Relación 5 - Aprendizaje Automático

Cuestiones

Cuestión 1. Definimos el tamaño de un árbol como el número de sus nodos, incluida las hojas. Dar dos árboles de tamaño distinto que devuelvan siempre la misma salida al clasificar una instancia.

Cuestión 2. En un problema de aprendizaje sobre el que vamos a aplicar el algoritmo ID3 se considera un conjunto de entrenamiento D con N ejemplos. Supongamos que hay un atributo Atr_1 que puede tomar N valores y que cada uno de los ejemplos de D toma en Atr_1 un valor distinto. Calcular $Ganancia(D, Atr_1)$.

Cuestión 3. Responder razonadamente si las siguientes afirmaciones son verdaderas o falsas. Si la respuesta **Verdadera** debes dar razones que apoyen tu decisión. Si la respuesta es **Falsa** debes dar un ejemplo en el que no se verifique la afirmación.

(a) Sea $D_1 = \{e_1, \ldots, e_n\}$ un conjunto de entrenamiento y sea D_2 el conjunto de entrenamiento formado a partir de D_1 donde cada ejemplo se considera dos veces, esto es, $D_2 = \{e_1, \ldots, e_n, e_{n+1}, \ldots, e_{2n}\}$ donde $\forall i \in \{1, \ldots, n\}\, e_i = e_{i+n}$. Afirmación: *El árbol de decisión obtenido mediante el algoritmo ID3 a partir de D_1 y D_2 son el mismo.*

(b) Tenemos un conjunto con $4n$ ejemplos y lo dividimos en dos partes: un conjunto de entrenamiento con $3n$ ejemplos y un conjunto de prueba con n ejemplos. En el conjunto de entrenamiento $2n$ ejemplos son positivos y n ejemplos son negativos. En el conjunto de prueba, los n ejemplos que hay son todos positivos. Construimos el árbol de decisión A a partir del conjunto de entrenamiento mediante el algoritmo ID3. Afirmación: *El árbol A contiene al menos un nodo tal que si podamos en ese nodo mejoramos el rendimiento respecto al conjunto de prueba.*

(c) Sean D_1 y D_2 dos conjuntos de entrenamiento para el mismo concepto. Entonces

$$\frac{Ent(D_1) + Ent(D_2)}{2} \leq Ent(D_1 \cup D_2)$$

(d) Sean D_1 y D_2 dos conjuntos de entrenamiento para el mismo concepto tales que $D_1 \cap D_2 \neq \emptyset$. Entonces $Ent(D_1 \cup D_2) < Ent(D_1) + Ent(D_2)$.

(e) Si sustituimos el log_2 de la definición de entropía por log_b siendo b cualquier número positivo, entonces el algoritmo ID3 devuelve el mismo árbol.

Cuestión 4. La ganancia de información en el algoritmo ID3 sólo sirve para hacer árboles más pequeños. Si aplicamos el algoritmo de árboles de decisión sobre un mismo conjunto de entrenamiento dos veces, la primera utilizando la ganacia de información para elegir el mejor atributo y la segunda vez utilizando otro criterio, entonces los dos árboles obtenidos pueden tener distinto tamaño, pero representan la misma hipótesis. **¿Verdadero o falso?**

Cuestión 5. Sean $C_1 = \{x_1, \ldots, x_n\} \subset \mathbb{R}^d$ y $C_2 = \{y_1, \ldots, y_n\} \subset \mathbb{R}^d$ dos *clusters* obtenidos después de aplicar el algoritmo de k-medias con $k = 2$ y la distancia euclídea. Sea m_1 el centro de C_1 y m_2 el centro de C_2.

Responder razonadamente si la siguiente afirmación es verdadera o falsa. Si la respuesta **Verdadera** debes dar razones que apoyen tu decisión. Si la respuesta es **Falsa** debes dar un ejemplo en el que no se verifique la afirmación.

La afirmación es:

PARA TODO $i \in \{1, \ldots, n\}$, $d(x_i, m_2) > d(y_i, m_2)$, SIENDO d LA DISTANCIA EUCLÍDEA.

Cuestión 6. Verdadero o Falso. Consideremos un conjunto de N datos $\{x_i\}_{i=1}^{i=N}$ y la distancia euclídea como función distancia. Si $k = N - 1$ y $\forall i \in \{1, \ldots, N\}\,(x_i \in \mathbb{R})$, entonces la salida del algoritmo de k-medias es siempre la misma cualesquiera que sean los k valores que se tomen como centros iniciales.

Si la respuesta es **Verdadero** tienes que dar razones que lo justifiquen. Si la respuesta es **Falso**, tienes que dar un ejemplo en el que esto no ocurre.

Problemas

Tabla de Entropías $Ent(X, Y)$

		Y									
		0	1	2	3	4	5	6	7	8	9
X	0	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000
	1	0.000	1.000	0.918	0.811	0.722	0.650	0.592	0.544	0.503	0.469
	2	0.000	0.918	1.000	0.971	0.918	0.863	0.811	0.764	0.722	0.684
	3	0.000	0.811	0.971	1.000	0.985	0.954	0.918	0.881	0.845	0.811
	4	0.000	0.722	0.918	0.985	1.000	0.991	0.971	0.946	0.918	0.890
	5	0.000	0.650	0.863	0.954	0.991	1.000	0.994	0.980	0.961	0.940
	6	0.000	0.592	0.811	0.918	0.971	0.994	1.000	0.996	0.985	0.971
	7	0.000	0.544	0.764	0.881	0.946	0.980	0.996	1.000	0.997	0.989
	8	0.000	0.503	0.722	0.845	0.918	0.961	0.985	0.997	1.000	0.998
	9	0.000	0.469	0.684	0.811	0.890	0.940	0.971	0.989	0.998	1.000

Ejercicio 1. Una empresa mantiene el historial sobre viajes ofrecidos a un cliente. El cliente a veces compra la oferta y a veces no. La siguiente tabla muestra las últimas ofertas de la empresa y si el cliente compró o no. Las ofertas dependen de los atributos *Continente, Lugar, Actividad* y *Precio*.

