ÁRBOLES DE DECISIÓN

Dado el siguiente conjunto de datos:

Edad	Antecedentes Familiares	Volumen Próstata (mm^3)	Valor PSA	Cáncer de Próstata
<45	N	< 60	< 4	N
<45	N	< 60	< 4	N
[45,70]	N	> 60	[4,10]	P
>70	N	> 60	> 10	P
>70	S	< 60	[4,10]	N
>70	S	< 60	>10	P
[45,70]	S	> 60	< 4	N
<45	N	< 60	[4,10]	N
<45	S	< 60	< 4	N
>70	S	< 60	< 4	N
<45	S	> 60	[4,10]	P
[45,70]	N	> 60	>10	P
[45,70]	S	< 60	[4,10]	P
>70	N	> 60	> 10	P

Tabla 1

Considerando los datos de la ***tabla 1***, se pretende construir un clasificador basado en árboles de decisión.
Se definen las siguientes medidas:

Entropía o Cantidad de Información:
Si un experimento puede tener *m* resultados distintos: $v_1, v_2, \cdots, v_m$ que se pueden producir con probabilidades $P(v_1), P(v_2), \cdots, P(v_m)$, la cantidad de información I que se obtiene al conocer el resultado real del experimento es:

$$I[P(v_1), P(v_2), \cdots, P(v_m)] = \sum_{i=1}^{m} - P(v_i) \log_2 P(v_i)$$

Entropía Residual:
La información residual respecto a un atributo A es la cantidad de entropía que se debe reducir para clasificar una instancia después de emplear el atributo A.

$$I_{RES}(A) = \sum_{v \in V(A)} P(v) \times I(E_v)$$

Ganancia de Información:
La ganancia de información es la reducción esperada en la entropía tras la partición de los ejemplos de acuerdo a un atributo A.

$$G(A) = I - I_{RES}(A)$$

Se deben calcular los siguientes elementos:

- Entropía inicial del problema considerando las clases de salida:

I = -P(Cáncer de Próstata=P) log2(P(Cáncer de Próstata=P))- P(Cáncer de Próstata=N) log2(P(Cáncer de Próstata=N)) = -(7/14) log2(7/14)- (7/14) log2(7/14) = 1

- Información residual y ganancia para cada atributo

Para el atributo Edad
Se determina en primer lugar las siguientes cantidades que se utilizarán para realizar los cálculos de entropías.
Valores del atributo: $V(Edad) = \{< 45, [45,70], > 70\}$
Distribución de los ejemplos por clase: $E_{TOT} = [7N \quad 7P]$
Distribución de los ejemplos para cada rango del atributo:
$E_{<45} = [4N \quad 1P]$
$E_{[45,70]} = [1N \quad 3P]$
$E_{>70} = [2N \quad 3P]$

I($E_{<45}$) = -P(Cáncer de Próstata=N | Edad<45) log2(P(Cáncer de Próstata=N|Edad<45))- P(Cáncer de Próstata=P | Edad<45) log2(P(Cáncer de Próstata=P|Edad<45))= -(4/5) log2(4/5)- (1/5) log2(1/5) = 0.72

I($E_{[45,70]}$) = -P(Cáncer de Próstata=N | Edad=[45,70]) log2(P(Cáncer de Próstata=N | Edad=[45,70]))- P(Cáncer de Próstata=P | Edad=[45,70]) log2(P(Cáncer de Próstata=P|Edad=[45,70]))= -(1/4) log2(1/4)- (3/4) log2(3/4) = 0.811

I($E_{>70}$) = -P(Cáncer de Próstata=N | Edad>70) log2(P(Cáncer de Próstata=N|Edad>70))- -P(Cáncer de Próstata=P | Edad>70) log2(P(Cáncer de Próstata=P|Edad>70))= -(2/5) log2(2/5)- (3/5) log2(3/5) = 0.97

$G(Edad) = I - I_{RES}(Edad) = \sum_{v \in V(Edad)} P(v) \times I(E_v)$ = 1 – (5/14) I($E_{<45}$) - (4/14) I($E_{[45,70]}$) - (5/14) I($E_{>70}$) = 1-(5/14).0.72 – (4/14).0.811-(5/14).0.97 = 0.165

Para el atributo Antecedentes Familiares
$V(AF) = \{N, S\}$
$E_{TOT} = [7N \quad 7P]$
$E_N = [3N \quad 4P]$
$E_P = [4N \quad 3P]$

I(E_N) = -P(Cáncer de Próstata=N | AF=N) log2(P(Cáncer de Próstata=N | AF=N))- P(Cáncer de Próstata=P | AF=N) log2(P(Cáncer de Próstata=P| AF=N))= -(3/7) log2(3/7)- (4/7) log2(4/7) = 0.985

