Universidad **Ricardo Palma**

RECTORADO

PROGRAMA DE ESPECIALIZACIÓN EN CIENCIA DE DATOS

Formamos seres humanos para una cultura de paz

PROGRAMA DE ESPECIALIZACIÓN DATA SCIENCE NIVEL I

R + Python

MÓDULO I

Validación de Modelos Supervisados I

A nuestro recordado Maestro

Dr. Erwin Kraenau Espinal, Presidente de la Comisión de Creación de la Maestría en Ciencia de los Datos

« Es increíble como puedes cambiar tu vida , cuando decides cambiar un pensamiento»

AGENDA

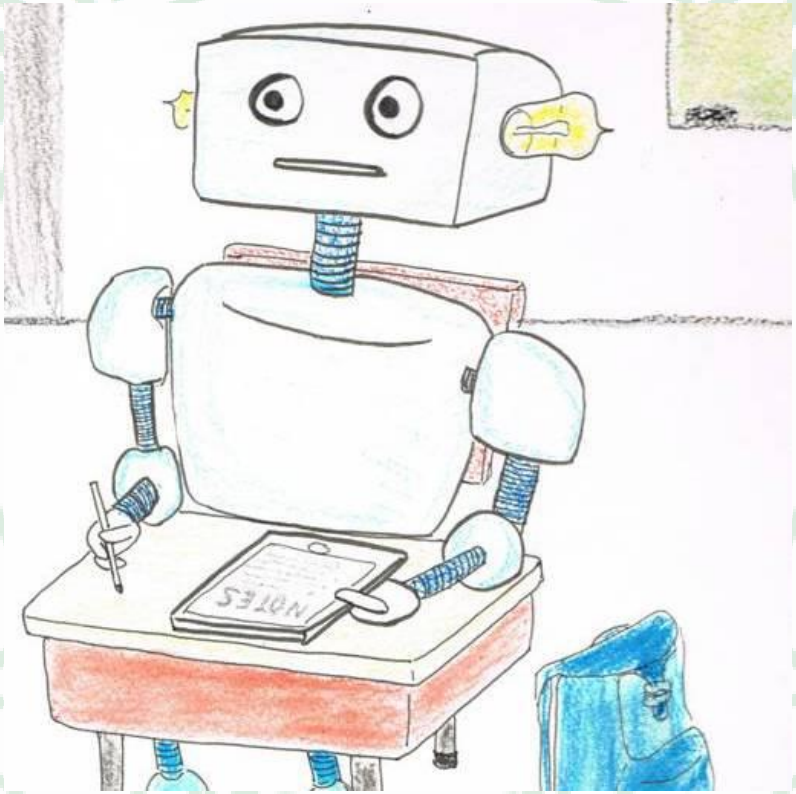
- Ideas Fundamentales
- Indicadores de Validación de Modelos.
- Validación Cruzada o CV.

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 o Precisión Global
  - ✓ Recall

- Asimismo existen otros indicadores que nos ayude a validar modelos como:
  - ✓ AUC (área bajo la curva)
  - ✓ GINI

