Informe Técnico Completo

Pipeline Bioestadístico para la Clasificación de Sedentarismo mediante Lógica Difusa y Clustering

Perspectiva Bioestadística, Clínica y Computacional

Luis Ángel Martínez

Universidad Autónoma de Chihuahua
Facultad de Medicina y Ciencias Biomédicas

Programa de Maestría en Ciencias de la Salud

22 de octubre de 2025

Resumen

El presente informe técnico documenta de manera exhaustiva el pipeline bioestadístico desarrollado para la clasificación objetiva del sedentarismo semanal utilizando datos biométricos de dispositivos wearables (Apple Watch). Este proyecto representa un estudio longitudinal con $N = 10$ participantes (5M/5H) que generaron 1,337 semanas válidas de datos continuos.

El pipeline integra tres perspectivas complementarias: **bioestadística** (modelado probabilístico robusto, reducción dimensional, clustering, validación), **clínica** (normalización antropométrica, interpretación fisiológica de variables derivadas, relevancia para ciencias del ejercicio), y **computacional** (arquitectura modular en Python, estrategias de imputación jerárquica, optimización de hiperparámetros).

Metodológicamente, el estudio pivotó de un enfoque supervisado inicial (predicción de Calidad de Vida mediante Redes Neuronales Artificiales, invalidado empíricamente) a un paradigma *data-driven* dual: (1) descubrimiento de patrones mediante clustering no supervisado (K-Means, $K = 2$, Silhouette= 0,232), empleado como **Verdad Operativa (GO)**, y (2) construcción de un Sistema de Inferencia Difusa Mamdani interpretable con 5 reglas expertas, validado contra la GO con $F1 = 0{,}840$, Recall= 0,976, MCC= 0,294.

Cada fase del pipeline se presenta bajo el marco riguroso de los **6 pasos del análisis estadístico**: planteamiento de hipótesis, selección del estadístico, regla de decisión, cálculos, decisión estadística y conclusión. Se incluyen ecuaciones matemáticas formales, pseudocódigo, referencias a figuras y tablas, y una justificación detallada de la decisión metodológica de *no* emplear un split Train/Test 80/20, reemplazado por validación cruzada Leave-One-User-Out (LOUO) y análisis de sensibilidad.

Palabras clave: Sedentarismo, Wearables, Apple Watch, Lógica Difusa, Clustering, K-Means, Imputación Jerárquica, Ingeniería de Características, Validación Cruzada, Python.

Índice general

Capítulo 1

Planteamiento del Problema e Hipótesis Inicial

1.1. Contexto Epidemiológico y Clínico

El comportamiento sedentario (CS), definido por la Organización Mundial de la Salud como cualquier actividad con gasto energético $\leq$ 1,5 METs en posición sentada o reclinada durante horas de vigilia, constituye un factor de riesgo independiente para enfermedades crónicas no transmisibles (ECNT), incluyendo obesidad, diabetes tipo 2, enfermedad cardiovascular y ciertos tipos de cáncer [1].

La medición objetiva del CS mediante acelerometría triaxial en dispositivos wearables de consumo masivo (e.g., Apple Watch, Fitbit, Garmin) ha revolucionado la epidemiología del comportamiento, permitiendo cuantificar patrones de actividad física en condiciones de "vida libre" con alta resolución temporal ($\geq$ 1 Hz) y sin el sesgo de auto-reporte característico de cuestionarios.

1.2. Hipótesis Inicial y Objetivo Primario

Paso 1: Planteamiento de Hipótesis

Hipótesis H_0 (inicial, posteriormente rechazada):
Existe una relación inversa, lineal y medible entre el comportamiento sedentario objetivo (CS_obj), cuantificado mediante métricas derivadas de acelerometría y fotopletismografía (PPG) del Apple Watch, y la percepción subjetiva de Calidad de Vida Relacionada con la Salud (CVRS), evaluada mediante el cuestionario SF-36.
Formalmente:

$$\text{CVRS}_{SF36} = f(\text{CS}_{\text{obj}}) + \epsilon, \quad \epsilon \sim \mathcal{N}(0, \sigma^2) \tag{1.1}$$

donde f sería una función lineal o no lineal modelable mediante Redes Neuronales Artificiales (ANN).

1.2.1. Objetivo Primario (Fase Inicial)

Desarrollar un modelo predictivo (ANN) capaz de cuantificar la CVRS a partir de métricas biométricas continuas, con $R^2 \geq 0{,}70$ y MAE ≤ 10 puntos en escala SF-36.

1.3. Marco de los 6 Pasos: Planteamiento

Paso 2: Selección del Estadístico/Método

Selección del método:
Se propuso inicialmente un análisis correlacional (Pearson/Spearman) seguido de modelado supervisado mediante ANN (arquitectura feedforward, activación ReLU, optimizador Adam).

Paso 3: Regla de Decisión

Regla de decisión:
Si $|r| \geq 0{,}60$ (correlación fuerte) y el modelo ANN alcanza $R^2 \geq 0{,}70$ en validación cruzada 5-fold, se aceptará la hipótesis de relación cuantificable.