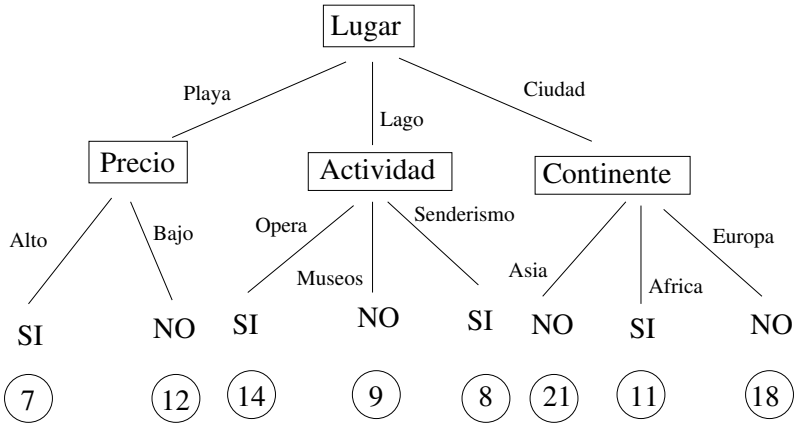
Ej.	CONTINENTE	LUGAR	ACTIVIDAD	PRECIO	*Compra*
1	Europa	Playa	Museos	Alto	Sí
2	Africa	Lago	Museos	Bajo	Sí
3	Africa	Ciudad	Opera	Bajo	Sí
4	Africa	Ciudad	Opera	Bajo	Sí
5	Africa	Playa	Museos	Bajo	No
6	Europa	Playa	Opera	Bajo	No
7	Asia	Lago	Senderismo	Bajo	No
8	Asia	Playa	Senderismo	Bajo	No
9	Europa	Playa	Opera	Alto	No
10	Africa	Playa	Opera	Bajo	No

(a) Calcula la ganancia de información que se obtendría si hiciéramos una partición del conjunto de entrenamiento mediante el atributo LUGAR.

(b) ¿Cuál sería la primera condición que deberíamos considerar para crear una regla que tuviera como conclusión COMPRA = NO usando el Algoritmo de Cobertura? En caso de igualdad, determina todas las condiciones posibles.

(c) Aplicar el clasificador *Naive Bayes* con *suavizado aditivo* usando $k = 1$ para clasificar la instancia *(Asia, Ciudad, Opera, Alto)*

(c) Consideremos ahora un problema distinto sobre los mismos datos. Supongamos ahora que hemos dividido un conjunto de 100 ejemplos en dos conjuntos. El primero, con 90 ejemplos lo hemos usado para crear un árbol de decisión y el segundo *Prueba*, con 10 ejemplos, es el que aparece en la tabla anterior. Supongamos que el árbol obtenido tras aplicar el algoritmo ID3 utilizando el conjunto de 90 ejemplos es

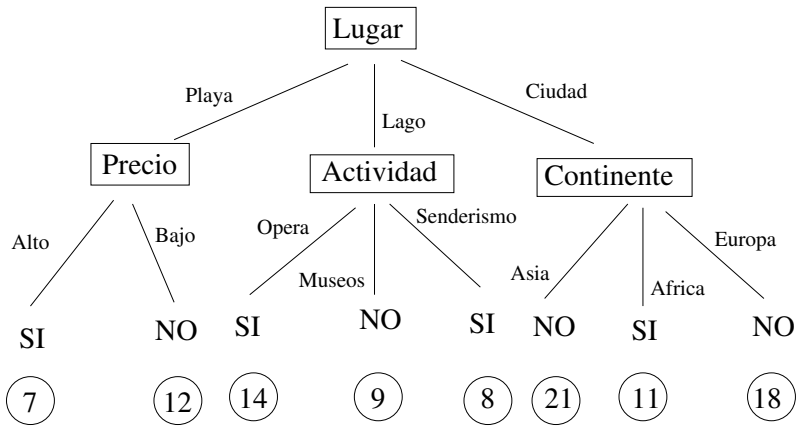


Junto a la clasificación de cada hoja aparece el número de elementos del conjunto D que verifica la condición. Se pide usar el ALGORITMO DE PODA PARA REDUCIR EL ERROR sobre el árbol usando el conjunto anterior como *Prueba*. **Aplicar el algoritmo hasta realizar la primera poda (si es necesario podar) Especificar claramente cuál es el árbol obtenido**.

Ejercicio 2. Una empresa mantiene el historial de viajes ofrecidos a un cliente. El cliente a veces compra la oferta y a veces no. Las ofertas dependen de los atributos *Continente, Lugar, Actividad* y *Precio*.

Ej.	CONTINENTE	LUGAR	ACTIVIDAD	PRECIO	*Compra*
1	Europa	Playa	Museos	Alto	Sí
2	Africa	Lago	Museos	Bajo	Sí
3	Africa	Ciudad	Opera	Bajo	Sí
4	Africa	Ciudad	Opera	Alto	Sí
5	Africa	Playa	Museos	Bajo	No
6	Europa	Playa	Opera	Bajo	No
7	Asia	Lago	Senderismo	Bajo	No
8	Asia	Playa	Senderismo	Bajo	No
9	Europa	Playa	Opera	Alto	No
10	Africa	Playa	Opera	Bajo	No

Supongamos que hemos dividido un conjunto de 100 ejemplos en dos conjuntos. El primero, con 90 ejemplos lo hemos usado para crear un árbol de decisión y el segundo *Prueba*, con 10 ejemplos, es el que aparece en la tabla anterior. Supongamos que el árbol obtenido tras aplicar el algoritmo ID3 utilizando el conjunto de 90 ejemplos es



Junto a la clasificación de cada hoja aparece el número de elementos del conjunto D que verifica la condición. Se pide usar el ALGORITMO DE PODA PARA REDUCIR EL ERROR sobre el árbol usando el conjunto anterior como *Prueba*. **Aplicar el algoritmo hasta realizar la primera poda (si es ecesario podar) Especificar claramente cuál es el árbol obtenido**.

Ejercicio 3. Construye el árbol de decisión aplicando el algoritmo *ID3 con valores continuos* a partir del siguiente conjunto de entrenamiento

Ej	Dolor	Temp	Clasif
Ej_1	SI	16	SI
Ej_2	SI	12	NO
Ej_3	SI	13	NO
Ej_4	SI	19	NO
Ej_5	SI	14	SI
Ej_6	NO	15	SI
Ej_7	NO	20	NO
Ej_8	NO	21	NO
Ej_9	NO	17	SI
Ej_{10}	NO	18	NO

Ejercicio 4. Considera el siguiente conjunto de entrenamiento

Ej	Atrib_1	Atrib_2	Clasif
Ej_1	3	4	SI
Ej_2	9	40	NO
Ej_3	5	12	NO
Ej_4	7	24	NO
Ej_5	8	15	SI
Ej_6	11	60	SI
Ej_7	12	35	NO
Ej_8	13	84	NO

Aplica el algoritmo k-NN **con rechazo** con $k = 7$, umbral $\mu = 5$ y distancia euclídea para clasificar $P = (0, 0)$ explicando los pasos realizados e indicando **explícitamente** la salida del algoritmo.

Ejercicio 5. Considera el siguiente conjunto de entrenamiento D y la distancia euclídea

	At_1	At_2	$Clas$
Ej_1	-10	0	1
Ej_2	0	0	1
Ej_3	0	24	0
Ej_4	7	0	1
Ej_5	70	0	0

(a) Sea $I = (18, 0)$ un ejemplo correspondiente al mismo problema de aprendizaje. Calcula la clasificación de I usando el conjunto de entrenamiento D mediante el método k-NN **con pesos**, con $k = 3$.