EVALUANDO UN ALGORITMO DE MACHINE LEARNING



EVALUANDO UN ALGORITMO DE MACHINE LEARNING

MUESTRA DE ENTRENAMIENTO Y VALIDACIÓN

UNIVERSO TOTAL DE CLIENTES DE LA EMPRESA 100 %

MUESTRA DE ENTRENAMIENTO 75 %

MUESTRA DE VALIDACIÓN 25 %

UNIVERSO TOTAL DE CLIENTES DE LA EMPRESA 100 %

MUESTRA DE ENTRENAMIENTO 50 %

MUESTRA DE COMPROBACIÓN 30 %

MUESTRA DE VALIDACIÓN 20 %

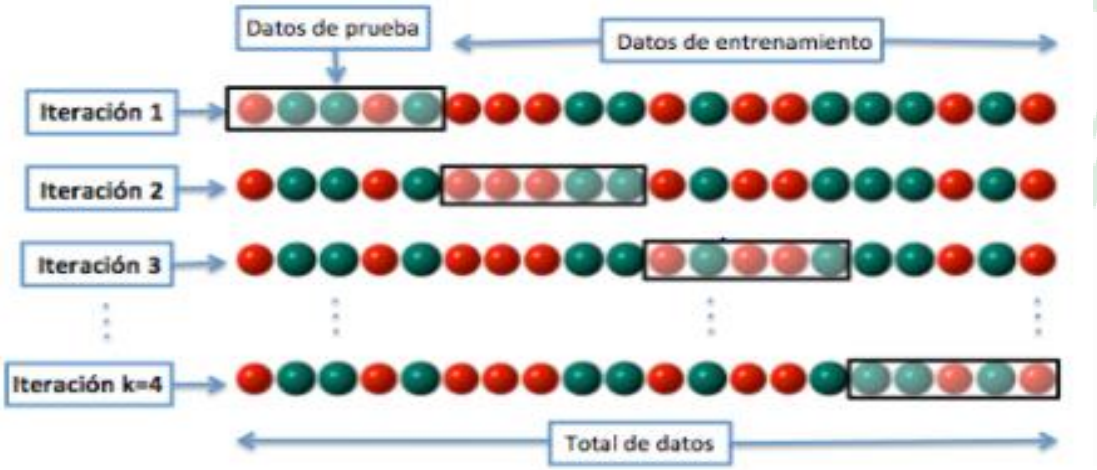
VALIDACIÓN CRUZADA

- La **validación cruzada o cross-validation** es una metodología utilizada para evaluar los resultados de un análisis y garantizar que son independientes de la partición entre datos de **entrenamiento y prueba**.

- Se utiliza en entornos donde el objetivo principal es la predicción y se quiere **estimar cómo de preciso es un modelo que se llevará a cabo a la práctica.**

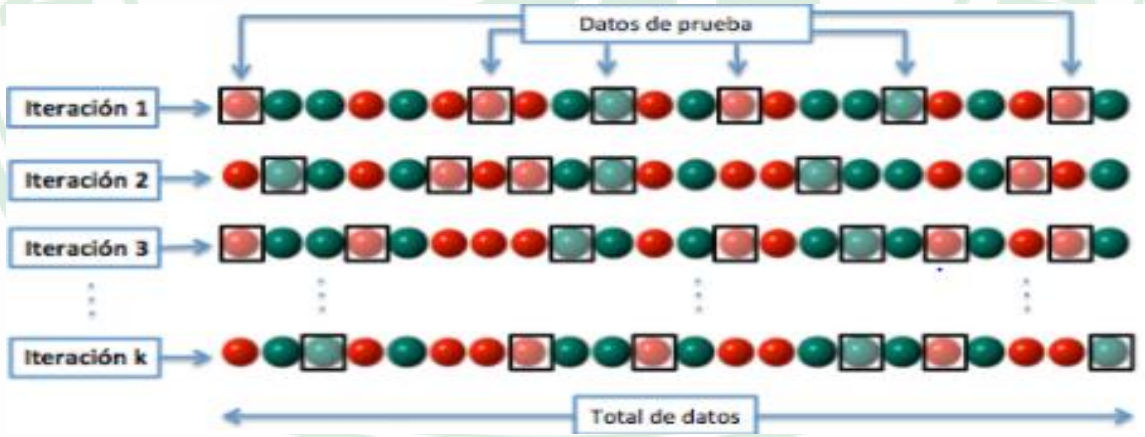
VALIDACIÓN CRUZADA DE K ITERACIONES O K-FOLD CROSS-VALIDATION.

- 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.



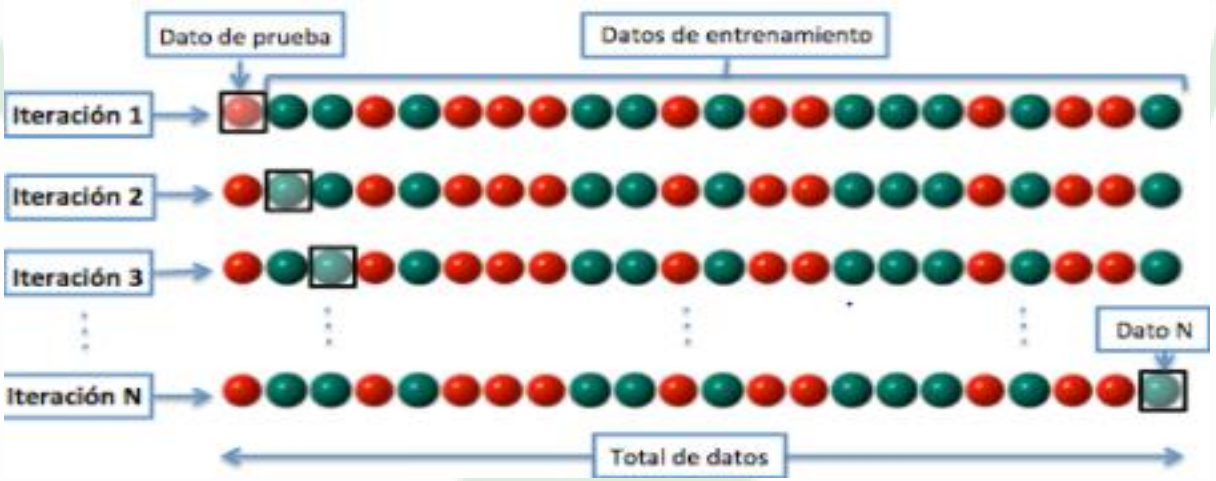
VALIDACIÓN CRUZADA ALEATORIA.

