Métodos Supervisados

Ing. Agustín Ullón Ramírez

CLASIFICACIÓN: DEFINICIÓN

- Dada una colección de registros (Conjunto de Entrenamiento) cada registro contiene un conjunto de variables (atributos) denominado x, con una variable (atributo) adicional que es la clase denominada y.

- El objetivo de la ***clasificación*** es encontrar un modelo (una función) para predecir la clase a la que pertenecería cada registro, esta asignación una clase se debe hacer con la mayor precisión posible.

- Un conjunto de prueba (tabla de testing) se utiliza para determinar la precisión del modelo. Por lo general, el conjunto de datos dado se divide en dos conjuntos al azar de el de entrenamiento y el de prueba.

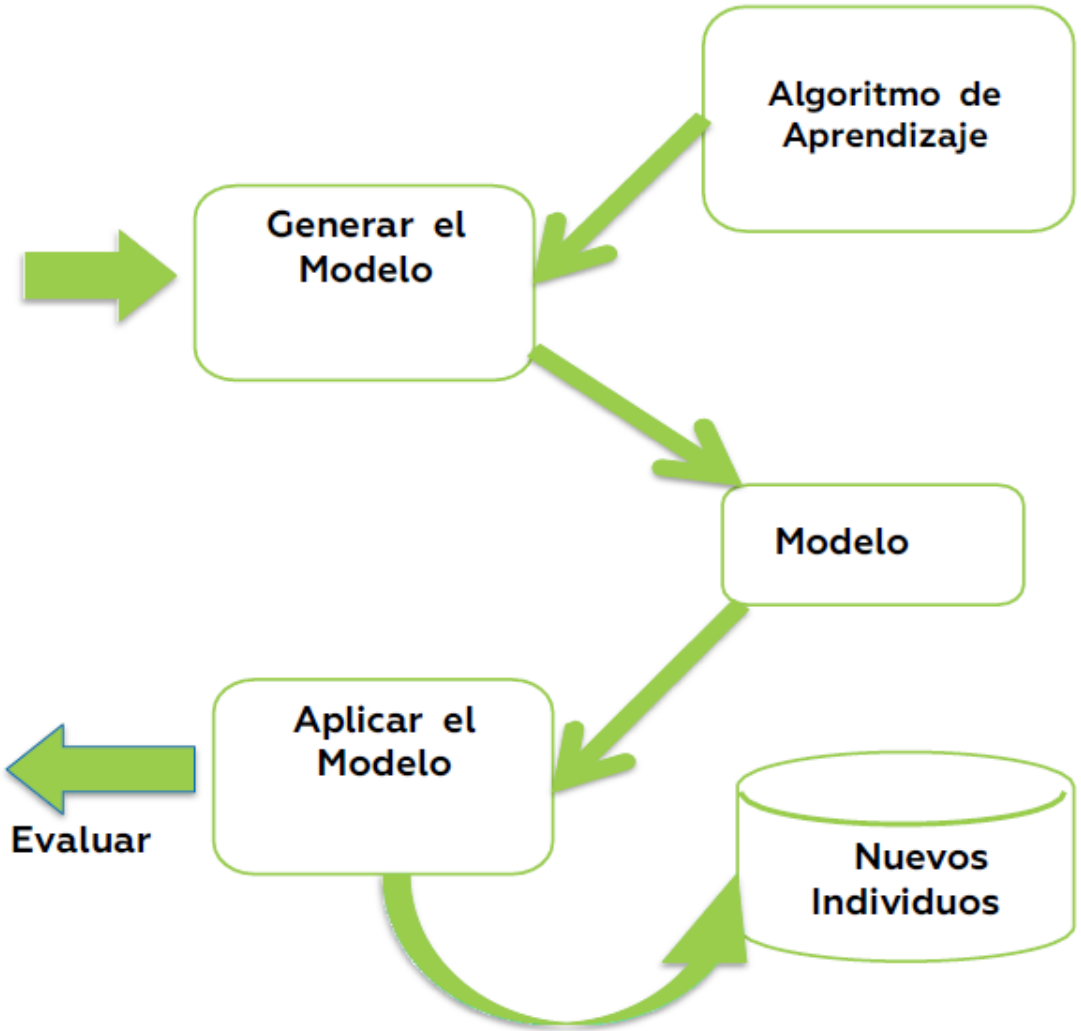
MODELO GENERAL DE LOS MÉTODOS DE CLASIFICACIÓN

ID	REEMBOLSO	ESTADO CIVIL	INGRESOS ANUALES	FRAUDE
1	SI	SOLTERO	S/ 1,000	NO
2	SI	CASADO	S/ 5,000	NO
3	NO	CASADO	S/ 3,500	SI
4	SI	VIUDO	S/ 4,500	NO
5	NO	SOLTERO	S/ 2,000	NO
6	NO	SOLTERO	S/ 1,500	SI

Tabla de Aprendizaje

ID	REEMBOLSO	ESTADO CIVIL	INGRESOS ANUALES	FRAUDE
7	SI	SOLTERO	S/ 4,000	NO
8	SI	CASADO	S/ 5,500	NO
9	NO	CASADO	S/ 6,500	SI

Tabla de Testing



Supervised Learning

- Género.