Paso 5: Decisión Estadística

Decisión preliminar:
Se decidió proceder con un diseño longitudinal que recolectaría datos biométricos continuos (Apple Watch) y evaluaciones periódicas del SF-36 para probar esta correlación.

Paso 6: Conclusión

Conclusión del planteamiento:
Existía suficiente justificación teórica (revisión de literatura: correlaciones reportadas entre actividad física y CVRS en el rango $r = 0{,}30 - 0{,}50$) para explorar esta vía, aunque con la precaución de que la relación podría ser más compleja de lo anticipado.

Capítulo 2

Selección del Dispositivo Wearable y Diseño de la Cohorte

2.1. Evaluación de Dispositivos Wearables

2.1.1. Criterios de Selección

Paso 1: Planteamiento de Hipótesis

Problema/Hipótesis:
Necesitábamos un dispositivo wearable que cumpliera simultáneamente:

- Alta penetración de mercado (facilitar reclutamiento BYOD)
- Sensores validados: acelerómetro 3-ejes ($\geq$ 50 Hz), PPG para FC/VFC
- Plataforma de exportación de datos crudos o agregados
- Consistencia inter-versión (minimizar heterogeneidad instrumental)

Hipótesis: El Apple Watch, por su ecosistema cerrado y validaciones previas en literatura (Stahl et al., 2016; Shcherbina et al., 2017), sería la opción preferente.

2.1.2. Análisis Comparativo

Tabla 2.1: Matriz de Decisión: Comparación de Dispositivos Wearables

Criterio	**Apple Watch**	**Fitbit**	**Garmin**	**Mi Band**
Penetración México	Alta	Media	Media-Baja	Alta
Sensores validados	Sí	Sí	Sí	Parcial
Exportación datos	HealthKit (XML)	API limitada	Garmin Connect	Propietaria
Consistencia HW	Alta	Media	Alta	Baja
Costo promedio (USD)	300-800	100-300	250-700	30-50
Score ponderado	**9.2**	7.5	7.8	5.1

Paso 2: Selección del Estadístico/Método

Método de evaluación:
Matriz de decisión multicriterio con pesos asignados según importancia para el estudio:

- Validez de sensores: 35 %
- Exportabilidad de datos: 30 %
- Consistencia: 20 %
- Penetración: 15 %

Paso 5: Decisión Estadística

Decisión:
Se seleccionó el **Apple Watch** (Series 3 o superior) como dispositivo estándar del estudio, adoptando un enfoque *Bring Your Own Device* (BYOD) para maximizar adherencia y minimizar el efecto Hawthorne.

2.2. Diseño de la Cohorte

2.2.1. Tamaño Muestral y Justificación

Paso 1: Planteamiento de Hipótesis

Planteamiento:
Dada la naturaleza longitudinal del estudio (objetivo: capturar variabilidad intra-sujeto durante ≥ 12 semanas), el tamaño muestral N se justificó por:

$$n_{\text{observaciones}} = N_{\text{sujetos}} \times T_{\text{semanas}} \geq 1000 \tag{2.1}$$

Con $N = 10$ y $T \approx 130$ semanas (promedio), se alcanzarían ≈ 1300 observaciones semanales, suficiente para:

- Modelado de clustering con $n/K \geq 500$ por grupo ($K = 2$)
- Optimización de hiperparámetros del sistema difuso
- Validación cruzada Leave-One-Subject-Out

2.2.2. Criterios de Inclusión/Exclusión

Tabla 2.2: Criterios de Elegibilidad de Participantes

Criterio	Inclusi
Edad	18-65 añ
Dispositivo	Propietario Apple Watch Series ≥
Uso previo	≥ 6 meses continu
Estado de salud	Ambulatorio, sin limitacion
Consentimiento	Informado por escr
Datos exportables	≥ 80 % días con dat

Paso 4: Cálculos

Cálculos de factibilidad:
Se convocó a 15 candidatos, de los cuales:

- 12 cumplieron criterios de inclusión
- 10 completaron el protocolo (2 abandonos por causas no relacionadas)
- Distribución final: 5 hombres, 5 mujeres
- Edad: $\bar{x} = 32{,}4$ años, $s = 8{,}7$ años
- IMC: $\bar{x} = 26{,}1$ kg/m^2, $s = 4{,}2$ kg/m^2

Paso 6: Conclusión

Conclusión metodológica:
Aunque no representativa poblacionalmente (muestra de conveniencia), la cohorte de $N = 10$ permite un análisis longitudinal profundo con potencia estadística adecuada para el descubrimiento de patrones intra-sujeto y validación de sistemas expertos interpretativos (objetivo secundario tras el pivote metodológico).

