TEMA 2: Técnicas de minería de datos

Alumno: Francisco Márquez
Actividad: Actividades Tema 2

Actividad 2.1. Realiza los cálculos para el resto de pasos del algoritmo hasta llegar a que todos los nodos sean de tipo hoja.

Solución:

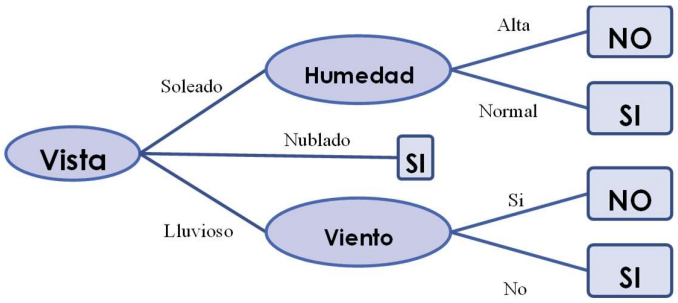
La tabla (2.1) representa un ejemplo sencillo de clasificación consistente en, a partir de los atributos que modelan el tiempo (vista, temperatura, humedad y viento), determinar si se puede o no jugar al tenis.

El ejemplo empleado tiene dos atributos, temperatura y humedad, que pueden emplearse como simbólicos o numéricos. Entre paréntesis se presentan sus valores numéricos.

Ejemplo	Vista	Temperatura	Humedad	Viento	Jugar
1	Soleado	Alta (85)	Alta (85)	No	No
2	Soleado	Alta(80)	Alta (90)	Si	No
3	Nublado	Alta (83)	Alta (86)	No	Si
4	Lluvioso	Media (70)	Alta (96)	No	Si
5	Lluvioso	Baja (68)	Normal (80)	No	Si
6	Lluvioso	Baja (65)	Normal (70)	Si	No
7	Nublado	Baja (64)	Normal (65)	Si	Si
8	Soleado	Media (72)	Alta (95)	No	No
9	Soleado	Baja (69)	Normal (70)	No	Si
10	Lluvioso	Media (75)	Normal (80)	No	Si
11	Soleado	Media (75)	Normal (70)	Si	Si
12	Nublado	Media (72)	Alta (90)	Si	Si
13	Nublado	Alta (81)	Normal (75)	No	Si
14	Lluvioso	Media (71)	Alta (91)	Si	No

Tabla 2.1. Ejemplo de clasificación

La Actividad consiste en completar los cálculos del algoritmo ID3 con el que obtiene la elaboración del siguiente árbol de decisión:



El algoritmo se describe a continuación:

- Seleccionar el atributo A que maximice la ganancia G(Ai)
- Crear un nodo para ese atributo con tantos sucesores como valores tenga. Introducir los ejemplos en los sucesores según el valor que tenga el atributo Ai.
- Por cada sucesor:
- Si sólo hay ejemplos de una clase, Ck, entonces etiquetarlo con Ck. Si no, llamar a ID3 con una tabla formada por los ejemplos de ese nodo, eliminado la columna del atributo Ai.

Y las formas de cálculo son las siguientes:

Dado un conjunto de eventos $A = \{A_1, A_2, \ldots, A_n\}$, con probabilidades $\{p_1, p_2, \ldots, p_n\}$, la información en el conocimiento de un suceso A_i (bits) se define en la ecuación (1), mientras que la información media de A (bits) se muestra en la ecuación (2).

$$I(A_i) = log_2\left(\frac{1}{p_i}\right) = -log_2(p_i) \tag{1}$$

$$I(A) = \sum_{i=1}^{n} p_i I(A_i) = -\sum_{i=1}^{n} p_i log_2(p_i) \tag{2}$$

El criterio de partición utilizado en el algoritmo ID3, denominado ganancia de información está basado en medir la ganancia de información asociada al usar el atributo A_i como:

$$G(A_i) = I - I(A_i) \tag{3}$$

Siendo I la información antes de utilizar el atributo e $I(A_i)$ la información después de usarlo. Sus expresiones vienen dadas por las ecuaciones (4) y (5).

$$I = -\sum_{c=1}^{nc} \frac{n_c}{n} log_2\left(\frac{n_c}{n}\right) \tag{4}$$

$$I(A_i) = \sum_{j=1}^{nv(A_i)} \frac{n_{ij}}{n} I_{ij} \; ; \; I_{ij} = -\sum_{k=1}^{nc} \frac{n_{ijk}}{n_{ij}} log_2\left(\frac{n_{ijk}}{n_{ij}}\right) \tag{5}$$

Donde,

nc: representa el número de clases.
n_c: el número de ejemplares de la clase c.
n: el número total de ejemplos.
$nv(a_i)$: el número de valores del atributo *A*.
n_{ij}: el número de ejemplos con el valor *j* en A_i.
n_{ijk}: el número de ejemplos con valor *j* en A_i y que pertenecen a la clase *k*.

