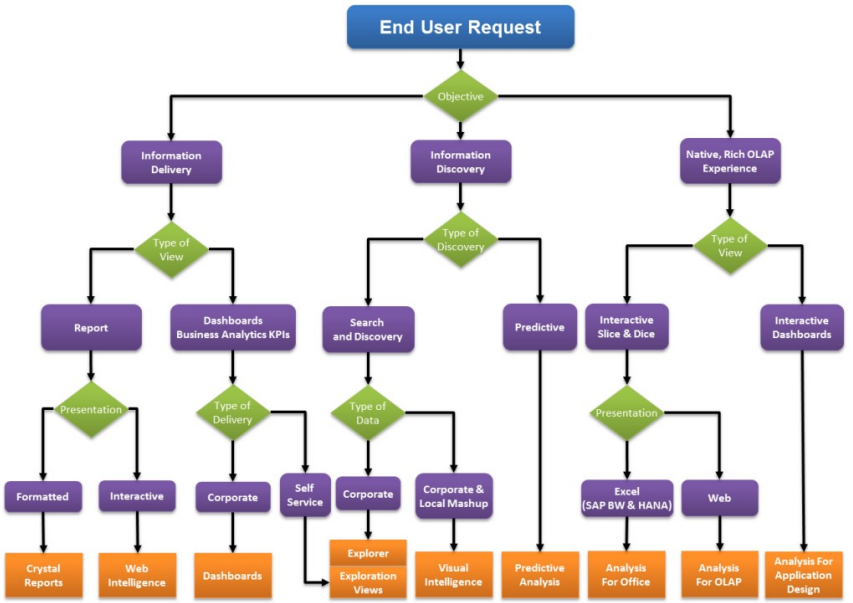
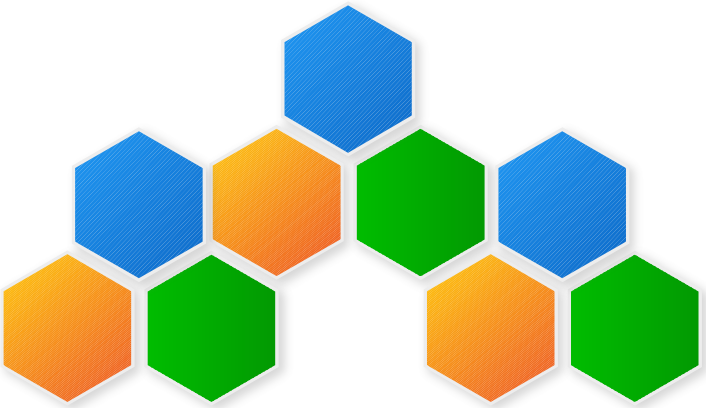
Árboles de decisión C4.5



Luis Casillas
DCC, CUCEI, UdeG
Julio - Noviembre 2015

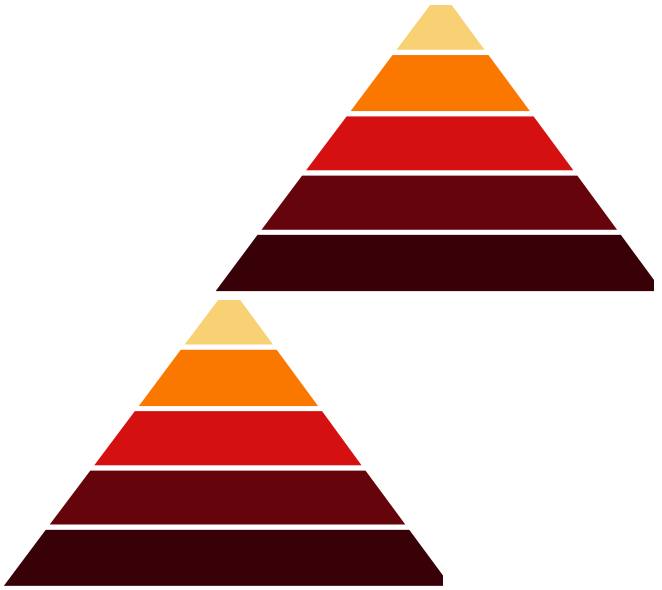
Árbol de Decisión

- Es una herramienta de apoyo en las decisiones.
- Tradicionalmente se han venido construyendo a mano.
- Provienen del mundo de la investigación de operaciones.
- Se usan principalmente para definir estrategias que permitan alcanzar metas.