Capítulo 3

Protocolo de Convocatoria, Recepción y Preprocesamiento de Datos

3.1. Protocolo de Recolección de Datos

3.1.1. Diseño del Protocolo

Paso 1: Planteamiento de Hipótesis

Planteamiento:
Para garantizar la integridad, trazabilidad y ética de los datos biométricos sensibles, se diseñó un protocolo estandarizado que incluye:

1. Consentimiento informado (aprobación comité ética institucional)
2. Instrucciones de exportación (HealthKit → archivo `export.zip`)
3. Aplicación del cuestionario SF-36 (versión mexicana validada)
4. Anonimización inmediata (códigos: u1, u2, ..., u10)
5. Almacenamiento seguro (servidor institucional, encriptación AES-256)

3.1.2. Estructura de Datos Crudos

Los datos exportados de Apple Health siguen el esquema XML:

```
<HealthData>
  <Record type="HKQuantityTypeIdentifierStepCount"
          sourceName="Apple Watch de Luis"
          value="1245"
          unit="count"
          startDate="2023-10-22 08:15:00"
          endDate="2023-10-22 08:16:00"/>
  ...
</HealthData>
```

Listing 3.1: Estructura XML de Apple Health Export

3.2. Pipeline de Preprocesamiento

3.2.1. Conversión XML → CSV

Paso 2: Selección del Estadístico/Método

Método:
Parseo XML mediante `ElementTree` (Python), con transformaciones:

- Filtrado por `sourceName` (solo datos Apple Watch, excluir iPhone)
- Conversión de timestamps a zona horaria local (UTC-6, Chihuahua)
- Agregación a nivel diario (suma/media según métrica)

Algorithm 1 Preprocesamiento XML a CSV Diario

```
Input: export.zip por participante
Output: DB_u{id}.csv con columnas [fecha, pasos, calorias, fc_reposo, hrv_sdnn, ...]

procedure ParseXML(xml_file, user_id)
    tree ← parse(xml_file)
    records ← tree.findall(Record")
    df ← empty_dataframe()
    for record in records do
        if record.sourceName contains .Apple Watch" then
            type ← record.type
            value ← record.value
            date ← record.startDate.date()
            df.append([date, type, value])
        end if
    end for
    df_pivot ← df.pivot(index=date, columns=type, values=value)
    df_pivot.to_csv(f"DB_u{user_id}.csv")
end procedure
```

Paso 4: Cálculos

Cálculos de agregación:
Para cada usuario y día:

$$\text{Pasos}_{\text{día}} = \sum_{t=0}^{23:59} \text{StepCount}(t) \tag{3.1}$$

$$\text{FC}_{\text{reposo}} = \min\{\text{HeartRate}(t) : t \in [02:00, 05:00]\} \tag{3.2}$$

$$\text{HRV\_SDNN}_{\text{día}} = \text{mean}\{\text{SDNN}(t) : t \in [00:00, 23:59]\} \tag{3.3}$$

3.2.2. Auditoría de Calidad de Datos

Tabla 3.1: Métricas de Completitud por Usuario (Fase Pre-Imputación)

Usuario	Días totales	Días válidos	Completitud (%)	Missing FC (%)	Missing HRV
u1	900	852	94.7	8.2	
u2	850	801	94.2	9.1	
u3	920	884	96.1	5.4	
...	...	...	...	...	
u10	880	831	94.4	7.8	
Media	**885**	**838**	**94.7**	**7.6**	

Paso 5: Decisión Estadística

Decisión:
La completitud general $> 94\,\%$ es aceptable para estudios observacionales de vida libre. Las variables cardiovasculares (FC, HRV) presentan mayor tasa de missingness (mecanismo: quitarse el reloj durante sueño/carga), requiriendo estrategia de imputación robusta (Capítulo 6).

Capítulo 4

Análisis Exploratorio de Datos (EDA) y Validación del SF-36

4.1. Caracterización de Variables Biométricas

4.1.1. Tipología y Distribuciones

Paso 1: Planteamiento de Hipótesis

Hipótesis:
Se esperaba que las variables biométricas diarias presentaran:

- Distribuciones asimétricas (pasos, minutos ejercicio: asimetría positiva)
- Alta variabilidad día-a-día (CV $> 50\,\%$)
- No-normalidad (rechazo de Shapiro-Wilk con $p < 0{,}05$)

Paso 2: Selección del Estadístico/Método

Métodos aplicados:

- Estadísticos descriptivos robustos: mediana, IQR, MAD
- Pruebas de normalidad: Shapiro-Wilk (si $n < 5000$), Kolmogorov-Smirnov (si $n \geq 5000$)
- Visualización: histogramas, Q-Q plots, boxplots por usuario

Tabla 4.1: Estadísticos Descriptivos de Variables Clave (Nivel Diario, $n = 8,380$ días)

Variable	Media	DE	Mediana	IQR	Min	Max	SW p-valor
Pasos	6,842	4,231	6,120	4,890	0	28,450	$< 0,001$
Calorías activas	385	287	342	298	0	1,892	$< 0,001$
FC reposo (lpm)	58.3	8.7	57.0	10.0	42	92	0,014
HRV SDNN (ms)	52.1	18.4	48.5	22.0	15	128	$< 0,001$
FC caminar (lpm)	95.8	12.3	94.0	15.0	65	145	0,082
Min sedentarios	678	142	702	185	120	1,320	$< 0,001$

Paso 5: Decisión Estadística

Decisión estadística:
Se rechaza la normalidad para todas las variables excepto FC_caminar ($p = 0,082$). Consecuencia: uso obligatorio de métodos no paramétricos o robustos (medianas, bootstrapping, Mann-Whitney U) en análisis posteriores.