(b) Considera ahora el mismo conjunto de entrenamiento *sin la columna de clasificación*. Aplica el algoritmo de K-MEDIAS hasta la primera modificación de los pesos, tomando como pesos iniciales $m_1 = (0, -24)$ y $m_2 = (70, 0)$

Ejercicio 6. Considera el siguente conjunto de puntos sobre la recta real $D = \{1, 3, 7, 12, 13\}$. Aplica *explicando cada uno de los pasos realizados*, el algoritmo de CLUSTERING JERÁRQUICO AGLOMERATIVO. Debes indicar *explícitamente* cuáles son los *clusters* considerados en cada uno de los pasos. Tomaremos como la distancia entre dos *clusters* C y D como la mínima distancia euclídea entre un punto de C y un punto de D.

Ejercicio 7. La siguiente tabla muestra 14 ejemplos de juguetes de madera. Consideraremos esa tabla como conjunto de entrenamiento D para el concepto *Pertenece a Juan*

Ej	Forma	Tamaño	Color	Superficie	Pertenece a Juan
E1	Círculo	Grande	Rojo	Lisa	Sí
E2	Triángulo	Medio	Rojo	Lisa	Sí
E3	Triángulo	Pequeño	Verde	Lisa	Sí
E4	Círculo	Pequeño	Verde	Rugosa	Sí
E5	Cuadrado	Pequeño	Verde	Lisa	Sí
E6	Triángulo	Medio	Verde	Lisa	Sí
E7	Cuadrado	Medio	Verde	Rugosa	Sí
E8	Círculo	Medio	Rojo	Rugosa	Sí
E9	Círculo	Grande	Verde	Lisa	Sí
E10	Cuadrado	Grande	Rojo	Lisa	No
E11	Cuadrado	Grande	Rojo	Rugosa	No
E12	Triángulo	Pequeño	Verde	Rugosa	No
E13	Cuadrado	Medio	Rojo	Lisa	No
E14	Triángulo	Medio	Rojo	Rugosa	No

(a) Construye el árbol de decisión mediante el algoritmo ID3 a partir del conjunto de entrenamiento D.

(b) Aplica el ALGORITMO DE PODA al árbol generado en el apartado anterior usando el siguiente conjunto de *Validación*

Ej	Forma	Tamaño	Color	Superficie	Pertenece a Juan
P1	Triángulo	Medio	Rojo	Rugosa	No
P2	Triángulo	Pequeño	Rojo	Lisa	Sí
P3	Círculo	Pequeño	Rojo	Rugosa	Sí
P4	Cuadrado	Pequeño	Verde	Rugosa	No
P5	Cuadrado	Medio	Rojo	Rugosa	No

Ejercicio 8. La siguiente tabla muestra los datos de 15 pacientes de los que conocemos si padecen o no diabetes, su temperatura y grado de glucemia en ayunas. Además conocmeos la recomendación del especialista para saber si en esas condiciones es necesario suministrar la medicina.

Paciente	DIABETES	TEMPERATURA	GLUCEMIA	MEDICINA
P1	SI	36.2	95	SI
P2	SI	36.4	73	SI
P3	SI	35.9	95	SI
P4	SI	38.8	97	SI
P5	SI	36.4	95	SI
P6	SI	38.3	96	SI
P7	NO	36.0	70	SI
P8	NO	36.5	71	SI
P9	NO	36.4	65	SI
P10	NO	38.3	95	SI
P11	NO	38.9	99	SI
P12	NO	36.6	80	NO
P13	NO	38.0	71	NO
P14	SI	36.7	75	NO
P15	SI	38.1	97	NO

Contruye un árbol de decisión a partir de estos datos para decidir si a un paciente hay que suministrarle la medicina o no. Ten en cuenta que los atributos Temperatura y Glucemia pueden tomar cualquier valor real.

Ejercicio 9. Considera el siguiente conjunto de entrenamiento

$Ej.$	Atr_1	Atr_2	Atr_3	$Clasif.$
E_1	1	1	0	SI
E_2	1	0	0	SI
E_3	1	0	1	SI
E_4	0	0	1	NO

(a) Aplicar el algoritmo k-NN con $k = 1$ para clasificar $P = (0{,}75, 0, 0)$ a partir del conjunto de entrenamiento anterior usando la distancia Manhattan.

(b) Considera ahora el conjunto de entrenamiento anterior sin la columna de clasificación y sean $m_1 = (1, 1, -1)$ y $m_2 = (0, -1, 1)$. Se pide aplicar el algoritmo de k-medias (k=2) con m_1 y m_2 como centros iniciales *hasta la primera modificación de los centros.*

Ejercicio 10. La siguiente tabla muestra la clasificación de la existencia de riesgo sanitario de 10 localizaciones turísticas con diferentes características. Tomaremos esa tabla como conjunto de entrenamiento D para el concepto *Riesgo Sanitario*.

Ej.	LUGAR	HOSPITAL	PAÍS	Riesgo S.
1	Playa	Cerca	Mozambique	SI
2	Playa	Cerca	Tanzania	SI
3	Ciudad	Cerca	Tanzania	SI
4	Playa	Lejos	Mozambique	SI
5	Playa	Lejos	Mozambique	SI
6	Ciudad	Lejos	Mozambique	NO
7	Campo	Cerca	Tanzania	NO
8	Campo	Cerca	Mozambique	NO
9	Ciudad	Lejos	Tanzania	NO
10	Campo	Cerca	Mozambique	NO

- Usar el clasificador *Naive Bayes* para clasificar la siguiente instancia

  (Ciudad, Cerca, Mozambique)

- Construye el árbol de decisión mediante el algoritmo ID3 y úsalo para clasificar la instancia (CIUDAD, CERCA, MOZAMBIQUE)

Ejercicio 11. Considera los puntos $P_1 = (0, 48)$, $P_2 = (0, 78)$, $P_3 = (36, 126)$, $P_4 = (36, 0)$ y los centros $m_1 = (20, 63)$ y $m_2 = (36, 63)$. Se pide aplicar el algoritmo de k-medias sobre los puntos $P_1, \ldots, P_4$ tomando m_1 y m_2 como centros iniciales ($k = 2$) *hasta la primera modificación de los centros.* Usar la distancia euclídea.

Ejercicio 12. Unos biólogos que exploraban la selva del Amazonas han descubierto una nueva especie de insectos, que bautizaron con el nombre de *lepistos*. Desgraciadamente, han desaparecido y la única información que disponemos del nuevo insecto viene dada por el siguiente conjunto de ejemplos encontrados en un cuaderno de notas, en los que se clasifican una serie de muestras de individuos en función de ciertos parámetros como su COLOR, el tener ALAS, su TAMAÑO y su VELOCIDAD:

Ejemplo	COLOR	ALAS	TAMAÑO	VELOCIDAD	LEPISTO
E_1	negro	si	pequeño	alta	si
E_2	amarillo	no	grande	media	no
E_3	amarillo	no	grande	baja	no
E_4	blanco	si	medio	alta	si
E_5	negro	no	medio	alta	no
E_6	rojo	si	pequeño	alta	si
E_7	rojo	si	pequeño	baja	no
E_8	negro	no	medio	media	no
E_9	negro	si	pequeño	media	no
E_{10}	amarillo	si	grande	media	no

Contestar a las siguientes cuestiones:

- ¿Cuál es la entropía del conjunto de ejemplos, respecto a la clasificación de los mismos que realiza el atributo LEPISTO?

