Big Data: Preprocesamiento y Calidad de Datos

Sergio M. Nava Muñoz

s3rgio.nava@gmail.com

CIMAT/INFOTEC

2025-02-19

Introducción

¿Qué es Big Data?

- Conjunto de datos masivos con características clave:
  - **Volumen**: Cantidad de datos enorme.
  - **Velocidad**: Generación y procesamiento rápido.
  - **Variedad**: Diversidad de formatos y fuentes.
  - **Veracidad**: Calidad y confiabilidad de los datos.
  - **Valor**: Información útil extraída.
- **Big Data vs. Smart Data:** Necesidad de filtrar y transformar datos en conocimiento útil.

> **Note**
>
> Big Data **sin preprocesamiento** genera modelos con bajo rendimiento.

Importancia del Preprocesamiento

Problemas en los datos masivos

- Ruido
- Valores perdidos
- Inconsistencias
- Datos redundantes o irrelevantes
- Alta dimensionalidad

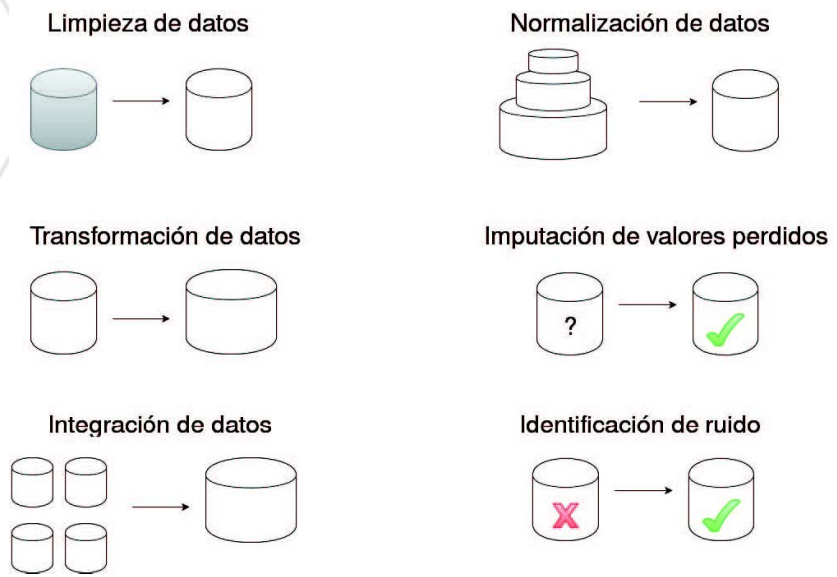
Objetivo del Preprocesamiento

- **Mejorar la calidad de los datos**
- **Reducir costos computacionales**
- **Optimizar la eficiencia de los modelos de minería de datos**

Técnicas de Preprocesamiento

1. Preparación de Datos

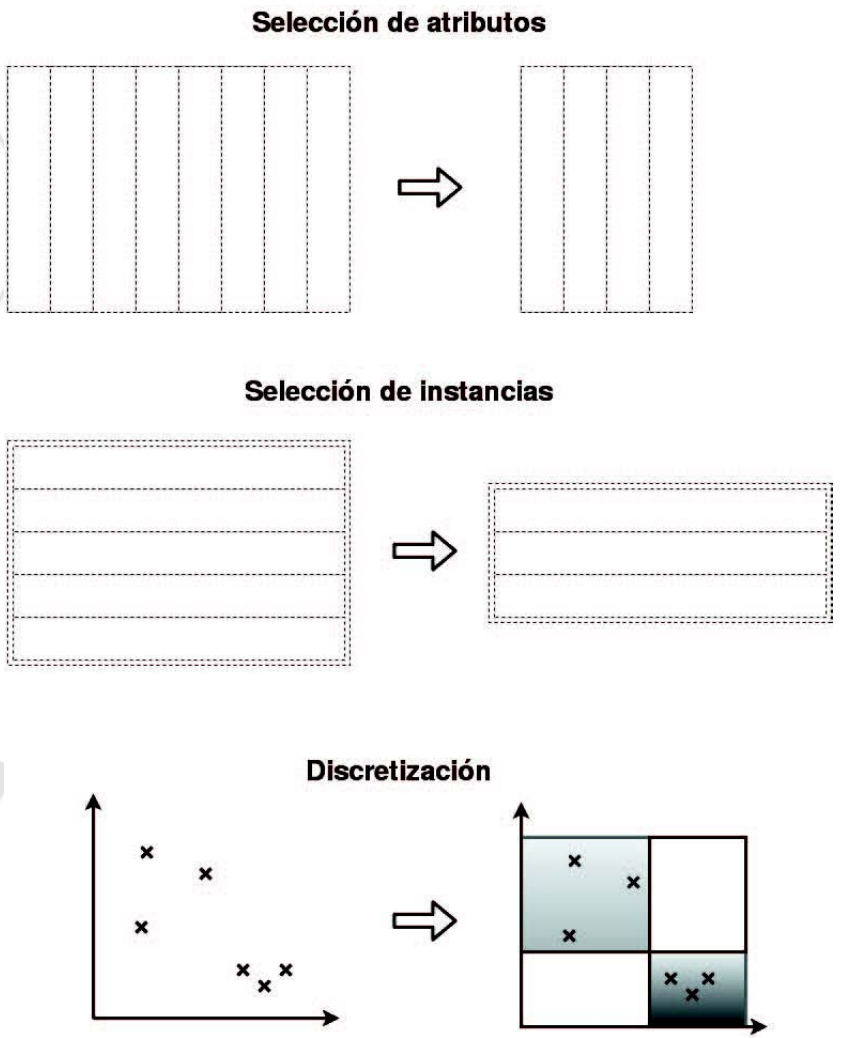
- **Limpieza de datos**: Eliminación de ruido y valores atípicos.
- **Normalización y transformación**: Ajuste de escalas y formatos.
- **Imputación de valores perdidos**: Métodos estadísticos y de machine learning.