4.1.2. Gráficos Exploratorios

Ver Figuras:

- `4 semestre_dataset/analisis_u/histogramas_variables_clave.png`
- `4 semestre_dataset/analisis_u/qqplots_normalidad.png`
- `4 semestre_dataset/analisis_u/boxplots_por_usuario.png`

4.2. Validación Psicométrica del SF-36

4.2.1. Estructura del Cuestionario

El SF-36 evalúa 8 dimensiones de CVRS mediante 36 ítems:

- Función Física (FF)
- Rol Físico (RF)
- Dolor Corporal (DC)
- Salud General (SG)
- Vitalidad (VT)
- Función Social (FS)
- Rol Emocional (RE)
- Salud Mental (SM)

Paso 2: Selección del Estadístico/Método

Métrica de fiabilidad:
Alfa de Cronbach por dimensión, criterio $\alpha \geq 0{,}70$ (aceptable).

$$\alpha = \frac{K}{K-1}\left(1 - \frac{\sum_{i=1}^{K} \sigma_i^2}{\sigma_{\text{total}}^2}\right) \tag{4.1}$$

donde K = número de ítems, σ_i^2 = varianza del ítem i.

Tabla 4.2: Fiabilidad del SF-36 en la Cohorte ($N = 10$)

Dimensión SF-36	**α Cronbach**	**Varianza**	**Decisión**
Función Física	0.82	145.3	Aceptable
Rol Físico	0.51	0.0	Rechazada (var=0)
Dolor Corporal	0.78	98.7	Aceptable
Salud General	0.73	112.4	Aceptable
Vitalidad	0.64	87.2	Marginal
Función Social	0.71	102.1	Aceptable
Rol Emocional	0.76	118.5	Aceptable
Salud Mental	0.80	134.2	Aceptable

Paso 5: Decisión Estadística

Decisión crítica:
La dimensión **Rol Físico** presenta varianza nula (todos los participantes reportaron el mismo valor, efecto techo/suelo), invalidando su uso. Vitalidad ($\alpha = 0{,}64$) está por debajo del umbral.
Consecuencia: Estos problemas psicométricos, sumados a correlaciones débiles con biométricos (siguiente sección), motivaron el rechazo de la hipótesis inicial y el pivote metodológico.

Paso 6: Conclusión

Conclusión EDA:

1. Los datos biométricos son ruidosos y no-normales, requiriendo métodos robustos.
2. El SF-36 presenta limitaciones en esta cohorte específica (tamaño, homogeneidad).
3. La alta variabilidad diaria (CV $> 100\,\%$ en ejercicio) justifica agregación temporal (semanal) para capturar patrones estables.

Capítulo 5

Pivote Metodológico: Del Enfoque Supervisado al Data-Driven

5.1. Análisis de Correlación SF-36 vs Biométricos

5.1.1. Hipótesis y Pruebas Iniciales

Paso 1: Planteamiento de Hipótesis

Hipótesis H_1 a probar:
Las métricas biométricas agregadas (media de 4 semanas) correlacionan significativamente ($|r| \geq 0{,}60$, $p < 0{,}01$) con los puntajes de CVRS del SF-36.

Paso 2: Selección del Estadístico/Método

Métodos:

- Correlación de Spearman (datos no-normales)
- Corrección Bonferroni para comparaciones múltiples ($\alpha^* = 0{,}05/32 = 0{,}0016$)
- Scatter plots con líneas de regresión LOWESS

Tabla 5.1: Matriz de Correlación: Biométricos Agregados vs SF-36 ($N = 10$)

	FF	**RF**	**DC**	**SG**	**VT**	**FS**	**RE**	**SM**
Pasos promedio	0.32	—	0.18	0.41	-0.05	0.27	0.14	0.09
Calorías promedio	0.38	—	0.22	0.45	-0.12	0.31	0.19	0.13
FC reposo promedio	-0.21	—	-0.14	-0.28	0.08	-0.18	-0.11	-0.06
HRV SDNN promedio	0.15	—	0.09	0.24	0.31	0.12	0.08	0.19
Min sedentarios	-0.29	—	-0.16	-0.35	-0.18	-0.24	-0.13	-0.11

Nota: RF excluido por varianza nula. Ninguna correlación alcanza $|r| \geq 0{,}60$ ni $p < 0{,}0016$.

Paso 5: Decisión Estadística

Decisión estadística:
Se rechaza H_1. Las correlaciones observadas son débiles a moderadas ($0{,}09 \leq |r| \leq 0{,}45$) y ninguna sobrevive la corrección Bonferroni. La asociación es insuficiente para justificar un modelo predictivo.