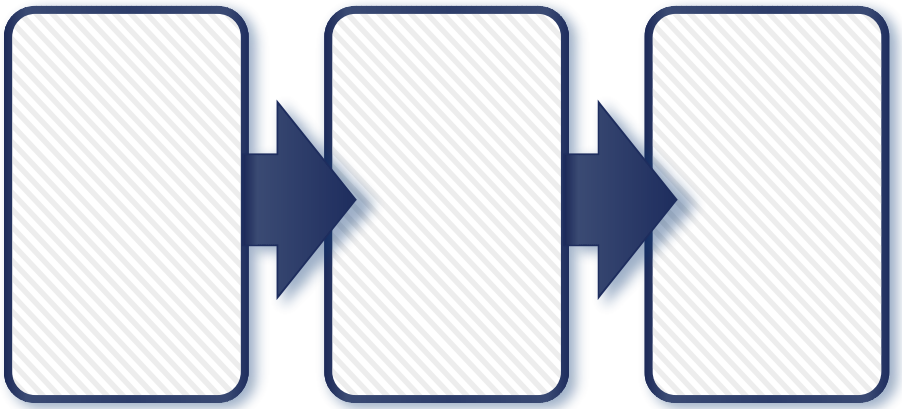
Construcción...

- El algoritmo C4.5 va eligiendo entre los atributos, el dato que tiene la mejor **carga de información**; para convertirlo en el nodo raíz del *subárbol* actual...



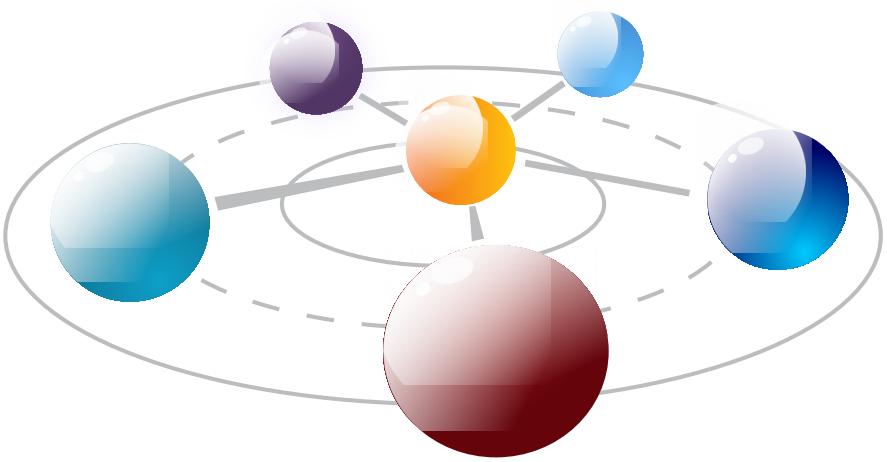
Construcción...

- La noción de **carga de información** se refiere a la capacidad de un dato para transmitir un mensaje con la menor pérdida...



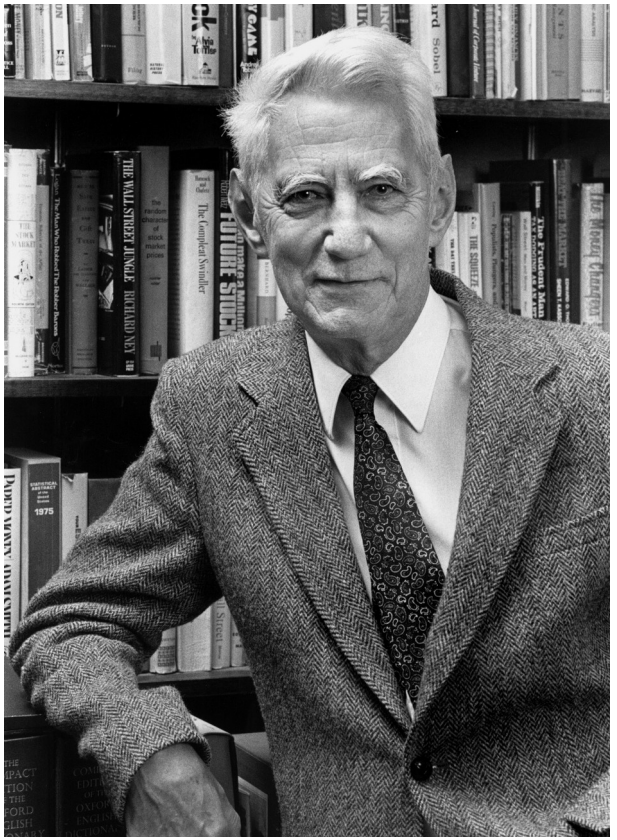
Construcción...

