

# **MODELOS DE PREDICCIÓN AVANZADOS**

**Mg Sc. Meza Rodríguez, Aldo Richard**



## TÉCNICAS SUPERVISADAS

### Árboles de decisión



### Supervised

- **Clasificación**

- Naive Bayes
- Logística
- Árboles de clasificación
- Bagging
- Adaboost
- Random Forest
- Redes neuronales, etc.

- **Regresión**

- Lineal
- Árboles de regresión
- Etc.



Label

### Unsupervised

- **Clustering**

- Kmean
- Clara
- Fanny
- Mona, etc.

- **Reglas de asociación**

- **Reducción de dimensiones**

- Análisis de compones principales
- SVD



## Indicadores para evaluación de modelos de clasificación

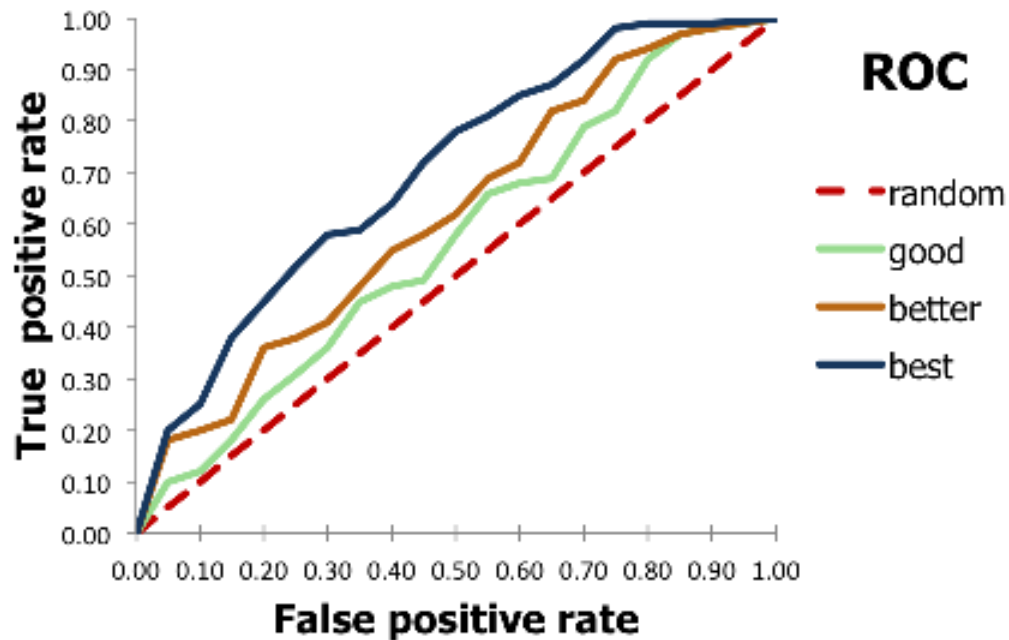
## MATRIZ DE CONFUSIÓN

		CLASE OBSERVADA (ACTUAL)	
		Positivo	Negativo
CLASE PREDICHA	Positivo	Verdadero Positivo (VP)	Falso Positivo (FP)
	Negativo	Falso Negativo (FN)	Verdadero Negativo (VN)

- ❑ **Accuracy:** Es la tasa de predicción correcta  $(VP + VN)/\text{Total de casos}$
- ❑ **Tasa de error:**  $(FP + FN)/\text{Total de casos} = 1 - \text{Accuracy}$
- ❑ **Sensibilidad:** Es la probabilidad de clasificar correctamente a un valor positivo.  $VP/(FN+VP)$
- ❑ **Especificidad:** Es la probabilidad de clasificar correctamente a un valor negativo.  $VN/(VN+FP)$
- ❑ **Precisión= (VP+):** Probabilidad de clasificar incorrectamente a un valor negativo =  $VP/(FP+VP)$
- ❑ **Valor Predictivo Negativo (VP-):** Probabilidad de clasificar incorrectamente a un valor positivo =  $VN/(VN+FN)$

