Repaso TA4

- ✅ Implementamos regresión lineal para predecir precios de casas
- ✅ Implementamos regresión logística para diagnóstico médico
- ✅ Diferenciamos ambos tipos de modelos
- ✅ Evaluamos modelos con múltiples métricas

Repaso TA4 – Regresión Lineal

🏠 Dataset: Boston Housing

Objetivo: Predecir precio de casas (valor continuo)

Métricas: MAE, MSE, RMSE, R^2, MAPE

Ejemplo: Precio real $25k vs Predicho $23k

Error: $2k (MAE)

Repaso TA4 - Regresión Logística

🩺 Dataset: Breast Cancer Diagnosis

Objetivo: Clasificar tumor (benigno/maligno)

Métricas: Accuracy, Precision, Recall, F1-Score

Ejemplo: Real=Benigno vs Predicho=Benigno

Accuracy: 95%

Repaso TA4 – Diferencias Clave

Aspecto	Regresión Lineal	Regresión Logística
Qué predice	Números continuos	Categorías
Ejemplo de uso	Precios, temperaturas	Diagnósticos, spam
Rango de salida	$-\infty$ a $+\infty$	0 a 1 (probabilidad)
Métrica principal	R^2, MAE	Accuracy, F1-Score

Problemas que Encontramos

- ¿Cómo sabemos si nuestro modelo es REALMENTE bueno?
- ¿Y si solo funciona bien con ESTOS datos específicos?
- ¿Cómo comparamos MÚLTIPLES modelos?
- ¿Hay errores ocultos en nuestro proceso?

Train/Test Split – ¿Suficiente?

- Problema: Solo una división de datos
- ¿Y si esa división es "fácil" o "difícil"?
- ¿Cómo saber si es suerte o habilidad?

- Necesitamos algo más robusto...

Validación y Selección de Modelos

Objetivos de Hoy

- Entender qué es la contaminacion de datos (DATA LEAKAGE) y cómo evitarlo
- Implementar validacion cruzada (CROSS-VALIDATION) robusto
- Comparar MÚLTIPLES modelos sistemáticamente
- Interpretar métricas de ESTABILIDAD
- Evitar errores que arruinan modelos en producción

Contaminacion de datos (Data Leakage)

Cuando el modelo "ve" información que NO debería tener durante el entrenamiento

Como hacer trampa en un examen:

- ❌ Ver las respuestas antes del examen
- ✅ Solo estudiar el material permitido

Contaminacion de datos (Data Leakage)

Ejemplo: Data Leakage en Acción

❌ INCORRECTO:

1. Preprocesar TODO el dataset
2. Dividir en train/test
3. Entrenar modelo

✅ CORRECTO:

1. Dividir en train/test
2. Preprocesar SOLO con datos de entrenamiento
3. Aplicar preproces ado a test

Contaminacion de datos (Data Leakage)

¿Por qué es tan Peligroso?

- Optimismo artificial: Métricas infladas
- Información del futuro: El modelo "hace trampa"
- Falla en producción: Rendimiento real muy bajo
- Decisiones erróneas: Seleccionas modelo malo

Contaminacion de datos (Data Leakage)

Solución: Pipelines

Pipeline = Secuencia automática de pasos

- Previene data leakage automáticamente
- Aplica transformaciones en orden correcto
- Garantiza proceso robusto

Validacion cruzada (Cross-Validation)

Problema del Train/Test Split

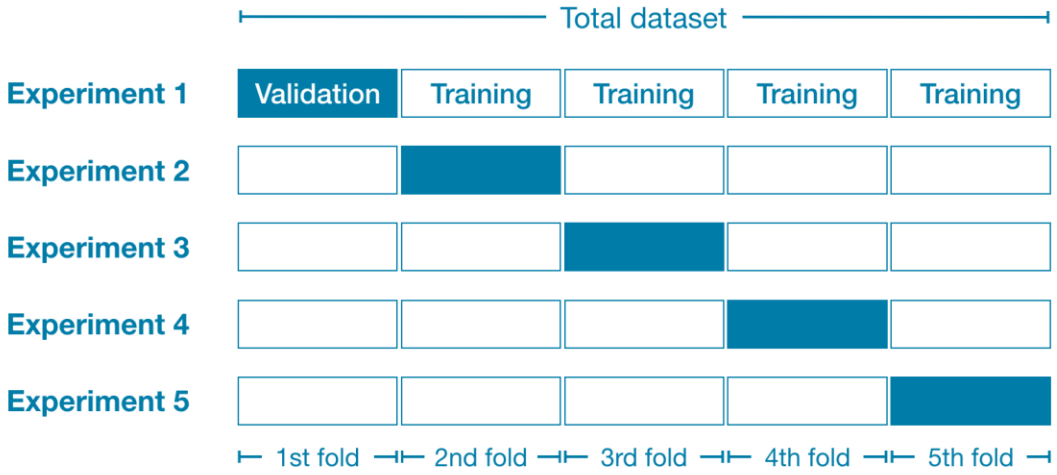
- Una sola división = Una sola "opinión"
- Resultados pueden variar por suerte
- ¿El modelo es bueno o tuvo suerte?

Necesitamos múltiples "opiniones"

Validacion cruzada (Cross-Validation)

¿Qué es Cross-Validation?

- Dividir datos en K partes (folds)
- Entrenar K veces, cada vez con diferente test
- Promedio de K resultados = estimación robusta
- Desviación estándar = estabilidad del modelo



Comparación de Modelos

¿Por qué Comparar Modelos?

- Diferentes algoritmos para diferentes problemas
- No hay "modelo perfecto universal"
- Competencia revela el mejor para TUS datos
- Combinar rendimiento + estabilidad

Comparación de Modelos

Candidatos Típicos

- Logistic Regression: Simple, interpretable
- Ridge Classifier: Con regularización
- Random Forest: Ensemble, robusto
- SVM: Fronteras complejas

Comparación de Modelos

Proceso de Comparación

1. Crear pipelines para cada modelo
2. Evaluar con cross-validation
3. Comparar accuracy promedio
4. Analizar estabilidad (desviación)
5. Seleccionar ganador

Comparación de Modelos

Métricas de Selección

- Rendimiento: ¿Cuál es más preciso?
- Estabilidad: ¿Cuál es más consistente?
- Velocidad: ¿Cuál entrena/predice más rápido?
- Memoria: ¿Cuál usa menos recursos?
- Interpretabilidad: ¿Cuál es más explicable?