencia entre dos variables, considerando tanto relaciones lineales como no lineales. MI-score detecta patrones más complejos. Un valor alto de MI-score implica que el predictor comparte una gran cantidad de información con la variable objetivo y, por tanto, es útil para modelos que capturan interacciones no lineales.

2.6. Eliminación Recursiva de Características (RFE)

El método RFE selecciona variables mediante un proceso iterativo que entrena un modelo repetidamente y elimina los predictores menos relevantes en cada iteración. El resultado final es un conjunto óptimo de características clasificadas por orden de importancia.

2.7. Modelos supervisados

Para cada variable objetivo se entrenaron dos modelos de regresión:

- **Regresión Lineal**: Busca modelar la relación entre una variable dependiente y (por ejemplo, calorías o ritmo promedio) y un conjunto de variables predictoras $x_1, x_2, \ldots, x_p$. El modelo general se expresa como:

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_p x_p + \varepsilon$$

  En forma matricial:

$$y = X\beta + \varepsilon$$

donde:

- $X \in \mathbb{R}^{n \times p}$: matriz de variables predictoras,
- $\beta \in \mathbb{R}^p$: vector de coeficientes,
- $y \in \mathbb{R}^n$: vector de valores observados,
- $\varepsilon \in \mathbb{R}^n$: términos de error.

- **Regresión de Bosques Aleatorios (Random Forest)**:Consiste en un ensamble de árboles de decisión, donde cada árbol $T_b(x)$ es entrenado utilizando una muestra bootstrap del conjunto de datos original y un subconjunto aleatorio de características.

  La predicción de un único árbol para una observación x se define como:

$$T_b(x) = \frac{1}{N_b} \sum_{i \in R_b(x)} y_i,$$

  donde $R_b(x)$ es la región terminal del árbol donde cae la observación x, y N_b es el número de muestras dentro de dicha región.

  La predicción final del Bosque Aleatorio se obtiene promediando las salidas de los B árboles:

$$\hat{y} = \frac{1}{B} \sum_{b=1}^{B} T_b(x).$$

  A continuación se describen detalladamente los elementos que componen esta expresión:

  - **Predicción del bosque** ($\hat{\boldsymbol{y}}$): es el valor predicho por todo el conjunto de árboles para una observación x. En este proyecto corresponde, por ejemplo, a las calorías estimadas o al ritmo promedio predicho a partir de las variables registradas durante una sesión de carrera.
  - **Vector de características** ($\boldsymbol{x}$): representa el conjunto de variables predictoras utilizadas para una observación. En este estudio, este vector puede escribirse como:

$$x = (\text{distancia\_km},\ \text{fc\_media},\ \text{fc\_max},\ \text{cadencia\_media},\ \text{zancada\_m},\ \text{pasos},\ \ldots)$$

  - **Número de árboles** ($\boldsymbol{B}$): indica cuántos árboles de decisión conforman el bosque. Este valor corresponde al parámetro `n_estimators` del modelo Random Forest. Por ejemplo,

en mi caso use `n_estimators = 200`, entonces $B = 200$. Un valor mayor de B suele producir predicciones más estables.

- **Índice del árbol** ($\boldsymbol{b}$): es un contador que recorre cada árbol dentro del bosque:

$$b = 1, 2, 3, \ldots, B$$

Matemáticamente permite expresar la idea de que cada árbol realiza su propia predicción sobre la observación x.

- **Predicción del árbol individual** ($\boldsymbol{T_b(x)}$): T_b representa el árbol número b. La expresión $T_b(x)$ es la predicción que realiza dicho árbol para la observación x. En un problema de regresión, esta predicción es un número real, por ejemplo:
  - 315,2 calorías predichas,
  - 485 segundos de ritmo promedio.

En conjunto, la fórmula indica que la predicción del Random Forest es el **promedio de las predicciones individuales de todos los árboles**, lo cual proporciona robustez, reduce la varianza del modelo y mejora su capacidad de generalización.

El conjunto de datos se dividió en entrenamiento y prueba utilizando una partición del 80 % para entrenamiento y 20 % para prueba.

La calidad de los modelos se evaluó mediante las métricas:

- **MAE (Mean Absolute Error)**: mide el error absoluto promedio entre los valores observados y los valores predichos. Indica cuánto se equivoca el modelo en promedio. Valores menores implican un mejor desempeño.

$$\text{MAE} = \frac{1}{n}\sum_{i=1}^{n} |y_i - \hat{y}_i| \,,$$

- **MSE (Mean Squared Error)**: promedio de los errores al cuadrado. Penaliza fuertemente los errores grandes, por lo que detecta modelos con predicciones muy alejadas del valor real.

$$\text{MSE} = \frac{1}{n}\sum_{i=1}^{n} (y_i - \hat{y}_i)^2,$$

- **RMSE (Root Mean Squared Error)**: raíz cuadrada del MSE. Expresa el error en las mismas unidades que la variable objetivo y representa la desviación promedio de las predicciones respecto a los valores reales.

$$\text{RMSE} = \sqrt{\text{MSE}},$$

- R^2 **(Coeficiente de determinación)**: mide la proporción de la variabilidad de los datos explicada por el modelo. Valores cercanos a 1 indican un buen ajuste, mientras que valores cercanos a 0 o negativos indican bajo desempeño.

$$R^2 = 1 - \frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{\sum_{i=1}^{n}(y_i - \bar{y})^2}.$$

2.8. Reducción de Dimensionalidad mediante Componentes Principales (PCA)

El Análisis de Componentes Principales (PCA) se aplicó con el propósito de proyectar las variables fisiológicas y biomecánicas en un subespacio de menor dimensión, preservando la mayor varianza posible del sistema. Matemáticamente, el procedimiento se desarrolló de la siguiente manera.

Sea $X \in \mathbb{R}^{n \times p}$ la matriz de datos que contiene n sesiones de entrenamiento y p variables fisiológicas estandarizadas. Primero, se centraron los datos en su media:

$$\tilde{X} = X - \mathbf{1}\mu^T,$$

donde μ es el vector de medias muestrales de cada variable y $\mathbf{1}$ es un vector columna de unos.

Posteriormente, se calculó la matriz de covarianza muestral:

$$S = \frac{1}{n-1}\tilde{X}^T\tilde{X}.$$

El PCA consiste en encontrar los vectores propios (autovectores) v_k y valores propios (autovalores) λ_k de la matriz S, tales que:

$$Sv_k = \lambda_k v_k, \quad k = 1, 2, \ldots, p.$$

Los valores propios se ordenan de mayor a menor:

$$\lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_p.$$

Cada autovalor λ_k representa la varianza explicada por el k-ésimo componente principal. La proporción de varianza explicada (PVE) se define como:

$$\text{PVE}_k = \frac{\lambda_k}{\sum_{j=1}^{p}\lambda_j}.$$

El k-ésimo componente principal se obtiene mediante la proyección:

$$Z_k = \tilde{X}v_k,$$

donde Z_k es la nueva variable latente que representa el componente principal. De esta forma, la matriz reducida de componentes principales está dada por:

$$Z = \tilde{X} V_r,$$

donde $V_r = [v_1\ v_2\ v_3]$ es la matriz de los tres autovectores seleccionados.

