DigitalHouse > Coding School

DATA SCIENCE

MÓDULO 4

Evaluación del ajuste de un modelo

1. Comprender los fundamentos de los métodos de evaluación de modelos de clasificación
2. Entender las métricas de: Accuracy, Recall, Precisión, F1 y curvas ROC
3. Implementar estas métricas de evaluación con Scikit-Learn

Evaluación del ajuste de un modelo

Los modelos de machine learning usados para clasificación se evalúan de manera diferente a las regresiones:

- En una **regresión** buscamos predecir una **variable continua**; en un **clasificador**, en cambio, el objetivo es predecir la pertenencia o la probabilidad de **pertenencia a una clase**.
- Existen varias maneras de **evaluar la performance de un clasificador.** Es importante elegir la adecuada para el problema en mano.

- Los outcomes en una clasificación en función de la tasa de acierto se pueden dividir en en **cuatro clases**. Pongamos como ejemplo un clasificador que determina si un individuo pertenece o no a la clase “enfermo”.
- Definiciones:
  - **Falsos Positivos (FP):** es una clase negativa que fue clasificada como positivo. Ejemplo: al individuo se lo clasificó como enfermo, pero estaba sano.
  - **Falsos Negativos (FN):** es una clase positiva que fue clasificada como negativa. Ejemplo: al individuo se lo clasificó como sano, pero estaba enfermo.
  - **Verdaderos Positivos (TP):** es una clase positiva clasificada correctamente.
  - **Verdaderos Negativos (TN)**: es una clase negativa clasificada correctamente.
- Aclaración: la noción de “positivos” o “negativos” es arbitraria y podría ser reemplazada por las de “presencia” o ”ausencia”.

La **matriz de confusión** es una tabla de doble entrada donde se describen los **resultados observados vs. resultados esperados** luego de haber aplicado del modelo.

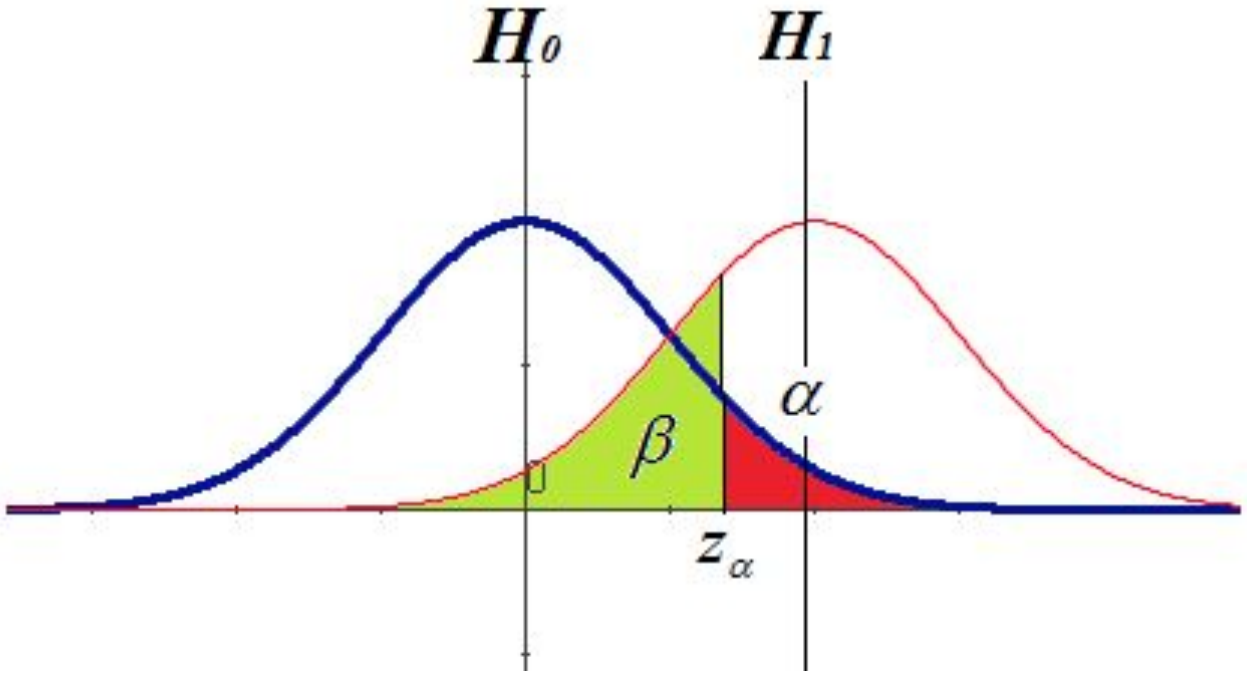
	Predice que está saludable (ŷ=0)	Predice que está enfermo (ŷ=1)
Está saludable (y=0)	46 **(TN)**	85 **(FP)**
Está enfermo (y=1)	168 **(FN)**	31 **(TP)**

- Nos permite **discernir entre los casos bien clasificados y los que fueron erróneamente clasificados** por el modelo.
- Es importante porque desde acá parten las categorías de **TP, TN, FP y FN**.

