Curso base

INTELIGENCIA ARTIFICIAL DESDE CERO

LECCIÓN 19
Matriz de confusión y métricas

Cómo interpretar accuracy, precisión, recall y F1 para entender los aciertos y los errores del clasificador

Lección	19	**Tema**	Matriz de confusión y métricas
Nombre del participante	_____	**Fecha**	_____

Actividad central

Actividad principal: análisis argumentado de errores de clasificación a partir de una matriz de confusión, métricas principales y dos ejemplos guiados.
Producto esperado: análisis corto, claro y argumentado que incluya al menos una matriz de confusión, las métricas principales y una explicación breve sobre cuál es el riesgo más importante del clasificador evaluado.

Propósito de la guía. Comprender cómo se evalúa un clasificador mediante la matriz de confusión y cuatro métricas muy importantes: accuracy, precisión, recall y F1. Al finalizar, el participante podrá leer los aciertos y los errores del modelo con más criterio, sin dejarse engañar por un porcentaje bonito que no siempre significa utilidad real.

Idea central

Cuando un modelo de clasificación produce predicciones, no basta con contar cuántas veces acertó. Dos clasificadores pueden tener una accuracy parecida y, sin embargo, comportarse de forma muy distinta en los casos que realmente importan. Por eso necesitamos mirar con más detalle qué tipo de errores aparecen y con qué frecuencia.

La matriz de confusión nos ayuda a ver el panorama completo. A partir de ella podemos calcular accuracy, precisión, recall y F1. En esta lección no buscamos memorizar fórmulas complicadas, sino desarrollar intuición y criterio para responder preguntas como estas: **¿el modelo detecta los casos importantes?, ¿lanza demasiadas alertas falsas?, ¿parece equilibrado?, ¿sirve realmente para apoyar decisiones?**

¿Qué vamos a hacer en esta guía?

- Reconocer los componentes básicos de una matriz de confusión binaria.
- Entender qué pregunta responde cada métrica: accuracy, precisión, recall y F1.
- Desarrollar un primer ejemplo guiado paso a paso con *aprobado / no aprobado*.
- Comparar un segundo ejemplo donde la accuracy sola puede engañar.
- Usar prompts de IA generativa para comprender, interpretar y redactar conclusiones.

Vocabulario mínimo para comenzar

Concepto	Interpretación sencilla
Verdadero positivo (VP)	El caso era positivo y el modelo también lo predijo como positivo.
Falso positivo (FP)	El modelo dijo positivo, pero en realidad el caso no lo era.
Falso negativo (FN)	El caso sí era positivo, pero el modelo no lo detectó.
Verdadero negativo (VN)	El caso era negativo y el modelo lo clasificó correctamente como negativo.

Estos cuatro resultados son la base de toda la lectura posterior. La matriz de confusión no es más que una forma ordenada de contar cuántos **VP**, **FP**, **FN** y **VN** produjo el clasificador.

¿Qué pregunta responde cada métrica?

Métrica	Pregunta que responde
Accuracy	¿Qué proporción total de predicciones fue correcta?
Precisión	Cuando el modelo dijo positivo, ¿con qué frecuencia acertó?
Recall	De todos los positivos reales, ¿cuántos logró detectar?
F1	¿Qué tan equilibrado es el desempeño entre precisión y recall?

Ejemplo guiado 1 - Aprobado / no aprobado

Empezaremos con un ejemplo cercano al mundo educativo. Supongamos que un clasificador intenta decidir si un estudiante aprobará o no aprobará. Tenemos etiquetas reales y predicciones del modelo. A partir de esa comparación construiremos la matriz de confusión y leeremos las métricas.

