**Flujo de datos de validación georreferenciada**

1. Ingreso de datos

* Carga de dataset anterior y nuevo
* Confirmación del rango de fechas
* Validación de estructura y formatos de entrada
* Ejecución de test automatizados

2. Limpieza y transformación

* Verificación y homologación de nombres de columnas
* Validación de datos por columna (valores esperados, duplicados, tipos)
* Eliminación de columnas irrelevantes o vacías
* Renombrado de columnas (estandarización semántica)
* Normalización de columnas (mayúsculas, espacios, separadores) y unificación de categorías entre datasets
* Normalización de valores dentro de columnas (mayúsculas, espacios, separadores)
* Conversión de tipos de datos críticos (string, datetime, numérico)
* Registro de logs de limpieza y errores detectados

3. Integración

* Concatenación vertical de los datasets limpios
* Creación de identificador único o índice secuencial
* Validación de integridad post-concatenación
* Ejecución de test automatizados de integridad

4. Filtrado de registros

* Eliminación de valores nulos en coordenadas
* Exclusión de coordenadas fuera del rango geográfico del Perú
* Retención únicamente de visitas “VD Efectiva”
* Depuración de duplicados por claves primarias

5. Cruce

* Merge con base maestro\_hogares (coordenadas hogares X/Y)
* Conteo de valores faltantes post-merge en coordenadas hogares
* Eliminación de registros sin coordenadas
* Cálculo de distancia Haversine (visita ↔ hogar) en metros y kilómetros

6. Clasificación y validación de visitas

* Asignación de categoría territorial (URBANO / ANDINO / AMAZÓNICO)
* Aplicación de reglas de validación de distancia: Urbano ≤ 0.5 km, Andino ≤ 2 km y Amazónico ≤ 5 km
* Generación de variable de consistencia: VALIDA / INCONSISTENTE

7. Agregación y clasificación de gestores

* Agrupación por gestor local (DNI\_GEL / GEL)
* Cálculo del número de visitas válidas e inconsistentes por gestor
* Cálculo del % de inconsistencia y validez
* Asignación de nivel de riesgo de inconsistencia: Bajo <30%, Medio 30–49%, Alto 50–69% y Crítico ≥70%

8. Análisis exploratorio (EDA)

* Tablas y gráfico de %válidas e inconsistentes por categoría territorial y total nacional
* Histograma y estadísticos descriptivos de distancias (p25, p50, p75, p90, p95)
* Evolución mensual de visitas válidas e inconsistentes
* Ranking de UT por %inconsistencias
* Histograma y estadísticos descriptivos del % de inconsistencias por gestor
* Tablas de gestores por nivel de riesgo (nacional y último periodo)
* Ranking de UT por gestores con nivel de riesgo elevado.
* Evolución diaria promedio de visitas totales y % inconsistentes.

9. Anonimización y exportación

* Eliminación de datos sensibles (DNI, nombres, coordenadas del hogar, etc)
* Generación del dataset anonimizado: df\_seguro.csv
* Validación final de integridad y consistencia

10. Visualización operativa (Streamlit Dashboard)

* Carga cacheada del dataset seguro
* Definición de periodos institucionales (Dic 2024 → Ago 2025) y filtrado de prioridad 4–5
* Filtros interactivos: Periodo/ UT / Distrito
* Semaforización del riesgo por gestor
* Tarjeta de resumen contextual (colores por nivel de riesgo)
* KPIs globales y ranking de gestores por riesgo (≥5 visitas)
* Tabla detallada de visitas por hogar con filtros dinámicos.