

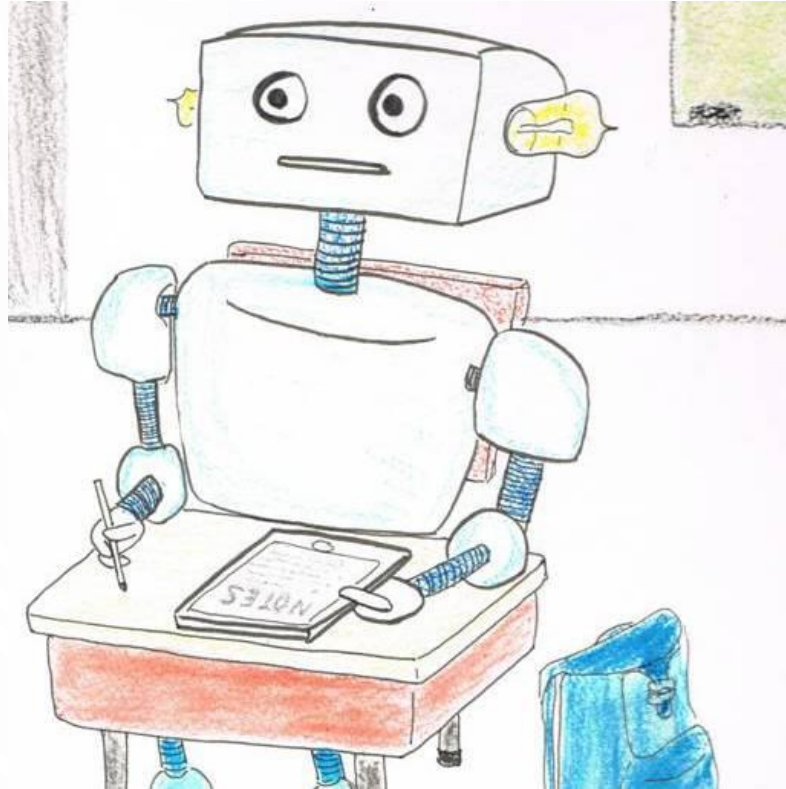
MACHINE LEARNING INMERSION

ANDRÉ OMAR CHÁVEZ PANDURO

« La **virtud de cada ser humano** , es saber mantener el **equilibrio ante sus victorias** y **no caerse** ante sus derrotas»



EVALUANDO UN ALGORITMO DE MACHINE LEARNING



IDEAS FUNDAMENTALES

- Existen medidas de error utilizadas para la evaluación de modelos de clasificación. Muchas de estas medidas se calculan en función de la matriz de confusión asociada al modelo, la que se define a continuación:
 - ✓ Error
 - ✓ Sensibilidad
 - ✓ Especificidad
 - ✓ Acierto
 - ✓ Youden
- Asimismo existen otros indicadores que nos ayude a validar modelos como:
 - ✓ AUC (área bajo la curva)
 - ✓ GINI
- Otro método es la de la Validación Cruzada