## CURVAS ROC

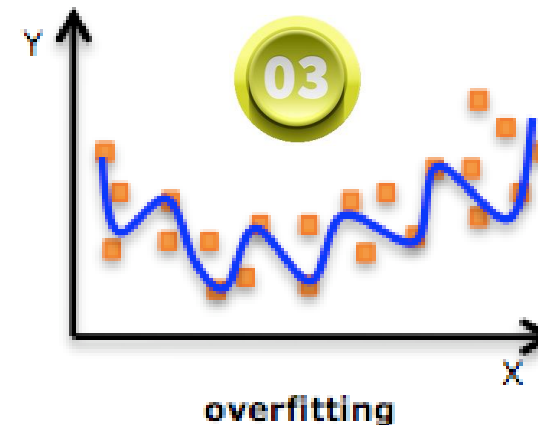
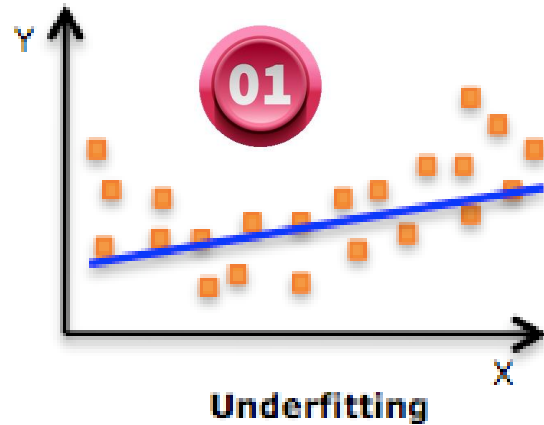
- Se forma con la tasa verdadera positiva (sensitividad) en función de la tasa falsa positiva (1-especificidad) para diferentes puntos de corte de un modelo.
- El clasificador tiene un mejor desempeño si la curva ROC se aleja más de la diagonal principal.
- El área bajo la curva (AUC) la probabilidad de que el clasificador pronostique correctamente a dos individuos. Uno que pertenece a la categoría  $Y=1$  y otro que pertenece a  $Y=0$ .



AUC ROC	Conclusión
0.5	No discrimina
0.6 - 0.7	Pobre
0.71 - 0.8	Aceptable
0.81 - 0.9	Excelente (bueno)
> 0.9	Muy bueno

## Métodos de Validación

## Problemas en los modelos

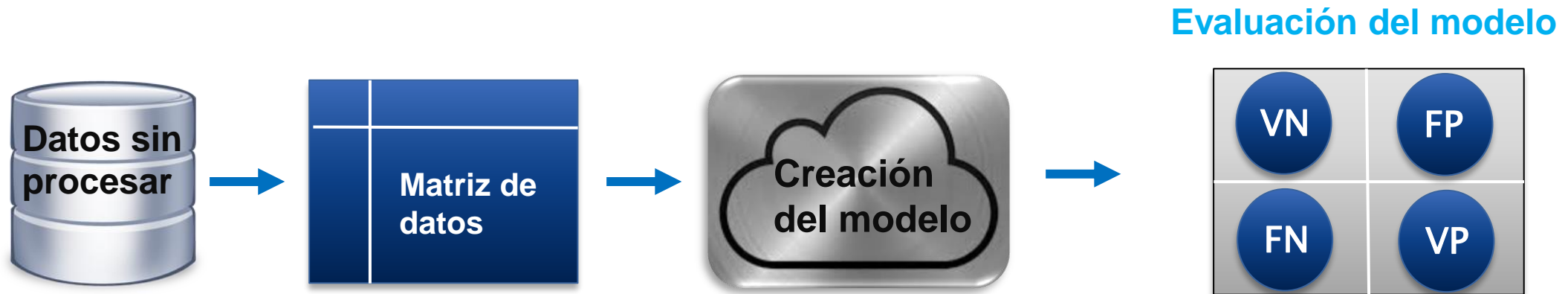


**1. Underfitting:** El modelo no captura la tendencia subyacente de los datos. La ecuación lineal tiene un alto error de los puntos de datos de entrenamiento. Esto no funcionará bien en la clasificación.

**2.** La relación entre las variables es correcta, el error de entrenamiento es bajo, existe buena generalización de los datos.

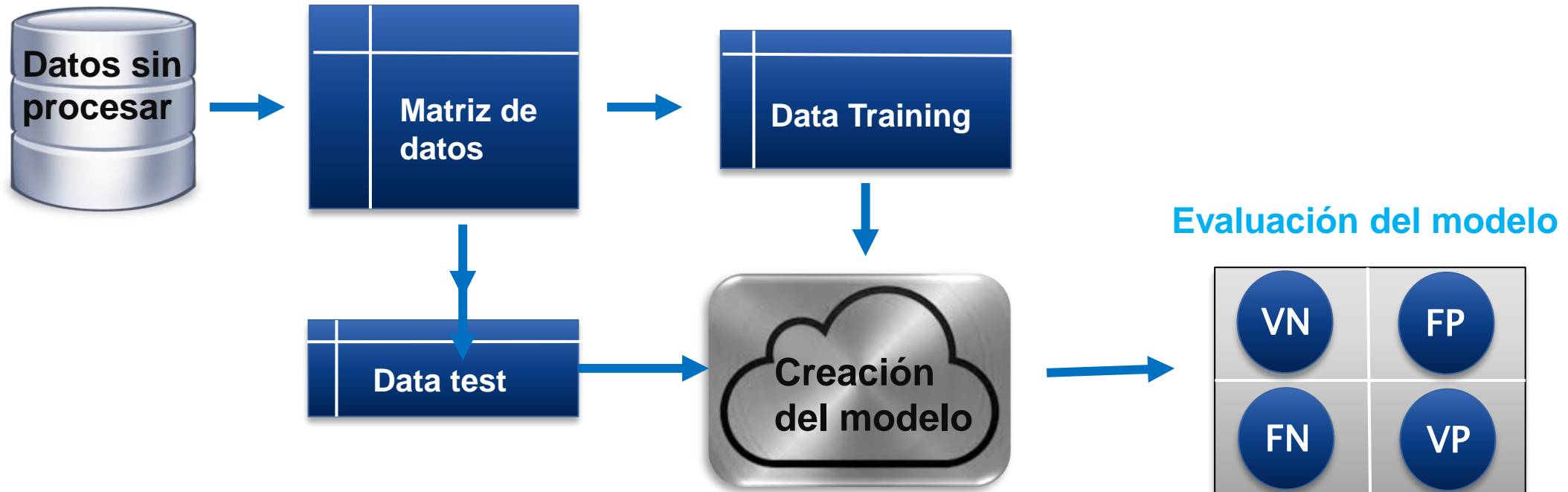
**3. Overfitting:** funciona bien en los datos de entrenamiento pero no tiene un buen rendimiento en datos no vistos. Usual en modelos **no paramétricos y no lineales**, sucede porque el modelo está memorizando la relación entre las variables  $X$  y  $Y$ , no puede generalizar bien los datos.

## VALIDACIÓN POR RESTITUCIÓN

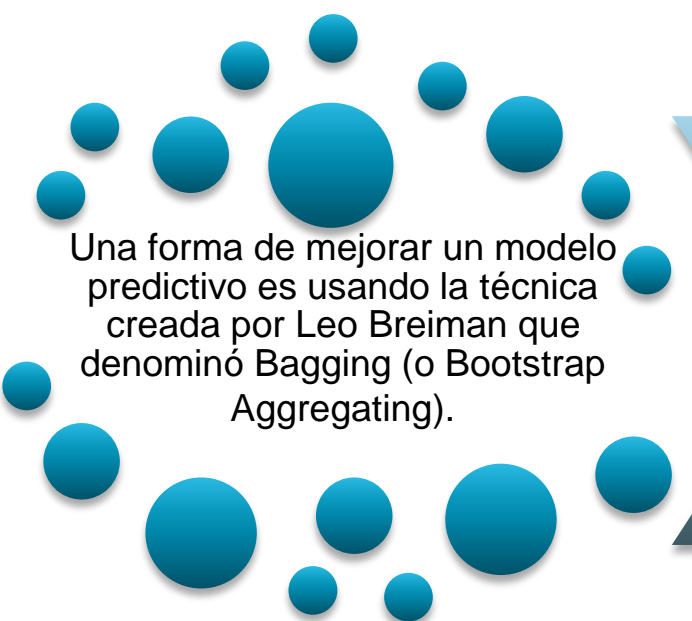




## VALIDACIÓN POR CONJUNTO DE PRUEBA



## BAGGING



Una forma de mejorar un modelo predictivo es usando la técnica creada por Leo Breiman que denominó Bagging (o Bootstrap Aggregating).

Esta técnica consiste en crear diferentes modelos usando muestras aleatorias con reemplazo y luego combinar o ensamblar los resultados.

