

Materiales y Métodos

Diseño y Datos: Se llevó a cabo un estudio observacional de datos biométricos recolectados mediante *wearables* (Apple Watch). A diferencia del diseño transversal originalmente propuesto (30 días, múltiples participantes), el enfoque ejecutado aprovechó datos longitudinales de un **cohorte piloto de 10 adultos** (5 mujeres, 5 hombres) con seguimiento multianual. Cada participante exportó sus registros de Apple Health (Apple Watch) incluyendo métricas diarias de actividad física y variables cardíacas. En total se analizaron **1,385 semanas-persona**, de las cuales 1,337 cumplieron los criterios de calidad tras filtrados. Este cambio de alcance respondió a consideraciones prácticas y permitió incorporar mayor resolución temporal por sujeto, manteniendo la coherencia con la **justificación clínica** (evaluación objetiva del sedentarismo y variabilidad fisiológica individual) y con los **objetivos específicos** redefinidos hacia la clasificación de sedentarismo y su validación interna.

Pipeline Metodológico: El procesamiento y análisis de los datos se implementó en **cinco fases secuenciales**: **(1)** Preprocesamiento diario de datos, **(2)** Agregación semanal de métricas, **(3)** Clustering no supervisado (*verdad operativa*), **(4)** Desarrollo del sistema de inferencia difusa, y **(5)** Validación del modelo. A continuación, se detalla cada fase y su fundamento técnico:

- 1. Preprocesamiento diario de datos:** Se aplicó una limpieza e imputación de los registros diarios priorizando la integridad y evitando sesgos temporales. Para los valores faltantes o intervalos sin registro por ausencia del dispositivo (*no-wear*), se implementó una **imputación jerárquica con gates** fisiológicos: no se imputa en casos de no-wear total (datos ausentes durante horas completas); en periodos de actividad extremadamente baja se asumieron valores basales (por ejemplo, usando la frecuencia cardíaca de reposo como estimación de FC durante inactividad); y en condiciones normales con pequeñas brechas de datos se aplicó **mediana móvil unidireccional** (utilizando solo datos previos del mismo usuario) para rellenar valores, evitando así cualquier fuga de información futura (*leakage*). Adicionalmente, se llevaron a cabo filtrados de calidad y **winsorización** de outliers a nivel mensual (limitando valores extremos al percentil 1 y 99) para mitigar la influencia de datos atípicos sin distorsionar las distribuciones reales. En esta fase también se **derivaron variables clave diarias** a partir de las mediciones brutas, con el fin de normalizar diferencias individuales: por ejemplo, la **Actividad_relativa** definida como el cociente entre minutos en movimiento y el tiempo de uso efectivo del dispositivo por día (minutos de movimiento / (60×horas monitorizadas)), la cual ajusta la actividad física por la exposición al wearable; y el **Superávit_calórico_basal**, calculado como (calorías activas diarias × 100) dividido entre la Tasa Metabólica Basal estimada (ecuación de Mifflin–St Jeor según sexo, peso, talla y edad). Este último indicador expresa el gasto calórico relativo al metabolismo basal en porcentaje, permitiendo comparaciones equitativas entre individuos de distinta antropometría (por ejemplo, 400 kcal tienen implicaciones distintas en un varón de mayor peso vs. una mujer petite, por lo que porcentualizar el gasto mejora la comparación). Estas transformaciones garantizan mediciones **objetivas y comparables** de la actividad/sedentarismo diarios, reduciendo sesgos por diferente tiempo de monitoreo o características fisiológicas entre participantes, acorde con la motivación de obtener indicadores normalizados clínicamente relevantes.