Caso	Etiqueta real	Predicción del modelo
1	Aprobado	Aprobado
2	Aprobado	Aprobado
3	Aprobado	No aprobado
4	No aprobado	No aprobado
5	No aprobado	Aprobado

6	Aprobado	Aprobado
7	No aprobado	No aprobado
8	Aprobado	Aprobado
9	No aprobado	No aprobado
10	No aprobado	No aprobado

Paso 1. Decidir cuál es la clase positiva

En este ejemplo tomaremos **Aprobado** como clase positiva. Esto es importante porque precisión y recall se calculan respecto a la clase que queremos vigilar con atención.

Paso 2. Identificar VP FP FN y VN

Tipo de resultado	Cantidad	Lectura breve
VP	4	Aprobado y el modelo también dijo aprobado.
FP	1	El modelo dijo aprobado, pero en realidad no aprobó.
FN	1	El estudiante sí aprobó, pero el modelo no lo detectó.
VN	4	No aprobado y el modelo también dijo no aprobado.

Paso 3. Leer la matriz de confusión

	Predijo no aprobado	Predijo aprobado
Real: no aprobado	4	1
Real: aprobado	1	4

Paso 4. Calcular e interpretar las métricas

Métrica	Cálculo	Lectura breve
Accuracy	$(4+4)/10 = 0.80$	Acertó en el 80 % de los casos.
Precisión	$4/(4+1) = 0.80$	Cuando predijo aprobado, acertó el 80 %.
Recall	$4/(4+1) = 0.80$	Detectó el 80 % de quienes realmente aprobaron.
F1	$2 \cdot (0.80 \cdot 0.80)/(0.80 + 0.80) = 0.80$	En este caso hay equilibrio entre precisión y recall.

Paso 5. Ejecutar el código básico

```
from sklearn.metrics import confusion_matrix, accuracy_score
from sklearn.metrics import precision_score, recall_score, f1_score

y_real = [1,1,1,0,0,1,0,1,0,0]
y_pred = [1,1,0,0,1,1,0,1,0,0]

matriz = confusion_matrix(y_real, y_pred)
print(matriz)

print("Accuracy:", round(accuracy_score(y_real, y_pred), 2))
print("Precision:", round(precision_score(y_real, y_pred), 2))
print("Recall:", round(recall_score(y_real, y_pred), 2))
print("F1:", round(f1_score(y_real, y_pred), 2))
```

Paso 6. Interpretar con palabras sencillas

En este primer ejemplo el clasificador parece razonablemente equilibrado. No solo tiene una accuracy aceptable, sino que también muestra precisión y recall similares. Eso significa que no está exagerando demasiado cuando dice *aprobado*, pero tampoco está dejando pasar demasiados casos positivos.

Prompts de IA generativa para acompañar el ejemplo 1

- "Explícame con palabras muy sencillas qué significa cada celda de una matriz de confusión en un problema de aprobado / no aprobado."
- "Ayúdame a redactar una interpretación breve de accuracy, precisión, recall y F1 para un estudiante principiante."
- "Revisa este código de matriz de confusión y métricas, y dime paso a paso si está correcto o si tiene errores."
- "Explícame la diferencia entre precisión y recall usando un lenguaje claro y un ejemplo educativo."

¿Cuándo puede engañar la accuracy?

La accuracy es útil, pero no debe leerse sola. Puede ocurrir que un modelo tenga una accuracy alta y, sin embargo, falle justo en los casos más importantes. Esto sucede con frecuencia cuando una clase aparece muy poco y la otra aparece mucho.

Imagine un conjunto de 100 casos donde solo 10 son positivos y 90 son negativos. Si un clasificador dijera siempre *negativo*, acertaría 90 veces y tendría 90 % de accuracy. Suena bien, pero en realidad no detectó ni un solo caso positivo. En un problema importante, ese comportamiento sería inaceptable.

Segundo ejemplo guiado - Accuracy alta, pero modelo poco útil

Ahora veremos un caso de cancelación de clientes. La empresa quiere detectar quiénes probablemente cancelarán el servicio. Supongamos que, en 100 clientes, solo 10 realmente cancelan y 90 no cancelan.