- Este dato es entonces usado como un mecanismo para dividir de forma más efectiva las muestras, en subconjuntos enriquecidos que se vinculan a las categorías semánticas de un modelo.



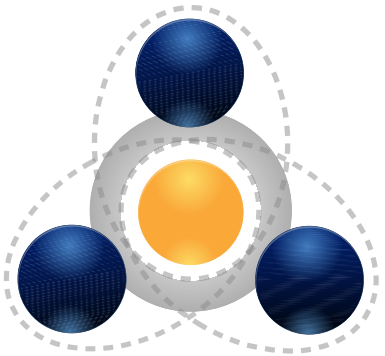
Normalización

- Mecanismo consistente en determinar el nivel de **entropía** vinculable a los datos del modelo.
- La noción **entropía**, en el contexto de la **Teoría de la Información** de **Claude Shannon** (1948), es el **valor promedio esperado** de información en cada mensaje recibido.



Normalización

- Se refiere a la **ganancia de información**, o bien la **diferencia en entropía**.
- El atributo con más alta **normalización** en la ganancia de información, es el elegido para decidir.
- Luego se recurre a las sublistas más pequeñas.



Entropía

- H(X) : H (Eta), la entropía de una variable X que es discreta y aleatoria.
- $H(X) = -\sum(P(X_i) * \log_b P(X_i))$
- Para:
  - i ∈ {1, 2, 3, … , numMuestras}
  - b ∈ {2, e, 10} : b=2 Ent. Shannon, b=e Ent. Natural, b=10 Ent. Hartley

Ganancia de Información

- Sea N un nodo que representa o guarda las tuplas de la partición D.
- El atributo con mayor ganancia de información es elegido para partir N.
- Este atributo minimiza la información necesaria para clasificar las tuplas en particiones.

Ganancia de Información

- La elección del atributo con mejor ganancia, implica:
  - Menor aleatoriedad o "impureza"
  - Minimiza el número de pruebas para clasificar una tupla dada.
  - Garantiza encontrar un árbol simple, aunque no el más simple.

Ganancia de Información

- Primero, la información que se espera para clasificar una tupla en D es dada por:

$$\mathbf{Info(D) = -\sum(p_i * log_2 p_i)}$$

Ganancia de Información

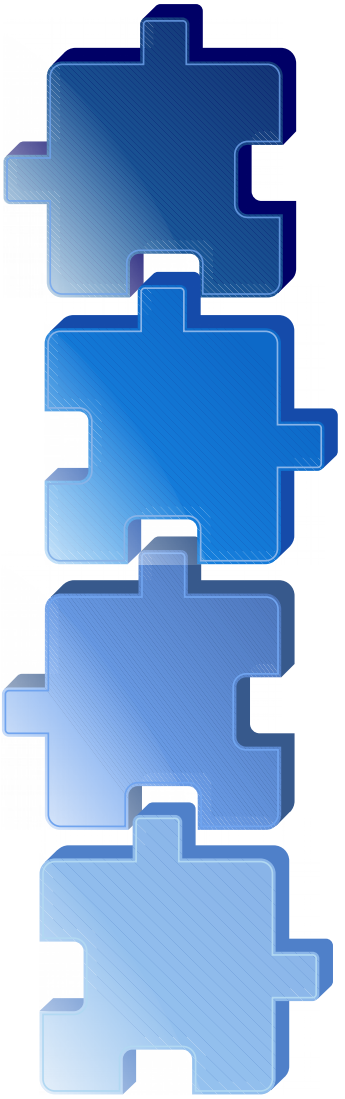
- Donde:
- $i \in \{1, 2, \ldots, numCasosClase\}$
- p_i: Prob≠0 de que una tupla arbitraria pertenezca al caso C_i de la clase. Se estima con $|C_i,D|/|D|$
- Se usa $\log_2$ debido a que la información se codifica como binaria.
- Info(D) es el monto promedio de información para identificar a una tupla en D como miembro de una clase. Info(D) se conoce también como la entropía de D.