Finalmente, los componentes principales Z fueron utilizados como entrada para los algoritmos de agrupamiento DBSCAN y Mean Shift, lo cual permitió operar sobre un espacio de menor colinealidad y mayor interpretabilidad.

2.9. Modelos no supervisados

Con el objetivo de identificar patrones y agrupar sesiones de entrenamiento con características similares, se aplicaron dos algoritmos de *clustering* sobre las variables estandarizadas `distancia_km`, `pasos`, `fc_media`, `fc_max`, `cadencia_media` y `zancada_m`:

- **DBSCAN** (*Density-Based Spatial Clustering of Applications with Noise*), con parámetros `eps = 0.7` y `min_samples = 3`.

  El algoritmo *DBSCAN* identifica grupos en los datos a partir de la densidad local de puntos en el espacio de características.

  El criterio fundamental consiste en evaluar cuántos puntos se encuentran dentro de un vecindario de radio ε alrededor de cada observación. Dos parámetros controlan el proceso: el radio máximo ε y el número mínimo de puntos requeridos MinPts para que un punto sea considerado suficientemente denso.

  Formalmente, un punto p es clasificado como **núcleo** si se cumple:

$$|\{q \in D \mid \text{dist}(p, q) \leq \varepsilon\}| \geq \text{MinPts}$$

- **Mean Shift**, con ancho de banda estimado automáticamente mediante `estimate_bandwidth`.

  El método utiliza un kernel de ventana, cuyo ancho de banda h determina el tamaño del vecindario para calcular el desplazamiento del punto hacia regiones de mayor densidad. En este trabajo, el ancho de banda se estimó automáticamente mediante la función `estimate_bandwidth`.

  El vector de desplazamiento (*mean shift*) para un punto x se define como:

$$m(x) = \frac{\sum\limits_{x_i \in N(x)} K\left(\frac{\|x - x_i\|^2}{h^2}\right) x_i}{\sum\limits_{x_i \in N(x)} K\left(\frac{\|x - x_i\|^2}{h^2}\right)} - x$$

donde:

- x es el punto actual.
- $N(x)$ es el vecindario dentro del radio determinado por h.
- $K(\cdot)$ es una función kernel (comúnmente el kernel gaussiano).
- $m(x)$ es el vector que señala la dirección del incremento de densidad.

La actualización iterativa del algoritmo es:

$$x_{t+1} = x_t + m(x_t),$$

DBSCAN permite identificar regiones densas y clasificar puntos aislados como ruido, mientras que Mean Shift localiza modos de densidad sin requerir especificar el número de clusters a priori.

3. Resultados

Cuadro 1: Selección de variables para la variable objetivo *calorías*.

Variable	**F-score**	**p-value**	**MI-score**	**RFE sel.**	**RFE rank**
distancia_km	4451.98	1.20e-46	1.189	True	1
pasos	3393.82	5.05e-44	1.131	True	1
fc_media	23.20	1.68e-05	0.287	True	1
fc_max	15.18	3.21e-04	0.183	False	2
cadencia_media	8.50	5.51e-03	0.143	True	1
zancada_m	2.20	1.44e-01	0.349	False	3
cadencia_max	0.032	8.58e-02	0.076	False	4

Cuadro 2: Selección de variables para la variable objetivo *ritmo_medio*.

Variable	**F-score**	**p-value**	**MI-score**	**RFE sel.**	**RFE rank**
fc_max	64.77	2.93e-10	0.480	True	1
cadencia_media	62.83	4.40e-10	0.376	True	1
fc_media	52.69	4.21e-09	0.586	False	2
zancada_m	35.38	3.77e-07	0.505	False	3
cadencia_max	27.13	3.57e-06	0.396	False	4
distancia_km	20.73	9.54e-05	0.239	False	2
pasos	5.12	2.84e-02	0.072	False	1

El análisis conjunto de estas métricas muestra qué variables aportan mayor información para cada modelo predictivo. Para la variable *calorías*, destacan `distancia_km`, `pasos`, `fc_media` y

`cadencia_media`. En el caso de *ritmo_medio*, las variables más relevantes son `fc_max`, `cadencia_media` y `fc_media`, lo cual coincide con la fisiología del esfuerzo y la mecánica de carrera.

3.1. Modelos supervisados

3.1.1. Predicción de calorías

Cuadro 3: Resultados de los modelos para la variable *calorías*.

Modelo	**MAE**	**MSE**	**RMSE**	R^2
Regresión Lineal	23.83	731.60	27.08	0.9231
Random Forest	**17.59**	**548.36**	**23.42**	**0.9423**

En el Cuadro 3 se observa que Random Forest presenta valores inferiores de MAE, MSE y RMSE, ademas de un coeficiente de determinación R2 ligeramente mayor. Esto indica una mayor capacidad para capturar la variabilidad del gasto energético en comparación con el modelo lineal.

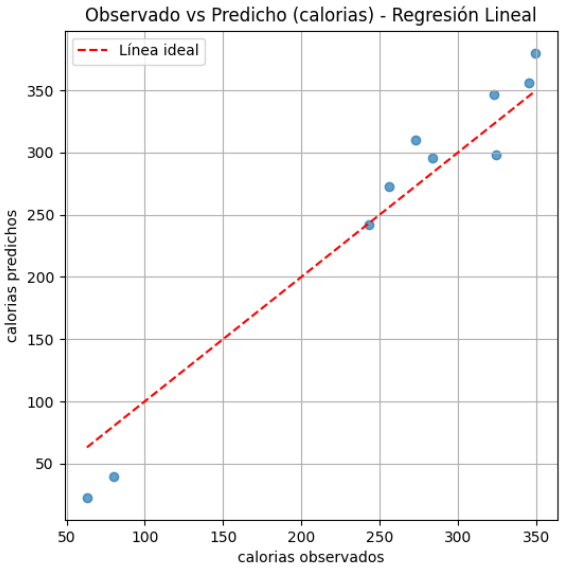

Figura 1: Observado vs Predicho (calorías) - Regresión Lineal.

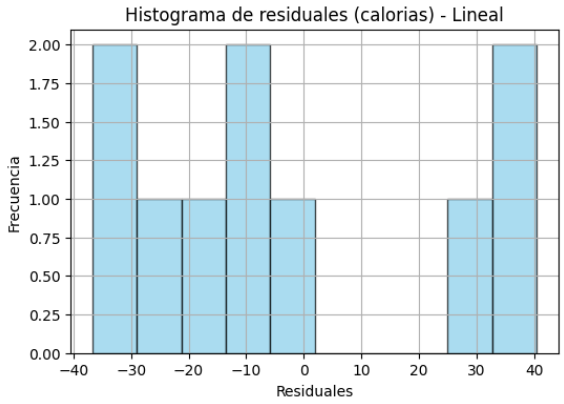

Figura 2: Histograma de residuales (calorias) - Regresión Lineal.