- 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.



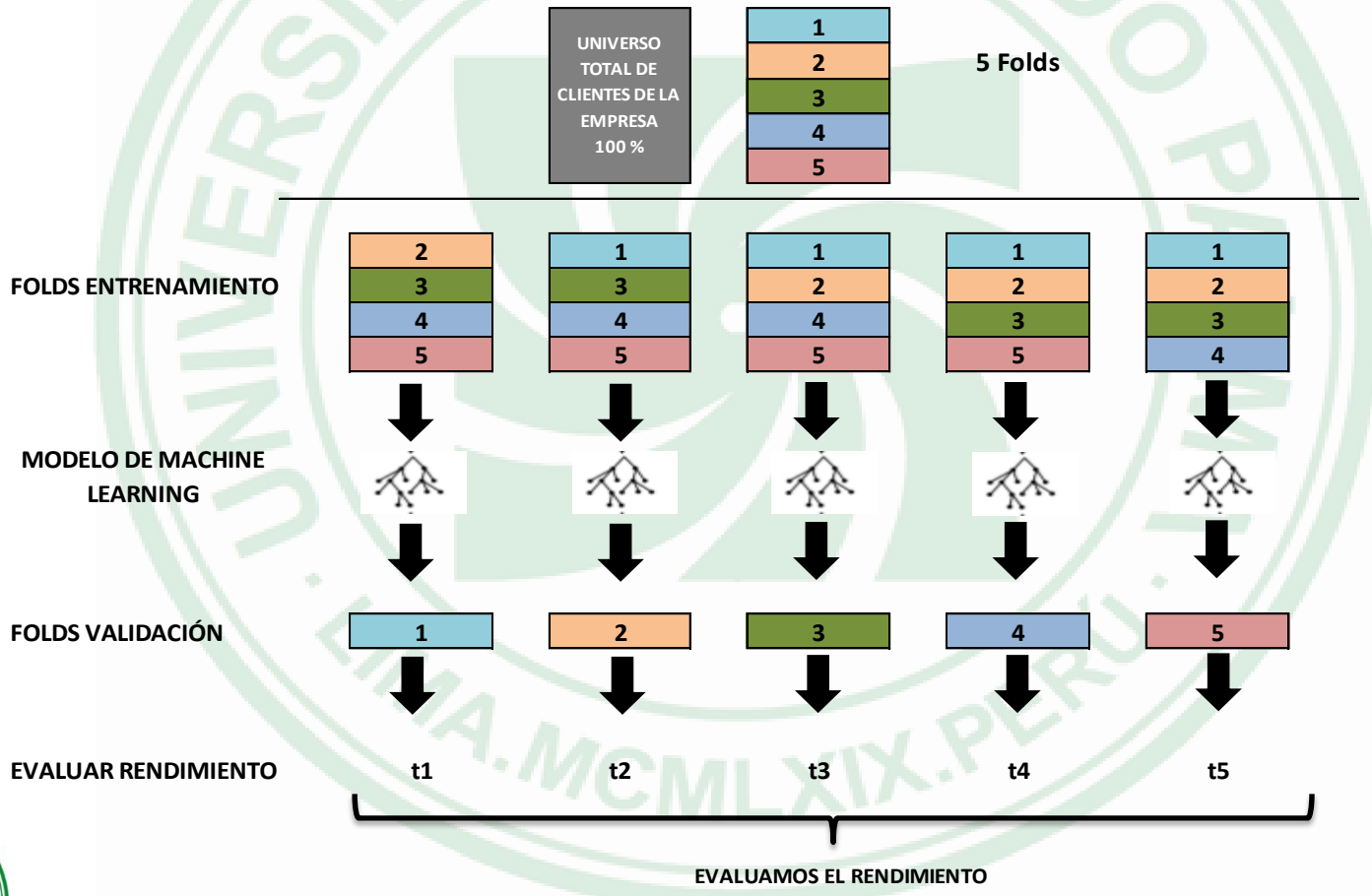
VALIDACIÓN CRUZADA DEJANDO UNO FUERA O LEAVE-ONE-OUT CROSS-VALIDATION (LOOCV).

