Resumen

Este trabajo de grado presenta el diseño e implementación de un sistema comparativo basado en la clusterización de atletas de CrossFit. A partir de un dataset público de rendimiento (variables antropométricas, levantamientos máximos y Benchmark WODs), se construyeron tres cohortes analíticas (mixta, masculina y femenina), se caracterizó el mecanismo de datos faltantes y se aplicaron dos estrategias de imputación (one-step clustering imputation y multiple imputation pooled). Sobre los datasets completos se evaluaron dieciocho escenarios experimentales (3 cohortes × 2 imputaciones × 3 algoritmos: K-Means, Agglomerative Ward y GMM), con barridos de $k = 2, \dots, 10$. Como corte productivo se adoptó la cohorte masculina ($n = 9.576$), imputación one-step y K-Means con $k = 6$, obteniendo seis perfiles interpretables de rendimiento. Sobre ese modelo se prototipó un módulo comparativo que asigna atletas a perfiles y facilita la identificación de fortalezas y debilidades relativas. Los resultados muestran que la estructura de agrupamiento es robusta frente a la estrategia de imputación y que K-Means ofrece mejor equilibrio entre separación y balance que los métodos de sensibilidad.

Palabras clave: clusterización, CrossFit, perfiles atléticos, imputación de datos, K-Means, aprendizaje no supervisado.

Summary

This undergraduate thesis presents the design and implementation of a comparative athletic-profile system based on clustering CrossFit athletes. Using a public performance dataset (anthropometric variables, maximal lifts, and Benchmark WODs), three analytical cohorts were built (mixed, male, and female), the missing-data mechanism was characterized, and two imputation strategies were applied (one-step clustering imputation and multiple imputation pooled). On the completed datasets, eighteen experimental scenarios were evaluated (3 cohorts × 2 imputations × 3 algorithms: K-Means, Agglomerative Ward, and GMM), with sweeps of $k = 2, \ldots, 10$. The productive cut adopted was the male cohort ($n = 9{,}576$), one-step imputation, and K-Means with $k = 6$, yielding six interpretable performance profiles. On that model, a comparative module was prototyped to assign athletes to profiles and support the identification of relative strengths and weaknesses. Results show that the clustering structure is robust to the imputation strategy and that K-Means provides a better trade-off between separation and cluster balance than the sensitivity methods.

Keywords: clustering, CrossFit, athletic profiles, data imputation, K-Means, unsupervised learning.

APROXIMACIÓN A UN SISTEMA COMPARATIVO BASADO EN TÉCNICAS DE CLUSTERIZACIÓN PARA ATLETAS DE CROSSFIT

GERMAN DAVID IRIARTE SÓLTAU

CÓDIGO 8953234

PROYECTO DE GRADO PARA OPTAR AL TÍTULO DE INGENIERO(A) DE SISTEMAS Y COMPUTACIÓN

DIRECTOR(A)

PHD. MARÍA CONSTANZA PABÓN

FACULTAD DE INGENIERÍA Y CIENCIAS INGENIERÍA DE SISTEMAS Y COMPUTACIÓN

SANTIAGO DE CALI, JULIO 1 DE 2026

ÍNDICE GENERAL

ÍNDICE DE TABLAS

Anexo A

ÍNDICE DE FIGURAS

1. INTRODUCCIÓN

El CrossFit exige el desarrollo simultáneo de múltiples capacidades físicas y habilidades técnicas. Esta amplitud genera un espectro diverso de destrezas entre atletas: no todos destacan en lo mismo, pero el ideal competitivo apunta al "atleta más completo", capaz de rendir de forma consistente en un conjunto amplio de tareas, aunque no sea el mejor absoluto en cada una.

En este contexto, la pregunta "¿cómo mejorar?" no tiene una única respuesta objetiva, menos aún en un deporte que involucra tantas variables. Una vía práctica —que se adoptó como supuesto de trabajo— consistió en priorizar las debilidades relativas para evitar que estas arrastraran el desempeño general. El reto fue definir el concepto de "debilidad" con un referente justo, ya que compararse con novatos o con la élite difícilmente constituye un buen punto de partida. La noción de fortaleza o debilidad depende del referente; por ello, se propuso comparar a cada atleta con pares realmente comparables, es decir, atletas con perfiles de rendimiento similares.

Este proyecto fundamentó un sistema comparativo entre pares comparables a partir de datos públicos (algunos PRs, benchmarks y ciertas métricas físicas), entendidos como aproximaciones y no como mediciones perfectas. Mediante técnicas de agrupamiento, se identificaron perfiles de rendimiento y, para cada atleta, se determinó su posición relativa dentro del grupo más parecido, destacando fortalezas y brechas prioritarias en las mismas variables que definieron la comparación.

Con los agrupamientos obtenidos, se implementó un prototipo capaz de asignar nuevos atletas al grupo más afín y contrastarlos con pares de características similares. Si bien esta versión inicial no proporcionó resultados altamente precisos debido a las limitaciones inherentes del conjunto de datos utilizado, permitió demostrar el potencial de esta estrategia comparativa. En particular, evidenció cómo, con datos más completos y de mayor calidad, cada deportista podría visualizar sus fortalezas y brechas bajo criterios homogéneos, orientando su entrenamiento

hacia mejoras específicas. En este sentido, el principal resultado del proyecto fue la construcción de una guía preliminar para la priorización del entrenamiento y la obtención de primeros insights sobre las capacidades del atleta, además del diseño de una base metodológica que, con insumos más robustos, podría evolucionar hacia un sistema comparativo significativamente más útil y aplicable en la práctica.

2. DEFINICIÓN DEL PROBLEMA

2.1. Planteamiento del problema

El CrossFit se ha consolidado en la última década como una de las disciplinas de acondicionamiento físico más completas y exigentes, integrando componentes de halterofilia, gimnasia y acondicionamiento metabólico. Esta diversidad de estímulos implica que el rendimiento de los atletas no depende de una sola capacidad, sino de un equilibrio complejo entre fuerza máxima, resistencia aeróbica, potencia anaeróbica, coordinación y habilidades gimnásticas [1].

Dicha complejidad plantea una dificultad recurrente para atletas y entrenadores: determinar qué factores priorizar en la preparación física para mejorar el desempeño general y, en consecuencia, estar más preparados al enfrentar los Workouts of the Day (WODs) o pruebas de CrossFit. En últimas, la pregunta de cómo mejorar es inherentemente compleja por la misma naturaleza del deporte.

Aunque en los últimos años el deporte de alto rendimiento ha experimentado una creciente adopción de tecnologías de análisis de datos para apoyar la toma de decisiones —por ejemplo, integrando métricas biométricas y modelos de machine learning— [2], en la práctica muchas decisiones de programación aún se basan en la experiencia del entrenador o en la percepción subjetiva del atleta. Se trata de una crítica parcial, pues estas estrategias continúan vigentes debido a la complejidad inherente del deporte y a la individualidad de cada atleta. Medidas como la escala de esfuerzo percibido, los diarios de carga o la observación cualitativa siguen siendo componentes esenciales del monitoreo habitual, como señalan estudios recientes sobre estrategias integradas de seguimiento deportivo [3]. No obstante, su persistencia y relevancia también evidencian que aún existe un amplio margen para desarrollar modelos más objetivos apoyados en tecnología, orientados a complementar y potenciar la interpretación de dichas métricas subjetivas dentro del

ámbito deportivo.

Así pues, el problema esbozado hasta ahora radica en la necesidad de desplazar progresivamente la toma de decisiones hacia modelos más objetivos, capaces de contribuir efectivamente al progreso de un atleta, considerando la gran cantidad de variables que deben evaluarse y atenderse. Una posible respuesta a esta dificultad consiste en adoptar una estrategia basada en la identificación de fortalezas y debilidades, con el fin de enfocar el trabajo en aquellas capacidades que hoy limitan el rendimiento. Sin embargo, esta misma estrategia plantea de inmediato una nueva dificultad metodológica, relacionada con la definición de qué constituye una fortaleza y una debilidad, así como con el establecimiento de criterios que permitan identificarlas de manera objetiva y consistente.

2.2. Formulación del problema

¿Cómo aplicar técnicas de clusterización sobre un dataset público de CrossFit para generar insumos que fundamenten un sistema comparativo orientado a la identificación de fortalezas y debilidades relativas en un atleta, a partir de su comparación con perfiles de rendimiento similar?

2.3 OBJETIVOS DEL PROYECTO

2.3.1. Objetivo general

Explorar técnicas de clusterización aplicadas a variables de rendimiento de un dataset público de atletas de CrossFit, con el fin de fundamentar un sistema comparativo que permita a cada atleta evaluar su posición relativa frente a perfiles de rendimiento similares e identificar sus principales fortalezas y debilidades.

2.3.2. Objetivos específicos

Estructurar un dataset público de atletas de CrossFit, integrando y depurando variables físicas y de rendimiento para su uso en análisis de clusterización.

Aplicar técnicas de clusterización con base en las variables de rendimiento seleccionadas en el dataset de atletas de CrossFit.

Validar la calidad y coherencia de los clusters generados mediante métricas de machine learning.

Caracterizar e interpretar los clusters obtenidos mediante el análisis de sus patrones de rendimiento y estructura interna, incorporando variables físicas como elementos contextuales complementarios.

La correspondencia entre cada objetivo específico y el desarrollo del documento se organiza así:

OE1. Estructurar un dataset público… → Cap. 4, secciones 4.2–4.5.

OE2. Aplicar técnicas de clusterización… → Cap. 4, secciones 4.6–4.9; Cap. 5, secciones 5.1–5.6.

OE3. Validar la calidad y coherencia de los clusters… → Cap. 5, secciones 5.3–5.5; detalle numérico en el Anexo A.

OE4. Caracterizar e interpretar los clusters… → Cap. 5, secciones 5.6–5.11.

Objetivo general (sistema comparativo) → Cap. 6 y Cap. 7.

Esta correspondencia no sustituye el desarrollo de cada capítulo; solo orienta la lectura según el objetivo que se desee verificar.

2.4. JUSTIFICACIÓN

El análisis del rendimiento en CrossFit enfrenta el reto de integrar la diversidad de capacidades involucradas — fuerza, resistencia, potencia y habilidades gimnásticas— dentro de modelos que permitan interpretaciones objetivas y comparables. Si bien en los últimos años se han incorporado herramientas de análisis de datos y aprendizaje automático en el ámbito deportivo, aún existe margen para

explorar enfoques que aborden de forma más integral las múltiples dimensiones del rendimiento y contribuyan a orientar las decisiones de entrenamiento, especialmente cuando se trata de aportar elementos útiles a una pregunta tan amplia y compleja como la de cómo progresar en este deporte. En este marco, la aplicación de técnicas de clusterización constituye una oportunidad metodológica para examinar patrones de similitud entre atletas y generar una base de referencia que habilite comparaciones más justas y precisas. El presente proyecto no pretende construir un modelo definitivo, sino proponer una estrategia que combine análisis estadístico y métodos de aprendizaje automático para enriquecer la comprensión del rendimiento en CrossFit desde una perspectiva exploratoria.

Desde el punto de vista práctico, esta aproximación podría facilitar en el futuro el desarrollo de herramientas que ayuden a atletas y entrenadores a identificar fortalezas y debilidades relativas, con base en perfiles comparables, y así contribuir a una planificación más personalizada. En el plano académico, el trabajo aporta una mirada experimental

sobre la aplicabilidad de técnicas de clusterización en entornos de entrenamiento funcional y alta variabilidad como el CrossFit.

3. MARCO TEÓRICO Y ANTECEDENTES

3.1 CrossFit y entrenamiento funcional

El CrossFit es una disciplina de entrenamiento funcional de alta intensidad que integra movimientos provenientes de la halterofilia, la gimnasia y el acondicionamiento metabólico con el propósito de desarrollar una condición física general e integral [1]. Su metodología se fundamenta en la ejecución de movimientos funcionales, constantemente variados y realizados a alta intensidad, buscando mejorar simultáneamente múltiples capacidades físicas, entre ellas la fuerza, la resistencia cardiovascular, la resistencia muscular, la potencia, la velocidad, la coordinación, la agilidad, el equilibrio, la flexibilidad y la precisión [1]. A diferencia de disciplinas deportivas especializadas, donde el rendimiento suele depender predominantemente de una capacidad específica, el CrossFit plantea un enfoque multidimensional en el que el desempeño resulta de la interacción de numerosos componentes físicos y técnicos.

Esta naturaleza multidimensional también se refleja en la estructura de las competiciones oficiales. Tanto el CrossFit Open como los CrossFit Games evalúan a los atletas mediante pruebas que combinan levantamientos olímpicos, ejercicios gimnásticos, carreras, remo, ciclismo, natación y otros movimientos funcionales, exigiendo un alto nivel de desempeño en dominios fisiológicos considerablemente diferentes [1]. En consecuencia, el rendimiento competitivo no puede describirse únicamente mediante indicadores aislados de fuerza, resistencia o velocidad, sino que requiere considerar simultáneamente múltiples dimensiones del acondicionamiento físico [2].

Uno de los principios fundamentales del deporte consiste precisamente en identificar al atleta con el mayor nivel de fitness global. Esto se evidencia en el sistema de puntuación empleado durante los CrossFit Games, donde el título de Fittest on Earth no necesariamente corresponde al atleta que obtiene el primer lugar en una prueba específica, sino a aquel que mantiene el mejor desempeño promedio a lo largo de todas las pruebas de la competencia.

Este criterio refleja la filosofía del CrossFit, según la cual la condición física debe entenderse como una capacidad integral para responder eficientemente a demandas físicas variadas e impredecibles, más que como la excelencia en una única habilidad [1].

Con el fin de evaluar objetivamente el rendimiento, la comunidad de CrossFit ha desarrollado diferentes indicadores estandarizados. Entre ellos se encuentran los Benchmark Workouts of the Day (Benchmark WODs), entrenamientos diseñados para medir capacidades específicas bajo condiciones reproducibles, y los Personal Records (PRs), que representan el mejor desempeño alcanzado por un atleta en levantamientos o pruebas determinadas. Adicionalmente, competencias como el CrossFit Open generan anualmente una de las mayores bases de datos públicas del deporte, al reunir información de cientos de miles de atletas que realizan exactamente los mismos entrenamientos bajo protocolos estandarizados. Esta disponibilidad de información ha permitido el desarrollo de investigaciones orientadas a identificar los factores que determinan el rendimiento competitivo y ha abierto la posibilidad de aplicar técnicas de análisis de datos y aprendizaje automático para estudiar patrones de desempeño entre atletas [2], [3].

Entre los indicadores más utilizados se encuentran levantamientos máximos como el snatch (arranque), el clean and jerk (envión) y el deadlift (peso muerto), así como Benchmark WODs como Fran, Helen o Grace, entrenamientos estandarizados cuyo tiempo o carga permiten comparar el rendimiento entre atletas.

Diversos estudios han demostrado que el rendimiento en CrossFit depende de la interacción entre variables antropométricas, capacidades de fuerza, resistencia muscular, potencia y desempeño en pruebas estandarizadas, sin que exista un único indicador capaz de explicar por sí solo el nivel competitivo de un atleta [2], [3]. Esta evidencia respalda la necesidad de emplear enfoques multivariados que permitan analizar simultáneamente múltiples características del deportista, identificar relaciones complejas entre ellas y reconocer perfiles de rendimiento con características similares. En este contexto, las técnicas de inteligencia artificial y aprendizaje automático constituyen una alternativa prometedora para transformar grandes volúmenes de datos deportivos en información útil para apoyar

procesos de evaluación, comparación y planificación del entrenamiento, aspectos que fundamentan el desarrollo de la presente investigación.

3.2. Inteligencia artificial y análisis de datos

El crecimiento en la disponibilidad de datos deportivos durante las últimas décadas ha transformado la manera en que se evalúa el rendimiento de los atletas. La incorporación de plataformas digitales, dispositivos de monitoreo y sistemas de registro de competencias ha permitido almacenar grandes volúmenes de información relacionados con variables antropométricas, indicadores fisiológicos, resultados deportivos y métricas de entrenamiento. Como consecuencia, el análisis de estos datos ha dejado de depender exclusivamente de la experiencia de entrenadores y especialistas, dando paso a metodologías computacionales capaces de extraer conocimiento de forma objetiva y reproducible [4].

En este contexto, la inteligencia artificial (IA) puede definirse como el conjunto de métodos y sistemas computacionales diseñados para realizar tareas que tradicionalmente requieren capacidades asociadas al razonamiento humano, como el aprendizaje, el reconocimiento de patrones, la toma de decisiones y la resolución de problemas [4], [5]. Dentro de este campo, el aprendizaje automático (Machine Learning) constituye una de las áreas de mayor desarrollo, al proporcionar algoritmos capaces de identificar relaciones complejas entre los datos y generar modelos a partir de la experiencia, sin necesidad de programar explícitamente todas las reglas que describen el problema [5].

De acuerdo con Mitchell [5], un programa aprende de la experiencia cuando mejora su desempeño en una determinada tarea como resultado de la información previamente observada. Esta capacidad ha convertido al aprendizaje automático en una herramienta ampliamente utilizada para analizar fenómenos donde la complejidad de las relaciones entre variables dificulta la construcción de modelos analíticos tradicionales. En el ámbito deportivo, estas técnicas permiten aprovechar simultáneamente información proveniente de múltiples dimensiones del rendimiento, identificando patrones que difícilmente podrían detectarse mediante análisis univariados o inspección visual.

Los algoritmos de aprendizaje automático suelen clasificarse en tres grandes categorías: aprendizaje supervisado, aprendizaje no supervisado y aprendizaje por refuerzo [4], [5]. El aprendizaje supervisado busca construir modelos capaces de predecir una variable objetivo conocida a partir de ejemplos previamente etiquetados, siendo ampliamente utilizado para tareas de clasificación y regresión. Por su parte, el aprendizaje por refuerzo se enfoca en el desarrollo de agentes capaces de aprender estrategias de decisión mediante la interacción con un entorno y la maximización de una recompensa acumulada. Finalmente, el aprendizaje no supervisado tiene como objetivo descubrir estructuras, patrones o relaciones existentes dentro de los datos cuando no se dispone de etiquetas o categorías previamente definidas.

En el contexto del análisis del rendimiento deportivo, el aprendizaje no supervisado resulta particularmente atractivo debido a que, en muchos casos, no existe una clasificación objetiva que permita definir diferentes tipos de atletas. Más bien, el propósito consiste en descubrir agrupaciones naturales a partir de las similitudes presentes entre múltiples variables físicas y de rendimiento. Este enfoque permite identificar perfiles de deportistas con características comunes, analizar diferencias entre grupos y generar información útil para apoyar procesos de evaluación y planificación del entrenamiento [6].

La presente investigación se enmarca dentro de este último paradigma. Dado que el objetivo no consiste en predecir el rendimiento futuro de un atleta, sino en identificar perfiles similares a partir de variables antropométricas, levantamientos máximos y resultados obtenidos en Benchmark WODs, el problema se aborda desde una perspectiva de aprendizaje no supervisado. En consecuencia, la técnica de clusterización constituye el eje metodológico del sistema propuesto, cuyos fundamentos teóricos se desarrollan en la siguiente sección.

3.3. Técnicas de clusterización

La clusterización, o análisis de conglomerados (*Cluster Analysis*), es una de las principales técnicas de aprendizaje no supervisado para descubrir estructuras naturales en un conjunto de datos. Su objetivo es agrupar observaciones de modo que las del mismo grupo sean similares entre sí y las de grupos distintos sean lo más

disímiles posible [6], [7]. A diferencia de los métodos supervisados, las categorías no se conocen de antemano: se identifican a partir de las características observadas.