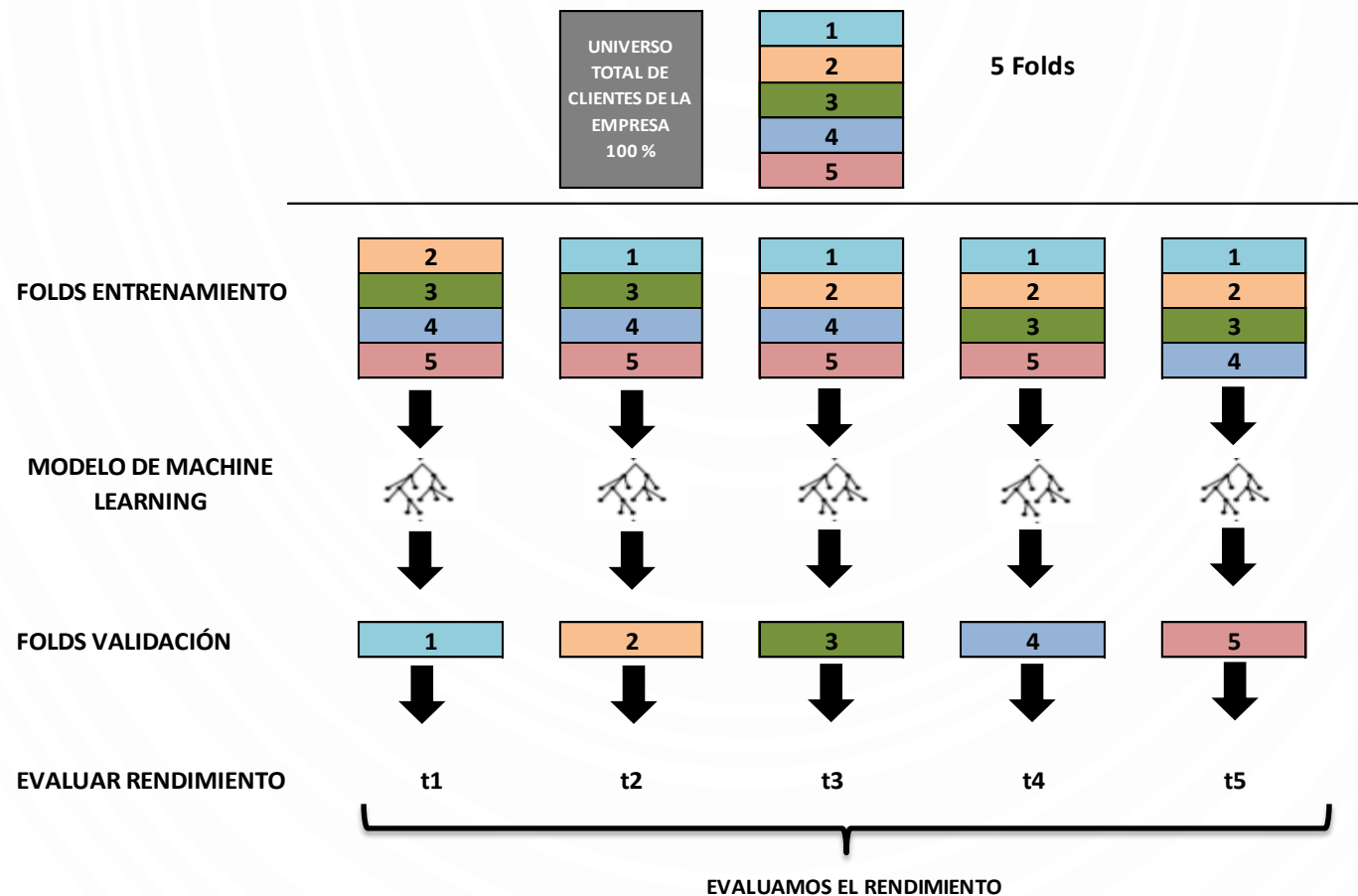
EVALUANDO UN ALGORITMO DE MACHINE LEARNING

MUESTRA DE ENTRENAMIENTO Y VALIDACIÓN



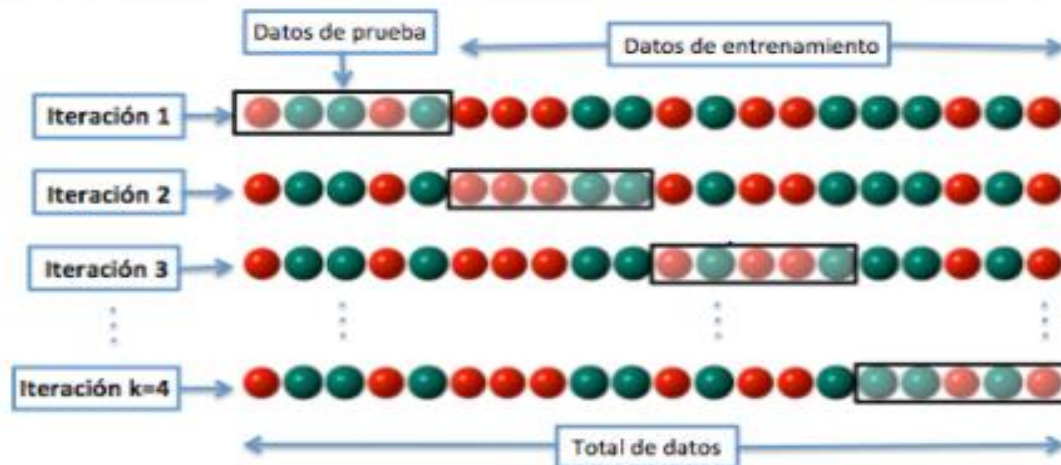
EVALUANDO UN ALGORITMO DE MACHINE LEARNING

