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.

Tablas de métricas por usuario

Nota: Estas tablas están listas para pegar en la tesis. Si ya tienes los archivos `analisis_u/fuzzy/fuzzy_output.csv` y `analisis_u/fuzzy/09_eval_fuzzy_vs_cluster.txt`, puedes rellenarlas automáticamente con el script `09_fuzzy_vs_clusters_eval.py` (bloque `per_user_summary`). Si prefieres hacerlo manualmente, utiliza las definiciones del anexo al final.

1) Métricas de clasificación (Fuzzy vs. Clusters) por usuario

Usuario	Semanas (N)	% Datos observados*	Accuracy	Precision	Recall	F1	MCC	τ usado	TP	FP	TN	FN
u1 – ale	—	—	—	—	—	—	—	0.30	—	—	—	—
u2 – brenda	—	—	—	—	—	—	—	0.30	—	—	—	—
u3 – christina	—	—	—	—	—	—	—	0.30	—	—	—	—
u4 – edson	—	—	—	—	—	—	—	0.30	—	—	—	—
u5 – esmeralda	—	—	—	—	—	—	—	0.30	—	—	—	—
u6 – fidel	—	—	—	—	—	—	—	0.30	—	—	—	—

Usuario	Semanas (N)	% Datos observados*	Accuracy	Precision	Recall	F1	MCC	τ usado	TP	FP	TN	FN
u7 – kevin	—	—	—	—	—	—	—	0.30	—	—	—	—
u8 – legarda	—	—	—	—	—	—	—	0.30	—	—	—	—
u9 – lmartinez	—	—	—	—	—	—	—	0.30	—	—	—	—
u10 – vane	—	—	—	—	—	—	—	0.30	—	—	—	—

Global (10 usuarios): Accuracy=0.74, Precision=0.737, Recall=0.976, F1=0.840, τ=0.30.

* `% Datos observados` se calcula como (1 − %imputación total) en la semana promedio del usuario.

2) Distribución de clusters por usuario

Usuario	Cluster Alto Sed (%)	Cluster Bajo Sed (%)	Diferencia absoluta (%)
u1 – ale	—	—	—
u2 – brenda	—	—	—
u3 – christina	—	—	—
u4 – edson	—	—	—
u5 – esmeralda	—	—	—
u6 – fidel	—	—	—
u7 – kevin	—	—	—
u8 – legarda	—	—	—
u9 – lmartinez	—	—	—
u10 – vane	—	—	—

3) Estadísticos semanales por usuario (p50 ± sd | IQR)

Usuario	Act_rel_p50 (±sd)	Act_rel_IQR	Superávit_p50 (±sd)	Superávit_IQR	HRV_SDNN_p50 (±sd)	HRV_IQR	ΔCard_p50 (±sd)
u1 – ale	—	—	—	—	—	—	—
u2 – brenda	—	—	—	—	—	—	—

Usuario	Act_rel_p50 (±sd)	Act_rel_IQR	Superávit_p50 (±sd)	Superávit_IQR	HRV_SDNN_p50 (±sd)	HRV_IQR	ΔCard_p50 (±sd)
u3 – christina	—	—	—	—	—	—	—
u4 – edson	—	—	—	—	—	—	—
u5 – esmeralda	—	—	—	—	—	—	—
u6 – fidel	—	—	—	—	—	—	—
u7 – kevin	—	—	—	—	—	—	—
u8 – legarda	—	—	—	—	—	—	—
u9 – lmartinez	—	—	—	—	—	—	—
u10 – vane	—	—	—	—	—	—	—

> Sugerencia de redacción: “En el Anexo X (Tabla 2) se reporta la proporción semanal de pertenencia al clúster de alto sedentarismo por usuario; en promedio la cohorte muestra ~70% de semanas en alto sedentarismo y ~30% en bajo, consistente con la estructura bimodal descubierta en K=2”.

Anexo: Definiciones operativas

Horizonte temporal - **Unidad de análisis:** Semana calendario (lunes–domingo). - **Semana válida:** ≥3 días con uso ≥8 h/día o <16 h sin registro; %imputación ≤60%.

Categorías de imputación - **Observado:** Datos reales del dispositivo. - **Rolling mediana (backward-only):** Mediana de ventanas históricas crecientes si faltan datos cercanos. - **Hard no-wear:** Días con uso <8 h o sin registro ≥16 h. - **Baseline fisiológico:** Sustitución conservadora (p.ej., FCr promedio de reposo) cuando no hay histórico suficiente.

Variables base (diario) - `Total_hrs_monitorizadas` (h/día): horas con señal válida. - `min_totales_en_movimiento` (min/día): minutos en anillo de movimiento. - `Gasto_calorico_activo` (kcal/día): calorías activas. - `HRV_SDNN` (ms): variabilidad cardiaca (mediana diaria de SDNN). - `FC_al_caminar_promedio_diario` (lpm): promedio de FC al caminar. - `FCr_promedio_diario` (lpm): frecuencia cardiaca en reposo.

Variables derivadas (diario) - **Actividad_relativa** = `min_totales_en_movimiento ÷ (60 × Total_hrs_monitorizadas)`; tasa de minutos en movimiento por hora monitorizada (corrige sesgo de exposición). - **TMB** (kcal/día): Mifflin–St Jeor (sexo/peso/estatura/edad). - **Superavit_calorico_basal** (%) = `(Gasto_calorico_activo × 100) ÷ TMB`. - **Delta_cardiaco** (lpm) = `FC_al_caminar_promedio_diario − FCr_promedio_diario`.

Agregación semanal (features para clustering/fuzzy) Para cada variable clave se calculan **p50** y **IQR** por semana; el set final de 8 features: - `Actividad_relativa_p50`, `Actividad_relativa_iqr` - `Superavit_calorico_basal_p50`, `Superavit_calorico_basal_iqr` - `HRV_SDNN_p50`, `HRV_SDNN_iqr` - `Delta_cardiaco_p50`, `Delta_cardiaco_iqr`

Estandarización para análisis - Z-robust por cohorte (mediana/MAD) en features semanales solamente (evita contaminar ratios con normalización previa a su construcción).

Clustering - Método: K-Means sobre 8 features estandarizadas. - Selección K: barrido K=2..6; óptimo **K=2** por Silhouette (≈0.23) y estabilidad. - Interpretación: Cluster 1 → **Alto sedentarismo**; Cluster 0 → **Bajo sedentarismo**.

Sistema de inferencia difusa - Entradas: `Actividad_relativa_p50`, `Superavit_calorico_basal_p50`, `HRV_SDNN_p50`, `Delta_cardiaco_p50`. - Funciones de membresía triangulares (baja/media/alta) definidas por percentiles de la cohorte (p10–p25–p40, p35–p50–p65, p60–p75–p90) respetando dirección fisiológica (higher_better vs lower_better). - Reglas (ejemplos): - R1: Si Actividad es **baja** y Superávit es **bajo** → Sedentarismo **alto**. - R2: Si HRV es **baja** y ΔCard es **alta** → Sedentarismo **alto**. - R3: Si Actividad es **alta** y Superávit es **alto** → Sedentarismo **bajo**. - Defuzzificación: centroide a score ∈ [0,1]. - Umbral operativo τ=0.30 (maximiza F1=0.84 y recall=0.976 frente a clusters).

Métricas globales de validación - Accuracy 0.74 | Precision 0.737 | Recall 0.976 | F1 0.84 | MCC 0.294. - Matriz de confusión total disponible en `analisis_u/fuzzy/plots/confusion_matrix.png`.

Criterios de trazabilidad y control - `apple_health_export/*` es **solo lectura** (respaldo). - Todo archivo de trabajo reside en `4_semestre_dataset/` con sufijo `_uN` por usuario. - Logs obligatorios por paso: `03_variabilidad_dual_log.txt`, `04_agregacion_semanal_log.txt`, `06_precluster_qc_log.txt`, `08_fuzzy_inference_log.txt`, `09_eval_fuzzy_vs_cluster.txt`.

Cómo rellenar automáticamente las tablas:

1) Ejecuta `python 09_fuzzy_vs_clusters_eval.py --per_user_csv per_user_metrics.csv`. 2) Abre `per_user_metrics.csv` y copia/pega los renglones en las tablas 1–3. 3) Verifica que τ=0.30 y los conteos TP/FP/TN/FN coincidan con la matriz global.