Modelo de evaluación del comportamiento sedentario mediante lógica difusa y datos biométricos

> Documento técnico para integrar al manuscrito de tesis. Incluye justificación clínica y data-driven, diseño metodológico, resultados, discusión crítica, limitaciones y próximos pasos.

---

1. Objetivo

Desarrollar y validar un **sistema de inferencia difusa** para clasificar el **sedentarismo semanal** a partir de biométricos de wearables, y contrastar su salida con una **verdad operativa** derivada de **clustering no supervisado**.

---

2. Población y datos

- 10 adultos (5 mujeres, 5 hombres), seguimiento multianual.
- **Unidad de análisis:** semana por usuario.
- **Dataset final semanal:** 1,385 semanas agregadas con estadísticas robustas (p25/p50/p75, IQR) por variable.
- **Variables base diarias:** minutos de movimiento, horas monitorizadas, gasto calórico activo, HRV_SDNN, FC reposo y FC al caminar, entre otras.
- **Variables derivadas clave (diarias):**
- **Actividad_relativa** = $\frac{\text{minutos en movimiento}}{60 \times \text{horas monitorizadas}}$
  Normaliza por exposición al uso del reloj.
- **TMB** (Mifflin–St Jeor) por sexo, peso, talla y edad.
- **Superávit_calórico_basal** = $\frac{\text{Gasto activo} \times 100}{\text{TMB}}$
  Ajusta por antropometría; permite comparaciones inter-sujeto.

---

3. Pipeline metodológico

1) **Preprocesamiento diario** y creación de derivadas (Actividad_relativa, TMB, Superávit_calórico_basal).
- Imputación jerárquica con *gates* (no-wear duro, actividad baja, normal) y **medianas móviles unidireccionales (pasado)** para evitar *leakage* temporal.
- Winsorización operativa p1–p99 por mes (limitaciones declaradas). 2) **Agregación semanal** con métricas robustas (p50 e IQR) de variables seleccionadas.
Variables semanales retenidas para modelado:
Actividad_relativa_p50, Actividad_relativa_IQR, Superávit_calórico_basal_p50, Superávit_IQR, HRV_SDNN_p50, HRV_SDNN_IQR, Delta_cardiaco_p50, Delta_cardiaco_IQR; donde **Delta_cardiaco** = FC_al_caminar_p50 − FCr_p50. 3) **Clustering no supervisado (verdad operativa):** K-means con *K-sweep*

(K=2..6), selección por *Silhouette* y estabilidad.
Resultado robusto: **K=2** con tamaños ~30% y ~70%. 4) **Sistema de inferencia difusa (screening interpretable):**
- **Inputs (4, p50):** Actividad_relativa, Superávit_calórico_basal, HRV_SDNN, Delta_cardiaco.
- **Funciones de pertenencia (MF):** triangulares por percentiles (p10–p25–p40; p35–p50–p65; p60–p75–p90) respetando la dirección clínica (**higher_better** o **lower_better**).
- **Reglas (5):** - R1: Actividad baja ∧ Superávit bajo → Sedentarismo alto. - R2: Actividad alta ∧ Superávit alto → Sedentarismo bajo. - R3: HRV baja ∧ Delta alto → Sedentarismo alto. - R4: Actividad media ∧ HRV media → Sedentarismo medio. - R5: Actividad baja ∧ Superávit medio → Sedentarismo medio-alto (peso 0.7). - **Salida:** *Sedentarismo_score* ∈ [0,1]. 5) **Validación cruzada:** búsqueda del **umbral τ** que maximiza F1 contra la partición K=2 del clustering.

---

4. Resultados

4.1. Pre-clustering QC

- **Multicolinealidad:** VIF ≤ 1.88 en todos los *features* (sin redundancia severa).
- **PCA:** PC1=26.5%, PC2=20.4% (≈46.9% acumulado) → estructura multidimensional; no se reduce dimensionalidad.
- **K-sweep:** Mejor **K=2** (Sil≈0.23); K≥5 inestable por *micro-clusters*.

4.2. Sistema difuso

- **Membresías:** 4 variables × 3 etiquetas (baja/media/alta) con percentiles de la muestra.
- **Distribución del score:** media 0.571 ± 0.235; rango [0.000, 1.000] → no degenerado.
- **Mapeo natural por cluster:**
- Cluster 1: *Sedentarismo_score* medio 0.621 → **Alto Sedentarismo**.
- Cluster 0: *Sedentarismo_score* medio 0.454 → **Bajo Sedentarismo**.