- Rangos de Edad.

- Ingresos.

- Estado Civil.

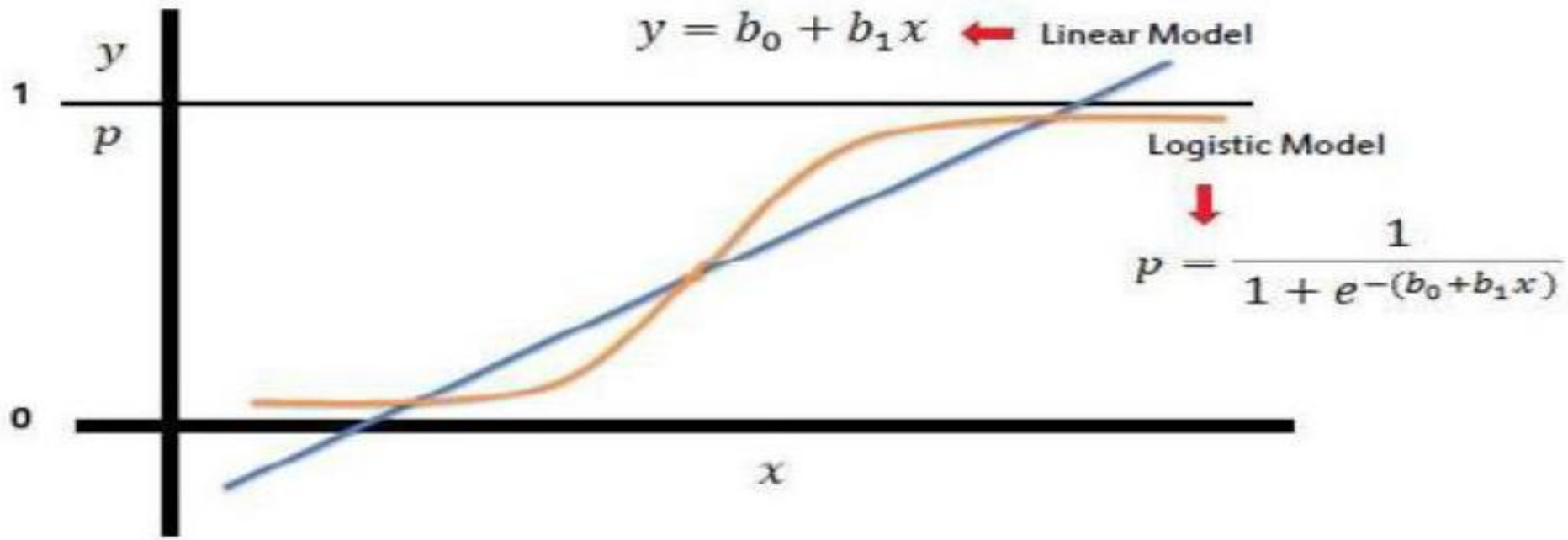
- Si Compra
- No Compra

Regresión Logística

DEFINICIÓN

- Es un modelo predictivo supervisado.
- La regresión logística es un modelo de elección discreta en el que la variable **dependiente es cualitativa binaria**.
- Es flexible en cuanto a la naturaleza de las **variables explicativas,** pues éstas pueden ser de **cuantitativas y categóricas**.
- Permite estudiar el **impacto que tiene cada una de las variables independientes** en la probabilidad de que ocurra el suceso de estudio.

Regresión Logística



MODELO DE REGRESIÓN LOGÍSTICA DICOTÓMICO

La variable Morosidad toma los siguientes valores:

"1" si el cliente es **moroso**.
"0" si el cliente es **no moroso**.

¿Es dicotómica?
¿Es cualitativa?
¿Es mutuamente excluyente?

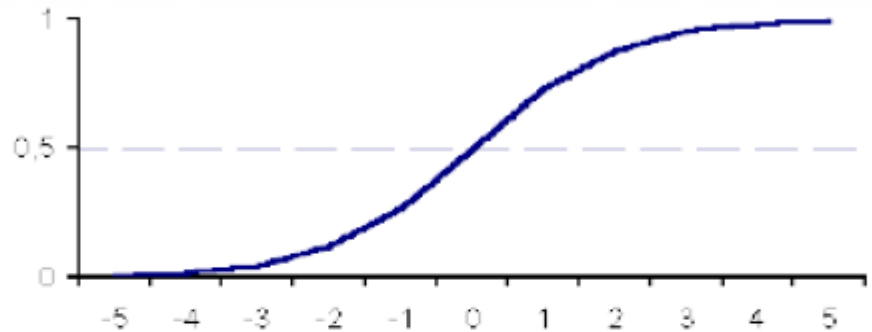
MODELO DE REGRESIÓN LOGÍSTICA DICOTÓMICO

$$p_i = f(\beta_0 + \beta_1 X_1)$$ Se calcula la función logit.

