

MACHINE LEARNING IMMERSION

ANDRÉ OMAR CHÁVEZ PANDURO

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



EXPOSITOR

André Omar Chávez Panduro
UNMSM

MSc in Data Science Candidate
Promotion “Erwin Kraenau Espinal”
Universidad Ricard Palma



**Senior Data
Scientist**



**Customer
Intelligence Analyst**



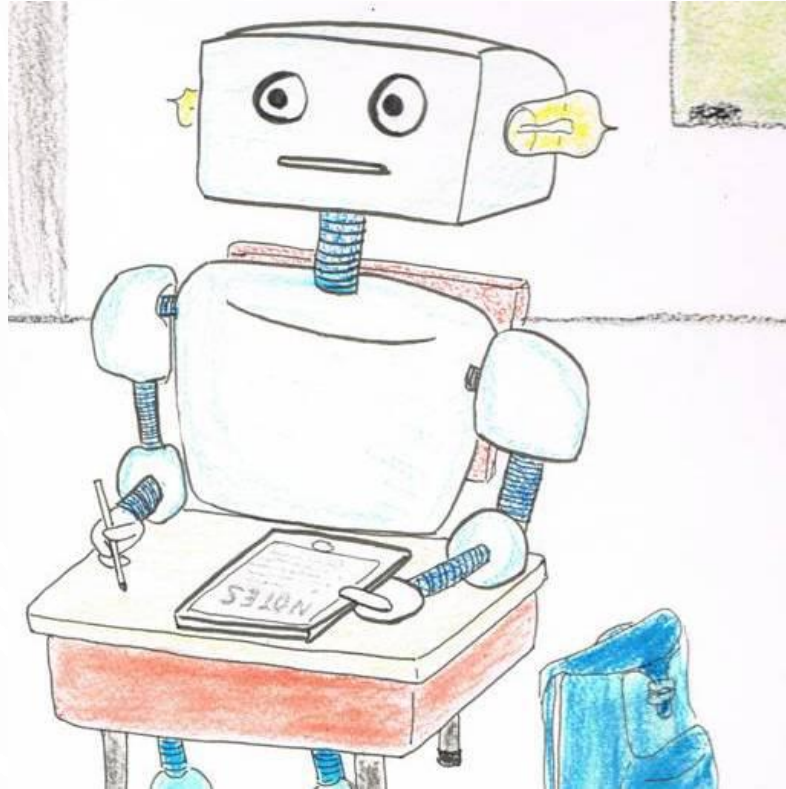