La Figura 1 muestra la relación entre los valores observados y predichos mediante Regresión Lineal. Aunque los puntos se alinean con la recta ideal y = x, se aprecia cierta dispersión alrededor de la linea, indicando que el modelo captura adecuadamente la tendencia general, pero presenta 11 errores moderados. Esta dispersión se confirma en la Figura 2, donde el histograma de residuales exhibe valores tanto positivos como negativos con una distribución relativamente amplia, lo que sugiere que existen patrones no lineales que el modelo no logra representar completamente.

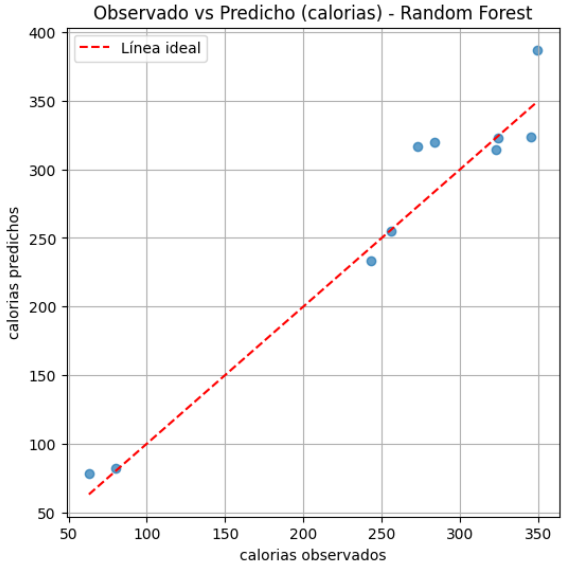

Figura 3: Observado vs Predicho (calorias) - Random Forest.

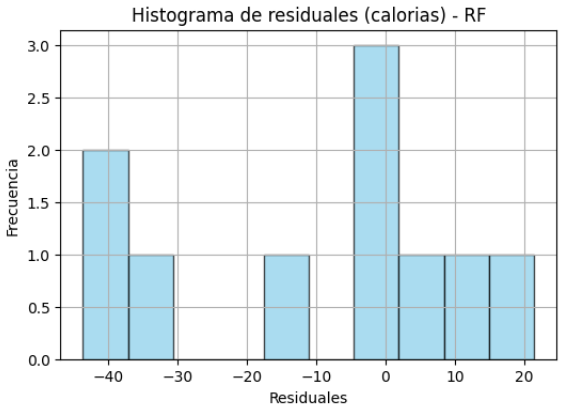

Figura 4: Histograma de residuales (calorias) -Random Forest.

En contraste, la Figura 3 evidencia que el modelo Random Forest produce predicciones más cercanas a la línea ideal, reflejando una mayor precisión en la estimación de las calorías quemadas. Su histograma de residuales (Figura 4) muestra errores más concentrados alrededor de cero y con menor dispersión, lo que indica un comportamiento más estable y robusto frente a variaciones en las variables predictoras.

3.1.2. Comparación de métricas para la variable *calorías*

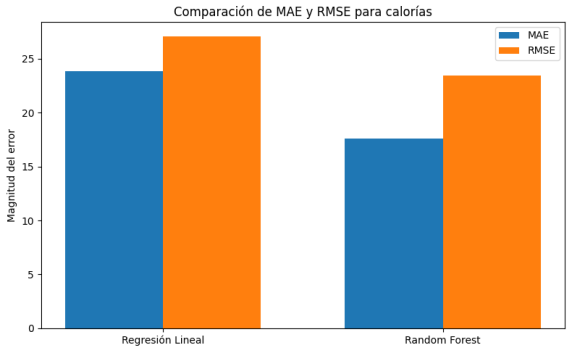

Figura 5: Comparación de los errores MAE y RMSE para la variable calorías entre Regresión Lineal y Random Forest.

Los resultados muestran que el modelo de Random Forest obtiene errores menores en ambas métricas de desviación: el MAE disminuye de aproximadamente 23,83 a 17,59 calorías y el RMSE pasa de 27,08 a 23,42 calorías. Esto indica que, en promedio, las predicciones del Bosque Aleatorio

se alejan menos de los valores observados que las de la regresión lineal, tanto en términos de error absoluto como de error cuadrático.

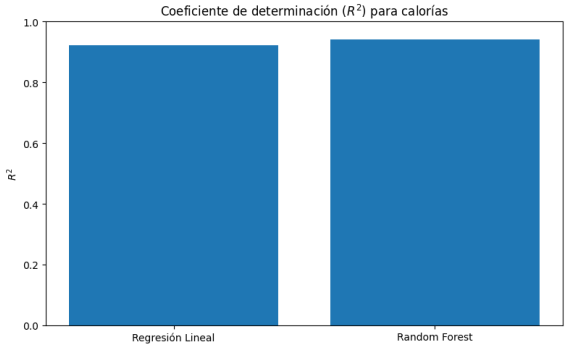

Figura 6: Coeficiente de determinación (R^2) para la variable calorías en los modelos de Regresión Lineal y Random Forest.

Por otro lado, el coeficiente de determinación también favorece ligeramente al Random Forest, con un $R^2 \approx 0{,}942$ frente a $R^2 \approx 0{,}923$ de la Regresión Lineal. Ambos modelos explican una proporción alta de la variabilidad del gasto energético, pero el Bosque Aleatorio captura un poco mejor las relaciones no lineales entre distancia, pasos y variables de frecuencia cardíaca.

En el Cuadro 3 y en las Figuras 5 y 6 se presenta la comparación del desempeño de la Regresión Lineal y el Random Forest para la variable *calorías*. En conjunto, las métricas y las gráficas comparativas confirman que el Random Forest es el modelo más preciso para estimar las calorías quemadas, aunque la Regresión Lineal sigue siendo una alternativa interpretable con un desempeño aceptable.

3.1.3. Predicción de ritmo promedio

A continuación se presentan los resultados obtenidos para la predicción de la variable *ritmo_medio* utilizando los modelos de Regresión Lineal y Random Forest. Las variables utilizadas como predictores fueron: `fc_max`, `cadencia_media`, `fc_media` y `zancada_m`, seleccionadas previamente mediante F-score, Información Mutua y RFE.

Cuadro 4: Resultados de los modelos para la variable *ritmo_medio*.