Al realizar un **test de hipótesis**, podemos incurrir en dos tipos de errores:

1. **Error de Tipo I**
2. **Error de Tipo II**

	No rechazar H0 ($\hat{y}$=0)	Rechazar H0 ($\hat{y}$=1)
H0 verdadera (y=0)	Decisión correcta (1 **-** α)	**Error de Tipo I (α)**
H0 falsa (y=1)	**Error de Tipo II (β)**	Decisión correcta (1 - β)


H_0
H_1
α
β
z_α

Explorando métricas de evaluación

Accuracy

- Proporción de **clases correctamente predichas**.
- Notar que esta métrica no "alerta" si no tengo casos positivos bien predichos.

	Predice que está saludable (y=0)	Predice que está enfermo (y=1)
Está saludable (y=0)	46 **(TN)**	85 **(FP)**
Está enfermo (y=1)	168 **(FN)**	31 **(TP)**

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

Recall (Sensibilidad o True Positive Rate):

- Proporción de **positivos correctamente predichos**.
- Si su valor es bajo, es porque hay presencia de falsos negativos. Por eso, esta medida es **sensible a los FN**.

	Predice que está saludable (y=0)	Predice que está enfermo (y=1)
Está saludable (y=0)	46 **(TN)**	85 **(FP)**
Está enfermo (y=1)	168 **(FN)**	31 **(TP)**

$$\text{Recall (Sensitivity)} = \frac{TP}{TP + FN}$$

Precisión:

- Cantidad de **verdaderos positivos sobre el total de predicciones positivas**.
- Si su valor es bajo, es porque hay presencia de falsos positivos. Por eso, esta medida es **sensible a los FP**.

	Predice que está saludable (y=0)	Predice que está enfermo (y=1)
Está saludable (y=0)	46 **(TN)**	85 **(FP)**
Está enfermo (y=1)	168 **(FN)**	31 **(TP)**

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

F1:

- Es la **media armónica de los scores de recall y precisión**.
- Como regla general, **cuanto mayor es esta métrica, mejor es el modelo**.

$$F1 = \frac{2}{\frac{1}{precision} + \frac{1}{recall}} = \frac{2 * (precision * recall)}{precision + recall}$$

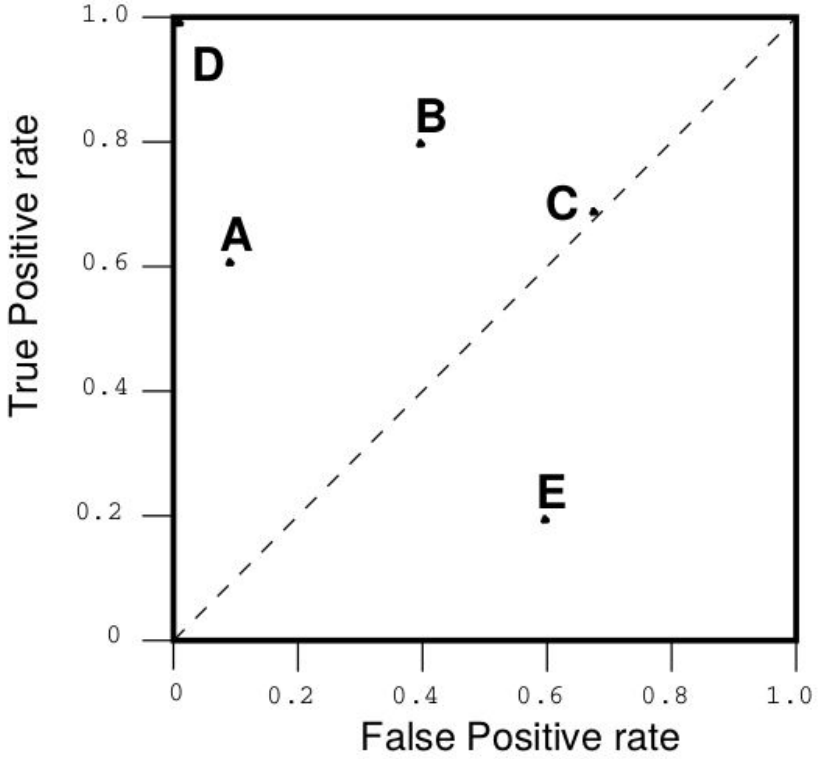
Fβ:

- Es la **media armónica ponderada** de los scores de recall y precisión**.**
- Con el parámetro β se puede regular la **importancia relativa de cada término**.