Para generar el árbol de decisión será necesario decidir en cada caso cual es el atributo que genera la rama. Para ello, seleccionaremos el atributo que maximice la ganancia.

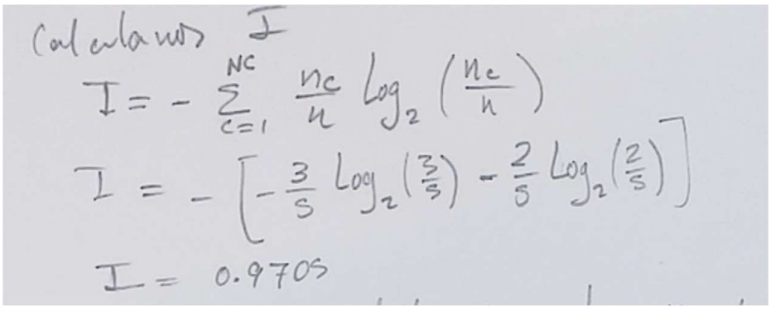
Tomemos en cuenta que los cálculos necesarios inician luego de formado el primer nodo, el cual fue decidido con el atributo 'Vista'. Este atributo tiene 3 categorías: soleado, nublado y lluvioso. Para cada uno es estos nodos tendremos que determinar el atributo que determinará la bifurcación del nodo, excepto por la categoría 'nublado' el cual es un nodo hoja porque todos los ya que todos los ejemplos de entrenamiento que llegan a dicho nodo son de clase SI.

Nodo 'soleado':

En primer lugar reducimos la tabla para gestionar los cálculos quedándonos sólo con las filas de la clase 'soleado' excluyendo el atributo 'vista'.

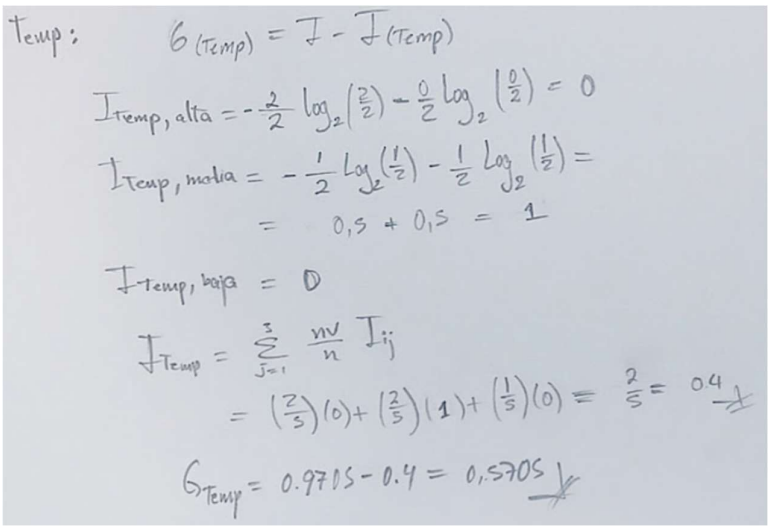
Ejemplo	Vista	Temperatura	Humedad	Viento	Jugar
1	Soleado	Alta (85)	Alta (85)	No	No
2	Soleado	Alta(80)	Alta (90)	Si	No
8	Soleado	Media (72)	Alta (95)	No	No
9	Soleado	Baja (69)	Normal (70)	No	Si
11	Soleado	Media (75)	Normal (70)	Si	Si

Luego repitiendo el procedimiento original, obtenemos I:

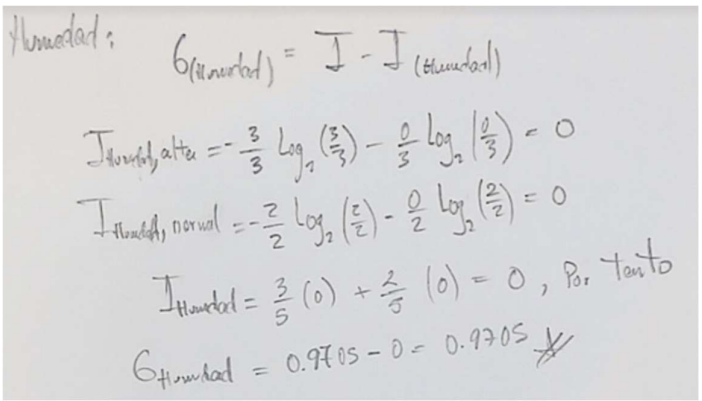


Seguidamente, Calculamos la Ganancia (G) para cada atributo: temperatura, humedad y viento:

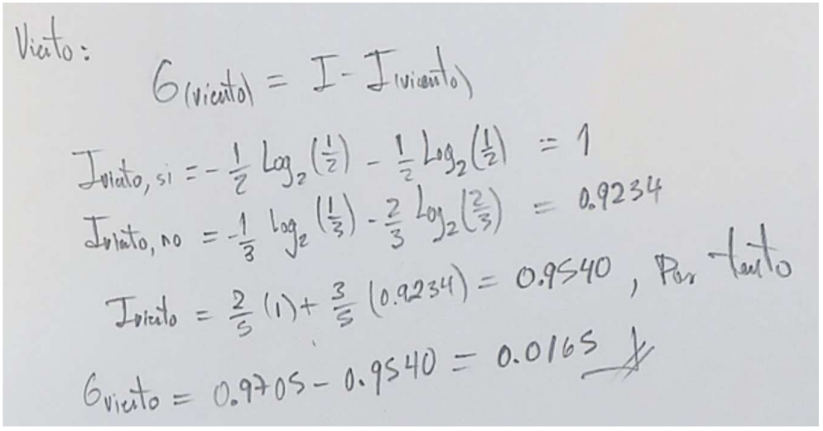
Temperatura:



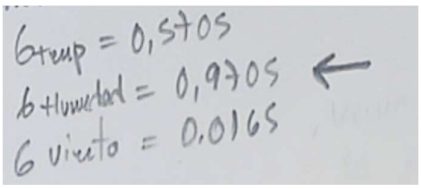
Humedad:



Viento:



Al comparar los resultados tenemos:



Vemos que para el segundo nodo del árbol a partir del nodo ‘soleado’, el mejor atributo para bifurcar el nodo es ‘humedad’. En la siguiente tabla podemos ver como los nodos a partir de Humedad terminan siendo nodos hoja, para *humedad = Alta* todos los ejemplos de entrenamiento son NO para *humedad = normal* todos los ejemplos de entrenamiento son SI. Esto se puede validar en las siguientes tablas:

soleado/humedad alta

Ejemplo	Vista	Temperatura	Humedad	Viento	Jugar
1	Soleado	Alta (85)	Alta (85)	No	No
2	Soleado	Alta(80)	Alta (90)	Si	No
8	Soleado	Media (72)	Alta (95)	No	No

soleado/humedad normal

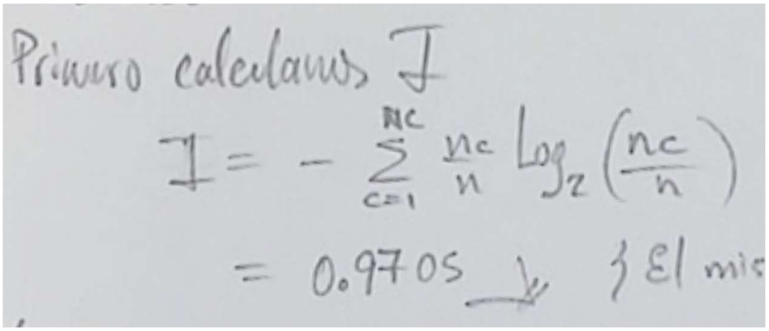
Ejemplo	Vista	Temperatura	Humedad	Viento	Jugar
9	Soleado	Baja (69)	Normal (70)	No	Si
11	Soleado	Media (75)	Normal (70)	Si	Si

Nodo ‘lluvioso’:

En primer lugar reducimos la tabla para gestionar los cálculos quedándonos sólo con las filas de la clase ‘lluvioso’ excluyendo el atributo ‘vista’.