Ganancia de Información

- Ahora, se busca particionar D en tuplas, basado en algún atributo A que tiene v valores distintos: {a1, a2, … ,av} a partir de datos de entrenamiento.
- Si A es discreto, el atributo A puede usarse para partir D en v particiones o subconjuntos: {D1, D2, … , Dv} donde Dj contiene las tuplas en D que entrega como salida aj de A.

Ganancia de Información

- Estas particiones corresponderían a las ramas del nodo N.
- Lo ideal es que cada partición sea pura, pero lo más probable es que serán impuras.
- **¿Cuánta información será necesaria (después del particionado) para alcanzar una clasificación exacta?**



Ganancia de Información

- Este monto es medido por:

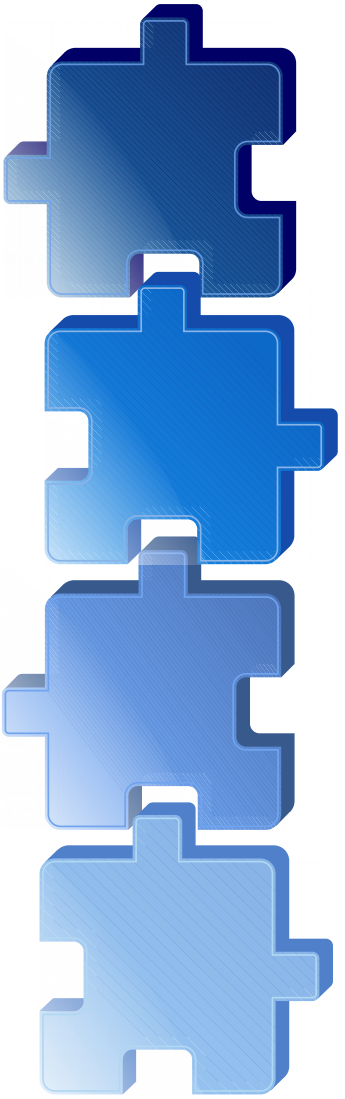
$$\mathbf{Info_A(D)=\sum((|D_j|/|D|)*Info(D_j))}$$

Donde:

$j \in \{1, 2, \ldots, numParticiones\}$

$|D_j|/|D|$: Peso j-ésima partición

$Info_A(D)$: Información para clasificar una tupla de D en el particionado por A



Ganancia de Información

- En la medida que la información esperada y requerida sea menor, mayor será la pureza de las particiones.

- Así, la Ganancia de Información es definida como la diferencia entre el requerimiento original de información (i.e. basado en la proporción justa de clases) y el nuevo requerimiento (i.e. obtenido luego de particionar respecto de A).

Ganancia de Información

- Esto es:

$$\textbf{Ganancia}(A) = \textbf{Info}(D) - \textbf{Info}_A(D)$$

La Ganancia(A) nos indica cuánto se ganaría ramificando respecto de A. Es la reducción esperada en la información requerida, que fue causada por conocer el valor de A.

El atributo A con la más alta ganancia, es el elegido para particionar N.