VALIDACIÓN CRUZADA



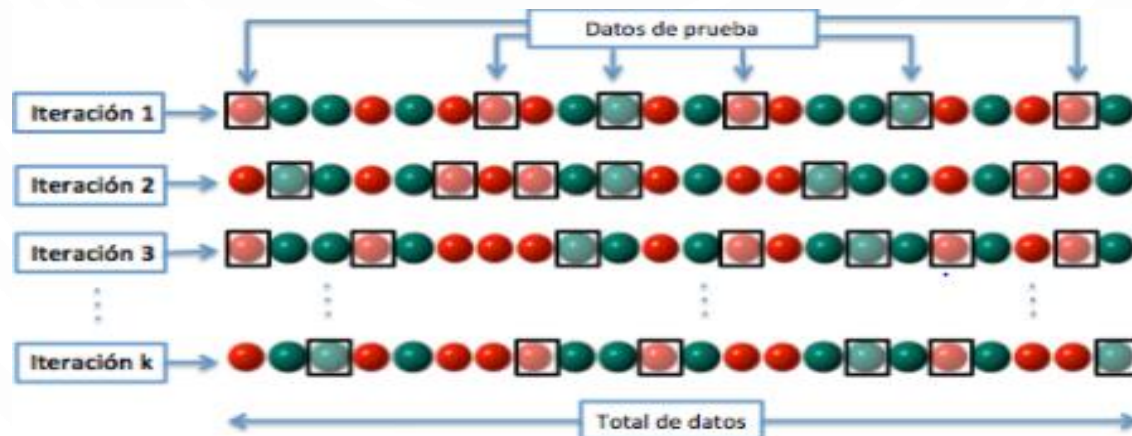
VALIDACIÓN CRUZADA DE K ITERACIONES O K-FOLD CROSS-VALIDATION.

- Los **datos** de muestra se dividen en K subconjuntos. Uno de los subconjuntos se utiliza como datos de prueba y el resto ($K-1$) como datos de entrenamiento. El proceso de validación cruzada es repetido durante k iteraciones, con cada uno de los posibles subconjuntos de datos de prueba. Finalmente se realiza la media aritmética de los resultados de cada iteración para obtener un único resultado.



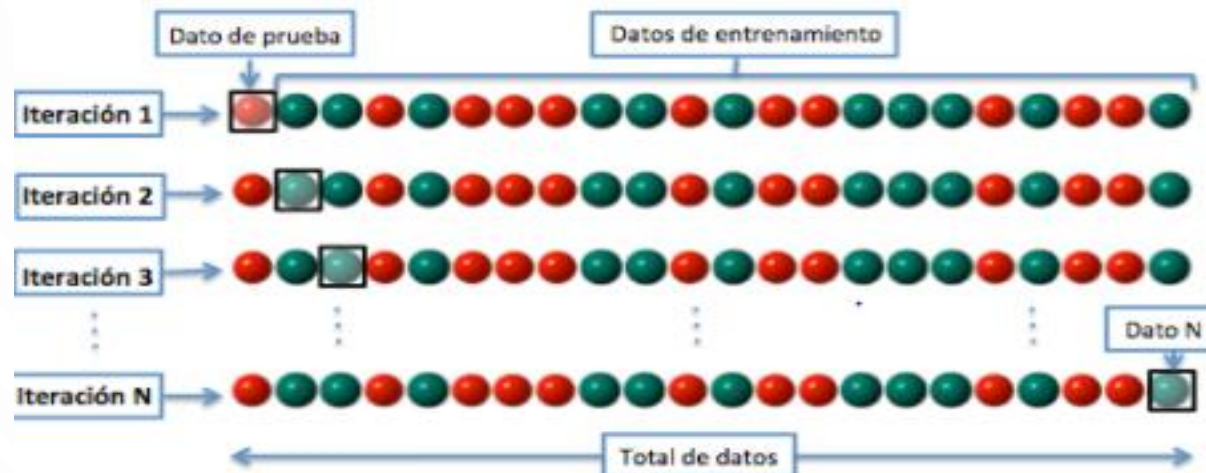
VALIDACIÓN CRUZADA ALEATORIA.

- La ventaja de este método es que la división de datos entrenamiento-prueba no depende del número de iteraciones. Pero, en cambio, con este método hay algunas muestras que quedan sin evaluar y otras que se evalúan más de una vez, es decir, los subconjuntos de prueba y entrenamiento se pueden solapar.