Modelo	**MAE**	**MSE**	**RMSE**	R^2
Regresión Lineal	21.80	728.16	26.98	0.9866
Random Forest	**21.23**	**599.83**	**24.49**	**0.9889**

Los resultados muestran que tanto la Regresión Lineal como Random Forest ofrecen un ajuste

sobresaliente para la predicción del *ritmo promedio*, con valores de R^2 superiores al 98 %. Sin embargo, se observan diferencias relevantes en la estabilidad y precisión de las predicciones.

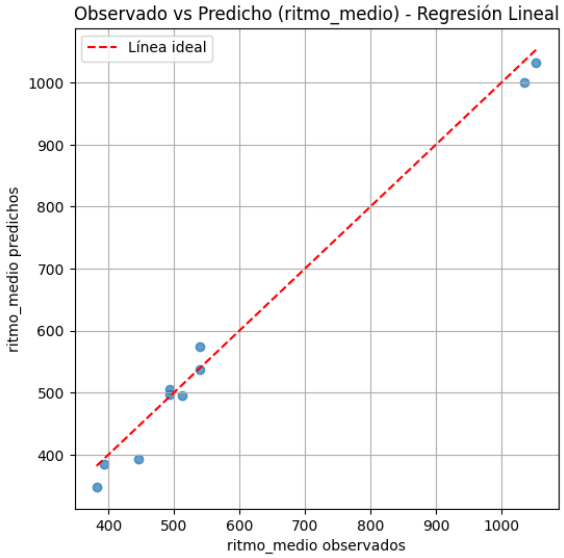

Figura 7: Observado vs Predicho (ritmo_medio) - Regresión Lineal

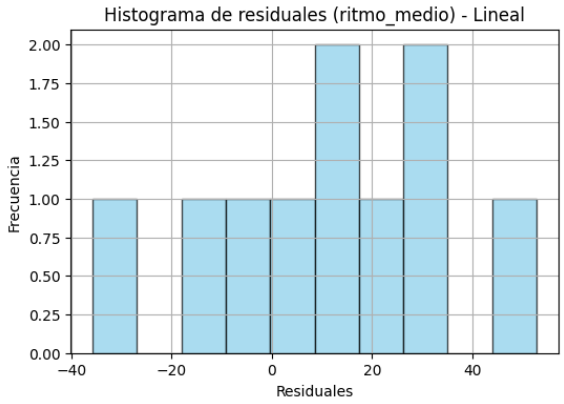

Figura 8: Histograma de residuales (ritmo_medio) - Regresión Lineal.

La (Figura 7) muestra que la Regresión Lineal captura correctamente la tendencia general, pero presenta desviaciones visibles respecto a la línea ideal, especialmente en los ritmos más altos y más bajos. El histograma de residuales en la (Figura 8) confirma esta variabilidad, mostrando una dispersión amplia de errores tanto positivos como negativos, lo que indica que el comportamiento

del ritmo no es estrictamente lineal.

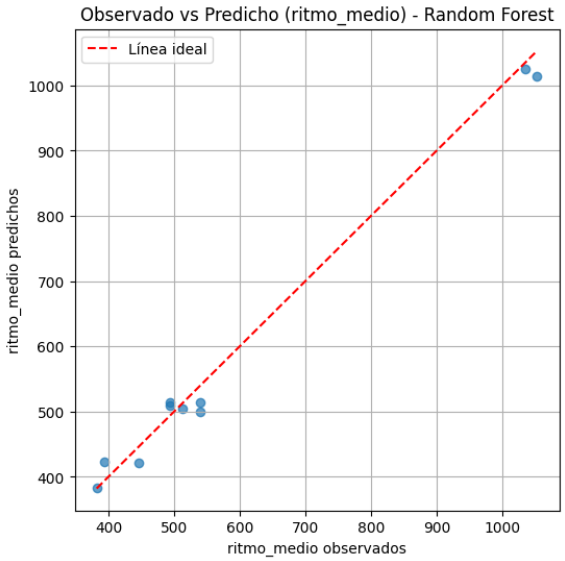

Figura 9: Observado vs Predicho (ritmo_medio) - Random Forest.

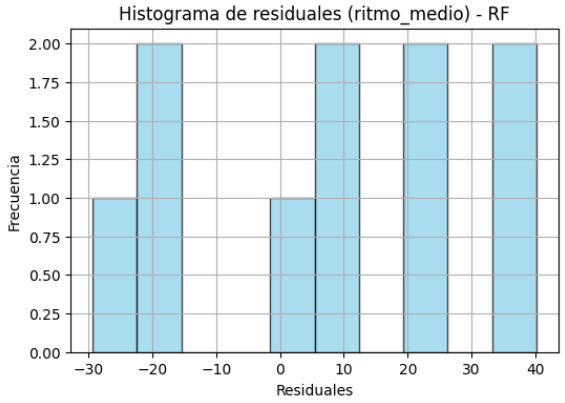

Figura 10: Histograma de residuales (ritmo_medio) - Random Forest.

En contraste, Random Forest muestra un ajuste más preciso (Figura 9), con puntos mejor alineados respecto a la línea ideal y un error global ligeramente menor en todas las métricas analizadas. El histograma de residuales en la (Figura 10) muestra errores distribuidos de forma más concentrada alrededor de cero y menor dispersión, lo que evidencia un modelo más robusto y estable.

3.1.4. Comparación de métricas para la variable *(ritmo_medio)*

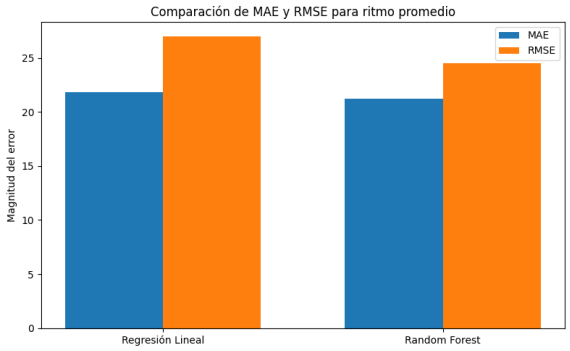

Figura 11: Comparación de los errores MAE y RMSE para la variable ritmo_medio entre Regresión Lineal y Random Forest.

En términos de error, ambos modelos presentan valores muy similares de MAE y RMSE, aunque el Random Forest logra ligeras mejoras: el MAE disminuye de aproximadamente 21,80 a 21,23 segundos por kilómetro y el RMSE se reduce de 26,98 a 24,49 segundos. Estas diferencias, aunque pequeñas, indican que el Bosque Aleatorio tiende a cometer errores algo menores al predecir el ritmo promedio de las sesiones de carrera.

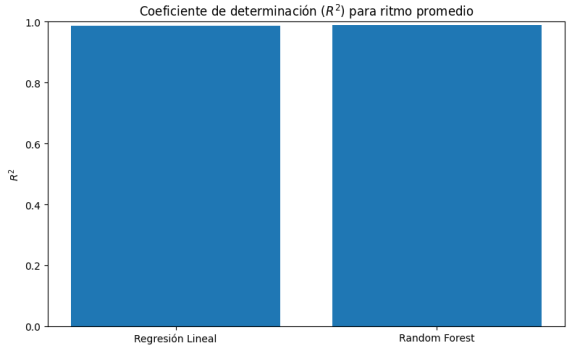

Figura 12: Coeficiente de determinación (R^2) para la variable ritmo_medio en los modelos de Regresión Lineal y Random Forest.

Respecto al coeficiente de determinación, los dos modelos alcanzan valores muy altos ($R^2 > 0{,}98$), lo que significa que explican prácticamente toda la variabilidad observada en el *ritmo_medio*. No obstante, el Random Forest obtiene un R^2 ligeramente superior ($\approx 0{,}9889$ frente a $\approx 0{,}9866$), lo que sugiere una capacidad de ajuste marginalmente mejor.

En el Cuadro 4 y en las Figuras 11 y 12 se resume la comparación del desempeño de la Regresión Lineal y el Random Forest para la variable *ritmo_medio*.

En conjunto, las métricas muestran que tanto la Regresión Lineal como el Random Forest describen de forma muy adecuada el comportamiento del ritmo promedio. Sin embargo, el Bosque Aleatorio ofrece un desempeño algo más sólido, especialmente al capturar posibles relaciones no lineales entre la intensidad (frecuencia cardiaca) y las variables biomecánicas (cadencia y zancada), lo que lo convierte en la opción preferente para la predicción del *ritmo_medio* en este conjunto de datos.

3.1.5. Discusión

Los resultados obtenidos muestran que la combinación de una adecuada selección de variables y el uso de modelos supervisados permitió estimar con precisión dos indicadores clave del rendimiento deportivo: *calorías* y *ritmo_medio*. En ambos casos, el modelo *Random Forest* superó a la Regresión Lineal, lo que evidencia la presencia de relaciones no lineales entre las variables fisiológicas y biomecánicas analizadas.

Para la variable *calorías*, las características más influyentes fueron la distancia recorrida, el número de pasos, la frecuencia cardiaca media y la cadencia media. Estos resultados coinciden con principios fisiológicos del ejercicio, donde el gasto energético depende del volumen total, la intensidad y la eficiencia mecánica del movimiento. El modelo Random Forest obtuvo menores errores y residuales más concentrados, indicando una mayor capacidad para capturar interacciones entre duración, intensidad y técnicas de carrera.

En cuanto al *ritmo_medio*, las variables más relevantes fueron la frecuencia cardiaca máxima, la cadencia media, la frecuencia cardiaca media y la longitud de zancada. Estas variables reflejan la eficiencia biomecánica y el nivel de esfuerzo del atleta. Nuevamente, Random Forest mostró mayor estabilidad en sus predicciones, especialmente en ritmos extremos, donde la Regresión Lineal presentó mayor dispersión.

Los resultados confirman que Random Forest es un modelo más robusto y preciso, presentando menor variabilidad en los errores y capturando de manera más completa la compleja estructura fisiológica del rendimiento deportivo.

3.2. Modelos no supervisados

3.3. Reducción de Dimensionalidad mediante Componentes Principales

Para mejorar la calidad del análisis no supervisado y facilitar la interpretación de los patrones presentes en los datos, se aplicó el método de Análisis de Componentes Principales (PCA).

La aplicación de PCA previo a los métodos de agrupamiento responde a dos objetivos principales: (1) eliminar la colinealidad existente entre las variables fisiológicas registradas por el dispositivo

Garmin, y (2) proyectar los datos en un espacio de menor dimensión donde la estructura de los grupos sea más evidente.

Los componentes principales obtenidos se interpretaron como factores latentes asociados al volumen del entrenamiento, la intensidad fisiológica y la mecánica de carrera.

Cuadro 5: Varianza explicada por los Componentes Principales

Componente	**Varianza Explicada**	**Varianza Acumulada**
PC1	0.500242	0.500242
PC2	0.225917	0.726160
PC3	0.131510	0.857670
PC4	0.059145	0.916815
PC5	0.044103	0.960919
PC6	0.033101	0.994020
PC7	0.002970	0.996991
PC8	0.002223	0.999214
PC9	0.000549	0.999769
PC10	0.000160	0.999929
PC11	0.000056	0.999986
PC12	0.000013	0.999999
PC13	0.0000007	1.000000
PC14	0.00000000043	1.000000

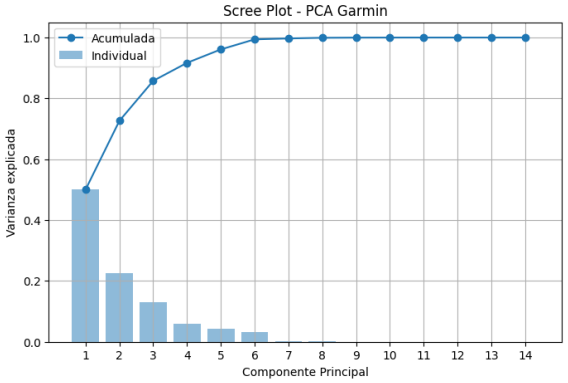

Figura 13: Scree Plot del análisis PCA aplicado a las variables fisiológicas derivadas del registro Garmin. Se observa que los tres primeros componentes explican el 85.7 % de la variabilidad total.

Cuadro 6: Cargas de los Componentes Principales (PC1–PC3)

Variable	**PC1**	**PC2**	**PC3**
distancia_km	0.363980	0.141103	0.017866
calorias	0.368040	0.114553	0.003881
Tiempo_Total	0.364993	0.120483	-0.025415
fc_media	0.264492	-0.230857	-0.010279
fc_max	0.235946	-0.269638	-0.347594
cadencia_media	0.128314	0.064824	0.645188
cadencia_max	-0.153932	0.252635	-0.013898
ritmo_medio	0.012931	-0.387012	-0.476851
zancada_m	-0.172083	0.414034	-0.061307
pasos	0.365983	0.120647	0.030358
num_vueltas	0.351650	0.097947	-0.139987
tiempo_mov	0.367089	0.112998	0.027360
alt_min	-0.098765	0.425898	-0.352429
alt_max	0.034186	0.467884	-0.291571

Cuadro 7: Cargas ordenadas para PC1(volumen)

Variable	**Carga PC1**
calorias	0.368040
tiempo_mov	0.367089
pasos	0.365983
Tiempo_Total	0.364993
distancia_km	0.363980
num_vueltas	0.351650
fc_media	0.264492
fc_max	0.235946
cadencia_media	0.128314
alt_max	0.034186
ritmo_medio	0.012931
alt_min	-0.098765
cadencia_max	-0.153932
zancada_m	-0.172083

Cuadro 8: Cargas ordenadas para PC2(intensidad)

