Métricas de Clasificación: Costos, Errores y Decisiones

HE2: Consultoría Económica con IA Responsable

Santiago Neira & Catalina Bernal

Universidad de los Andes
Departamento de Economía

Febrero 2026

Agenda de hoy

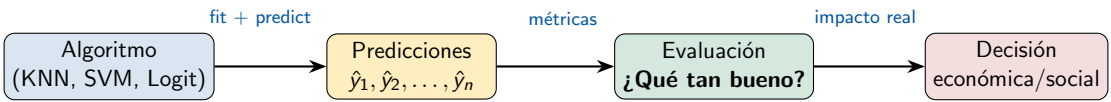
1. De la Matriz de Confusión a los Costos Reales
2. Métricas de Clasificación: Más Allá del Accuracy
3. ¿A Quién Afectan los Errores? IA Responsable
4. Feature Engineering vs Hyperparameter Tuning
5. Framework: Clasificación en Consultoría

Agenda de hoy

Recordatorio: ¿Dónde Estamos?

Clase pasada: Algoritmos de clasificación (KNN, SVM, Logit)

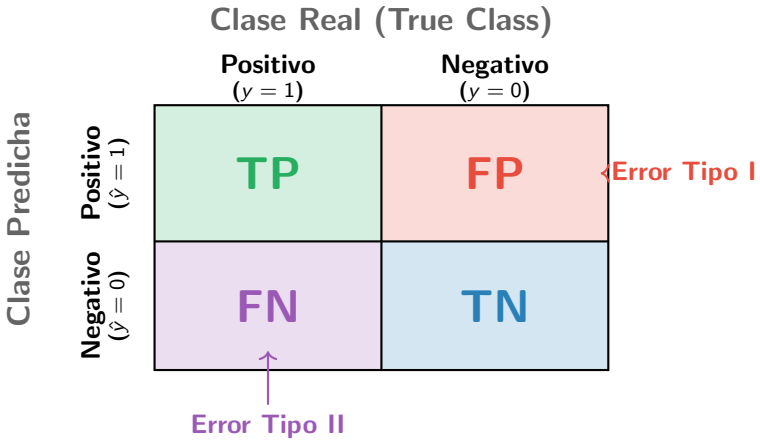
Hoy: ¿Cómo evaluamos si un clasificador es *bueno*?



Preguntas centrales de hoy

- ¿Cómo mejorar la definición de qué tan bueno es un modelo?
- Equivocarse no es gratis. **¿Cuánto cuesta cada tipo de error?**

La Matriz de Confusión



¿Por qué no basta con el Accuracy?

El **accuracy** mide la proporción de predicciones correctas:

$$\text{Accuracy} = \frac{TP + TN}{TP + TN + FP + FN}$$

El problema: Accuracy es una métrica ingenua

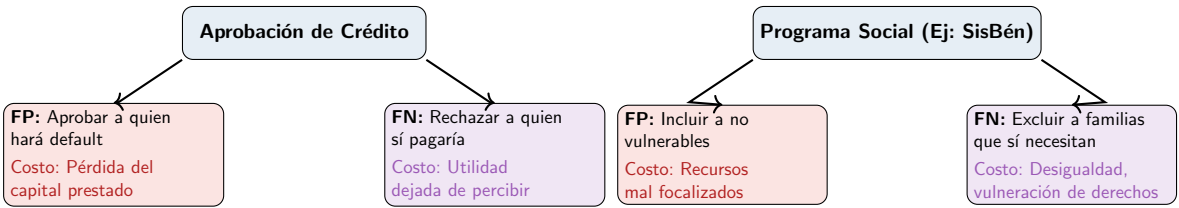
Asume que todos los errores cuestan lo mismo

Pero en la práctica:

- FP y FN tienen **costos diferentes** según el contexto
- Depende de **a quién** afecta el error
- Depende de las **consecuencias** de cada tipo de error

La métrica que elijamos debe reflejar nuestras prioridades

Cada Error Tiene un Costo: Ejemplos Reales



Mismo algoritmo, misma accuracy, pero los costos de equivocarse son radicalmente diferentes.
En crédito, FP $\gg$ FN (pérdida directa). En programas sociales, FN $\gg$ FP (costo humano).

Formalización: Función de Costo Asimétrica

Idea económica: no todos los errores cuestan lo mismo.