Es un procedimiento general que se puede usar para reducir la varianza para aquellos algoritmos que tienen alta varianza. Un algoritmo que tiene alta varianza son árboles de decisión, como árboles de clasificación y regresión (CART).

## CONCEPTOS SOBRE BAGGING



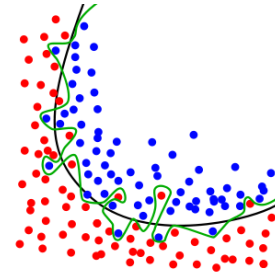
Disminuye la varianza de un data set al realizar remuestreo con reemplazo

01



Mientras más inestable es un modelo ((data set con mucha varianza), mejor será la predicción al usar Bagging.

02



Reduce overfitting o sobre entrenamiento de modelos (ningún modelo tiene todos los datos de entrenamiento)

03



outliers

Reduce el ruido de los outliers, ya los outliers no pueden estar presentes en todos los modelos.

04



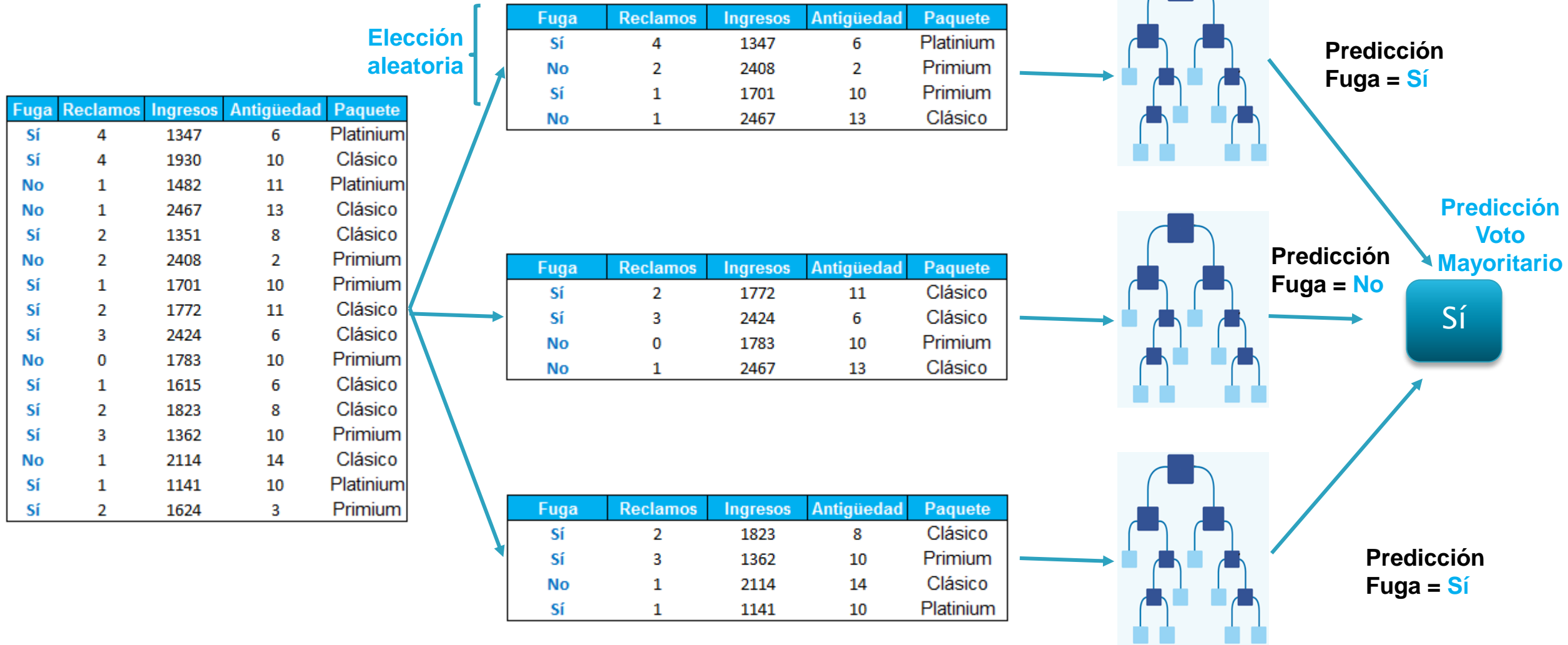
Mejora la predicción, ya que lo que no detecta un modelo lo detectan los otros.

05

## PASOS DEL BAGGIN

1. **Divide el set de Entrenamiento en distintos sub set de datos**, obteniendo como resultado diferentes muestras aleatorias con las siguientes características:
  - Muestra uniforme (misma cantidad de individuos en cada set)
  - Muestras con reemplazo (los individuos pueden repetirse en el mismo set de datos)
  - El tamaño de la muestra es igual al tamaño del set de entrenamiento, pero no contiene a todos los individuos ya que algunos se repiten.
  - Si se usan muestras sin reemplazo, suele elegirse el 50% de los datos como tamaño de muestra
2. **Luego se crea un modelo predictivo con cada set**, obteniendo modelos diferentes
3. **Luego se construye o ensambla un único modelo predictivo**, que es el promedio de todos los modelos.

## CLASIFICACIÓN Bagging



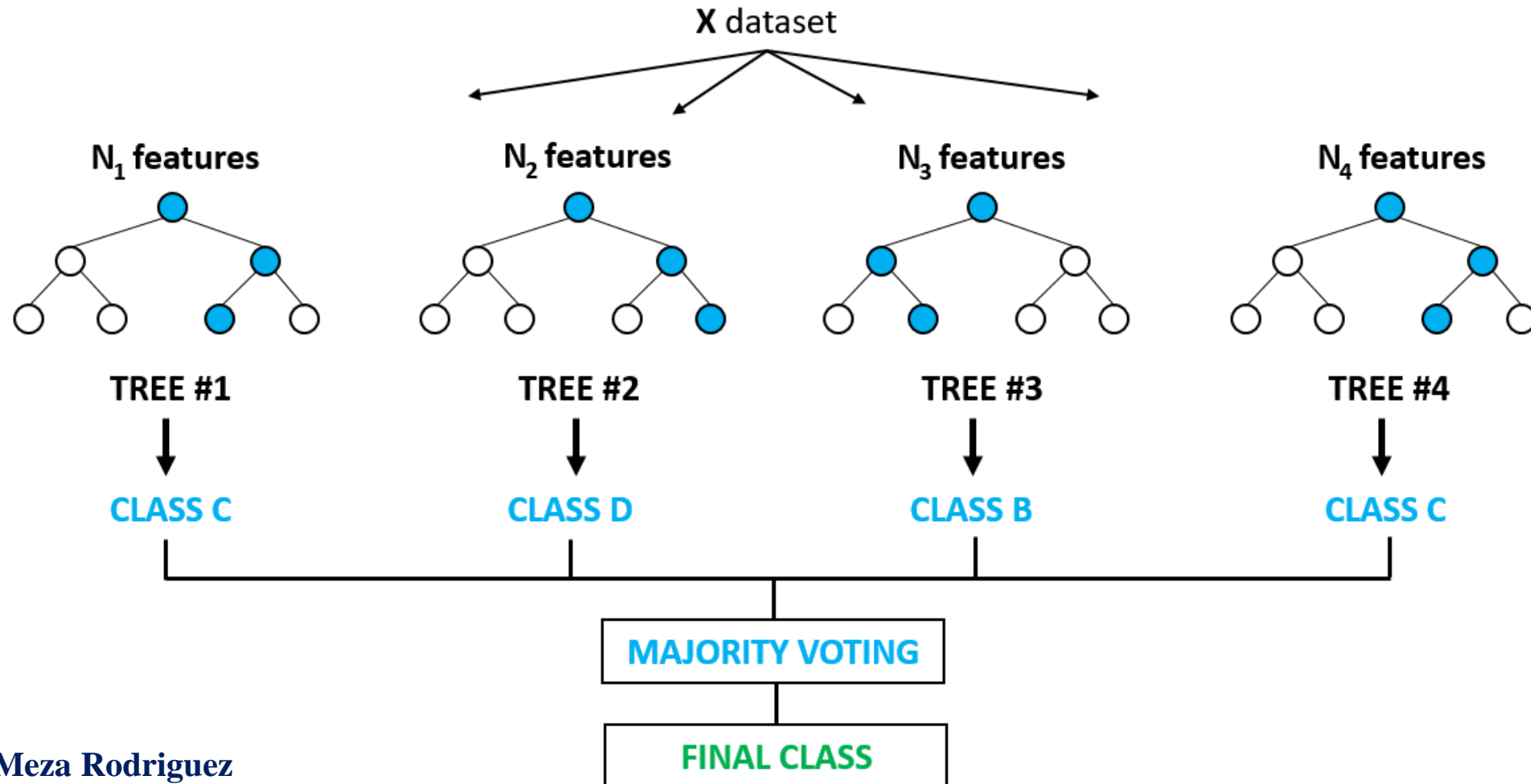
## RANDOM FOREST





## RANDOM FOREST

los parámetros en el bosque aleatorio se ajustan para aumentar el poder de predicción o para hacer que el modelo sea más rápido.



