Ciudad de México a 4 de abril de 2022

**Ejercicio de clase:**

Grupo: 5CDM1 Equipo No. 1

Integrantes:

De Luna Ocampo Yanina

Medina Barreras Daniel Ivan

Ángeles Lomelí Felipe Alberto

García Rodríguez Diana Itzel

Bhumika Gupta, Aditya Rawat, Akshay Jain, Arpit Arora, Naresh Dhami. (2017). Analysis of Various Decision Tree Algorithms for Classification in Data Mining. *International Journal of Computer Applications* (0975-8887). Volume 163 - No 8, April 2017.

Leer el artículo mencionado y responder las siguientes preguntas:

## Ejercicio No. 1

**a)** Describa la minería de datos

Comprende la extracción de información de un dataset y los transforma en una estructura que sea entendible. Es el proceso computacional que descubre patrones en largos datasets que envuelven métodos de inteligencia artificial, aprendizaje máquina, estadística y sistemas de bases de datos.
Te ayuda a reducir el ruido de los datos y te ayuda a entender la información relevante para evaluar todos los resultados probables y posibles.
Hay 6 clases existentes:

- detección anómala (AS)
- aprendizaje de reglas de asociación (AS)
- clustering (ANS)
- clasificación (AS)
- regresión (ANS)

**b)** Explique las razones por las cuales se utilizan los árboles de decisión

1. Son fáciles de visualizar, de entenderlas e interpretarlas.
2. Necesitan muy poca preparación de los datos y elimina los valores en blanco.
3. El costo de usar el árbol es logarítmico en el número de puntos de datos usados para

entrenar el árbol.
4. Puede manejar datos categóricos y numéricos.
5. Maneja problemas de múltiples salidas.
6. Puede ser explicado fácilmente con lógica booleana cuando hay dos salidas.
7. Funciona bien incluso si las suposiciones son violadas por el conjunto de datos del que se toman los mismos.

**c)** ¿Cuál es la diferencia entre un árbol de clasificación y un árbol de regresión?

El árbol de clasificación es en el que el análisis del resultado predicho es la clase a la que pertenecen los datos. Y el árbol de regresión es en el que el análisis del resultado previsto puede ser considerado un número real.

## Ejercicio No. 2

Considerando los algoritmos: ID3, C4.5, CART y Random Forest. Realice un cuadro comparativo que considere los siguientes aspectos: descripción del algoritmo, criterio de partición que utiliza, si utiliza poda o no, tipo de datos que utiliza, ventajas y desventajas.