$$p_i = \frac{e^{(\beta_0+\beta_1 X_1)}}{1+e^{(\beta_0+\beta_1 X_1)}} = \frac{1}{1+e^{-(\beta_0+\beta_1 X_1)}}$$

$$1-p_i = \frac{1}{1+e^{(\beta_0+\beta_1 X_1)}}$$

MODELO DE REGRESIÓN LOGÍSTICA DICOTÓMICO



La representación matemática del modelo es la siguiente:

$$z_i = \log \frac{P_i}{1-P_i} = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \ldots + \beta_p x_{ip}$$

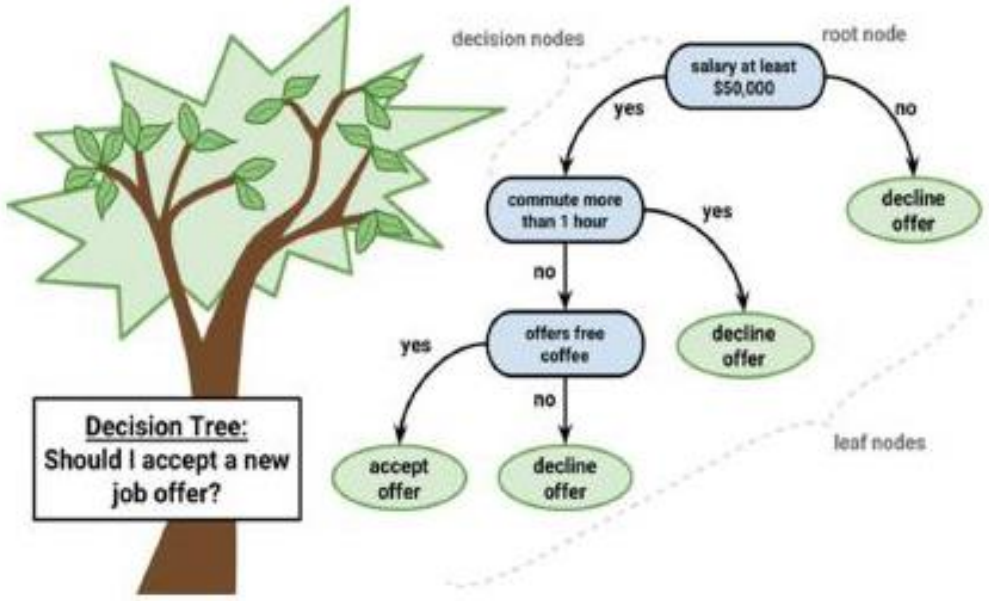
z_i : Variable dependiente del modelo: "Moroso" y " No Moroso"

p_i : Probabilidad de que el cliente sea "Moroso"

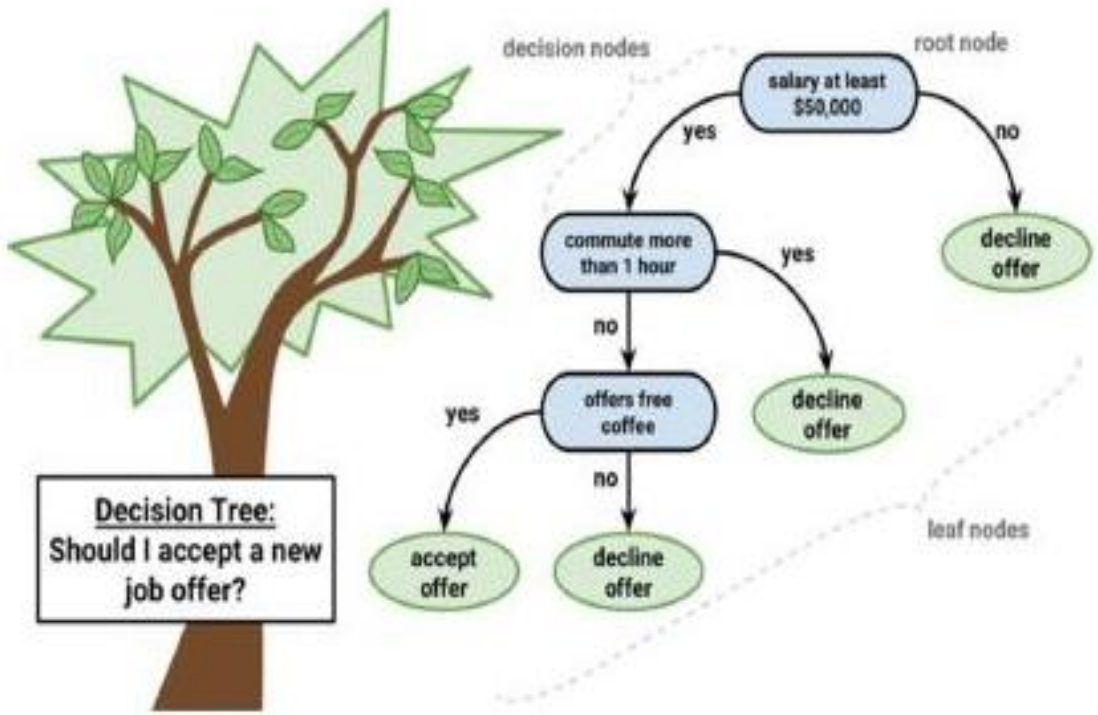
β_i : Coeficientes del modelo (parámetros a estimar)