- ¿Qué atributo proporciona mayor ganancia de información?
- Aplicar (detallando cada uno de los pasos realizados) el **algoritmo ID3** para encontrar, a partir de este conjunto de entrenamiento, un árbol que nos permita decidir si un determinado individuo es un lepisto o no.
- Obtener un conjunto de reglas a partir del árbol obtenido en el apartado anterior.
- Según el concepto aprendido, ¿hay algún atributo que sea irrelevante para decidir si un individuo es un lepisto?

Ejercicio 13. Una entidad bancaria concede un préstamo a un cliente en función de una serie de parámetros: su EDAD (puede ser `joven`, `mediano` o `mayor`), sus INGRESOS (`altos`, `medios` o `bajos`), un INFORME sobre su actividad financiera (que puede ser `positivo` o `negativo`) y, finalmente, si tiene OTRO PRÉSTAMO a su cargo o no. La siguiente tabla presenta una serie de ejemplos en los que se especifica la concesión o no del préstamo en función de estos parámetros:

Ejemplo	EDAD	INGRESOS	INFORME	OTRO PRÉSTAMO	CONCEDER
E_1	joven	altos	negativo	no	no
E_2	joven	altos	negativo	si	no
E_3	mediano	altos	negativo	no	si
E_4	mayor	medios	negativo	no	si
E_5	mayor	bajos	positivo	no	si
E_6	mayor	bajos	positivo	si	no
E_7	mediano	bajos	positivo	si	si
E_8	joven	medios	negativo	no	no
E_9	joven	bajos	positivo	si	si
E_{10}	mayor	medios	positivo	no	si
E_{11}	joven	medios	positivo	si	si
E_{12}	mediano	medios	negativo	si	si
E_{13}	mediano	altos	positivo	no	si
E_{14}	mayor	medios	negativo	si	no

Supongamos que modificamos el **algoritmo ID3** de manera que el criterio para obtener el "mejor" atributo que clasifica un conjunto de ejemplos es el de *menor* ganancia de información. En esta situación se pide:

- En caso de ausencia de ruido, ¿obtendría el algoritmo modificado un árbol de decisión consistente con los ejemplos del conjunto de entrenamiento?, justifica la respuesta.
- ¿Qué sesgo tendría el algoritmo modificado?, justifica la respuesta.
- Aplicar (detallando cada uno de los pasos realizados) el algoritmo modificado para encontrar, a partir de este conjunto de entrenamiento, un árbol que nos permita decidir sobre la concesión de préstamos.

Ejercicio 14. Aplicar el **algoritmo ID3** para construir un árbol de decisión consistente con los siguientes ejemplos, que nos ayude a decidir si comprar o no un CD nuevo.

Ejemplo	Cantante	Discográfica	Género	Precio	Tienda	Comprar
E_1	Queen	Emi	rock	30	Mixup	si
E_2	Mozart	Emi	clásico	40	Virgin	no
E_3	Anastacia	Corazón	soul	20	Virgin	si
E_4	Queen	Sony	rock	20	Virgin	si
E_5	Anastacia	Corazón	soul	30	Mixup	si
E_6	Queen	Sony	rock	30	Virgin	si
E_7	Wagner	Sony	clásico	30	Mixup	no
E_8	Anastacia	Corazón	soul	30	Virgin	no
E_9	Queen	Emi	rock	40	Virgin	no
E_{10}	Mozart	Sony	clásico	40	Mixup	si

Considerar los siguientes ejemplos como conjunto de prueba y obtener la medida de rendimiento del árbol obtenido.

Ejemplo	Cantante	Discográfica	Género	Precio	Tienda	Comprar
E_{11}	Queen	Emi	rock	30	Virgin	si
E_{12}	Anastacia	Corazón	soul	20	Virgin	no
E_{13}	Queen	Sony	rock	20	Virgin	no
E_{14}	Anastacia	Corazón	soul	30	Virgin	no
E_{15}	Queen	Sony	rock	40	Virgin	no
E_{16}	Mozart	Sony	clásico	40	Mixup	si

Ejercicio 15. La siguiente tabla muestra ejemplos de plantas, indicando si sobrevivieron más de un año o no después de ser compradas, en función de su tamaño (`grande`, `medio` o `pequeño`), de su ambiente adecuado (`interior` o `exterior`), de si tienen flores y de la estación en la que se compró.

Ejemplo	Tamaño	Flores	Ambiente	Estación	Sobrevive
E_1	grande	si	interior	verano	no
E_2	grande	si	interior	verano	no
E_3	grande	si	exterior	primavera	no
E_4	grande	si	exterior	invierno	no
E_5	grande	no	interior	otoño	no
E_6	grande	no	exterior	primavera	no
E_7	medio	si	interior	verano	si
E_8	medio	si	interior	verano	si
E_9	medio	no	interior	primavera	si
E_{10}	medio	no	exterior	otoño	no
E_{11}	medio	no	exterior	verano	no
E_{12}	pequeño	si	interior	invierno	no
E_{13}	pequeño	si	exterior	verano	si
E_{14}	pequeño	no	interior	primavera	no
E_{15}	pequeño	no	interior	verano	si
E_{16}	pequeño	no	exterior	otoño	no

1. Vamos a utilizar el **algoritmo de cobertura** para encontrar reglas que nos permitan deducir que cierta planta sí sobrevive más de un año. Detallar los primeros pasos de dicho algoritmo, sólo hasta el momento en que el algoritmo encuentra la segunda regla. ¿El algoritmo pararía en ese momento, o continuaría? (responder explicando el motivo)

2. Aplicar (detallando cada uno de los pasos realizados) el **algoritmo ID3** para encontrar un árbol de decisión consistente con el conjunto de entrenamiento $\{E_1, \ldots, E_{16}\}$ que permita decidir si una planta sobrevivirá más de un año o no después de ser comprada. Suponer que se elige para el nodo raíz el atributo TAMAÑO, y continuar la ejecución del algoritmo a partir de ahí.

3. Consideremos la siguiente tabla de ejemplos como conjunto de validación

Ejemplo	TAMAÑO	FLORES	AMBIENTE	ESTACIÓN	SOBREVIVE
E_{17}	grande	no	exterior	verano	no
E_{18}	medio	no	interior	otoño	si
E_{19}	medio	no	exterior	primavera	no
E_{20}	medio	si	exterior	verano	no
E_{21}	pequeño	si	interior	verano	no
E_{22}	pequeño	si	interior	invierno	no
E_{23}	pequeño	no	interior	verano	no
E_{24}	pequeño	no	exterior	otoño	no

   a) Calcular el rendimiento del árbol de decisión obtenido en el apartado anterior.

   b) Aplicar un proceso de poda sobre dicho árbol.