5.2. Modelado con Redes Neuronales Artificiales (ANN)

5.2.1. Arquitectura y Entrenamiento

A pesar de las correlaciones débiles, se procedió a entrenar ANNs como prueba definitiva:

Algorithm 2 Entrenamiento de ANN para CVRS

Input: $X \in \mathbb{R}^{10\times16}$ (16 features biométricos), $y \in \mathbb{R}^{10\times7}$ (7 dimensiones SF-36 válidas)
Output: Modelo ANN, métricas de desempeño

Arquitectura: [16 inputs] $\rightarrow$ [32 ReLU] $\rightarrow$ [16 ReLU] $\rightarrow$ [7 Linear]
Optimizador: Adam ($\alpha = 0{,}001$, $\beta_1 = 0{,}9$, $\beta_2 = 0{,}999$)
Función de pérdida: MSE
Validación cruzada: 5-fold
Épocas: 500 con early stopping (patience=50)

Paso 4: Cálculos

Resultados del entrenamiento:

Métrica	Train	Validación	Test	Criterio
R^2	0.92	-0.18	-0.34	$\geq 0{,}70$
MAE	5.2	18.7	21.3	≤ 10
RMSE	7.8	24.1	27.9	≤ 15

Tabla 5.2: Desempeño del modelo ANN (peor de 20 configuraciones probadas)

Observación crítica: R^2 negativo en validación/test indica que el modelo es *peor que predecir la media*, evidenciando sobreajuste severo y ausencia de relación generalizable.

Paso 5: Decisión Estadística

Decisión metodológica CRÍTICA:
Se rechaza definitivamente la hipótesis inicial y el enfoque supervisado. Las causas identificadas:

1. $N = 10$ es insuficiente para ANN (regla de oro: $\geq 10 \times$ parámetros; aquí: $\approx 1,000$ parámetros)
2. Relación CS-CVRS es multifactorial, confundida por variables psicosociales no capturadas
3. SF-36 carece de sensibilidad a variaciones diarias/semanales de actividad en población joven-adulta sana

5.3. Reformulación: Nuevo Enfoque Data-Driven

5.3.1. Nueva Hipótesis

Paso 1: Planteamiento de Hipótesis

Hipótesis H_2 (reformulada):
Los datos biométricos contienen patrones latentes que permiten clasificar objetivamente semanas como "alto sedentarismo" vs "bajo sedentarismo", independientemente de la percepción subjetiva de CVRS.
Enfoque dual propuesto:

1. **Descubrimiento empírico**: Clustering no supervisado (K-Means) para identificar grupos naturales en los datos $\rightarrow$ *Verdad Operativa (GO)*
2. **Sistema experto interpretable**: Lógica Difusa (Mamdani) con reglas basadas en conocimiento fisiológico $\rightarrow$ *Modelo Clínico*
3. **Validación cruzada**: Concordancia entre ambos métodos independientes

Paso 2: Selección del Estadístico/Método

Métricas de éxito reformuladas:

- F1-Score $\geq 0,80$ (balance precisión-recall)
- Matthews Correlation Coefficient (MCC) $\geq 0,30$ (manejo desbalanceo)
- Interpretabilidad clínica de las reglas difusas

Paso 6: Conclusión

Conclusión del pivote:
Este cambio paradigmático transforma el estudio de *predictivo supervisado* a *descriptivo-clasificatorio data-driven*, más apropiado para la naturaleza exploratoria de los datos y el tamaño muestral. Los capítulos siguientes desarrollan este nuevo enfoque.

Capítulo 6

Estrategia de Imputación Jerárquica para Datos Faltantes

6.1. Diagnóstico de Missingness

6.1.1. Mecanismos de Datos Faltantes

Paso 1: Planteamiento de Hipótesis

Hipótesis sobre mecanismos:
Los datos faltantes en wearables no son MCAR (Missing Completely At Random), sino:

- **MAR (Missing At Random)**: FC/HRV ausentes durante actividades acuáticas (no resistance device)
- **MNAR (Missing Not At Random)**: Dispositivo quitado intencionalmente durante eventos sedentarios prolongados (e.g., cine, sueño extendido)

Paso 2: Selección del Estadístico/Método

Pruebas aplicadas:

- Test de Little MCAR: $\chi^2 = 487{,}3$, $p < 0{,}001 \rightarrow$ Rechazo MCAR
- Patrones de missingness visualizados con matrices de co-ocurrencia
- Análisis temporal: ACF/PACF de indicadores de missingness

Ver Figuras:

- `4 semestre_dataset/analisis_u/missingness_y_acf/missingness_matrix_u1.png`
- `4 semestre_dataset/analisis_u/missingness_y_acf/acf_plots/acf_u1.png`
- `4 semestre_dataset/analisis_u/missingness_y_acf/pacf_plots/pacf_u1.png`

6.2. Estrategia de Imputación Jerárquica

6.2.1. Principios de Diseño

1. **Sin fuga temporal**: Imputación *forward-only* (día t usa solo info $\leq t-1$)
2. **Plausibilidad fisiológica**: Valores imputados dentro de rangos clínicos
3. **Jerarquía de métodos**: De específico a general
4. **Transparencia**: Marcar columnas con sufijo `_imp` y registrar tasa

6.2.2. Algoritmo de Imputación

Algorithm 3 Imputación Jerárquica para Variables Cardiovasculares

Input: DataFrame diario con columnas [fecha, FC_caminar, FC_reposo, HRV_SDNN, ...]
Output: DataFrame con valores imputados y flags

for variable **in** [FC_caminar, FC_reposo, HRV_SDNN] **do**
    for row_idx **in** missing_indices(variable) **do**
        usuario ← row_idx.usuario
        fecha ← row_idx.fecha