2. **Agregación semanal de métricas:** Definida la semana-calendario como unidad de análisis, se resumieron los datos diarios imputados en **indicadores semanales robustos**. Para cada participante y semana se calcularon **estadísticos resistentes** a valores extremos: principalmente la mediana (p50) y el rango intercuartílico (IQR) de las distribuciones diarias de cada variable de interés. Esta agregación semanal cumple un doble propósito: (a) reducir el ruido e inestabilidad de mediciones diarias aisladas (p. ej., variaciones puntuales de un día atípico) enfocando el análisis en patrones de comportamiento sostenido, y (b) alinear la temporalidad con la perspectiva clínica, dado que las recomendaciones de actividad/sedentarismo suelen evaluarse en bloques semanales. **Entre las variables semanales calculadas se incluyeron las derivadas normalizadas antes descritas – Actividad_relativa_p50, Superávit_calórico_basal_p50** (medianas semanales) junto con sus IQR-, así como métricas cardiovasculares semanales: la **HRV_SDNN_p50** (mediana semanal de la variabilidad de frecuencia cardíaca, indicador del tono autonómico) con su IQR, y las frecuencias cardíacas de reposo y en actividad. Con estas últimas se definió una variable compuesta **Delta_cardiaco** por semana, calculada como la diferencia entre la FC al caminar vs. la FC de reposo medianas de esa semana ($FC_{caminata_p50} - FC_{r_p50}$). Este delta resume la **carga cardíaca relativa al esfuerzo** físico semanal. El resultado de la fase de agregación fue un conjunto de datos semanal consolidado (1 fila = 1 semana de un sujeto) con aproximadamente una decena de *features* por semana, que sirvió como base para las fases de modelado subsecuentes.
3. **Clustering no supervisado (verdad operativa):** Para obtener una referencia objetiva del nivel de sedentarismo sin depender de umbrales predefinidos, se realizó una **clusterización K-Means** sobre el espacio de características semanales. Se exploraron diferentes números de clúster (un **K-sweep** de 2 a 6) evaluando la calidad de cada partición mediante el coeficiente de **Silhouette** y la estabilidad de los grupos (consistencia en múltiples inicializaciones y comparación de asignaciones). El **mejor agrupamiento** resultó ser de **K=2** clústeres, con un Silhouette promedio moderado (≈ 0.23) pero estable, indicando la presencia de **dos patrones distintos de comportamiento semanal**. Al examinar los centroides, un clúster (aprox. 30% de las semanas) presentó mayor actividad relativa y mayor superávit calórico porcentual, mientras el otro clúster ($\sim 70\%$ de las semanas) mostró indicadores opuestos (menor actividad, menor gasto relativo y generalmente menor HRV). Estos grupos fueron interpretados, respectivamente, como **“Bajo Sedentarismo”** versus **“Alto Sedentarismo”** semanales. Cabe destacar que esta distribución estuvo sesgada hacia el sedentarismo alto (consistentemente, más semanas caen en el grupo de conducta sedentaria elevada), reflejando quizá la realidad de una población mayoritariamente inactiva. Para fines del estudio, la asignación de cada semana a **clúster 0 (bajo sedentarismo)** o **clúster 1 (alto sedentarismo)** se consideró como **“verdad operativa”** o referencia *data-driven* contra la cual validar el modelo difuso. En otras palabras, los rótulos binarios provenientes del clustering no supervisado funcionan como criterio objetivo obtenido desde los datos mismos, capturando la distinción fundamental entre semanas activas vs. sedentarias de forma libre de sesgo humano.
4. **Sistema de inferencia difusa (modelo Mamdani):** Con la clasificación no supervisada como contexto, se procedió a diseñar un **sistema de lógica difusa** capaz de estimar el nivel de sedentarismo semanal de forma transparente e interpretable. Se optó por un modelo difuso tipo **Mamdani**, compuesto por un conjunto de reglas lingüísticas SI-ENTONCES que imitan el razonamiento de un experto, combinando las variables continuas de entrada en una salida categórica difusa. **Entradas:** Se seleccionaron **cuatro features semanales** como insumos del sistema, representativas de diferentes dimensiones fisiológicas del sedentarismo: (a) **Actividad_relativa** (mediana semanal) – porcentaje de tiempo monitorizado en movimiento, (b)

