Padrón: _____ Nombre y Apellido:_____ Grupo :_____

Parcial 16-10-2025

Ejercicio 1

a. Explique la diferencia entre un problema de regresión y uno de clasificación. De un ejemplo concreto de un *dataset* para cada caso y qué métrica usarías para evaluar los modelos.

- **Regresión**: predecir una variable **continua**. Ejemplo: predecir el precio de una casa (variable: precio en $). Métrica típica: **MAE** (error absoluto medio) o **RMSE** (raíz del error cuadrático medio).

- **Clasificación**: predecir una variable **discreta / categórica**. Ejemplo: clasificar correos como *spam* o *no spam*. Métrica típica: **accuracy**, **F1-score**, **AUC-ROC**; la elección depende del balance de clases y del costo de errores.

b. Explique con un ejemplo la diferencia entre *precision* y *recall*. ¿Qué rol cumple la matriz de confusión en la evaluación de modelos de clasificación?

- **Precision** = TP / (TP + FP). Si el clasificador marca 100 correos como spam y 80 son realmente spam , entonces precision = 0.8. Mide qué fracción de positivos predichos son correctos.

- **Recall (sensibilidad)** = TP / (TP + FN). Si hay 200 spams reales y detectamos 80 entonces recall = 0.4. Mide qué fracción de positivos reales detectamos.

- **Matriz de confusión :** Es la tabla que muestra TP, FP, TN, FN. Permite calcular precision, recall, accuracy, specificity, F1, y entender qué tipo de errores comete el modelo (p. ej. muchos falsos negativos vs falsos positivos). Es fundamental para evaluar clasificación, sobre todo con clases desbalanceadas.

Ejercicio 2

En el proceso de ingeniería de características, un estudiante propone aplicar **one-hot encoding** a una variable categórica con 200 categorías diferentes.

a. ¿Qué problemas podría traer esto?

Genera **dimensionalidad muy alta** (200 columnas nuevas), lo que provoca: sparsity (muchos ceros), aumento del coste computacional y memoria, mayor probabilidad de overfitting, y modelos menos interpretables. Además, algunos algoritmos pueden verse afectados por la alta dimensionalidad.

b. ¿Qué alternativas hay para tratar este tipo de variables?

Padrón: _____ Nombre y Apellido:_____ Grupo :_____

- **Agrupar categorías**: combinar categorías poco frecuentes en una categoría "otros".
- **Embeddings**: aprender representaciones densas de baja dimensión (por ejemplo, embeddings con redes neuronales o usando técnicas como entity embeddings).
- Usar otro tipo de encoding (label, target, mean, etc).

Ejercicio 3

a. ¿Qué diferencias conceptuales hay entre *bagging* y *boosting*? ¿Cómo se reflejan esas diferencias en Random Forest y XGBoost?

- **Bagging (Bootstrap Aggregating)**: construye múltiples modelos de forma **independiente** y los **promedia** (o vota). Reduce varianza; cada modelo se entrena en una muestra bootstrap distinta.

  En Random Forest: muchos árboles independientes, cada árbol se entrena con bootstrap y con selección aleatoria de features por split; la predicción es promedio (regresión) o voto mayoritario (clasificación).

- **Boosting**: construye modelos **secuencialmente**, cada nuevo modelo corrige los errores de los anteriores (pone más peso en las observaciones mal predichas). Reduce el sesgo y puede reducir la varianza; es más propenso a sobreajuste si no se regulariza. Ejemplo: XGBoost, LightGBM, AdaBoost.

  En XGBoost: boosting basado en gradiente; cada árbol se ajusta a los residuos (gradiente negativo) del ensemble anterior y usa regularización (L1/L2), shrinkage (learning rate), pruning, etc.

b. ¿Cual es la diferencia principal entre ensambles híbridos y homogéneos?

- **Homogéneo**: todos los modelos del ensamble son del **mismo tipo** (p. ej. 100 árboles de decisión en Random Forest).

- **Híbrido (heterogéneo)**: combina **modelos distintos** (p. ej. un Random Forest + un SVM + una red neuronal) y luego los combina (stacking, cascading, etc). Desventaja: mayor complejidad y necesidad de combinar salidas.

Ejercicio 4

a. ¿Qué ventajas tiene usar una técnica de reducción de la dimensionalidad?

Reduce el **ruido** y la **colinealidad**, mejora eficiencia computacional (menos tiempo y memoria), reduce riesgo de **overfitting**, facilita visualización (p. ej. 2D/3D) y puede acelerar/converger mejor modelos de aprendizaje.

b. Mencione tres técnicas que conozca de reducción de la dimensionalidad.

PCA (Análisis de Componentes Principales) , t-SNE, ISOMAP

Ejercicio 5

a. ¿Cuál es la limitación principal del perceptrón simple?

Solo puede separar clases **linealmente separables**. Si las clases no son separables por una hiperplano (p. ej. XOR), el perceptrón no converge.

b. ¿Qué mecanismo matemático utiliza el método de Backpropagation?

Utiliza la **regla de la cadena** para calcular gradientes de la función de pérdida respecto a los pesos en una red multicapa y luego aplica descenso del gradiente (o variantes) para actualizar los pesos.

c. ¿Qué es un optimizador? Nombre tres que conozca.

Un **optimizador** es el algoritmo que actualiza los parámetros del modelo a partir de los gradientes de la función de pérdida (determina la regla de actualización y su dinámica). **Es una implementación concreta del algoritmo de backpropagation** Ejemplos: SGD, **RMSprop,Adam**, **Adamax**, Nadam, etc

d. ¿Para qué utilizaría una red SOM?

SOMs son redes no supervisadas para **visualización y clustering** de datos de alta dimensión en un mapa 2D. Útiles para explorar estructuras, agrupar datos, y crear mapas que muestran similitudes entre observaciones (por ejemplo, segmentación de clientes, análisis exploratorio).

Ejercicio 6

En el ejercicio de **clustering** con el *dataset* de Spotify (Trabajo Práctico 1), ¿cuántos grupos eligieron finalmente al aplicar **K-Means** y qué representan esos grupos?

DEPENDE DE CADA GRUPO