| Criterio | ID3 | C4.5 | CART | Random Forest |
|---|---|---|---|---|
| Descripción | Es un algoritmo de aprendizaje que pretende modelar los datos mediante un árbol, llamado árbol de decisión. En este árbol los nodos intermedios son atributos de los ejemplos presentados, las ramas representan valores de dichos atributos y los nodos finales son los valores de la clase, como ya vimos al hablar de los árboles de decisión binarios.<br>Su principal aplicación son los problemas de decisión. Su empleo se centra en los llamados problemas de clasificación. | Genera un árbol de decisión a partir de los datos mediante particiones realizadas recursivamente. Antes de cada partición de datos, el algoritmo considera todas las pruebas posibles que pueden dividir el conjunto de datos y selecciona la prueba que resulta en la mayor ganancia de información o en la mayor proporción de ganancia de información. Para cada atributo discreto, se considera una prueba con n resultados, siendo n el número de valores posibles que puede tomar el atributo. Para cada atributo continuo se realiza una prueba binaria sobre cada uno de los valores que toma el atributo en los datos. | Se trata de un algoritmo basado en árbol que funciona examinando muchas diversas maneras de particionar o dividir localmente los datos en segmentos más pequeños con base en diferentes valores y combinaciones de predictores. CART selecciona las divisiones de mejor rendimiento y luego repite este proceso de forma recursiva hasta encontrar el conjunto óptimo. El resultado es un árbol de decisión representado por una serie de divisiones binarias que conducen a nodos terminales que pueden ser descritos por un conjunto de reglas específicas.<br>Funciona para problemas de clasificación y regresión. | Un bosque aleatorio es un conjunto de árboles de decisión que se ha creado a partir del mismo conjunto de datos que “vota” conjuntamente para producir un modelo mejor que un árbol individual. Los árboles se crean seleccionando aleatoriamente un subconjunto de registros de visitas con sustitución (conocidos como empacados) y seleccionando aleatoriamente un subconjunto de los atributos para que el bosque esté formado de árboles de decisión ligeramente distintos. Con este método se introducen pequeñas variaciones en los árboles que se crean en el bosque aleatorio. Al añadir esta cantidad controlada de varianza, la precisión predictiva del algoritmo mejora. |
| Criterio partición | Es un método “divide y vencerás” y está basado en criterios de partición derivados de la ganancia de información. | Es un método “divide y vencerás”, utiliza la estrategia de profundidad-primero (depth-first). Y está basado en criterios de partición derivados de la ganancia (GainRatio) | Para saber cuál de estas dos particiones es la mejor el algoritmo CART define una función de costo que asigna un puntaje al nodo padre, usando el promedio ponderado de los índices Gini individuales de sus nodos hijos.<br>Elige la que tenga el menor valor posible para la función de costo, lo que indica un menor nivel de impureza y por tanto una mejor clasificación | Es un método “divide y vencerás” y está basado en criterios de partición derivados de la ganancia de información. |

| | | | | |
|---|---|---|---|---|
| Poda | No aplica ningún procedimiento de poda. | Poda basada en errores. Elimina las ramas que no contribuyen a la precisión y las reemplaza. | Uno de los métodos más usados es el de poda de complejidad de costos, que consiste en definir un hiperparámetro alpha que controla el nivel de overfitting: con un alpha igual a cero tendremos el árbol de decisión sin ningún recorte, y por tanto con un alto overfitting, mientras que a medida que aumenta alpha se eliminarán algunos nodos del árbol hasta lograr un balance adecuado entre la precisión con el set de entrenamiento y la que se logra con el de validación | No aplica ningún procedimiento de poda. |
| Tipos de datos que utiliza | Datos categóricos. | Datos categóricos y continuos. Puede manejar el ruido y los datos faltantes. | Datos nominales y continuos. Puede manejar datos faltantes. | Datos numéricos y categóricos. |

| Criterio | ID3 | C4.5 | CART | Random Forest |
|---|---|---|---|---|
| Ventajas | Los datos de entrenamientos son utilizados para crear reglas de predicción entendibles.<br>Se construye rápido y se puede obtener un árbol pequeño.<br>Utiliza todo el dataset para crear el árbol.<br>El tiempo de cálculo es la función lineal del producto de la característica y el número de nodo | Es fácil de implementar construye modelos fáciles de interpretar<br>Puede manejar datos categóricos y continuos.<br>lidia con el ruido y datos faltantes. | Puede manejar datos vacíos automáticamente.<br>Utiliza cualquier combinación de variables continuas o discretas.<br>Ejecuta automáticamente la selección de variables<br>Puede establecer interacciones entre las variables. | Reconoce datos fuera de lo común y anomalías.<br>Es uno de los algoritmos de aprendizaje más exactos.<br>Da un estimado de las variables importantes en clasificación. |
| Desventajas | Para un dataset pequeño, se pueden obtener resultados sobre puestos o sobre clasificados.<br>Solo un atributo es revisado al instante por lo que se consume mucho tiempo para la toma de decisiones. | Pequeñas variaciones en los datos pueden guiar a diferentes árboles de decisión.<br>No funciona para pequeños conjuntos de datos | Puede tener árboles de decisión inestables.<br>Se divide solo por una variable.<br>Es no parametrizado. | Se puede producir un sobreajuste de datos, con lo que un solo árbol difícilmente predeciría datos futuros que no se hubieran utilizado para crear el árbol inicial.<br>A veces las clasificaciones pueden ser difíciles de interpretar. |