Variable	Carga PC2
alt_max	0.467884
alt_min	0.425898
zancada_m	0.414034
cadencia_max	0.252635
distancia_km	0.141103
pasos	0.120647
Tiempo_Total	0.120483
calorias	0.114553
tiempo_mov	0.112998
num_vueltas	0.097947
cadencia_media	0.064824
fc_media	-0.230857
fc_max	-0.269638
ritmo_medio	-0.387012

Cuadro 9: Cargas ordenadas para PC3(tecnica)

Variable	Carga PC3
cadencia_media	0.645188
pasos	0.030358
tiempo_mov	0.027360
distancia_km	0.017866
calorias	0.003881
fc_media	-0.010279
cadencia_max	-0.013898
Tiempo_Total	-0.025415
zancada_m	-0.061307
num_vueltas	-0.139987
alt_max	-0.291571
fc_max	-0.347594
alt_min	-0.352429
ritmo_medio	-0.476851

3.3.1. Interpretación final del PCA

El primer componente principal (PC1) agrupó variables como calorías, tiempo total, distancia recorrida, número de pasos y tiempo en movimiento, por lo que fue interpretado como un **factor de volumen del entrenamiento** o **carga externa**. Este componente refleja la magnitud global de cada sesión y distingue claramente los entrenamientos más largos y demandantes en términos de duración y gasto energético.

El segundo componente principal (PC2) mostró altas cargas positivas en la altitud mínima,

altitud máxima, longitud de zancada y cadencia máxima, mientras que presentó cargas negativas en la frecuencia cardiaca y el ritmo medio. Este patrón sugiere la presencia de un **factor de intensidad y exigencia impuesta por el terreno**, donde sesiones con mayor variación altimétrica y zancada más amplia tienden a exigir un mayor esfuerzo, el cual se refleja en un ritmo más rápido y en una frecuencia cardíaca más elevada.

Finalmente, el tercer componente principal (PC3) estuvo dominado principalmente por la cadencia media, acompañado de contribuciones moderadas de otras variables biomecánicas. Por ello, este componente se interpretó como un **factor de mecánica o técnica de carrera**, asociado con la eficiencia del paso, la frecuencia del movimiento y la dinámica del corredor durante la sesión.

3.4. Algoritmo DBSCAN

En este trabajo, DBSCAN se aplicó sobre el espacio reducido generado por los tres primeros componentes principales del PCA, con el fin de trabajar en un entorno de menor dimensionalidad y libre de colinealidad. El algoritmo depende de dos hiperparámetros fundamentales: el radio de vecindad ε y el número mínimo de puntos *min_samples* necesarios para considerar una región como densa.

3.4.1. Exploración de parámetros

Con el objetivo de seleccionar una configuración adecuada, se evaluó una rejilla de valores para $\varepsilon \in \{0{,}3,\ 0{,}5,\ 0{,}7,\ 1{,}0,\ 1{,}3\}$ y *min_samples* $\in \{2,\ 3,\ 4\}$. Para cada combinación se registró el número de clústeres detectados (n_{clusters}, sin considerar la etiqueta -1) y el número de observaciones clasificadas como ruido (n_{ruido}). Los resultados se resumen en el Cuadro 11.

Cuadro 10: Exploración de parámetros para DBSCAN sobre los tres primeros componentes principales

ε	*min_samples*	n_{clusters}	n_{ruido}
0.3	2	0	14
0.3	3	0	14
0.3	4	0	14
0.5	2	1	12
0.5	3	0	14
0.5	4	0	14
0.7	2	1	11
0.7	3	1	11
0.7	4	0	14
1.0	2	1	10
1.0	3	1	10
1.0	4	1	10
1.3	2	1	10
1.3	3	1	10
1.3	4	1	10

Como se observa, para valores pequeños de ε (por ejemplo, $\varepsilon = 0{,}3$) el algoritmo considera a todas las observaciones como ruido, es decir, no se identifica ningún clúster denso. A medida que ε aumenta, DBSCAN comienza a formar un único clúster acompañado de varios puntos aislados, pero en ningún caso se logran dos o más grupos bien definidos.

3.4.2. Conclusión sobre DBSCAN

En resumen, la aplicación de DBSCAN sobre los componentes principales del conjunto de datos Garmin no permitió identificar estructuras densas naturales que dieran lugar a múltiples clústeres. La mayoría de las combinaciones de parámetros produjeron únicamente ruido, y aquellas que generaron algún agrupamiento solo formaron un clúster principal acompañado de observaciones aisladas. Estos resultados sugieren que las sesiones de entrenamiento analizadas son relativamente homogéneas en el espacio reducido por PCA y no presentan patrones de alta densidad que puedan ser explotados por este tipo de algoritmo de agrupamiento.

3.5. Algoritmo Mean Shift

El algoritmo *Mean Shift* se utilizó como segundo enfoque de agrupamiento no supervisado, con el objetivo de identificar modos de densidad en el espacio reducido por los componentes principales. Mean Shift no requiere fijar de antemano el número de clústeres y detecta regiones de alta concentración de puntos desplazando iterativamente cada observación hacia las zonas de mayor densidad.

En este estudio, Mean Shift se aplicó sobre los tres primeros componentes principales obtenidos del PCA, los cuales explican el 85.7 % de la variabilidad total. El parámetro de suavizamiento (*bandwidth*) se estimó de manera automática mediante la función *estimate_bandwidth*.

3.5.1. Centros de clúster en el espacio de componentes principales

La Tabla 11 muestra las coordenadas de los centros de cada clúster en el espacio definido por los tres primeros componentes principales, así como el número de observaciones asignadas a cada uno.

Cuadro 11: Centros de clúster obtenidos mediante Mean Shift en el espacio de los tres primeros componentes principales

Cluster	**PC1**	**PC2**	**PC3**	n **observaciones**
0	-0.565	-0.699	0.419	11
1	-2.972	2.927	0.448	2
2	7.516	1.551	0.215	1

3.5.2. Visualización de los clústeres en el espacio reducido

En la Figura 14 se representa la distribución de las observaciones en el plano definido por los dos primeros componentes principales, coloreadas de acuerdo con la etiqueta asignada por Mean Shift. Los marcadores en forma de "X" indican la posición de los centros de cada clúster.