Data Analyst



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



EVALUANDO UN ALGORITMO DE MACHINE LEARNING

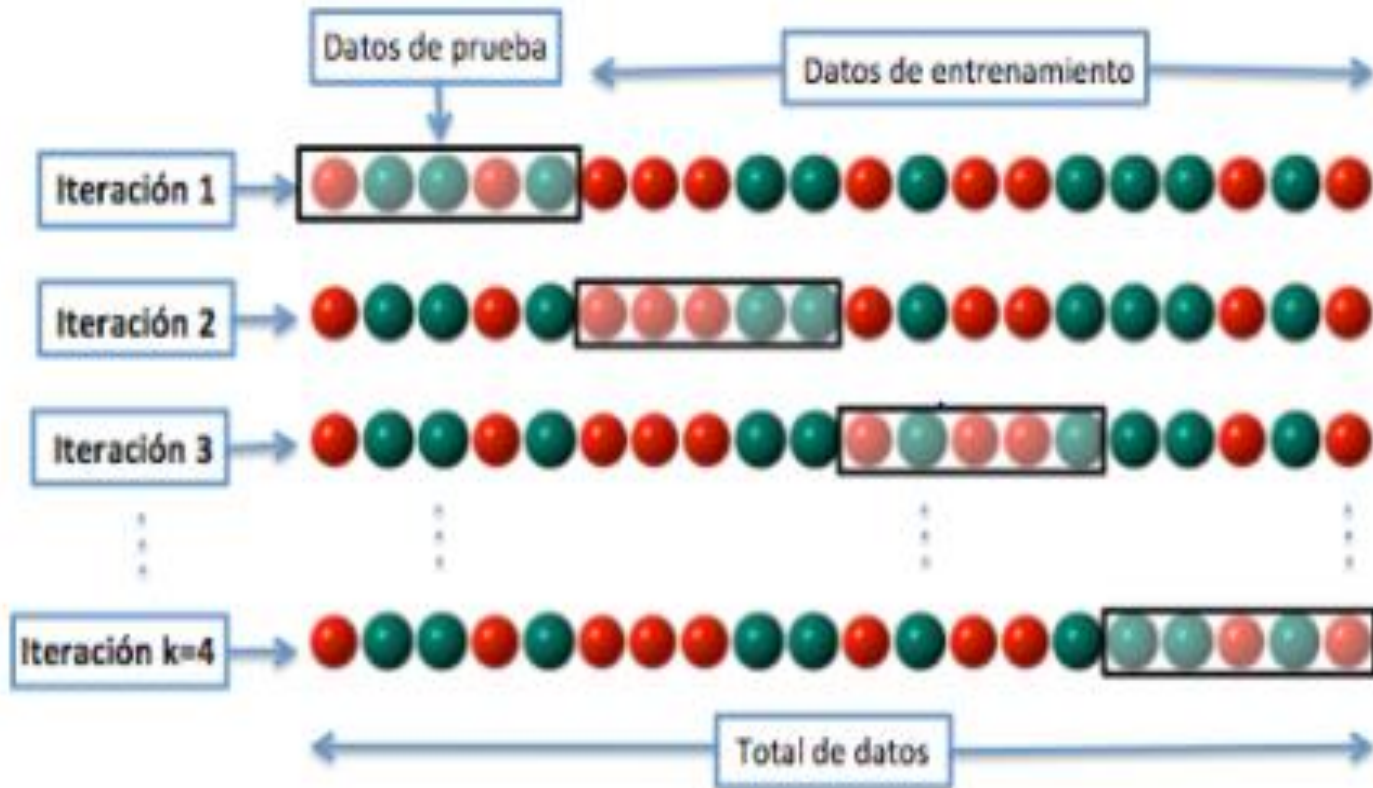
MUESTRA DE ENTRENAMIENTO Y VALIDACIÓN



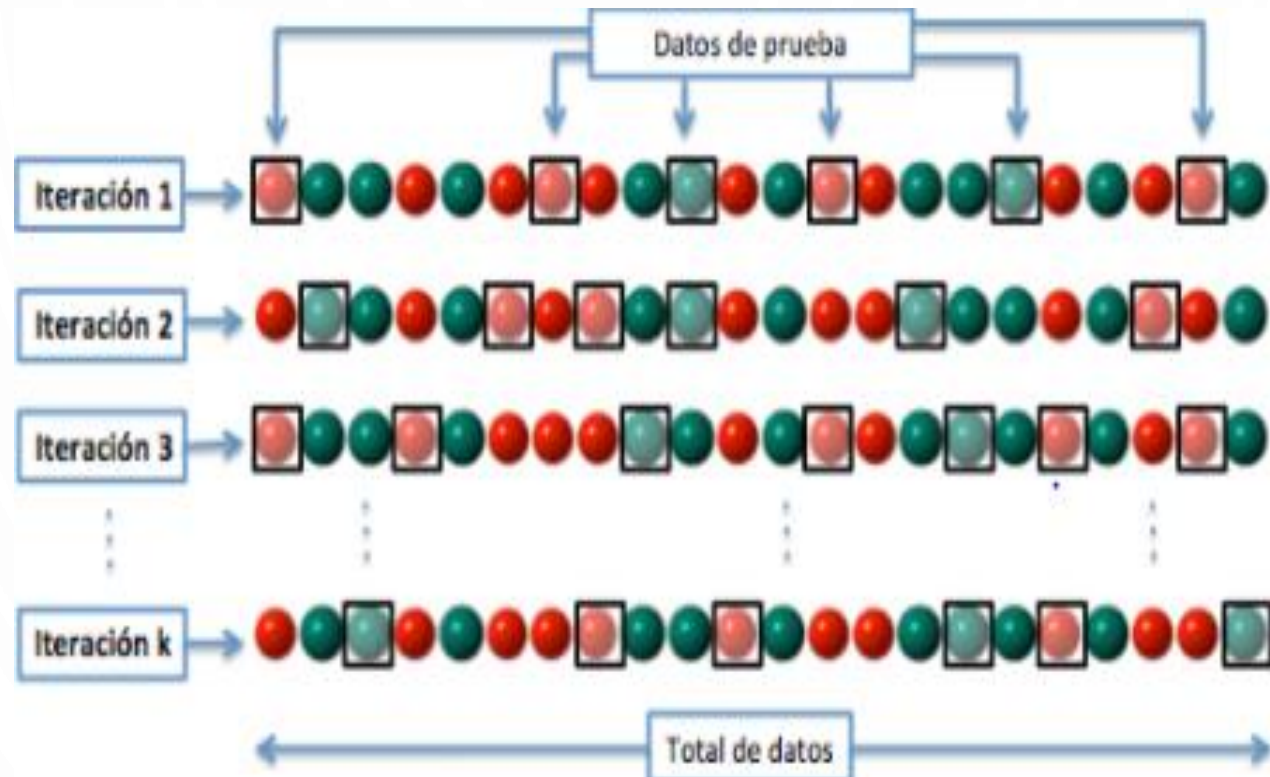
VALIDACIÓN CRUZADA

- 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**.