## Ejercicio 3

Considerando los siguientes criterios de selección de atributo para particionamiento: *Entropy (Information Gain), Gain Ratio* and *Gini Index*. Realice una descripción con sus propias palabras de cada uno de ellos.

| Criterio | Descripción |
|---|---|
| Entropy (Information Gain) | La entropía es el grado de incertidumbre, impureza o desorden de una variable aleatoria, o una medida de pureza. Caracteriza la impureza de una clase arbitraria de ejemplos.<br>La entropía es la medida de las impurezas o la aleatoriedad de los puntos de datos.<br>En este caso, si todos los elementos pertenecen a una única clase, se denomina "Pura", y si no, la distribución se denomina "Impura".<br>Se calcula entre 0 y 1, pero dependiendo del número de grupos o clases presentes en el conjunto de datos, puede ser superior a 1, pero representando el mismo significado, es decir, un nivel extremo de desorden. |
| Gain Ratio | La Relación de Ganancia intenta disminuir el sesgo de la Ganancia de Información en los predictores altamente ramificados introduciendo un término normalizador llamado Información Intrínseca.<br>La Información Intrínseca se define como la entropía de las proporciones del subconjunto de datos. En otras palabras, es difícil que nos resulte difícil adivinar en qué rama se encuentra una muestra seleccionada al azar. |
| Gini Index | El índice de gini, o coeficiente de gini, o impureza de gini, calcula el grado de probabilidad de que una variable específica esté mal clasificada cuando se elige al azar y una variación del coeficiente de gini. Funciona en variables categóricas, proporciona resultados ya sea "éxito" o "fracaso" y por lo tanto lleva a cabo la división binaria solamente. |

## Ejercicio 4

Considerando el ejercicio de evaluación del artículo (laptop), aplique el proceso de cálculo de medidas de evaluación al conjunto de datos de carros (accesible, no accesible).
Realizar el proceso de los cinco (*visto en clase*) pasos de los cálculos para la elección del atributo de particionamiento.

| | A | B | C | D | E | F | G |
|---|---|---|---|---|---|---|---|
| 1 | precioCompra | costoManto | noPuertas | NoPasajeros | tamCajuela | Seguridad | evalcarro |
| 2 | vhigh | med | dos | cuatro | small | low | acc |
| 3 | vhigh | med | dos | tres | med | high | acc |
| 4 | high | med | dos | cuatro | big | med | acc |
| 5 | med | low | dos | more | med | high | acc |
| 6 | high | med | dos | more | big | med | acc |
| 7 | vhigh | med | tres | cuatro | small | low | acc |
| 8 | high | low | tres | dos | med | high | acc |
| 9 | vhigh | med | tres | cuatro | big | med | acc |
| 10 | med | med | cuatro | cuatro | small | high | acc |
| 11 | high | med | cuatro | cuatro | med | med | acc |
| 12 | vhigh | low | cuatro | more | big | med | acc |
| 13 | vhigh | med | cuatro | more | big | high | acc |
| 14 | med | med | 5more | cuatro | small | high | acc |
| 15 | high | med | 5more | cuatro | med | med | acc |
| 16 | vhigh | med | 5more | cuatro | med | high | acc |
| 17 | vhigh | vhigh | dos | dos | small | low | unacc |
| 18 | vhigh | vhigh | dos | tres | small | med | unacc |
| 19 | vhigh | vhigh | dos | more | small | high | unacc |
| 20 | high | low | cuatro | tres | med | high | unacc |
| 21 | high | low | tres | dos | big | low | unacc |
| 22 | | | | | | | |

