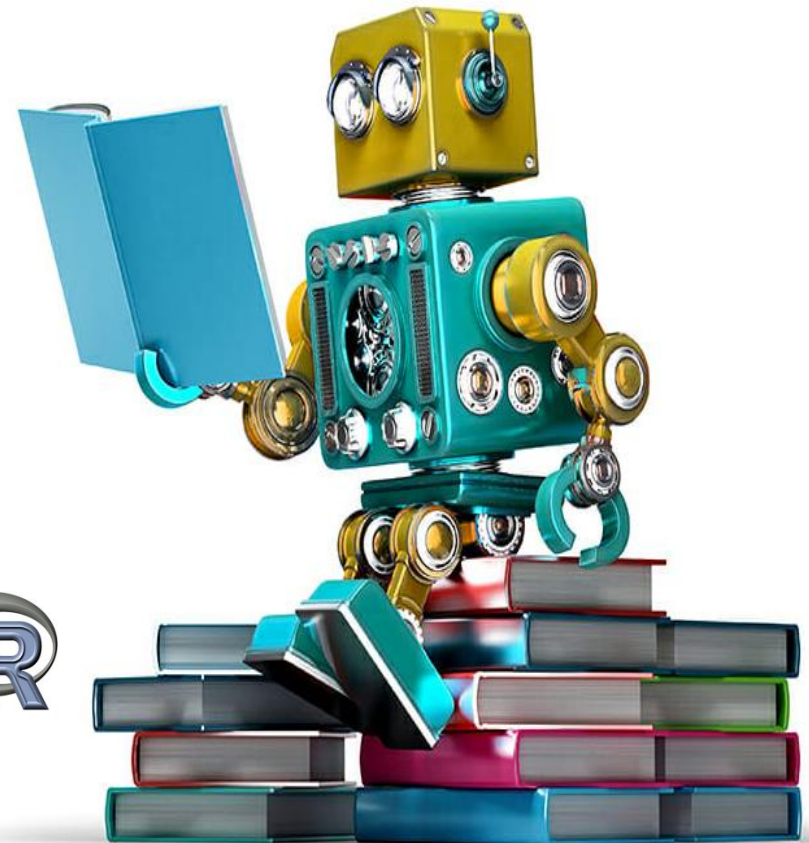
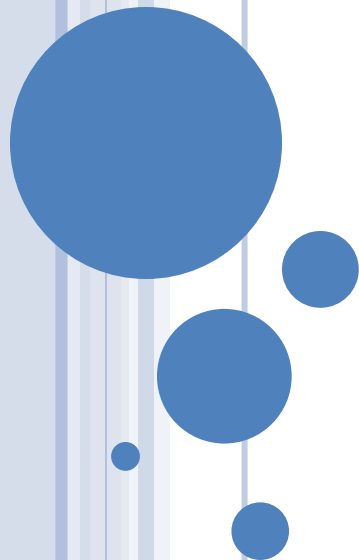




ALGORITMOS DE CLASIFICACIÓN : ÁRBOLES CHAID – CART – C50



«Lo llaman suerte, pero es constancia.
Lo llaman casualidad, pero es
disciplina. Lo llaman genética pero es
sacrificio. Ellos hablan , tú estudia.»»



SAN MARCOS DATA SCIENCE COMMUNITY

MENTORES



José Antonio Cárdenas Garro
ESTADÍSTICA
UNMSM

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



André Omar Chávez Panduro
ESTADÍSTICA
UNMSM

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



Predictive Modelling Specialist



Especialidades : Predictive Modeling | Machine Learning | Advanced Statistics | Risk Modeling | Data Science

Data Scientist



Especialidades : Predictive Modeling | Machine Learning | Advanced Statistics | Business Analytics | Data Science



Aldo Ray Chávez Panduro
INGENIERÍA DE SISTEMAS
UNMSM

Student of MSc in Data Science
Universidad Ricard Palma



Risk Management Specialist



Especialidades : Big Data | Machine Learning | Programming | Risk Specialist – IFRS9 | Data Science