Reduccion de datos

2. Reducción de Datos

- **Selección de atributos (Feature Selection)**
- **Selección de instancias**
- **Discretización**



Reduccion de datos

Tecnologías para Big Data

Plataformas

- **Hadoop**: Sistema de archivos distribuido con MapReduce.
- **Spark**: Procesamiento en memoria, más rápido que Hadoop.
- **Flink**: Procesamiento de flujos de datos en tiempo real.

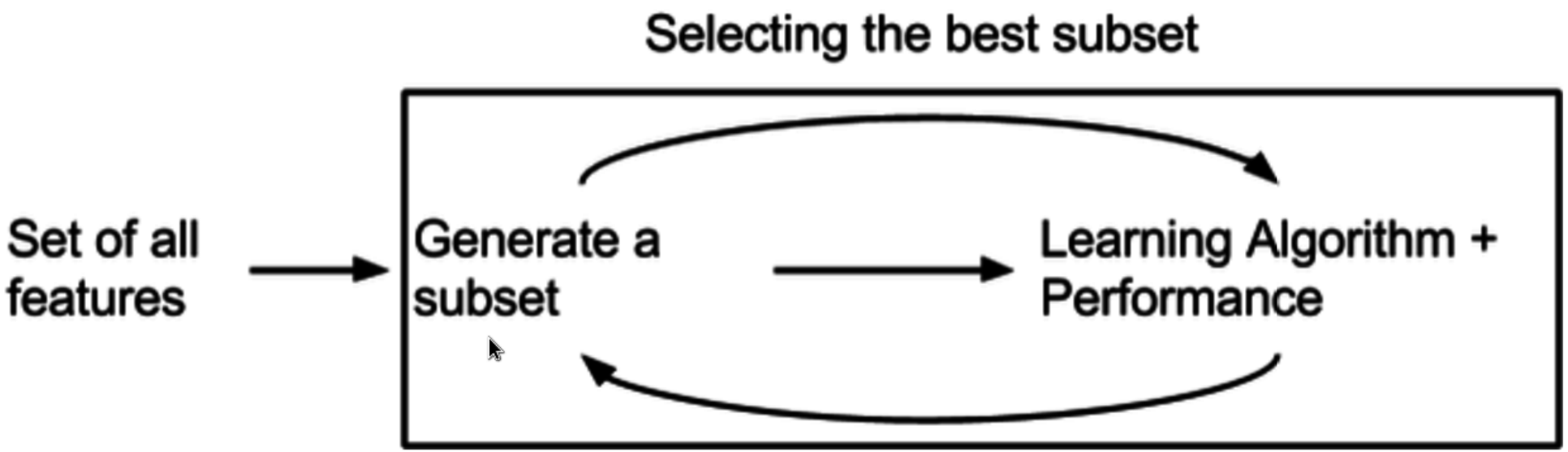
Herramientas de Analítica

- **MLlib**: Librería de Machine Learning en Spark.
- **Mahout**: Algoritmos en Hadoop.
- **FlinkML**: Aprendizaje automático en Flink.
- **H2O**: Modelado avanzado, incluyendo Deep Learning.

Algoritmos de Preprocesamiento

Métodos de Selección de Atributos

- **Chi-cuadrado**
- **PCA (Análisis de Componentes Principales)**
- **Fast-mRMR (Minimum Redundancy Maximum Relevance)**



Proceso de Selección de Atributos

Conclusiones y Retos Futuros

Conclusiones

- El preprocesamiento es **clave** en Big Data.
- Avances hacia **Smart Data**: datos limpios y optimizados.

Retos Futuros

- Desarrollo de nuevos algoritmos escalables.
- Integración con inteligencia artificial para procesamiento autónomo.
- Mayor eficiencia en la imputación de valores faltantes y reducción de dimensionalidad.

> ⚠ **Warning**
>
> “¡Decisiones de calidad requieren datos de calidad!”

Referencias

- García et al. (2016). “Big Data: Preprocesamiento y Calidad de Datos”.
- Wu et al. (2014). “Data Mining with Big Data”.
- Dean & Ghemawat (2004). “MapReduce: Simplified Data Processing”.
- Meng et al. (2016). “MLlib: Machine Learning in Apache Spark”.

¡Gracias por su atención! 🎯