1. cálculo de la Entropía

$$E(s) = \sum_{i=1}^{c} - pi\ Log_2(pi)$$

| eval carro | |
|---|---|
| acc | unacc |
| 15 | 5 |

$E(eval\ carro) = E(acc, unacc) = E(15, 5)$
$= (-\ 15/20\ Log_2(15/20)) + (-\ 5/20\ Log_2(5/20))$
$= 0.0937 + 0.1505 = 0.2442 \rightarrow$ Entropía total

2. Dividir el conjunto de datos en los diversos atributos.

| **Atributo objetivo** | eval carro |
|---|---|
| | |
| **Atributo** | **Dominio** |
| precioCompra | med, high, vhigh |
| costoManto | low, med, vhigh |
| noPuertas | dos, tres, cuatro, 5more |
| NoPasajeros | dos, tres, cuatro, more |
| tamCajuela | small, med, big |
| Seguridad | low, med, high |

3. Se calcula la entropía en cada rama y se suman proporcionalmente para calcular la entropía total

$E(T, X) = \sum_{c \in x} p(c)E(C)$ ; $Gain(T, X) = E(T) - E(T, X)$

**Para precioCompra**

| | | | | Count |
|---|---|---|---|---|
| precioCompra | med | eval carro | acc | 3 |
| | | | unacc | 0 |
| | high | eval carro | acc | 5 |
| | | | unacc | 2 |
| | vhigh | eval carro | acc | 7 |
| | | | unacc | 3 |

$E(eval\ carro,\ precioCompra) = P(med) * E(3,0) + P(high) * E(5,2) + P(vhigh) * E(7,3)$

-med: P(3/20) = 0.15 E(3,0) = 0
-high: P(7/20) = 0.35 E(5,2) = $(-\ 5/20\ Log_2(5/20)) + (-\ 2/20\ Log_2(2/20))$ = 0.1505 + 0.1 = 0.2505
-vhigh: P(10/20) = 0.5 E(7,3) = $(-\ 7/20\ Log_2(7/20)) + (-\ 3/20\ Log_2(3/20))$ = 0.1595 + 0.1235 = 0.283
$E(eval\ carro,\ precioCompra)$ = 0.15 * 0 + 0.35 * 0.2505 + 0.5 * 0.283 = 0.23

4. Ganancia de información

GAIN = 0.24 - 0.23 = 0.01

**Para costoManto**

| | | | | Count |
|---|---|---|---|---|
| costoManto | low | eval carro | acc | 3 |
| | | | unacc | 2 |
| | med | eval carro | acc | 12 |
| | | | unacc | 0 |
| | vhigh | eval carro | acc | 0 |
| | | | unacc | 3 |

$E(eval\ carro,\ costoManto) = P(low) * E(3,2) + P(med) * E(12,0) + P(vhigh) * E(0,3)$

-low: P(5/20) = 0.25 E(3,2) = $(-\ 3/20\ Log_2(3/20)) + (-\ 2/20\ Log_2(2/20))$ = 0.1235 + 0.1 = 0.2235

-med: P(12/20) = 0.6 E(12,0) = 0

-vhigh: P(3/20) = 0.15 E(0,3) = 0

$E(eval\ carro,\ costoManto)$ = 0.25 * 0.2235 + 0.6 * 0 + 0.15 * 0 = 0.5

4. Ganancia de información

GAIN = 0.24 - 0.5 = - 0.26

**Para noPuertas**

| | | | | Count |
|---|---|---|---|---|
| noPuertas | dos | eval carro | acc | 5 |
| | | | unacc | 3 |
| | tres | eval carro | acc | 3 |
| | | | unacc | 1 |
| | cuatro | eval carro | acc | 4 |
| | | | unacc | 1 |
| | 5more | eval carro | acc | 3 |
| | | | unacc | 0 |

$E(eval\ carro,\ noPuertas) = P(dos) * E(5,3) + P(tres) * E(3,1) + P(cuatro) * E(4,1) + P(5more) * E(3,0)$

-dos: P(8/20) = 0.4 E(5,3) = $(-\ 5/20\ Log_2(5/20)) + (-\ 3/20\ Log_2(3/20))$ = 0.1505 + 0.1235 = 0.274

-tres: P(4/20) = 0.2 E(3,1) = $(-\ 3/20\ Log_2(3/20))\ +\ (-\ 1/20\ Log_2(1/20))$ = 0.1235 + 0.0650 = 0.1885

-cuatro: P(5/20) = 0.25 E(4,1) = $(-\ 4/20\ Log_2(4/20))\ +\ (-\ 1/20\ Log_2(1/20))$ = 0.1397 + 0.0650 = 0.2047

-5more: P(3/20) = 0.15 E(3,0) = 0

$E(eval\ carro,\ noPuertas)$ = 0.4 * 0.274 + 0.2 * 0.1885 + 0.25 * 0.2047 + 0.15 * 0 = 0.1096 + 0.0377 + 0.0511 + 0 = 0.205

4. Ganancia de información

GAIN = 0.24 - 0.205 = 0.035

**Para NoPasajeros**

| | | | | Count |
|---|---|---|---|---|
| NoPasajeros | dos | eval carro | acc | 1 |
| | | | unacc | 2 |
| | tres | eval carro | acc | 1 |
| | | | unacc | 2 |
| | cuatro | eval carro | acc | 9 |
| | | | unacc | 0 |
| | more | eval carro | acc | 4 |
| | | | unacc | 1 |

$E(eval\ carro,\ NoPasajeros) = P(dos)\ *\ E(1,2)\ +\ P(tres)\ *\ E(1,2)\ +\ P(cuatro)\ *\ E(9,0)\ +\ P(more)\ *\ E(4,1)$

-dos: P(3/20) = 0.15 E(1,2) = $(-\ 1/20\ Log_2(1/20))\ +\ (-\ 2/20\ Log_2(2/20))$ = 0.0650 + 0.1 = 0.165

-tres: P(3/20) = 0.15 E(1,2) = $(-\ 1/20\ Log_2(1/20))\ +\ (-\ 2/20\ Log_2(2/20))$ = 0.0650 + 0.1 = 0.165

-cuatro: P(9/20) = 0.45 E(9,0) = 0

-more: P(5/20) = 0.25 E(4,1) = $(-\ 4/20\ Log_2(4/20))\ +\ (-\ 1/20\ Log_2(1/20))$ = 0.1397 + 0.0650 = 0.2047

$E(eval\ carro,\ NoPasajeros)$ = 0.15 * 0.165 + 0.15 * 0.165 + 0 + 0.25 * 0.2047 = 0.0247 + 0.0247 + 0 + 0.0511 = 0.1005

4. Ganancia de información

GAIN = 0.24 - 0.1005 = 0.1395

**Para tamCajuela**

| | | | | Count |
|---|---|---|---|---|
| tamCajuela | small | eval carro | acc | 4 |
| | | | unacc | 3 |
| | med | eval carro | acc | 6 |
| | | | unacc | 1 |
| | big | eval carro | acc | 5 |
| | | | unacc | 1 |

$E(eval\ carro,\ tamCajuela) = P(small) * E(4,3) + P(med) * E(6,1) + P(big) * E(5,1)$

-small: P(7/20) = 0.35 E(4,3) = $(-\ 4/20\ Log_2(4/20)) + (-\ 3/20\ Log_2(3/20))$ = 0.1397 + 0.1235 = 0.2632
-med: P(7/20) = 0.35 E(6,1) = $(-\ 6/20\ Log_2(6/20)) + (-\ 1/20\ Log_2(1/20))$ = 0.1568 + 0.0650 = 0.2218
-big: P(6/20) = 0.3 E(5,1) = $(-\ 5/20\ Log_2(5/20)) + (-\ 1/20\ Log_2(1/20))$ = 0.1505 + 0.0650 = 0.2155
$E(eval\ carro,\ tamCajuela)$ = 0.35 * 0.2632 + 0.35 * 0.2218 + 0.3 * 0.2155 = 0.0921 + 0.0776 + 0.0646 = 0.2343

4. Ganancia de información

GAIN = 0.24 - 0.23 = 0.01

**Para Seguridad**

| | | | | Count |
|---|---|---|---|---|
| Seguridad | low | eval carro | acc | 2 |
| | | | unacc | 2 |
| | med | eval carro | acc | 6 |
| | | | unacc | 1 |
| | high | eval carro | acc | 6 |
| | | | unacc | 2 |

$E(eval\ carro,\ Seguridad) = P(low) * E(2,2) + P(med) * E(6,1) + P(high) * E(6,2)$

-low: P(4/20) = 0.2 E(2,2) = $(-\ 2/20\ Log_2(2/20)) + (-\ 2/20\ Log_2(2/20))$ = 0.1 + 0.1 = 0.2
-med: P(7/20) = 0.35 E(6,1) = $(-\ 6/20\ Log_2(6/20)) + (-\ 1/20\ Log_2(1/20))$ = 0.1568 + 0.0650 =

0.2218
-high: P(8/20) = 0.4 E(6,2) = $(-\ 6/20\ Log_2(6/20))\ +\ (-\ 2/20\ Log_2(2/20))$ = 0.1568 + 0.1 = 0.2568

$E(eval\ carro,\ Seguridad)$ = 0.2 * 0.2 + 0.35 * 0.2218 + 0.4 * 0.2568 = 0.4 + 0.0776 + 0.1027 = 0.5803

4. Ganancia de información

GAIN = 0.24 - 0.6 = -0.36

5. Elegir el atributo con mayor ganancia de información.

| Variable | Ganancia | (acc, unacc) |
|---|---|---|
| precioCompra | 0.01 | med = (3,0), high= (5,2), vhigh= (7,3) |
| costoManto | -0.26 | low= (3,2), med= (12,0), vhigh= (0,3) |
| noPuertas | 0.035 | dos= (5,3), tres= (3,1), cuatro= (4,1), 5more= (3,0) |
| noPasajeros | 0.1395 | dos= (1,2), tres= (1,2), cuatro= (9,0), more= (4,1) |
| tamCajuela | 0.01 | small= (4,3), med= (6,1), big= (5,1) |
| Seguridad | -0.36 | low= (2,2), med= (6,1), high= (6,2) |

**noPasajeros** es la variable que brinda la mayor ganancia de información, por tal, será la primera en ser elegida.

## Ejercicio 5

Plantee un conjunto de datos, con 15 registros y 5 atributos, cuyo atributo objetivo sea dicotómico y aplique las actividades que realizó en el ejercicio número 4 de esta guía.

| | raza | tamaño | color | tieneCasa | comió |
|---|---|---|---|---|---|
| **1** | Pura | chico | miel | si | no |
| **2** | única | chico | blanco | no | no |
| **3** | única | grande | negro | no | no |
| **4** | pura | chico | miel | si | si |
| **5** | única | mediano | miel | no | si |
| **6** | pura | chico | negro | si | si |
| **7** | única | grande | blanco | no | no |

| | | | | | |
|---|---|---|---|---|---|
| **8** | única | mediano | miel | no | si |
| **9** | única | chico | blanco | si | si |
| **10** | pura | chico | negro | no | si |
| **11** | única | grande | miel | si | si |
| **12** | única | chico | miel | si | si |
| **13** | pura | mediano | negro | no | no |
| **14** | pura | mediano | blanco | no | si |
| **15** | única | grande | miel | si | no |

1. cálculo de la Entropía

$$E(s) = \sum_{i=1}^{c} - pi\ Log_2(pi)$$

| **comió** | |
|---|---|
| si | no |
| 9 | 6 |

$E(comió) = E(si, no) = E(9, 6)$
$= (-\ 9/15\ Log_2(9/15)) + (-\ 6/15\ Log_2(6/15))$
$= 0.4421 + 0.5287 = 0.9708 \rightarrow$ Entropía total

