



Universidad **Ricardo Palma**

RECTORADO
PROGRAMA DE ESPECIALIZACIÓN EN CIENCIA DE DATOS

Formamos seres humanos para una cultura de paz

TALLER DE ESTADÍSTICA PARA LA CIENCIA DE DATOS



R + Python



MÓDULO 5 : ANALISIS MULTIVARIADO II

Arboles de Decisión



A nuestro recordado Maestro

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



TALLER DE ESPECIALIZACIÓN “STATISTICAL SCIENCE INTRODUCTION”

TALLER DE ESTADÍSTICA PARA CIENCIA DE DATOS

EXPOSITORES



José Antonio Cárdenas Garro
UNMSM

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



André Omar Chávez Panduro
UNMSM

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

**Predictive Modelling
Specialist**



Data Scientist



**Portfolio and
Consumption Analyst**



**Customer Intelligence
Analyst**



Data Analyst



Data Analyst



Correo : josecardenasgarro@gmail.com
LinkedIn : www.linkedin.com/in/jos%C3%A1-antonio-c%C3%A1rdenas-garro-599266b0

Correo : andrecp38@gmail.com /
09140205@unmsm.edu.pe
LinkedIn : www.linkedin.com/in/andr%C3%A9-ch%C3%A1vez-a90078b9



TALLER DE ESPECIALIZACIÓN "STATISTICAL SCIENCE INTRODUCTION"

« Divide las dificultades que examinas en tantas partes como sea posible , para su mejor solución»



Agenda

- Introducción a los Árboles de Clasificación.
- Árbol de Clasificación **CHAID**.
- Fase de Fusión, División y Reglas de parada.



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.



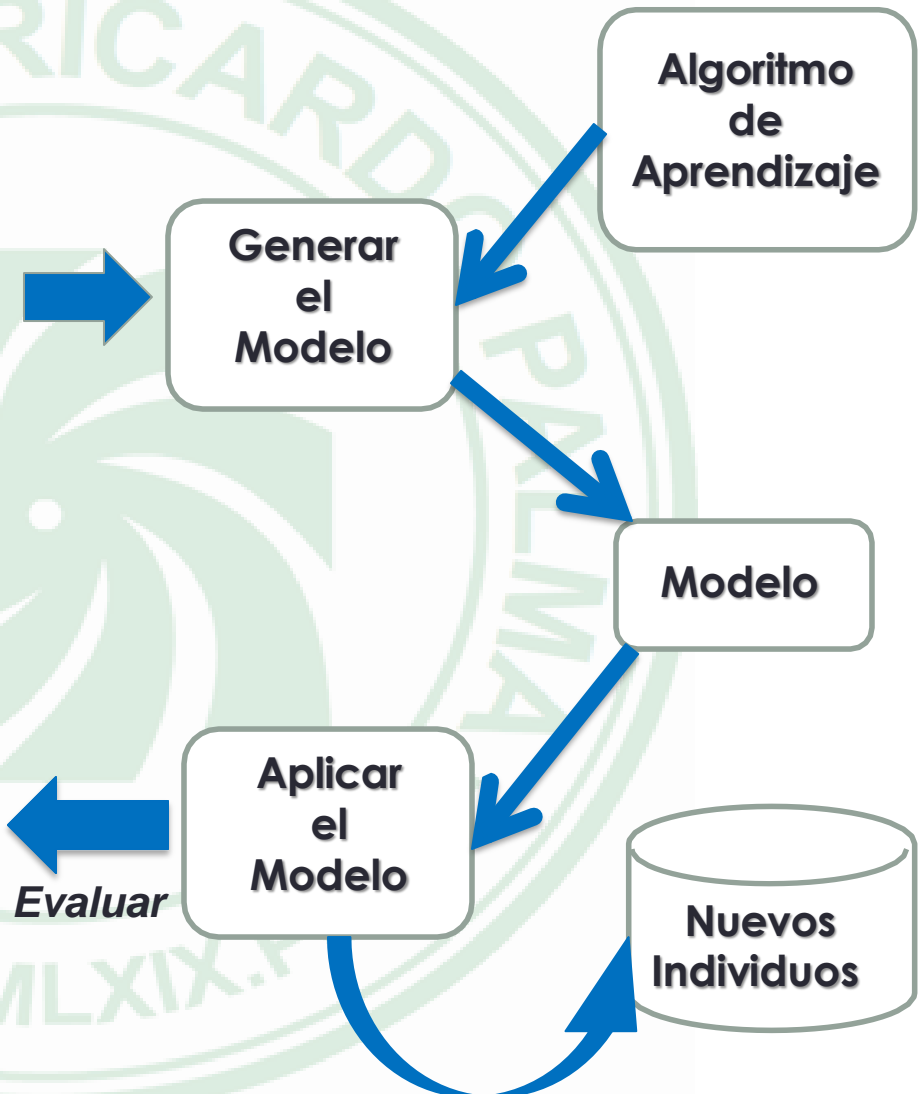
MODELO GENERAL DE LOS MÉTODOS DE CLASIFICACIÓN