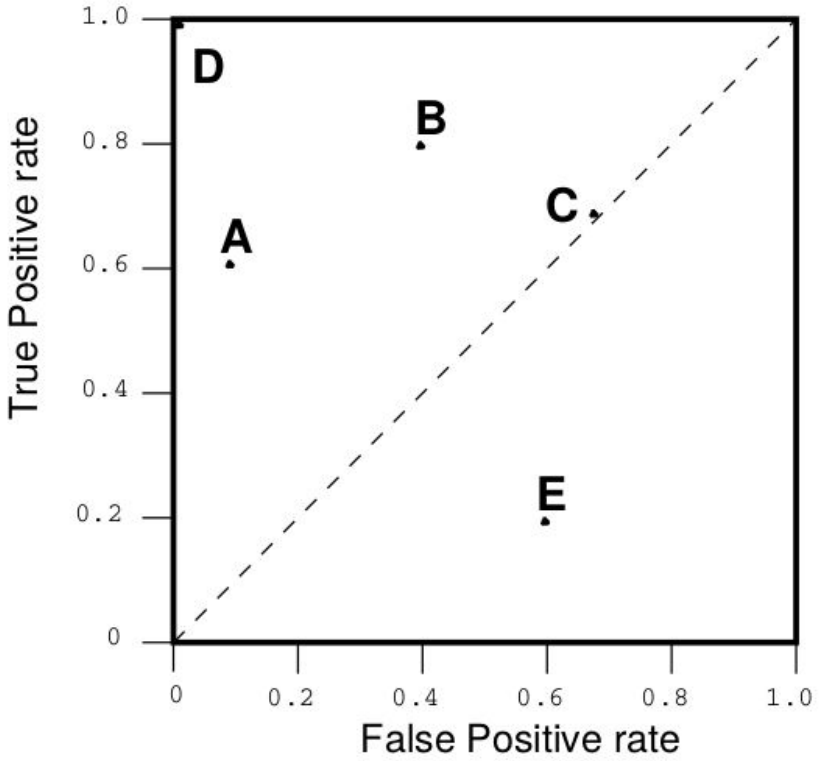
$$F_\beta = (1 + \beta^2) \cdot \frac{\text{precision} \cdot \text{recall}}{\beta^2 \cdot \text{precision} + \text{recall}}$$

- Es una forma de representar los métricas en un modelo de clasificación binaria:
  - True Positive Ratio = **TP / (TP + FN)**
  - False Positive Ratio = **FP / (FP + TN)**
- Mundo ideal: mi modelo debería tener una Sensibilidad (TPR) de 100% y una FPR de 0%.

- Cada punto es un modelo representado por su relación entre las TPR y FPR.
- ¿Cuál es el mejor modelo? ¿Por qué?
- ¿Qué puede decirse del punto **D**?
- ¿Y Del punto **E**?



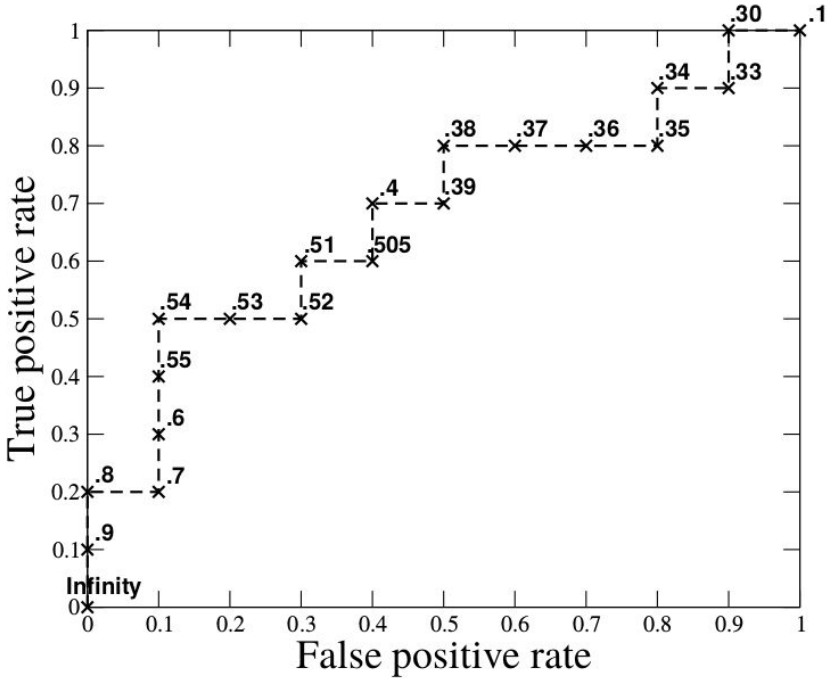
- Cuanto más al **noroeste del gráfico**, **mejor**: la TPR es alta y la FPR, baja.
- **Clasificadores cerca del eje X => "conservadores"**: clasificaciones positivas solamente con fuerte evidencia, obtiene pocos FP pero también pocos TP.
- **Clasificadores al noreste del gráfico => "liberales"**: clasificaciones positivas con poca evidencia, obtiene muchos TP pero también muchos FP.
- La **diagonal => "random guess"**