2. Dividir el conjunto de datos en los diversos atributos.

| **Atributo objetivo** | comió |
|---|---|
| | |
| **Atributo** | **Dominio** |
| raza | pura, única |
| tamaño | chico, mediano, grande |
| color | miel,blanco,negro |
| tieneCasa | si, no |
| comió | si, no |

3. Se calcula la entropía en cada rama y se suman proporcionalmente para calcular la entropía total

$E(T, X) = \sum_{c \epsilon x} p(c)E(C)$ ; $Gain(T, X) = E(T) - E(T, X)$

**Para raza**

| | | | | Count |
|---|---|---|---|---|
| raza | pura | comió | si | 4 |
| | | | no | 2 |
| | única | comió | si | 5 |
| | | | no | 4 |

$E(comió, raza) = P(pura) * E(4, 2) + P(única) * E(5, 4)$

-pura: P(6/15) = 0.4 E(4,2) = $(-\ 4/15\ Log_2(4/15)) + (-\ 2/15\ Log_2(2/15))$ = 0.5085 + 0.3875 = 0.8955
-única: P(9/15) = 0.6 E(5,4) = $(-\ 5/15\ Log_2(5/15)) + (-\ 4/15\ Log_2(4/15))$ = 0.5283 + 0.5085 = 1.0368
$E(comió, raza)$ = 0.4 * 0.8955 + 0.5 * 1.0368 = 0.8766

4. Ganancia de información

GAIN = 0.9708 - 0.8766 = 0.0942

**Para tamaño**

| | | | | Count |
|---|---|---|---|---|
| tamaño | chico | comió | si | 5 |
| | | | no | 2 |
| | median o | comió | si | 3 |
| | | | no | 1 |
| | grande | comió | si | 1 |
| | | | no | 3 |

$E(comió, tamaño) = P(chico) * E(5, 2) + P(mediano) * E(3, 1) + P(grande) * E(1, 3)$

-chico: P(7/15) = 0.46 E(5,2) = $(-\ 5/15\ Log_2(5/15)) + (-\ 2/15\ Log_2(2/15))$ = 0.5283 + 0.3857 = 0.914
-mediano: P(4/15) = 0.2 E(3,1) = $(-\ 3/15\ Log_2(3/15)) + (-\ 1/15\ Log_2(1/15))$ = 0.4643 + 0.2604 = 0.7247

-grande: P(4/15) = 0.2  E(1,3) = $(-\ 1/15\ Log_2(1/15))\ +\ (-\ 3/15\ Log_2(3/15))$ = 0.2604 + 0.4643 = 0.7247

$E(comió,\ tamaño)$ = 0.46 * 0.914 + 0.2 * 0.7247 + 0.2 * 0.7247 = 0.7103

4. Ganancia de información

GAIN = 0.9708 - 0.7103 = 0.2605

**Para color**

| | | | | Count |
|---|---|---|---|---|
| color | miel | comió | si | 5 |
| | | | no | 2 |
| | blanco | comió | si | 2 |
| | | | no | 2 |
| | negro | comió | si | 2 |
| | | | no | 2 |

$E(comio,\ color)\ =\ P(miel)\ *\ E(5,2)\ +\ P(blanco)\ *\ E(2,2)\ +\ P(negro)\ *\ E(2,2)$

-miel: P(7/15) = 0.46  E(5,2) = $(-\ 5/15\ Log_2(5/15))\ +\ (-\ 2/15\ Log_2(2/15))$ = 0.5283 + 0.3857 = 0.914

-blanco: P(4/15) = 0.2  E(2,2) = $(-\ 2/15\ Log_2(2/15))\ +\ (-\ 2/15\ Log_2(2/15))$ = 0.3857 + 0.3857 = 0.7714

-negro: P(4/15) = 0.2  E(2,2) = $(-\ 2/15\ Log_2(2/15))\ +\ (-\ 2/15\ Log_2(2/15))$ = 0.3857 + 0.3857 = 0.7714