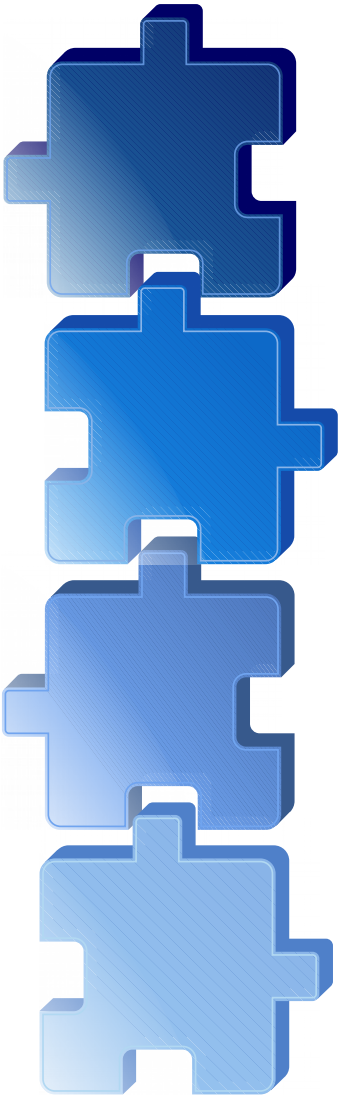
AllElectronics: Base de Clientes

IdR	Edad	Ingreso	Estudiante	Evaluación Crédito	Clase: Compra Computadora
1	Joven	Alto	No	Justo	No
2	Joven	Alto	No	Excelente	No
3	Media	Alto	No	Justo	Sí
4	Mayor	Medio	No	Justo	Sí
5	Mayor	Bajo	Sí	Justo	Sí
6	Mayor	Bajo	Sí	Excelente	No
7	Media	Bajo	Sí	Excelente	Sí
8	Joven	Medio	No	Justo	No
9	Joven	Bajo	Sí	Justo	Sí
10	Mayor	Medio	Sí	Justo	Sí
11	Joven	Medio	Sí	Excelente	Sí
12	Media	Medio	No	Excelente	Sí
13	Media	Alto	Sí	Justo	Sí
14	Mayor	Medio	No	Excelente	No

Encontrar Atributo para Partición

- **En este ejemplo:**

- Cada atributo es evaluado de forma discreta. Los valores continuos han sido generalizados.
- La clase **compra computadora** tiene dos valores: **Sí** y **No**. Es decir, que hay dos clases distintas (i.e. m=2).
- Sea C1=Sí y C2=No.
- Hay nueve tuplas de la clase **Sí** y cinco de la clase **No**



Encontrar Atributo para Partición

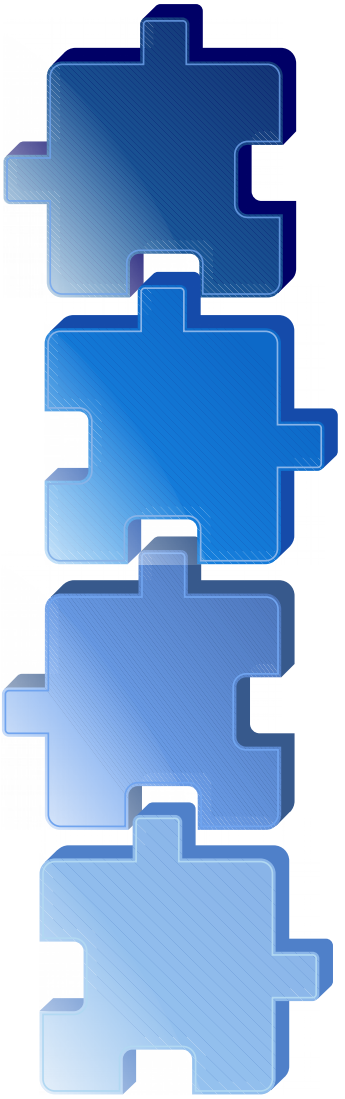
- Para encontrar el criterio de partición, debe calcularse la Ganancia de Información de cada atributo.
- Primero se calcula la ganancia necesaria para clasificar una tupla en D.

$$\mathbf{Info(D) = -(9/14)*\log_2(9/14) - (5/14)*\log_2(5/14)}$$

$$\mathbf{Info(D) = 0.940}$$

Encontrar Atributo para Partición

- Ahora, se calcula el requerimiento de información esperada para cada atributo.
- Iniciamos con **Edad**. Se observa la distribución de tuplas **Sí** y **No**, para cada categoría de **Edad**. Para la categoría **Joven** en **Edad**, hay dos tuplas **Sí** y tres **No**. Para **Media** en **Edad**, hay cuatro tuplas **Sí** y cero **No**. Para **Mayor** en **Edad**, hay tres tuplas **Sí** y dos **No**.