Desde una perspectiva geométrica, cada observación se representa como un punto en un espacio multidimensional cuyas coordenadas son las variables del análisis. Los algoritmos buscan regiones de mayor proximidad mediante medidas de distancia o disimilitud (p. ej. euclidiana, Manhattan o Mahalanobis) [7]. La calidad del agrupamiento depende del algoritmo y de la representación del espacio de características.

La utilidad de la técnica radica en sintetizar grandes volúmenes de información en un conjunto reducido de perfiles representativos. Ha sido aplicada en biología, medicina, mercadeo, minería de datos y ciencias del deporte [7], [8], por ejemplo para identificar perfiles de atletas, estilos de juego o patrones de rendimiento cuando no existe una clasificación objetiva previa [8].

Entre las familias más utilizadas se encuentran los métodos particionales (K-Means, K-Medoids), los jerárquicos, los basados en modelos probabilísticos (GMM) y los difusos (*Fuzzy Clustering*) [7], [9]. A continuación se formalizan las técnicas de preparación, agrupamiento y evaluación de balance empleadas en este estudio. Su aplicación concreta sobre las cohortes se desarrolla en Metodología.

Estandarización de variables numéricas. Dado que las variables se expresan en unidades heterogéneas, se emplea la estandarización *z-score*. Para cada variable numérica x_j, con media muestral μ_j y desviación estándar s_j:

$$z_{ij} = \frac{x_{ij} - \mu_j}{s_j}$$

donde i indexa observaciones y j variables. Así se evita que magnitudes mayores dominen la distancia euclídea. Los perfiles interpretativos se reportan en unidades originales.

Codificación de variables categóricas. Las categóricas se representan mediante *One-Hot Encoding*. Si una categoría c de una variable con C niveles genera un vector indicador $\mathbf{e}_c \in \{0,1\}^{C-1}$ (omitiendo una categoría de referencia), cada observación queda $\mathbf{x}_i = [\mathbf{z}_i \mid \mathbf{e}_i]$.

Algoritmo K-Means. Particiona n observaciones en k grupos minimizando la inertia:

$$J = \sum_{i=1}^{n} \| \mathbf{x}_i - \boldsymbol{\mu}_{c(i)} \|_2^2$$

donde $c(i) \in \{1, \dots, k\}$ es la etiqueta asignada a i y $\boldsymbol{\mu}_j$ el centroide del cluster j. El algoritmo alterna asignación al centroide más cercano y actualización de centroides hasta convergencia.

Clustering aglomerativo con enlace de Ward. En cada paso se fusionan los dos clusters cuyo unión produce el menor incremento de la suma de cuadrados intra-grupo. Si A y B son disjuntos:

$$d(A,B) = \sqrt{\frac{|A| \ |B|}{|A|+|B|}} \ \| \bar{\mathbf{x}}_A - \bar{\mathbf{x}}_B \|_2$$

La partición en k grupos se obtiene cortando el dendrograma. A diferencia de K-Means, no depende de una inicialización aleatoria de centroides.

Modelos de mezcla gaussiana (GMM). Con k componentes, la densidad de $\mathbf{x}_i$ es:

$$p(\mathbf{x}_i) = \sum_{j=1}^{k} \pi_j \ \mathcal{N}(\mathbf{x}_i \mid \boldsymbol{\mu}_j, \boldsymbol{\Sigma}_j)$$

con $\sum_j \pi_j = 1, \pi_j \geq 0$. Los parámetros se estiman por EM. Para comparar k se usan:

$$\text{AIC} = 2p - 2\ln(\hat{L}), \text{BIC} = p\ln(n) - 2\ln(\hat{L})$$

donde p es el número de parámetros libres, n el tamaño muestral y $\hat{L}$ la verosimilitud maximizada. Valores menores indican mejor compromiso ajuste–complejidad.

Reducción de dimensionalidad. Técnicas como PCA o t-SNE proyectan a menor dimensión. En este estudio no forman parte del pipeline de clusterización, a fin de preservar la interpretabilidad variable a variable; su uso, si existe, se limita a visualización exploratoria.

Balance de clusters. No se aplican técnicas de balanceo de clases (SMOTE, *under/oversampling*). Se evalúa el balance estructural de la partición mediante:

$$\text{max\_cluster\_share} = \max_j \frac{|\ C_j\ |}{n}, \text{min\_cluster\_size} = \min_j |\ C_j\ |$$

para detectar un cluster dominante o micro-clusters inestables.

3.4 Los datos faltantes como problema para la clusterización

En aplicaciones de aprendizaje no supervisado, la calidad de los resultados depende directamente de la información disponible para describir cada observación. En particular, los algoritmos de clusterización parten del supuesto de que todos los individuos pueden representarse mediante un mismo conjunto de variables, de manera que sea posible calcular

medidas de similitud o distancia entre ellos. Cuando este supuesto no se cumple debido a la presencia de datos faltantes, la estructura geométrica del conjunto de datos se ve afectada y, en consecuencia, también lo hace la calidad de los conglomerados obtenidos [10], [11].

Desde una perspectiva geométrica, cada observación puede representarse como un punto dentro de un espacio multidimensional cuyas dimensiones corresponden a las variables consideradas en el análisis. La distancia entre dos observaciones refleja el grado de similitud existente entre ellas y constituye el principal criterio utilizado por la mayoría de algoritmos de agrupamiento para construir los conglomerados. Sin embargo, cuando diferentes observaciones presentan valores faltantes en distintas variables, dichas distancias dejan de calcularse sobre un espacio común y pasan a depender únicamente del subconjunto de variables compartidas por cada par de individuos.

Considérese, por ejemplo, un conjunto de datos conformado por tres variables: snatch, deadlift y el tiempo obtenido en el benchmark fran. Si un atleta únicamente reporta su marca en snatch, otro únicamente registra deadlift y un tercero solo dispone del tiempo en Fran, las comparaciones entre ellos se realizan sobre variables diferentes. Mientras dos atletas pueden compararse exclusivamente mediante un indicador de fuerza máxima, otros únicamente pueden hacerlo a partir de una prueba metabólica. Como resultado, las distancias calculadas dejan de representar una medida homogénea de similitud entre todos los individuos del conjunto de datos.

Esta situación tiene varias implicaciones para el análisis de conglomerados. En primer lugar, modifica la geometría del espacio de representación, ya que las observaciones dejan de ocupar posiciones comparables dentro de un mismo espacio de características. En segundo lugar, debilita la representación del nivel de condición física (fitness) de cada atleta, al construirse sobre información parcial y heterogénea. Finalmente, puede impedir que el algoritmo identifique la estructura real presente en los datos. Es posible, por ejemplo, que los atletas se agrupen naturalmente según una combinación de capacidades de fuerza y acondicionamiento metabólico; sin embargo, si una proporción importante de ellos únicamente reporta variables relacionadas con uno de estos dominios, el algoritmo puede ser

incapaz de reconstruir dichos patrones aun cuando realmente existan.

La problemática se intensifica cuando el porcentaje de datos faltantes es elevado. En estos escenarios disminuye el solapamiento de variables observadas entre pares de individuos, fenómeno conocido como observed overlap. Como consecuencia, las distancias calculadas se vuelven progresivamente más inestables y aumenta la probabilidad de obtener conglomerados influenciados por los propios patrones de ausencia de información, en lugar de reflejar similitudes reales entre los atletas [11].

En consecuencia, el tratamiento de los datos faltantes constituye una etapa previa e indispensable antes de aplicar técnicas de clusterización. No obstante, la selección de una estrategia adecuada depende de la naturaleza del mecanismo que origina dichas ausencias. Por esta razón, el primer paso metodológico consiste en caracterizar el mecanismo generador de los datos faltantes, aspecto que se desarrolla en la siguiente subsección.

3.4.1 Mecanismos de datos faltantes

Una vez reconocido que los datos faltantes pueden afectar significativamente la estructura geométrica del espacio de representación y, por ende, la calidad de los conglomerados obtenidos, el siguiente paso consiste en comprender cómo se generan dichas ausencias. La literatura especializada coincide en que la elección de una estrategia para tratar los datos faltantes no puede realizarse de manera independiente del mecanismo que los origina. En consecuencia, antes de seleccionar una técnica de imputación o un algoritmo capaz de manejar valores ausentes, es necesario caracterizar el denominado mecanismo de datos faltantes (missing data mechanism) [10], [11].

El concepto fue introducido por Rubin [12], quien plantea que cada valor presente en un conjunto de datos posee una determinada probabilidad de ser observado o faltar. El proceso probabilístico que gobierna dichas probabilidades recibe el nombre de mecanismo de datos faltantes y constituye el elemento central para determinar la validez de los métodos estadísticos que posteriormente pueden emplearse. En otras palabras, dos conjuntos de datos con el mismo porcentaje de valores faltantes pueden requerir estrategias completamente

diferentes si las causas que originan dichas ausencias son distintas.

Rubin clasificó estos mecanismos en tres categorías fundamentales. La primera corresponde a Missing Completely at Random (MCAR), situación en la que la probabilidad de que un dato esté ausente es completamente independiente tanto del valor faltante como de las demás variables observadas. Bajo este supuesto, las observaciones con datos faltantes constituyen, en esencia, una muestra aleatoria del conjunto de datos. Un ejemplo clásico corresponde a la pérdida de información ocasionada por un fallo completamente aleatorio en un instrumento de medición, como una balanza que deja de registrar algunas observaciones debido al agotamiento inesperado de su batería [11], [12].

La segunda categoría corresponde a Missing at Random (MAR). En este caso, la probabilidad de ausencia depende únicamente de variables que sí han sido observadas en el conjunto de datos, pero no del valor faltante una vez dichas variables son conocidas. Por ejemplo, podría ocurrir que atletas con menor experiencia o pertenecientes a determinadas categorías competitivas presenten una mayor probabilidad de no reportar algunos Benchmark WODs. En este escenario, la ausencia de información puede modelarse utilizando las variables disponibles, permitiendo aplicar técnicas de imputación que aprovechan las relaciones existentes entre las observaciones [10], [11].

Finalmente, el mecanismo Missing Not at Random (MNAR) describe situaciones donde la probabilidad de que un dato esté ausente depende del propio valor faltante o de factores no observados. Un ejemplo frecuente en el contexto deportivo sería que algunos atletas decidan no reportar sus resultados en determinados levantamientos debido a que consideran que su desempeño fue inferior al esperado. En este caso, la ausencia de información contiene por sí misma información sobre el valor no observado, haciendo que el problema no pueda resolverse únicamente a partir de los datos disponibles [10], [12].

La importancia de esta clasificación radica en que el mecanismo de datos faltantes condiciona directamente las técnicas que pueden aplicarse durante el preprocesamiento de la información. Mientras que bajo el supuesto MCAR pueden emplearse procedimientos relativamente simples sin introducir sesgos importantes, los escenarios MAR suelen requerir

métodos de imputación que aprovechen las relaciones entre variables observadas, como la imputación múltiple. Por el contrario, cuando se sospecha un mecanismo MNAR, la información disponible resulta insuficiente para identificar de manera única los valores faltantes, siendo necesario incorporar supuestos adicionales, información externa o realizar análisis de sensibilidad para evaluar el impacto de diferentes mecanismos de ausencia [10]–[12].

En consecuencia, la caracterización del mecanismo de datos faltantes constituye una decisión metodológica fundamental, ya que determina tanto la validez de los procedimientos de imputación como la confiabilidad de los análisis posteriores. Sobre esta base, la siguiente subsección presenta las principales estrategias propuestas en la literatura para tratar datos incompletos antes de aplicar técnicas de clusterización.

3.4.2 Estrategias para el tratamiento de datos faltantes

Una vez caracterizado el mecanismo de datos faltantes, el siguiente paso consiste en seleccionar una estrategia adecuada para tratar la información incompleta antes de realizar el análisis. La literatura propone diferentes enfoques para abordar este problema, cuya conveniencia depende tanto del mecanismo de ausencia identificado como de las características del conjunto de datos y del método analítico que se pretende aplicar. En términos generales, estas estrategias pueden agruparse en tres grandes familias: imputación previa de los datos, algoritmos capaces de trabajar directamente con valores faltantes y modelos probabilísticos o de variables latentes [10], [11].

La primera estrategia consiste en realizar una **imputación previa**, es decir, completar el conjunto de datos antes de aplicar cualquier algoritmo de análisis. Bajo este enfoque, los valores faltantes son estimados a partir de la información disponible utilizando diferentes procedimientos estadísticos o computacionales. Entre los métodos más empleados se encuentran la imputación mediante la media o mediana, la imputación basada en vecinos más cercanos (*K-Nearest Neighbors Imputation*), la imputación múltiple (*Multiple

Imputation*) y diversos métodos de completación de matrices (*Matrix Completion*). La principal ventaja de esta estrategia radica en que, una vez completado el conjunto de datos, puede utilizarse prácticamente cualquier algoritmo convencional de clusterización. Sin embargo, una imputación inadecuada puede introducir relaciones artificiales entre las variables y modificar la estructura real de los datos, afectando posteriormente la formación de los conglomerados [10], [11].

Una segunda familia corresponde a los denominados algoritmos missing-aware, diseñados para incorporar explícitamente la presencia de valores faltantes durante el proceso de agrupamiento sin necesidad de realizar una imputación previa. Estos métodos adaptan el cálculo de las medidas de similitud o distancia utilizando únicamente la información disponible entre cada par de observaciones. Dentro de este grupo se encuentran enfoques basados en la distancia de Gower, algoritmos jerárquicos construidos sobre matrices de disimilitud parciales y variantes de K-Medoids capaces de operar con datos incompletos. Aunque estos métodos evitan imputar información inexistente, su desempeño puede deteriorarse cuando el porcentaje de variables compartidas entre observaciones es reducido, produciendo estimaciones inestables de las distancias y, en consecuencia, agrupamientos menos confiables [7], [11].

Finalmente, una tercera estrategia corresponde a los modelos probabilísticos o de variables latentes, cuyo objetivo consiste en modelar explícitamente la estructura generadora de los datos. En lugar de completar directamente las observaciones, estos enfoques asumen que los datos observados son manifestaciones de una estructura latente de menor dimensión que puede estimarse incluso en presencia de información incompleta. Dentro de esta categoría se encuentran los modelos de mezcla gaussianos estimados mediante el algoritmo EM (*Expectation-Maximization*), el análisis de componentes principales probabilístico (PPCA), los métodos de factorización matricial y técnicas de completación de matrices como *SoftImpute*. Estos enfoques suelen ofrecer buenos resultados cuando los datos presentan una estructura latente bien definida; sin embargo, requieren supuestos estadísticos más fuertes y una mayor complejidad computacional para su implementación [6], [10].

La selección entre estas tres estrategias no constituye una decisión independiente del mecanismo de datos faltantes descrito en la subsección anterior. Por el contrario, ambas decisiones se encuentran estrechamente relacionadas. Cuando el mecanismo puede asumirse como MCAR, procedimientos relativamente simples de imputación suelen producir resultados satisfactorios. Bajo el supuesto MAR, la literatura recomienda emplear métodos que aprovechen las relaciones existentes entre las variables observadas, siendo la imputación múltiple una de las alternativas más ampliamente aceptadas. En contraste, cuando se sospecha un mecanismo MNAR, el problema deja de ser identificable únicamente a partir de los datos observados, por lo que resulta necesario modelar explícitamente el mecanismo de ausencia o complementar el análisis mediante estudios de sensibilidad [10]–[12].

En consecuencia, la elección de una estrategia para tratar los datos faltantes debe entenderse como una decisión metodológica condicionada por el mecanismo de generación de las ausencias y por las características particulares del problema de investigación. A partir de estas consideraciones, a continuación se presentan formalmente dos estrategias pertenecientes a la familia de imputación previa: la imputación one-step basada en clusters y la imputación múltiple (MICE) con pooling. Su aplicación concreta sobre las cohortes del estudio se detalla en la sección de Metodología.

3.5. Imputación one-step e imputación múltiple

Dentro de la familia de **imputación previa**, dos estrategias resultan particularmente relevantes para el análisis de datos incompletos previo a la clusterización: la imputación one-step basada en clusters y la imputación múltiple mediante MICE.

Imputación one-step basada en clusters. La imputación one-step (Steinley & Henson) alterna de forma iterativa una etapa de clusterización y una etapa de imputación condicionada al grupo. El procedimiento conceptual es el siguiente: (1) se realiza una imputación inicial bootstrap, típicamente con la mediana global en variables numéricas y la moda en variables categóricas; (2) se estandarizan las variables y se aplica K-Means con un número fijo k de clusters; (3) cada valor originalmente faltante se reemplaza por la mediana

(o moda) del cluster al que fue asignada la observación, sin sobrescribir valores observados; (4) se repiten los pasos (2)–(3) hasta que las etiquetas de cluster se estabilizan. Formalmente, para un valor faltante x_{ij} de la observación i en la variable j:

$$\hat{x}_{ij} = \text{mediana}\{x_{\ell j} : c(\ell) = c(i),\ x_{\ell j} \text{ observado}\}$$

donde $c(\cdot)$ denota la asignación de cluster. La principal ventaja de este enfoque es que la imputación se condiciona a perfiles similares, en lugar de a una única tendencia global. Su limitación principal es que produce una sola versión completada del conjunto de datos y, por tanto, no representa explícitamente la incertidumbre asociada a la ausencia.

Imputación múltiple (MICE) y pooling. La imputación múltiple genera M versiones completas del conjunto de datos, reconociendo que un valor faltante admite múltiples estimaciones plausibles. En la implementación MICE (*Multiple Imputation by Chained Equations*), cada variable con ausencias se modela de forma iterativa condicionada al resto de variables. Tras obtener las M réplicas, una forma habitual de consolidar las estimaciones numéricas es el *pooling* por media:

$$\bar{x}_j^{(\text{pooled})} = \frac{1}{M} \sum_{m=1}^{M} x_j^{(m)}$$

donde $x_j^{(m)}$ es el valor imputado de la variable j en la réplica m. Este enfoque permite, además, evaluar la sensibilidad de los resultados posteriores —por ejemplo, de una partición en clusters— frente a distintas completaciones del dataset. Una forma natural de comparar particiones obtenidas tras one-step y tras imputación múltiple es el índice de Rand ajustado (ARI), cuya definición formal se presenta en la siguiente subsección.

3.6. Métricas de evaluación de modelos de agrupamiento

La evaluación de un agrupamiento en aprendizaje no supervisado no dispone de una

etiqueta verdadera externa. Por ello, se emplean métricas internas de calidad, índices de concordancia entre particiones y criterios de información, complementados con medidas de balance estructural.

Coeficiente de silueta. Para cada observación i se define la cohesión intra-cluster $a(i)$ como la distancia media a las observaciones del mismo cluster, y la separación inter-cluster $b(i)$ como la distancia media al cluster vecino más cercano. El coeficiente individual es:

$$s(i) = \frac{b(i) - a(i)}{\max\{a(i), b(i)\}}$$

y la silueta global $\bar{s} = \frac{1}{n}\sum_{i=1}^{n} s(i)$, con rango teórico $[-1,1]$. Valores cercanos a 1 indican buena separación; valores cercanos a 0 sugieren solapamiento. En datos multivariados reales, la silueta suele emplearse para comparar de forma relativa distintos valores de k, no como criterio único de decisión.

Inertia. En K-Means, la inertia coincide con la función objetivo J definida en la subsección de técnicas de clusterización: la suma de distancias euclídeas al cuadrado de cada punto a su centroide. Decrece al aumentar k y sirve como diagnóstico de compactación interna, pero no basta por sí sola para seleccionar el número de clusters.