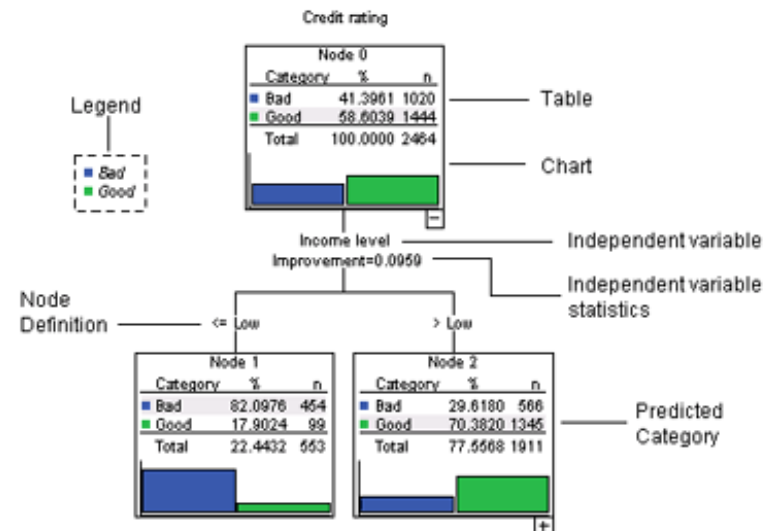
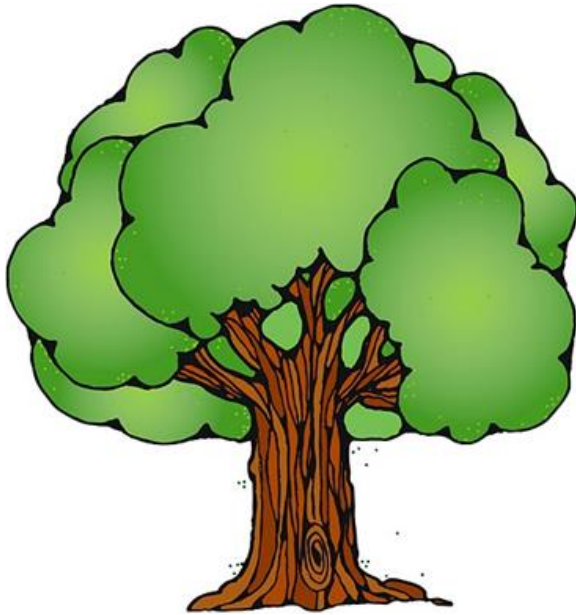


AGENDA

- Regresión **Logística Binaria**.
- Aplicaciones.
- Introducción a los Arboles de Clasificación.
- Árbol de Clasificación **CHAID**.
- Fase de Fusión, División y Reglas de parada.
- Árbol de Clasificación **CART**.
- Criterios de Partición y Reglas de parada.
- **Poda** de un Árbol.
- Árbol C50.
- Aplicaciones



ÁRBOLES DE CLASIFICACIÓN



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 CLASIFICACIÓN

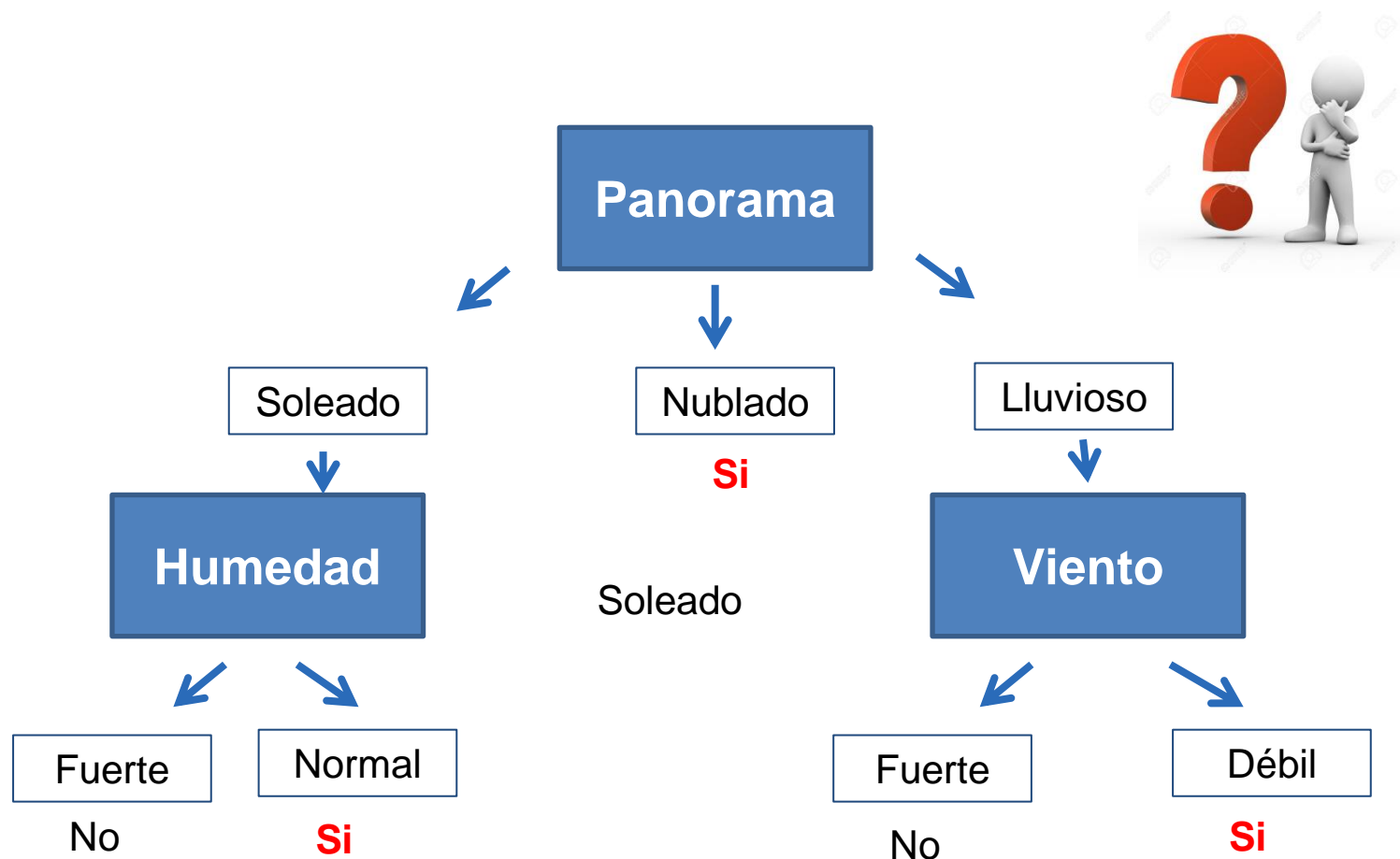
- Se clasifican las instancias desde la raíz (**Nodo padre**) hacia las hojas (**Nodos hijos**), las cuales proveen la clasificación.
- Cada nodo especifica el test (Composición de la VD) de algún atributo.
- Ejemplo: Si 
(Panorama= Soleado, Temperatura = Calurosa, Humedad = Alta, Viento= Fuerte)