$E(comio,\ color)$ = 0.46 * 0.914 + 0.2 * 0.7714 + 0.2 * 0.7714 = 0.729

4. Ganancia de información

GAIN = 0.9708 - 0.729 = 0.2418

**Para tieneCasa**

| | | | | Count |
|---|---|---|---|---|
| tieneCasa | si | comió | si | 5 |
| | | | no | 2 |
| | no | comió | si | 4 |

| | | | | |
|---|---|---|---|---|
| | | | no | 4 |

$E(comió, tieneCasa) = P(si) * E(5,2) + P(no) * E(4,4)$

-si: P(7/15) = 0.46 E(5,2) = $(-\ 5/15\ Log_2(5/15)) + (-\ 2/15\ Log_2(2/15))$ = 0.5283 + 0.3857 = 0.914

-no: P(8/15) = 0.53 E(4,4) = $(-\ 4/15\ Log_2(4/15)) + (-\ 4/15\ Log_2(4/15))$ = 0.5085 + 0.5085 = 1.017

$E(comió,\ tieneCasa)$ = 0.46 * 0.914 + 0.53 * 1.017 = 0.9594

4. Ganancia de información

GAIN = 0.9708 - 0.9594 = 0.0113

5. Elegir el atributo con mayor ganancia de información.

| Variable | Ganancia | (acc, unacc) |
|---|---|---|
| raza | 0.0942 | pura = (4, 2), única = (5. 4) |
| tamaño | 0.2605 | chico = (5, 2), mediano = (3, 1), grande = (1, 3) |
| color | 0.2418 | miel = (5, 2), blanco = (2, 2), negro = (2, 2) |
| tieneCasa | 0.0942 | si = (5, 2), no = (5, 4) |

**tamaño** es la variable que brinda la mayor ganancia de información, por tal, será la primera en ser elegida.

## Ejercicio 6

Suponga que tiene la siguiente matriz de confusión de la evaluación de carros.

| Original/Predicción | Accesible | No accesible |
|---|---|---|
| **Accesible** | 40 | 5 |
| **No accesible** | 2 | 3 |
| | 42 | 8 |

Calcule y explique las diversas medidas que puede generar con base en los datos de la matriz de confusión.

| Medida | Cálculo | Explicación |
|---|---|---|
| Negativo verdadero | 40 | El valor real es negativo y la prueba predijo también que el resultado era negativo |
| Positivo falso | 5 | El valor real es negativo, y la prueba predijo que el resultado es positivo. |
| Negativo falso | 2 | El valor real es positivo, y la prueba predijo que el resultado es negativo |

| | | |
|---|---|---|
| Positivo verdadero | 3 | El valor real es positivo y la prueba predijo también que era positivo. |
| Tasa de exactitud | (40+3)/(40+5+2+3) = 43/50 = 0.86 | La cantidad de predicciones positivas que fueron correctas fue del 86% |

| | | |
|---|---|---|
| Tasa de error | (5+2)/(40+5+2+3) = 7/50 = 0.14 | La cantidad de predicciones que fueron incorrectas es del 14% |
| Precisión | 3/(5+3) = 3/8 = 0.375 | El porcentaje de casos positivos detectados fue 37.5% |
| Sensibilidad (*Recall*) | 3/(2+3) = 3/5 = 0.6 | En este caso la sensibilidad apenas es capaz de detectar correctamente, con un porcentaje del 60% |
| Tasa de positivos falsos | 5/(40+5) = 5/45 = 0.11 | La probabilidad de que se produzca una falsa alarma: que se dé un resultado positivo cuando el valor verdadero sea negativo es del 11% |
| Tasa de negativos falsos | 2/(2+3) = 2/5 = 0.4 | La probabilidad de que la prueba pase por alto un verdadero positivo es del 4% |
| Especificidad | 40/(40+5) = 40/45 = 0.88 | La probabilidad de los casos negativos que el algoritmo ha clasificado correctamente es del 88% |