

Guía de Mejora de Datos — Portal Ortodoncia

Fecha: 06-sep-2025

Autor: Santiago Tupper

Contexto: Proyecto de segmentación de pacientes con clustering (K=3) — foco en reproducibilidad, interpretabilidad y privacidad.

1) Objetivo

Alinear reglas y prioridades para **mejorar la calidad de datos** que alimenta el análisis (Tab_Clientes y Prestaciones), con el fin de: - aumentar la confiabilidad de métricas (p. ej., % cumplimiento, recencia), - facilitar segmentaciones accionables (retención, reactivación, depuración), - y reducir retrabajo entre 'dashboard.pbix' y Analytics.

2) Alcance

- **Incluye:** Fechas (histórico vs planificación), `DiasDesdeUltimaVisita`, KPIs de presupuesto, presencias 15d-6m, geografía (Comuna/Región), Empresa/Convenio, vínculo ppto↔pago (Prestaciones).
 - **No incluye (por ahora):** detalle clínico por prestación, rendimiento por profesional, costos.
-

3) Hallazgos de calidad (resumen)

Fechas - Histórico en ISO; planificación en **DD/MM/AAAA** con **fechas futuras** (baja incidencia).

- Se crearon flags de planificación y `dias_hasta` (mediana ≈ 65 días; máx ≈ 114).
- Acción: mantener doble parsing y flags; consolidar diccionario de campos fecha.

DiasDesdeUltimaVisita - Validado vs recálculo; **offset sistemático +11 días** (100% positivo en distribución).

- Decisión: usar el **valor del sistema** y documentar offset; revisar origen junto a TI.

KPIs de presupuestos - Colas largas (outliers) en montos → se usa **log1p** .

- TicketPromPpto incluido; **NaN→0 solo para el modelo** (interpretación "sin actividad").
- Acción: estandarizar reglas de KPIs y denominadores protegidos.

Presencias de atención (uso real) - Ventanas con mejor señal: **15d, 1m, 3m, 6m**.

- Acción: mantener solo estas cuatro; descartar 1a/2a en el set de modelado.

Geografía - Comuna_grp (Top-N + Otras/Infreq + Sin Comuna) y Region con dummies.

- >95% de pacientes en **Región Metropolitana**; C2 tiene **registros incompletos** (Sin Región 30%).
- Acción: reforzar captura de Región/Comuna y mapeo maestro.

Empresa/Convenio - Cobertura desigual; **no aportó separación adicional** en clustering (A/B).

- Acción: mantener como metadato; no forzar su uso en el modelo si no mejora calidad.

Prestaciones (cohortes 2025) - ppto=4.564, pago=4.862, ambos=3.951 → **tasa 91%**.

- Acción: fijar reglas de estados válidos (DIAGNOSTICADA/INICIADA) y consistencia ppto↔OT.

4) Reglas acordadas (versión aplicable)

- **Fechas:** parseo dual (ISO vs DD/MM/AAAA) + flags de planificación; conservar dias_hasta .
- **DiasDesdeUltimaVisita:** usar valor del sistema; registrar offset +11 en documentación técnica.
- **KPIs:** log1p en montos; salvaguardas de denominador; NaN→0 solo para features de modelado.
- **Ventanas:** solo presencias **15d, 1m, 3m, 6m**.
- **Geografía:** mantener Comuna_grp y Region (one-hot); controlar "Sin Comuna/Región".
- **Empresa/Convenio:** no se usa para separar clusters (respaldo A/B); se conserva como contexto.
- **Prestaciones:** cohortes por año; estados válidos DIAGNOSTICADA/INICIADA; controles ppto↔OT.

5) Recomendaciones priorizadas

Prioridad	Recomendación	Detalle	Éxito (cómo se mide)
MUST	Normalizar fechas	Diccionario de campos + validación automática de formato	0% errores de parseo; flags coherentes
MUST	Revisar origen de offset +11 días	Revisión con TI del cálculo de <code>DiasDesdeUltimaVisita</code>	Offset documentado/ajustado
MUST	Estandarizar KPIs	Reglas compartidas con 'dashboard.pbix' (denominadores, <code>log1p</code>)	Discrepancias < $\pm 1\%$
SHOULD	Completar geografía	Reducción de "Sin Comuna/Región"; maestro de comunas	<2% "Sin Comuna/Región"
SHOULD	Consolidar vínculo ppto↔pago	QA por OT y por paciente/año	Consistencias > 98%
COULD	Reevaluar Empresa/Convenio	Mejoras de calidad + nueva A/B	Sólo se usa si mejora métricas

6) Roadmap sugerido

- Semana 1–2:** Normalización de fechas y documentación del offset de `DiasDesdeUltimaVisita`.
- Semana 3:** Reglas de KPIs consensuadas entre Analytics y 'dashboard.pbix' (one-pager).
- Semana 4–5:** Limpieza geográfica y maestro de comunas/regiones.
- Semana 6:** QA ppto↔pago (Prestaciones) y reporte de consistencia.

7) Métricas de éxito

- Errores de parseo de fechas = 0%.
- Discrepancia de KPIs ('dashboard.pbix' vs Analytics) < $\pm 1\%$.
- "Sin Región/Comuna" < 2%.
- Consistencia ppto↔pago > 98%.
- Trazabilidad reproducible de IDs (RutBeneficiario, RutTitular).

8) Anexos

- **Metodología:** pipeline mixto de escalado (Standard/Robust/passthrough).
 - **Clustering:** K=3; baseline basta (Empresa no agrega separación).
 - **Archivos relacionados:**
 - /reports/nb04_insights_executive.pdf
 - /reports/nb04_dashboard_vs_clustering.csv
 - /reports/nb06_prestaciones_validacion.csv
-