Encontrar Atributo para Partición

Edad / Clase	Joven	Media	Mayor
Clase: Sí	2	4	3
Clase: No	3	0	2
Σ	5	4	5

Ingreso / Clase	Alto	Medio	Bajo
Clase: Sí	2	4	3
Clase: No	2	2	1
Σ	4	6	4

Estudiante / Clase	Sí	No
Clase: Sí	6	3
Clase: No	1	4
Σ	7	7

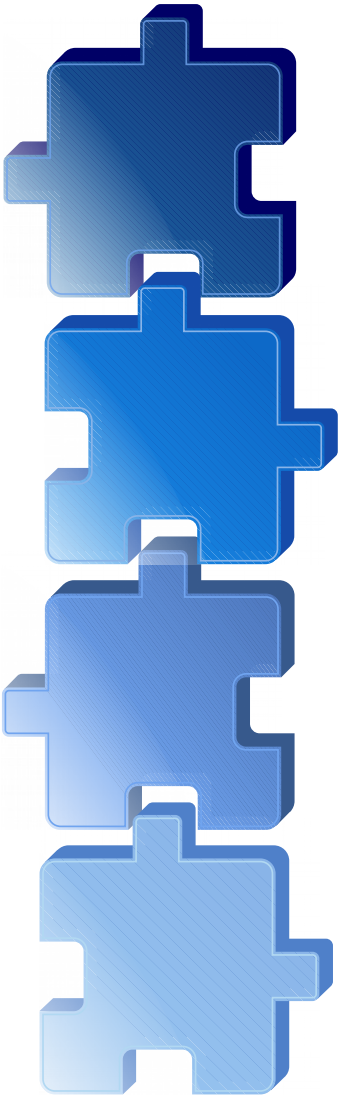
Eval.Crédito / Clase	Excelente	Justo
Clase: Sí	3	6
Clase: No	3	2
Σ	6	8

Encontrar Atributo para Partición

- Así, se obtiene:

$$\text{Info}_{\text{Edad}}(D) = 5/14 * (-(2/5)*\log_2(2/5) - (3/5)*\log_2(3/5)) + 4/14 * (-(4/4)*\log_2(4/4)) + 5/14 * (-(3/5)*\log_2(3/5) - (2/5)*\log_2(2/5))$$

$$\text{Info}_{\text{Edad}}(D) = 0.694 \text{ bits}$$

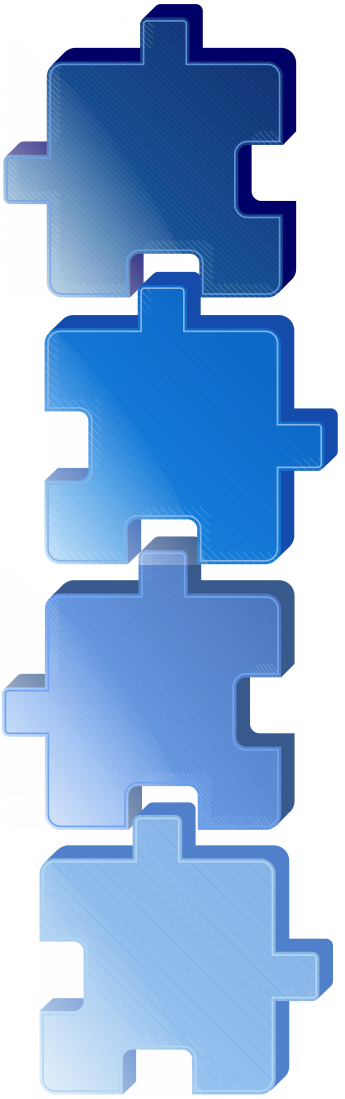


Encontrar Atributo para Partición

- La ganancia de información de tal particionado sería:

$$\textbf{Ganancia(Edad)} = \textbf{Info(D)} - \textbf{Info}_{\textbf{Edad}}\textbf{(D)}$$

$$\textbf{Ganancia(Edad)} = 0.940 - 0.694$$
$$= 0.246 \text{ bits}$$



Encontrar Atributo para Partición

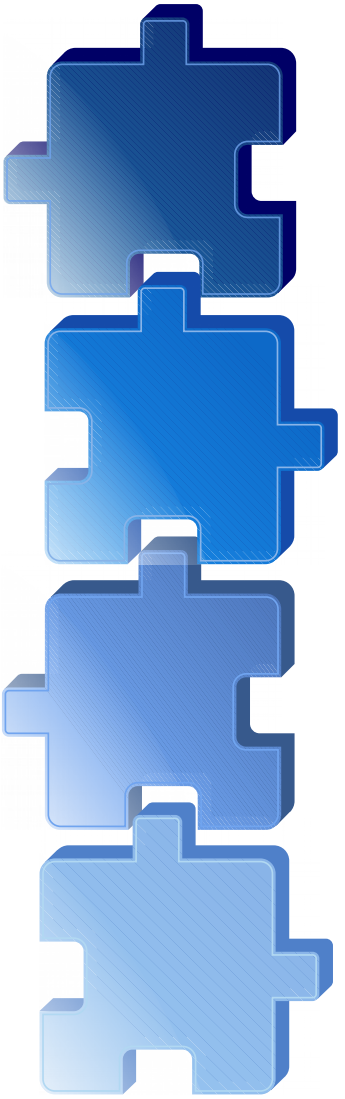
- Del mismo modo se calculan:

Ganancia(Ingreso) = 0.029 bits

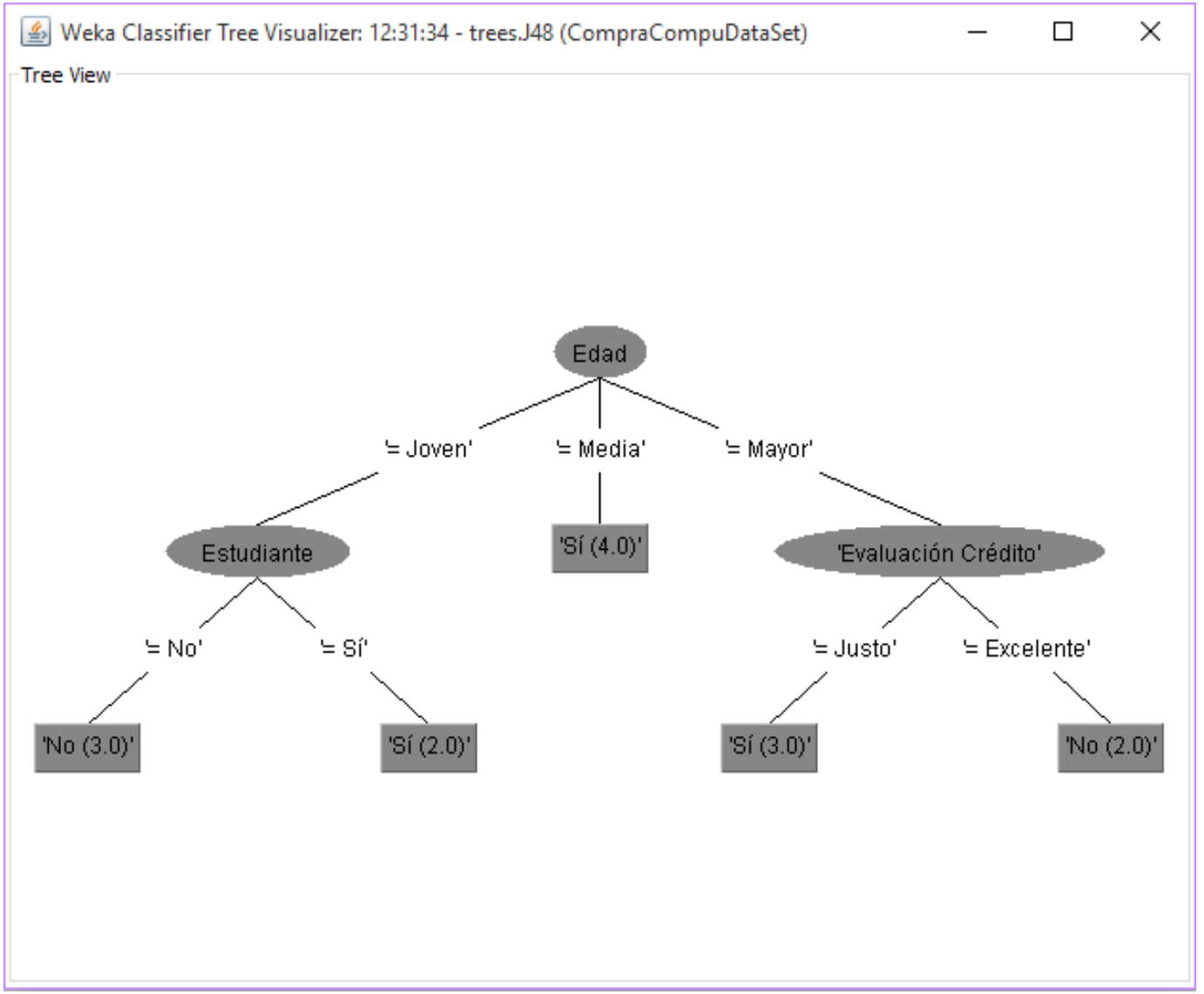
Ganancia(Estudiante) = 0.151 bits

Ganancia(EvalCrédito) = 0.048 bits

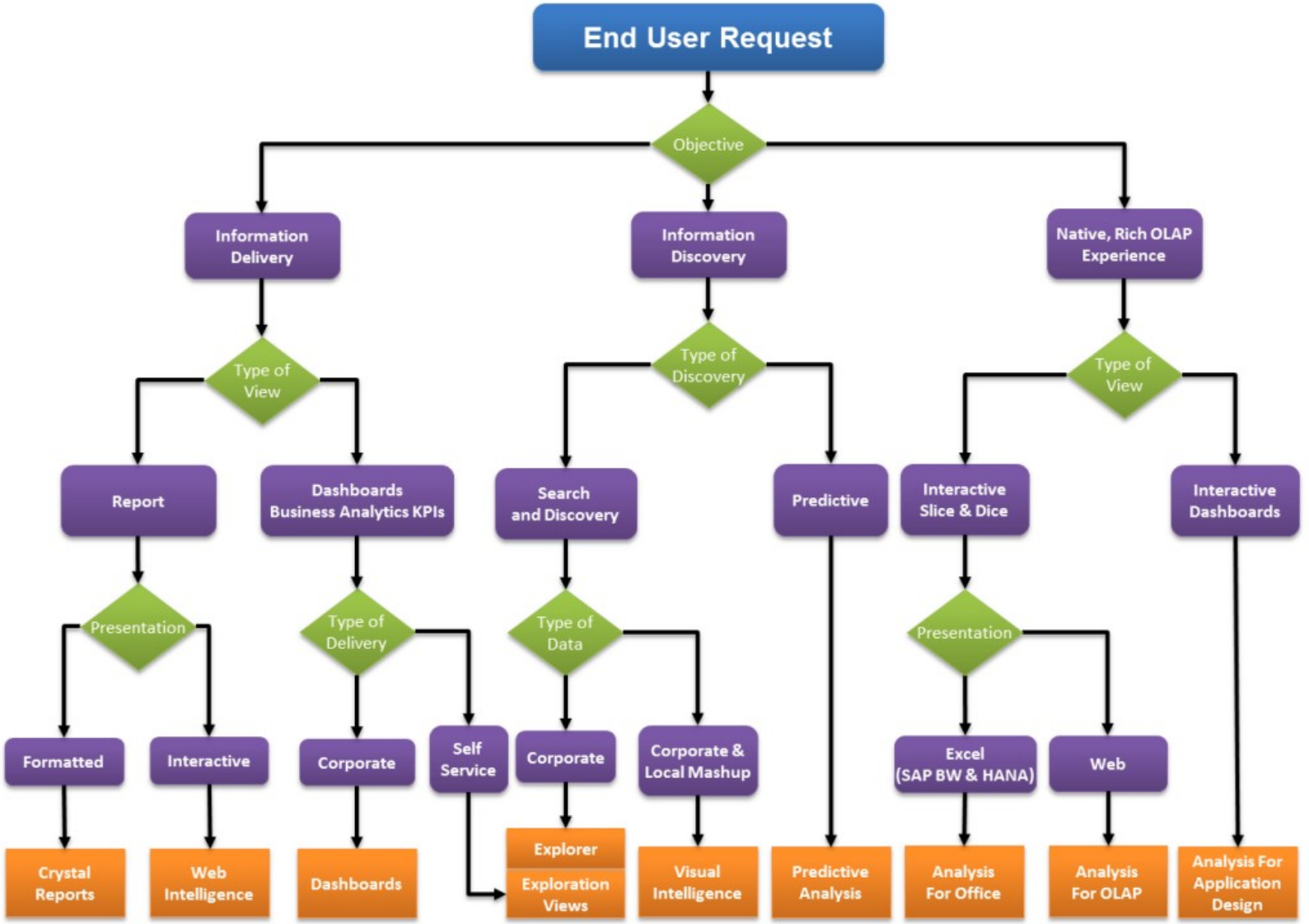
** El atributo **Edad** es el que tiene la mayor ganancia y con éste se realiza el el primer particionado. De forma recurrente se particiona con el resto de atributos.*



Árbol J48 Weka



Gracias :)


End User Request
Objective
Information Delivery
Information Discovery
Native, Rich OLAP Experience
Type of View
Type of Discovery
Type of View
Report
Dashboards Business Analytics KPIs
Search and Discovery
Predictive
Interactive Slice & Dice
Interactive Dashboards
Presentation
Type of Delivery
Type of Data
Presentation
Formatted
Interactive
Corporate
Self Service
Corporate
Corporate & Local Mashup
Excel (SAP BW & HANA)
Web
Crystal Reports
Web Intelligence
Dashboards
Explorer
Exploration Views
Visual Intelligence
Predictive Analysis
Analysis For Office
Analysis For OLAP
Analysis For Application Design