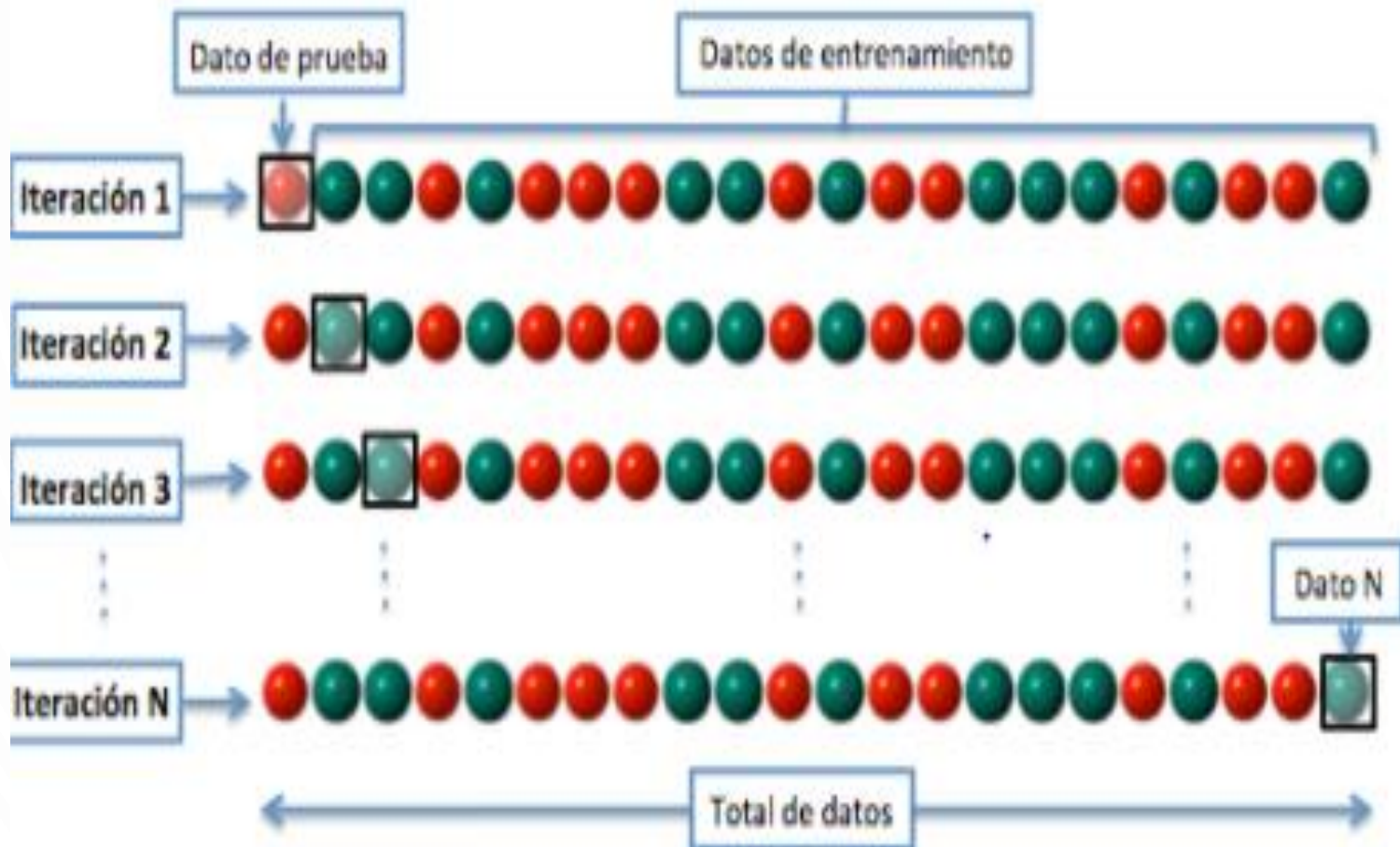
VALIDACIÓN CRUZADA DE K ITERACIONES O K-FOLD CROSS-VALIDATION.