Costo esperado de un clasificador:

$$\text{Costo Total} = C_{FP} \cdot FP + C_{FN} \cdot FN \tag{1}$$

Ejemplo — aprobación de crédito

- **Falso positivo (FP):** el banco aprueba un crédito a alguien que hará default $\Rightarrow$ pierde el capital prestado
- **Falso negativo (FN):** el banco rechaza a alguien que sí pagaría $\Rightarrow$ pierde utilidad potencial

Supuestos:

- Crédito promedio: \$10M COP $\rightarrow$ $C_{FP} = \$10\text{M}$
- Utilidad por crédito: \$1.5M COP $\rightarrow$ $C_{FN} = \$1{,}5\text{M}$
- Ratio: $C_{FP}/C_{FN} \approx 6{,}7$

Un FP cuesta $\sim$7 veces más que un FN. El banco tolera rechazar varios buenos clientes para evitar aprobar uno malo.

Agenda de hoy

Accuracy: Útil Pero Peligrosa

$$\text{Accuracy} = \frac{TP + TN}{TP + TN + FP + FN} \qquad (2)$$

¿Cuándo funciona bien?

- Clases balanceadas ($\approx$50/50)
- Ambos errores cuestan igual

¿Cuándo engaña?

- Clases desbalanceadas
- Costos asimétricos

Accuracy: Útil Pero Peligrosa

$$\text{Accuracy} = \frac{TP + TN}{TP + TN + FP + FN} \qquad (2)$$

¿Cuándo funciona bien?

- Clases balanceadas ($\approx$50/50)
- Ambos errores cuestan igual

¿Cuándo engaña?

- Clases desbalanceadas
- Costos asimétricos

Detección de Fraude

(1 % fraude, 99 % legítimo)

	Fraude	Legítimo
	0	**0**
	10	**990**

Accuracy = $\frac{990}{1000}$ = **99 %**

Accuracy: Útil Pero Peligrosa

$$\text{Accuracy} = \frac{TP + TN}{TP + TN + FP + FN} \qquad (2)$$

¿Cuándo funciona bien?

- Clases balanceadas ($\approx$50/50)
- Ambos errores cuestan igual

¿Cuándo engaña?

- Clases desbalanceadas
- Costos asimétricos

Detección de Fraude

(1 % fraude, 99 % legítimo)

Fraude	Legítimo
0	**0**
10	**990**

Accuracy $= \frac{990}{1000} =$ **99 %**

Pero NO detecta ningún fraude!

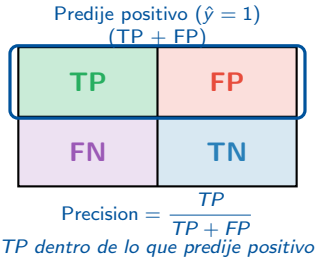
Precision: "De Lo Que Predije Positivo, ¿Cuánto Acerté?"

$$\text{Precision} = \frac{TP}{TP + FP} \tag{3}$$

Intuición: Mide la *calidad* de las predicciones positivas. ¿Qué tan confiable es cuando el modelo dice "sí"?

Importa cuando FP es costoso:

- **Crédito:** Aprobar a un moroso → pérdida
- **Spam:** Marcar email legítimo como spam → usuario pierde info
- **Consultoría:** Recomendar inversión riesgosa al cliente



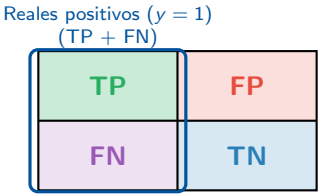
Recall (Sensibilidad): "De Los Reales Positivos, ¿Cuántos Encontré?"

$$\text{Recall} = \frac{TP}{TP + FN} \tag{4}$$

Intuición: Mide la *cobertura*. ¿Qué proporción de los positivos reales capturé?

Importa cuando FN es costoso:

- **Salud pública:** No detectar un caso de tuberculosis → contagio
- **Programas sociales:** Excluir familia vulnerable → desprotección
- **Fraude fiscal:** No detectar evasión → pérdida de recaudo



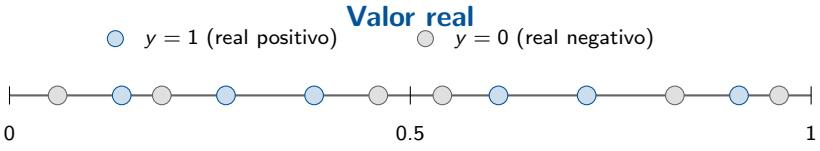
Clasificación = Probabilidades + Umbral

Idea clave:
Los modelos producen probabilidades $\hat{p}$.