Modelo	VP	FP	FN	VN	Accuracy	Recall	Lectura rápida
Modelo A	0	0	10	90	0.90	0.00	Parece bueno por accuracy, pero no detecta ninguna cancelación.
Modelo B	7	11	3	79	0.86	0.70	Tiene menor accuracy, pero detecta muchos más casos importantes.

Paso 1. Comparar qué clase importa más

En este escenario la clase positiva es **cancelará**. Si la empresa quiere anticiparse y tomar medidas, lo más importante es no dejar pasar demasiados clientes que sí van a cancelar. Por eso el **recall** empieza a ganar protagonismo.

Paso 2. Entender por qué accuracy sola no basta

El Modelo A tiene 90 % de accuracy, pero recall igual a 0. Eso significa que no detectó a ninguno de los clientes que sí cancelarían. El Modelo B tiene una accuracy algo menor, pero recall de 0.70, lo cual lo vuelve mucho más útil si el objetivo principal es detectar cancelaciones.

Paso 3. Pensar en el costo del error

Toda métrica debe leerse según el contexto. Si para la empresa es muy grave perder clientes sin advertencia, entonces un **falso negativo** pesa mucho. En ese caso, recall y los falsos negativos importan más que una accuracy llamativa.

Prompts de IA generativa para acompañar el ejemplo 2

- "Compárame dos clasificadores: uno con mayor accuracy pero recall muy bajo, y otro con menor accuracy pero recall más alto. Explícalo para principiantes."
- "Ayúdame a escribir un comentario donde se vea por qué accuracy sola puede engañar en un problema desbalanceado."
- "Explícame qué tipo de error sería más grave en un problema de cancelación de clientes y por qué."
- "Redáctame una conclusión prudente sobre cuál modelo elegiría si me interesa detectar más casos positivos."

Buenas prácticas para leer métricas de clasificación

Recomendación	¿Por qué importa?
No leer accuracy sola	Puede ocultar problemas graves cuando las clases están desbalanceadas.
Mirar la matriz de confusión	Permite ver de qué tipo son los errores y no solo cuántos aciertos hubo.
Pensar en el contexto	La métrica importante cambia según el costo de cada tipo de error.
Comparar precisión y recall	Ayuda a decidir si el clasificador detecta bien o si exagera demasiado.
Usar IA generativa con criterio	Sirve para comprender, revisar y redactar mejor, pero no reemplaza el análisis del estudiante.

Mini ruta de trabajo sugerida

- Observe primero cuál es la clase positiva y qué error sería más grave.
- Identifique correctamente VP, FP, FN y VN antes de mirar cualquier métrica.
- Calcule o verifique accuracy, precisión, recall y F1.
- Compare las métricas entre sí y no se quede con un solo número.
- Escriba una interpretación breve donde explique qué hace bien y qué hace mal el clasificador.

Producto esperado

El producto esperado es un análisis corto, claro y argumentado de errores de clasificación. Debe incluir al menos una matriz de confusión, las métricas principales y una explicación breve sobre cuál es el riesgo más importante del clasificador evaluado.

Lista de verificación

Revisión	Sí / No
☐	¿Se identificaron correctamente VP, FP, FN y VN?
☐	¿Se calcularon o verificaron accuracy, precisión, recall y F1?
☐	¿Se explicó con palabras qué significa cada métrica en el ejemplo estudiado?
☐	¿Se comparó al menos un caso donde accuracy sola no alcanza?
☐	¿Se escribió una conclusión clara sobre el error que más preocupa?

Cierre

La matriz de confusión y las métricas de clasificación nos enseñan a mirar más allá del porcentaje total de aciertos. Ese paso es fundamental para usar la inteligencia artificial con criterio. Un clasificador útil no es el que solo luce bien en un número global, sino el que responde de manera razonable en los casos que realmente importan. Cuando el estudiante comprende esta idea, empieza a evaluar modelos con más madurez y menos ingenuidad.