

Asignatura	
Técnicas de Inteligencia Artificial	Claudia Villalonga Palliser

## Actividades resueltas

### ID3 – Árbol de decisión

#### Descripción de la actividad

Una óptica tiene la siguiente base de datos con registros de los clientes a los que se les ha prescrito lentes de contacto duras o blandas.

	Edad	Anomalía	Astigmatismo	Lentes de contacto
E1	joven	miopía	no	blandas
E2	joven	miopía	sí	blandas
E3	joven	hipermetropía	no	blandas
E4	joven	hipermetropía	sí	duras
E5	pre-presbicia	miopía	no	blandas
E6	pre-presbicia	miopía	sí	duras
E7	pre-presbicia	hipermetropía	no	blandas
E8	presbicia	miopía	sí	duras
E9	presbicia	hipermetropía	no	blandas

Los atributos de los pacientes y los correspondientes posibles valores para estas características son:

1. Edad: joven, pre-presbicia, presbicia
2. Anomalía: miopía, hipermetropía
3. Astigmatismo: no, sí
4. Lentes de contacto: duras, blandas

A partir de esta base de datos se quiere construir un modelo predictor del tipo de lentes de contacto (duras o blandas) que ha de llevar un paciente. La predicción del tipo de lentes de contacto para los pacientes que se presenten en un futuro en la óptica se va a efectuar en base a los valores de los atributos edad, anomalía y astigmatismo.

Se ejecuta el **algoritmo ID3** para crear el modelo predictor. Al aplicar el método de selección de atributos se selecciona el atributo **Astigmatismo** para que sea el nodo raíz del árbol de decisión.

Asignatura	
Técnicas de Inteligencia Artificial	Claudia Villalonga Palliser

Explica por qué el algoritmo ID3 ha escogido el atributo Astigmatismo para la raíz del árbol. No es necesario que lo pruebes haciendo los cálculos, sólo que expliques razonadamente el por qué.

Continúa con la aplicación del algoritmo ID3 para desarrollar el árbol de decisión a partir del nodo raíz Astigmatismo. Describe claramente los pasos que se realizan en la ejecución del algoritmo y además dibuja el árbol de decisión.

### Resolución de la actividad

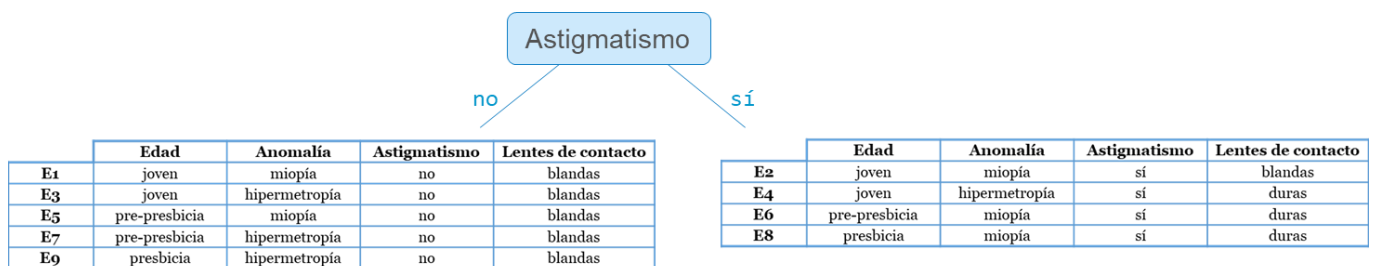
El algoritmo ID3 utiliza un método de selección de atributos basado en la ganancia de información. Entonces podemos asegurar que si ID3 ha escogido el atributo Astigmatismo para la raíz del árbol es porque de los tres atributos (Edad, Anomalía y Astigmatismo) es el que tiene mayor ganancia de información. La ganancia de información mide la reducción de la entropía al distribuir los ejemplos de acuerdo a los valores de un atributo. Por lo tanto el atributo que tiene mayor ganancia de información es el que aporta mayor información y por lo tanto el más útil para particionar los ejemplos y el más efectivo en la clasificación.