Regla de decisión:

$$\hat{y} = \begin{cases} 1 & \text{si } \hat{p} \geq \tau \\ 0 & \text{si } \hat{p} < \tau \end{cases}$$

Mover τ cambia FP y FN.

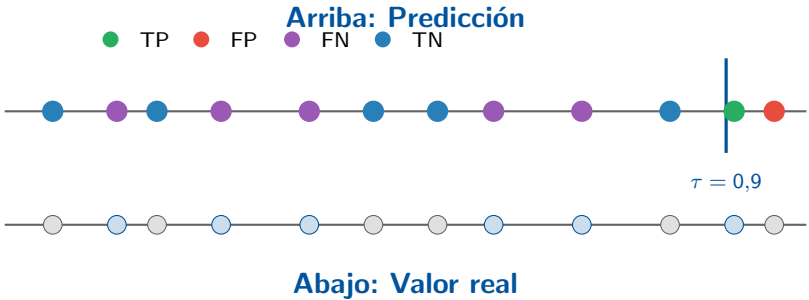


Trade-off: Umbral Alto $t = 0{,}9$ (Modelo Conservador)

Modelo conservador

- FP $\downarrow$
- FN $\uparrow$

Precision $\uparrow$ Recall $\downarrow$

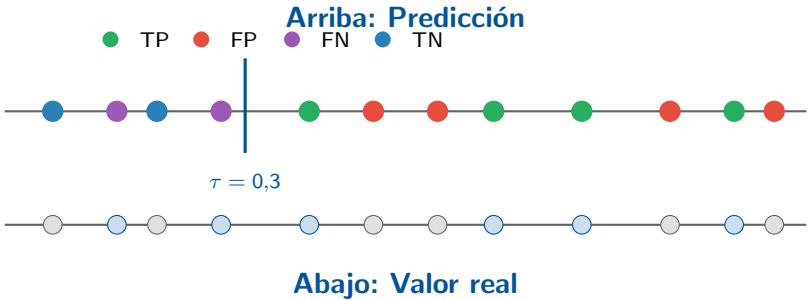


Trade-off: Umbral Bajo $t = 0{,}3$ (Modelo Agresivo)

Modelo agresivo

- FN $\downarrow$
- FP $\uparrow$

Recall $\uparrow$ Precision $\downarrow$

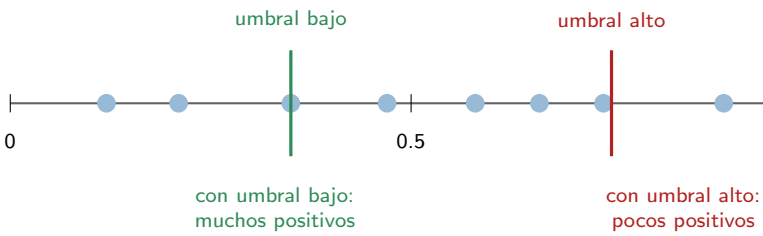


¿Qué Hace Realmente Mover el Umbral? Tradeoff entre Precision y Recall

Mover τ cambia:

- qué casos se clasifican como positivos
- cuántos errores FP/FN aparecen
- precision vs recall

Esto determina la importancia de saber escoger la métrica o el modelo



Specificity: El Recall de los Negativos

$$\text{Specificity} = \frac{TN}{TN + FP} \tag{5}$$

Intuición: De todos los negativos reales, ¿cuántos identifiqué correctamente como negativos?

Importa cuando queremos proteger a los negativos:

- **Sistema judicial:** No condenar inocentes → Specificity alta
- **Test médico:** No alarmar innecesariamente → Specificity alta
- **Selección de personal:** No descartar buenos candidatos → Specificity alta

Relación clave

Specificity $= 1 -$ Tasa de Falsos Positivos (FPR).
Recall y Specificity son complementarios: uno mide qué tan bien detectamos los positivos, el otro los negativos.