Índice de Rand ajustado (ARI). Dados dos particionamientos $\mathcal{U}$ y $\mathcal{V}$ del mismo conjunto de n objetos, el índice de Rand (RI) mide la proporción de pares de observaciones en los que ambas particiones coinciden (juntos o separados). El ARI corrige dicho índice por el acuerdo esperado al azar:

$$\mathrm{ARI} = \frac{\mathrm{RI} - \mathbb{E}[\mathrm{RI}]}{\max(\mathrm{RI}) - \mathbb{E}[\mathrm{RI}]}$$

Valores cercanos a 1 indican alta concordancia; valores cercanos a 0 indican

concordancia comparable al azar. Esta métrica resulta útil para comparar estabilidad entre distintos valores de k, entre semillas de inicialización o entre particiones obtenidas tras distintas estrategias de imputación.

Criterios AIC y BIC. En modelos de mezcla gaussiana, AIC y BIC —definidos en [5.3.1]— permiten comparar configuraciones con distinto número de componentes, penalizando la complejidad del modelo.

Métricas de balance. Además de la calidad geométrica, interesa que los clusters tengan tamaños interpretables. Se emplean:

$$\text{max\_cluster\_share} = \frac{\max_j \mid C_j \mid}{n}, \text{min\_cluster\_size} = \min_j \mid C_j \mid$$

Un valor elevado de max_cluster_share indica un cluster dominante; un min_cluster_size demasiado bajo sugiere micro-clusters poco estables.

Score compuesto para la selección de k. Dado que maximizar únicamente la silueta puede favorecer particiones demasiado gruesas y desbalanceadas, puede definirse un puntaje compuesto que combine calidad geométrica, preferencia por mayor granularidad y balance estructural. Una formulación general es:

$$\text{score} = w_1\, s_{\text{norm}} + w_2\, k_{\text{norm}} + w_3\, \text{size\_ok} + w_4\, \text{share\_ok} + w_5\, \text{floor\_ok}$$

donde s_{norm} es la silueta normalizada en el barrido de k, k_{norm} favorece valores de k más altos dentro del rango evaluado, y los términos restantes son indicadores o puntuaciones asociadas al tamaño mínimo de cluster y a la ausencia de un grupo dominante. Los pesos w_i (con $\sum_i w_i = 1$) reflejan la prioridad metodológica del análisis. La parametrización concreta empleada en el estudio se detalla en Metodología.

3.7. Estadística descriptiva para la caracterización de perfiles

Tras obtener una partición, la interpretación de los conglomerados requiere resumir el comportamiento de cada variable dentro de cada grupo y comparar dichos resúmenes entre clusters. En este marco se emplean medidas de tendencia central agregadas por cluster, contrastes relativos entre perfiles y un índice de separación intergrupal. Su aplicación concreta sobre la cohorte de atletas se desarrolla en la sección de resultados.

Mediana por cluster. Para una variable j y un cluster c, sea $\{x_{ij}: i \in C_c\}$ el conjunto de valores observados (o ya imputados) de esa variable en el grupo. La mediana del cluster se denota $\tilde{x}_{jc}$. La mediana es menos sensible que la media a valores extremos y resulta adecuada cuando las distribuciones de rendimiento son asimétricas, situación frecuente en tiempos de benchmarks y cargas máximas.

Z-score de medianas entre clusters. Para comparar perfiles en una escala común, se construye un contraste relativo a partir de las k medianas de cada variable. Sea

$$\bar{\tilde{x}}_j = \frac{1}{k}\sum_{c=1}^{k} \tilde{x}_{jc}\,, s_{\tilde{x}_j} = \sqrt{\frac{1}{k-1}\sum_{c=1}^{k}(\tilde{x}_{jc} - \bar{\tilde{x}}_j)^2}$$

el promedio y la desviación estándar de las medianas de la variable j a lo largo de los k clusters. El z-score del perfil del cluster c en esa variable se define como

$$z_{jc} = \frac{\tilde{x}_{jc} - \bar{\tilde{x}}_j}{s_{\tilde{x}_j}}.$$

Valores de z_{jc} positivos indican una mediana relativamente alta frente al conjunto de perfiles; valores negativos, una mediana relativamente baja. En variables de tiempo, una mediana menor suele asociarse a mejor rendimiento; en cargas, repeticiones o puntuaciones, una mediana mayor suele asociarse a mejor rendimiento. Este z-score no describe la dispersión de cada atleta individual, sino la posición relativa del perfil agregado del cluster.

Ratio de separación. Para cuantificar qué variables discriminan más entre grupos, se define

$$R_j = \frac{\max_c \tilde{x}_{jc} - \min_c \tilde{x}_{jc}}{\mathrm{SD}_{\mathrm{pooled},j}},$$

donde $\mathrm{SD}_{\mathrm{pooled},j}$ es una medida de dispersión de referencia de la variable j en el conjunto analizado (p. ej. desviación estándar global o agrupada). Valores altos de R_j indican mayor contraste entre el cluster de mediana máxima y el de mediana mínima, en relación con la variabilidad típica de la variable. Este índice permite ordenar los indicadores según su capacidad para separar perfiles y orientar la interpretación atlética de la partición.

En conjunto, la mediana describe cada grupo en unidades originales, el z-score de medianas facilita la comparación relativa entre perfiles y el ratio de separación identifica las variables más discriminantes. Estas herramientas complementan las métricas geométricas de calidad del agrupamiento (silueta, inercia, ARI, balance), orientadas a evaluar la partición, con un aparato descriptivo orientado a interpretar los conglomerados.

3.8. Antecendentes

La aplicación de técnicas de aprendizaje automático al análisis del rendimiento deportivo ha experimentado un crecimiento sostenido durante las últimas décadas, impulsada por la disponibilidad de grandes volúmenes de datos y por la necesidad de comprender el desempeño de los atletas desde una perspectiva multidimensional. Dentro de este contexto, las técnicas de aprendizaje no supervisado y, particularmente, la clusterización, se han consolidado como herramientas para descubrir patrones latentes, segmentar poblaciones deportivas e identificar perfiles de rendimiento sin requerir categorías previamente definidas [4]–[7].

Desde el punto de vista metodológico, la clusterización constituye uno de los

problemas clásicos de la minería de datos y del aprendizaje no supervisado. Su propósito consiste en agrupar observaciones de manera que los elementos pertenecientes a un mismo conglomerado presenten una elevada similitud entre sí y una marcada diferencia respecto a los integrantes de otros grupos [6], [7]. Jain [9] señala que, pese a más de cincuenta años de desarrollo, la clusterización continúa siendo un área activa de investigación debido a desafíos como la selección del número óptimo de conglomerados, la evaluación objetiva de la calidad de las particiones, el tratamiento de datos heterogéneos y la interpretación de los grupos obtenidos. Estas dificultades han motivado el desarrollo de nuevos algoritmos y criterios de validación cada vez más robustos.

La consolidación de estas técnicas también se refleja en las ciencias del deporte. Fernández, Casals, Oliver, Plensa y Manisera realizaron una revisión panorámica (scoping review) de 278 investigaciones publicadas entre 1996 y 2023 sobre el uso de técnicas de agrupamiento en este campo [8]. Los autores evidenciaron un incremento sostenido en la utilización de métodos de clusterización durante los últimos catorce años, especialmente en disciplinas como fútbol, baloncesto y tenis. Asimismo, identificaron que los algoritmos más empleados fueron el clustering jerárquico (31.6 %) y k-means (26.3 %), aunque también señalaron importantes deficiencias metodológicas, entre ellas la ausencia de criterios estandarizados para seleccionar el número de conglomerados, la escasa descripción de los procedimientos de validación y la limitada reproducibilidad de numerosos estudios. Como conclusión, proponen fortalecer el rigor metodológico mediante lineamientos comunes para el uso y reporte de técnicas de clusterización en investigación deportiva.

Más allá de documentar el crecimiento de estas técnicas, investigaciones recientes han desarrollado metodologías específicamente orientadas al análisis de datos deportivos. D'Urso, De Giovanni y Vitale propusieron el algoritmo FCMd-MD-NC (Fuzzy C-Medoids Clustering for Mixed Data with Noise Cluster), diseñado para agrupar jugadores de fútbol considerando simultáneamente variables cuantitativas, categóricas, composicionales y posicionales [12]. El método incorpora funciones de

distancia específicas para cada tipo de dato, ponderación automática de variables y un clúster de ruido para reducir la influencia de observaciones atípicas. Su aplicación sobre información de 397 futbolistas profesionales permitió identificar perfiles de juego altamente estables, demostrando que la incorporación de estrategias robustas para datos mixtos mejora la caracterización de atletas complejos.

En una línea complementaria, Akhanli y Hennig desarrollaron un marco metodológico para la selección y validación objetiva de clusterizaciones en jugadores de fútbol [13]. Su propuesta combina medidas de disimilitud diseñadas específicamente para datos deportivos con un índice compuesto que integra criterios de homogeneidad interna, separación entre grupos, estabilidad, entropía y preservación de la estructura original de los datos. Aplicado sobre una base de 1501 futbolistas pertenecientes a ocho ligas europeas, el método produjo agrupamientos que mostraron una alta concordancia con la evaluación realizada por expertos deportivos. Este trabajo evidencia que la utilidad práctica de una clusterización depende no solo del algoritmo utilizado, sino también de la rigurosidad de los procedimientos empleados para validar e interpretar los conglomerados obtenidos.

Aunque la aplicación de técnicas de aprendizaje no supervisado ha aumentado considerablemente en diversas disciplinas deportivas, el CrossFit continúa siendo un área relativamente poco explorada desde esta perspectiva. La mayor parte de la literatura disponible se ha orientado a identificar los factores fisiológicos, antropométricos y de condición física asociados al rendimiento competitivo mediante análisis estadísticos tradicionales o modelos predictivos supervisados.

En este contexto, Meier, Schlie y Schmidt realizaron una revisión sistemática sobre los predictores del rendimiento en CrossFit [2]. Los autores concluyen que el desempeño competitivo depende de una interacción compleja entre fuerza máxima, potencia, resistencia aeróbica y anaeróbica, composición corporal, experiencia deportiva y habilidades gimnásticas, sin que exista una única variable capaz de

explicar consistentemente el rendimiento. Esta naturaleza multidimensional respalda el empleo de metodologías capaces de analizar simultáneamente múltiples atributos, como ocurre con las técnicas de clusterización.

Resultados similares fueron obtenidos por Mangine et al. [3], quienes analizaron los factores asociados al rendimiento en el CrossFit Open. Sus resultados mostraron que variables relacionadas con la fuerza máxima, la capacidad metabólica y las características antropométricas presentan asociaciones significativas con el desempeño competitivo, evidenciando que el éxito deportivo emerge de la interacción entre múltiples capacidades físicas y no del predominio de un único componente. Estos hallazgos sugieren que el estudio del rendimiento mediante perfiles integrales puede aportar una comprensión más completa del desempeño de los atletas que el análisis aislado de cada variable.

Otro desafío metodológico relevante corresponde al tratamiento de los datos faltantes. Las bases de datos deportivas suelen presentar porcentajes importantes de información incompleta debido a registros voluntarios, diferencias en los protocolos de medición o ausencia de determinadas pruebas físicas. Little y Rubin [10] demuestran que la eliminación directa de observaciones incompletas puede reducir el tamaño muestral e introducir sesgos significativos en los resultados, mientras que van Buuren [11] presenta estrategias modernas de imputación capaces de preservar la información disponible y producir estimaciones estadísticamente más consistentes. Este aspecto adquiere especial relevancia cuando se trabaja con bases de datos de gran dimensión y múltiples variables de rendimiento, como ocurre en el presente estudio.

En conjunto, los antecedentes revisados muestran una evolución clara en tres dimensiones. En primer lugar, la clusterización constituye actualmente una metodología consolidada para el descubrimiento de patrones en datos complejos [6], [7], [9]. En segundo lugar, las ciencias del deporte han incorporado progresivamente estas técnicas para construir perfiles de atletas mediante procedimientos cada vez

más robustos y rigurosamente validados [8], [12], [13]. Finalmente, la evidencia disponible sobre CrossFit confirma que el rendimiento competitivo es un fenómeno inherentemente multidimensional que requiere integrar simultáneamente múltiples indicadores físicos y deportivos [2], [3].

No obstante, persiste un vacío en la literatura respecto a la integración de estas tres líneas de investigación. Aunque existen estudios que analizan predictores del rendimiento en CrossFit, revisiones sobre la aplicación de técnicas de clusterización en las ciencias del deporte y metodologías avanzadas para el análisis de agrupamientos, aún son escasas las investigaciones que combinen procedimientos sistemáticos de imputación de datos faltantes, evaluación rigurosa de múltiples algoritmos de clusterización y caracterización estadística de perfiles aplicados específicamente a atletas de CrossFit. La presente investigación busca contribuir a este vacío mediante el desarrollo de un marco metodológico reproducible que integre dichas etapas para construir perfiles comparativos de rendimiento a partir de información deportiva multidimensional e incompleta.

4. METODOLOGÍA Y DESARROLLO DEL PROCESO ANALÍTICO

4.1 Enfoque metodológico, pipeline y redefinición del alcance de la fuente

El presente proyecto se desarrolló como un estudio exploratorio-aplicado, de enfoque cuantitativo, sobre datos reales de atletas de CrossFit. Su propósito fue aplicar, comparar y validar técnicas de clusterización en un conjunto con alta proporción de valores faltantes, a fin de fundamentar un sistema comparativo de rendimiento. En este trabajo, el objetivo de la clusterización no es etiquetar o clasificar al deportista con una categoría fija ("tipo de atleta"), sino operativo: dado un atleta, identificar un conjunto de atletas con características similares (su cluster) y, a continuación, compararlo con ese mismo grupo de referencia para estimar fortalezas y debilidades relativas en variables de rendimiento. Los clusters son, por tanto, vecindarios comparables, no rótulos finales del sistema.

El problema se abordó mediante aprendizaje no supervisado (capítulo 3): no se predice un resultado futuro ni se clasifica bajo categorías prefijadas, sino que se buscan agrupaciones a partir de variables antropométricas, levantamientos máximos y resultados en *Benchmark WODs*. La clusterización fue la técnica central, precedida de preprocesamiento y tratamiento de faltantes, y seguida de evaluación de modelos y de un prototipo funcional.

El trabajo se organizó en un pipeline de seis etapas, alineado con los objetivos de la sección 2.3:

- Recolección y exploración del dataset.
- Preprocesamiento y construcción de cohortes.
- Imputación bajo el supuesto operativo MAR.
- Clusterización post-imputación.

- Selección del modelo y validación de perfiles.
- Implementación del sistema comparativo.

El desarrollo fue incremental, en ciclos de unas dos semanas, con entregables verificables (datasets, scripts, métricas o módulos) y revisiones con la directora del proyecto; no se pretendió una adopción formal de SCRUM en equipo.

Desde el punto de vista experimental, no se aplicó un único algoritmo a un único conjunto. Se combinaron tres cohortes (mixta, masculina y femenina), dos estrategias de imputación (one-step y MI *pooled*) y tres familias de agrupamiento (K-Means, Agglomerative Ward y GMM). En la clusterización post-imputación, cada combinación se evaluó con un barrido de k (2…10), lo que generó dieciocho escenarios principales, más tablas de métricas y visualizaciones de soporte. La validación integró criterios cuantitativos (silueta, inertia, BIC, AIC y ARI) con balance e interpretabilidad deportiva, porque un óptimo geométrico no es útil si concentra a casi todos los atletas en uno o dos grupos o produce microgrupos inestables.

Tras ese contraste, la configuración operativa del sistema fue: cohorte masculina, imputación one-step y K-Means con $k = 6$. El resto de cohortes, algoritmos e imputación múltiple se conservó como evidencia de sensibilidad (capítulo de resultados). La implementación se realizó en Python (pandas, NumPy, scikit-learn, Matplotlib), con scripts y notebooks modulares, exportación en CSV y semillas fijas (random_state = 42) para favorecer la reproducibilidad.

Ese diseño, sin embargo, solo pudo ejecutarse tras redefinir la fuente de datos. En el anteproyecto se contempló información pública de la plataforma oficial de CrossFit Games (incluidos resultados asociados al Open), eventualmente mediante extracción automatizada. Durante el desarrollo, CrossFit Games actualizó sus políticas y prohibió la extracción, manipulación y uso externo de esos datos; una consulta a la organización confirmó que no era viable emplearlos fuera de su

ecosistema. Por ello se adoptó un dataset alternativo en Kaggle, cambio de alcance presentado y aprobado por el Comité de Programa de Ingeniería de Sistemas y Computación. La metodología de este documento corresponde a ese diseño autorizado.

En síntesis, el enfoque combina rigor analítico, comparación sistemática de alternativas y orientación a un producto concreto. La sección siguiente describe la fuente finalmente utilizada, la unidad de análisis y la estructura del dataset con el que se ejecutó el pipeline.

4.2 Fuente de datos, estructura del dataset e interpretación de variables

Tras la redefinición del alcance, el pipeline se ejecutó sobre un dataset público de atletas de CrossFit. Esta sección fija la fuente utilizada, la unidad de análisis, la organización de las variables y el sentido en que se interpretan para la clusterización.

El conjunto empleado proviene de Kaggle, repositorio *CrossFit Athletes* de Ulrik Thyge Pedersen (**https://www.kaggle.com/datasets/ulrikthygepedersen/crossfit-athletes**), bajo licencia Creative Commons Attribution 4.0 International (CC BY 4.0), que permite uso y adaptación con el crédito al autor. Conserva una estructura cercana a la prevista en el anteproyecto: antropometría, *Benchmark WODs* y marcas máximas de fuerza, reportadas de forma autodeclarada. El archivo bruto se almacenó en el proyecto como data/raw/rawAthletes.csv (CSV, UTF-8).

Cada fila corresponde a un único atleta: no hay series temporales ni múltiples registros por deportista. La trazabilidad se garantiza con athlete_id. El archivo crudo contiene 423 006 registros y 27 columnas, agrupadas como sigue:

Tabla 1. Variables de identificación y contexto

Variable	Tipo de dato	Unidad	Descripción
athlete_id	Identificador numérico	—	Identificador único del atleta en la fuente original.
name	Texto	—	Nombre del atleta.
region	Texto / categórica	—	Región geográfica asociada al registro del atleta.
team	Texto	—	Equipo reportado por el atleta, cuando existe.
affiliate	Texto	—	Box o afiliado de CrossFit asociado al atleta (gimnasio reportado; no es un catálogo cerrado).

Tabla 2. Variables demográficas y antropométricas

Variable	Tipo de dato	Unidad	Descripción
gender	Categórica	—	Sexo reportado del atleta (`Male` / `Female`).
age	Numérica	años	Edad del atleta.
height	Numérica	pulgadas (in)	Estatura del atleta.
weight	Numérica	libras (lb)	Peso corporal del atleta.