Juego al tenis?

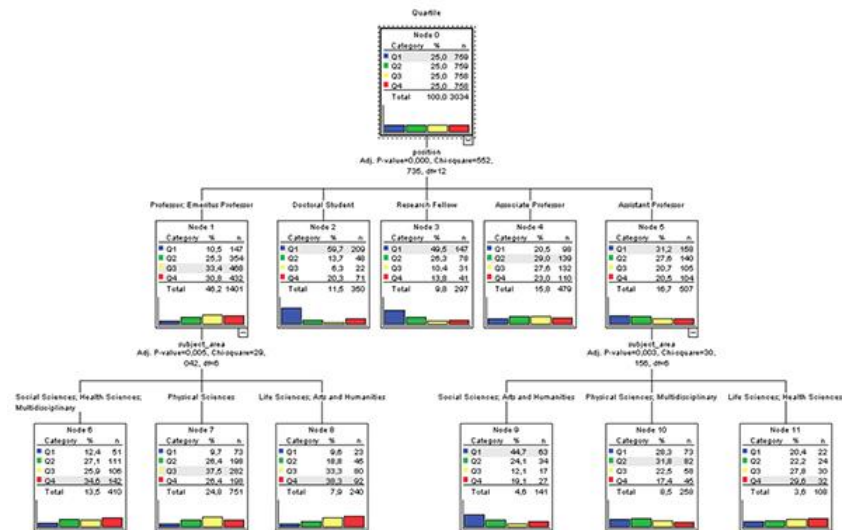


LA PREGUNTA A RESPONDER ES : JUEGO TENNIS ?



ALGORITMO DE ÁRBOL DE CLASIFICACIÓN CHAID

- Chi-Square Automatic Interaction Detector (Detector Automático de Interacciones mediante Chi-cuadrado).
- Kass,G.,1980. An exploratory technique for investigating large quantities of categorical data. Applied Statistics, 29:2, 199-127.



ALGORITMO DE ÁRBOL DE CLASIFICACIÓN CHAID

- Procede del ámbito de la Inteligencia artificial. Desarrollado por Kass a principios de los años 80.
- Asume que las variables explicativas son categóricas u ordinales. Cuando no lo son, se discretizan.
- Inicialmente se diseñó para el caso de variable respuesta Y categórica. Posteriormente se extendió a variables continuas.
- Utiliza contrastes de la χ^2 de Pearson y la F de Snedecor.
- El corte en cada nodo es multi-vía.



FUNCIONAMIENTO : PRUEBA CHI- CUADRADO

REGIÓN	SITUACIÓN CREDITICIA		TOTAL
	MOROSO	NO MOROSO	
NORTE	40	60	100
CENTRO	30	70	100
SUR	50	50	100
TOTAL	120	180	100

- H0: La situación crediticia es independiente de la región.
- H1: La situación crediticia es dependiente de la región.



χ^2 elevado , p – value (Sig) muy pequeño.



FUNCIONAMIENTO : ¿SI TENGO 2 VARIABLES INDEPENDIENTES O FEATURES, CUÁL ES MÁS IMPORTANTE?

IDEA INTUITIVA : FASE SPLIT

GÉNERO	SITUACIÓN CREDITICIA		TOTAL
	MOROSO	NO MOROSO	
MASCULINO	40	60	100
FEMENINO	30	70	100
TOTAL	70	130	200
TOTAL %	35%	65%	100%

GÉNERO	SITUACIÓN CREDITICIA		TOTAL
	MOROSO	NO MOROSO	
JÓVENES	65	35	100
ADULTOS	5	95	100
TOTAL	70	130	200
TOTAL %	35%	65%	100%

- ¿ El género o la categoría de edad discrimina mejor la situación crediticia ?



FUNCIONAMIENTO : ¿SI TENGO UNA VARIABLE INDEPENDIENTE O FEATURE CON MÁS DE UNA CATEGORÍA, TODAS LAS CATEGORÍAS SERÁN IGUALMENTE IMPORTANTES?

IDEA INTUITIVA : FASE MERGE

VARIABLE	SITUACIÓN CREDITICIA		TOTAL
	MOROSO	NO MOROSO	
A	20	80	100
B	25	75	100
C	60	40	100
D	65	35	100
TOTAL	170	230	400
TOTAL %	43%	57%	100%

VARIABLE	SITUACIÓN CREDITICIA		TOTAL
	MOROSO	NO MOROSO	
A - B	45	155	200
C - D	125	75	200
TOTAL	170	230	400
TOTAL %	43%	57%	100%



¿ Cuando paramos de fusionar?