VALIDACIÓN CRUZADA ALEATORIA.

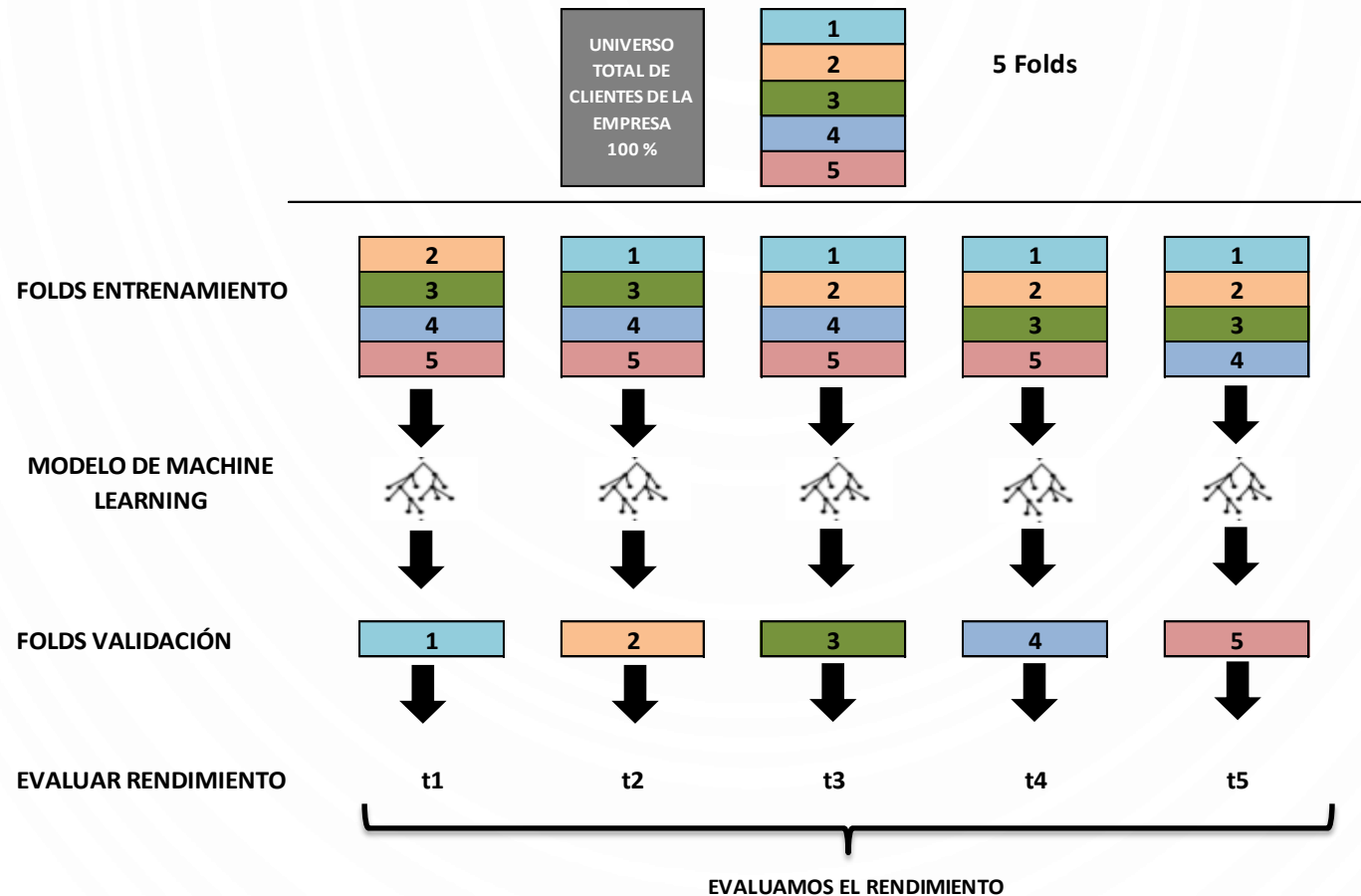


VALIDACIÓN CRUZADA DEJANDO UNO FUERA O LEAVE-ONE-OUT CROSS-VALIDATION (LOOCV).