x_i : Variables explicativas del modelo

Árboles de Decisión



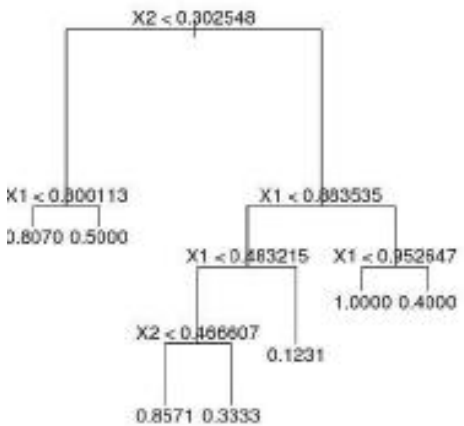
- Los métodos basados en árboles consisten en **segmentar el espacio** de predictores en varias regiones.
- Dentro de cada región, se utiliza la media o la moda de las observaciones de entrenamiento en esa región para hacer la predicción.
- Se dice que son "métodos basados en árboles" porque las reglas que se utilizan para dividir el espacio de predictores pueden ser representadas en forma de diagrama de árbol.
- El método más sencillo es el **árbol de decisión binario**.
- Los métodos de árboles se pueden utilizar **tanto en clasificación como en regresión**.

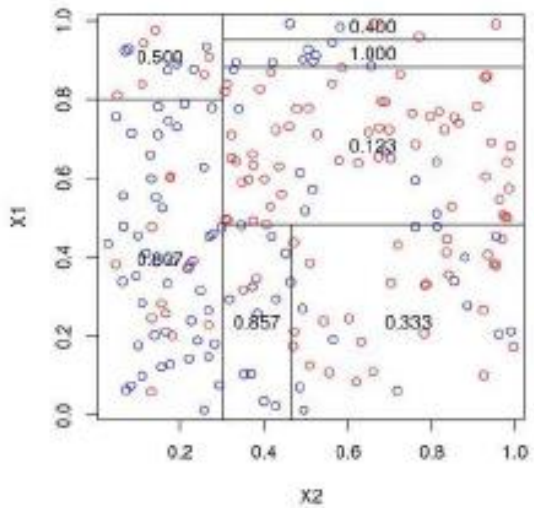


Modelos basados en árboles

Árbol de regresión

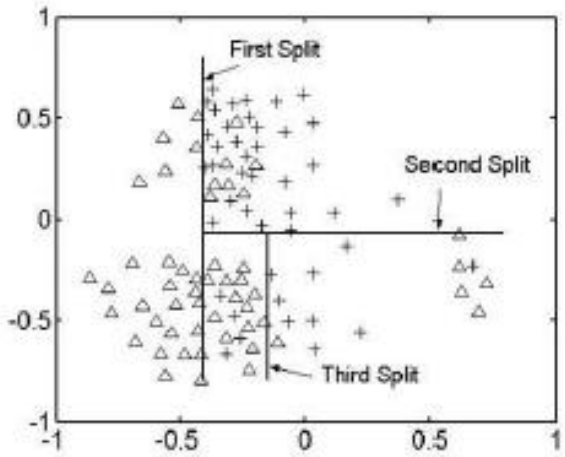
- Las hojas del árbol se corresponden con una **variable objetivo numérica** (regresión)
- La **salida** del modelo de árbol consiste en la **media** de todos los valores de cada hoja

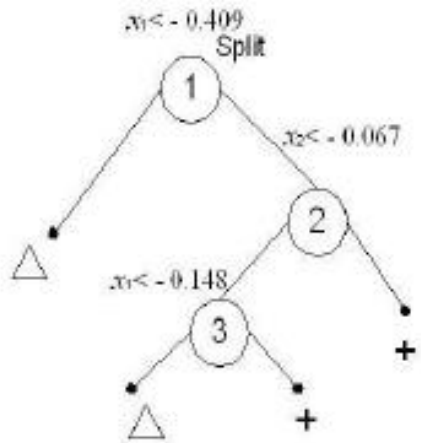




Árbol de clasificación

- Las hojas del árbol se corresponden con una **variable objetivo categórica** (clasificación)
- La **salida** del modelo de árbol consiste en la **proporción** de cada una de las clases de cada hoja



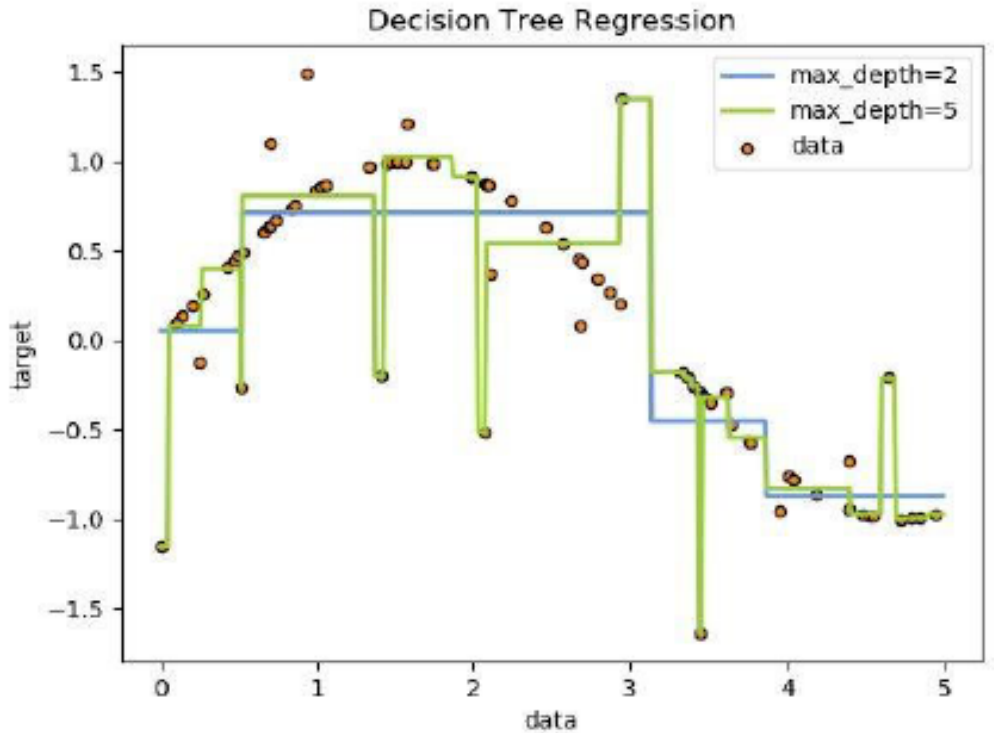


Árbol de decisión básico

Hiperparámetros de un árbol

En un modelo de árbol entran en juego numerosos hiperparámetros con los que podemos especificar su construcción. Los más habituales son los siguientes.

- **Profundidad del árbol:** Define cuántos niveles queremos en la construcción del árbol. Cuanto más alto sea este valor, más complejo será el árbol y más posibilidades de sobreajuste.
- **Número máximo de hojas:** Podemos limitar el número de hojas.
- **Número mínimo de muestras en cada hoja:** Limitar la cantidad mínima de registros en una hoja.



ÁRBOLES DE CLASIFICACIÓN

- **Entrada:**

Objetos caracterizables mediante propiedades.

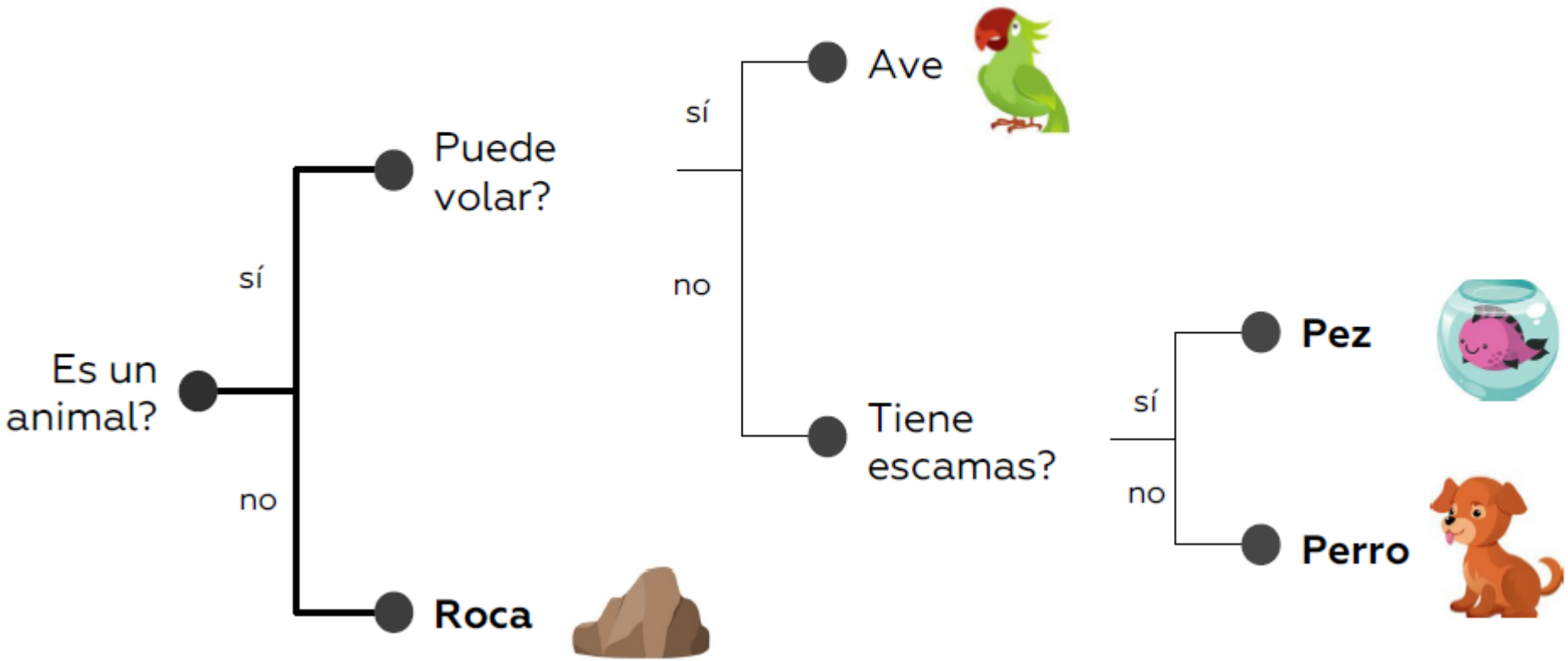
Variables o Features.