CARACTERÍSTICAS

- Es el algoritmo de árbol de clasificación más conocido.
- No es binario, es decir se pueden generar más de 2 categorías en cualquier nivel del árbol.
- Tiende a crear un árbol más ancho que los métodos de desarrollo binario.
- Aprovecha los valores perdidos, tratándolos como una categoría válida individual.



ALGORITMO

- Las categorías de cada predictor (variable independiente) se funden si no son significativamente distintos respecto a la variable dependiente. **FASE DE FUSIÓN O MERGE.**
- En cada paso, se elige la variable independiente (predictora) que presenta la interacción más fuerte con la variable dependiente. **FASE DE DIVISIÓN O SPLIT.**
- El proceso se repite hasta que se cumplan las reglas de parada establecidas.



EJEMPLO : FASE DE FUSIÓN

TARGET	CATEGORÍAS		TOTAL
	A	B	
COMPRA	40%	50%	35%
NO COMPRA	60%	50%	65%
TOTAL	100%	100%	100%

	Valor	gl	Sig. (p-value)
Chi-Cuadrado	42,56	1	0.0000

TARGET	CATEGORÍAS		TOTAL
	A	C	
COMPRA	33%	12%	20%
NO COMPRA	67%	88%	80%
TOTAL	100%	100%	100%

	Valor	gl	Sig. (p-value)
Chi-Cuadrado	16,74	1	0.0000

TARGET	CATEGORÍAS		TOTAL
	B	C	
COMPRA	25%	20%	18%
NO COMPRA	75%	80%	72%
TOTAL	100%	100%	100%

	Valor	gl	Sig. (p-value)
Chi-Cuadrado	1,54	1	0,1200

α fusión = 0,05 = α merge

- H0: La compra del producto es independiente de las categorías.
- H1: La compra del producto es dependiente de las categorías.



EJEMPLO : FASE DE FUSIÓN

- Si se ha fusionado un par de categorías, se procede a realizar nuevas fusiones de los valores del pronosticador.
- El proceso se acaba cuando no se pueden realizar más fusiones porque los χ^2 ofrecen resultados significativos.

TARGET	CATEGORÍAS		
	A	B - C	TOTAL
COMPRA	35%	20%	25%
NO COMPRA	65%	80%	75%
TOTAL	100%	100%	100%

	Valor	gl	Sig. (p-value)
Chi-Cuadrado	48236.00	1	0,00000

- H0: La compra del producto es independiente de las categorías A y B-C.
- H1: La compra del producto es dependiente de las categorías A y B-C.



EJEMPLO : FASE DE FUSIÓN

Fase merge

Paso 1. Encontrar el emparejamiento de categorías que conducen al mayor p-valor $-p^*$ para el test de la χ^2 o F

Paso 2. Comparar p^* con el umbral establecido α_{merge}

- Si $p^* > \alpha_{merge}$ agrupar las dos categorías en una sola. Volver a paso 1
- Si $p^* < \alpha_{merge}$ ir a paso 3

Paso 3. Ajustar el p-valor utilizando el multiplicador de Bonferroni:

$$p_{adj} = p^* \cdot B, \text{ siendo } B = \sum_{i=0}^{r-1} (-1)^i \frac{(r-i)^c}{i!(r-i)!}, \text{ } c \text{ el número original}$$

de categorías y r el número de categorías tras el agrupamiento



EJEMPLO : FASE DE DIVISIÓN

- Primera segmentación. Selección de la variable que mejor prediga la variable dependiente.

GÉNERO			
TARGET	MASCULINO	FEMENINO	TOTAL
COMPRA	40%	50%	35%
NO COMPRA	60%	50%	65%
TOTAL	100%	100%	100%

	Valor	gl	Sig. (p-value)
Chi-Cuadrado	23,78	1	0.0000

INGRESO			
TARGET	<5000	>=5000	TOTAL
COMPRA	33%	12%	20%
NO COMPRA	67%	88%	80%
TOTAL	100%	100%	100%

	Valor	gl	Sig. (p-value)
Chi-Cuadrado	12,06	1	0,00345

α split = 0,05 = α división

- H0: La compra del producto es independiente de la variable independiente.
- H1: La compra del producto es dependiente de la variable independiente.



EJEMPLO : FASE DE DIVISIÓN

Fase split

Paso 1. Encontrar la variable predictora con el menor p-valor ajustado p^\dagger

Paso 2. Comparar p^\dagger con el umbral establecido α_{split}

- Si $p^\dagger < \alpha_{split}$ particionar el nodo utilizando el agrupamiento de categorías obtenido en la fase merge
- Si $p^\dagger > \alpha_{split}$ declarar el nodo terminal



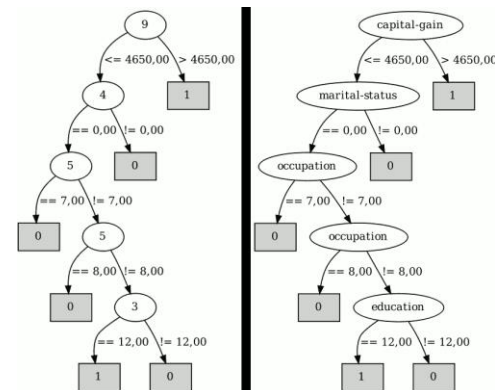
REGLAS DE PARADA

- Todos los casos en un nodo tengan valores idénticos en todos los predictores.
- El nodo se vuelve puro; esto es todos sus casos tienen el mismo valor en la variable criterio.
- La **profundidad del árbol** ha alcanzado su valor máximo preestablecido.
- El número de casos que constituyen el nodo es menor que el tamaño mínimo preestablecido para un nodo parental.
- 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.