En el enunciado se indica que no es necesario probar porque se ha elegido el atributo Astigmatismo sin embargo si hiciéramos los cálculos veríamos que la ganancia de información de este atributo es de 0.5577, la ganancia de información del atributo Edad es de 0.0294 y la ganancia de información del atributo Anomalía es de 0.0183.

Continuamos con la aplicación del algoritmo ID3 para desarrollar el árbol de decisión a partir del nodo raíz Astigmatismo. Como el método de selección de atributos ha decidido que el nodo raíz es Astigmatismo debemos eliminar este atributo de la lista de atributos, quedando `Lista_atributos = {Edad, Anomalía}`.

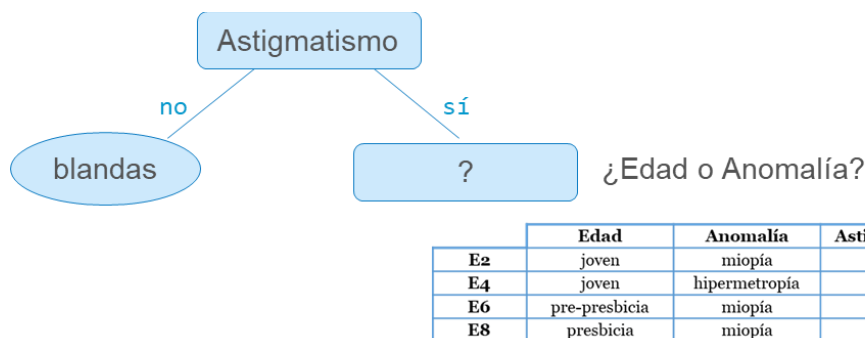
Asignatura	
Técnicas de Inteligencia Artificial	Claudia Villalonga Palliser

Partiendo del nodo raíz etiquetado con el atributo Astigmatismo lo primero que se hace es distribuir los ejemplos en base a los posibles valores del atributo Astigmatismo, es decir entre Astigmatismo=no y Astigmatismo=sí. En la rama Astigmatismo=no caen los ejemplos E1, E3, E5, E7 y E9 y en la rama Astigmatismo=sí los ejemplos E2, E4, E6, E8.



En la rama Astigmatismo=no se observa que todos los ejemplos pertenecen a la clase 'blandas' y por lo tanto se crea un nodo hoja etiquetado con dicha clase.

En la rama Astigmatismo=sí el conjunto de ejemplos  $E = \{E2, E4, E6, E8\}$  contiene ejemplos que pertenecen a ambas clases. Por lo tanto debemos crear un nuevo nodo y aplicar el método de selección de atributos para determinar que atributo debe ir en dicho nodo, Edad o Anomalia.



El método de selección de atributos calcula la ganancia de información para cada uno de los dos atributos.

Primero se calcula la entropía del conjunto  $E = \{E2, E4, E6, E8\}$ , que contiene 4 ejemplos de los cuales 1 pertenece a la clase 'blandas' y 3 a la clase 'duras':

$$Entropia(E) = -\frac{1}{4} \log_2 \frac{1}{4} - \frac{3}{4} \log_2 \frac{3}{4} = 0.8113$$

Asignatura	
Técnicas de Inteligencia Artificial	Claudia Villalonga Palliser

Para calcular la ganancia de información del atributo Edad necesitamos calcular las entropías resultantes de particionar el conjunto de ejemplos (E) respecto a los posibles valores de este atributo: joven, pre-presbicia, presbicia.

Cuando el atributo Edad toma el valor joven tenemos un subconjunto de ejemplos con dos elementos.

	Edad	Anomalía	Astigmatismo	Lentes de contacto
<b>E2</b>	joven	miopía	sí	blandas
<b>E4</b>	joven	hipermetropía	sí	duras

Este subconjunto  $E_{\text{joven}} = \{E2, E4\}$  contiene 1 ejemplo de la clase 'blandas' y 1 ejemplo de la clase 'duras' por lo tanto tiene heterogeneidad total y su entropía es 1.

$$Entropia(E_{\text{joven}}) = 1$$

Cuando el atributo Edad toma el valor pre-presbicia tenemos un subconjunto de ejemplos con un único elemento.