- **Salida:**

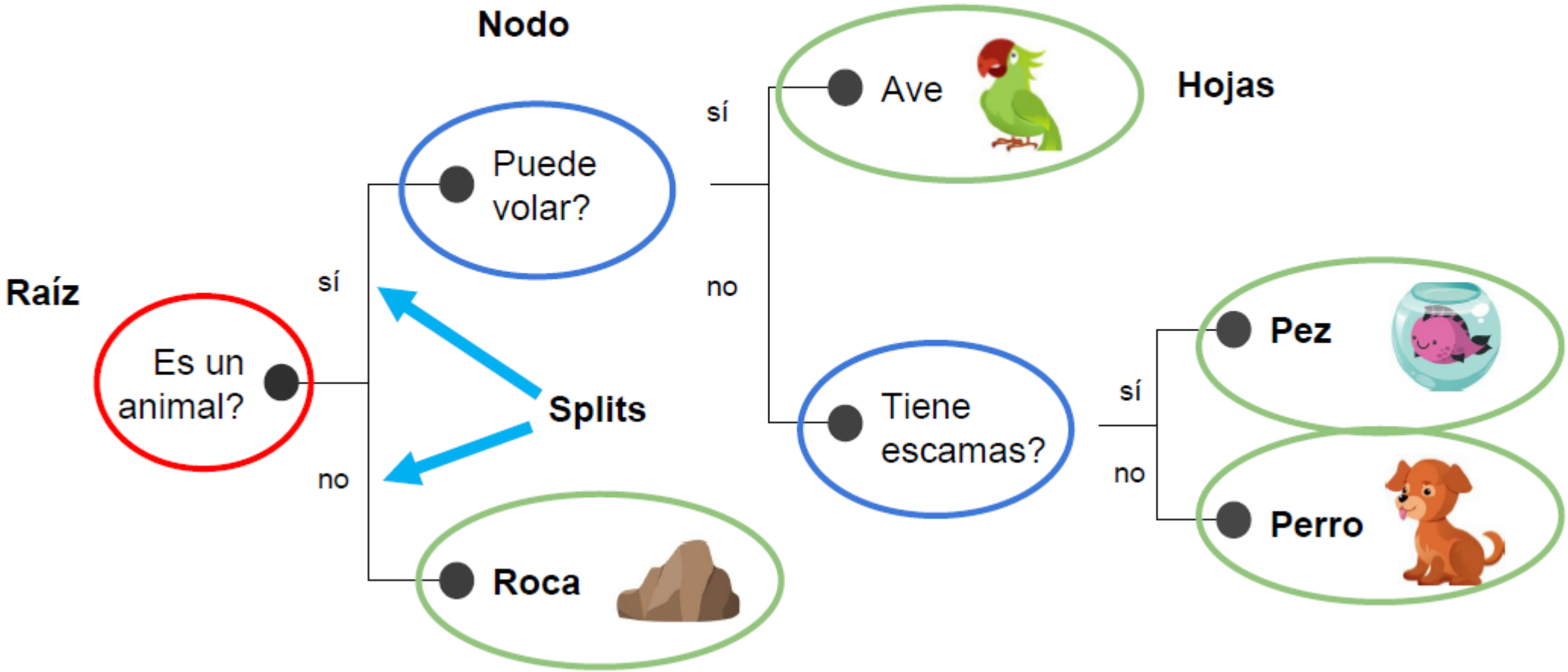
En árboles de clasificación: **una decisión** (sí o no)

Conjunto de **reglas**.

ÁRBOLES DE DECISIÓN

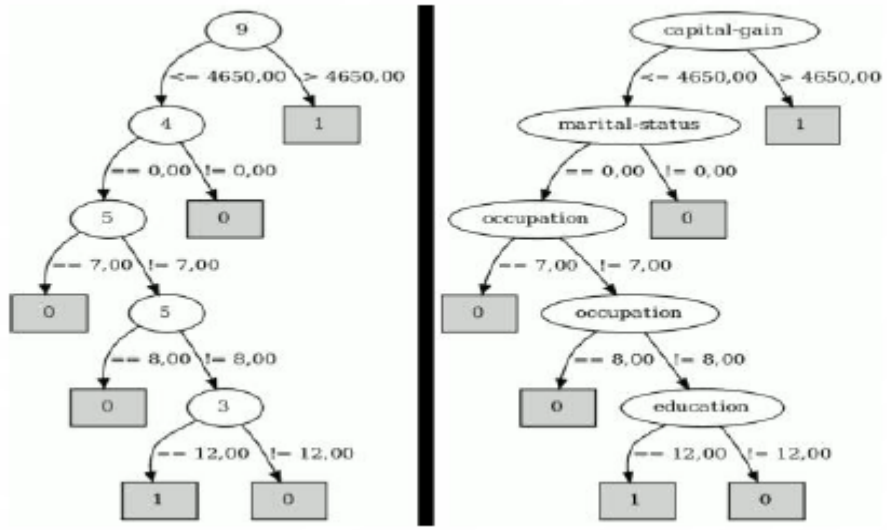


ÁRBOLES DE DECISIÓN



ALGORITMO: CART (CLASSIFICATION AND REGRESSION TREES)

