Guia para el examen

Esta guía está diseñada para ayudarte a **pensar y actuar como un científico de datos**. No se trata simplemente de resolver ejercicios, sino de **aprender a aplicar un enfoque sistemático** que puedas reutilizar una y otra vez con distintos problemas.

Objetivos del control

Ser capaz de:

1. Leer y explorar conjuntos de datos (.csv).
2. Mostrar gráficamente las muestras para problemas con dos características.
3. Aplicar técnicas de preprocesamiento adecuadas.
4. Entrenar y comparar distintos clasificadores:
   - SVM, KNN, Naive Bayes, Árboles de Decisión, Random Forest, Stacking.
5. Evaluar modelos con distintas métricas y estrategias:
   - Hold-out y validación cruzada.
   - Métricas: exactitud, precisión, recall, F1.
6. Ajustar hiperparámetros manualmente o con GridSearchCV / RandomizedSearchCV.
7. Extraer conclusiones sólidas y justificar decisiones.

Pasos recomendados

1. Elige un conjunto de datos

- Puedes usar alguno con los que hemos trabajado (Iris, el de las caras, Breast Cancer...) o descargar alguno nuevo de Kaggle o similares.
  - https://www.kaggle.com/
  - https://archive.ics.uci.edu/
- Asegúrate de que tenga 2 o más características numéricas para facilitar la visualización y que la clasificación sea interesante.

---

2. Explora y visualiza los datos

- Usa `.head()`, `.info()`, `.describe()`.

---

3. Preprocesa

- Escala los datos si es necesario (`MinMaxScaler`).
  - Sobre todo para el KNN y el SVC.
- Verifica si hay valores faltantes y cómo tratarlos (Muy opcional)
  - Todos los datasets que os vamos a poner van a estar listo para hacer clasificación, pero es buena practica hacer esto.
- Divide en entrenamiento y prueba, por ejemplo 80% - 20% pero puedes probar otras divisiones como 70% - 30%.

---

4. Entrena distintos modelos

- Prueba con:
  - **SVM**
  - **KNN**
  - **Naive Bayes**

- **Árbol de Decisión**
- **Random Forest**
- **Stacking**

---

5. Evalúa rendimiento

- Usa:
  - **Hold-out** (train/test split)
  - **Validación cruzada**
- Mide:
  - **Exactitud (accuracy)**
  - **Precisión (precision)**
  - **Recall**
  - **F1-score**

---

6. Ajusta hiperparámetros

- Primero manualmente (con el bucle): cambia algunos parámetros y observa el efecto.
- Luego usa:
  - `GridSearchCV`
  - `RandomizedSearchCV`

Compara resultados y comenta las mejoras.

---

7. Extrae conclusiones

Reflexiona sobre:

- Qué modelos rindieron mejor y por qué.
- Cómo afectó el preprocesamiento.
- Qué impacto tuvo el ajuste de hiperparámetros.
- Qué aprendiste sobre los datos y los clasificadores.

Repite con otros conjuntos de datos

> ⚠️ **Recomendación fundamental:**
> Una vez termines este proceso con un dataset, **hazlo otra vez con otro... y luego con otro**.
> La **clave para entender realmente el aprendizaje automático** no está en memorizar pasos, sino en ver cómo se **comportan los métodos en escenarios distintos**. A medida que los repites, empezarás a notar **cuál herramienta es más útil para cada situación**.

Piensa en esto como construir tu propia caja de herramientas:

- La única forma de saber qué herramienta usar es **haberla probado en distintos contextos**.
- Así, cuando tengas un nuevo problema real, sabrás cómo abordarlo de forma efectiva.