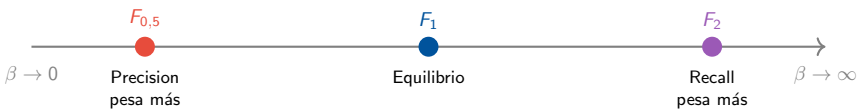
F1 Score y F_β: Combinando Precision y Recall

F1 Score: Media armónica de Precision y Recall

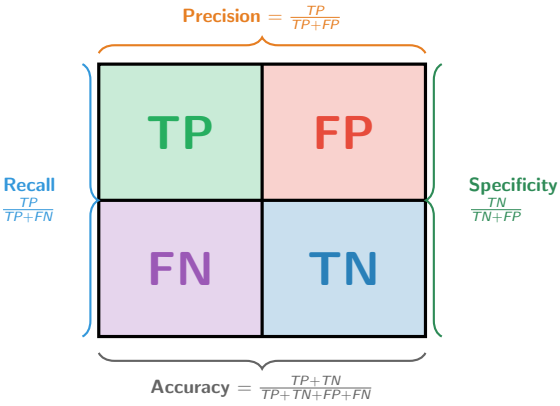
$$F_1 = 2 \cdot \frac{\text{Precision} \cdot \text{Recall}}{\text{Precision} + \text{Recall}} \tag{6}$$

F_β **Score:** Versión generalizada con peso β

$$F_\beta = (1 + \beta^2) \cdot \frac{\text{Precision} \cdot \text{Recall}}{\beta^2 \cdot \text{Precision} + \text{Recall}} \tag{7}$$



Resumen Visual de Métricas



Regla rápida:

- **Precision** → fila superior
- **Recall** → columna izquierda
- **Specificity** → columna derecha
- **Accuracy** → diagonal

F1 combina precision y recall

Agenda de hoy

Los Errores No Son Simétricos: Quién Paga

Distribución del costo del error

La empresa asume el costo

- Banco aprueba mal crédito (FP)
- Empresa contrata mal candidato
- Campaña de marketing a público equivocado

Consecuencia: Pérdida financiera para la firma

El individuo asume el costo

- Familia excluida de programa (FN)
- Paciente sin diagnóstico
- Persona rechazada de crédito injustamente

Consecuencia: Daño social, vulneración de derechos

IA Responsable: La métrica que optimizamos refleja *a quién* priorizamos. Optimizar solo accuracy puede maximizar utilidad de la empresa a costa del individuo.

Caso 1: Consultoría para Banco — Scoring Crediticio

Contexto: Una firma de consultoría diseña un modelo de scoring para aprobar lineas de tarjetas de crédito "premium". El banco tiene 100,000 solicitudes mensuales. El CFO está preocupado por la coyuntura actual y tiene énfasis en no aprobar "malos" clientes

Modelo A: Optimiza Accuracy

800	500
200	98,500

Accuracy = 99.3 %
Precision = 61.5 %
500 malos créditos

Modelo B: Optimiza $F_{0,5}$ (Precision)

200	50
800	98,950

Accuracy = 99.1 %
Precision = 80.0 %
Solo 50 créditos malos aprobados

Caso 1: Consultoría para Banco — Scoring Crediticio

Contexto: Una firma de consultoría diseña un modelo de scoring para aprobar lineas de tarjetas de crédito "premium". El banco tiene 100,000 solicitudes mensuales. El CFO está preocupado por la coyuntura actual y tiene énfasis en no aprobar "malos" clientes

Modelo A: Optimiza Accuracy

800	500
200	98,500

Accuracy = 99.3 %
Precision = 61.5 %
500 malos créditos

Modelo B: Optimiza $F_{0,5}$ (Precision)

200	50
800	98,950

Accuracy = 99.1 %
Precision = 80.0 %
Solo 50 créditos malos aprobados

Impacto económico

Con $100M COP por crédito: Modelo A pierde $50,000M en defaults. Modelo B pierde $5,000M. La diferencia es $45,000M mensuales solo por elegir la métrica correcta.

Caso 2: Focalización de Programas Sociales

Contexto: El gobierno quiere identificar familias en pobreza extrema para un subsidio de vivienda. Hay 50,000 familias candidatas.

Modelo con alta Precision ($F_{0,5}$)

3,000	200
2,000	44,800

Precision = 93.8 %
Recall = 60 %
2,000 familias vulnerables excluidas

Modelo con alto Recall (F_2)

4,700	1,500
300	43,500

Precision = 75.8 %
Recall = 94 %
Solo 300 familias vulnerables excluidas

Perspectiva de IA Responsable

Aquí FN $\gg$ FP en costo social. 1,500 subsidios "desperdiciados" son mucho menos graves que 2,000 familias sin acceso a vivienda.

Caso 2: Focalización de Programas Sociales

Contexto: El gobierno quiere identificar familias en pobreza extrema para un subsidio de vivienda. Hay 50,000 familias candidatas.

Modelo con alta Precision ($F_{0,5}$)

3,000	200
2,000	44,800

Precision = 93.8 %
Recall = 60 %
2,000 familias vulnerables excluidas

Modelo con alto Recall (F_2)

4,700	1,500
300	43,500

Precision = 75.8 %
Recall = 94 %
Solo 300 familias vulnerables excluidas

Perspectiva de IA Responsable

Aquí FN $\gg$ FP en costo social. 1,500 subsidios "desperdiciados" son mucho menos graves que 2,000 familias sin acceso a vivienda.¿Si ustedes fueran hacedores de política, estarían de acuerdo con esta propuesta?

Caso 3: Consultoría Estratégica — Detección de Churn

Contexto económico: Una telco quiere predecir qué clientes cancelarán su plan (churn) para ofrecer retención personalizada. Base analizada: 200,000 clientes.

Estructura de costos por cliente:

- Oferta de retención: \$50K COP
- Valor de vida del cliente (CLV): \$2M COP

Interpretación económica de errores:

- **Falso Positivo (FP):** Ofrecemos retención a alguien que no se iba a ir → costo innecesario = \$50K
- **Falso Negativo (FN):** No detectamos churn real → pérdida total = \$2M

$$\frac{C_{FN}}{C_{FP}} = \frac{2M}{50K} = 40$$

Cada FN cuesta **40×** más que un FP.

Escenario ilustrativo

Supongamos que el modelo evita:

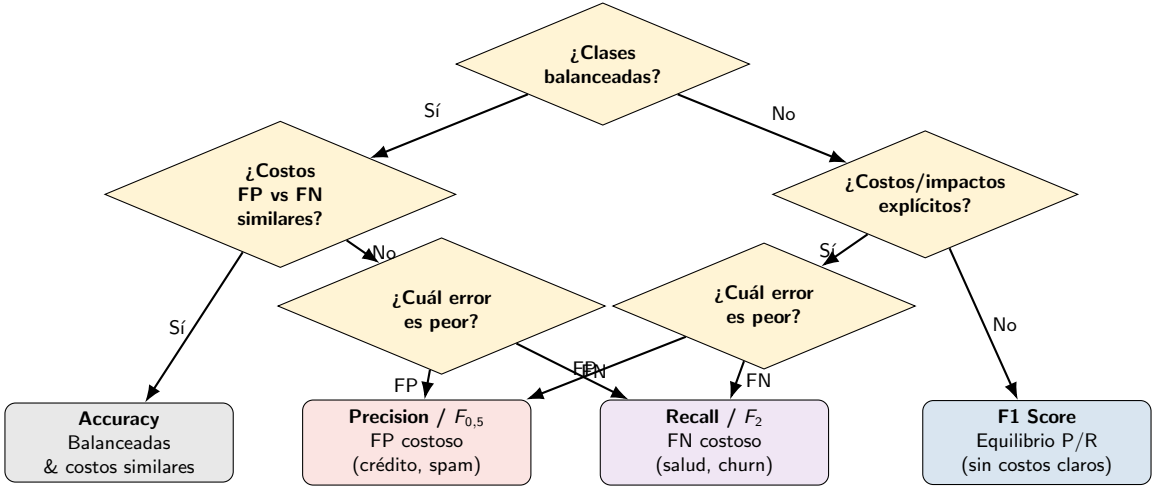
- 100 FN → ahorro = \$200M
- pero genera 1,000 FP → costo = \$50M

Impacto económico neto:

$$200M - 50M = \mathbf{+150M}$$

Conclusión estratégica:
Es racional tolerar muchos FP si eso reduce FN.
Optimizar → **Recall** (o F_2)

Guía: ¿Qué Métrica Elegir?



Regla de oro: La métrica la define el negocio/contexto.
Pregunta clave: *"¿Qué pasa si el modelo se equivoca en cada dirección?"*

Agenda de hoy

¿Qué Mejora Más un Modelo?

Feature Engineering

Δ F1 $\approx$ 5–15 %

Hyperparameter Tuning

Δ F1 $\approx$ 1–3 %

Feature Engineering:

- Crear variables informativas
- Transformar variables existentes
- Interacciones, ratios, lags
- Conocimiento del dominio

"Garbage in, garbage out"
No importa qué tan bien tuneemos si las features no capturan la señal.

