# TALLER PREPROCESAMIENTO DE DATOS

## Objetivo

Desarrollar un proceso integral de exploración, caracterización, preprocesamiento y estructuración de datos, a partir de al menos tres fuentes distintas, con el propósito de dejarlos preparados para su posterior uso en técnicas de aprendizaje supervisado o no supervisado.

**Objetivos específicos:**

- Comprender la importancia del preprocesamiento en el ciclo de minería de datos.
- Analizar críticamente la calidad y estructura de los datos.
- Diseñar un esquema de almacenamiento adecuado.
- Documentar técnicamente el proceso realizado.

## Fechas

- Entrega informe: 4 de marzo (23:59)
- Sustentación: 5 de marzo (11:00)

## Planteamiento: Desarrollo del Taller

### *Paso 1.* Selección del dominio y fuentes de datos

Teniendo en cuenta las fuentes de datos, repositorios y dataset revisados en clases previas:

- Seleccionar una temática o área de estudio.
- Identificar al menos tres fuentes de datos diferentes.
- El conjunto total debe contener mínimo 10.000 registros/instancias.
- Las fuentes pueden ser repositorios abiertos, APIs, portales institucionales, bases públicas, etc.

Debe justificarse la selección del dominio y las fuentes.

### *Paso 2.* Exploración inicial y contextualización

Hacer un proceso de exploración inicial de los datos y de la documentación asociada. Pueden responder a preguntas como las siguientes para guiar este proceso:

- A qué tipo de datos corresponden: ¿transaccionales, series temporales, espaciales, matriciales, documentos, grafos, etc?
- ¿Son datos temporales? ¿Desde qué año/periodo tengo registro de datos?
- ¿Estos datos cuentan con diccionario de datos o metadatos?
- ¿El propietario de los datos es conocido? ¿Tendría como comunicarme con él?

- ¿Tengo dominio medio o avanzado en el área o temática seleccionada?
- ¿Qué tipo de análisis creo que podrían ser de interés con estos datos? Plantear mínimo tres preguntas que se puedan responder a partir de estos datos.

***Nota:*** Estas preguntas son una guía, pero la exploración inicial no se debe limitar a ellas. Pueden ampliar. Este paso debe evidenciar comprensión estructural y contextual del dataset.

***Paso 3.*** **Caracterización estructural y diseño de almacenamiento**

Teniendo en cuenta la documentación (diccionarios de datos) y la exploración inicial, determinar:

- Variables o atributos (es necesario describirlos todos)
- Tipo de dato (numérico, categórico, ordinal, binario, fecha, texto, etc.).
- Formatos de dichos atributos
- Nivel de granularidad (son datos a nivel de individuos, de institución, de municipio, de departamento)
- Escalas
- Fuentes

Adicionalmente, proponer un esquema de almacenamiento para los datos (ej. datawarehouse). Hacer el diseño del modelo de datos y justificar técnicamente la selección.

***Paso 4.*** **Preprocesamiento y limpieza de datos**

Identificar y solucionar problemas en los datos. Hacer un preprocesamiento detallado en el que se identifiquen y propongan soluciones a problemas como:

- Datos faltantes
- Datos atípicos
- Datos duplicados
- Formato de datos incorrecto
- Problemas con la ortografía o digitación
- Caracteres especiales
- Problemas de codificación
- Transformaciones necesarias (estandarización, codificación categórica, etc.).

***Nota:*** este paso puede ser realizado apoyados en alguna herramienta de análisis o lenguaje de programación. Se debe especificar: Herramientas utilizadas, criterios de decisión aplicados y automatización implementada (si aplica).

***Paso 5.*** **Análisis descriptivo**

Hacer un análisis descriptivo de los datos seleccionados (medidas de tendencia central y de dispersión, distribuciones, visualizaciones pertinentes). Con este análisis revisar y proponer algunas reglas básicas de filtrado o rangos válidos y confiables, posibles restricciones de calidad de los datos, que posteriormente sirvan para seguir pre-procesando los datos de este dominio.

***Paso 6.*** **Reflexión técnica y recomendaciones**

Documentar todo el proceso llevado a cabo y dar recomendaciones atendiendo a preguntas como:

- ¿Es posible establecer un rango de datos confiables para cada una de las variables o atributos presentes en los datos?
- ¿Dentro del proceso llevado a cabo, cuál fue el paso que más demandante y por qué?
- ¿Lograron automatizar alguna parte del proceso?
- Si en este momento se les pidiese hacer el procesamiento para una nueva fuente de este dominio, ¿sería igual de complejo? ¿O por qué disminuiría el esfuerzo? ¿Qué tan escalable sería el proceso ante nuevas fuentes?
- Recomendaciones para una futura aplicación de modelos de aprendizaje automático.

**Notas generales:**

- Tendrán puntos extra (0.3 en la nota computada de la entrega grupal y la sustentación individual) los equipos que adicionalmente a proponer el esquema de almacenamiento:
    - Implementen el esquema de almacenamiento en un motor de base de datos.
    - Realicen carga efectiva de los datos procesados.

**Informe ejecutivo**

El informe debe presentarse en formato técnico y contener:

1. Portada
2. Resumen ejecutivo (máximo 1 página)
3. Introducción
4. Descripción del dominio y fuentes
5. Exploración inicial
6. Caracterización estructural de los datos
7. Diseño del modelo de almacenamiento (con diagrama)
8. Proceso de preprocesamiento (metodología aplicada)
9. Análisis descriptivo
10. Resultados del preprocesamiento

11. Reflexión técnica y recomendaciones
12. Conclusiones
13. Referencias
14. Anexos (código, scripts, tablas extendidas, etc.)

Debe evidenciar rigor técnico, trazabilidad del proceso y claridad argumentativa.

**Sustentación**

Se realizará de forma individual el día 5 de marzo en el horario habitual de la clase (11-13)