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

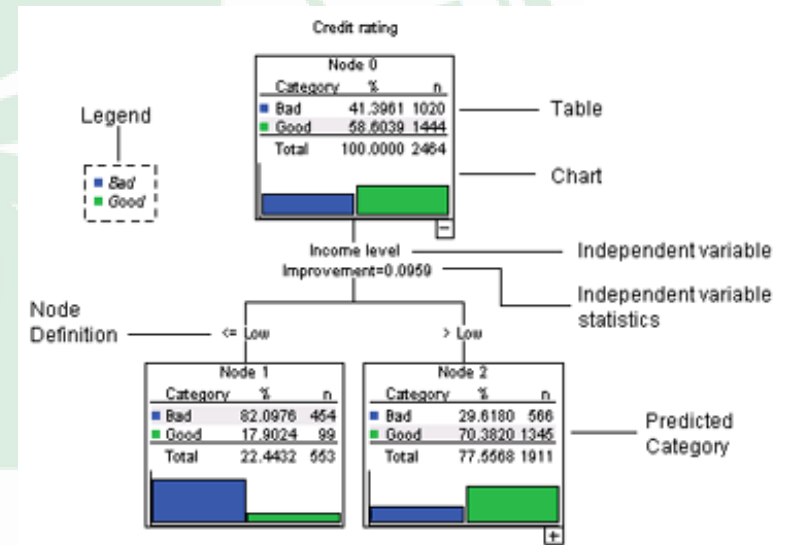
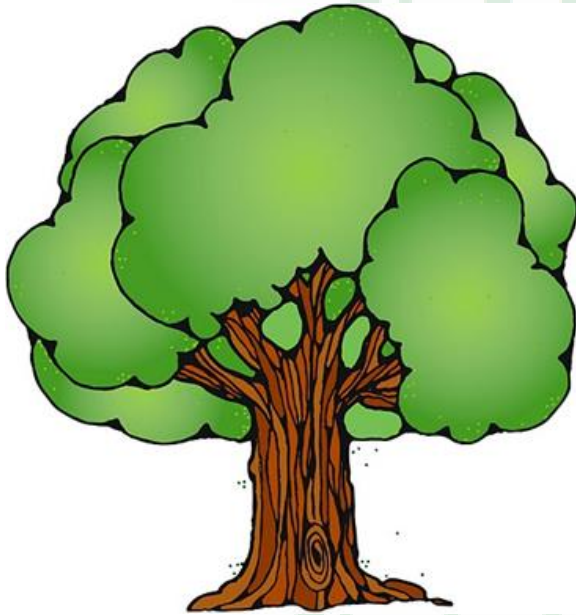


DEFINICIÓN DE CLASIFICACIÓN

- Dada una base de datos $D = \{t_1, t_2, \dots, t_n\}$ de tuplas o registros (individuos) y un conjunto de clases $C = \{C_1, C_2, \dots, C_m\}$, el **problema de la clasificación** es encontrar una función $f: D \rightarrow C$ tal que cada t_i es asignada una clase C_j .
- $f: D \rightarrow C$ podría ser una Red Neuronal, un Árbol de Decisión, un modelo basado en Análisis Discriminante, o una Red Bayesiana.



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**.