Ejercicio 16. Una empresa de material deportivo quiere hacer un estudio de mercado para encontrar las características principales de sus potenciales clientes. En una primera fase, las características a estudiar son las siguientes: la EDAD (joven o adulto), ser deportista PROFESIONAL, el nivel de INGRESOS (altos, medios o bajos) y el SEXO. Para ello, se realiza un cuestionario a 21 personas, obteniendo los resultados que se reflejan en la siguiente tabla:

Ejemplo	EDAD	PROFESIONAL	INGRESOS	SEXO	INTERESADO
E_1	joven	si	bajos	hombre	si
E_2	joven	si	altos	hombre	si
E_3	joven	no	altos	mujer	no
E_4	joven	si	bajos	mujer	si
E_5	joven	no	medios	mujer	no
E_6	adulto	si	altos	hombre	no
E_7	adulto	no	altos	mujer	no
E_8	adulto	si	altos	mujer	no
E_9	adulto	no	medios	mujer	no
E_{10}	adulto	si	bajos	mujer	no
E_{11}	adulto	no	medios	mujer	no
E_{12}	adulto	si	medios	hombre	no
E_{13}	adulto	no	altos	hombre	si
E_{14}	joven	si	altos	mujer	si
E_{15}	joven	si	medios	hombre	si
E_{16}	adulto	no	medios	hombre	no
E_{17}	adulto	no	bajos	hombre	no
E_{18}	joven	no	medios	hombre	no
E_{19}	joven	no	bajos	mujer	no
E_{20}	adulto	si	medios	mujer	no
E_{21}	joven	si	medios	mujer	si

Se pide:

- Aplicar el **algoritmo ID3** (desarrollándolo paso a paso) para obtener un árbol de decision que sirva para decidir si un cliente potencial está interesado o no en el producto que ofrece la empresa. Tomar como conjunto de entrenamiento los primeros 15 ejemplos de la tabla.
- Tomando ahora como conjunto de validación los ejemplos del 16 al 21 de la tabla, calcular el rendimiento del árbol de decisión obtenido en el apartado anterior. Usando ese conjunto, aplicar un proceso de **podado a posteriori** sobre el árbol de decisión. Expresar mediante reglas el árbol obtenido tras la poda ¿Qué rendimiento tiene este árbol sobre el conjunto de validación? ¿Y sobre el conjunto de entrenamiento?
- Usando ahora **todos los ejemplos de la tabla** como conjunto de entrenamiento, predecir mediante un clasificador *Naive Bayes* si un hombre joven con ingresos bajos que no es deportista profesional, va a estar interesado en los productos de la empresa.

Ejercicio 17. Se han encontrado una gran cantidad de obras de arte realizadas por dos artistas **A** y **B**, pero sólo para un pequeño número de obras se ha podido asegurar cuál de los dos es el autor.

La siguiente tabla muestra los datos de dichas obras, indicando el autor en función del TIPO de obra (`grabado`, `óleo` o `acuarela`), del LUGAR donde se encontró la obra (`España`, `Portugal` o `Francia`), de su ESTILO (`clásico` o `moderno`), y de si tienen MARCO o no.

Ejemplo	TIPO	LUGAR	ESTILO	MARCO	AUTOR
E_1	grabado	España	clásico	no	B
E_2	grabado	España	moderno	no	B
E_3	grabado	Portugal	moderno	no	B
E_4	grabado	Francia	clásico	si	B
E_5	grabado	Francia	moderno	no	B
E_6	grabado	Francia	moderno	si	B
E_7	óleo	España	clásico	si	A
E_8	óleo	España	clásico	no	A
E_9	óleo	Francia	moderno	no	A
E_{10}	óleo	Portugal	moderno	si	B
E_{11}	óleo	España	moderno	si	B
E_{12}	acuarela	Francia	clásico	no	B
E_{13}	acuarela	España	clásico	si	A
E_{14}	acuarela	Francia	moderno	no	B
E_{15}	acuarela	España	moderno	no	A
E_{16}	acuarela	Portugal	moderno	si	B

1. Aplicar (detallando cada uno de los pasos realizados) el **algoritmo ID3** para encontrar un árbol de decisión consistente con el conjunto de entrenamiento $\{E_1, \ldots, E_{16}\}$ que permita decidir si una obra de arte fue realizada por **A** o por **B**.
2. Consideremos la siguiente tabla de ejemplos como conjunto de validación

Ejemplo	Tipo	Lugar	Estilo	Marco	Autor
E_{17}	grabado	España	moderno	si	A
E_{18}	óleo	Portugal	moderno	no	A
E_{19}	óleo	Francia	moderno	si	B
E_{20}	óleo	España	moderno	no	A
E_{21}	acuarela	España	clásico	no	A
E_{22}	acuarela	Francia	clásico	si	B
E_{23}	acuarela	España	moderno	si	A
E_{24}	acuarela	Portugal	clásico	si	B

a) Calcular el rendimiento del árbol de decisión obtenido en el apartado anterior.

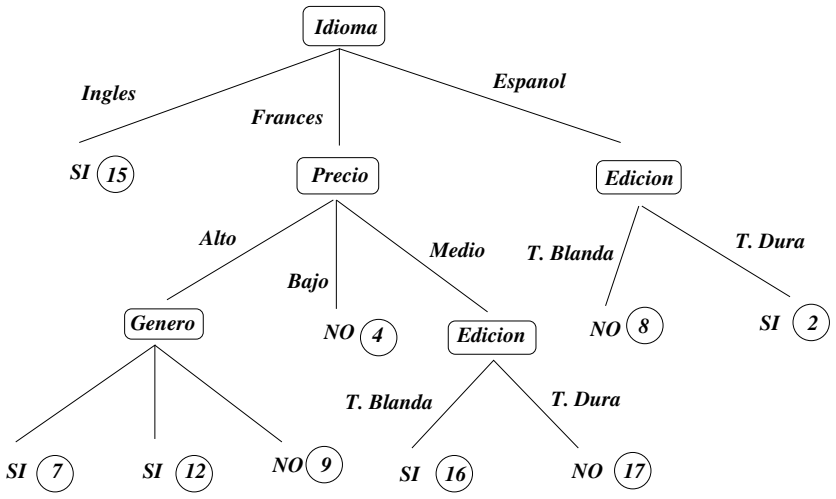
b) Aplicar (detallando los pasos realizados) un proceso de poda sobre dicho árbol.

Ejercicio 18.