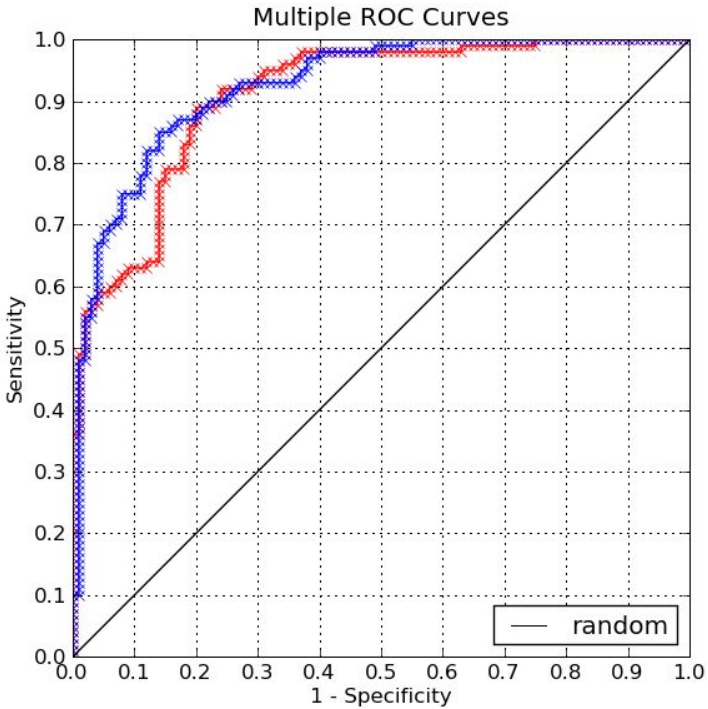
- La relación entre TPR y FPR depende del **umbral de decisión**: una **probabilidad mínima** definida a partir de la cual se realiza una clasificación positiva.
- Existen modelos (NB, regresión logística, etc.) que predicen de forma natural scores o **probabilidades de pertenencia**, en lugar de etiquetas de clase.
- Tales probabilidades se contrastan contra un umbral **T** de decisión:
    - Si la probabilidad de pertenencia a la clase positiva se encuentra por encima del umbral **T**, el caso es clasificado como positivo.
    - Si está por debajo, es clasificado como negativo.
- La **performance de un modelo varía conforme al umbral T.** Una buena práctica consiste en modificar este valor y analizar cómo impacta en los resultados.

A medida que modificamos **T**, el modelo performa diferente:

1. **T** = 0.90 => FPR: 0.00 y TPR: 0.10
2. **T** = 0.80 => FPR: 0.00 y TPR: 0.20
3. **T** = 0.70 => FPR: 0.10 y TPR: 0.20
4. **T** = 0.60 => FPR: 0.10 y TPR: 0.30
5. **T** = 0.50 => FPR: 0.40 y TPR: 0.60
6. **T** = 0.40 => FPR: 0.40 y TPR: 0.70
7. **T** = 0.30=> FPR: 0.90 y TPR: 1.00



- Teniendo en cuenta esto…
    - ¿Cuál de los siguientes modelos es mejor?
- Una buena medida es el **área debajo de la curva ROC**:
    - Cuanto mayor sea el área… mejor será el modelo. ¿Por qué?



- Es posible generar **métricas** para evaluar modelos de clasificación **a partir de la matriz de confusión**.
- Esas métricas permiten **discernir entre casos bien y mal clasificados** por el modelo.
- En ocasiones, una métrica de performance general como el **Accuracy** no resulta suficiente para evaluar modelos que no toleran la presencia de falsos negativos o positivos. Para esos casos, utilizamos el **Recall** y la **Precisión**, respectivamente.
- Las **curvas ROC** son una buena herramienta para **visualizar la performance general** del modelo.