Tabla 3. Variables de rendimiento físico

Variable	Tipo de dato	Unidad	Descripción
fran	Numérica	segundos	Tiempo en Fran (21-15-9 *thrusters* y dominadas). Menor tiempo = mejor rendimiento.
helen	Numérica	segundos	Tiempo en Helen (3 rondas: 400 m, *kettlebell swings* y dominadas). Menor tiempo = mejor rendimiento.
grace	Numérica	segundos	Tiempo en Grace (30 *clean and jerk*). Menor tiempo = mejor rendimiento.
filthy50	Numérica	segundos	Tiempo en Filthy 50 (circuito de 10 movimientos × 50 repeticiones). Menor tiempo = mejor rendimiento.
fgonebad	Numérica	puntos	Puntuación en Fight Gone Bad (3 rondas, 5 estaciones). Mayor puntuación = mejor rendimiento.
run400	Numérica	segundos	Tiempo en 400 m. Menor tiempo = mejor rendimiento.
run5k	Numérica	segundos	Tiempo en 5 km. Menor tiempo = mejor rendimiento.
candj	Numérica	libras (lb)	1RM en *Clean & Jerk* (cargada y envión). Mayor carga = mejor rendimiento.
snatch	Numérica	libras (lb)	1RM en *Snatch* (arranque). Mayor carga = mejor rendimiento.
deadlift	Numérica	libras (lb)	1RM en *Deadlift* (peso muerto). Mayor carga = mejor rendimiento.
backsq	Numérica	libras (lb)	1RM en *Back Squat* (sentadilla trasera). Mayor carga = mejor rendimiento.
pullups	Numérica	repeticiones	Máximo de dominadas consecutivas. Mayor número = mejor rendimiento.

Tabla 4. Variables de cuestionario (texto libre)

Variable	Tipo de dato	Unidad	Descripción
eat	Texto	—	Hábitos alimenticios (respuesta abierta).
train	Texto	—	Hábitos o modalidad de entrenamiento (respuesta abierta).
background	Texto	—	Antecedentes deportivos reportados.
experience	Texto	—	Experiencia en CrossFit u otras disciplinas.
schedule	Texto	—	Frecuencia o agenda de entrenamiento.
howlong	Texto	—	Tiempo de práctica en CrossFit.

A diferencia de la propuesta inicial, el dataset final no incluye resultados históricos del CrossFit Open. El análisis deja de orientarse a explicar el rendimiento en pruebas oficiales concretas y se centra en patrones de similitud entre atletas a partir de las variables disponibles.

Esas variables se entienden como aproximaciones parciales al *fitness* general: en CrossFit la condición integral no se observa de forma directa, sino mediante pruebas complementarias (benchmarks, 1RM, pruebas metabólicas y gimnásticas). Ninguna variable basta por sí sola; en conjunto permiten una representación multidimensional del rendimiento, adecuada para el agrupamiento planteado en el marco teórico.

Con la fuente y el significado de las variables definidos, la sección siguiente caracteriza la calidad del dataset —en particular la magnitud de los datos faltantes— antes del preprocesamiento.

4.3 Exploración inicial, almacenamiento y limitaciones de la fuente

Antes del preprocesamiento conviene caracterizar la calidad del dataset, fijar cómo se almacenó la información y explicitar las limitaciones de la fuente, porque

condicionan las decisiones posteriores de limpieza, cohortes e imputación.

Se realizó un análisis exploratorio (EDA) sobre el archivo crudo. El hallazgo dominante fue la alta ausencia en variables de rendimiento: varios benchmarks superaron el 90 % de faltantes —*filthy50* (95,4 %), *run400* (94,7 %), *fgonebad* (93,0 %), *helen* (92,8 %) y *run5k* (91,5 %)—, mientras que *fran*, *pullups* y los levantamientos máximos se situaron entre aproximadamente 73 % y 87 %. Las variables antropométricas y demográficas estaban más completas, aunque aún con ausencias relevantes (*height* 62,2 %, *weight* 45,7 %). Las de cuestionario (74–78 % de ausencia) se descartaron después del pipeline principal por ser texto libre, de alta cardinalidad y poco útiles para clusterización numérica.

El EDA también reveló valores imposibles o inconsistentes. Para tratarlos de forma sistemática se definió, por variable numérica, un rango atlético plausible; lo que quedó fuera se anuló (pasó a faltante) en el preprocesamiento, sin eliminar el registro completo. Los límites y su justificación se resumen en la Tabla 5.

Tabla 5. Rangos de validez por variable y criterio de anulación

Variable	Unidad	Condición anulada	Justificación
age	años	> 62	Edades por encima del rango de atletas considerado; el 99 % de los datos se concentraba por debajo de este umbral; valores cercanos o superiores a 100 años se asociaron a errores de registro.
height	pulgadas	< 40 o > 96	Fuera del rango de estatura humana plausible en adultos (≈ 1,02 m a 2,44 m).
weight	libras	< 75 o > 450	Pesos corporales fisiológicamente inviables para un atleta adulto (≈ 34 kg a 204 kg).
fran	segundos	< 70 o > 2700	WOD corto; tiempos < 70 s son inferiores a la marca de élite y > 45 min son incompatibles con la prueba.

Variable	Unidad	Condición anulada	Justificación
helen	segundos	< 320 o > 3600	WOD con carrera; por debajo de ~5,3 min es humanamente inalcanzable y por encima de 60 min es incompatible.
grace	segundos	< 35 o > 1200	WOD corto (30 *clean and jerk*); tiempos < 35 s o > 20 min no son plausibles.
run400	segundos	< 43,03 o > 300	Límite inferior próximo al récord mundial de 400 m; > 5 min es incompatible con la prueba.
run5k	segundos	< 755,36 o > 7200	Límite inferior cercano al récord de 5 km; > 2 h es inconsistente para la prueba.
candj	libras	< 50 o > 500	1RM de *Clean & Jerk* fuera del rango físicamente razonable.
snatch	libras	< 45 o > 400	1RM de *Snatch* fuera del rango físicamente razonable.
deadlift	libras	< 45 o > 650	1RM de *Deadlift* fuera del rango físicamente razonable.
backsq	libras	< 45 o > 600	1RM de *Back Squat* fuera del rango físicamente razonable.
pullups	repeticiones	< 5 o > 110	Número de dominadas consecutivas fuera del máximo razonable reportado.
filthy50	segundos	< 600 o > 5400	Circuito largo (50 repeticiones × 10 movimientos); duración plausible entre ~10 y ~90 min.
fgonebad	puntos	< 50 o > 600	Puntuación (repeticiones acumuladas) fuera del rango alcanzable en Fight Gon

Esos rangos no fueron arbitrarios. Se apoyaron en tres criterios: (i) límites físicos anclados en marcas oficiales —por ejemplo, el piso de *run400* (43,03 s) y de *run5k* (755,36 s) corresponden a récords mundiales; los techos de *snatch*, *candj*, *deadlift* y *backsq* se fijaron por encima de marcas de élite en

halterofilia/powerlifting—; (ii) plausibilidad biológica en adultos para *age*, *height* y *weight*; y (iii) estructura de cada *Benchmark WOD* (*fran*, *helen*, *grace*, *filthy50*), que acota tiempos mínimos y máximos razonables.

La información se integró en un único archivo analizable (rawAthletes.csv), conservando el identificador por atleta y las unidades originales (segundos, libras, pulgadas; puntos en Fight Gone Bad). En esta etapa no hubo normalización ni imputación: el objetivo fue preservar la estructura original para la limpieza, las cohortes y el tratamiento de faltantes.

Las limitaciones de la fuente condicionaron el diseño. Al ser registros **autodeclarados**, hay sesgo de autorreporte, actualización irregular y ausencia selectiva de marcas; por eso se validó por rangos (Tabla 5) y se descartó asumir MCAR de forma operativa, orientando el análisis hacia MAR (con MNAR como posibilidad reconocida). La magnitud de los faltantes impidió usar el conjunto completo tal cual: hubo que definir cohortes por completitud e imputar (one-step y, como contraste, MI). Además, sin resultados del CrossFit Open ni composición corporal detallada, los perfiles se construyen solo con antropometría básica, 1RM y benchmarks, en clave de rendimiento funcional.

Aun así, el volumen (> 420 000 atletas) y la diversidad de variables físicas disponibles bastaron para clusterizar, identificar perfiles comparables y fundamentar el sistema comparativo. La sección siguiente describe la limpieza y el preprocesamiento aplicados sobre esta base.

4.4 Limpieza, variables del pipeline y salidas del preprocesamiento

Integrado el archivo bruto, esta etapa depura inconsistencias, delimita las variables utilizables y caracteriza la ausencia de información sin completar el dataset ni llevarlo aún a una escala común. El resultado es una base coherente para cohortes, imputación y clusterización.

Sobre cada variable se examinaron distribución, proporción de faltantes, atípicos y la relación entre compl

Tabla 6. Proporción de datos faltantes por variable (dataset crudo, n = 423 006)

Variable	% faltante	Variable	% faltante
athlete_id	0,00 %	fgonebad	92,97 %
name	21,72 %	run400	94,74 %
region	40,60 %	run5k	91,47 %
Team	63,32 %	candj	75,31 %
Affiliate	42,81 %	snatch	77,00 %
gender	21,72 %	deadlift	72,74 %
age	21,72 %	backsq	73,87 %
height	62,21 %	Pullups	88,04 %
weight	45,65 %	eat	77,79 %
fran	86,90 %	train	74,98 %
helen	92,84 %	background	76,61 %
grace	90,37 %	experience	75,19 %
filthy50	95,42 %	schedule	76,86 %
		howlong	74,18 %

Tabla 7. Distribución de valores por variable numérica (antes de la limpieza)

Variable	Mínimo	Mediana	Máximo
age	13	31	125
height	0	69	8.388.607
weight	1	170	20.175
fran	1	290	8.388.607
helen	1	595	8.388.607
grace	−60	193	8.388.607
filthy50	1	1.550	8.388.607
fgonebad	0	294	8.388.607
run400	1	71	8.388.607
run5k	1	1.380	8.388.607
candj	−45	195	8.388.607
snatch	0	145	8.388.607
deadlift	−500	345	8.388.607
backsq	−7	275	8.388.607
pullups	−6	27	2.147.483.647

La comparación entre mediana y extremos mostró valores imposibles: mínimos negativos o nulos en variables físicamente positivas (p. ej. *deadlift* = −500, *grace* = −60) y máximos que coinciden con **centinelas de desbordamiento entero** —8.388.607 y 2.147.483.647— repetidos en 12 variables numéricas. Esa coincidencia no es rendimiento real, sino artefacto de codificación o valor por defecto cuando el dato no se registró bien.

En el crudo, 8.388.607 aparece en 33 celdas de 21 atletas (0,005 %); en 8 de esos registros el centinela se repite en dos o más benchmarks del mismo atleta, lo que refuerza un perfil incompleto más que una marca extrema. El valor

2.147.483.647 aparece en una celda (*pullups*). La Tabla 8 detalla, por variable, la frecuencia del centinela y el máximo observado inmediatamente inferior.

Tabla 8. Valores centinela de desbordamiento numérico en el dataset crudo

variable	n	%	Máx. previo	Centinela
run5k	8	0,0019	1.200.018 s	8.388.607
deadlift	5	0,0012	18.000 lb	8.388.607
fran	4	0,0009	2.400.001 s	8.388.607
backsq	4	0,0009	787.987 lb	8.388.607
fgonebad	4	0,0009	1.000.000	8.388.607
helen	2	0,0005	531.255 s	8.388.607
height	1	0,0002	7.087	8.388.607
grace	1	0,0002	5.184.000 s	8.388.607
candj	1	0,0002	250.235	8.388.607
snatch	1	0,0002	1.000.000 lb	8.388.607
run400	1	0,0002	272.220 s	8.388.607
filthy50	1	0,0002	600.039 s	8.388.607
pullups	1	0,0002	9.972.638	2.147.483.647

La fuente no documenta si esos enteros significan negativa a informar, “no aplica” o error de captura. En el pipeline se trataron como inválidos y se convirtieron a NaN, junto con otros extremos fuera de rango, sin imputar ni interpretarlos como rendimiento. Aunque son muy pocos (<0,002 % por variable), sin tratarlos dominarían los máximos del EDA y falsearían los *outliers* reales.

Tabla 9. Registros atípicos por variable (fuera del rango de validez)

Variable	n fuera	% obs.	Ejemplos
age	8	0,00 %	125 años
height	1.677	1,05 %	0 pulg; 8.388.607 pulg
weight	688	0,30 %	1 lb; 20.175 lb
fran	136	0,25 %	1 s; 8.388.607 s
helen	206	0,68 %	1 s
grace	148	0,36 %	−60 s
filthy50	153	0,79 %	1 s
fgonebad	3.285	11,05 %	0; 1; 8.388.607 pts
run400	495	2,23 %	1 s
run5k	271	0,75 %	1 s
candj	3.655	3,50 %	−45 lb
snatch	3.739	3,84 %	0 lb
deadlift	3.949	3,42 %	−500 lb
backsq	3.777	3,42 %	−7 lb
pullups	5.158	10,19 %	−6; 2.147.483.647 reps

Tabla 10. Completitud por sexo (% de valores informados dentro de cada grupo)

Variable	Hombres (%)	Mujeres (%)
age	100,00	100,00
weight	76,22	59,70
height	55,06	38,58
deadlift	40,01	27,42
backsq	38,44	26,14

Variable	Hombres (%)	Mujeres (%)
candj	36,21	24,86
snatch	34,08	22,66
fran	21,94	9,29
pullups	19,52	9,22
grace	15,75	7,38
run5k	13,67	6,93
helen	11,89	5,22
fgonebad	11,00	6,09
run400	8,92	3,57
filthy50	7,40	3,62

En todas las variables de rendimiento la completitud fue mayor en hombres que en mujeres. Ese patrón respaldó un supuesto operativo MAR y motivó cohortes separadas por sexo.

De las 27 columnas crudas, **16** entraron al pipeline principal de imputación y clusterización (COHORT_COLS): *gender*, *age*, *height*, *weight* y las de rendimiento *fran*, *helen*, *grace*, *filthy50*, *fgonebad*, *run400*, *run5k*, *candj*, *snatch*, *deadlift*, *backsq* y *pullups*. Se conservaron por ser cuantitativas (o categóricas de estratificación) y comparables entre atletas. Las 11 restantes se excluyeron según la Tabla 11.

Tabla 11. Variables excluidas del pipeline y justificación

Variable	Tipo	% NA	Motivo de exclusión
athlete_id	Identificador	0,00 %	Clave técnica sin contenido atlético; no describe rendimiento ni permite comparar perfiles.
name	Identificador personal	21,72 %	Nombre propio; identifica al individuo, no su desempeño. Incluirlo no aporta similitud física.
region	Categórica geográfica (17 niveles)	40,60 %	Describe dónde compite o se registra el atleta, no cómo rinde. Incluirla podría inducir conglomerados por ubicación (estructura trivial respecto al objetivo). Se reconoce que podría enriquecer un análisis contextual; quedó fuera del alcance del perfil de rendimiento.
team	Categórica organizacional (~4.500 niveles)	63,32 %	Nombre del equipo asociado al registro. La alta cardinalidad no es, por sí sola, causa de exclusión (podría reducirse con encodings o embeddings). Se excluyó porque el constructo es organizacional, no de capacidad física: los grupos podrían reflejar afiliación a equipos y no perfiles atléticos comparables.
affiliate	Categórica organizacional (~9.800 niveles)	42,81 %	Identifica el box afiliado a CrossFit (gimnasio/sede certificada donde entrena el atleta). Igual que team, es contexto de pertenencia; su uso orientaría la clusterización a “quién entrena dónde”, no a “quién se parece en rendimiento”. La reducción de dimensionalidad sería viable, pero cambiaría el objeto del sistema comparativo.
eat	Categórica de autorreporte (opciones de hábitos alimentarios; ~47 valores)	77,79 %	No es una métrica de rendimiento; describe hábitos declarados. Aunque existen métodos de análisis de texto/categorización, integrarlos exige un diseño distinto (NLP o tipologías de hábitos) ajeno al alcance de perfiles Benchmark/1RM. La elevada ausencia agrava la factibilidad, pero no es el argumento único.
train	Categórica de autorreporte (hábitos de entrenamiento; ~83 valores)	74,98 %	Ídem: información de práctica declarada, no de resultado medido en WOD o levantamiento. Útil para un estudio de hábitos; fuera del constructo de rendimiento del presente trabajo.
background	Categórica de autorreporte (antecedente deportivo; ~43 valores)	76,61 %	Contexto biográfico auto-reportado. Puede correlacionar con rendimiento, pero no lo mide; su inclusión mezclaría tipologías de trayectoria con perfiles de capacidad actual.
experience	Categórica de autorreporte (~84 valores)	75,19 %	Experiencia declarada (formación/cursos/vivencias CrossFit), no desempeño cuantificado. Mismo criterio de alcance.

Variable	Tipo	% NA	Motivo de exclusión
schedule	Categórica de autorreporte (~134 valores)	76,86 %	Organización del entrenamiento (frecuencia/descanso), no resultado atlético. Pertenece a un eje de hábitos, no al de benchmarks.
howlong	Categórica de autorreporte (~30 valores)	74,18 %	Antigüedad aproximada en CrossFit. Puede ser covariable interesante, pero no es una medición de rendimiento; se priorizó comparar atletas por marcas y antropometría observadas.

En síntesis, la exclusión respondió a: (i) identificadores o contexto sin medir rendimiento (*athlete_id*, *name*, *region*, *team*, *affiliate*); (ii) texto libre de alta cardinalidad en el cuestionario (*eat*, *train*, *background*, *experience*, *schedule*, *howlong*); y (iii) elevada ausencia, crítica en el cuestionario (74–78 %) y en *team* (63 %). La variable *gender* no se validó por rangos; su uso fue en el filtrado de cohortes.

Resultado de la limpieza. Tras las reglas de validez, unas 27.345 celdas pasaron a NaN (sobre todo en fuerza y rendimiento: *pullups*, *deadlift*, *backsq*, *snatch*, *candj*, *fgonebad*). El número de filas se mantuvo en 423.006: no se borraron atletas, sino valores imposibles o corruptos. Se prefirió ausencia explícita frente a retener datos que distorsionarían las distancias del clustering.

Transformaciones no aplicadas aquí. Para no sesgar temprano la estructura original, en esta fase no hubo imputación, estandarización/escalado, transformaciones logarítmicas ni reducción de dimensionalidad. Los NaN (de origen o por limpieza) se conservaron hasta la etapa de imputación, y las unidades originales no se modificaron (segundos, libras, pulgadas, puntos).

Salida. Quedó un dataset limpio de 423.006 atletas y 27 columnas, con un subconjunto analítico de 16 variables aptas para construir cohortes. Esa base alimenta la sección siguiente: definición formal de variables de análisis y construcción de las cohortes masculina, femenina y mixta por completitud mínima.

4.5 Construcción de cohortes analíticas

Con el subconjunto de 16 variables (COHORT_COLS) ya delimitado, el siguiente paso fue definir quién entra al análisis. No bastaba usar los 423.006 registros: la ausencia residual impediría distancias fiables. Por ello se construyeron tres cohortes —masculina, femenina y mixta— con criterios explícitos de completitud por atleta y por variable.

La selección combinó dos reglas. Por atleta: se contó cuántas de las 16 variables tenían valor observado; el registro era apto solo si ese conteo era ≥ k. Se buscó un k que maximizara el tamaño de la cohorte sin violar la completitud por columna (sin pretender un óptimo global exhaustivo). Por variable: en la cohorte final, cada una de las 16 columnas debía tener al menos 60 % de valores observados (a lo sumo 40 % de faltantes). Ese umbral se apoya en Aschenbruck et al. [11], quienes reportan buen comportamiento de imputaciones basadas en prototipos de clusterización con ausencias de hasta ~40 %.

En las tres cohortes se adoptó k = 14: es el umbral más inclusivo que aún garantiza ≥ 60 % de completitud en las 16 variables; al bajar a k = 13, la variable más débil cae por debajo de ese mínimo. Cada atleta incluido tiene, por tanto, información en al menos 14 de 16 variables (como máximo dos huecos por fila antes de imputar). El barrido del umbral por atleta se resume en las Tablas 12–14.