Una empresa suministra a un cliente información sobre novelas, las cuales a veces compra y a veces no. Supongamos que hemos dividido un conjunto de 100 ejemplos en dos subconjuntos. El primero, con 90 ejemplos, lo hemos usado para crear un árbol de decisión y el segundo, con 10 ejemplos, es el que aparece en la siguiente tabla, donde se muestran ofertas de la empresa y si el cliente compró o no. Las ofertas dependen de los atributos *Idioma, Género, Precio* y *Edición*.

Ej.	Idioma	Género	Precio	Edición	*Compra*
1	Español	Aventuras	Alto	Tapa dura	Sí
2	Francés	Policíaco	Alto	Tapa blanda	Sí
3	Inglés	Aventuras	Medio	Tapa blanda	No
4	Francés	Histórico	Bajo	Tapa blanda	Sí
5	Francés	Aventuras	Alto	Tapa blanda	No
6	Español	Aventuras	Bajo	Tapa blanda	No
7	Francés	Histórico	Bajo	Tapa blanda	Sí
8	Inglés	Policíaco	Medio	Tapa blanda	No
9	Español	Aventuras	Bajo	Tapa dura	No
10	Francés	Aventuras	Bajo	Tapa blanda	No

Supongamos que el árbol obtenido tras aplicar el algoritmo ID3 utilizando el conjunto de 90 ejemplos es el siguiente:



Junto a la clasificación de cada hoja aparece el número de elementos del conjunto D que verifica la condición, esto es, hay 4 ejemplos con el Idioma *Francés* y Precio *Bajo*, todos ellos con la clasificación `No` y hay 2 ejemplos con Idioma *Español* y Edición en `Tapa Dura`, ambos con la clasificación `Sí`. Se pide usar el ALGORITMO DE PODA PARA REDUCIR EL ERROR sobre el árbol usando como conjunto de validación el mostrado en la tabla anterior.

Ejercicio 19. La siguiente tabla muestra información sobre setas, indicando si son comestibles o no en función de algunas características: el COLOR, el TAMAÑO del pie, la FORMA del sombrero, el ENTORNO en el que se presenta y la forma en que se AGRUPA con otras setas:

Ejemplo	COLOR	TAMAÑO	FORMA	ENTORNO	AGRUPACIÓN	COMESTIBLE
E_1	rojo	mediano	plana	pinar	aislada	si
E_2	blanco	grande	plana	pinar	racimo	si
E_3	rojo	pequeño	plana	pradera	racimo	si
E_4	blanco	pequeño	cóncava	cueva	grupo	si
E_5	marrón	grande	convexa	pinar	grupo	si
E_6	rojo	grande	cóncava	cueva	grupo	no
E_7	marrón	mediano	convexa	pradera	aislada	no
E_8	marrón	mediano	plana	pinar	racimo	no
E_9	blanco	mediano	convexa	pinar	aislada	no
E_{10}	rojo	pequeño	convexa	cueva	grupo	no

Se pide:

- Aplicar (detallando cada uno de los pasos realizados) el **algoritmo de cobertura** para encontrar, a partir de este conjunto de entrenamiento, un conjunto de reglas que nos permita clasificar nuevas instancias. Obtener las reglas para clasificar tanto instancias positivas como negativas. Según lo aprendido ¿hay algún atributo irrelevante para realizar esta clasificación?
- Clasificar las siguientes instancias utilizando el conjunto de reglas aprendido:

	COLOR	TAMAÑO	FORMA	ENTORNO	AGRUPACIÓN
I_1	blanco	pequeño	cóncava	pradera	racimo
I_2	rojo	mediano	plana	cueva	aislada
I_3	marrón	mediano	plana	pinar	racimo
I_4	rojo	mediano	cóncava	pinar	grupo

- Usando todos los ejemplos de la tabla como conjunto de entrenamiento, utilizar un clasificador *Naive Bayes* para clasificar las instancias anteriores.

Ejercicio 20. La siguiente tabla muestra ejemplos de situaciones en las que comprar o no un ordenador, en función de su PRECIO (`alto`, `medio` o `bajo`), su PROCESADOR (`AMD` o `Intel`), si tiene tarjeta ETHERNET y si el monitor es TFT (se supone que el resto de características es común).

Ejemplo	Precio	Ethernet	Procesador	TFT	Comprar
E_1	alto	si	AMD	si	no
E_2	alto	si	AMD	no	no
E_3	alto	si	Intel	si	no
E_4	alto	si	Intel	no	no
E_5	alto	no	AMD	no	no
E_6	alto	no	Intel	no	no
E_7	medio	si	AMD	si	si
E_8	medio	si	AMD	no	si
E_9	medio	no	AMD	si	si
E_{10}	medio	no	Intel	si	no
E_{11}	medio	no	Intel	no	no
E_{12}	bajo	si	AMD	si	no
E_{13}	bajo	si	Intel	no	si
E_{14}	bajo	no	AMD	si	no
E_{15}	bajo	no	AMD	no	si
E_{16}	bajo	no	Intel	si	no

Aplicar (detallando cada uno de los pasos realizados) el **algoritmo de cobertura** para encontrar, a partir de este conjunto de entrenamiento, un conjunto de reglas permita decidir sobre la compra de un ordenador, tanto afirmativa como negativamente. Según las reglas aprendidas ¿deberíamos comprar un ordenador con monitor TFT si el precio es bajo? ¿hay algún atributo irrelevante?

Ejercicio 21. La siguiente tabla muestra información sobre si un alumno aprueba o no la asignatura de IA2, en función de su nota en la asignatura de Lógica Informática (LI), si tiene internet en casa, si usa la bibliografía recomendada y de su afición preferida:

Ejemplo	LI	Internet	Bibliografía	Afición	IA2
E_1	sobresaliente	si	no	cine	si
E_2	aprobado	si	no	música	si
E_3	aprobado	si	si	deporte	si
E_4	sobresaliente	no	si	deporte	si
E_5	notable	si	si	deporte	si
E_6	aprobado	si	si	música	si
E_7	notable	no	si	música	si
E_8	sobresaliente	no	no	música	si
E_9	sobresaliente	si	si	cine	si
E_{10}	notable	si	no	cine	no
E_{11}	notable	no	no	cine	no
E_{12}	aprobado	no	si	deporte	no
E_{13}	notable	si	no	música	no
E_{14}	aprobado	no	no	música	no

Aplicar (detallando cada uno de los pasos realizados) el **algoritmo de cobertura** para encontrar, a partir de este conjunto de entrenamiento, un conjunto de reglas que nos permita decidir si un determinado alumno va a aprobar la asignatura IA2 o no.