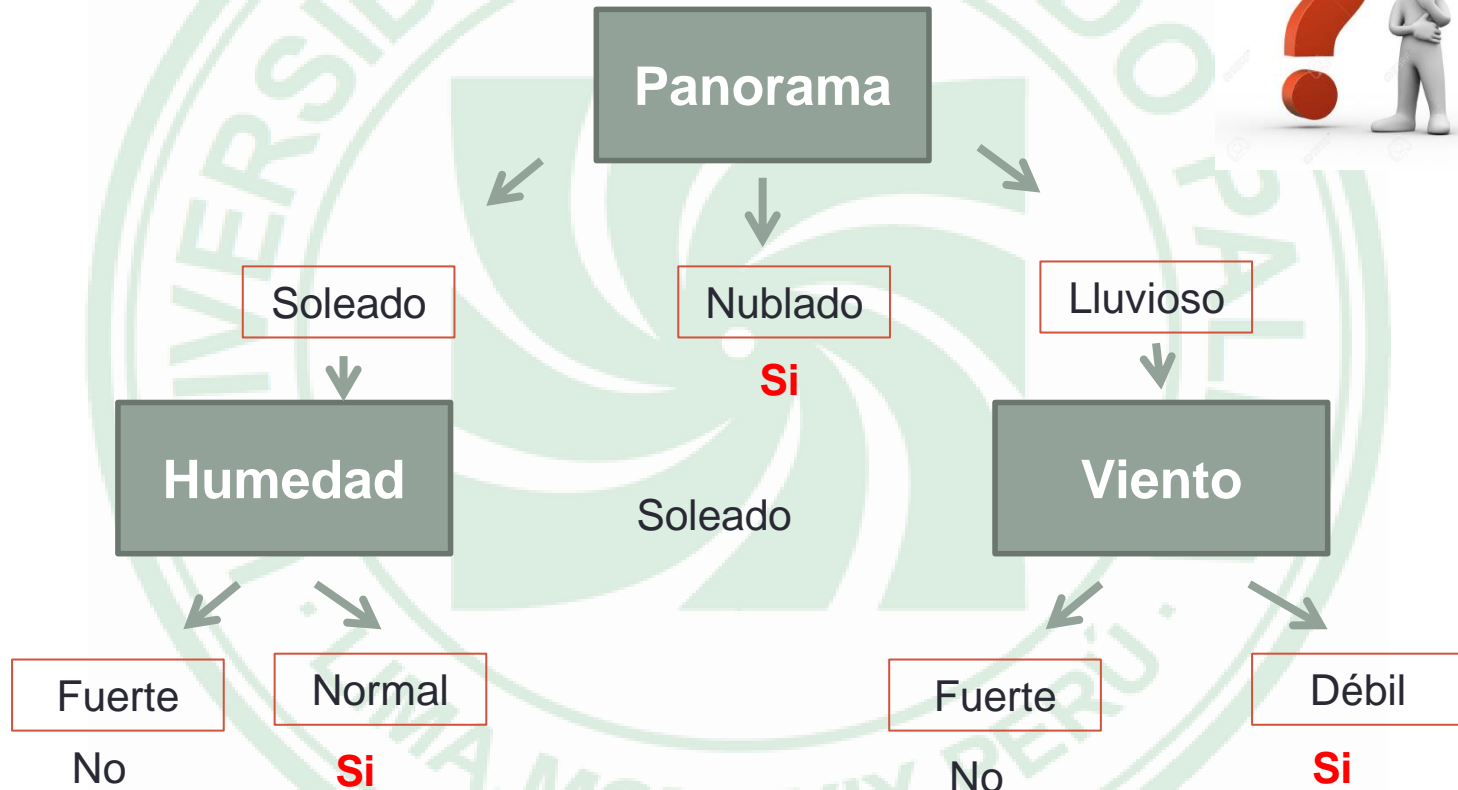
Árboles de Clasificación

- 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?

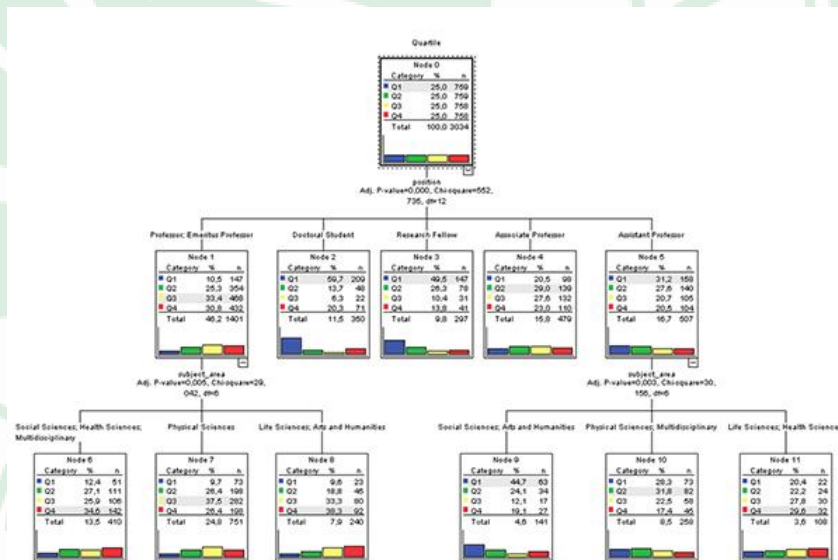


La Pregunta a responder es : Juego Tennis ?



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

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

SITUACIÓN CREDITICIA			
GÉNERO	MOROSO	NO MOROSO	TOTAL
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

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

SITUACIÓN CREDITICIA			
VARIABLE	MOROSO	NO MOROSO	TOTAL
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		
	A	B	TOTAL
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		
	A	C	TOTAL
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		
	B	C	TOTAL
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.





¡Gracias!

TALLER DE ESPECIALIZACIÓN “STATISTICAL SCIENCE INTRODUCTION”