- 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

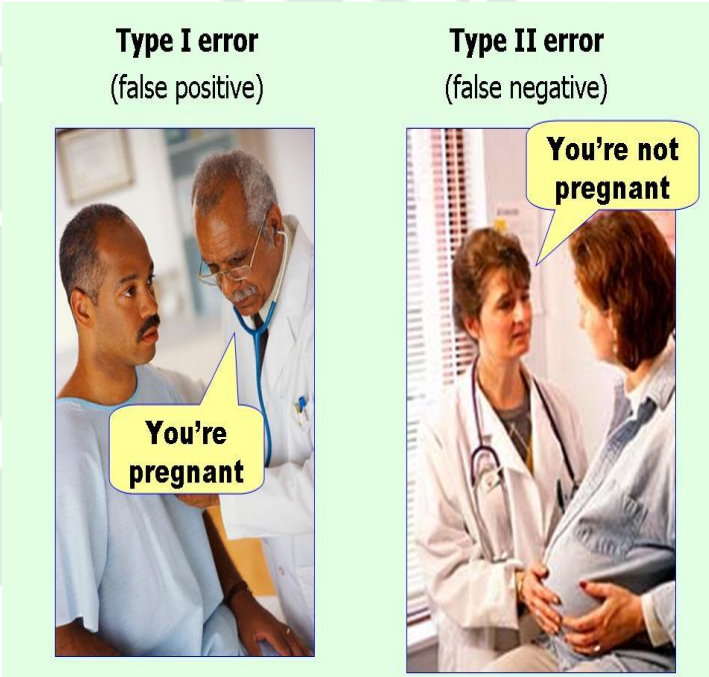
VALIDACIÓN CRUZADA : K - FOLDS



Evaluando un Algoritmo de Machine Learning

MATRIZ DE CONFUSIÓN Y MATRIZ DE COSTOS

MATRIZ DE CONFUSIÒN		PREDICCIÒN	
		NO MOROSOS	MOROSOS
REALIDAD	NO MOROSOS	DECISIÓN CORRECTA VN	**FP**
	MOROSOS	**FN**	DECISIÓN CORRECTA VP



PRECISIÓN = (VN + VP) / (VN + VP + FP + FN)

SENSIBILIDAD = VP / (VP + FN)

ESPECIFICIDAD = VN / (VN + FP)

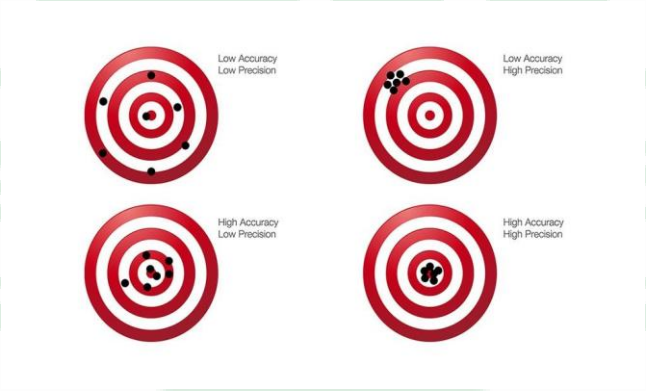
F-SCORE = 2 *((VP/ VP + FP) * (VP/ VP + FN)) / ((VP/ VP + FP) + (VP/ VP + FN))