Hyperparameter Tuning:

- GridSearchCV, RandomizedCV
- Ajustar K, C, γ, λ
- Optimización fina

Importante pero *secundario*.

Feature Engineering: Ejemplos en Economía

Contexto: Predicción de default crediticio

Features originales

- Ingreso mensual
- Monto deuda total
- Edad
- Meses en empleo
- N° de créditos

Feature Eng.

Features construidas

- Ratio deuda/ingreso
- Ingreso per cápita
- log(ingreso)
- Edad^2 (efecto cuadrático)
- Estabilidad laboral (bins)
- Créditos $\times$ deuda (interacción)

Sin FE:
Accuracy: 78 %
F1: 0.71

Con FE:
Accuracy: 87 %
F1: 0.84

+13 pp en F1

Con FE + Tuning:
Accuracy: 88 %
F1: 0.86

+2 pp adicionales

Feature Engineering: El Conocimiento del Dominio es Rey

Ejemplos por industria:

Banca / Crédito

- Ratio deuda/ingreso
- Volatilidad de ingresos (últimos 12 meses)
- Días promedio de mora histórica
- Utilización de cupo (%)
- Antigüedad bancaria

Política Pública

- Índice de hacinamiento
- Acceso a servicios (score compuesto)
- Distancia a centros de salud
- Ingreso per cápita del hogar
- Nivel educativo máximo del hogar

Marketing / Churn

- Tendencia de consumo (pendiente)
- Días desde última compra
- Ratio quejas/transacciones
- Cambio en frecuencia de uso
- Engagement score compuesto

Para el consultor económico

Su ventaja competitiva frente a un data scientist "puro" es el conocimiento del dominio económico. Saber qué variables crear y por qué es más valioso que dominar 20 algoritmos.

GridSearchCV y RandomizedSearchCV: Repaso Rápido

Una vez que las features son buenas, tuneamos hiperparámetros:

GridSearchCV

- Prueba *todas* las combinaciones
- Exhaustivo pero lento
- Ejemplo: $K \in \{3, 5, 7, 11\}$ y métrica $\in$ {euclidiana, manhattan}
- $4 \times 2 = 8$ combinaciones

Usar cuando: Pocos hiperparámetros, espacio de búsqueda pequeño.

RandomizedSearchCV

- Muestrea n combinaciones al azar
- Más rápido, escalable
- Ejemplo: $C \in [0{,}01, 100]$, $\gamma \in [0{,}001, 10]$
- Prueba 50 combinaciones aleatorias

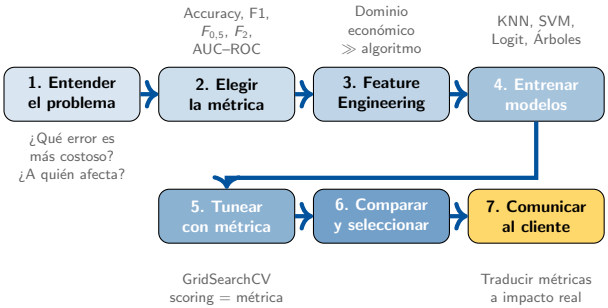
Usar cuando: Espacio grande, variables continuas (SVM con RBF).

Clave: ¿Con qué métrica tuneamos?

En `GridSearchCV(scoring=...)`, la métrica que usamos **debe alinearse con el objetivo de negocio**. No siempre es `'accuracy'`. Puede ser `'f1'`, `'recall'`, `'precision'`, o `'roc_auc'`.

Agenda de hoy

El Framework Completo



Resumen: Ideas Clave

1. **Accuracy no basta:** Con clases desbalanceadas o costos asimétricos, accuracy engaña.
2. **Cada error tiene un precio:** FP y FN tienen costos diferentes según el contexto. La métrica elegida refleja nuestras prioridades.
3. **IA Responsable:** Preguntarse siempre *"¿a quién afecta el error?"*. En política pública, FN suelen ser más graves (excluir vulnerables).
4. **Feature Engineering $\gg$ Tuning:** Crear buenas variables con conocimiento del dominio tiene más impacto que optimizar hiperparámetros.
5. **El scoring parámetro importa:** En GridSearchCV, tunear con la métrica correcta alinea el modelo con el objetivo de negocio.

Próxima clase: Árboles de decisión y Random Forests.

¡Gracias!

s.neira10@uniandes.edu.co