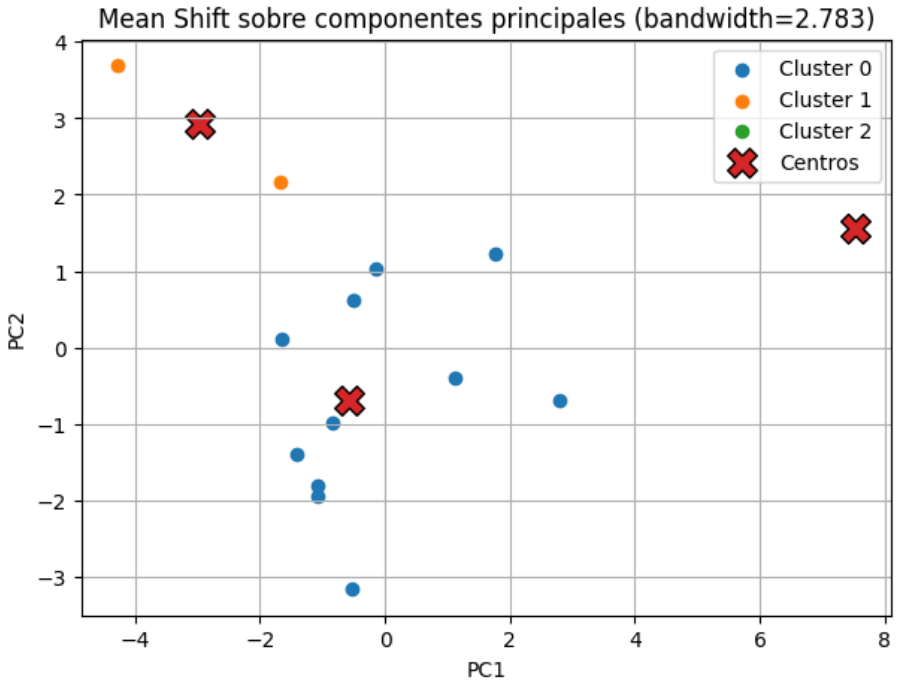

Figura 14: Clústeres identificados por Mean Shift en el plano PC1–PC2. Se observa un clúster dominante (Cluster 0) que agrupa la mayoría de las sesiones, y dos clústeres adicionales asociados a actividades con características más extremas en el espacio de componentes principales.

3.5.3. Clústeres en el espacio de variables originales

Para facilitar la interpretación en términos de rendimiento deportivo, la Figura 15 muestra los mismos clústeres proyectados sobre el espacio de dos variables originales: distancia recorrida (km) y calorías consumidas. Esta visualización permite relacionar cada grupo con patrones concretos de volumen de entrenamiento.

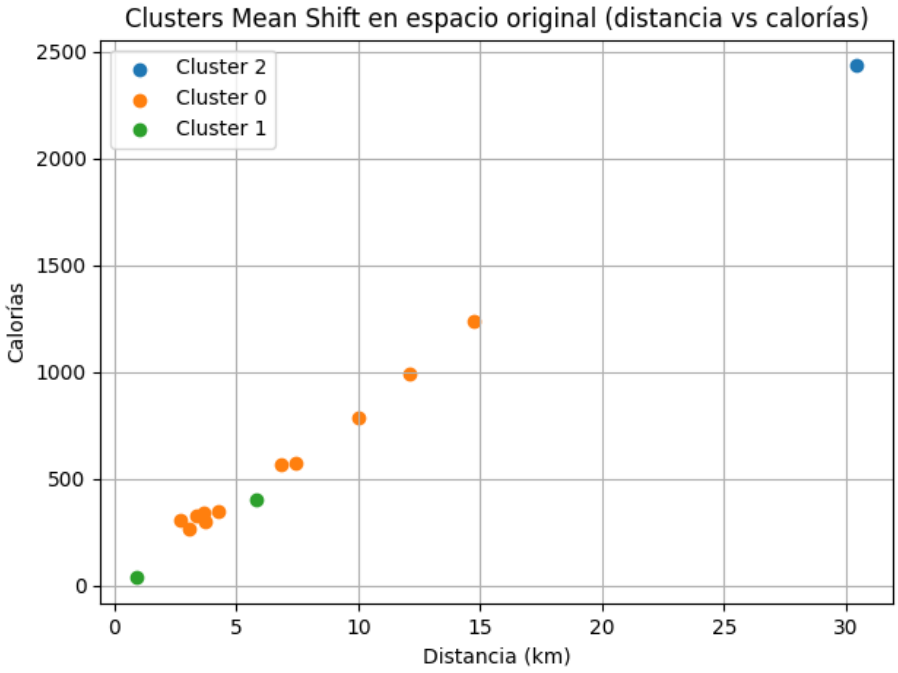

Figura 15: Clústeres de Mean Shift proyectados en el espacio original distancia–calorías. El Cluster 0 concentra la mayoría de las sesiones con distancias y gastos energéticos moderados, mientras que los Clusters 1 y 2 corresponden a entrenamientos atípicos en términos de volumen.

3.5.4. Conclusión sobre Mean Shift

Los resultados de Mean Shift muestran la presencia de tres modos de densidad en el espacio reducido por PCA. El **Cluster 0**, que agrupa 11 de las 14 sesiones analizadas, representa el patrón de entrenamiento más frecuente, caracterizado por volúmenes e intensidades moderadas. El **Cluster 1** reúne 2 sesiones con valores más extremos en los componentes asociados a la intensidad y a la altimetría, lo que sugiere entrenamientos relativamente más exigentes o realizados en recorridos con mayores variaciones de desnivel. Finalmente, el **Cluster 2** corresponde a una única sesión con valores muy elevados en el primer componente principal, lo cual indica un entrenamiento atípico de alto volumen y demanda global.

El uso de Mean Shift sobre los componentes principales permite identificar no sólo la modalidad típica de los entrenamientos del corredor, sino también aquellas sesiones que se comportan como *outliers* en términos de volumen e intensidad.Mientras que DBSCAN no encontró múltiples regiones densas bien definidas, Mean Shift reveló la existencia de un clúster dominante y de un pequeño subconjunto de sesiones con características inusuales.

3.6. Diseño de Experimentos para el Problema de Estudio

El problema de estudio se centra en analizar cómo las características de los entrenamientos de carrera registrados con el dispositivo Garmin influyen en el rendimiento del corredor rumbo a un maratón. Para ello, se propone un diseño de experimentos conceptual que permita organizar las sesiones en función de factores de interés relacionados con el volumen y la intensidad del entrenamiento.

Objetivo del diseño experimental

Identificar combinaciones de volumen e intensidad de las sesiones de carrera que se asocien con cargas de entrenamiento adecuadas para la preparación de un maratón, utilizando como base los datos fisiológicos registrados (distancia, ritmo, frecuencia cardiaca, calorías, cadencia, altimetría, etc.).

Factores y niveles

Se consideraron dos factores principales, definidos a partir de las variables registradas por el reloj Garmin:

- **Factor A: Volumen de la sesión (distancia recorrida).**
  - Nivel A_1: Sesión corta (≤ 5 km)
  - Nivel A_2: Sesión media (5–10 km)
  - Nivel A_3: Sesión larga (> 10 km)
- **Factor B: Intensidad relativa de la sesión (ritmo medio).**
  - Nivel B_1: Ritmo suave ($>$ 6:30 min/km)
  - Nivel B_2: Ritmo moderado (5:30–6:30 min/km)
  - Nivel B_3: Ritmo rápido ($<$ 5:30 min/km)

Este esquema define un diseño factorial 3×3, donde cada combinación de niveles representa un tipo de sesión de entrenamiento potencialmente relevante para la preparación del maratón.