VALIDACIÓN CRUZADA DEJANDO UNO FUERA O LEAVE-ONE-OUT CROSS-VALIDATION (LOOCV).

- Se separan los datos de forma que para cada iteración tengamos una sola muestra para los datos de prueba y todo el resto conformando los datos de entrenamiento. La evaluación viene dada por el error, y en este tipo de validación cruzada el error es muy bajo, pero en cambio, a nivel computacional es muy costoso, puesto que se tienen que realizar un elevado número de iteraciones, tantas como N muestras tengamos y para cada una analizar los datos tanto de entrenamiento como de prueba.



EVALUANDO UN ALGORITMO DE MACHINE LEARNING

MATRIZ DE CONFUSIÓN Y MATRIZ DE COSTOS

MATRIZ DE CONFUSIÓN		PREDICCIÓN	
		NO MOROSOS	MOROSOS
REALIDAD	NO MOROSOS	DECISIÓN CORRECTA VN	FP
	MOROSOS	FN	DECISIÓN CORRECTA VP

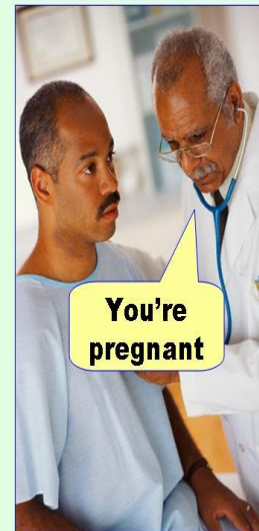
$$\text{PRECISIÓN} = (VN + VP) / (VN + VP + FP + FN)$$

$$\text{SENSIBILIDAD} = VP / (VP + FN)$$

$$\text{ESPECIFICIDAD} = VN / (VN + FP)$$

$$\text{F-SCORE} = 2 * ((VP / (VP + FP)) * (VP / (VP + FN))) / ((VP / (VP + FP)) + (VP / (VP + FN)))$$

Type I error
(false positive)



Type II error
(false negative)



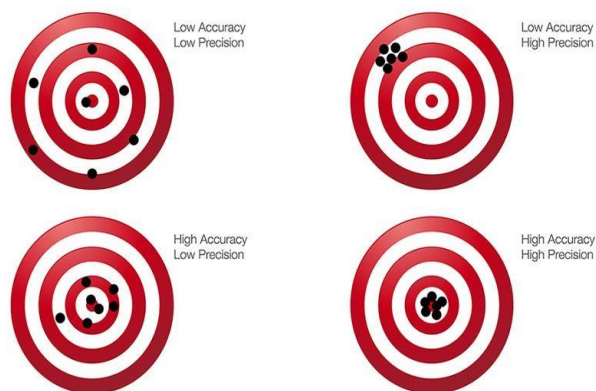
APLICACIÓN DE MACHINE LEARNING

Caso práctico: Clasificación del cáncer

- Casos en los que el nº de ejemplos negativos es mucho mayor que el de ejemplos positivos
- Ejemplo:

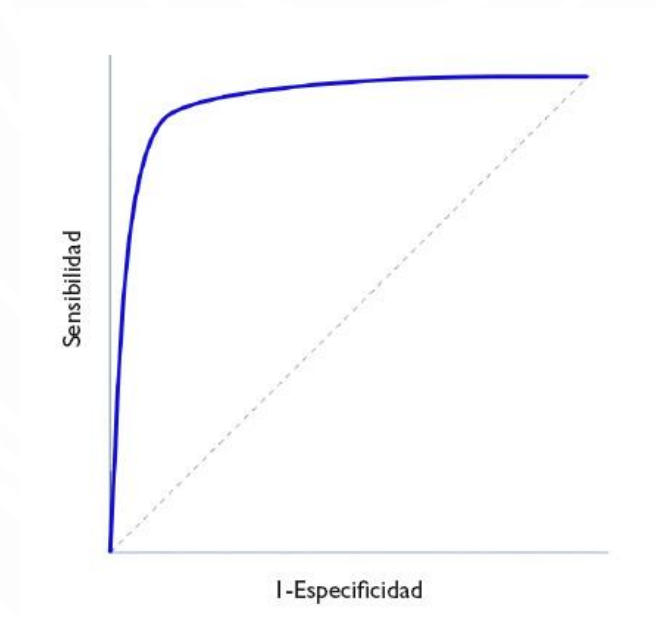
- Modelo regresión logística
 - $y = 1$ *cáncer*
 - $y = 0$ *no cáncer*
- Se tiene un 1 % de error en el set de test (99 % de diagnósticos correctos)
- Sólo el 0,5 % de los pacientes tiene cáncer

Exactitud vs. Precisión (Accuracy vs. Precision)



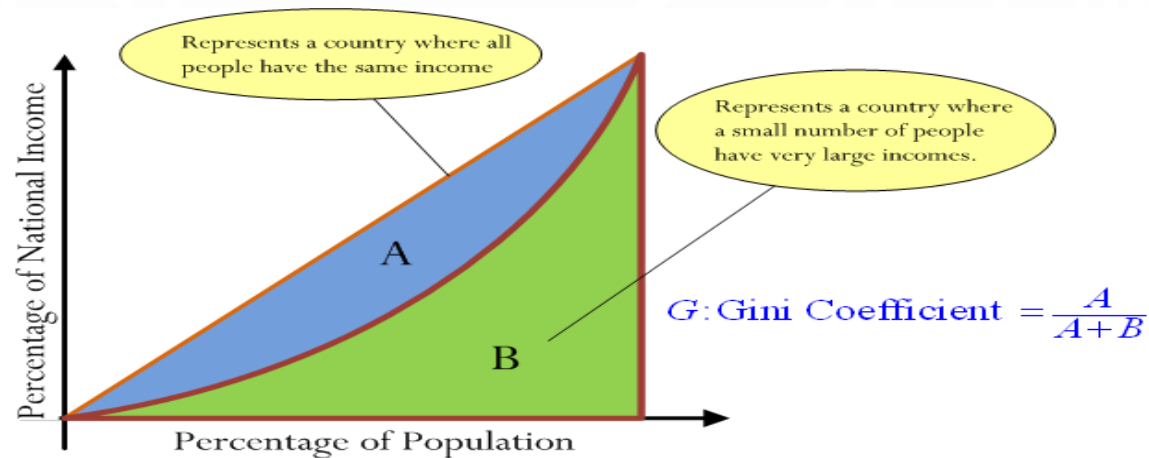
INDICADORES

Curva de ROC: Una curva ROC es una representación gráfica de la sensibilidad en función de los falsos positivos (complementario de la especificidad) para distintos puntos de corte. Un parámetro para evaluar la bondad de la prueba es el área bajo la curva que tomará valores entre 1 (prueba perfecta) y 0,5 (prueba inútil).



ÍNDICE DE GINI. $Gini = 2 * (ROC - 0.5)$

Si el valor del Gini se encuentra entre 0 y 0.25, decimos que el modelo predictivo tiene una clasificación "Baja"; si el valor del Gini se encuentra entre 0.25 y 0.45, tiene una clasificación "Aceptable"; si el valor del Gini se encuentra entre 0.45 y 0.6, tiene una clasificación "Buena", y finalmente, si el valor del Gini es mayor a 0.5, el modelo tiene una clasificación de "Muy buena".



ESTADISTICO KAPPA

- El estadístico **Kappa** es similar a la accuracy, pero tiene en cuenta la posibilidad de una predicción correcta por casualidad.
- Los valores de Kappa oscilan entre 0 y un máximo de 1, lo que indica una concordancia perfecta entre las predicciones del modelo y los valores reales. Los valores inferiores a uno indican una concordancia imperfecta.

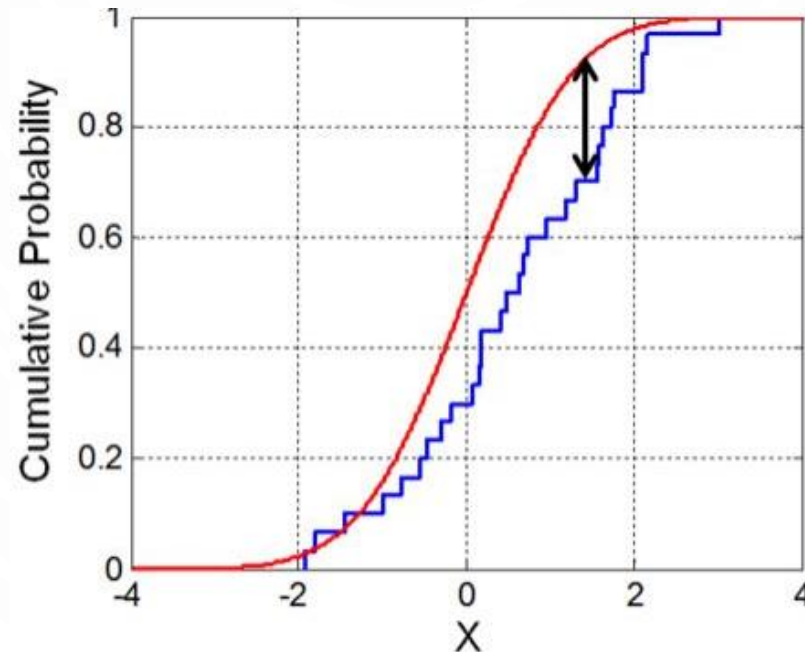
$$\text{Kappa} = \frac{\text{Concordancia}_{\text{Observada}} - \text{Concordancia}_{\text{Aleatoria}}}{1 - \text{Concordancia}_{\text{Aleatoria}}}$$