        // Método 1: Media móvil 7 días previos
        ventana ← [fecha−7, fecha−1]
        if count(ventana) ≥ 4 **then**
            impute median(ventana) ▷ Robusto a outliers
            continue
        end if

        // Método 2: Media del mismo día de semana (último mes)
        mismo_dia ← filter(fecha.weekday == dia_semana, fecha ∈ [fecha−28, fecha−1])
        if count(mismo_dia) ≥ 2 **then**
            impute median(mismo_dia)
            continue
        end if

        // Método 3: Mediana histórica del usuario
        historico ← filter(usuario == usuario, fecha < fecha)
        if count(historico) ≥ 10 **then**
            impute median(historico)
            continue
        end if

        // Método 4: Estimación por ecuaciones de Tanaka (FC_reposo)
        if variable == FC_reposo **and** edad disponible **then**
            impute 220 − edad × 0,7 ▷ FC reposo estimado
            continue
        end if

        // Método 5 (último recurso): Mediana global
        impute median_global(variable)
    end for
end for

6.2.3. Resultados de Imputación

Tabla 6.1: Tasa de Imputación por Variable y Método

Variable	Missing (%)	M1 (%)	M2 (%)	M3 (%)	M4 (%)	M5 (%)
FC_caminar	7.6	68.2	21.3	8.9	0.0	1.6
FC_reposo	4.2	72.1	18.7	6.5	2.1	0.6
HRV_SDNN	14.8	61.5	24.8	10.3	0.0	3.4

Paso 4: Cálculos

Validación de plausibilidad:
Post-imputación, se verificó que todos los valores cumplan:

$$40 \leq \text{FC}_{\text{reposo}} \leq 100 \text{ lpm} \tag{6.1}$$
$$60 \leq \text{FC}_{\text{caminar}} \leq 160 \text{ lpm} \tag{6.2}$$
$$15 \leq \text{HRV\_SDNN} \leq 150 \text{ ms} \tag{6.3}$$

Violaciones detectadas: 3 outliers extremos (0.04 %), reemplazados por mediana del usuario.

Paso 5: Decisión Estadística

Decisión:
La estrategia jerárquica logró reducir missingness de 14.8 % (HRV) a 0 %, con > 90 % de valores imputados mediante métodos específicos del usuario (M1-M3), garantizando consistencia individual.

Paso 6: Conclusión

Conclusión:
La imputación jerárquica sin fuga temporal preserva la integridad de series temporales para análisis posteriores (ACF/PACF, agregación semanal). El análisis de variabilidad dual (Capítulo 8) confirmará que la imputación no distorsiona las distribuciones originales.

Capítulo 7

Ingeniería de Características: Variables Derivadas con Normalización Antropométrica

7.1. Problema de Comparabilidad Inter-Sujeto

7.1.1. Heterogeneidad Antropométrica

Paso 1: Planteamiento de Hipótesis

Problema:
Variables brutas (pasos, calorías, FC) no son directamente comparables entre individuos con diferente:

- Masa corporal (IMC: 19.8 – 32.4 kg/m^2 en la cohorte)
- Tasa Metabólica Basal (TMB: función de sexo, edad, peso, altura)
- Tiempo de uso del dispositivo (6.2 – 23.8 h/día)

Consecuencia: Un usuario pesado quemará más calorías en reposo que uno liviano; ignorar esto induce sesgo en clustering.

7.2. Variable 1: Actividad Relativa

7.2.1. Definición y Justificación

Paso 2: Selección del Estadístico/Método

Derivación matemática:

$$\text{Actividad\_relativa}_{\text{día}} = \frac{\text{Pasos}}{\text{Horas\_con\_datos}} \times \frac{1}{1000} \tag{7.1}$$

Unidades: *kilopasos por hora de monitoreo*

Justificación clínica: Normaliza por exposición al dispositivo. Un usuario con 10,000 pasos en 10 horas (1.0 kph) es *más activo* que uno con 10,000 pasos en 20 horas (0.5 kph).

7.2.2. Distribución y Validación

Tabla 7.1: Comparación: Pasos Brutos vs Actividad Relativa

Variable	Usuario	Media	DE	CV (%)	Mediana	IQR
Pasos	u1 (IMC 22.1)	8,542	3,921	45.9	8,120	4,650
	u5 (IMC 29.8)	5,234	2,814	53.8	5,010	3,210
	u9 (IMC 24.5)	7,892	3,654	46.3	7,650	4,120
Act_rel (kph)	u1	0.62	0.28	45.2	0.59	0.31
	u5	0.58	0.31	53.4	0.55	0.35
	u9	0.65	0.30	46.2	0.63	0.34

Paso 5: Decisión Estadística

Decisión:
Actividad_relativa reduce la varianza inter-sujeto atribuible a diferencias en tiempo de uso (CV similar, pero medianas más homogéneas), permitiendo clustering más justo.