Tratamientos

Cada combinación de volumen e intensidad se considera un tratamiento experimental:

$$T_{ij} = \{\text{Sesión con nivel } A_i \text{ de volumen y nivel } B_j \text{ de intensidad}\},$$

de forma que:

- T_{11}: Sesiones cortas y suaves
- T_{12}: Sesiones cortas y moderadas
- T_{13}: Sesiones cortas y rápidas
- T_{21}: Sesiones medias y suaves
- T_{22}: Sesiones medias y moderadas
- T_{23}: Sesiones medias y rápidas
- T_{31}: Sesiones largas y suaves
- T_{32}: Sesiones largas y moderadas
- T_{33}: Sesiones largas y rápidas

Las sesiones reales registradas en Garmin se asignan a cada tratamiento de acuerdo con su distancia y su ritmo medio, lo que permite analizar cómo se distribuyen los entrenamientos dentro de este espacio de diseño.

Variables de respuesta

Para cada tratamiento se consideran como variables de respuesta:

- Calorías consumidas por sesión.
- Frecuencia cardiaca media y máxima.
- Cadencia media y longitud de zancada.
- Carga externa global (primer componente principal del PCA).
- Clúster asignado por los algoritmos de agrupamiento (DBSCAN y Mean Shift).

Estas variables permiten cuantificar el impacto de cada combinación de volumen e intensidad sobre la carga de entrenamiento y la respuesta fisiológica del corredor.

Hallazgos encontrados

El diseño experimental propuesto no implica la manipulación controlada de los entrenamientos en un laboratorio, sino una *experimentación observacional* basada en los datos reales registrados por el dispositivo. Sin embargo, la estructura factorial 3×3 facilita:

- Identificar qué tipos de sesiones (tratamientos) se presentan con mayor frecuencia en la preparación actual del corredor.
- Relacionar cada tratamiento con la carga externa e intensidad fisiológica resumidas por el PCA.
- Analizar si los clústeres identificados por DBSCAN y Mean Shift corresponden a combinaciones específicas de volumen e intensidad (por ejemplo, sesiones largas y rápidas frente a sesiones cortas y suaves).

Este diseño de experimentos proporciona un marco sistemático para interpretar los patrones de entrenamiento rumbo al maratón y para proponer ajustes en la distribución de sesiones (volumen e intensidad) con el fin de optimizar el rendimiento.

4. Discusión

Los resultados sugieren que mantengo una carga de entrenamiento consistente y predominantemente moderada, con sesiones puntuales que representan esfuerzos significativos. La integración de modelos supervisados y no supervisados permite describir tanto las relaciones predictivas entre variables como los patrones estructurales que emergen dentro del conjunto de sesiones.

5. Conclusión General

El presente estudio integró técnicas supervisadas y no supervisadas para analizar mi rendimiento durante su preparación rumbo a un maratón, utilizando datos reales registrados por un reloj Garmin. Los modelos supervisados permitieron identificar las variables predictoras más relevantes en el gasto energético y en la dinámica del entrenamiento, destacando la distancia, los pasos, la frecuencia cardiaca media y la cadencia. Estos resultados refuerzan la importancia del control del volumen y la intensidad como elementos centrales en la planificación de sesiones orientadas a eventos de larga duración.

Los métodos no supervisados complementaron esta perspectiva al revelar la estructura oculta en los datos. El PCA permitió condensar adecuadamente la información fisiológica en componentes significativos, mientras que los algoritmos de agrupamiento evidenciaron patrones de entrenamiento y la presencia de sesiones atípicas. La identificación de clústeres mediante Mean Shift proporciona información valiosa para monitorear la distribución de cargas y detectar sesiones que representan esfuerzos inusualmente altos, lo cual resulta esencial para evitar sobreentrenamiento y optimizar la progresión del volumen.

El diseño de experimentos observacional propuesto permitió organizar las sesiones dentro de un marco sistemático basado en volumen e intensidad, facilitando la interpretación de los clústeres y la carga fisiológica. Este enfoque ofrece una herramienta práctica para evaluar la consistencia del entrenamiento y su alineación con las demandas de un maratón.

Los resultados muestran que el uso integrado de algoritmos supervisados y no supervisados es una estrategia efectiva para caracterizar el rendimiento, monitorear la evolución del entrenamiento y orientar la toma de decisiones. Este análisis constituye un primer paso hacia la construcción de modelos más complejos que puedan predecir el desempeño en carreras largas y apoyar de manera personalizada mi preparación deportiva.

Agradecimientos

Expreso mis agradecimientos al profesor José Alberto Benavides Vázquez por su disposición y enseñanza fueron fundamentales para la realización de este trabajo.

Asimismo, se agradece al entrenador David Sánchez por su acompañamiento en la preparación deportiva, cuyas sesiones de entrenamiento y retroalimentación práctica resultaron esenciales para comprender y analizar mi rendimiento. Su experiencia y profesionalismo contribuyeron de manera significativa al enfoque aplicado de esta investigación.

Referencias

[1] Runner's World. (2023). *How to fuel your body during marathon training*. Recuperado de `https://www.runnersworld.com/uk/nutrition/diet/a776033/how-to-fuel-your-body-best-during-marathon-training/`

[2] Academia Española de Nutrición y Dietética. (2023). *Dieta para maratón*. Recuperado de `https://www.academianutricionydietetica.org/nutricion-deportiva/dieta-maraton/`

[3] IBM. (2023). *Random Forest: How it works*. Recuperado de `https://www.ibm.com/mx-es/think/topics/random-forest`

[4] DataCamp. (2023). *DBSCAN Clustering Algorithm Tutorial*. Recuperado de `https://www.datacamp.com/es/tutorial/dbscan-clustering-algorithm`

[5] Pythoneers / Medium. (2022). *Fully explained Mean Shift clustering with Python*. Recuperado de `https://medium.com/pythoneers/fully-explained-mean-shift-clustering-with-python-51aef7a17c5d`

[6] Hernández Zamora, R. I. (2025). *Sesión 6: Componentes Principales*. Facultad de Ciencias Físico Matemáticas, Universidad Autónoma de Nuevo León.

[7] Cortes Cepeda, S. (2025). *Dataset personal de sesiones de entrenamiento registradas con Garmin*. Archivo: run_activitiesAA.csv. Datos no publicados.

[8] Brownlee, J. (2020). *Recursive Feature Elimination (RFE) for Feature Selection in Python*. Machine Learning Mastery. Recuperado de `https://machinelearningmastery.com/rfe-feature-selection-in-python/`

[9] Wood, T. (s.f.). *F-score*. DeepAI. Recuperado de `https://deepai.org/machine-learning-glossary-and-terms/f-score`