Taller Práctico – Ejercicios Adicionales de Introducción a Machine Learning

Este documento resume y enriquece los 10 ejercicios adicionales del taller de Machine Learning con el dataset Iris.
Incluye explicaciones, notas importantes y conclusiones, además de conceptos complementarios para estudiantes que inician en Aprendizaje Automático.

1. Estadísticas Descriptivas

Consigna: Calcular media, mediana y desviación estándar de sepal length (cm).

Conceptos importantes:
- Media: tendencia central, sensible a valores extremos.
- Mediana: valor central, robusta a outliers.
- Desviación estándar: mide la dispersión respecto a la media.

Conclusión: En Iris, las medidas son consistentes y reflejan un dataset limpio y balanceado.

2. Filtrado y Agrupación

Consigna: Filtrar solo flores virginica y calcular el promedio de petal width (cm).

Conceptos importantes:
- El filtrado permite comparar subconjuntos.
- El agrupamiento (groupby) organiza datos por categorías.

Conclusión: Virginica tiene pétalos más anchos → variable discriminante clave para modelos.

3. Creación de Histogramas

Consigna: Graficar histograma de sepal width (cm) y variar bins.

IMPORTANTE: ¿Qué es un histograma?
- Representa la distribución de una variable numérica.
- El eje X son intervalos (bins), el eje Y la frecuencia.
- Muestra patrones de normalidad, sesgo o concentración.

Conclusión: En Iris, sepal width es casi normal. Cambiar bins altera la percepción de la distribución.

4. Visualización con Scatter Plot

Consigna: Scatter plot de sepal length vs sepal width, coloreado por especie.

Conceptos importantes:
- Cada punto = un registro (X=variable1, Y=variable2).
- Detecta agrupamientos, correlaciones y outliers.

Conclusión: Setosa se distingue fácilmente; Versicolor y Virginica se solapan → mayor dificultad de clasificación.

5. Búsqueda por Condición

Consigna: Contar flores con petal length > 6.0 y crear un DataFrame con ellas.

Conceptos importantes:
- Filtrado lógico (df[df['col'] > valor]).
- Útil para detectar valores extremos.

Conclusión: Los casos con petal length > 6 son mayormente Virginica, confirmando su dominio en valores grandes.

6. Re-división de Datos

Consigna: train_test_split con test_size=0.33, random_state=100.

Conceptos importantes:
- Test size: porcentaje de datos para evaluar.
- Random state: asegura reproducibilidad.

Conclusión: Un 33% en test implica menos datos para entrenar. La reproducibilidad es fundamental en experimentos.

7. Ingeniería de Características Simple

Consigna: Crear columna petal_area = petal length * petal width y calcular media.

Conceptos importantes:
- Feature engineering: creación de nuevas variables a partir de las existentes.
- Aporta más poder predictivo a los modelos.

Conclusión: El área del pétalo es más discriminante que cada dimensión por separado.

8. Análisis de Correlación

Consigna: Calcular matriz de correlación de las 4 características y detectar la más alta.

Conceptos importantes:
- La correlación mide la relación entre variables (-1 a 1).
- Cercano a 1 = relación positiva fuerte.

Conclusión: petal length y petal width tienen correlación $\approx 0.96 \rightarrow$ transmiten información muy similar.

9. Efecto de la Estratificación

Consigna: Hacer train_test_split sin stratify=y y comparar distribuciones.

Conceptos importantes:
- La estratificación mantiene proporciones de clases en train/test.
- Sin ella, se pueden generar conjuntos desbalanceados.

Conclusión: Estratificar es clave para datasets pequeños o con clases balanceadas $\rightarrow$ evita sesgos de evaluación.

10. Generalizando el Proceso (Dataset Wine)

Consigna: Cargar load_wine y explorar con .head(), .info(), .describe().

Conceptos importantes:
- El dataset Wine tiene 13 variables (más complejo que Iris).
- Practicar exploración en datasets distintos es esencial para generalizar habilidades.

Conclusión: Explorar, visualizar y preparar datos antes de modelar es un paso universal en ML.

Síntesis Final

Estos 10 ejercicios fortalecen los pilares de la ciencia de datos:

1. Exploración: estadísticas, histogramas, scatter plots.
2. Preparación: divisiones train/test, estratificación.
3. Transformación: filtrados, agregaciones, feature engineering.
4. Generalización: aplicar la misma metodología en diferentes datasets.

Mensaje clave: Antes de entrenar un modelo, el paso más importante es conocer y preparar los datos.