¿Cómo se deben preparar los datos para clasificadores en Machine Learning?

Cuando se desea aplicar algoritmos de clasificación (como regresión logística, KNN, árboles de decisión, etc.), es fundamental que los datos estén organizados y transformados de una manera que los modelos puedan entender. A continuación se explican los conceptos clave.

1. Variables de entrada (features) numéricas o codificadas

Los clasificadores tradicionales no aceptan directamente texto o cadenas. Por eso, toda variable de entrada debe estar expresada en forma numérica discreta o continua. Para lograrlo, se pueden transformar las variables categóricas a través de:

a) Categorías (Categorical)

Una categoría es un valor cualitativo que representa un grupo, clase o tipo. Por ejemplo:
- "Junior", "Mid", "Senior"
- "FT", "PT", "Contract"

b) Etiquetas codificadas (Label Encoding)

Consiste en asignar un número a cada categoría. Ejemplo:
- "Junior" → 0
- "Mid" → 1
- "Senior" → 2

Esto convierte texto en números, aunque introduce un orden implícito que no siempre es deseado.

c) Variables Dummies (One-Hot Encoding)

En vez de representar la categoría con un solo número, se crean varias columnas binarias (0 o 1) que indican si una observación pertenece a esa categoría. Ejemplo:

FT	PT	Contract
1	0	0
0	1	0
0	0	1

2. Variables objetivo (target) categóricas o discretas

Los modelos de clasificación predicen clases discretas. Por eso, la variable que se quiere predecir (objetivo) debe tener un número finito de categorías, por ejemplo:

- Binaria: 0 o 1 (como "no" o "sí")
- Multiclase: "bajo", "medio", "alto" (3 clases)

No se puede usar una variable continua (como salario exacto) a menos que se transforme en categorías.

3. Consideraciones adicionales

- Evitar datos faltantes: deben ser imputados o eliminados.
- Balance de clases: si una clase es mucho más frecuente que otra, se deben aplicar técnicas de balanceo (como SMOTE o pesos ajustados).
- Escalado: algunos modelos (como KNN o SVM) requieren que los valores numéricos estén en la misma escala.

Resumen

Requisito	Descripción
Variables de entrada	Deben ser numéricas (reales o categóricas codificadas)
Variables categóricas	Usar Label Encoding o One-Hot Encoding (dummies)
Variable objetivo (target)	Discreta (binaria o multiclase)
Datos faltantes	No deben existir o deben imputarse
Balance de clases	Se debe revisar y corregir si es necesario