	Edad	Anomalía	Astigmatismo	Lentes de contacto
<b>E6</b>	pre-presbicia	miopía	sí	duras

Este subconjunto  $E_{\text{pre-presbicia}} = \{E6\}$  es homogéneo y su entropía es 0.

$$Entropia(E_{\text{pre-presbicia}}) = 0$$

Cuando el atributo Edad toma el valor presbicia tenemos un subconjunto de ejemplos con un único elemento.

	Edad	Anomalía	Astigmatismo	Lentes de contacto
<b>E8</b>	presbicia	miopía	sí	duras

Este subconjunto  $E_{\text{presbicia}} = \{E8\}$  es homogéneo y su entropía es 0.

$$Entropia(E_{\text{presbicia}}) = 0$$

Asignatura	
Técnicas de Inteligencia Artificial	Claudia Villalonga Palliser

Al haber calculado ya las entropías resultantes de particionar el conjunto de ejemplos (E) respecto a los posibles valores del atributo Edad se puede calcular la ganancia de información de este atributo:

$$\begin{aligned}
 \text{Ganancia}(E, \text{Edad}) \\
 &= \text{Entropia}(E) - \frac{2}{4} \text{Entropia}(E_{\text{joven}}) - \frac{1}{4} \text{Entropia}(E_{\text{pre-presbicia}}) \\
 &\quad - \frac{1}{4} \text{Entropia}(E_{\text{presbicia}}) = 0.3113
 \end{aligned}$$

Procedemos de la misma forma para calcular la ganancia de información del atributo Anomalía.

Cuando el atributo Anomalía toma el valor miopía tenemos un subconjunto de ejemplos con tres elementos.

	Edad	Anomalía	Astigmatismo	Lentes de contacto
<b>E2</b>	joven	miopía	sí	blandas
<b>E6</b>	pre-presbicia	miopía	sí	duras
<b>E8</b>	presbicia	miopía	sí	duras

Este subconjunto  $E_{\text{miopía}} = \{E2, E6, E8\}$  contiene 1 ejemplo de la clase 'blandas' y 2 ejemplos de la clase 'duras' y su entropía se calcula de la siguiente forma:

$$\text{Entropia}(E_{\text{miopía}}) = -\frac{1}{3} \log_2 \frac{1}{3} - \frac{2}{3} \log_2 \frac{2}{3} = 0.9183$$

Cuando el atributo Anomalía toma el valor hipermetropía tenemos un subconjunto de ejemplos con un único elemento.

	Edad	Anomalía	Astigmatismo	Lentes de contacto
<b>E4</b>	joven	hipermetropía	sí	duras

Este subconjunto  $E_{\text{hipermetropía}} = \{E4\}$  es homogéneo y su entropía es 0.

$$\text{Entropia}(E_{\text{hipermetropía}}) = 0$$

Asignatura	
Técnicas de Inteligencia Artificial	Claudia Villalonga Palliser

Al haber calculado ya las entropías resultantes de particionar el conjunto de ejemplos (E) respecto a los posibles valores del atributo Anomalía se puede calcular la ganancia de información de este atributo:

$$\begin{aligned}
 \text{Ganancia}(E, \text{Anomalía}) \\
 &= \text{Entropía}(E) - \frac{3}{4} \text{Entropía}(E_{\text{miopía}}) - \frac{1}{4} \text{Entropía}(E_{\text{hipermetropía}}) \\
 &= 0.1226
 \end{aligned}$$

Una vez calculadas las ganancias de información para los dos atributos  $\text{Ganancia}(E, \text{Edad}) = 0.3113$  y  $\text{Ganancia}(E, \text{Anomalía}) = 0.1226$  el método de selección de atributos decide quedarse con el atributo que tiene mayor ganancia, en este caso el atributo Edad.