Ejercicio 22. Una asociación juvenil de geología propone a sus miembros una excursión a Sierra Mágina (Jaén) para buscar restos de meteoritos. Para distinguirlos de las demás

piedras se tienen en cuenta diferentes factores entre los que se encuentran los siguientes: la presencia de corteza de fusión (el COLOR de su superficie), la DENSIDAD, el MAGNETISMO y la apariencia INTERIOR (`metálica`, `cristalina` o `pétrea`). Una vez recogidas las muestras, el Museo de Geología de Sevilla determina cuáles son restos de meteoritos y cuáles no. Los datos se recogen en la siguiente tabla:

Ejemplo	COLOR	DENSIDAD	MAGNETISMO	INTERIOR	METEORITO
E_1	negro	alta	alto	metálico	si
E_2	blanco	baja	bajo	pétreo	no
E_3	blanco	baja	bajo	cristal	no
E_4	gris	alta	medio	metálico	si
E_5	negro	baja	medio	metálico	no
E_6	marrón	alta	alto	metálico	si
E_7	marrón	alta	alto	cristal	no
E_8	negro	baja	medio	pétreo	no
E_9	negro	alta	alto	pétreo	no
E_{10}	blanco	alta	bajo	pétreo	no

Se pide:

- Aplicar el **algoritmo de cobertura** para obtener un conjunto de reglas que ayuden a decidir si una muestra es un resto de meteorito o no.
- Aplicar el **algoritmo ID3** para obtener un árbol de decisión que igualmente nos ayude a decidir si una muestra es un resto de meteorito o no.
- A la vista de los resultados, ¿existe algún atributo que sea irrelevante a la hora de tomar la decisión?
- Extraer un conjunto de reglas a partir del árbol construido por el algoritmo ID3 ¿es el mismo conjunto de reglas que el obtenido en el primer apartado?

Ejercicio 23. El departamento de Biología Marina quiere analizar las características que hacen que los peces de tamaño inferior a 20 cm sobrevivan en cautividad. Para ello han recogido datos de 20 muestras en las que se indica la adaptación a la CAUTIVIDAD en función del TAMAÑO (`pequeño`, `mediano` o `grande`), la TEMPERATURA del habitat natural (`fría`, `templada` o `cálida`), la SALINIDAD del habitat natural (agua `dulce` o `salada`) y la SOCIABILIDAD de la especie (`solitario`, `pareja` o `grupo`)

Ejemplo	TAMAÑO	TEMPERATURA	SALINIDAD	SOCIABILIDAD	CAUTIVIDAD
E_1	pequeño	fría	dulce	solitario	no
E_2	pequeño	fría	salada	pareja	si
E_3	pequeño	fría	dulce	grupo	no
E_4	pequeño	templada	salada	solitario	si
E_5	pequeño	templada	dulce	grupo	no
E_6	pequeño	cálida	salada	grupo	si
E_7	mediano	fría	dulce	solitario	si
E_8	mediano	fría	salada	pareja	si
E_9	mediano	templada	dulce	grupo	no
E_{10}	mediano	templada	salada	solitario	no
E_{11}	mediano	cálida	dulce	solitario	si
E_{12}	mediano	cálida	salada	grupo	no
E_{13}	mediano	cálida	dulce	solitario	no
E_{14}	grande	fría	salada	pareja	si
E_{15}	grande	fría	dulce	grupo	si
E_{16}	grande	templada	salada	solitario	no
E_{17}	grande	templada	dulce	grupo	si
E_{18}	grande	templada	salada	grupo	si
E_{19}	grande	cálida	dulce	solitario	no
E_{20}	grande	cálida	salada	pareja	no

1. Al aplicar el algoritmo de aprendizaje inductivo **ID3**, ¿cúal es el nodo elegido como raíz del árbol de aprendizaje?

2. A partir del nodo raíz elegido en el apartado anterior, desarrolla todo el subárbol correspondiente al nodo hijo de mayor entropía.

3. Argumenta razonadamente a favor o en contra de la siguiente afirmación: "A mayor entropía mayor es la profundidad del árbol de decisión"

4. Aplicar el **algoritmo de cobertura** para encontrar reglas con el menor número de condiciones posible, que nos permitan deducir si una especie sobrevive en cautividad a partir de sus características. Detallar los pasos de dicho algoritmo hasta el momento en que se completa la primera regla. En este punto, ¿por qué es necesario seguir aplicando el algoritmo de cobertura?

Ejercicio 24. La siguiente tabla muestra una serie de datos acerca de ejemplos de personas que han sufrido quemaduras solares, junto con los datos acerca de su color de PELO, su ALTURA, su PESO y si usaban PROTECCIÓN o no.

Ejemplo	PELO	ALTURA	PESO	PROTECCIÓN	QUEMADURA
E_1	rubio	medio	bajo	no	si
E_2	rubio	alto	medio	si	no
E_3	moreno	bajo	medio	si	no
E_4	rubio	bajo	medio	no	si
E_5	rojo	medio	alto	no	si
E_6	moreno	alto	alto	no	no
E_7	moreno	medio	alto	no	no
E_8	rubio	bajo	bajo	si	no

Aplicar (detallando cada uno de los pasos realizados) el **algoritmo de cobertura** para encontrar, a partir de este conjunto de entrenamiento, un conjunto de reglas que nos permita decidir situaciones en las que se producirá quemadura solar. Según lo aprendido ¿hay algún atributo irrelevante para decidir si se producirá quemadura solar?

Ejercicio 25.

- Describir en pseudocódigo el algoritmo k-medias ¿Para qué se usa? ¿Por qué puede verse como un algoritmo de búsqueda local? ¿Qué se busca y qué se trata de optimizar?
- Considérese el siguiente conjunto de puntos en R^2: $(0, 2)$, $(2, 2)$, $(2, 0)$, $(6, 5)$ y $(7, 2)$, que se desea dividir en dos grupos. Aplicar **un paso** del algoritmo k-medias, suponiendo que los centros de inicio son $(0, 0)$ y $(3, 3)$ ¿Parará el algoritmo tras ese único paso? Justificar la respuesta
- ¿Por qué el algoritmo k-medias puede verse como un algoritmo de búsqueda local? ¿Qué se busca y qué se trata de optimizar? (usado para clustering)

Ejercicio 26.

Una empresa está promocionando tres servicios (**s1**, **s2** y **s3**) y quiere crear una herramienta automática que decida para cada cliente cuál es el producto que se ajusta mejor a su perfil (se supone que cada cliente contratará sólo un servicio). Para ello, los asesores de marketing han elaborado una encuesta con cuatro preguntas de tipo si/no (**A**, **B**, **C** y **D**) de manera que conociendo las respuestas de un cierto cliente ya se dispone de información suficiente para decidir qué producto ofrecerle.

A partir de los datos de ventas anteriores, disponemos de las respuestas de 5000 clientes, y sabemos lo siguiente:

- 1500 contrataron el servicio **s1**, y mostramos sus respuestas en la siguiente tabla:

A=sí	B=sí	C=sí	D=sí
750	1000	500	1350

- 1000 contrataron el servicio **s2**, y mostramos sus respuestas en la siguiente tabla:

A=sí	B=sí	C=sí	D=sí
500	900	750	100

- 2500 contrataron el servicio **s3**, y mostramos sus respuestas en la siguiente tabla:

A=sí	B=sí	C=sí	D=sí
50	2000	1500	500

Se pide:

1. Plantear este problema como un problema de clasificación según un modelo **Naive Bayes**, y dibujar la red bayesiana correspondiente. ¿Qué relaciones de independencia condicional se están suponiendo?