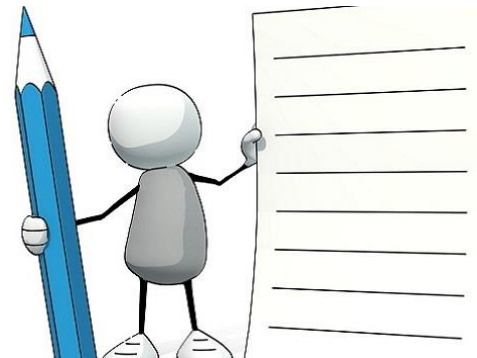
ALGORITMO: **CART** (CLASSIFICATION AND REGRESSION TREES)

- 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'.



CARACTERÍSTICAS PRINCIPALES

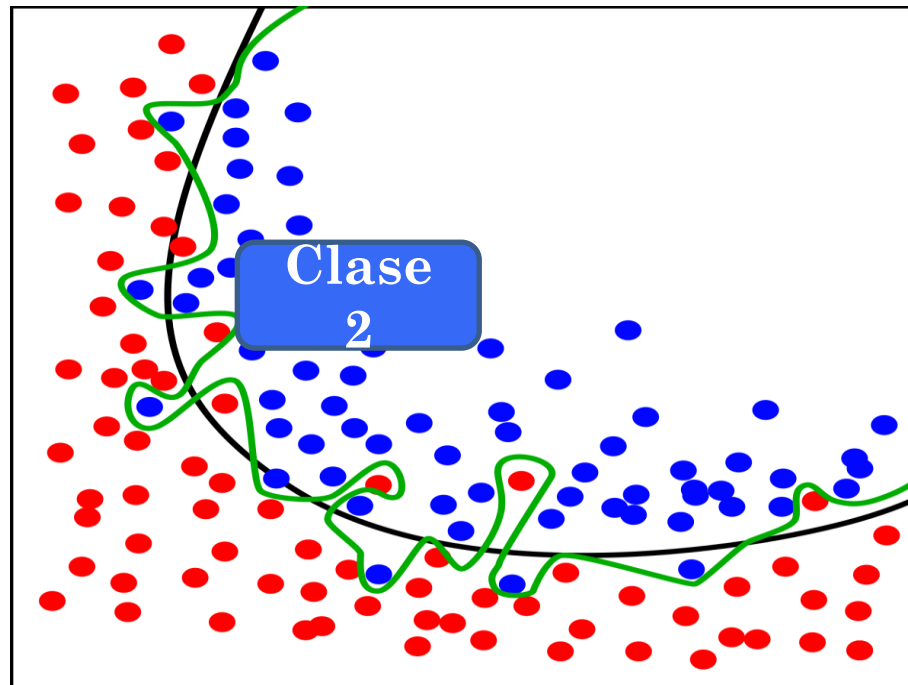
- Uno de los métodos de aprendizaje **supervisado no paramétrico** más utilizado.
- Realizar **sucesivas divisiones binarias** en un conjunto de datos guiado por un criterio.
- Para cada nodo selecciona a la variable independiente que proporciona el mejor desempeño en el criterio para particionar los datos.



