

Preprocesamiento de Datos en Machine Learning: EDA y ETL

Abraham Jain Jiménez

Necesidad del preprocesamiento de datos

El preprocesamiento es una etapa fundamental en el Machine Learning, pues es donde se limpian y transforman los datos, adecuándolos para el correcto funcionamiento de los modelos que los reciben.

Recordemos que para implementar modelos de ML con la biblioteca `scikit-learn`, estos requieren bases de datos:

- Sin *missing values*.
- En formato numérico.
- Almacenadas en un `DataFrame` de `Pandas` o un arreglo `NumPy`.

Además, un preprocesamiento adecuado evita que los algoritmos puedan producir resultados erróneos, poco robustos o con bajo rendimiento.

Tareas Comunes de Preprocesamiento (EDA y ETL)

- **Manejo de valores nulos:** eliminación o imputación de datos faltantes.
- **Escalamiento de datos:** normalizar magnitudes para mejorar el desempeño de los modelos.
- **Normalización:** ajustar las distribuciones para mejorar la estabilidad numérica.
- **Imputación:** rellenar valores faltantes mediante estrategias estadísticas.
- **Estandarización:** transformar los datos para que tengan media cero y varianza uno.
- **One Hot Encoding:** convertir variables categóricas en variables binarias.
- **Binary Encoding:** codificación alternativa para categorías con muchas clases.
- **Tratamiento de outliers:** detección y mitigación de valores extremos.
- **Reducción de cardinalidad y redundancia:** simplificación de variables categóricas.
- **Limpieza de estructura:** tratamiento de columnas mal formateadas o filas repetidas.
- **Combinación de variables:** ingeniería de características para mejorar el poder predictivo.
- **Binning Encoding:** discretización de variables continuas en intervalos.
- **Documentación de insights:** descripción de hallazgos clave durante el análisis exploratorio.