4.3. Validación vs clusters (verdad operativa)

- **Umbral óptimo: τ = 0.30** (máx F1).
- **Métricas globales (N=1337):**
  Accuracy 0.74 · F1 0.84 · Precision 0.737 · Recall 0.976 · MCC 0.294.
- **Matriz de confusión:** TN=77, FP=325, FN=22, TP=913.
- **Concordancia por usuario:** media 70% (rango 27.7% – 99.3%). Casos con baja concordancia: u3, u2, u8 (revisión dirigida).

---

5. Interpretación clínica y fisiológica

1) **Alta sensibilidad (Recall 97.6%)**: adecuado para **cribado**; minimiza falsos negativos (seguridad del paciente).
2) **Trade-off esperado**: falsos positivos en τ=0.30; preferible en screening con confirmación clínica posterior.
3) **Roles fisiológicos de inputs**:
- **Actividad_relativa** (exposición-normalizada) y **Superávit_calórico_basal** (ajustado por TMB) separan

perfiles **activo–gastador** vs **sedente–conservador**.
- **HRV_SDNN** y **Delta_cardiaco** capturan eficiencia autonómica y carga cardiovascular durante marcha.
4) **Heterogeneidad inter-sujeto**: discordancias concentradas en usuarios con **alta variabilidad intra-semanal**; sugiere explorar **τ personalizado** o reglas moduladas por **IQR**.

---

6. Fortalezas metodológicas

- **Convergencia supervisado–no supervisado:** fuzzy (interpretable) ≈ clustering (data-driven) con **F1=0.84**.
- **MF por percentiles**: anclaje robusto a la distribución observada; fácil recalibración por cohorte.
- **Trazabilidad completa**: desde insumos diarios hasta auditorías de imputación y *logs* por paso.

---

7. Limitaciones y mitigación

1) **Falsos positivos** (FP=325): mantener τ=0.30 por política de sensibilidad; reportar **zona intermedia (0.40–0.60)** y usar confirmación clínica.
2) **Heterogeneidad por usuario**: revisar `discordancias_top20` y considerar **τ por usuario** o **pesos por IQR** en R5.
3) **Silhouette moderado** del clustering (≈0.23): aceptado por interpretabilidad K=2; no usar K≥5.
4) **Escalado global**: recalibración anual o por cohorte para evitar arrastre por valores extremos históricos.

---

8. Reproducibilidad (archivos clave)

- **Configuración fuzzy:** `fuzzy_config/fuzzy_membership_config.yaml` y `feature_scalers.json` (funciones de pertenencia y escalado).
- **Salidas fuzzy:** `analisis_u/fuzzy/fuzzy_output.csv`, `08_fuzzy_inference_log.txt`.
- **Evaluación vs clusters:** `09_eval_fuzzy_vs_cluster.txt`, `plots/` (PR curve, histograma, matriz de confusión, distribución por cluster).
- **Semanal consolidado:** `weekly_consolidado.csv` y `cluster_inputs_weekly.csv`.

---

9. Implicaciones y aplicación

- **Clínica:** herramienta de **screening** poblacional del sedentarismo con reglas auditables.
- **Salud pública/laboral:** monitoreo longitudinal y detección temprana de empeoramiento conductual.
- **Investigación:** marco reproducible para integrar nuevas variables (sueño, dieta, estrés) sin perder interpretabilidad.

---

10. Próximos pasos

1) **Personalización de umbral τ** por usuario o subpoblaciones.
2) **Reglas moduladas por variabilidad (IQR)** para capturar intermitencia.
3) **Validación externa** en nueva cohorte y análisis de sensibilidad de MF.
4) **Reporte clínico**: generar *dashboard* y resúmenes por usuario/semana con alertas.

Agradecimientos

A los participantes y al equipo de análisis por su colaboración sostenida.

Notas para el manuscrito - Incluir 6 figuras: MF (4), PR-curve, matriz de confusión, distribución por cluster.
- Incluir 2 tablas: métricas globales y concordancia por usuario.
- Anexar rutas y nombres de archivos para asegurar reproducibilidad.