Consideraciones para la Selección de Datos en Modelos de Regresión y Clasificación

Introducción

La calidad y adecuación de los datos son factores determinantes en el rendimiento de los modelos de aprendizaje automático. La correcta selección del conjunto de datos no solo afecta la precisión del modelo, sino también su capacidad de generalización, su eficiencia en entrenamiento y su aplicabilidad real. Este documento describe los criterios clave para seleccionar datasets apropiados para tareas de regresión y clasificación, considerando el tipo de variables, tamaño del conjunto de datos, número de características y otros aspectos relevantes.

1. Tipo de Problema: Regresión vs Clasificación

- Regresión: El objetivo es predecir un valor numérico continuo. Por lo tanto, la variable objetivo (target) debe ser de tipo numérico (entero o decimal).
- Clasificación: El objetivo es predecir una clase o categoría. La variable objetivo es categórica y puede ser binaria (dos clases) o multiclase (tres o más clases).

2. Tipos de Datos Recomendados

Regresión:
- Variables predictoras: preferiblemente numéricas continuas o discretas (pueden incluir categóricas transformadas).
- Variable objetivo: numérica continua.
- Ejemplos: precios, temperaturas, tiempo, edad.

Clasificación:
- Variables predictoras: pueden ser numéricas o categóricas.
- Variable objetivo: categórica.
- Ejemplos: diagnósticos médicos, tipos de productos, clases de imágenes.

3. Tamaño del Conjunto de Datos

- Se recomienda tener al menos 10 veces más instancias que el número de características.
- Regresión: idealmente de 500 a 1000 instancias o más.
- Clasificación: puede funcionar con menos datos, pero requiere balance entre clases.

4. Número y Tipo de Columnas (Características)

- Buena proporción entre número de columnas y filas.
- Variables relevantes y no redundantes.
- Evitar o tratar valores faltantes y normalizar si es necesario.
- Variables categóricas deben codificarse (one-hot o label encoding).

5. Calidad de los Datos

- Ausencia o tratamiento de valores nulos.
- Control de valores atípicos (outliers).
- Balance en las clases (para clasificación).
- Distribución adecuada de la variable objetivo.

6. Consideraciones Éticas y Contextuales

- Representatividad de la muestra.
- Evaluar y mitigar sesgos presentes.
- Cumplir normativas de privacidad y uso de datos.