Superávit_calórico_basal (mediana semanal) – porcentaje de gasto calórico activo relativo al metabolismo basal, (c) **HRV_SDNN** (mediana semanal) – variabilidad de frecuencia cardíaca como indicador de balance autonómico, y (d) **Delta_cardiaco** (mediana semanal) – diferencial de frecuencia cardíaca esfuerzo vs. reposo como proxy de la respuesta cardiovascular al ejercicio. Cada una de estas entradas se **fuzzificó** mediante **tres conjuntos difusos** o etiquetas lingüísticas (**Bajo, Medio, Alto**): se definieron funciones de pertenencia de forma triangular basadas en percentiles de la distribución muestral (aproximadamente p10–p25–p40 para el conjunto “Bajo”, p35–p50–p65 para “Medio” y p60–p75–p90 para “Alto”, con traslapes suaves), ajustando cada rango según la **direccionalidad clínica** de la variable. Esto significa que, por ejemplo, en variables donde un valor más alto implica menor riesgo (como actividad física o HRV), la pertenencia al conjunto “Alto” corresponde a valores numéricos elevados de dicha métrica; inversamente, en variables donde valores altos implican mayor carga o peor condición (ej. Delta_cardiaco, que aumenta cuando el sujeto tiene una respuesta desproporcionada de FC al caminar), se diseñaron funciones de pertenencia donde “Alto” cubre rangos altos que representan mayor sedentarismo. **Base de Reglas Difusas:** Con las entradas fuzzificadas, se construyó una base de conocimiento de **5 reglas difusas** del tipo “**SI** \<premisas> **ENTONCES** \<conclusión>”, formuladas en conjunto con criterio experto y apoyadas en los hallazgos del clustering. Estas reglas capturan escenarios prototípicos de sedentarismo. Por ejemplo:

5. *Regla 1: SI Actividad_relativa es **Baja** Y Superávit_calórico_basal es **Bajo**, ENTONCES el Sedentarismo es **Alto**.* (Interpretación: muy poca actividad física combinada con escaso gasto calórico relativo – perfil claramente sedentario).
6. *Regla 2: SI Actividad_relativa es **Alta** Y Superávit_calórico_basal es **Alto**, ENTONCES Sedentarismo **Bajo**.* (Perfil opuesto: individuo muy activo, con alto consumo calórico proporcional, por tanto bajo nivel de sedentarismo).
7. *Regla 3: SI HRV_SDNN es **Baja** Y Delta_cardiaco es **Alto**, ENTONCES Sedentarismo **Alto**.* (Interpretación: variabilidad cardíaca deprimida junto con respuesta cardíaca elevada al esfuerzo – indica posible descondicionamiento físico asociado a sedentarismo).
8. *Regla 4: SI Actividad_relativa es **Media** Y HRV_SDNN es **Media**, ENTONCES Sedentarismo **Medio**.* (Caso intermedio equilibrado en actividad y modulación autonómica).
9. *Regla 5: SI Actividad_relativa es **Baja** Y Superávit_calórico_basal es **Medio**, ENTONCES Sedentarismo **Medio-Alto** (con menor peso).* (Esta regla se diseñó con un peso de 0.7 para matizar su influencia, representando un caso borderline donde la actividad es baja pero el gasto calórico no tan extremo; al reducir su peso, se evita que esta condición intermedia domine la inferencia).