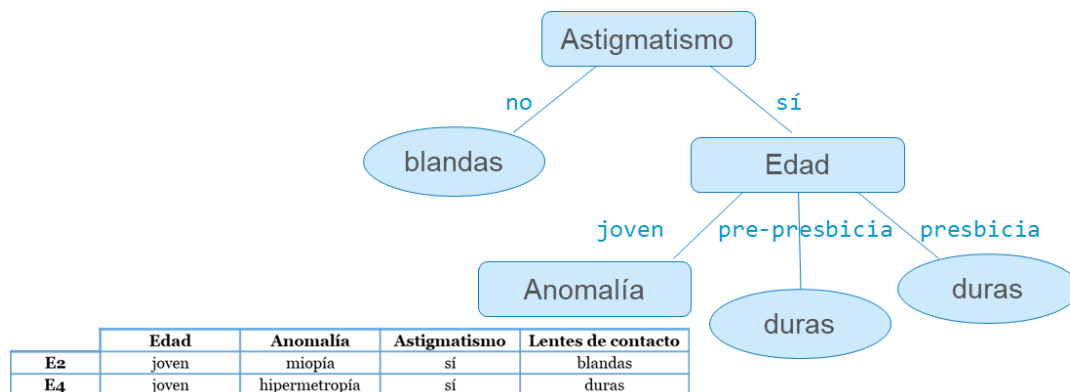
Por lo tanto se etiqueta el nodo con el atributo Edad, se elimina este atributo de la lista de atributos, quedando `Lista_atributos = {Anomalía}` y se distribuyen los ejemplos en base a los posibles valores del atributo Edad.

En la rama Edad=pre-presbicia cae sólo el ejemplo E6 que pertenece a la clase 'duras' y por lo tanto se crea un nodo hoja etiquetado con dicha clase.

En la rama Edad=presbicia también cae sólo el ejemplo E8 que pertenece a la clase 'duras' y por lo tanto se crea un nodo hoja que también se etiqueta con dicha clase.

En la rama Edad=joven caen los ejemplos E2, E4 que pertenecen uno a cada clase. Por lo tanto debemos crear un nuevo nodo y aplicar el método de selección de atributos para determinar que atributo debe ir en dicho nodo. Sin embargo comprobamos que ya sólo queda un posible atributo, Anomalía, por lo que se etiqueta directamente el nodo con el atributo Anomalía éste se elimina de la lista de atributos, quedando ésta vacía.

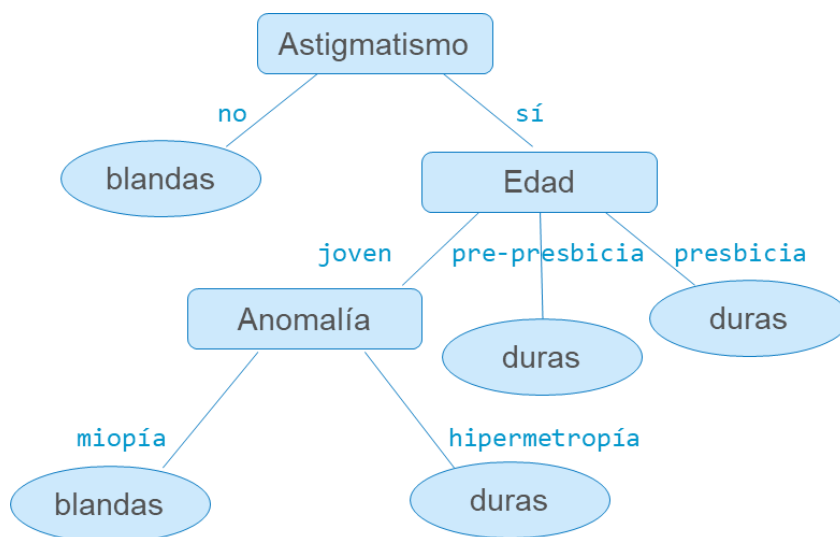
Asignatura	
Técnicas de Inteligencia Artificial	Claudia Villalonga Palliser



Ahora se distribuyen los ejemplos en base a los posibles valores del atributo Anomalía.

En la rama Anomalía=miopía cae sólo el ejemplo E2 que pertenece a la clase 'blandas' y por lo tanto se crea un nodo hoja etiquetado con dicha clase.

En la rama Anomalía=hipermetropía también cae sólo el ejemplo E4 que pertenece a la clase 'duras' y por lo tanto se crea un nodo hoja que también se etiqueta con dicha clase.



Hemos construido un árbol de decisión donde todas las ramas acaban en un nodo hoja por lo tanto se finaliza con la ejecución del algoritmo ID3.