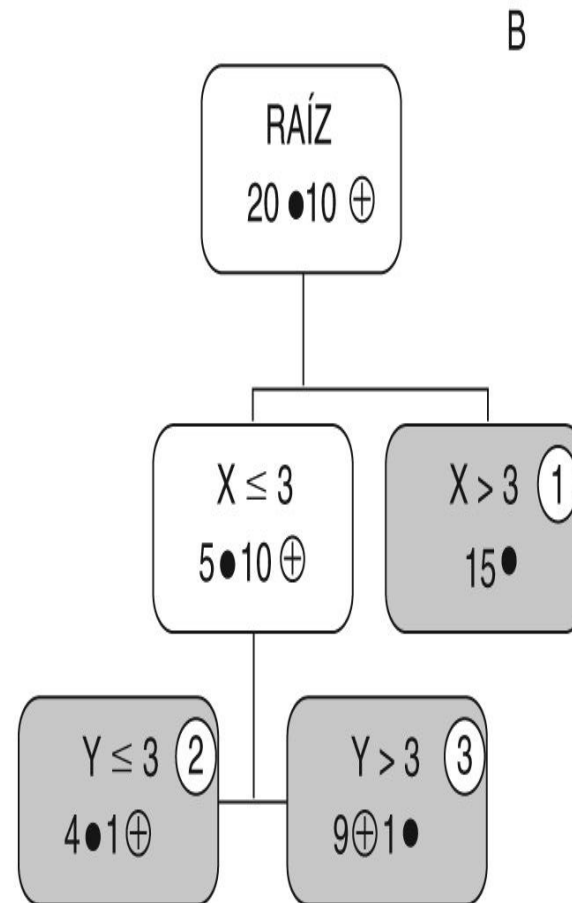
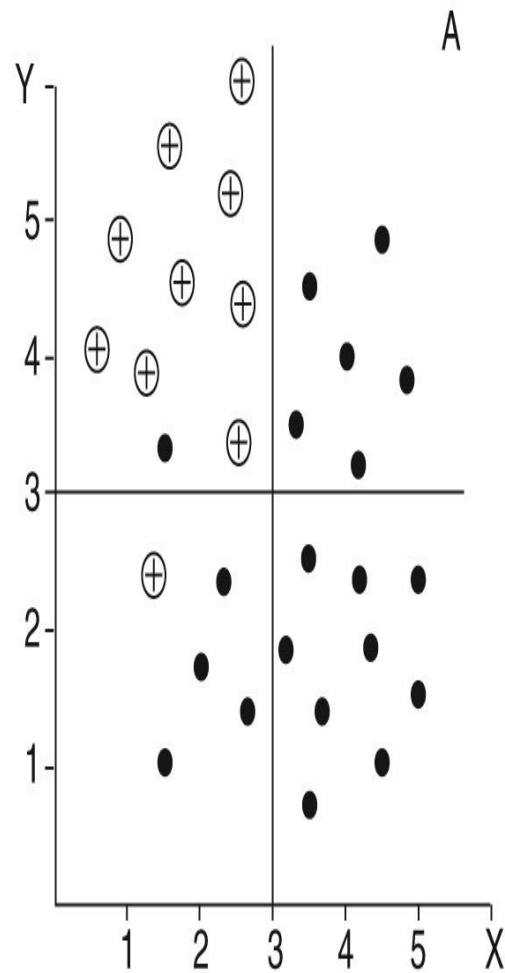
INTRODUCCIÓN

- Imaginemos que tenemos 2 variables predictoras X e Y .
- Se tienen 160 observaciones divididas en 2 clases.

**Clase
1**

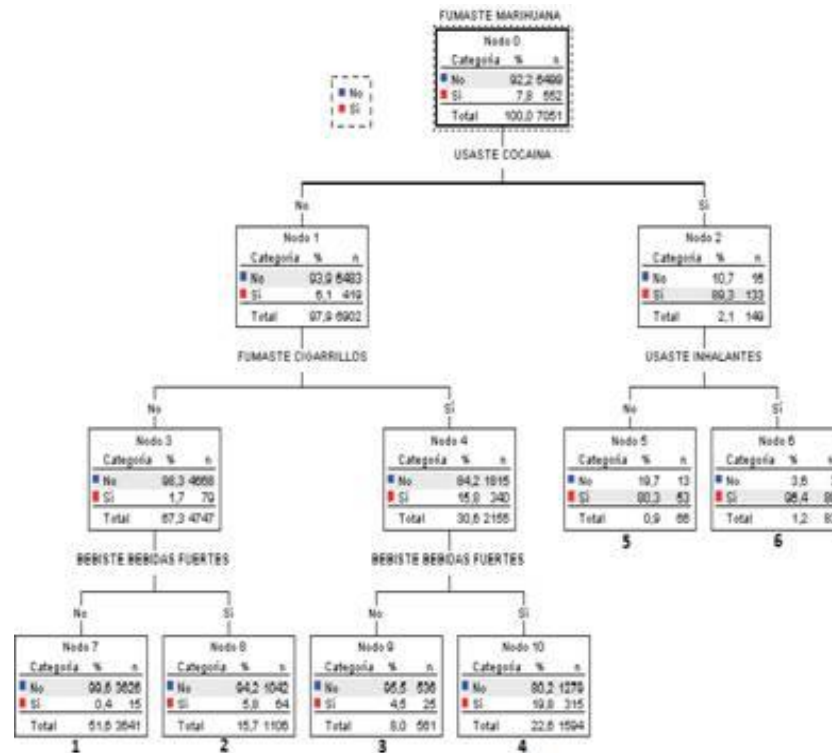


INTRODUCCIÓN : IDEA INTUITIVA



INTRODUCCIÓN : IDEA INTUITIVA

- ¿Qué representa un árbol de clasificación?
- Un árbol de clasificación representa una participación recursiva realizada en base a un conjunto de individuos.



LÓGICA DE UN ÁRBOL CART

- El método consiste en buscar la variable independiente X_i que mejor explique la variable dependiente Y_i .
- Esta variable define la primera división de la muestra en dos subconjuntos llamados segmentos.
- Después se reitera el procedimiento en cada uno de los segmentos buscando la segunda mejor variable y así sucesivamente hasta que se tengan **subconjuntos con elementos de una sola clase** o se cumpla alguna regla de parada.
- Éste método no considera simultáneamente al conjunto de variables explicativas, sino que las examina una a una.
- Sin embargo, la relación entre las variables explicativas se tiene en cuenta en las diferentes etapas del árbol.



PUNTOS CLAVES EN LA CONSTRUCCIÓN DE UN ÁRBOL DE CLASIFICACIÓN CART

1.¿ De qué forma se hacen las particiones y se selecciona la mejor de entre las posibles en cada momento?

¿ Cómo se formulan las preguntas?

¿Qué partición es la mejor?

2.¿Cuál es el criterio para determinar que un nodo es homogéneo? O ¿Cuándo se debe declarar un nodo como terminal o por el contrario, continuar su división?.

3.¿Cómo asignar una etiqueta a un nodo terminal?



SELECCIÓN DE LAS PARTICIONES

- Una partición divide un conjunto de individuos en conjuntos disjuntos.
- En CART las particiones son binarias.
- Objetivo de una partición: Incrementar la **homogeneidad** (en términos de clase) de los subconjuntos resultantes. Que sean más **puros** que el conjunto originario.



FORMULACIÓN DE LAS PREGUNTAS: VARIABLES INDEPENDIENTES O FEATURES CUALITATIVOS

1. Cada partición depende de un único atributo.
2. Si X_i es un atributo categórico, que toma valores en {Soltero, Casado, Divorciado} se incluyen las preguntas:

¿ $X_i = \text{Soltero}$?

donde cada valor es una categoría de entre los valores posibles que puede tomar la variable.

Por ejemplo. Si X_2 toma valores en { A , B , C } , ¿ $X_2=\{A\}$? , ¿ $X_2=\{B\}$? , ¿ $X_2=\{C\}$?.



FORMULACIÓN DE LAS PREGUNTAS: VARIABLES INDEPENDIENTES O FEATURES CUANTITATIVOS

3. Si X_i es un atributo continuo, se incluyen las preguntas:

$$¿ X_i \leq v ?$$

donde v es valor real, teóricamente cualquiera . En CART , v es el punto medio de los valores consecutivos de X_i .

Por ejemplo . Si X_1 es real, con valores 0.1 , 0.5 , 1.0

$$¿ X_1 \leq (0.1 + 0.5) / 2 ? , ¿ X_1 \leq (0.5 + 1.0) / 2 ?$$



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**



ÍNDICE DE DIVERSIDAD DE GINI

$$G(A_i) = \sum_{j=1}^{M_i} p(A_{ij}) G(C|A_{ij})$$
$$G(C|A_{ij}) = - \sum_{k=1}^J p(C_k|A_{ij}) p(\neg C_k|A_{ij}) =$$
$$= 1 - \sum_{k=1}^J p^2(C_k|A_{ij})$$

- A_i es el atributo para ramificar el árbol.
- M_i es el número de valores diferentes del atributo A_i .
- $p(A_{ij})$ es la probabilidad de que A_i tome su j -ésimo valor ($1 \leq j \leq M_i$).
- $p(C_k|A_{ij})$ es la probabilidad de que un ejemplo pertenezca a la clase C_k cuando su atributo A_i toma su j -ésimo valor.



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.



ALGORITMO DE ÁRBOL DE CLASIFICACIÓN C50

- C5.0 es el algoritmo sucesor de C4.5, ambos publicados por Quinlan, con el objetivo de crear árboles de clasificación.
- Entre sus características, destacan la capacidad para generar árboles de predicción simples, modelos basados en reglas, *ensembles* basados en ***boosting*** y asignación de distintos **pesos a los errores**.



HISTORIA

- A fines de la década de 1970, Ross Quinlan estaba desarrollando modelos basados en árboles como ID3.
- En la década de 1980 estos métodos evolucionaron en un modelo de árbol de clasificación llamado C4.5 (Quinlan, 1993).
- Aunque Quinlan publicó muy poco sobre este modelo después de su libro, él fue quien desarrolló continuamente el árbol de clasificación y los modelos basados en reglas, su última publicación llamada C5.0
- Kuhn y Johnson (2013) tienen una descripción más completa de C5.0 y otro modelo inédito llamado **Cubist**.



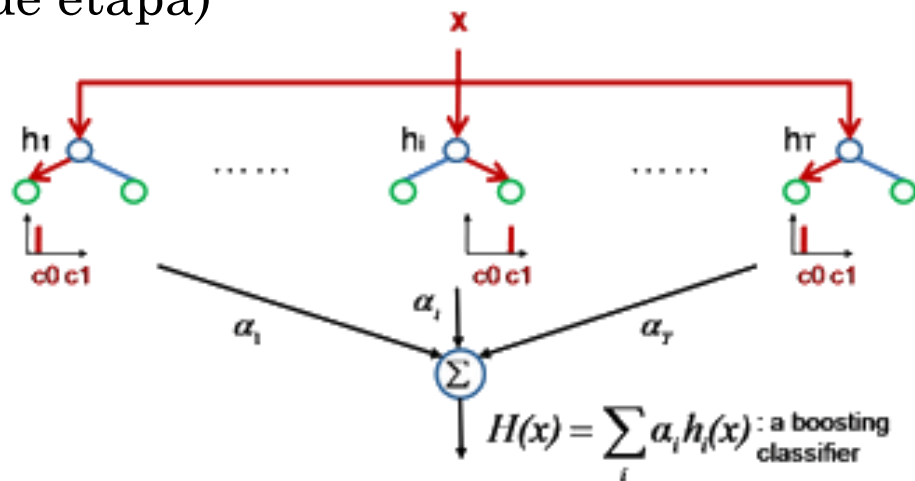
ALGORITMO C50 VS CART

- Se usa como medida de impureza la **entropía**.
- La poda de los árboles se realiza mediante «poda pessimistic».
- Los valores faltantes se manejan mediante el envío de muestras fraccionadas a nodos subsecuentes.
- Las principales diferencias entre los dos algoritmos son: **boosting**, **winnowing** y **costos asimétricos** para errores específicos.



BOOSTING

- El algoritmo C5.0 hace algo similar a **adaboost**, es decir después de que se crea el primer árbol, se determinan los pesos y subsecuentes iteraciones crean árboles o conjuntos de reglas ponderados.
- Los árboles posteriores (o conjuntos de reglas) están limitados a tener aproximadamente el mismo tamaño que el modelo inicial.
- La predicción final es un promedio simple de las probabilidades de clase generadas de cada árbol o conjunto de reglas (es decir, sin pesos de etapa)



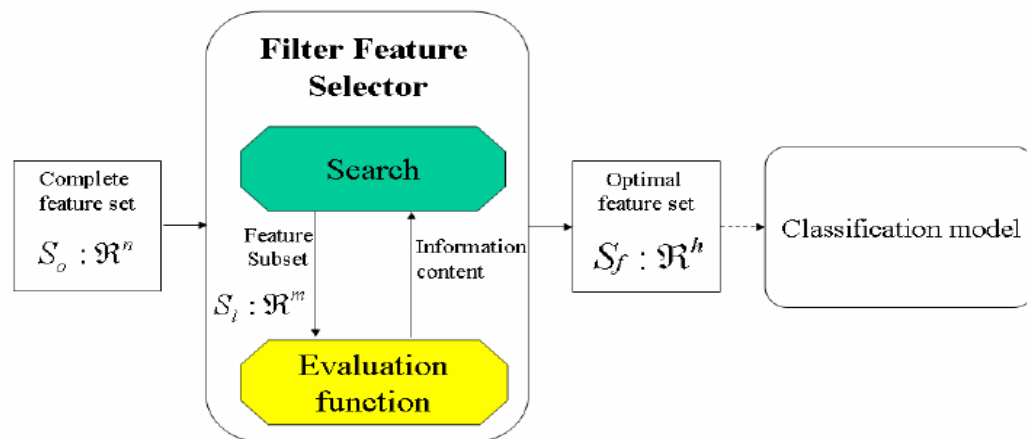
WINNOWING : FEATURE SELECTION

- Se divide aleatoriamente el set de entrenamiento en dos mitades. Con una de ellas se ajusta un árbol (llamado ***winnowing tree***) cuya finalidad es evaluar la utilidad de los predictores.
- Los predictores que no se han utilizado en ninguna división del ***winnowing tree*** se identifican como no útiles.
- La otra mitad de las observaciones de entrenamiento, no utilizadas en el ajuste del ***winnowing tree***, se emplean para estimar el error del árbol. Se inicia un proceso iterativo en el que se eliminan (de uno en uno) cada uno de los predictores que forman el ***winnowing tree***. Si, como resultado de la eliminación, el error del árbol disminuye, se considera que el predictor solo aporta ruido y se identifica como no útil.



WINNOWING : FEATURE SELECTION

- Una vez que todos los predictores han sido identificados como útiles o no útiles, se reajusta el *winnowing tree* utilizando únicamente los predictores útiles. Si al hacerlo, el error del árbol (calculado con la otra mitad del set de entrenamiento) no se reduce, el proceso de *winnowing* se descarta y no se excluye ningún predictor. Si por lo contrario, el error sí se reduce, se lanza el ajuste C5.0 convencional, empleando todas las observaciones de entrenamiento pero solo los predictores identificados como útiles.



MATRIZ DE PESOS PARA LOS ERRORES

- En muchos escenarios, las consecuencias de cometer un error de clasificación son distintas dependiendo el tipo que sean. Por ejemplo, en el ámbito biomédico, no es igual de grave confundir un tumor maligno con uno benigno que viceversa, ya que, en el primer caso, la vida del paciente puede estar en peligro.
- El algoritmo C5.0 permite asignar diferentes penalizaciones a cada tipo de error, lo que fuerza al modelo (árbol) a minimizar determinados tipos de error.

	Predicted	
	Positive	Negative
Actual True	TP	FN
Actual False	FP	TN



APLICACIÓN DE LA TECNOLOGÍA

- Los Árboles de Decisión se pueden utilizar para modelizar problemas de

- **Clasificación**

- Binaria (fraude vs no fraude)
- Multiclase (niveles de satisfacción: completamente, bastante, poco satisfecho, totalmente insatisfecho)

- **Regresión**

- Pagos que realiza una compañía de seguros
- Gasto mensual de los clientes de una cadena de supermercados



SECTORES DE APLICACIÓN

- **Industria del seguro** (modelización del riesgo)
- **Banca** (credit scoring, detección de fraude, fuga de clientes, modelos de cross-selling)
- **Sector retail** (fidelización y captación de clientes)
- **Telecos** (optimización de campañas, modelos de propensión de compra)
- **Muchos otros.....**





¡Gracias!



San Marcos Data Science Community

Auspicio : Escuela Académica Profesional de Estadística
San Marcos Data Science Community.



¿PREGUNTAS?

REALICEMOS EL TALLER