Estas reglas –que combinan condiciones *fisiológicamente plausibles*– conforman la **base de conocimiento difusa** del modelo. **Inferencia y Defuzzificación:** El motor de inferencia difusa aplica operadores lógicos difusos sobre las premisas de cada regla (se utilizó la *t-norma* mínima para el operador “Y”/conjunción, y *s-norma* máxima para combinaciones disyuntivas cuando aplicaban, además del complemento estándar para negaciones). A partir de las premisas, cada regla produce un consecuente difuso (etiqueta de Sedentarismo con cierto grado de verdad). Mediante el método de inferencia Mamdani, se agregan las salidas de todas las reglas en un conjunto difuso de salida acumulado. Finalmente, para obtener un valor numérico utilizable, este conjunto difuso resultante se **defuzzificó** usando el método del *centroide*. Así, el sistema proporciona para cada semana un **índice continuo de sedentarismo** denominado *Sedentarismo_score* $\in [0, 1]$, donde 0 corresponde teóricamente al perfil menos sedentario (más activo) y 1 al más sedentario. Este enfoque permite traducir múltiples señales biométricas complejas en una **clasificación interpretable** del

nivel de sedentarismo, respaldada por reglas auditables en lenguaje natural –una ventaja importante frente a métodos de “caja negra”–.

1. **Validación del modelo difuso:** Para evaluar la eficacia del sistema de inferencia difusa, se comparó su salida con la clasificación obtenida vía clustering no supervisado (tomada como *ground truth* operacional). Cada semana tiene tanto un *Sedentarismo_score* continuo (modelo difuso) como una etiqueta binaria de referencia (clúster 0 = bajo sedentarismo, clúster 1 = alto sedentarismo). Se realizó una búsqueda exhaustiva del **umbral óptimo (τ)** sobre el score difuso para discriminar entre *sedentarismo bajo* vs. *alto*. El criterio de selección fue maximizar el **F1-Score** (media armónica de precisión y sensibilidad) de la clasificación resultante, de modo que el modelo difuso replicara lo mejor posible la partición de clusters. Este análisis arrojó un **$\tau \approx 0.30$** como umbral que optimiza el balance; con dicho punto de corte, el modelo presentó un desempeño **robusto** al identificar semanas de sedentarismo alto. En particular, se logró una **sensibilidad ≈ 97 –98%** (recall del modelo respecto al clúster de sedentarismo alto), minimizando los falsos negativos –lo cual es crítico dado el enfoque de *cribado* sanitario, donde es preferible detectar todos los casos de riesgo aunque aumenten los avisos falsos. La **precisión** obtenida fue $\sim 74\%$, reflejando algunos falsos positivos esperados por la alta sensibilidad elegida. Globalmente, el modelo difuso alcanzó un **F1-Score ≈ 0.84** , indicando una **alta concordancia** con la clasificación data-driven. También se calculó la **exactitud global** ($\sim 74\%$ de semanas correctamente clasificadas) y el coeficiente de correlación Matthews (MCC ~ 0.29), que evidenció una concordancia moderada positiva considerando el desequilibrio de clases. Adicionalmente, se examinó la concordancia por individuo: en promedio un **70%** de las semanas de cada sujeto fueron categorizadas igual por el fuzzy y por el clustering, con rangos desde $\sim 28\%$ hasta $\sim 99\%$. Se observó que la menor concordancia se concentró en aquellos participantes con **alta variabilidad intra-semanal** (IQR semanales amplios en sus métricas), lo cual sugiere que en casos de comportamiento muy irregular el modelo difuso –calibrado sobre estadísticos medianos– puede sobre-simplificar. Esto plantea posibles ajustes futuros, como personalizar el umbral τ por usuario o incluir la variabilidad (IQR) como modulador en las reglas. En conjunto, esta etapa de validación confirmó que el **sistema difuso propuesto reproduce con fiabilidad la distinción entre sedentarismo alto y bajo** presente en los datos y, significativamente, lo hace con un sesgo deliberado hacia la sensibilidad clínica. Dicho de otro modo, el modelo cumple con el objetivo de **detectar la mayoría de los casos de sedentarismo relevante** (potencial factor de riesgo cardiovascular) proporcionando al mismo tiempo una herramienta interpretable para apoyo en la toma de decisiones de salud poblacional. Este enfoque metodológico alineado con la justificación clínica garantiza que el modelo no solo tenga un sólido sustento técnico, sino también pertinencia médica al priorizar la identificación temprana de individuos en alto riesgo por comportamiento sedentario.