- Árboles de **clasificación**: predicen categorías de objetos.
- Árboles de **regresión**: predicen valores continuos.
- Partición binaria recursiva.
- En cada iteración se selecciona la variable predictiva y el punto de separación que mejor reduzcan la 'impureza'.



Criterios de Partición

- Cada partición tiene asociada una medida de pureza.
- Se trata de incrementar la homogeneidad de los subconjuntos resultantes de la partición.
- Que sean más puros que el conjunto originario. Existen criterios de impureza tales como :

Medida de Entropía

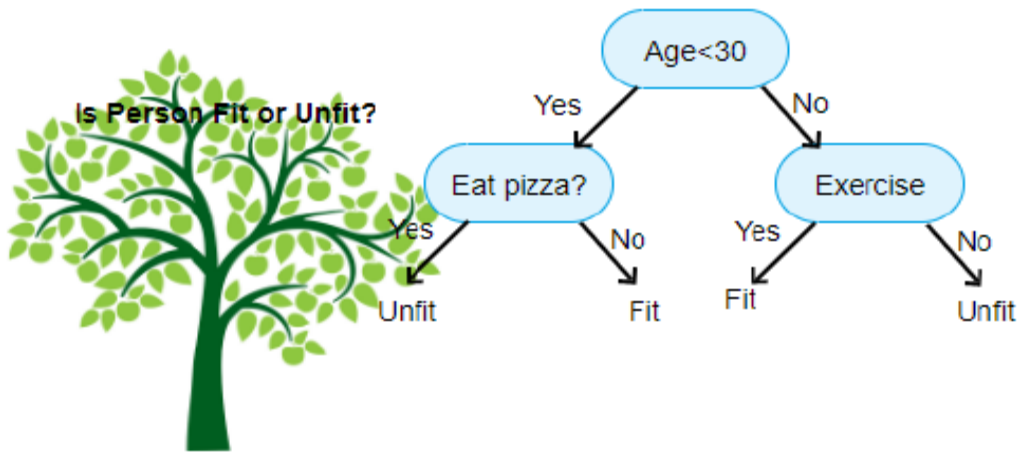
Índice de Gini

Criterio de Parada

- Un nodo se declarará terminal si el nodo es puro.
- Un nodo se declarará terminal si el nodo parental no tiene el mínimo establecido.
- Un nodo se declarará terminal si cualquier otra subdivisión no da una mejora mayor que la obtenida en el nodo padre.
- La división del nodo tiene como resultado un nodo hijo cuyo número de casos es menor que el tamaño mínimo preestablecido para un nodo hijo.
- La profundidad del árbol ha alcanzado su valor máximo preestablecido.

Poda de un Árbol

- En la primera fase, se construye un árbol que tenga cientos de nodos.
- En la segunda fase, el árbol es podado eliminando las ramas innecesarias hasta dar con el árbol adecuado.
- Este proceso compara simultáneamente todos los posibles subárboles resultado de podar en diferente grado el árbol original.



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

\\DROBO-FS\QuickDrops\JB\PPTX NG\Droplets\LightingOverlay.png

MUESTRA DE COMPROBACIÓN 30 %

MUESTRA DE VALIDACIÓN 20 %

- 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

Evaluación de modelos de clasificación

Clasificación binaria

Verdaderos positivos: número de elementos **positivos** clasificados como **positivos**.

TP

FN

Falsos negativos: número de elementos **positivos** clasificados como **negativos**.

Falsos positivos: número de elementos **negativos** clasificados como **positivos**.

FP

TN

Verdaderos negativos: número de elementos **negativos** clasificados como **negativos**.