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

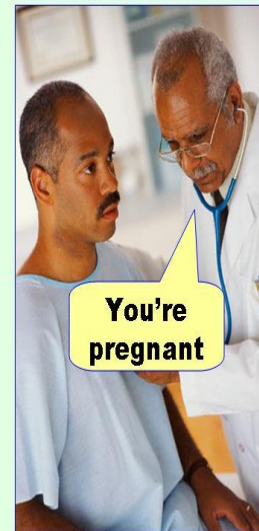
$$\text{PRECISIÓN} = (VN + VP) / (VN + VP + FP + FN)$$

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

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

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

Type I error
(false positive)



Type II error
(false negative)



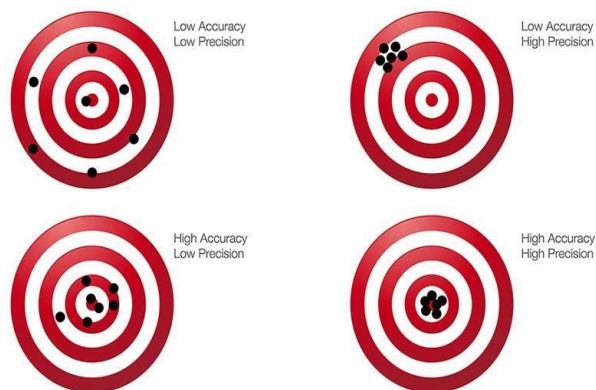
APLICACIÓN DE MACHINE LEARNING

Caso práctico: Clasificación del cáncer

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

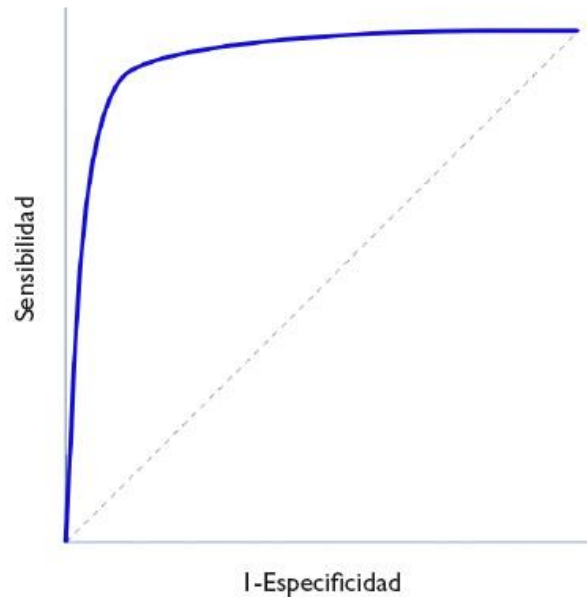
- Modelo regresión logística
 - $y = 1$ *cáncer*
 - $y = 0$ *no cáncer*
- Se tiene un 1 % de error en el set de test (99 % de diagnósticos correctos)
- Sólo el 0,5 % de los pacientes tiene cáncer

Exactitud vs. Precisión (Accuracy vs. Precision)



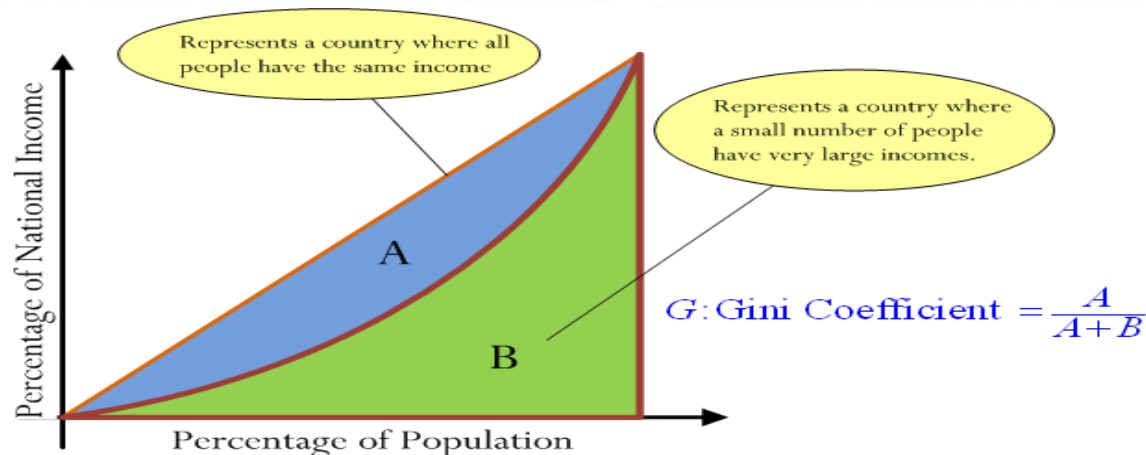
INDICADORES

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”.



**GRACIAS
POR SU ATENCIÓN**