## RANDOM FOREST

En forma resumida sigue este proceso:

1. Selecciona individuos al azar (usando muestreo con reemplazo) para crear diferentes set de datos.
2. Crea un árbol de decisión con cada set de datos, obteniendo diferentes arboles, ya que cada set contiene diferentes individuos y diferentes variables en cada nodo.
3. Al crear los arboles se eligen variables al azar en cada nodo del árbol, dejando crecer el árbol en profundidad (es decir, sin podar).
4. Predice los nuevos datos usando el "voto mayoritario", donde clasificará como "positivo" si la mayoría de los arboles predicen la observación como positiva.



## RANDOM FOREST

Elección aleatoria de variables {

Elección aleatoria de individuos {

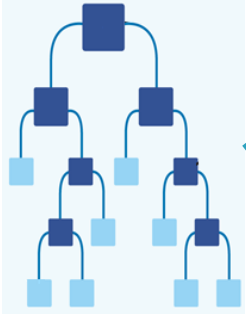
Fuga	Reclamos	Ingresos	Antigüedad	Paquete
Sí	4	1347	6	Platinum
Sí	4	1930	10	Clásico
No	1	1482	11	Platinum
No	1	2467	13	Clásico
Sí	2	1351	8	Clásico
No	2	2408	2	Primium
Sí	1	1701	10	Primium
Sí	2	1772	11	Clásico
Sí	3	2424	6	Clásico
No	0	1783	10	Primium
Sí	1	1615	6	Clásico
Sí	2	1823	8	Clásico
Sí	3	1362	10	Primium
No	1	2114	14	Clásico
Sí	1	1141	10	Platinum
Sí	2	1624	3	Primium

Fuga	Ingresos	Antigüedad	Paquete
Sí	1823	8	Clásico
Sí	1362	10	Primium
No	2114	14	Clásico
Sí	1141	10	Platinum

Fuga	Reclamos	Ingresos	Antigüedad
Sí	2	1823	8
No	2	2408	2
Sí	1	1701	10
Sí	1	1141	10

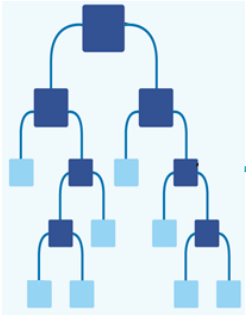
Fuga	Reclamos	Antigüedad	Paquete
Sí	4	6	Platinum
Sí	4	10	Clásico
No	1	13	Clásico
Sí	1	10	Platinum

Árbol sin podar



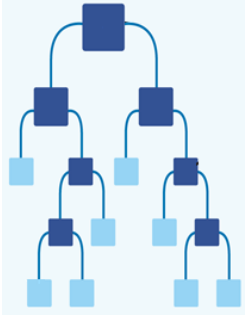
Predicción  
Fuga = Sí

Árbol sin podar



Predicción  
Fuga = No

Árbol sin podar



Predicción  
Fuga = Sí

Predicción  
Voto  
Mayoritario

Sí

## MEDICIONES DEL RANDOM FOREST



## RANDOM FOREST

### Ventajas de Random Forest

- 1.El algoritmo de Random Forest evita el sobreajuste
- 2.Para tareas de clasificación y regresión, se puede usar el mismo algoritmo de bosque aleatorio
- 3.El algoritmo Random Forest se puede utilizar para identificar las características más importantes del conjunto de datos de capacitación.

### Desventajas de Random Forest

- 1.Random Forest es difícil de interpretar. Debido a promediar los resultados de muchos árboles se vuelve difícil para nosotros descubrir por qué un bosque al azar está haciendo predicciones de la manera que es.
- 2.Random Forest tarda más tiempo en crear. Es computacionalmente costoso en comparación con un Árbol de decisiones.



# FIN

