Hoja de resumen – Preprocesamiento y EDA en ML

(Fundamentos de Machine Learning – caso Iris)

Jean Carlo Londoño Ocampo

🎯 Objetivo del laboratorio

Construir una **buena representación de los datos**:

(X,y)(X, y)(X,y)

antes de entrenar cualquier modelo.

👉 El modelo aprende sobre **X**, no sobre la realidad.

1️⃣ Carga del dataset

Se construye:

- **X** → variables (features)
- **y** → etiqueta (clase)

En Iris:

- X: medidas de sépalo y pétalo
- y: especie

2️⃣ Exploración inicial (EDA)

Sirve para entender la forma de la distribución de X.

Incluye:

- `info()`

- `describe()`
- valores únicos

👉 No es estética, es para detectar problemas de calidad.

3️⃣ Calidad de datos (checks básicos)

Se revisa principalmente:

- ☑ Completitud → valores nulos
- ☑ Validez → rangos y tipos
- ☑ Consistencia básica

Esto garantiza que:

$$(X,y) \approx \text{fenómeno real observado}$$

4️⃣ Visualización

Se usan gráficas para detectar:

✔ **patrones predominantes**

✔ **separaciones naturales**

✔ **estructuras**

Ejemplos usados:

- scatter 3D
- pairplot
- mapa de correlación

👉 Las visualizaciones ayudan a ver si el problema es separable con la representación actual.

5️⃣ Detección básica de outliers

Se usa z-score:

z=x−μσz = \frac{x-\mu}{\sigma}z=σx−μ

Sirve para encontrar:

- posibles anomalías
- errores de medición

👉 No siempre se eliminan.

---

6 Estandarización

Se aplica:

x′=x−μσx' = \frac{x-\mu}{\sigma}x′=σx−μ

Porque:

- muchos modelos son sensibles a escala
- mejora la geometría del espacio de X
- acelera el aprendizaje

👉 Es una transformación de representación:

X→X′X \rightarrow X'X→X′

---

7 PCA

No es un clasificador.

Sirve para:

- encontrar direcciones de máxima varianza
- visualizar patrones dominantes

Formalmente:

Z=XWZ = XWZ=XW

👉 Permite ver si existen estructuras internas en los datos.

8️⃣ Partición train / test

Se separan los datos en:

- entrenamiento
- prueba

Para evitar:

❌ evaluar el modelo con los mismos datos con los que aprendió.

9️⃣ Exportación

Se guardan los datasets limpios para:

- desacoplar el preprocesamiento
- reutilizar en distintos modelos

🔑 Idea central del laboratorio

Este laboratorio no mejora un modelo.

Mejora la **calidad de la representación**.

📌 Relación directa con lo visto en clase

Todo el flujo es:

datos crudos→ϕ(X)→X′→fθ(X′)\text{datos crudos} \rightarrow \phi(X) \rightarrow X' \rightarrow f_\theta(X')datos crudos→ϕ(X)→X′→fθ(X′)

Donde:

- la mayor ganancia suele venir de **φ (preprocesamiento)**,
- no de cambiar el modelo.