$$\text{Concordancia}_{\text{Observada}} = (a + d) / (a + b + c + d)$$

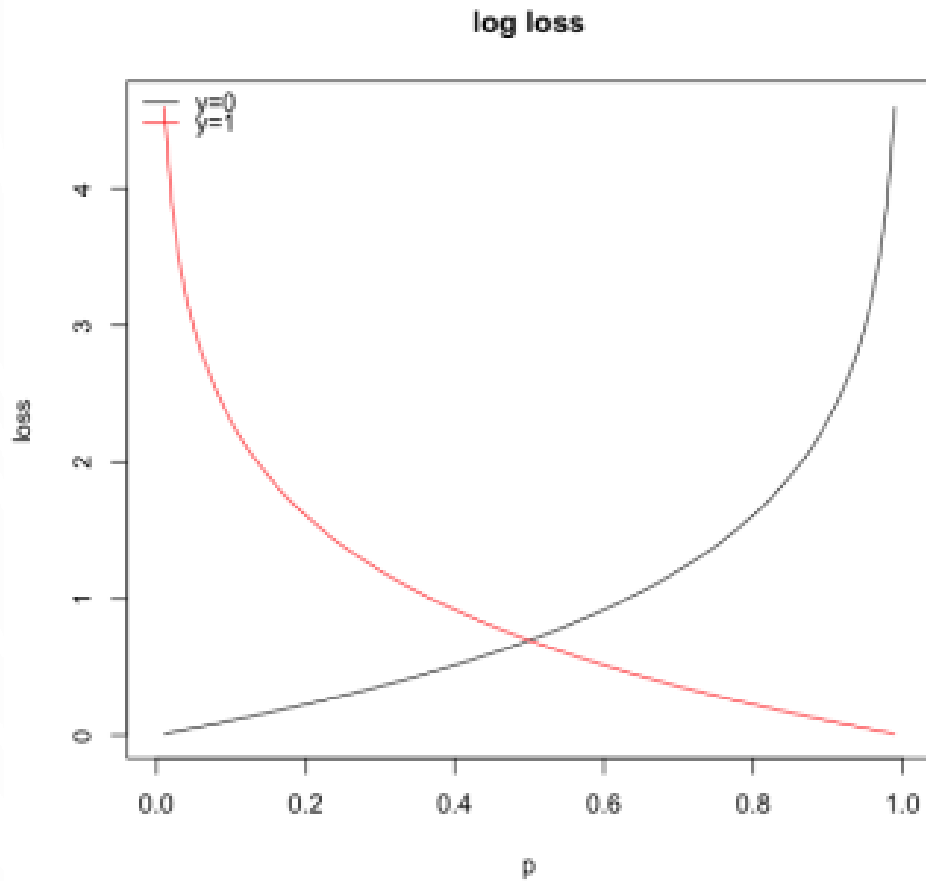
$$\begin{aligned} \text{Concordancia}_{\text{Aleatoria}} &= [a / (a + b)] * [a / (a + c)] + \\ (\text{esperada}) &[d / (c + d)] * [d / (b + d)] \end{aligned}$$

ESTADISTICO K-S

- El estadístico KS ayuda a entender, qué porción de la población debe ser dirigida para obtener la mayor tasa de respuesta.
- El KS se usa para tomar decisiones como: ¿Cuántos clientes deberíamos apuntar para dirigir una campaña de marketing?.



LOG - LOSS



**GRACIAS
POR SU ATENCIÓN**