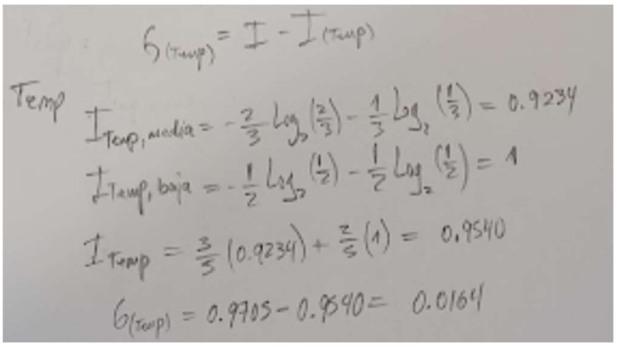
Ejemplo	Vista	Temperatura	Humedad	Viento	Jugar
4	Lluvioso	Media (70)	Alta (96)	No	Si
5	Lluvioso	Baja (68)	Normal (80)	No	Si
6	Lluvioso	Baja (65)	Normal (70)	Si	No
10	Lluvioso	Media (75)	Normal (80)	No	Si
14	Lluvioso	Media (71)	Alta (91)	Si	No

Luego repitiendo el procedimiento original, obtenemos I. Que resultó ser el mismo cálculo que para ‘soleado’:

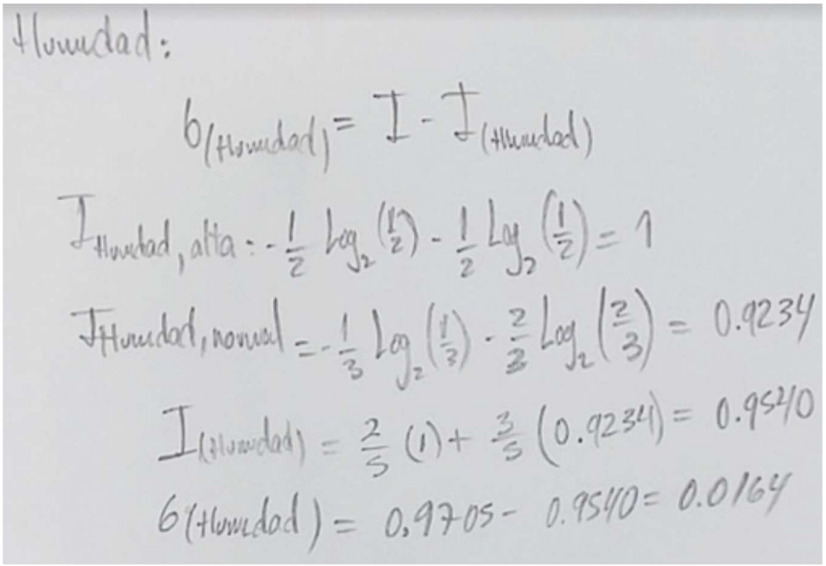


Seguidamente, Calculamos la Ganancia (G) para cada atributo: temperatura, humedad y viento:

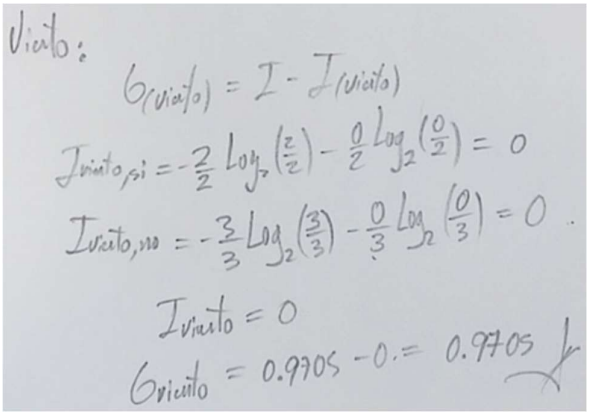
Temperatura:



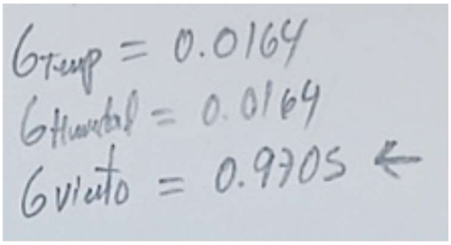
Humedad:



Viento:



Al comparar los resultados tenemos:



Vemos que para el segundo nodo del árbol a partir del nodo ‘lluvioso’, el mejor atributo para bifurcar el nodo es ‘viento’. En la siguiente tabla podemos ver como los nodos a partir de viento terminan siendo nodos hoja, para *viento = sí* todos los ejemplos de entrenamiento son NO para *viento = no* todos los ejemplos de entrenamiento son SI. Esto se puede validar en las siguientes tablas:

lluvioso/viento si

Ejemplo	Vista	Temperatura	Humedad	Viento	Jugar
6	Lluvioso	Baja (65)	Normal (70)	Si	No
14	Lluvioso	Media (71)	Alta (91)	Si	No

lluvioso/viento no

Ejemplo	Vista	Temperatura	Humedad	Viento	Jugar
4	Lluvioso	Media (70)	Alta (96)	No	Si
5	Lluvioso	Baja (68)	Normal (80)	No	Si
10	Lluvioso	Media (75)	Normal (80)	No	Si