Tabla 12. Barrido del umbral k por atleta (cohorte masculina, base n = 194 926).

	n	Mín. compl.	Más débil	≥ 60 %
16	2.078	100,0 %	gender	Sí
15	5.044	85,4 %	run400	Sí
14	9.576	71,2 %	run400	Sí (elegido)
13	15.522	59,0 %	filthy50	No
12	22.848	48,1 %	filthy50	No

Tabla 13. Barrido del umbral k por atleta (cohorte femenina, base $n = 136\,128$).

	n	Mín. compl.	Más débil	≥ 60 %
16	406	100,0 %	gender	Sí
15	1.069	83,3 %	run400	Sí
14	2.183	67,4 %	run400	Sí (elegido)
13	3.827	54,9 %	run400	No
12	6.045	44,9 %	run400	No

Tabla 14. Barrido del umbral $\boldsymbol{k}$ por atleta (cohorte mixta, base $\boldsymbol{n = 331\,054}$).

	n	Mín. compl.	Más débil	≥ 60 %
16	2.484	100,0 %	gender	Sí
15	6.113	85,0 %	run400	Sí
14	11.759	70,5 %	run400	Sí (elegido)
13	19.349	59,3 %	run400	No
12	28.893	48,9 %	filthy50	No

Un patrón equivalente aparece en femenina y mixta: $k = 14$ es el mayor umbral que preserva el 60 % en todas las columnas. Así se excluyen perfiles casi vacíos (uno o dos benchmarks) sin exigir completitud total, que habría colapsado el tamaño muestral.

Resultado. La aplicación conjunta de ambos criterios produjo: 9.576 atletas (masculina), 2.183 (femenina) y 11.759 (mixta). Esas cohortes son la base de la imputación y la clusterización.

Algoritmo de búsqueda (resumen operativo). Para cada cohorte: (i) limpiar variables y filtrar por sexo; (ii) probar tamaños mínimos min_rows decrecientes

desde 20.000 en pasos de 250; (iii) para cada min_rows, evaluar *k* de 16 a 1 y retener atletas con al menos *k* variables informadas; (iv) aceptar la pareja (min_rows, *k*) solo si se cumple el tamaño mínimo y el 60 % por columna; (v) elegir un min_rows factible y, para ese umbral, el mayor *k* compatible.

Definidas las cohortes, aún queda un faltante residual (~8 % de celdas). La sección siguiente caracteriza el mecanismo de ausencia y justifica el supuesto adoptado para imputar.

4.6 Mecanismo de datos faltantes y supuesto operativo MAR

Tras armar las cohortes, aún quedó un faltante residual de unas 7,9–8,3 % de celdas en las variables numéricas, pese a que cada atleta tenía al menos 14 de 16 variables informadas. La estrategia de imputación no puede elegirse sin caracterizar el mecanismo de esas ausencias. Esta sección descarta MCAR, reconoce la plausibilidad de MNAR y adopta MAR como supuesto operativo, con sus implicaciones de diseño.

Descarte de MCAR. *Missing Completely at Random* se descartó temprano. En un dataset de autorreporte voluntario, la ausencia casi nunca es independiente del individuo ni del contexto: reportar o no un *benchmark*, un 1RM o una variable antropométrica depende de hábitos de actualización, experiencia y del tipo de prueba. Es razonable que atletas más experimentados o competitivos informen con más frecuencia variables exigentes (*snatch*, *deadlift*, ciertos WOD), mientras perfiles ocasionales muestren más huecos. Esa dependencia sistemática invalida la ausencia "completamente al azar".

Plausibilidad de MNAR. También es concebible un mecanismo *Missing Not at Random*: omitir una marca mala, no actualizar el perfil, no haber hecho la prueba, o no reportar una variable por no considerarla representativa. Ante un NaN en *snatch*, por ejemplo, coexisten escenarios distintos (vergüenza, falta de actualización, prueba no realizada, marca alta no reportada) que generan

el mismo patrón observado, pero procesos distintos. El dataset puede tener componentes MNAR, porque motivación o sesgo de autorrepresentación no están modelados en las variables observadas.

Por qué no se adoptó MNAR como hipótesis de trabajo. Bajo MNAR la ausencia depende de lo no observado: el problema deja de ser identificable solo con los datos disponibles; distintos mecanismos producen el mismo patrón. Asumirlo habría exigido hipótesis externas no verificables (p. ej. "los huecos en levantamientos son siempre marcas bajas"), y los resultados dependerían más de esas hipótesis que de lo observado.

Adopción de MAR. Por ello se adoptó *Missing at Random* como aproximación operativa: la probabilidad de ausencia se modela a partir de relaciones entre variables observadas dentro de cada cohorte, lo que habilita imputaciones alineadas con la literatura del marco teórico (capítulo 3). Esto no afirma que el mecanismo real sea estrictamente MAR; es una decisión defendible dadas las limitaciones del dataset y la necesidad de evitar supuestos no contrastables sobre lo no reportado.

Implicaciones para el diseño. Bajo MAR: (i) se descartó el análisis por casos completos como vía principal (pérdida masiva de n y sesgo hacia perfiles casi totales); (ii) se justificó imputar antes de clusterizar, en línea con Aschenbruck et al. [11]; (iii) se implementaron dos estrategias —one-step y MI *pooled*— para medir sensibilidad al completado; (iv) se conservaron las tres cohortes para ver si la estructura de ausencias y los resultados variaban por subconjunto. En la práctica, MAR permite completar huecos vía prototipos de cluster (one-step) o modelos iterativos (MI), sin inventar el valor "verdadero" no observado.

La sección siguiente describe esas estrategias de imputación y la selección de k en la etapa A.

4.7 Estrategias de imputación (one-step y MI)

Con las cohortes definidas y el supuesto MAR fijado (sección 4.6), se completaron los huecos residuales antes de la clusterización post-imputación. Imputar —en lugar de quedarse solo con casos completos— respondió a la alta ausencia del dataset crudo y a la pérdida de tamaño muestral (y el sesgo) que habría implicado eliminar atletas con información incompleta. La literatura de agrupamiento con datos incompletos privilegia imputar frente a eliminar masivamente cuando se buscan estructuras latentes [11]. Se evaluaron dos vías: (i) imputación one-step basada en clusterización y (ii) imputación múltiple con pooling (MI pooled).

Tras las cohortes, cada atleta tenía al menos 14 de 16 variables informadas, pero aún había entre 7,9 % y 8,3 % de celdas numéricas ausentes. El análisis por casos completos habría reducido mucho el tamaño muestral y sesgado la selección hacia perfiles casi totales. La imputación permitió conservar 11.759 atletas en la cohorte mixta, 9.576 en la masculina y 2.183 en la femenina, completando los huecos de forma condicionada a la información observada, en coherencia con el supuesto MAR.

La primera estrategia fue la imputación one-step clustering imputation, un esquema iterativo que alterna agrupamiento e imputación hasta estabilizar las etiquetas de cluster. El procedimiento fue el siguiente: (1) relleno provisional con la mediana global de cada columna numérica (y la moda en categóricas), solo para habilitar una primera ejecución de K-Means; (2) escalado de las variables numéricas; (3) aplicación de K-Means; (4) actualización únicamente de las celdas que eran originalmente NaN, usando la mediana (o moda) de la variable dentro del cluster asignado; (5) repetición de los pasos 3 y 4 hasta que las etiquetas dejaran de cambiar entre iteraciones consecutivas o se alcanzara un máximo de 25 iteraciones. Los valores originalmente observados nunca se sobrescribieron: la imputación se aplicó solo sobre celdas ausentes.

La selección del número de clusters en la etapa de imputación (etapa A) fue independiente de la selección de k en la clusterización post-imputación (etapa B). Para cada cohorte se realizó un barrido con k = 2, 3, …, 10 y se registraron silueta, inertia, proporción del cluster mayor, tamaño del cluster menor e iteraciones hasta convergencia (Tablas 15–18).

No se eligió el k de máxima silueta: aunque k = 2 da las siluetas más altas en las tres cohortes, concentra aproximadamente entre el 50 % y el 60 % de la cohorte en un grupo y comprime la variabilidad del relleno. El criterio fue multicriterio: silueta en meseta aceptable; evitar cluster dominante (cuando el grupo mayor supera aproximadamente el 45 %); tamaño mínimo útil para estabilizar medianas de imputación; y diversidad de tipologías dentro de esa banda de compromiso. Bajo estos criterios se adoptaron k = 7 (mixta), k = 6 (masculina) y k = 8 (femenina), resumidos en la Tabla 15.

Tabla 15. Valor de k seleccionado para imputación one-step por cohorte

Cohorte	n	k	**Sil.**	**Inertia**	**Máx. (%)**	**Mín. (n)**
Mixta	11.759	7	0,137	80.654	23,9	506
Masculina	9.576	6	0,125	75.472	26,2	719
Femenina	2.183	8	0,095	15.816	17,8	82

Tabla 16. Barrido de k — cohorte femenina ($n = 2.183$)

K	**Sil.**	**Máx. (%)**	**Mín. (n)**	**It.**	**Candidato**
2	0,211	50,2	1.087	3	No (muy grueso)
3	0,143	46,0	561	4	No
4	0,132	34,1	472	5	No
5	0,113	25,8	236	4	Marginal
6	0,101	26,2	177	6	Marginal

K	Sil.	Máx. (%)	Mín. (n)	It.	Candidato
7	0,095	20,4	135	4	Sí
8	0,095	17,8	82	6	Elegido
9	0,093	16,5	81	5	Alternativa
10	0,087	13,3	75	6	Alternativa (más fino)

Tabla 17. Barrido de k — cohorte masculina ($n = 9.576$)

	Sil.	Inertia	Máx. (%)	Mín. (n)	It.	Candidato
2	0,225	105.031	51,3	4.666	4	No (máx. sil., peor balance)
3	0,152	92.050	46,5	2.137	5	No
4	0,145	84.308	32,8	1.726	4	No
5	0,123	78.953	28,0	777	5	Cercano
6	0,125	75.472	26,2	719	5	Elegido
7	0,108	72.188	25,5	31	7	No (micro-clusters)
8	0,096	69.643	20,5	33	6	No
9	0,088	67.397	17,7	31	5	No
10	0,089	65.190	18,0	11	17	No

Tabla 18. Barrido de k — cohorte mixta ($n = 11.759$)

	Sil.	Máx. (%)	Mín. (n)	It.	Candidato
2	0,271	59,9	4.716	3	No
3	0,199	41,3	2.316	4	No
4	0,156	38,0	2.027	5	No
5	0,140	33,0	785	4	Marginal
6	0,140	25,5	656	7	Alternativa

	Sil.	Máx. (%)	Mín. (n)	It.	Candidato
7	0,137	23,9	506	6	Elegido
8	0,124	21,8	39	5	No (micro-clusters)
9	0,114	19,1	40	5	No
10	0,112	15,8	37	18	No

La lectura conjunta de los barridos confirma un patrón común: la silueta es máxima en valores bajos de k, pero asociada a un cluster dominante; al aumentar k, la silueta desciende y el reparto mejora hasta una zona intermedia; en valores altos aparecen microgrupos (por ejemplo, en hombres, mínimos de 31–11 atletas) poco útiles para medianas de imputación.

Como estrategia de contraste se implementó imputación múltiple, aproximando un enfoque MICE (Multiple Imputation by Chained Equations) [13]. El procedimiento consistió en: generar cinco réplicas imputadas independientes, cada una con semilla distinta; modelar las relaciones entre variables observadas para estimar los valores faltantes (usando la moda en categóricas cuando correspondía); y agregar las cinco réplicas mediante pooling por promedio en las columnas numéricas. A diferencia del one-step, la imputación múltiple no incorporó clusterización durante el relleno: el agrupamiento se aplicó únicamente después, sobre los datasets ya completos.

Ambas estrategias se aplicaron sobre las tres cohortes. La Tabla 19 verifica la completitud de los datasets resultantes (0 % de valores faltantes residuales en las variables de análisis), condición necesaria para algoritmos que no admiten ausencias. La Tabla 20 no busca declarar un método superior por proximidad numérica celda a celda: one-step y MI parten de lógicas distintas (medianas por cluster frente a modelos iterativos con réplicas), por lo que es esperable que las celdas originalmente ausentes difieran entre estrategias. Lo que aporta MI en esta

etapa, y que one-step no ofrece por sí solo, es una cuantificación de incertidumbre entre las cinco réplicas. En cambio, one-step produce un completado determinista en una ejecución reproducible. La comparación de si ambas estrategias conducen a estructuras de agrupamiento y perfiles análogos se evalúa en la clusterización post-imputación (etapa B).

Tabla 19. Completitud tras imputación (ambos métodos)

Cohorte	n	% NA antes	% NA OS	% NA MI
Mixta	11.759	7,93	0,0	0,0
Masculina	9.576	7,85	0,0	0,0
Femenina	2.183	8,28	0,0	0,0

Tabla 20 Contraste entre métodos (solo celdas que eran NA

Cohorte	k OS	Diff. med.	Disp. MI
Mixta	7	26,16	54,08
Masculina	6	25,55	53,84
Femenina	8	20,10	55,75

Los resultados mostraron que ambas estrategias alcanzaron 0 % de valores faltantes tras la imputación. No obstante, los valores imputados no coincidieron celda a celda: la diferencia media absoluta en las celdas que eran originalmente NaN se situó entre 59 y 68 unidades en la escala mixta de las variables, con una diferencia mediana entre 20 y 26 unidades. Esta divergencia era esperable, dado que one-step condiciona la imputación a prototipos de cluster, mientras que MI modela relaciones entre variables sin agrupamiento previo.

A pesar de esas diferencias a nivel celda, el análisis posterior de perfiles (etapa

B) evidenció que ambos métodos producían estructuras ampliamente comparables en la cohorte masculina, con medianas de separación entre clusters prácticamente equivalentes (ratio de separación 2,42 para one-step frente a 2,46 para MI pooled). En consecuencia, la imputación múltiple funcionó como validación de sensibilidad y no como contradicción del enfoque principal.

Con los datasets ya completos, la sección siguiente aborda la clusterización post-imputación (etapa B) y el diseño de los 18 escenarios analíticos.

4.8 Clusterización post-imputación y diseño de 18 escenarios

Con los datasets completos de la etapa de imputación (sección 4.7), se aplicó la clusterización post-imputación —etapa B— para identificar perfiles atléticos interpretables. A diferencia de la etapa A, en la que K-Means participó en el relleno de faltantes, la etapa B no modifica los datos: solo particiona a los atletas según su desempeño en las variables de cohorte.

El objetivo no fue maximizar una métrica de separación aislada, sino obtener agrupaciones con balance razonable, tamaños mínimos defendibles y perfiles útiles para el sistema comparativo. Operativamente, la clusterización se entiende así: dado un atleta, se identifica el conjunto de atletas similares (su cluster) y, a continuación, se le compara con ese mismo grupo de referencia para estimar fortalezas y debilidades relativas. Es decir, construye el vecindario comparable; el sistema usa ese vecindario como base de contraste, no a toda la población ni a un ideal absoluto.

La clusterización intervino en dos momentos del pipeline con propósitos distintos, resumidos en la Tabla 21.

Tabla 21. Roles de K-Means y de k en las etapas A y B

Aspecto	Etapa A (imputación)	Etapa B (perfiles)
Entrada	Cohortes con NA	Datasets imputados (0 % NA)
Rol de K-Means	Imputar por mediana de cluster	Asignar etiquetas de perfil
Selección de k	Estabilidad de la imputación	Interpretabilidad de perfiles

Los tres algoritmos de la etapa B compartieron el mismo preprocesamiento, para que las diferencias se deban al criterio de agrupamiento y no a una preparación distinta. Se usaron las 16 variables de análisis: gender, age, height, weight y los doce indicadores de rendimiento. En las cohortes masculina y femenina, gender se excluyó del espacio de clusterización por ser constante; en la mixta se incorporó con codificación categórica. Las quince variables numéricas se estandarizaron (media 0, varianza 1) sobre cada cohorte ya imputada, porque K-Means y Ward usan distancia euclídea y, sin escalado, variables de mayor magnitud dominarían el espacio. No se aplicó reducción de dimensionalidad, a fin de preservar la interpretabilidad en unidades atléticas. En la mixta, gender se representó con indicadores Male/Female. Aunque los datasets imputados debían tener 0 % de NaN, se mantuvo una salvaguarda: cualquier residual numérico se completaba con la mediana global y las categóricas con la moda.

Se diseñó una malla experimental de tres dimensiones: cohortes mixta (11.759), masculina (9.576) y femenina (2.183); imputación one-step y MI pooled; algoritmos K-Means, Agglomerative Ward y GMM. Ello produjo 18 combinaciones (3 × 2 × 3), cada una con barrido de k = 2, 3, …, 10 sobre el dataset imputado correspondiente, con métricas de calidad, balance y, cuando correspondía, datasets

etiquetados.

Como línea base se empleó K-Means, por su coherencia con la imputación one-step, su eficiencia en cohortes de miles de atletas y la interpretabilidad de los centroides. El procedimiento asigna cada atleta al centroide más cercano y actualiza centroides hasta convergencia (fundamento en el marco teórico), con múltiples inicializaciones. Como sensibilidad se usó Agglomerative con enlace de Ward (fusión que minimiza el incremento de varianza intra-grupo; sin dependencia de semilla de centroides) y GMM (componentes gaussianas más flexibles; asignación dura; BIC/AIC como apoyo), también sobre la misma matriz y el mismo barrido de k.

En la cohorte masculina con one-step, el patrón fue común a los tres algoritmos: la silueta máxima aparece en k bajos, pero con un cluster dominante (aproximadamente entre 52 % y 75 %), demasiado grueso para un comparativo de perfiles; al subir k mejora el reparto hasta una zona intermedia; en k altos aparecen microgrupos (en K-Means, mínimos de 31–11 atletas; en Ward, grupos de hasta 7). Por ello se adoptó k = 6 como valor de evaluación transversal en la línea base: silueta en meseta (aproximadamente 0,134), cluster mayor acotado (aproximadamente 26 %) y mínimo aún estable (719 atletas), sin maximizar silueta. Un patrón análogo se observó en mixta y femenina. El detalle numérico de los barridos de los 18 escenarios se presenta en el Anexo A; la justificación del modelo operativo se desarrolla en el capítulo de resultados.

La sección siguiente fija las métricas de comparación, el corte transversal en k = 6 (Tablas 41A–41C) y el enfoque para describir perfiles.

4.9 Evaluación, corte en k = 6 y caracterización de perfiles

Para cada combinación escenario–k se calcularon: coeficiente de silueta; inertia (solo K-Means); BIC y AIC (solo GMM); balance (proporción del grupo mayor y tamaño del menor); índice de Rand ajustado (ARI) al variar k; y un ranking compuesto que combina silueta, preferencia por k más altos dentro de una banda

útil, tamaño mínimo y penalización de grupos dominantes. Las definiciones formales están en el marco teórico; aquí se usan como criterios operativos.

Sobre cada cohorte imputada se barre k = 2…10 para los tres algoritmos. El objetivo no fue maximizar una sola métrica, sino encontrar valores de k con separación razonable, balance y utilidad para perfiles, evitando el k de máxima silueta cuando produce un grupo dominante y evitando microgrupos. Los barridos completos de los 18 escenarios (antes Tablas 22–40) están en el Anexo A. En este capítulo se conserva el resumen transversal en k = 6 (Tablas 41A–41C).