Evaluación de modelos de clasificación

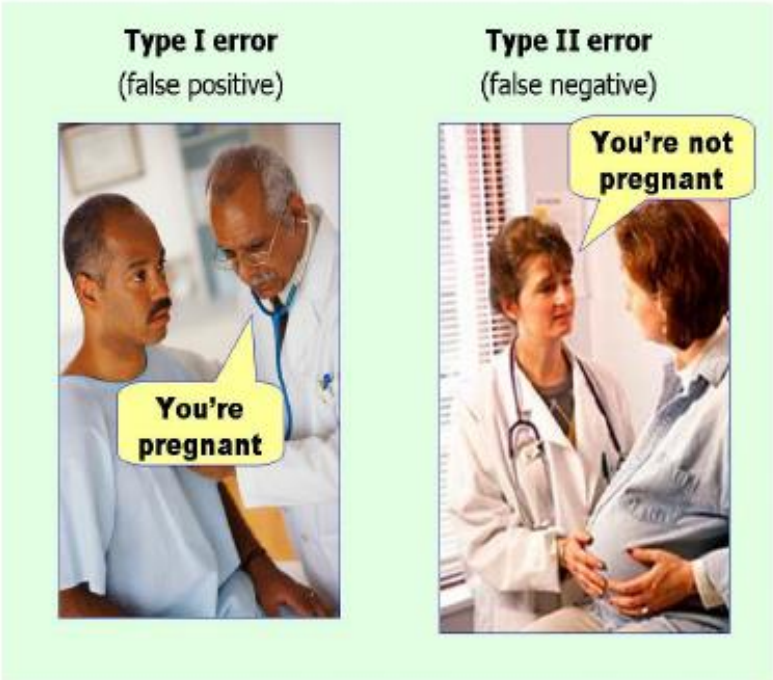
Matriz de Confusión

	Predicted 0	Predicted 1
Actual 0	TN	FP
Actual 1	FN	TP

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))

Evaluación de modelos de clasificación

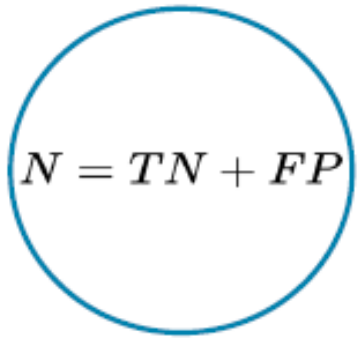
Métricas

Accuracy

Frecuencia de predicciones correctas

Positivos

Número total de casos positivos en la muestra



Negativos

Número total de casos negativos en la muestra

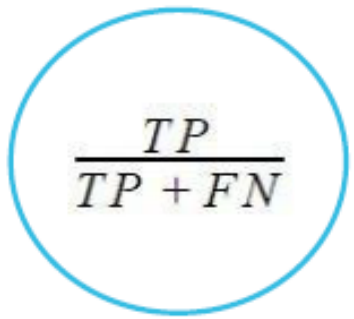
Evaluación de modelos de clasificación

Métricas

Especificidad (TNR)

Ratio de predicciones negativas acertadas sobre el total de las negativas reales

$$TNR = 1 - FPR$$



Recall (TPR)

También conocida como sensitivity.
Ratio de predicciones positivas acertadas, sobre el total de positivas reales

Precision

Ratio de predicciones positivas acertadas, sobre el total de positivas predichas

F1-score

Combina precisión y recall en una misma métrica como la media armónica de ambas

Evaluación de modelos de clasificación

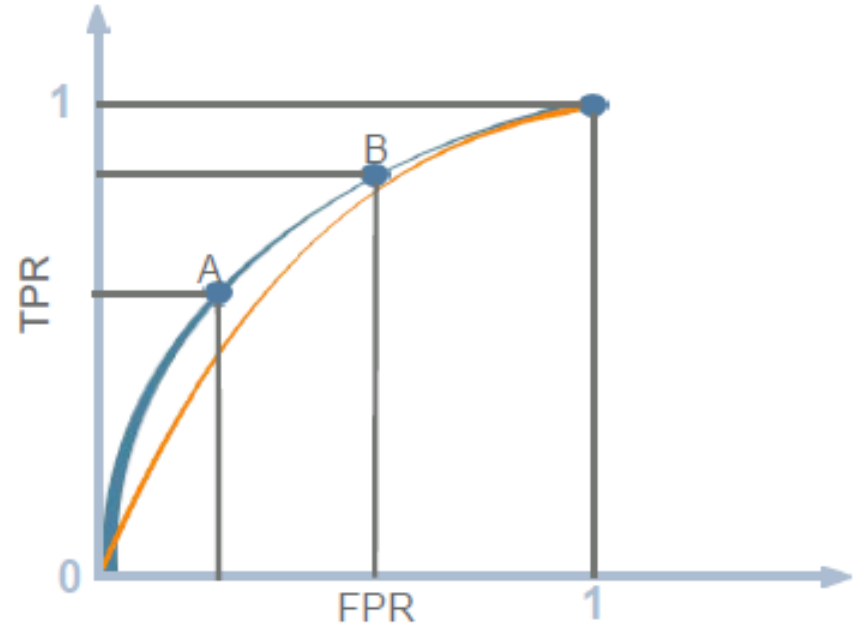
Clasificación binaria

La **curva ROC** (Receiver Operating Characteristic) constituye una de las herramientas más utilizadas para la evaluación y comparación de modelos de clasificación binaria.

Representa la tasa de falsos positivos frente a la tasa de verdaderos positivos (Recall) para **todos los posibles umbrales de decisión** en el conjunto de test

La curva ROC es una función, por tanto para interpretarla se emplea el **área bajo la curva** (AUC).

El **AUC ROC** representa la *probabilidad de que una muestra positiva se sitúe con un score superior al de una muestra negativa*. Cuanto más alto, mejor es el modelo.



False Positive Rate $FPR = \frac{FP}{P}$

False Negative Rate $FNR = \frac{FN}{N}$

True Positive Rate $TPR = 1 - FNR$

GRACIAS…