2. Dado un cliente que responde "no" a todas las preguntas, ¿qué servicio le aconsejaría el modelo?

Ejercicio 27. Supongamos que queremos aprender a clasificar un correo como SPAM a partir de tres características X_1, X_2 y X_3 (cada una de ellas puede tomar un valor verdadero o falso). Para ello, disponemos de un conjunto de 1000 correos de los cuales 750 están clasificados como "no SPAM" y 250 como "SPAM". De los "no SPAM", la mitad tienen la característica X_1, la cuarta parte la característica X_2 y 225 tienen la característica X_3. Y de los "SPAM", la cuarta parte tiene característica X_1, la mitad la característica X_2 y 100 tienen la característica X_3. Se pide:

- Asumir un modelo **Naive Bayes** para este problema, y dibujar la red bayesiana correspondiente
- Estimar, según los datos tomados del conjunto de entrenamiento, las tablas de probabilidad de la red anterior ¿Cuál es la propiedad fundamental de esa estimación?
- Dado un nuevo correo que no tiene la característica X_1 pero que sí tiene las otras dos ¿cómo se clasificaría según este modelo **Naive Bayes**?

Ejercicio 28.

Supongamos que queremos aprender a clasificar una solicitud de VPO como "apta" o "no apta" a partir de tres características X_1, X_2 y X_3 que posea el solicitante (cada una de ellas puede tomar un valor verdadero o falso). Para ello, disponemos de un conjunto de 1000 solicitudes de las cuales 800 están clasificadas como "no aptas" y 200 como "aptas". De los "no aptas", la mitad tienen la característica X_1, la cuarta parte la característica X_2 y 500 tienen la característica X_3. Y de los "aptos", la cuarta parte tiene característica X_1, la mitad la característica X_2 y 50 tienen la característica X_3. Se pide:

- Asumir un modelo **Naive Bayes** para este problema, y dibujar la red bayesiana correspondiente
- Estimar, según los datos tomados del conjunto de entrenamiento, las tablas de probabilidad de la red anterior
- Dado un nuevo solicitante que no tiene la característica X_3 pero que sí tiene las otras dos ¿cómo se clasificaría según este modelo **Naive Bayes**?

Ejercicio 29. Considérese el conjunto de entrenamiento en el problema de los "lepistos" (ejercicio de la relación del tema "Aprendizaje Inductivo") y plantearlo como un modelo probabilístico **Naive Bayes**. Calcular las tablas de probabilidad del modelo a partir del conjunto de entrenamiento. Usando el modelo planteado, decidir si un animal amarillo, pequeño, sin alas y con velocidad alta es un lepisto. Compárese el resultado con la clasificación que se obtendría con el árbol de decisión obtenido con ID3 o con el conjunto de reglas obtenido por el algoritmo de cobertura. En el algoritmo ID3 para este problema aparece un atributo irrelevante ¿qué ocurre con este atributo en el modelo probabilístico?

Ejercicio 30. Supongamos que queremos aprender a clasificar un correo como SPAM a partir de tres características X_1, X_2 y X_3 (cada una de ellas puede tomar un valor verdadero o falso). Para ello, disponemos de un conjunto de 1000 correos de los cuales

750 están clasificados como "no SPAM" y 250 como "SPAM". De los "no SPAM", la mitad tienen la característica X_1, la cuarta parte la característica X_2 y 225 tienen la característica X_3. Y de los "SPAM", la cuarta parte tiene característica X_1, la mitad la característica X_2 y 100 tienen la característica X_3. Se pide:

- Asumir un modelo **Naive Bayes** para este problema, y dibujar la red bayesiana correspondiente
- Estimar, según los datos tomados del conjunto de entrenamiento, las tablas de probabilidad de la red anterior ¿Cuál es la propiedad fundamental de esa estimación?
- Dado un nuevo correo que no tiene la característica X_1 pero que sí tiene las otras dos ¿cómo se clasificaría según este modelo **Naive Bayes**?

Ejercicio 31.

Supongamos que queremos aprender a clasificar una solicitud de VPO como "apta" o "no apta" a partir de tres características X_1, X_2 y X_3 que posea el solicitante (cada una de ellas puede tomar un valor verdadero o falso). Para ello, disponemos de un conjunto de 1000 solicitudes de las cuales 800 están clasificadas como "no aptas" y 200 como "aptas". De los "no aptas", la mitad tienen la característica X_1, la cuarta parte la característica X_2 y 500 tienen la característica X_3. Y de los "aptos", la cuarta parte tiene característica X_1, la mitad la característica X_2 y 50 tienen la característica X_3. Se pide:

- Asumir un modelo **Naive Bayes** para este problema, y dibujar la red bayesiana correspondiente
- Estimar, según los datos tomados del conjunto de entrenamiento, las tablas de probabilidad de la red anterior
- Dado un nuevo solicitante que no tiene la característica X_3 pero que sí tiene las otras dos ¿cómo se clasificaría según este modelo **Naive Bayes**?

Ejercicio 32. Considérese el conjunto de entrenamiento en el problema de los "lepistos" (ejercicio de la relación del tema "Aprendizaje Inductivo") y plantearlo como un modelo probabilístico **Naive Bayes**. Calcular las tablas de probabilidad del modelo a partir del conjunto de entrenamiento. Usando el modelo planteado, decidir si un animal amarillo, pequeño, sin alas y con velocidad alta es un lepisto. Compárese el resultado con la clasificación que se obtendría con el árbol de decisión obtenido con ID3 o con el conjunto de reglas obtenido por el algoritmo de cobertura. En el algoritmo ID3 para este problema aparece un atributo irrelevante ¿qué ocurre con este atributo en el modelo probabilístico?

Ejercicio 33. Se sabe que la miopía es un defecto visual que tiene cierta componente hereditaria, ya que tener un padre miope influye, de manera probabilística, en su aparición. Para realizar un modelo probabilístico de tal circunstancia, consideraremos dos variables aleatorias booleanas $Miope$ y $PadreMiope$ que representan, respectivamente, el tener miopía y el tener un progenitor con miopía. Se pide:

- Modelar la situación mediante una red bayesiana

- Supongamos que realizamos un cuestionario a 1000 personas, obteniendo la siguiente información: 400 tenían un progenitor miope, de los cuales 300 eran a su vez miopes; otros 100 eran miopes sin tener progenitor miope. Con estos datos, obtener una estimación de máxima verosimilitud de los parámetros de la red bayesiana del apartado anterior ¿Qué significa exactamente que dichas estimaciones sean de máxima verosimilitud? Demostrar (con el desarrollo matemático correspondiente) que lo son.