$I(E_P)$ = -P(Cáncer de Próstata=N | AF=P) log2(P(Cáncer de Próstata=N | AF=P))-P(Cáncer de Próstata=P | AF=P) log2(P(Cáncer de Próstata=P| AF=P))= -(4/7) log2(4/7)- (3/7) log2(3/7) = 0.985

$$G(\text{AntecedentesFamiliares}) = I - I_{RES}(\text{AntecedentesFamiliares}) = \sum_{v \in V(\text{AF})} P(v) \times I(E_v)$$

= 1 – (7/14) $I(E_N)$ - (7/14) $I(E_P)$ = 1-(7/14).0.985 -(7/14).0. 985 = 0.0147

Para el atributo Volumen Próstata

$V(VP) = \{< 60, > 60\}$

$E_{TOT} = [7N \quad 7P]$

$E_{<60} = [6N \quad 2P]$

$E_{>60} = [1N \quad 5P]$

$I(E_{<60})$ = -P(Cáncer de Próstata=N | VP<60) log2(P(Cáncer de Próstata=N| VP<60))-P(Cáncer de Próstata=P | VP<60) log2(P(Cáncer de Próstata=P| VP<60))= -(6/8) log2(6/8)- (2/8) log2(2/8) = 0.811

$I(E_{>60})$ = -P(Cáncer de Próstata=N | VP>60) log2(P(Cáncer de Próstata=N| VP>60))--P(Cáncer de Próstata=P | VP>60) log2(P(Cáncer de Próstata=P| VP>60))= -(1/6) log2(1/6)- (5/6) log2(5/6) = 0.650

$G(VP) = I - I_{RES}(VP) = \sum_{v \in V(VP)} P(v) \times I(E_v)$ = 1 – (8/14) $I(E_{<60})$ - (6/14) $I(E_{>60})$ = 1-(8/14).0.811 – (6/14).0.650 = 0.258

Para el atributo PSA

$V(PSA) = \{< 4, [4{,}10], > 10\}$

$E_{TOT} = [7N \quad 7P]$

$E_{<4} = [5N \quad 0P]$

$E_{[4,10]} = [2N \quad 3P]$

$E_{>10} = [0N \quad 4P]$

$I(E_{<4})$ = -P(Cáncer de Próstata=N | PSA<4) log2(P(Cáncer de Próstata=N| PSA <4))- P(Cáncer de Próstata=P | PSA <4) log2(P(Cáncer de Próstata=P| PSA <4))= -(5/5) log2(5/5)- (0/5) log2(0/5) = 0

$I(E_{[4,10]})$ = -P(Cáncer de Próstata=N | PSA =[4,10]) log2(P(Cáncer de Próstata=N| PSA =[4,10]))- P(Cáncer de Próstata=P | PSA =[4,10]) log2(P(Cáncer de Próstata=P| PSA =[4,10]))= -(2/5) log2(2/5)- (3/5) log2(3/5) = 0.97

$I(E_{>10})$ = -P(Cáncer de Próstata=N | PSA >70) log2(P(Cáncer de Próstata=N| PSA >10))- -P(Cáncer de Próstata=P | PSA >10) log2(P(Cáncer de Próstata=P| PSA >10))= -(0/4) log2(0/4)- (4/4) log2(4/4) = 0

$G(PSA) = I - I_{RES}(\text{PSA}) = \sum_{v \in V(\text{PSA})} P(v) \times I(E_v) = 1 - (5/14)\ I(E_{<4}) - (5/14)\ I(E_{[4,10]}) - (4/14)\ I(E_{>10}) = 1-(5/14).0 - (5/14).0.97-(4/14).0 = 0.346$

Considerando los valores de ganancia de información para cada atributo el algoritmo ID3 elegirá PSA como primera variable de separación.

El proceso de seleccionar un nuevo atributo y particionar los ejemplos de entrenamiento es repetido ahora para cada nodo descendiente no-terminal, usando esta vez sólo los ejemplos de entrenamiento asociados con este nodo. Los atributos que han sido incorporados más alto en el árbol son excluidos y por consiguiente, cualquier atributo puede aparecer a lo sumo una vez en cualquier paso del árbol.

El proceso continúa para cada nuevo nodo hoja, hasta que alguna de las siguientes 2 condiciones es alcanzada:
1. Todo atributo ya ha sido incluido en este paso a través del árbol.
2. Todos los ejemplos de entrenamiento asociados con este nodo hoja tienen el mismo valor de atributo objetivo (entropía = 0).

El árbol resultante es el siguiente: