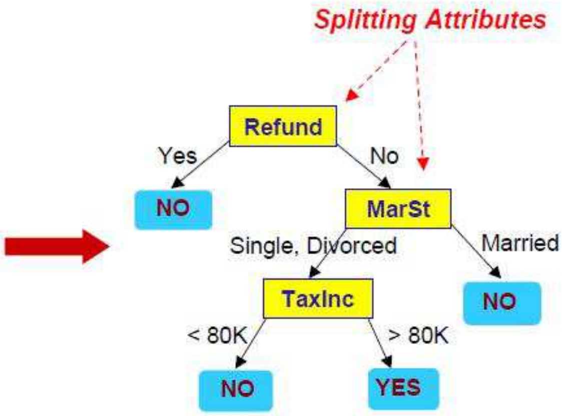
Ejercicio 1.- En la siguiente figura se muestra un árbol de decisión para predecir ejemplos acerca de si la persona miente. En el árbol de decisión, la raíz y los nodos internos contienen atributos con condiciones de prueba para separar los registros que tienen características diferentes. A todo nodo terminal se le asigna la clase Sí o No.
Refund = Devolución, Marital Status: Estado Civil, Taxable Income = Renta gravable en miles, Cheat= Miente. Splitting Attribute: Atributos de división

	categorical	categorical	continuous	class
Tid	Refund	Marital Status	Taxable Income	Cheat
1	Yes	Single	125K	No
2	No	Married	100K	No
3	No	Single	70K	No
4	Yes	Married	120K	No
5	No	Divorced	95K	Yes
6	No	Married	60K	No
7	Yes	Divorced	220K	No
8	No	Single	85K	Yes
9	No	Married	75K	No
10	No	Single	90K	Yes

Training Data



Model: Decision Tree

1) ¿Coincide el árbol de decisión con el construido utilizando la reducción de entropia?

$$H(C)=-\sum_{c}^{n} p(c)\log_2 p(c) = -(7/10)*\log_2(7/10)-(3/10)*\log_2(3/10)=0,8813$$

Refund

$$H(C\,|\,\mathrm{R}) = -\sum_c\sum_r p(\mathrm{r,c})\log_2 p(c\,|\,\mathrm{r}) = \text{-P(NR,NC)}\log_2\text{P(NC|NR)-}$$
$$\text{-P(NR,SC)}\log_2\text{P(SC|NR)-P(SC,NR)}\log_2\text{P(NC|SR)}$$
$$\text{-P(SR,SC)}\log_2\text{P(SC|SR)=-(4/10)}\log_2\text{(4/7)-(3/10)}\log_2\text{(3/7)-}$$
$$\text{-(3/10)}\log_2\text{(3/3)-(0/10)}\log_2\text{(0/3)=0,6896}$$

Marital Status

$$H(C\,|\,\mathrm{MS}) = -\sum_c\sum_r p(\mathrm{ms,c})\log_2 p(c\,|\,\mathrm{ms}) = -P(\mathrm{S},NC)\log_2 P(\mathrm{NC|S}) -$$
$$-P(\mathrm{S,SC})\log_2 P(\mathrm{SC|S}) - P(M,\mathrm{NC})\log_2 P(\mathrm{NC|M})$$
$$-P(M,\mathrm{SC})\log_2 P(SC|\mathrm{M}) - P(D,\mathrm{NC})\log_2 P(\mathrm{NC|D})$$
$$-P(D,\mathrm{SC})\log_2 P(SC|\mathrm{D}) = \text{-(2/10)}\log_2\text{(2/4)-(2/10)}\log_2\text{(2/4)-}$$
$$\text{-(4/10)}\log_2\text{(4/4)-(0/10)}\log_2\text{(0/4) -(1/10)}\log_2\text{(1/2)-(1/10)}\log_2\text{(1/2)=}\mathbf{0,6}$$

Taxable Income

$$H(C\,|\,\mathrm{TI}) = -\sum_c\sum_r p(\mathrm{ti,c})\log_2 p(c\,|\,\mathrm{ti}) = -P(<80,NC)\log_2 P(\mathrm{NC|<80}) -$$
$$-P(<80,\mathrm{SC})\log_2 P(\mathrm{SC|<80}) - P(>80,\mathrm{NC})\log_2 P(\mathrm{NC|>80})$$
$$-P(>80,\mathrm{SC})\log_2 P(SC|\mathrm{>80}) = \text{-(3/10)}\log_2\text{(3/3)-(0/10)}\log_2\text{(0/3)}$$
$$\text{-(4/10)}\log_2\text{(4/7)-(3/10)}\log_2\text{(3/7)=0,6896}$$

El nodo raíz es Marital Status dado que es el que menos entropía condicional tiene. De esta forma la nueva base de datos a analizar es por una parte

Entrenamiento	Refund	Marital Status	Taxable Income	Cheat
1	Yes	Single	>80K	No
3	No	Single	<80K	No
8	No	Single	>80K	Yes
10	No	Single	>80K	Yes

Ahora

$$H(C)=-\sum_{c}^{n}p(c)\log_2 p(c)=-(2/4)*\log_2\left(2/4\right)-(2/4)*\log_2\left(2/4\right)=1$$

Refund

$$\begin{aligned}H(C\,|\,\mathrm{R})&=-\sum_{c}\sum_{r}p(\mathrm{r},\mathrm{c})\log_2 p(c\,|\,\mathrm{r})=-P(NR,NC)\log_2 P(\mathrm{NC|NR})-\\&-P(NR,\mathrm{SC})\log_2 P(\mathrm{SC|NR})-P(\mathrm{SR},\mathrm{NC})\log_2 P(\mathrm{NC|SR})\\&-P(\mathrm{SR},\mathrm{SC})\log_2 P(SC|\mathrm{SR})=\text{-(1/4)}\log_2(1/3)\text{-(2/4)}\log_2(2/3)-\\&\text{-(1/4)}\log_2(1)\text{-(0/4)}\log_2(0/1)=(3/4)\log_2 3-(1/2)\log_2 2\end{aligned}$$

Taxable Income

$$\begin{aligned}H(C\,|\,\mathrm{TI})&=-\sum_{c}\sum_{r}p(\mathrm{ti},\mathrm{c})\log_2 p(c\,|\,\mathrm{ti})=-P(<80,NC)\log_2 P(\mathrm{NC}|<80)-\\&-P(<80,\mathrm{SC})\log_2 P(\mathrm{SC}|<80)-P(>80,\mathrm{NC})\log_2 P(\mathrm{NC}|>80)\\&-P(>80,\mathrm{SC})\log_2 P(SC|>80)=\text{-(1/4)}\log_2(1/1)\text{-(0/4)}\log_2(0/1)\\&\text{-(1/4)}\log_2(1/3)\text{-(2/4)}\log_2(2/3)=(3/4)\log_2 3-(1/2)\log_2 2\end{aligned}$$

Luego el siguiente nodo por la rama de la izquierda puede ser cualquiera de los dos.
Si Consideramos la rama de la derecha del árbol, tenemos

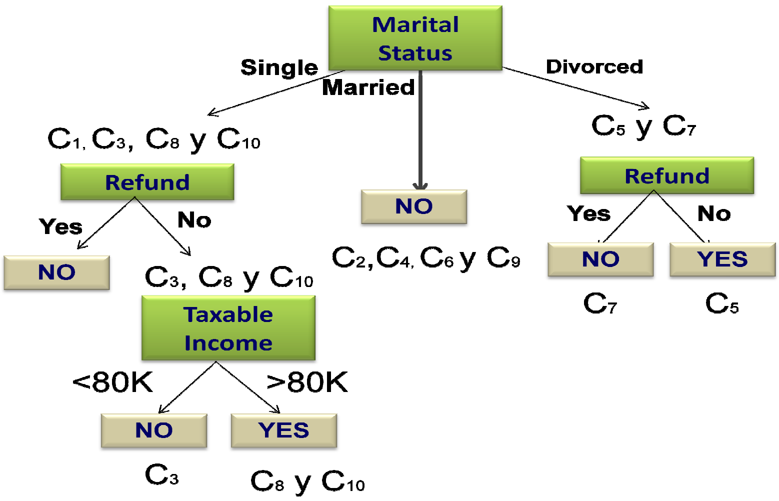
Entrenamiento	Refund	Marital Status	Taxable Income	Cheat
5	No	Divorced	>80K	Yes
7	Yes	Divorced	>80K	No

Ahora $$H(C)=-\sum_{c}^{n}p(c)\log_2 p(c)=-(1/2)*\log_2\left(1/2\right)-(1/2)*\log_2\left(1/2\right)=1$$

$$\begin{aligned}H(C\,|\,\mathrm{R})&=-\sum_{c}\sum_{r}p(\mathrm{r},\mathrm{c})\log_2 p(c\,|\,\mathrm{r})=-P(NR,NC)\log_2 P(\mathrm{NC|NR})-\\&-P(NR,\mathrm{SC})\log_2 P(\mathrm{SC|NR})-P(\mathrm{SR},\mathrm{NC})\log_2 P(\mathrm{NC|SR})\\&-P(\mathrm{SR},\mathrm{SC})\log_2 P(SC|\mathrm{SR})=\text{-(0/2)}\log_2(0/1)\text{-(1/2)}\log_2(1/1)-\\&\text{-(1/2)}\log_2(1/1)\text{-(0/2)}\log_2(0/1)=0\end{aligned}$$

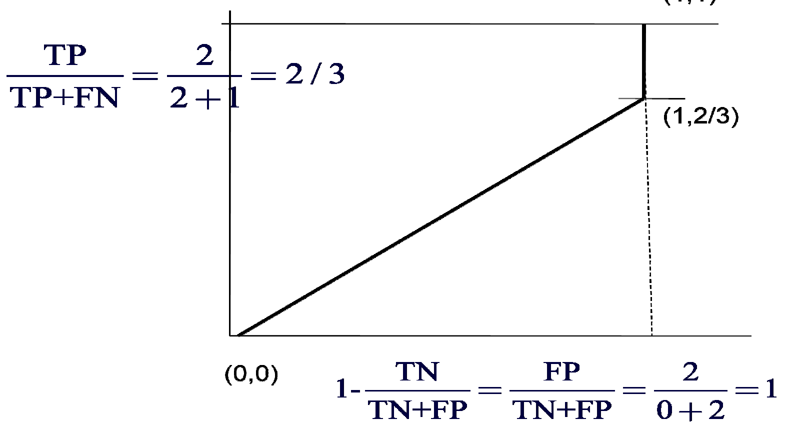
$$\begin{aligned}H(C\,|\,\mathrm{TI})&=-\sum_{c}\sum_{r}p(\mathrm{ti},\mathrm{c})\log_2 p(c\,|\,\mathrm{ti})=-P(<80,NC)\log_2 P(\mathrm{NC}|<80)-\\&-P(<80,\mathrm{SC})\log_2 P(\mathrm{SC}|<80)-P(>80,\mathrm{NC})\log_2 P(\mathrm{NC}|>80)\\&-P(>80,\mathrm{SC})\log_2 P(SC|>80)=\text{-(0/2)}\log_2(0/0)\text{-(0/2)}\log_2(0/0)\\&\text{-(1/2)}\log_2(1/2)\text{-(1/2)}\log_2(1/2)=1\end{aligned}$$

luego en este caso el siguiente nodo es Refund, y el árbol puede ser



Ejercicio 2.- Con el clasificador de la figura, clasificar los patrones del conjunto de test con las siguientes características
P1.- Refund = No, Marital Status: Married, Taxable income = 80K, Cheat=No
P2.- Refund = Si, Marital Status: Married, Taxable income = 180K, Cheat=Si
P3.- Refund = No, Marital Status: Divorced, Taxable income = 90K, Cheat=No
P4.- Refund = No, Marital Status: Married, Taxable income = 50K, Cheat=Si
P5.- Refund = Si, Marital Status: Single, Taxable income = 60K, Cheat=No
3) Construir la Matriz de Confusión y a partir de ella construir el área bajo la curva ROC, AUC. En función del resultado cómo calificaría al clasificador.
Solución.- Los patrones P1, P2, P4 y P5 el clasificador los clasifica en la clase No, a P3 en la clase Si. La matriz de confusión es de la forma.

$$\begin{array}{cc} & \text{C. asignada} \\ \text{C. real} & \begin{pmatrix} 2 & 1 \\ 2 & 0 \end{pmatrix} \end{array} = \begin{pmatrix} TP & FN \\ FP & TN \end{pmatrix}$$



Luego el AUC=2/6=1/3, por lo que el clasificador es peor que un clasificador aleatorio.

Ejercicio 3.- Un médico con experiencia está construyendo un sistema de razonamiento basado en casos para automatizar una tarea de diagnóstico. Los casos se corresponden con personas individuales, donde sus datos se componen de una serie de características que describen los posibles síntomas y la parte de solución representa el diagnóstico (clasificación de la enfermedad). La base de casos contiene 10 casos que se pueden ver en la siguiente tabla.

Entrenamiento	Fiebre	Vómitos	Diarrea	Clasificación
C1	alta	si	si	Into. por Salmonella, (IS)
C2	media	no	si	Into. por Salmonella, (IS)
C3	no	si	si	Infección intestinal, (II)
C4	media	si	no	Infección intestinal, (II)
C5	no	no	si	Infección intestinal, (II)
C6	alta	si	si	Into. por Salmonella, (IS)
C7	media	no	no	Infección intestinal, (II)
C8	alta	si	si	Into. por Salmonella, (IS)
C9	alta	no	si	Into. por Salmonella, (IS)
C10	alta	no	si	Into. por Salmonella, (IS)

Construya el árbol de decisión y clasifique un patrón con valores de las características q = (alta; si; si).

Solución.- La entropía asociada a la clase es

$$H(C) = -\sum_{c}^{n} p(c)\log_2 p(c) = -(6/10)*\log_2\left(6/10\right) - (4/10)*\log_2\left(4/10\right) = 0{,}97$$

Las entropias condicionadas a diferentes características de los pacientes son:

Fiebre

$$\begin{aligned}H(C\,|\,\text{Fiebre}) &= -\sum_{c}\sum_{x} p(x,c)\log_2 p(c\,|\,x) = -P(Alta,\text{IS})\log_2 P(IS|\text{Alta}) - \\ &- P(\text{Alta,II})\log_2 P(II|\text{Alta}) - P(Media,\text{IS})\log_2 P(IS|\text{Media}) - \\ &- P(Media,\text{II})\log_2 P(II|\text{Media}) - P(No,\text{IS})\log_2 P(IS|\text{No}) - \\ &- P(No,\text{II})\log_2 P(II|\text{No}) = \text{-}(5/10)\log_2(5/5)\text{-}(0/10)\log_2(0/5)\text{-} \\ &\text{-}(1/10)\log_2(1/3)\text{-}(2/10)\log_2(2/3)\text{-}(2/10)\log_2(2/2)\text{-}(0/10)\log_2(0/2)=\mathbf{0,27}\end{aligned}$$

Vómitos

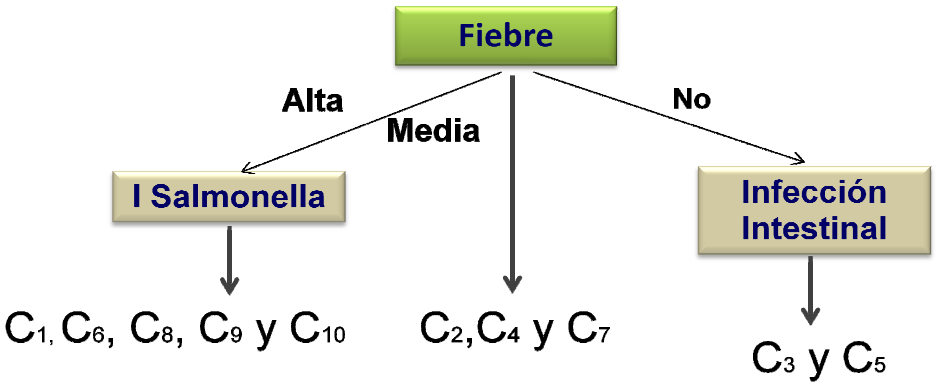
$$\begin{aligned}H(C|\text{Vómitos}) = &\ \text{-}P(SI,IS)\log_2 P(IS|SI)\text{-}P(SI,II)\log_2 P(II|SI) \\ &\text{-}P(NO,IS)\log_2 P(IS|NO)\text{-}P(NO,II)\log_2 P(II|NO)= \\ &\text{-}(3/10)\log_2(3/5)\text{-}(2/10)\log_2(2/5)\text{-}(3/10)\log_2(3/5) \\ &\text{-}(2/10)\log_2(2/5)=0{,}97\end{aligned}$$

Diarrea

$$\begin{aligned}H(C|\text{Diarrea}) = &\ \text{-}P(SI,IS)\log_2 P(IS|SI)\text{-}P(SI,II)\log_2 P(II|SI) \\ &\text{-}P(NO,IS)\log_2 P(IS|NO)\text{-}P(NO,II)\log_2 P(II|NO)= \\ &\text{-}(6/10)\log_2(6/8)\text{-}(2/10)\log_2(2/8)\text{-}(0/10)\log_2(0/2) \\ &\text{-}(2/10)\log_2(2/2)=0{,}66\end{aligned}$$

Entrenamiento	Fiebre	Vómitos	Diarrea	Clasificación
C2	media	no	si	Into. por Salmonella, (IS)

C4	media	si	no	Infección intestinal, (II)
C7	media	no	no	Infección intestinal, (II)



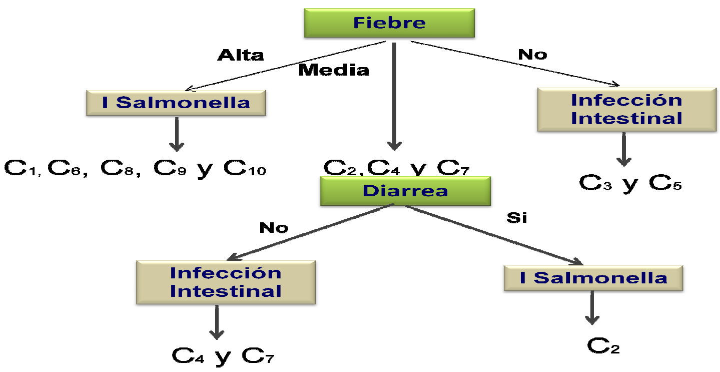
Ahora volvemos a rehacer los cálculos, pero con la nueva estructura de datos

$$H(C) = -\sum_{c}^{n} p(c)\log_2 p(c) = -(1/3)*\log_2(1/3) - (2/3)*\log_2(2/3) = 0,91$$

$$\begin{aligned}H(C|\text{Vómitos}) = & -P(SI,II)\log_2 P(II|SI) \\ & -P(NO,IS)\log_2 P(IS|NO) - P(NO,II)\log_2 P(II|NO) = \\ & -(1/3)\log_2(1/2) - (1/3)\log_2(1/2) - (1/3)\log_2(1) = 0,66\end{aligned}$$

$$\begin{aligned}H(C|\text{Diarrea}) = & -P(SI,IS)\log_2 P(IS|SI) - P(NO,II)\log_2 P(II|NO) = \\ & -(1/3)\log_2(1) - (2/3)\log_2(1) = 0\end{aligned}$$

De nuevo se elige el que tiene menos entropía ya que hará máxima la cantidad de información mutua, en este caso Diarrea, por lo que el árbol de decisión será (Ver figura) y el patrón q = (alta; si; si) se clasificara en I Salmonella dado que su primera característica al ser Alta lo sitúa en la citada clase



Ejercicio 4.- Un médico con experiencia está construyendo un sistema de razonamiento basado en casos para automatizar una tarea de diagnóstico. Los casos se corresponden con personas individuales, donde sus datos se componen de una serie de características que describen los posibles síntomas y la parte de solución representa el diagnóstico

(clasificación de la enfermedad). La base de casos contiene 10 casos que se pueden ver en la siguiente tabla.

Entrenamiento	Fiebre	Vómitos	Clasificación
C_1	alta	si	Into. por Salmonella, (IS)
C_2	media	no	Into. por Salmonella, (IS)
C_3	no	si	Otros (O)
C_4	media	si	Infección intestinal, (II)
C_5	no	no	Infección intestinal, (II)
C_6	media	no	Infección intestinal, (II)
C_7	alta	si	Into. por Salmonella, (IS)
C_8	alta	no	Otros (O)
C_9	no	si	Into. por Salmonella, (IS)
C_{10}	alta	no	Into. por Salmonella, (IS)

Construya el árbol de decisión y clasifique un patrón con valores de las características q = (alta; no).

Solución.-

La entropía asociada a la clase es

$$H(C) = -\sum_{c}^{n} p(c)\log_2 p(c) = -(5/10)*\log_2\left(5/10\right)$$
$$-(2/10)*\log_2\left(2/10\right)-(3/10)*\log_2\left(3/10\right) = 2{,}701$$

Fiebre

$$H(C\,|\,\mathrm{F}) = -\sum_{c}\sum_{r} p(\mathrm{f,c})\log_2 p(c\,|\,\mathrm{f}) = -P(alta, IS)\log_2 P(IS|\mathrm{alta}) -$$
$$-P(\mathrm{media}, \mathrm{IS})\log_2 P(IS|\mathrm{media}) - P(no,\mathrm{IS})\log_2 P(IS|\mathrm{no})$$
$$-P(alta,\mathrm{O})\log_2 P(O|\mathrm{alta}) - P(media,\mathrm{O})\log_2 P(O|\mathrm{media})$$
$$-P(\mathrm{no},\mathrm{O})\log_2 P(O|\mathrm{no}) - P(alta,\mathrm{II})\log_2 P(II|\mathrm{alta})$$
$$-P(\mathrm{media},\mathrm{II})\log_2 P(II|\mathrm{media}) \; - P(no,\mathrm{II})\log_2 P(II|\mathrm{no}) =$$
$$-(3/10)\log_2(3/4) - (1/10)\log_2(1/3) - (1/10)\log_2(1/3) - (1/10)\log_2(1/4)$$
$$-0-(1/10)\log_2(1/3) - 0 - (2/10)\log_2(2/3) - (1/10)\log_2(1/3) = \mathbf{1{,}34}$$

Vómitos

$$H(C\,|\,\mathrm{V}) = -\sum_{c}\sum_{r} p(\mathrm{v,c})\log_2 p(c\,|\,\mathrm{v}) = -P(\mathrm{Si}, IS)\log_2 P(IS|\mathrm{Si}) -$$
$$-P(\mathrm{Si},\mathrm{O})\log_2 P(\mathrm{O}|\mathrm{Si}) - P(\mathrm{Si},\mathrm{II})\log_2 P(\mathrm{II}|\mathrm{Si})$$
$$-P(\mathrm{No},\mathrm{IS})\log_2 P(\mathrm{IS}|\mathrm{No}) - P(\mathrm{No},\mathrm{O})\log_2 P(\mathrm{O}|\mathrm{No})$$
$$-P(\mathrm{No},\mathrm{II})\log_2 P(\mathrm{II}|\mathrm{No}) = \text{-}(3/10)\log_2(3/5)\text{-}(1/10)\log_2(1/5)\text{-}$$
$$\text{-}(1/10)\log_2(1/5)\text{-}(2/10)\log_2(2/5)\ \text{-}(1/10)\log_2(1/5)\text{-}(2/10)\log_2(2/5)=1{,}445$$

El nodo raíz es Fiebre dado que es el que menos entropía condicional tiene. De esta forma la nueva base de datos a analizar es para **Fiebre alta**

Entrenamiento	Fiebre	Vómitos	Clasificación
C_1	**alta**	**si**	**Into. por Salmonella, (IS)**
C_7	**alta**	**si**	**Into. por Salmonella, (IS)**
C_8	alta	no	Otros (O)
C_{10}	alta	no	Into. por Salmonella, (IS)

Para fiebre media es

Entrenamiento	Fiebre	Vómitos	Clasificación
C_2	media	no	Into. por Salmonella, (IS)
C_4	**media**	**si**	**Infección intestinal, (II)**
C_6	media	no	Infección intestinal, (II)

y para fiebre NO es

Entrenamiento	Fiebre	Vómitos	Clasificación
C_3	no	si	Otros (O)
C_5	**no**	**no**	**Infección intestinal, (II)**
C_9	no	si	Into. por Salmonella, (IS)