En todos los escenarios el patrón cualitativo es el mismo: silueta máxima en k bajos con cluster dominante; al subir k mejora el reparto hasta una zona intermedia; en k altos aparecen microgrupos o un balance poco útil. Sobre esa evidencia se fijó k = 6 como punto de comparación transversal entre cohortes, imputaciones y algoritmos: no porque maximice la silueta en todos los casos, sino porque, en la línea base (K-Means), combina silueta en meseta, cluster mayor acotado y tamaño mínimo aún estable. Las Tablas 41A–41C resumen los 18 escenarios exactamente en ese k.

Tabla 41A — Cohorte masculina (*n* = 9 576; *k* = 6)

Imputación	Algoritmo	Silhouette	Menor (n)	Mayor (%)
One-step	K-Means	0,134	719	26,2
MI pooled	K-Means	0,127	646	22,6
One-step	Ward	0,084	436	36,6
MI pooled	Ward	0,087	664	31,0
One-step	GMM	0,060	226	39,7
MI pooled	GMM	0,035	230	27,6

Modelo operativo: one-step + K-Means

Tabla 41B — Cohorte femenina (*n* = 2 183; *k* = 6)

Imputación	Algoritmo	Silhouette	Menor (n)	Mayor (%)
One-step	K-Means	0,130	87	29,2
MI pooled	K-Means	0,121	169	26,3
One-step	Ward	0,087	135	32,2
MI pooled	Ward	0,088	129	25,8
One-step	GMM	0,035	38	29,5
MI pooled	GMM	0,050	37	29,3

Tabla 41C — Cohorte mixta (*n* = 11 759; *k* = 6)

Imputación	Algoritmo	Silhouette	Menor (n)	Mayor (%)
One-step	K-Means	0,162	940	24,8
MI pooled	K-Means	0,160	740	25,3
One-step	Ward	0,123	1.301	31,8
MI pooled	Ward	0,121	917	35,5
One-step	GMM	0,119	125	32,6
MI pooled	GMM	0,124	121	35,4

Con el mismo k de evaluación se compararon K-Means, Ward y GMM bajo one-step y MI pooled, usando silueta, balance y, en GMM, criterios de información. K-Means fue la línea base; Ward y GMM, análisis de sensibilidad.

Tras obtener una partición, los perfiles se describen en unidades originales con estadísticos agregados por cluster —en particular, medianas por variable—, complementados con medidas de separación relativa y con la distribución intra-cluster de variables clave. La interpretación atlética detallada del modelo retenido se desarrolla en el capítulo de resultados y validación.

5. RESULTADOS, VALIDACIÓN Y CARACTERIZACIÓN DE PERFILES

5.1 Resultados de la construcción de cohortes

A continuación, se resumen los valores seleccionados para cada cohorte.

Tabla 42. Cohortes analíticas obtenidas

Cohorte	n	k mín.	Filas mín.	NA (%)
Mixta	11.759	14/16	11.750	7,9
Masculina	9.576	14/16	9.500	7,9
Femenina	2.183	14/16	2.000	8,3

Los resultados mostraron que las tres cohortes compartieron el mismo umbral de completitud individual (k = 14), pero difirieron en tamaño muestral según el universo disponible por sexo. La cohorte mixta alcanzó el mayor número de atletas (11 759), seguida por la cohorte masculina (9 576) y la femenina (2 183). En el caso de las mujeres, el algoritmo debió reducir el umbral min_rows hasta 2 000 para mantener simultáneamente el criterio de completitud por columna y un tamaño cohorte mínimamente viable.

Dentro de las cohortes finales, la variable que más limitó el cumplimiento del criterio del 60 % de completitud fue run400, con aproximadamente 70–71 % de valores presentes en las cohortes masculina y mixta. No obstante, todas las columnas superaron el umbral mínimo exigido.

5.2 Selección de la cohorte masculina como núcleo principal

Aunque las tres cohortes fueron conservadas para análisis comparativos y de sensibilidad, la cohorte masculina fue seleccionada como núcleo principal del sistema comparativo. Esta decisión respondió a un mejor equilibrio entre tamaño

muestral, homogeneidad poblacional e interpretabilidad de los perfiles generados.

La cohorte mixta, aunque más numerosa, presentaba el riesgo metodológico de que la variable gender dominara estructuras de agrupamiento triviales, dada la diferencia sistemática de rendimiento entre hombres y mujeres en variables de fuerza y potencia. Por su parte, la cohorte femenina resultó metodológicamente válida, pero con un tamaño considerablemente menor, lo que limitaba la estabilidad de perfiles y la granularidad del análisis.

En consecuencia, la cohorte masculina ofreció aproximadamente 9 576 atletas con reglas de completitud equivalentes a las otras cohortes, sin mezclar diferencias biológicas y de rendimiento asociadas al sexo. Los resultados obtenidos sobre las cohortes mixta y femenina se reportan en el capítulo de resultados como evidencia complementaria de sensibilidad metodológica.

5.3 Lectura transversal de los 18 escenarios y decisiones del modelo operativo

Las Tablas 41A–41C del capítulo 4 resume los 18 escenarios analíticos en $k = 6$, corte transversal motivado por los barridos $k = 2, \dots, 10$. A partir de esa evidencia se formalizan las decisiones de cohorte, imputación y algoritmo.

En cuanto a la cohorte, la masculina ($n = 9\,576$) ofrece un tamaño suficiente para perfiles estables —del orden de diez mil atletas— sin incurrir en el riesgo de la cohorte mixta, donde la variable de sexo puede inducir separaciones triviales por diferencias sistemáticas de rendimiento en fuerza y potencia. La cohorte femenina es metodológicamente válida y sus barridos se documentaron en el capítulo 4, pero su menor n y la aparición de clusters pequeños en varios escenarios (p. ej., mínimos de 37–87 atletas en $k = 6$) limitan la granularidad y la estabilidad del comparativo. Por ello se adoptó la cohorte masculina como núcleo del modelo operativo, conservando mixto y mujeres como evidencia de sensibilidad.

En cuanto a la imputación, one-step y MI pooled producen métricas cercanas

dentro de cada cohorte–algoritmo (p. ej., en hombres con K-Means: siluetas 0,134 frente a 0,127; en mixto: 0,162 frente a 0,160). Ello indica que la estructura de agrupamiento no depende de forma crítica del método de completado. One-step se privilegia por su menor coste computacional y su carácter determinista en ejecuciones reproducibles, frente al esquema estocástico de múltiples réplicas; MI se conserva como contraste compatible con el supuesto MAR. La equivalencia sustantiva entre ambos se confirma, además, en la caracterización de perfiles de las secciones siguientes.

Respecto al algoritmo, K-Means obtiene de forma sistemática mejor silueta y, en la cohorte masculina, mejor balance que Ward y GMM en $k = 6$ (silueta ≈ 0,134; cluster mayor ≈ 26 %; cluster menor = 719 atletas). Ward tiende a concentrar más masa en el cluster mayor; GMM registra siluetas más bajas y, en varios escenarios, clusters pequeños. Por ello se adoptó K-Means como línea operativa del prototipo, y Ward y GMM como análisis de sensibilidad.

En conjunto, el núcleo productivo se sitúa en cohorte masculina + imputación one-step + K-Means + $k = 6$. Las secciones siguientes detallan la elección independiente de k en las etapas A y B, y presentan la partición resultante.

5.4 Selección de *k*: etapa A (imputación) y etapa B (perfiles)

El número de clusters k se eligió de forma independiente en dos momentos del pipeline, porque responden a preguntas distintas:

Tabla 43. Separación de k en imputación y clusterización.

Etapa	Pregunta	Rol de k
A — Imputación one-step	¿Con cuántos grupos se rellenan de forma estable los valores faltantes?	Define los prototipos (medianas) usados solo para imputar
B — Clusterización post-imputación	¿En cuántos grupos se describen los atletas ya con datos completos?	Define los perfiles del comparativo

No se exigió que ambos k coincidieran. En la cohorte masculina, tras aplicar en cada etapa el mismo tipo de criterios (silueta relativa, balance y rechazo de microgrupos), ambos barridos condujeron **a** $k = 6$. Esa coincidencia es una convergencia empírica, no un requisito metodológico. Además, la partición generada durante la imputación (etapa A) **no** se reutilizó como mapa de perfiles: los perfiles productivos corresponden exclusivamente a la partición de la etapa B.

[5.4.1] k en la etapa A (imputación, cohorte masculina)

En el barrido one-step de hombres ($k = 2 \dots 10$), $k = 2$ maximizó la silueta pero concentró más de la mitad de la cohorte en un solo grupo, comprimiendo la variabilidad del relleno. Valores $k \geq 7$ generaron microgrupos (31–11 atletas), inadecuados para medianas de imputación estables. $k = 6$ equilibró silueta (~0,125), cluster mayor (~26 %) y tamaño mínimo (719 atletas), y fue el valor adoptado para completar la cohorte masculina.

[5.4.2] k en la etapa B (clusterización post-imputación)

En la etapa B el k ya no define prototipos de imputación, sino el número de perfiles del comparativo. Sobre la cohorte masculina ya completada, y con K-Means

como línea base, se volvió a aplicar un criterio multicriterio: no se adoptó el k de máxima silueta. Aunque k = 2 alcanzó la silueta más alta (0,241), concentró el 51,7 % de los atletas en un solo grupo, lo que resulta demasiado grueso para un sistema comparativo. En el otro extremo, a partir de k ≈ 9 aparecieron microgrupos (31–11 atletas), poco útiles para interpretar medianas de referencia estables.

En la zona intermedia (k = 5–7) la silueta se mantuvo en meseta (aprox. 0,134–0,144), con un cluster mayor acotado y un tamaño mínimo aún operativo. El ranking compuesto situó a k = 6 en primer lugar (score 0,675), por delante de k = 5 (0,657): seis perfiles aportan granularidad adicional sin caer en microgrupos. Por ello se adoptó k = 6 como número de perfiles del modelo operativo (hombres + one-step + K-Means).

Esta elección es independiente del k de la etapa A. La coincidencia k = 6 en ambas etapas, en la cohorte masculina, es una convergencia empírica, no un requisito metodológico. Además, la partición de la imputación no se reutilizó como mapa de perfiles: los perfiles del sistema corresponden solo a la partición de la etapa B.

El mismo k = 6 se usó además como punto de comparación transversal entre los 18 escenarios (cohortes × imputación × algoritmo), no porque maximice la silueta en todos ellos, sino para contrastarlos bajo un k común ya justificado en la línea base.

Por ello se adoptó $k = 6$ como número de perfiles del modelo operativo.

5.5 Marco de decisión en cascada (resumen)

Una vez aclaradas ambas elecciones de k, la selección del modelo final se resume así:

Tabla 44. Secuencia de decisiones del pipeline comparativo.

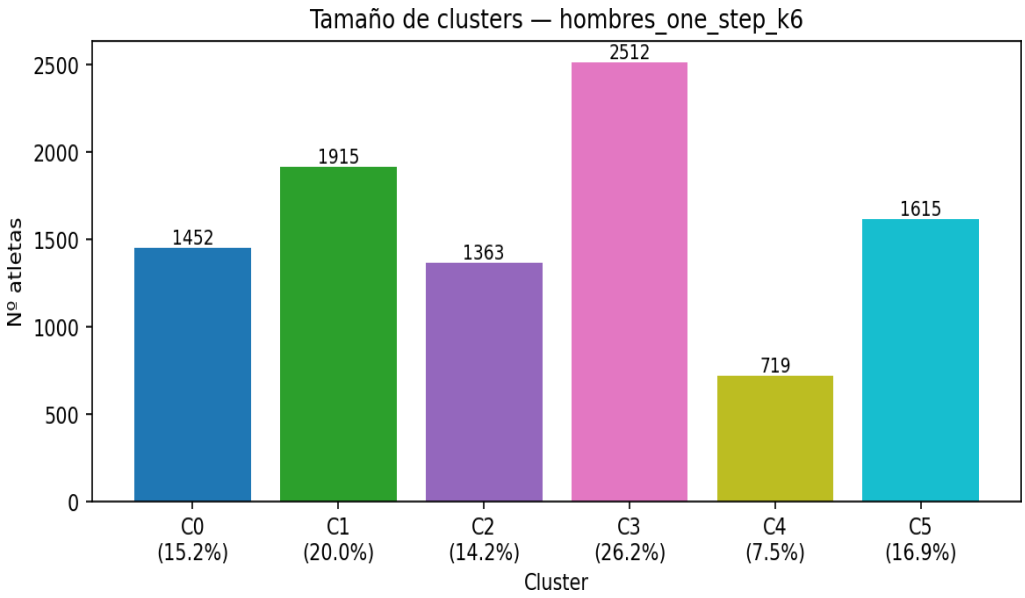
Orden	Decisión	Alternativas evaluadas	Selección
1	Cohorte analítica	Mixta, masculina, femenina	Masculina (9.576 atletas)
2	Estrategia de imputación	One-step, MI pooled	One-step (k=6 en etapa A)
3	Algoritmo de clusterización	K-Means, Agglomerative Ward, GMM	K-Means (etapa B)
4	Número de clusters (perfiles)	Barrido k=2…10	k=6 en etapa B

5.6 Resultados de la clusterización

Con el modelo operativo ya definido —cohorte masculina, imputación one-step y K-Means con $k = 6$—, esta sección presenta los resultados de la partición obtenida. Se reporta, en primer lugar, la composición de los seis grupos (tamaño y proporción sobre la cohorte) y, a continuación, las figuras y tablas que resumen el comportamiento de las variables entre clusters. La interpretación detallada de cada perfil se desarrolla en los apartados siguientes.

La Figura A muestra el número de atletas asignados a cada uno de los seis clusters (C0–C5) y su proporción respecto a la cohorte masculina ($n = 9.576$). El grupo más numeroso es C3 (2.512 atletas; 26,2 %) y el más pequeño es C4 (719 atletas; 7,5 %). Todos los clusters se sitúan dentro de un rango de participación compatible con un comparativo de perfiles (entre ~7,5 % y ~26,2 %), sin un cluster dominante por encima del 35 % ni microgrupos de decenas de atletas.

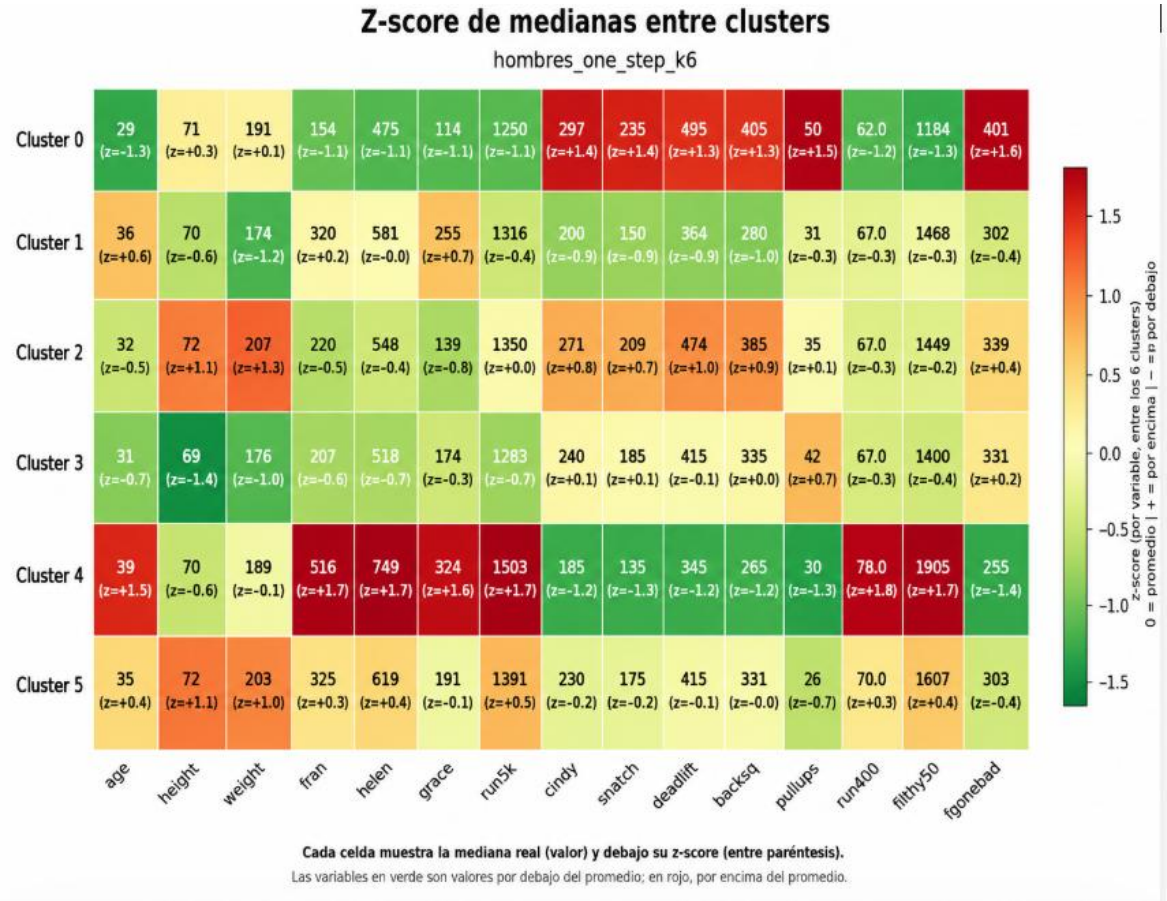
Figura 1. Distribución del tamaño de los clusters (cohorte masculina, one-step, K-Means



La Figura 1 muestra la distribución del tamaño de los seis clusters obtenidos en la cohorte masculina con imputación one-step y K-Means ($k = 6, n = 9.576$). Cada barra corresponde a un cluster (C0–C5); la altura indica el número de atletas asignados a ese grupo y la etiqueta sobre la barra reporta, además, la proporción respecto al total de la cohorte. El grupo más numeroso es C3 (2.512 atletas; 26,2 %) y el más pequeño es C4 (719 atletas; 7,5 %). El resto se sitúa entre ~14 % y ~20 % (C0: 15,2 %; C1: 20,0 %; C2: 14,2 %; C5: 16,9 %). Esta composición indica una partición utilizable para un comparativo de perfiles: ningún cluster concentra más del ~26 % de la cohorte y no aparecen microgrupos de decenas de atletas.

La Figura 2 informa únicamente cuántos atletas hay en cada grupo. Para interpretar diferencias atléticas hace falta examinar cómo se comportan las variables entre clusters. La Figura B presenta, para cada variable, la mediana por cluster y su posición relativa (z-score) respecto a los demás grupos, lo que permite identificar fortalezas y debilidades relativas de cada perfil.

Figura 2. Z score medianas entre clusters



La Figura 2 resume el perfil de cada cluster: en cada celda se reporta la **mediana en unidades originales** y, entre paréntesis, el **z-score** de esa mediana respecto a la distribución de las seis medianas del mismo indicador. El color refleja el z-score. En variables de tiempo, medianas más bajas (y z más negativos) indican, en general, mejor rendimiento; en levantamientos, *pullups* y *fgonebad*, medianas más altas (y z más positivos) indican mejor rendimiento. La figura no muestra el tamaño de cada grupo —ya presentado en la Figura A—, sino las diferencias de perfil entre clusters. Sobre esta evidencia se construye la interpretación de los ejes y de cada etiqueta.

5.7 Distribución de atletas por perfil

A partir de la composición de la partición (Figura A) y del contraste multivariado entre clusters (Figura 2), la Tabla 21 resume el reparto de la cohorte masculina ($n = 9.576$) y asigna a cada grupo una etiqueta interpretativa, justificada en los apartados siguientes.

Tabla 45. Reparto por cluster — cohorte masculina, one-step, K-Means, k=6

Cluster	n atletas	% cohorte	Etiqueta interpretativa
0	1.452	15,2 %	Alto rendimiento equilibrado
1	1.915	20,0 %	Intermedio ligero
2	1.363	14,2 %	Potencia y fuerza elevada
3	2.512	26,2 %	Perfil central
4	719	7,5 %	Perfil en consolidación
5	1.615	16,9 %	Masa elevada, rendimiento intermedio

5.8 Marco interpretativo

5.8.1 Primer eje — rendimiento global (C0 frente a C4). C0 concentra el mejor desempeño relativo en el conjunto de pruebas: tiempos más bajos en fran, helen, grace, run5k, run400 y filthy50, y valores más altos en candj, snatch, deadlift, backsq, pullups y fgonebad. C4 ocupa el extremo opuesto: tiempos más altos y cargas/repeticiones más bajas en esas mismas variables. Ese contraste define el gradiente de rendimiento más claro de la partición.

5.8.2 Segundo eje — variación morfofuncional (masa, fuerza y eficiencia).

Sobre ese gradiente aparece una segunda fuente de diferenciación, no idéntica al "mejor vs. peor" global. Se asocia especialmente a:

masa corporal: weight (y, en menor medida, height);

fuerza absoluta: candj, snatch, deadlift, backsq;

eficiencia en pruebas metabólicas y gimnásticas: tiempos de WOD (fran, helen, grace, filthy50) y pullups.

En este segundo eje, el contraste más nítido es C2 frente a C3: C2 tiende a mayor masa y mayor fuerza absoluta, mientras C3 se sitúa más cerca del centro de la cohorte, con menor peso relativo y un perfil más equilibrado. C1 y C5 ocupan posiciones intermedias en ese mismo eje (C1 más ligero y con fuerza relativa menor; C5 con masa elevada y rendimiento intermedio), sin alcanzar los extremos de C0–C4 ni el contraste C2–C3.

En resumen: C0 vs. C4 ordena a los atletas por nivel global de rendimiento; C2 vs. C3 (con C1 y C5 en el medio) ordena por composición corporal y tipo de fortaleza, no solo por "más rápido / más lento".

5.8.3 Eje de rendimiento global

Cluster 0 — Alto rendimiento equilibrado (n = 1.452; 15,2 %)

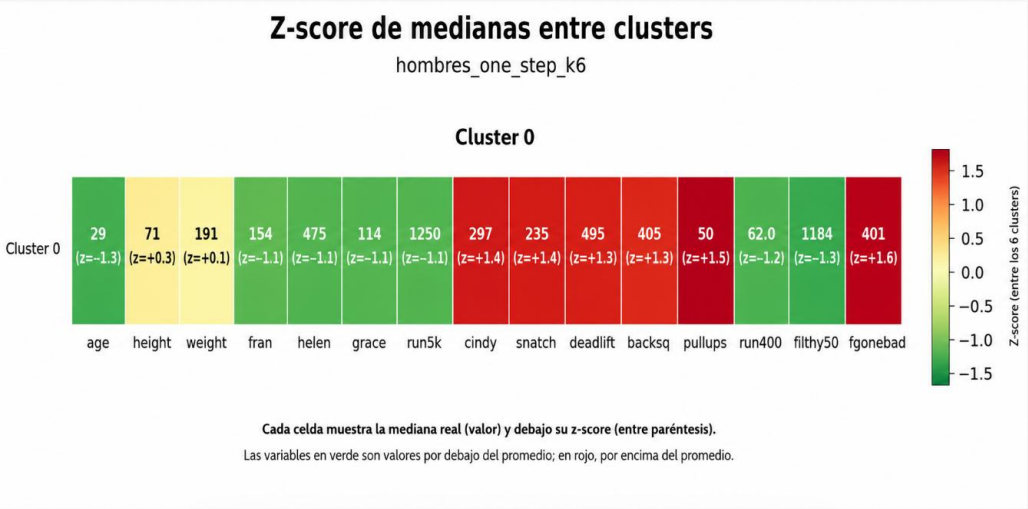


Figura 3 Detalle de medianas y z-scores del Cluster 0 (extraído de la Figura 2).

Como se observa en la Figura B.0, C0 combina, de forma simultánea, buen desempeño en las pruebas de tiempo y en las de fuerza. En las variables de tiempo —fran, helen, grace, run5k, run400 y filthy50— presenta medianas relativas bajas (z

negativos), lo que indica tiempos más rápidos que el centro de los seis clusters. En las variables de levantamiento —candj, snatch, deadlift y backsq— presenta medianas relativas altas (z positivos), junto con valores elevados en pullups y fgonebad. Así, el perfil integra fuerza absoluta elevada, buen desempeño metabólico y alta capacidad gimnástica, sin especializarse en una sola dimensión.

Su masa corporal es intermedia (191 lb): no es el grupo más pesado (C2: 207 lb) ni el más liviano (C1: 174 lb; C3: 176 lb). Representa el perfil más cercano a atletas avanzados o competitivos dentro del dataset.

5.8.4 Cluster 4 — Rezagado global (n = 719; 7,5 %) C4

Concentra el patrón opuesto en la Figura B: tiempos relativos altos y fuerza/gimnasia relativas bajas, además de la mayor edad mediana (39 años).

Tabla 46. Medianas del cluster 0 (C0)

Variable	Mediana C0
fran	154 s
helen	475 s
run400	62 s
deadlift	495 lb
pullups	50
age	29 años
weight	191 lb

Tabla 47. Medianas del cluster 4 (C4) y contraste con C0

Variable	Mediana C4	Contraste con C0
fran	516 s	+362 s
helen	749 s	+274 s
filthy50	1.905 s	+721 s
run400	78 s	+16 s
pullups	20	−30 reps
deadlift	345 lb	−150 lb
age	39 años	+10 años

Sugiere un perfil globalmente menos competitivo, en el que pueden converger mayor edad relativa y menor acondicionamiento. La oposición C0–C4 constituye el eje principal de la estructura.

Cluster 1 — Intermedio desbalanceado (n = 1.915; 20,0 %) Segundo grupo más numeroso. Presenta metcon (**condicionamiento metabólico**; p. ej. fran = 320 s) relativamente lento (fran = 320 s), cargas moderadas-bajas (deadlift = 364 lb), bajo peso corporal (174 lb) y pullups = 31. No sobresale de forma marcada en ninguna dimensión; se interpreta como un perfil recreativo o híbrido sin especialización clara.

Cluster 5 — Masa elevada con rendimiento intermedio (n = 1.615; 16,9 %) Peso elevado (203 lb) con rendimiento intermedio (fran = 325 s, deadlift = 415 lb, pullups = 26). Se sitúa entre el núcleo (C3) y los extremos C0/C4, sin alcanzar la fuerza de C2 ni la eficiencia metabólica de C3. En la Figura B se distingue de C2 por un perfil de fuerza y metcon menos favorable a masa corporal similar.

5.8.5 Eje morfofuncional: masa corporal y eficiencia

Cluster 2 — Pesado con alta fuerza ($n = 1.363$; 14,2 %)

Tabla 48. Medianas del cluster 2 (C2)

Variable	Mediana C2
height	72 in
weight	207 lb
deadlift	474 lb
fran	220 s
filthy50	1.449 s

Atletas más altos y pesados, con buena fuerza absoluta, pero sin la eficiencia metabólica de C0 o C3. Corresponde a un patrón de masa corporal funcional orientada a la fuerza.

5.8.6 Cluster 3 — Núcleo estructural ($n = 2.512$; 26,2 %)

tabla 49. Medianas del cluster 3 (C3)

Variable	Mediana C3
weight	176 lb
fran	207 s
deadlift	415 lb
pullups	42
run400	67 s
helen	518 s

Es el grupo más numeroso. Con menor masa que C2, muestra mejor equilibrio relativo entre metcon y fuerza para ese rango de peso, y concentra el perfil "típico" de la cohorte.

Tabla 50. Comparación C2 vs C3

Variable	C2 (207 lb)	C3 (176 lb)	Ventaja	Lectura
deadlift	474 lb	415 lb	C2	Fuerza absoluta
grace	139 s	174 s	C2	Estímulo de C&J
fran	220 s	207 s	C3	Metcon / gimnasia
helen	548 s	518 s	C3	Carrera + swings + pull-ups
run5k	1.350 s	1.283 s	C3	Carrera
pullups	35	42	C3	Fuerza relativa
filthy50	1.449 s	1.400 s	C3	WOD largo

La diferencia de peso mediano es de ~31 lb (~14 kg). Sin inferir causalidad, el patrón es compatible con que mayor masa acompañe mayor fuerza absoluta y, a la vez, penalice pruebas cíclicas o de fuerza relativa.

5.9 Interpretación transversal

C2 y C3 ilustran la tensión fuerza–eficiencia por debajo de C0. El cluster más competitivo (C0) no es el más pesado ni el más liviano (191 lb) y, aun así, combina fuerza alta, metcon rápido y buena gimnasia, lo que sugiere un equilibrio funcional no reducido a la masa corporal sola.

5.10 Variables discriminantes

Tabla 51. Principales variables discriminantes entre clusters

Orden	Variable	Ratio de separación	Rango de medias entre clusters
1	fran	3,09	158 – 541 s
2	fgonebad	2,82	245 – 408

Orden	Variable	Ratio de separación	Rango de medias entre clusters
3	helen	2,73	474 – 797 s
4	filthy50	2,73	1.143 – 2.056 s
5	candj	2,71	183 – 298 lb
6	snatch	2,66	136 – 236 lb
7	grace	2,63	116 – 378 s
8	pullups	2,42	19 – 53
9	backsq	2,34	267 – 414 lb
10	deadlift	2,28	343 – 492 lb

Las variables de metcon (*fran*, *helen*, *filthy50*) y de fuerza olímpica (*candj*, *snatch*) encabezan la separación entre perfiles, en coherencia con el contraste C0–C4 observado en la Figura B.

5.11 Resumen de hallazgos

En la cohorte masculina ($n = 9.576$), con imputación one-step y K-Means ($k = 6$), se obtuvieron seis perfiles de tamaño estable (entre 7,5 % y 26,2 % de la cohorte), sin clusters dominantes ni microgrupos.

El contraste principal es de rendimiento global: el cluster 0 (élite equilibrada) concentra el mejor desempeño conjunto en fuerza, metcon y gimnasia, mientras que el cluster 4 (rezagado global) concentra el peor, con mayor edad mediana. Entre ambos se sitúan perfiles intermedios (clusters 1 y 5).

Un segundo patrón es morfofuncional: el cluster 2 (más pesado y fuerte en términos absolutos) se diferencia del cluster 3 (más liviano y eficiente en pruebas metabólicas y de fuerza relativa), que además concentra el mayor volumen de atletas (núcleo estructural). El cluster más competitivo (C0) no coincide con el más pesado ni con el más liviano, lo que sugiere un equilibrio entre masa, fuerza y eficiencia.

Las variables que más separan los grupos son, sobre todo, benchmarks metabólicos (fran, helen, filthy50) y levantamientos olímpicos (candj, snatch), en coherencia con esa doble lectura de rendimiento global y compromiso fuerza–eficiencia.

6. IMPLEMENTACIÓN DEL SISTEMA COMPARATIVO

Este capítulo describe la implementación del prototipo asociado al modelo operativo definido en el capítulo 5 (cohorte masculina, imputación one-step, K-Means con $k = 6$). El énfasis no está en reanalizar los perfiles atléticos, sino en documentar la arquitectura del software, la organización del código, la reproducibilidad del pipeline y el funcionamiento del módulo que permite comparar a un atleta frente al perfil de su grupo.

6.1 Objetivo y alcance del prototipo

El prototipo tiene como objetivo permitir el registro de un atleta de la cohorte masculina, completar sus valores no informados, asignarlo a uno de los seis grupos comparables (clusters) del modelo operativo y contrastar su rendimiento con la mediana de su grupo. De este modo, el sistema traduce la partición en una comparación relativa, no en una etiqueta definitiva del deportista.

El alcance del prototipo v1 se limita a:

atletas interpretados bajo el espacio de variables de la cohorte masculina;

asignación a los clusters 0–5 del modelo hombres_one_step_k6;

comparación variable a variable frente a las medianas del cluster asignado;

operación local mediante una interfaz web servida por un servidor en Python.

Quedan fuera de este alcance el reentrenamiento interactivo del modelo, la extensión automática a cohortes femenina o mixta, y la sustitución del corte productivo por Agglomerative o GMM.

6.2 Arquitectura general

El flujo de ejecución del prototipo se organiza en cuatro pasos:

Entrada: el usuario suministra valores disponibles del atleta (antropometría, levantamientos y *benchmarks*).

Completado de faltantes: las variables no informadas se estiman mediante imputación múltiple ajustada sobre la cohorte de referencia con valores ausentes.

Asignación de cluster: el vector ya completo se proyecta al mismo espacio de características del modelo productivo y se asigna a un cluster mediante K-Means con $k = 6$, entrenado sobre el dataset masculino imputado por one-step.

Comparación: cada variable observada o imputada se contrasta con la mediana del cluster asignado, generando fortalezas, debilidades y textos descriptivos del perfil.

Esta separación es deliberada: el mapa de grupos comparables proviene del corte one-step + K-Means del capítulo 7, mientras que la imputación del atleta nuevo usa MI por conveniencia operativa (completa un único registro sin reejecutar el ciclo iterativo one-step). La coherencia se mantiene porque la asignación ocurre en el mismo espacio y con el mismo modelo de clusters del corte productivo.

6.3 Requerimientos del sistema

Los requerimientos se derivaron del objetivo general del proyecto —aproximar un sistema comparativo basado en clusterización para atletas de CrossFit— y del modelo operativo adoptado en el capítulo 5 (cohorte masculina, one-step, K-Means con $k = 6$). Se organizaron en requerimientos funcionales (RF) y no funcionales (RNF).

Tabla 52. Requerimientos funcionales

ID	Requerimiento	Descripción
RF-01	Ingreso de datos del atleta	El usuario ingresa nombre, identificador y hasta 15 variables numéricas de antropometría y rendimiento. No es obligatorio completar todas; se aceptan registros parciales.
RF-02	Validación de cohorte	El prototipo opera sobre el modelo masculino: el sexo informado debe corresponder a la cohorte de hombres.
RF-03	Imputación de valores faltantes	Ante variables no informadas, el sistema estima valores mediante imputación múltiple ajustada sobre la cohorte masculina. Los valores ingresados por el usuario no se sobrescriben.
RF-04	Asignación a perfil (cluster)	Tras la imputación, el atleta se asigna al cluster K-Means más cercano (k=6) en el espacio escalado del modelo productivo.
RF-05	Descripción del perfil	El sistema presenta el título y la descripción narrativa del cluster asignado (perfiles C0–C5), alineada con el análisis del capítulo 7.
RF-06	Comparación vs medianas del cluster	Para cada variable con valor (observado o imputado), se compara al atleta con la mediana de su cluster e identifica fortalezas y debilidades relativas (tiempos: menor es mejor; cargas/reps: mayor es mejor).
RF-07	Vista previa sin registro	El usuario puede ejecutar la asignación y comparación sin persistir el resultado.
RF-08	Registro de atletas	El usuario puede guardar el resultado de la evaluación en un registro persistente.
RF-09	Consulta de atletas registrados	El sistema permite buscar un atleta por nombre o identificador y listar los registros guardados.
RF-10	Catálogo de perfiles	los seis clusters con sus descripciones interpretativas

Tabla 53. Requerimientos no funcionales

ID	Requerimiento	Descripción
RNF-01	Ejecución local	El prototipo opera como aplicación web local, sin dependencia de servicios en la nube.
RNF-02	Reproducibilidad de asignación	La asignación de clusters utiliza semilla fija y múltiples inicializaciones, alineada con el pipeline analítico.
RNF-03	Trazabilidad	Cada registro incluye metadatos: variables proporcionadas vs imputadas, método de imputación, cluster asignado y marca temporal.
RNF-04	Independencia de frameworks web pesados	El servidor HTTP se implementó con la biblioteca estándar de Python, minimizando dependencias externas.
RNF-05	Tiempo de respuesta aceptable	La primera ejecución carga y prepara los modelos en memoria; las peticiones siguientes reutilizan esa carga para reducir latencia.

6.4 Organización del repositorio

El código se estructuró de forma modular, separando el pipeline analítico del módulo de producto:

Tabla 54. Componentes del repositorio y su rol

Componente	Rol
`src/`	Lógica reutilizable (preprocesamiento, imputación, clustering, motor del comparativo)
`scripts/`	Puntos de entrada para regenerar artefactos y levantar el prototipo
`data/raw/`	Dataset original
`data/processed/`	Cohortes, datasets imputados, etiquetados y métricas
`data/comparativo/`	Persistencia de atletas registrados por el prototipo
`notebooks/`	Exploración y documentación interactiva del análisis
`docs/`	Documentación metodológica y de referencia

Dentro del módulo comparativo, los componentes principales son:

el motor de imputación del atleta nuevo, asignación de cluster y registro;

el catálogo de perfiles (textos asociados a C0–C5);

el cálculo de medianas por cluster para la comparación;

el servidor HTTP y los archivos estáticos de la interfaz.

[8.5] Pipeline reproducible

Para favorecer la reproducción de resultados, el pipeline se ejecutó mediante scripts con parámetros documentados y semillas fijas en los algoritmos estocásticos. La secuencia lógica de regeneración es:

limpieza y construcción de cohortes;

imputación one-step (y, como contraste, MI pooled);

clusterización post-imputación y exportación del dataset etiquetado;

generación de descriptivos por cluster (medianas y métricas de separación);

arranque del módulo comparativo sobre los artefactos ya producidos.

Los artefactos de entrada mínimos del prototipo son:

el dataset masculino imputado por one-step (base del ajuste de K-Means);

el dataset etiquetado hombres_one_step_k6_labeled.csv;

la tabla descriptiva de medianas por cluster;

el dataset crudo (o cohorte con ausencias) para ajustar el imputador del atleta

nuevo.

El código fuente del pipeline analítico y del prototipo comparativo está disponible en el repositorio: https://github.com/GermanIriarte/clusterizacionPerfilesRendimientoCrossfit

6.5 Modelo operativo embebido en el prototipo

El prototipo incorpora de forma fija el escenario declarado en el capítulo 7:

Tabla 55. Configuración del modelo productivo

Elemento	**Configuración**
Cohorte	Masculina (n=9.576)
Imputación del mapa de perfiles	One-step (k=6 en etapa A)
Algoritmo de asignación	K-Means
Número de perfiles	k=6 (etiquetas 0–5)
Identificador	`hombres_one_step_k6`

La descripción textual de cada perfil en la interfaz se alinea con la caracterización del capítulo 7 (élite equilibrada, intermedio desbalanceado, pesado con alta fuerza, núcleo estructural, rezagado global y masa elevada con rendimiento intermedio).

6.6 Módulo comparativo: asignación y contraste

Tras asignar el cluster, el sistema compara cada variable con la mediana del grupo:

En variables de tiempo (*fran*, *helen*, etc.), un valor **por debajo** de la mediana se interpreta como fortaleza; en cargas y repeticiones (*deadlift*, *snatch*, *pullups*, etc.), un valor **por encima** de la mediana se interpreta como fortaleza; en edad,

altura y peso se reporta la posición relativa respecto a la mediana, sin etiqueta de fortaleza/debilidad.

La salida incluye el cluster asignado, el resumen del perfil, las medianas de referencia, y las listas de fortalezas y debilidades. El usuario puede ejecutar una **vista previa** (asignar sin guardar) o un **registro** persistente del atleta.

Por defecto, la imputación del atleta nuevo usa una variante rápida de MI; de forma opcional puede solicitarse el esquema de varias réplicas con *pooling*, a costa de mayor tiempo de cómputo.

[8.8] Forma de uso

El prototipo se ejecuta en local mediante un servidor Python que expone la interfaz en el navegador. La aplicación no está diseñada para abrirse como archivo estático aislado: requiere el servidor activo para invocar el pipeline de imputación y asignación.

La interfaz permite:

- consultar el catálogo de perfiles;
- ingresar o completar datos de un atleta;
- previsualizar la asignación;
- registrar atletas y recuperar registros previos.

6.7 Limitaciones técnicas

Entre las limitaciones del prototipo se destacan:

opera sobre el espacio masculino del modelo v1; no generaliza automáticamente a mujeres o a la cohorte mixta;

la imputación del atleta nuevo (MI) no es idéntica al procedimiento one-step

usado para construir el mapa de perfiles, aunque la asignación utiliza el mismo modelo de clusters;

la calidad de la comparación depende de cuántas variables informe el usuario: a mayor ausencia, mayor peso de la imputación;

el sistema no actualiza en línea los centroides ni reestima k; cualquier cambio de modelo exige regenerar artefactos del pipeline;

la interpretación de fortalezas/debilidades es relativa al cluster asignado, no una predicción de rendimiento competitivo futuro.

7. CONCLUSIONES

El presente trabajo aproximó un sistema comparativo para atletas de CrossFit basado en técnicas de clusterización, a partir de un dataset real de registros autodeclarados y de un pipeline analítico reproducible. A continuación se sintetizan los principales hallazgos, el grado de cumplimiento de los objetivos y las limitaciones que condicionan la interpretación de los resultados.

7.1 Cumplimiento de los objetivos

En conjunto, el proyecto cumplió su objetivo general de aproximar un sistema comparativo basado en clusterización. En particular:

Se caracterizó un dataset real de atletas de CrossFit, documentando su estructura, la elevada proporción de valores faltantes, la presencia de registros inconsistentes y las implicaciones de estas limitaciones para el diseño metodológico.

Se diseñó y ejecutó un pipeline reproducible de preprocesamiento, construcción de cohortes, imputación y clusterización, aplicado de forma sistemática a las cohortes mixta, masculina y femenina.

Se compararon estrategias de imputación (one-step y MI pooled) y familias de algoritmos de agrupamiento (K-Means, Agglomerative Ward y GMM), y se justificó la selección del modelo operativo con criterios multidimensionales (separación, balance, estabilidad e interpretabilidad), no únicamente con la maximización de una métrica aislada.

Se identificaron e interpretaron seis perfiles atléticos en la cohorte masculina bajo un marco de dos ejes de variación: un eje principal de rendimiento global (contraste entre el perfil de élite y el rezagado) y un eje morfofuncional asociado a la tensión entre masa corporal, fuerza absoluta y eficiencia metabólica o gimnástica.

Se entregó un prototipo de software que traduce los resultados analíticos en

una herramienta de asignación de perfil y de comparación de un atleta frente a las medianas de su grupo.

7.2 Hallazgos principales

En la cohorte masculina ($n = 9.576$), con imputación one-step y K-Means ($k = 6$), los seis clusters presentaron tamaños estables y diferencias interpretables en variables de fuerza y benchmarks. El contraste más claro opone un perfil de alto desempeño multivariado a un perfil globalmente rezagado; entre ambos emergen tipologías intermedias y, de manera relevante, la distinción entre un grupo más pesado y fuerte en términos absolutos y un núcleo más liviano y eficiente en pruebas cíclicas. Las variables más discriminantes fueron, en lo esencial, indicadores metabólicos y levantamientos olímpicos, coherentes con esa doble lectura.

La equivalencia práctica entre one-step y MI pooled en la estructura de perfiles, junto con el peor balance o menor utilidad de Agglomerative y GMM en el corte evaluado, respaldó la adopción de one-step y K-Means como línea operativa del prototipo, conservando las demás alternativas como análisis de sensibilidad.

7.3 Limitaciones del estudio

Los resultados deben interpretarse a la luz de las siguientes limitaciones:

El dataset proviene de registros voluntarios en una plataforma pública y no constituye una muestra representativa de la población de practicantes de CrossFit.

La clusterización opera sobre variables observadas y no captura factores como experiencia de entrenamiento, técnica, composición corporal ni nivel competitivo formal.

El sistema comparativo v1 opera exclusivamente sobre la cohorte masculina; las cohortes mixta y femenina se conservaron como análisis de sensibilidad.

Las interpretaciones sobre la relación entre peso corporal y rendimiento son

exploratorias y no implican causalidad.

El prototipo no fue evaluado con usuarios finales en un entorno de producción.

La imputación en tiempo de ejecución del comparativo (MI) difiere de la imputación usada para entrenar el mapa de perfiles (one-step), aunque la equivalencia de estructuras en $k = 6$ fue contrastada en el análisis.

En síntesis, el trabajo aporta una aproximación metodológica y un prototipo funcional para la comparación de atletas de CrossFit mediante perfiles derivados de clusterización, con evidencia reproducible y con límites explícitos que orientan las líneas de continuación del estudio.

7.4 Trabajos futuros

A partir de los resultados y limitaciones identificadas, se proponen las siguientes líneas de continuación:

7.4.1 Extensión a otras cohortes

Desarrollar modelos de clusterización y perfiles específicos para la cohorte femenina (n = 2 183), evaluando si conviene relajar criterios de completitud o ajustar k para compensar el menor tamaño muestral. Explorar estrategias de estratificación por sexo en el comparativo, evitando mezclar perfiles masculinos y femeninos en una misma asignación.

7.4.2 Enriquecimiento del dataset y variables

Incorporar variables no disponibles en el dataset actual: composición corporal, años de entrenamiento, frecuencia semanal, resultados de competencias o datos de Open/Games cuando las políticas de acceso lo permitan. Evaluar si nuevas dimensiones modifican la estructura de clusters o refuerzan los dos ejes interpretativos identificados (rendimiento global y morfofuncional).

7.4.3 Validación causal y de dominio

Contrastar las hipótesis exploratorias —especialmente la tensión entre masa corporal y eficiencia metabólica observada en C2 vs C3— con estudios que incluyan medidas de composición corporal y evaluación por entrenadores o atletas. Realizar validación cualitativa de las etiquetas de perfil con expertos en CrossFit.

7.4.4 Mejoras del sistema comparativo

Desplegar el prototipo en un servidor accesible para pruebas con usuarios reales.

Referencias

[1] CrossFit, Inc. (2019). *Guía de entrenamiento de nivel 1 de CrossFit* (4.ª ed.). CrossFit, Inc. https://library.crossfit.com/free/pdf/CFJ_Level1_Spanish_Latin_American.pdf

[2] Meier, N., Schlie, J., & Schmidt, A. (2023). CrossFit®: 'Unknowable' or Predictable?—A Systematic Review on Predictors of CrossFit® Performance. *Sports*, *11*(6), 112. https://doi.org/10.3390/sports11060112

[3]Mangine, G. T., Tankersley, J. E., McDougle, J. M., et al. (2020). Predictors of CrossFit Open Performance. Sports, 8(7), 102.

[4] Russell, S., & Norvig, P. (2021). *Artificial Intelligence: A Modern Approach* (4th ed.). Pearson.

[5] Mitchell, T. M. (1997). *Machine Learning*. McGraw-Hill.

[6] Hastie, T., Tibshirani, R., & Friedman, J. (2021). *The Elements of Statistical Learning: Data Mining, Inference, and Prediction* (2nd ed., corr. printing). Springer.

[7] Tan, P.-N., Steinbach, M., Karpatne, A., & Kumar, V. (2019). *Introduction to Data Mining* (2nd ed.). Pearson.

[8] Fernández, J., Casals, M., Oliver, J., Plensa, J., & Manisera, M. (2024). *Reporting of Clustering Techniques in Sports Sciences: A Scoping Review*. *Journal of Sports Sciences*.

[9] Jain, A. K. (2010). *Data Clustering: 50 Years Beyond K-Means*. *Pattern Recognition Letters*, 31(8), 651–666.

[10] Little, R. J. A., & Rubin, D. B. (2019). *Statistical Analysis with Missing Data* (3rd ed.). Wiley.

[11] van Buuren, S. (2018). *Flexible Imputation of Missing Data* (2nd ed.). Chapman & Hall/CRC.

[12] D'Urso, P., De Giovanni, L., & Vitale, V. (2023). A Robust Fuzzy C-Medoids Clustering Model for Mixed Data with Noise Cluster: An Application to Soccer Players' Performance Analysis. Mathematics, 11.

[13] Akhanli, S., & Hennig, C. (2020). Using Cluster Validation Indexes with Custom Dissimilarities: A Soccer Player Clustering Study. Journal of Classification, 37(3), 503–543. Rubin, D. B. (1976). *Inference and Missing Data*. *Biometrika*, 63(3), 581–592.

ANEXOS

Anexo A. Barridos de *k* por escenario analítico

Este anexo documenta los barridos de $k = 2…10$ correspondientes a las 18 combinaciones de cohorte, estrategia de imputación y algoritmo de clusterización. Su propósito es permitir la verificación numérica de las métricas de separación y balance. La síntesis interpretativa y las decisiones del modelo operativo se desarrollan en los capítulos 4 y 5 (Tablas 41A–41C y sección 5.3).

Tabla A1. K-Means (hombres, one-step)

	Sil.	Inertia	Mín. (n)	Máx. (%)
2	0,241	104.544	4.623	51,7
3	0,167	92.984	2.000	46,0
4	0,161	85.049	1.637	32,4
5	0,144	79.024	771	28,2
6	0,134	75.471	719	26,2
7	0,137	72.399	31	26,1
8	0,138	70.130	11	26,1
9	0,108	68.227	11	20,1
10	0,096	66.945	31	14,6

Tabla A2. Agglomerative Ward (hombres, one-step)

	Sil.	Mín. (n)	Máx. (%)
2	0,228	3.524	63,2
3	0,112	2.543	36,8
4	0,115	1.329	36,6
5	0,109	436	36,6
6	0,084	436	36,6
7	0,063	436	19,0

	Sil.	Mín. (n)	Máx. (%)
8	0,066	7	19,0
9	0,067	7	19,0
10	0,054	7	18,4

Tabla A3. GMM (hombres, one-step)

	Sil.	BIC	AIC	Mín. (n)	Máx. (%)
2	0,220	283.159	281.217	2.391	75,0
3	0,144	274.659	271.742	608	47,0
4	0,088	272.268	268.377	248	42,0
5	0,069	271.283	266.417	264	42,1
6	0,060	271.102	265.261	226	39,7
7	0,029	270.896	264.080	190	30,9
8	0,027	271.102	263.311	91	29,1
9	0,024	271.267	262.502	129	31,3
10	0,008	271.543	261.803	126	25,5

Tabla A4. — Comparación en $k = 6$ (hombres)

Algoritmo	Imp.	Sil.	Máx. (%)	Mín. (n)
K-Means	OS	0,134	26,2	719
K-Means	MI	0,127	22,6	646
Ward	OS	0,084	36,6	436
Ward	MI	0,087	31,0	664
GMM	OS	0,060	39,7	226
GMM	MI	0,035	27,6	230

Tabla A5. — K-Means (mixta, one-step, n=11.759)

	Sil.	Inertia	Mín. (n)	Máx. (%)
2	0,287	124.023	4.688	60,1
3	0,216	107.473	2.318	42,7
4	0,195	97.479	1.292	40,7
5	0,163	90.205	1.025	35,5
6	0,162	84.620	940	24,8
7	0,162	80.654	506	23,9
8	0,164	77.889	12	23,9
9	0,151	75.897	35	21,4
10	0,136	74.009	35	17,5

Tabla A6. — Agglomerative Ward (mixta, one-step, n=11.759)

	Sil.	Mín. (n)	Máx. (%)
2	0,308	2.701	77,0
3	0,164	2.173	58,6
4	0,151	1.301	58,6
5	0,121	1.301	44,0
6	0,123	1.301	31,8
7	0,130	581	31,8
8	0,133	43	31,8
9	0,134	7	31,8
10	0,119	7	31,8

Tabla A7. — GMM (mixta, one-step, n=11.759)

k	Sil.	BIC	AIC	Mín. (n)	Máx. (%)
2	0,303	87.063	84.549	2.456	79,1
3	0,218	60.357	56.582	2.190	59,3
4	0,159	32.531	27.495	625	38,3
5	0,125	30.934	24.638	134	36,6
6	0,119	29.625	22.068	125	32,6
7	0,108	28.091	19.274	80	33,7
8	0,066	27.965	17.887	77	27,9
9	0,065	28.136	16.797	72	25,6
10	0,033	25.250	12.651	69	20,3

Tabla A8. — K-Means (mixta, MI pooled, n=11.759)

k	Sil.	Inertia	Mín. (n)	Máx. (%)
2	0,289	123.141	4.725	59,8
3	0,220	106.104	2.337	42,9
4	0,198	96.262	1.282	39,2
5	0,158	89.413	906	29,9
6	0,160	84.181	740	25,3
7	0,144	80.617	846	23,3
8	0,140	77.728	132	22,7
9	0,131	75.190	140	18,3
10	0,130	72.968	133	15,6

Tabla A9. — Agglomerative Ward (mixta, MI pooled, n=11.759)

k	Silhouette	Cluster menor (n)	Cluster mayor (%)
2	0,203	4.169	64,5 %
3	0,189	2.250	45,4 %
4	0,130	2.250	35,5 %
5	0,121	917	35,5 %
6	0,121	917	35,5 %
7	0,083	917	21,6 %
8	0,086	193	21,6 %
9	0,087	7	21,6 %
10	0,080	7	21,4 %

Tabla A10. — GMM (mixta, MI pooled, n=11.759)

k	Silhouette	BIC	AIC	Cluster menor (n)	Cluster mayor (%)
2	0,303	78.308	75.794	2.344	80,1 %
3	0,237	33.832	30.057	2.070	63,8 %
4	0,167	26.047	21.012	471	41,1 %
5	0,138	24.675	18.379	124	37,6 %
6	0,124	23.619	16.063	121	35,4 %
7	0,090	21.682	12.865	266	29,3 %
8	0,096	22.529	12.451	162	36,1 %
9	0,066	22.785	11.446	35	26,2 %
10	0,065	23.194	10.595	68	30,0 %

Tabla A11. — K-Means (femenina, one-step, n=2.183)

k	Silhouette	Inertia	Cluster menor (n)	Cluster mayor (%)
2	0,228	23.768	1.085	50,3 %
3	0,156	20.965	578	44,1 %
4	0,150	19.192	467	33,8 %
5	0,136	18.133	163	28,1 %
6	0,130	17.294	87	29,2 %
7	0,125	16.459	82	28,0 %
8	0,120	15.816	82	17,8 %
9	0,110	15.440	82	16,6 %
10	0,107	15.074	81	14,0 %

Tabla A12. — Agglomerative Ward (femenina, one-step, n=2.183)

k	Silhouette	Cluster menor (n)	Cluster mayor (%)
2	0,182	650	70,2 %
3	0,131	337	54,8 %
4	0,116	337	32,2 %
5	0,094	135	32,2 %
6	0,087	135	32,2 %
7	0,088	68	32,2 %
8	0,057	68	32,2 %
9	0,056	68	20,2 %
10	0,049	68	14,6 %

Tabla A13. — GMM (femenina, one-step, n=2.183)

k	Silhouette	BIC	AIC	Cluster menor (n)	Cluster mayor (%)
2	0,186	68.749	67.207	708	67,6 %
3	0,104	68.313	65.998	362	49,3 %
4	0,081	68.740	65.651	278	36,3 %
5	0,060	69.116	65.254	34	36,9 %
6	0,035	69.476	64.840	38	29,5 %
7	0,043	70.085	64.675	23	30,4 %
8	0,027	70.736	64.553	22	26,9 %
9	0,015	67.203	60.246	21	23,7 %

Tabla A14. — K-Means (femenina, MI pooled, n=2.183)

k	Silhouette	Inertia	Cluster menor (n)	Cluster mayor (%)
2	0,233	23.541	1.078	50,6 %
3	0,158	20.740	586	44,5 %
4	0,151	18.951	484	32,0 %
5	0,134	17.841	205	26,3 %
6	0,121	17.091	169	26,3 %
7	0,110	16.441	109	24,3 %
8	0,110	15.847	108	17,2 %
9	0,109	15.389	86	15,7 %
10	0,104	14.982	85	13,0 %

Tabla A15. — Agglomerative Ward (femenina, MI pooled, n=2.183)

k	Sil.	Mín. (n)	Máx. (%)
2	0,210	693	68,3
3	0,129	374	51,1
4	0,117	353	35,0
5	0,112	129	35,0
6	0,088	129	25,8
7	0,069	129	25,8
8	0,061	129	23,8
9	0,047	129	18,0
10	0,048	82	18,0

Tabla A16. — GMM (femenina, MI pooled, n=2.183)

k	Silhouette	BIC	AIC	Cluster menor (n)	Cluster mayor (%)
2	0,185	67.453	65.911	688	68,5 %
3	0,109	67.206	64.890	384	51,5 %
4	0,079	67.576	64.487	286	36,7 %
5	0,060	68.032	64.170	124	34,5 %
6	0,050	68.385	63.749	37	29,3 %
7	0,035	69.121	63.712	49	27,6 %
8	0,036	69.762	63.579	21	29,6 %
9	0,021	70.500	63.543	56	23,5 %
10	0,026	71.133	63.402	20	23,5 %

Tabla A17. — K-Means (hombres, MI pooled, n=9.576)

k	Silhouette	Inertia	Cluster menor (n)	Cluster mayor (%)
2	0,247	103.667	4.422	53,8 %
3	0,169	91.527	2.109	44,6 %
4	0,168	83.578	1.640	32,8 %
5	0,141	78.237	792	28,4 %
6	0,127	75.037	646	22,6 %
7	0,127	71.999	44	22,3 %
8	0,117	69.654	41	21,3 %
9	0,106	67.530	37	17,7 %
10	0,105	65.838	10	15,8 %

Tabla A18. — Agglomerative Ward (hombres, MI pooled, n=9.576)

k	Silhouette	Cluster menor (n)	Cluster mayor (%)
2	0,195	4.156	56,6 %
3	0,147	2.399	43,4 %
4	0,138	839	43,4 %
5	0,094	839	31,0 %
6	0,087	664	31,0 %
7	0,089	52	31,0 %
8	0,060	52	22,8 %
9	0,062	7	22,8 %
10	0,054	7	22,8 %

Tabla A19. — GMM (hombres, MI pooled, n=9.576)

k	Silhouette	BIC	AIC	Cluster menor (n)	Cluster mayor (%)
2	0,233	278.288	276.345	2.059	78,5 %
3	0,152	270.200	267.283	562	49,0 %
4	0,144	269.911	266.019	428	48,0 %
5	0,075	267.346	262.479	264	41,5 %
6	0,035	267.166	261.325	230	27,6 %
7	0,031	267.301	260.486	190	27,8 %
8	0,020	267.762	259.971	177	25,3 %
9	0,000	267.999	259.233	93	19,1 %
10	0,006	268.301	258.561	10	18,7 %