Aplicación de Machine Learning

Caso práctico: Clasificación del cáncer

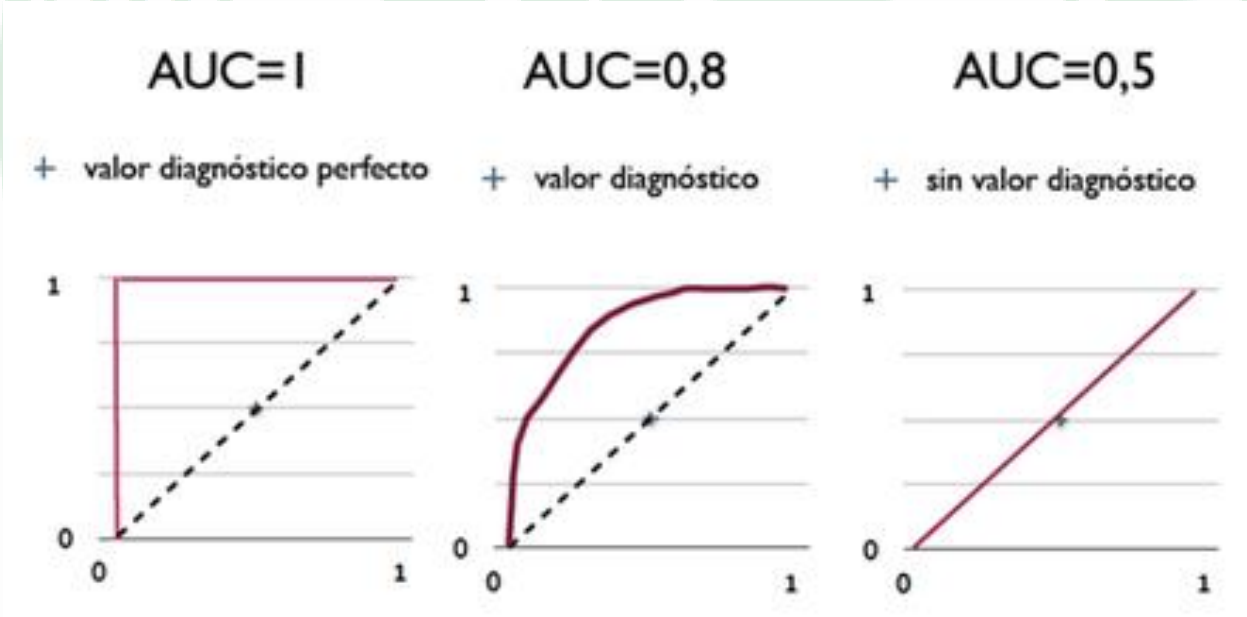
- Casos en los que el nº de ejemplos negativos es mucho mayor que el de ejemplos positivos
- Ejemplo:
  - Modelo regresión logística $y = 1$ $cáncer$ / $y = 0$ $no\ \ cáncer$
  - Se tiene un 1 % de error en el set de test (99 % de diagnósticos correctos)
  - Sólo el 0,5 % de los pacientes tiene cáncer

Exactitud vs. Precisión (Accuracy vs. Precision)



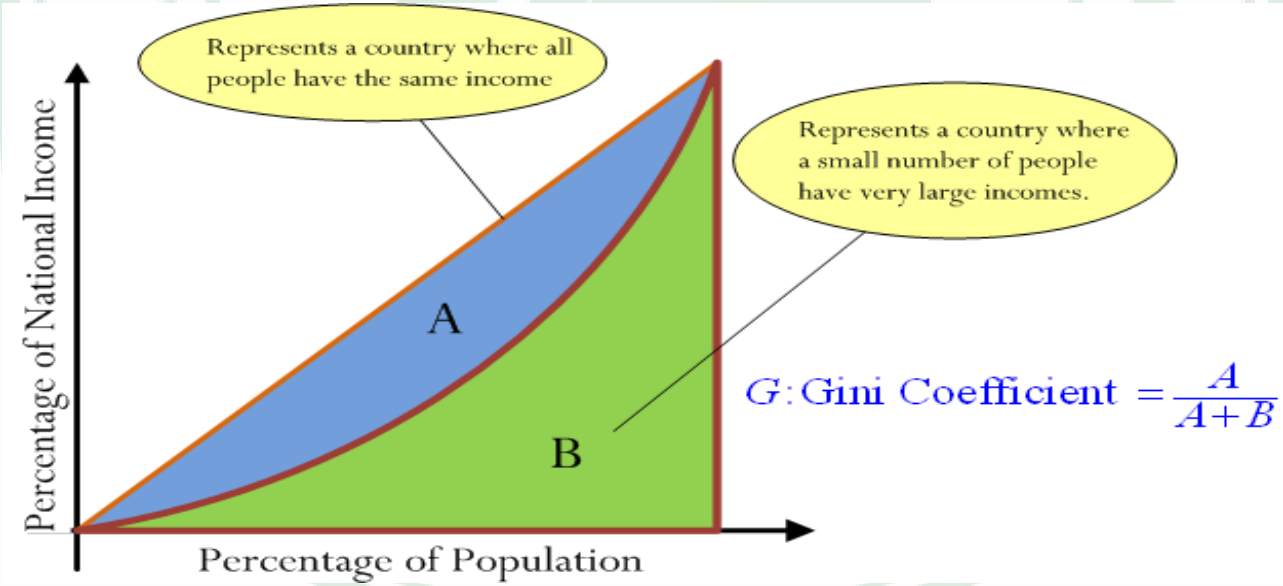
INDICADORES

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.

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



¡Gracias!

Comunidad Data Science Perú

Comunidad Data Science Perú