7.3. Variable 2: Superávit Calórico Basal

7.3.1. Cálculo de TMB

Paso 2: Selección del Estadístico/Método

Ecuación de Harris-Benedict (revisada):
Para hombres:

$$\text{TMB}_\text{h} = 88{,}362 + (13{,}397 \times \text{peso\_kg}) + (4{,}799 \times \text{altura\_cm}) - (5{,}677 \times \text{edad}) \quad (7.2)$$

Para mujeres:

$$\text{TMB}_\text{m} = 447{,}593 + (9{,}247 \times \text{peso\_kg}) + (3{,}098 \times \text{altura\_cm}) - (4{,}330 \times \text{edad}) \quad (7.3)$$

7.3.2. Definición de Superávit

$$\text{Superávit\_calórico\_basal}_{\text{día}} = \frac{\text{Calorías\_activas}}{\text{TMB}} \times 100\,\% \quad (7.4)$$

Interpretación clínica:

- $< 20\,\%$: Gasto activo muy bajo (sedentarismo)
- $20 - 50\,\%$: Actividad ligera-moderada
- $> 50\,\%$: Actividad vigorosa o deportiva

Tabla 7.2: TMB y Superávit Calórico por Usuario

Usuario	Sexo	IMC	TMB (kcal/día)	Sup. p50 (%)
u1	M	22.1	1,742	28.3
u2	F	24.3	1,521	31.7
u3	M	26.8	1,865	25.9
...	...	...	...	...
u10	F	23.5	1,498	34.2

7.4. Variables 3 y 4: Perfiles Cardiovasculares

7.4.1. Delta Cardíaco

$$\text{Delta\_cardiaco}_{\text{día}} = \text{FC\_caminar} - \text{FC\_reposo} \quad (7.5)$$

Relevancia fisiológica: Mayor delta indica mejor reserva cardiovascular (respuesta rápida del sistema nervioso autónomo a demanda metabólica).

7.4.2. HRV SDNN

La Variabilidad de la Frecuencia Cardíaca (HRV), específicamente SDNN (Standard Deviation of NN intervals), es un biomarcador del tono vagal:

- SDNN > 50 ms: Buena modulación autonómica
- SDNN < 30 ms: Posible fatiga, sobreentrenamiento, o estrés crónico

Paso 4: Cálculos

Correlación entre variables derivadas:

	Act_rel	**Sup_cal**	**HRV**	**Delta_card**
Act_rel	1.00	0.68	0.12	0.24
Sup_cal	0.68	1.00	0.09	0.31
HRV	0.12	0.09	1.00	0.18
Delta_card	0.24	0.31	0.18	1.00

Tabla 7.3: Matriz de Correlación (Spearman, $n = 8,380$ días)

Observación: Correlación moderada Act_rel – Sup_cal (esperada: ambas reflejan volumen de actividad), pero baja con variables cardiovasculares, confirmando que capturan dominios distintos.

Paso 6: Conclusión

Conclusión:
Las 4 variables derivadas son:

1. Antropométricamente normalizadas (comparabilidad)
2. Fisiológicamente interpretables (relevancia clínica)
3. Relativamente independientes ($r < 0{,}70$, evitando multicolinealidad severa)

Estas formarán la base para la agregación semanal (siguiente capítulo) y posterior modelado.

Capítulo 8

Agregación Temporal y Análisis Dual de Variabilidad

8.1. Justificación de la Agregación Semanal

Paso 1: Planteamiento de Hipótesis

Hipótesis:
Los datos diarios presentan una variabilidad excesiva (CV $> 50\,\%$) atribuible a:

- Comportamientos esporádicos (ejercicio intenso 1 día, sedentarismo el siguiente)
- Ruido de medición (errores de sensor, eventos atípicos)
- Ciclos semanales (diferencias fin de semana vs días laborales)

La agregación a nivel semanal (7 días continuos) utilizando estadísticos robustos (mediana, IQR) capturará el *patrón habitual* de comportamiento, reduciendo ruido y mejorando estabilidad para clustering/modelado.

8.1.1. Ventana de Agregación

$$\text{Semana } k: \quad \text{fecha\_inicio} = \text{Lunes}, \quad \text{fecha\_fin} = \text{Domingo} \tag{8.1}$$

Criterio de validez: Semana incluida si ≥ 5 días tienen datos completos ($71\,\%$ completitud).

8.2. Estadísticos Calculados por Semana

Para cada una de las 4 variables derivadas:

$$x_{\text{p50}}^{(k)} = \text{median}\{x_{\text{día}_1}, x_{\text{día}_2}, \ldots, x_{\text{día}_7}\} \tag{8.2}$$
$$x_{\text{IQR}}^{(k)} = Q_3(x) - Q_1(x) \tag{8.3}$$
$$x_{\text{p10}}^{(k)} = \text{percentil}_{10}(x) \tag{8.4}$$
$$x_{\text{p90}}^{(k)} = \text{percentil}_{90}(x) \tag{8.5}$$

Resultado: Dataset semanal con $n_{\text{semanas}} = 1,337$ (válidas) y 16 features (4 variables $\times$ 4 estadísticos).

8.3. Análisis Dual de Variabilidad

8.3.1. Definición de Variabilidad Observada vs Operativa

Paso 2: Selección del Estadístico/Método

Variabilidad Observada (datos crudos, sin imputar):
Cuantifica la fluctuación natural día-a-día medida directamente por el sensor.

$$\text{CV}_{\text{obs}}^{(u,v)} = \frac{\sigma_{\text{obs}}(v,u)}{\mu_{\text{obs}}(v,u)} \times 100\,\% \tag{8.6}$$

donde v = variable, u = usuario.
Variabilidad Operativa (datos post-imputación):
Refleja la variabilidad utilizada en el análisis final.

$$\text{CV}_{\text{op}}^{(u,v)} = \frac{\sigma_{\text{op}}(v,u)}{\mu_{\text{op}}(v,u)} \times 100\,\% \tag{8.7}$$

8.3.2. Comparación Observada vs Operativa

Tabla 8.1: Coeficiente de Variación: Observado vs Operativo (promedio 10 usuarios)

Variable	CV obs (%)	CV op (%)	ΔCV (%)	Dir.	Efecto impute
Pasos	62.3	59.8	-2.5	↓	Suaviza
Actividad_relativa	58.7	56.4	-2.3	↓	Suaviza
Calorías_activas	74.5	71.2	-3.3	↓	Suaviza
Superávit_calórico	68.9	66.1	-2.8	↓	Suaviza
FC_reposo	14.2	13.8	-0.4	↓	Mínimo
FC_caminar	11.8	13.1	+1.3	↑	Leve aumento
HRV_SDNN	35.4	32.7	-2.7	↓	Suaviza
Delta_cardiaco	15.6	16.2	+0.6	↑	Leve aumento

Paso 5: Decisión Estadística

Decisión:
La imputación tiene un impacto moderado ($|\Delta CV| < 5\,\%$), tendiendo a *reducir* ligeramente la dispersión (efecto de regresión a la media en métodos basados en medianas). El aumento en FC_caminar y Delta_cardiaco es marginal ($< 2\,\%$) y aceptable.
Conclusión: La imputación no distorsiona dramáticamente las distribuciones; los datos operativos son representativos de los observados.

8.3.3. Gráficos de Variabilidad

Ver Figuras:

- `4 semestre_dataset/variabilidad_operativa_vs_observada.png`: Comparación global
- `4 semestre_dataset/variabilidad_por_usuario_boxplot.png`: Distribución por individuo
- `4 semestre_dataset/heatmap_cv_usuario_variable.png`: Mapa de calor CV
- `4 semestre_dataset/analisis_u/variabilidad/CV_por_usuario_u1.png`: Desglose usuario 1

8.4. Agregación Semanal: Resultados Finales

Paso 4: Cálculos

Dataset semanal generado:

- Archivo: `DB_usuarios_consolidada_con_actividad_relativa.csv`
- Dimensiones: $1,337 \times 18$ (16 features + usuario_id + semana_inicio)
- Completitud: 100 % (post-imputación y agregación)

Estadísticos de las 4 variables p50 (para clustering/fuzzy):

Variable p50	Mediana global	IQR global	Min	Max
Actividad_relativa	0.58	0.31	0.02	1.87
Superávit_calórico	29.4	18.7	1.2	98.5
HRV_SDNN	48.2	21.5	18.3	112.7
Delta_cardiaco	36.8	14.2	8.5	78.4

Tabla 8.2: Estadísticos del Dataset Semanal (n=1,337 semanas)

Paso 6: Conclusión

Conclusión del capítulo:

1. La agregación semanal reduce efectivamente el ruido diario.
2. El análisis dual de variabilidad confirma que la imputación no introduce artefactos severos.
3. El dataset semanal con 4 variables p50 + 4 IQRs está listo para el clustering (Capítulo 9) y modelado difuso (Capítulo 10).

Capítulo 9

Análisis de Correlación, Multicolinealidad y Reducción Dimensional (PCA)

Capítulo 10

Clustering No Supervisado: Verdad Operativa (K-Means, K=2)

Capítulo 11

Sistema de Inferencia Difusa Mamdani

Capítulo 12

Validación Cruzada y Análisis de Robustez

Capítulo 13

Justificación Metodológica: Por Qué NO Split Train/Test 80/20

Bibliografía

[1] World Health Organization. (2020). *WHO guidelines on physical activity and sedentary behaviour*. Geneva: World Health Organization.

[2] Stahl, S. E., et al. (2016). How accurate are the wrist-based heart rate monitors during walking and running activities? Are they accurate enough? *BMJ Open Sport & Exercise Medicine*, 2(1), e000106.

[3] Shcherbina, A., et al. (2017). Accuracy in wrist-worn, sensor-based measurements of heart rate and energy expenditure in a diverse cohort. *Journal of Personalized Medicine*, 7(2), 3.

[4] Little, R. J. (1988). A test of missing completely at random for multivariate data with missing values. *Journal of the American Statistical Association*, 83(404), 1198-1202.

[5] Pedregosa, F., et al. (2011). Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12, 2825-2830.

[6] Zadeh, L. A. (1965). Fuzzy sets. *Information and Control*, 8(3), 338-353.

[7] Mamdani, E. H., & Assilian, S. (1975). An experiment in linguistic synthesis with a fuzzy logic controller. *International Journal of Man-Machine Studies*, 7(1), 1-13.