UNIVERSIDAD TECNOLÓGICA DE CHIHUAHUA

DESARROLLO Y GESTIÓN DE SOFTWARE

REPORTE DE LIMPIEZA DE DATOS

EXTRACCIÓN DE CONOCIMIENTO EN BASES DE DATOS

PRESENTA:

KARLA ALEJANDRA DE LA CRUZ ZEA

DOCENTE:

ING. LUIS ENRIQUE MASCOTE CANO

4 de octubre de 2025

Contenido

Introducción

En este reporte mostrare la procedencia de los datos, su clasificación por tipos y fuentes, así como las técnicas de limpieza mas utilizadas generalmente en los proyectos de análisis de datos. Se presentará también un caso práctico para identificar problemas comunes en conjuntos de datos y proponer acciones de resolución.

1. Procedencia de los datos

Los datos pueden provenir de múltiples orígenes.

Datos biométricos

Los datos biométricos son mediciones de características físicas o biológicas de personas (huellas, reconocimiento facial, ritmo cardíaco).

Ejemplos

Dispositivos wearables (smartwatches), sistemas de control de acceso, laboratorios clínicos.

Máquina a máquina (M2M)

El lenguaje de máquina a máquina se refiere a los datos intercambiados entre dispositivos sin intervención humana directa.

Ejemplos

Sensores IoT, telemetría de equipos industriales, logs de sistemas embebidos.

Datos de transacciones

Se le llama datos de transacciones a los registros de operaciones comerciales o financieras que documentan eventos (ventas, pagos, transacciones bancarias).

Ejemplos

Sistemas POS, bases de datos financieras, logs de comercio electrónico.

Datos generados por humanos

Se dice que los datos generados por humanos es aquella información introducida deliberadamente por personas (formularios, encuestas, informes).

Ejemplos

Formularios web, registros de atención al cliente, notas manuales.

Datos web

Los datos web son contenido obtenido de sitios web, APIs y portales (estructurado o no).

Ejemplos

APIs públicas (datos abiertos), páginas HTML, archivos descargados.

Medios sociales

Los medios sociales son datos procedentes de plataformas sociales (redes sociales, publicaciones, comentarios, reacciones etc).

Ejemplos

Twitter/X API, Facebook Graph API, posts y comentarios públicos.

2. Tipos y clasificación de los datos

Categoría	Tipo (cuant./cual.)	Formato (estruc./no estruc.)	Ejemplo (procedencia)
Numéricos continuos	Cuantitativo	Estructurado	PM2.5 de sensores IoT (M2M)
Numéricos discretos	Cuantitativo	Estructurado	Conteo de transacciones (Transaccional)
Categóricos nominales	Cualitativo	Estructurado	Género, categoría de producto (Formularios)
Categóricos ordinales	Cualitativo	Estructurado	Niveles de satisfacción (encuestas)
Texto libre	Cualitativo	No estructurado	Comentarios en redes sociales (Social)
Imágenes/Audio/Señales	Cualitativo	No estructurado	Imágenes médicas, fotos de vigilancia (Biométricos/Web)
Logs/Telemetría	Mixto	Semi-estructurado	Registros de máquinas (M2M)

3. Técnicas de limpieza de datos

Los conjuntos de datos reales suelen presentar problemas como valores nulos, duplicados, errores de formato y valores atípicos. Veamos entonces las técnicas más relevantes y acciones recomendadas aplicadas a un caso de estudio simulado.

Valores nulos (missing values)

Conteo por columna (isnull(), describe).

Causas comunes: fallas en la captura, sensores offline, campos opcionales.

Acciones correctivas: imputación (media/mediana para numéricos; moda o categoría "Desconocido" para categóricos), eliminación de registros si la proporción es alta, usar modelos de imputación (KNN, MICE) cuando sea crítico.

Duplicados

Identificación: detección por claves primarias o huella (hash) del registro.

Causas: reenvíos, integración de fuentes múltiples.

Acciones: eliminar duplicados exactos; consolidar duplicados parciales (deduplicación basada en reglas o clustering por similitud) conservando la versión más completa.

Errores de formato y consistencia

Identificación: validación de tipos, patrones (fechas, códigos).

Acciones: normalizar formatos (ISO-8601 para fechas), conversión de unidades, estandarización de categorías (mapping), validaciones con expresiones regulares.

Valores atípicos (outliers)

Identificación: métodos estadísticos (IQR, z-score), visualización (boxplots, scatter).

Acciones: investigar origen (error sensor vs evento real), truncamiento o winsorizing, modelado robusto o mantener con flag de anomalía si es informativo.

Datos faltantes por diseño y datos implícitos

Considerar cuando los valores faltan por decisión (por ejemplo, campo no aplicable).

Acciones: documentar y diferenciar "missing at random" vs "missing not at random"; imputación condicionada o uso de variables indicadoras (missing flag).

Limpieza de texto y datos no estructurados

Técnicas: tokenización, lematización, eliminación de stopwords, corrección ortográfica, normalización Unicode.
Herramientas comunes: NLTK, spaCy, regex; para deduplicación de texto usar similitud coseno o fuzzy matching.

4. Ejemplo práctico

Un conjunto de datos de 10,000 registros de sensores ambientales (PM2.5, temperatura, humedad) + 2,000 registros de reportes ciudadanos (texto). Problemas detectados: 8% valores nulos en PM2.5, duplicados por reenvío (1.2%), fechas en formatos mixtos, y picos atípicos en 0.5% de lecturas.

Pasos propuestos:

- Auditoría inicial: reportes de calidad por columna (missing, unique, min/max).
- Normalización de fechas y unidades.
- Imputación: mediana para PM2.5 y temperatura; modelo KNN si la imputación simple afecta tendencias.
- Eliminación y consolidación de duplicados, manteniendo timestamp más reciente.
- Detección de outliers: marcar eventos extremos y verificar con sensores vecinos; conservar si se confirma evento real.
- Preprocesamiento de texto: limpieza, lematización y vectorización antes de análisis de sentimiento.

Conclusión

En este documento reconocimos que los datos pueden provenir de múltiples fuentes y que su clasificación, es necesaria para decidir cómo analizarlos. También destacamos que la calidad de los datos son comunes : problemas como valores nulos, duplicados o inconsistencias afectan directamente los resultados y requieren aplicar técnicas de limpieza adecuadas.

Entonces podemos decir que un buen análisis de datos no depende solo de algoritmos, sino que también de una gestión adecuada de la procedencia, clasificación y depuración de la información. Todo esto para garantizar una mayor confiabilidad en los resultados y aportar valor real en proyectos de análisis, inteligencia artificial y big data.

Fuentes

Rahm, E., & Do, H. H. (2000). Data Cleaning: Problems and Current Approaches. Retrieved from https://www.researchgate.net/publication/220282831_Data_Cleaning_Problems_and_Current_Approaches

Batini, C., & Scannapieco, M. (2006). Data Quality: Concepts, Methodologies and Techniques. Springer.

Kwak, S.-K., & Kim, J. H. (2017). Statistical data preparation: management of missing values and outliers. PMC article. Retrieved from https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5548942/

IBM. (2024). What is Data Provenance? IBM. Retrieved from https://www.ibm.com/think/topics/data-provenance