

Ciudad de México a 25 de octubre de 2021

**Ejercicio de clase:**

Grupo: 3CV19

Equipo No. 6

Integrantes:

Medina Granados Alan Alejandro

Guzman Gutierrez Manuel

Castro Cruces Jorge Eduardo

Bhumika Gupta, Aditya Rawat, Akshay Jain, Arpit Arora, Naresh Dharmi. (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

Es el proceso de clasificación de grandes conjuntos de datos para hallar patrones, información relevante y útil; involucran métodos en la intersección de inteligencia, aprendizaje automático, estadísticas y base de datos sistemas.

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

- Son fáciles de comprender e interpretar.
- Rápida presentación.
- Tienen un costo no tan alto
- Se pueden usar tanto para datos numéricos como categóricos.
- Maneja problemas de múltiple salida.
- Se explican fácilmente mediante lógica de boole.

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

En el árbol de clasificación las variables dependientes son categóricas y no ordenadas

En el árbol de regresión las variables dependientes son continuas u ordenadas (valores completos).

**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.

<b>Criterio</b>	<b>ID3</b>	<b>C4.5</b>	<b>CART</b>	<b>Random Forest</b>
Descripción	Utiliza una búsqueda codiciosa de arriba hacia abajo a través del conjuntos dados, donde cada atributo en cada nodo del árbol se prueba para seleccionar el atributo que sea mejor para la clasificación de un conjunto dado	Es mucho más rápido, más eficiente en memoria y se utiliza para Construye árboles de decisión más pequeños y produce interpretaciones más intuitivas.	Significa árboles de clasificación y regresión	Un bosque aleatorio es una colección de árboles simples predictores, de modo que cada árbol produce una respuesta cuando un conjunto de los valores predictores se dan como entrada. También funciona tanto para clasificación y problemas de regresión.
Criterio partición	Las instancias de resultado que son posibles se examinan si pertenecen a la misma clase o no. Para los casos de la misma clase, se utiliza una clase de un solo nombre para denotar, de lo contrario, el las instancias se clasifican sobre la base del atributo de división.	Para la división de atributos categóricos, C4.5 sigue el enfoque similar a los algoritmos ID3. Continuo los atributos siempre generan divisiones binarias. Seleccionar el atributo con la relación de ganancia más alta. Estos pasos se aplican repetidamente a nuevas ramas de árboles.	El árbol de clasificación es construido por CART mediante la división binaria del atributo. La función de regresión de CART se puede utilizar al pronosticar una variable dependiente dado un conjunto de predictores durante un periodo de tiempo determinado	Resuelve la clasificación problemas, la respuesta o el resultado aparece en forma de un membresía de clase, que asocia o clasifica, un conjunto de valores de predictores independientes con la categoría coincidente presente en la variable dependiente.
Poda	El atributo con la mayor ganancia de información se puede seleccionar como atributo de prueba del nodo actual. ID3 es basado en la navaja de Occam.	La ganancia de información sesga el atributo con más número de valores. Por lo tanto, C4.5 usa Gain Ratio que es un criterio de selección menos sesgado.	El índice de Gini se utiliza para seleccionar el atributo de división.	Cada árbol produce una respuesta cuando un conjunto de los valores predictores se dan como entrada.
Tipos de datos que utiliza	Categóricos	Continuos y discritos	Admite tanto continuo como	Admite tanto continuo como

## DATA MINING

			datos de atributos nominales	datos de atributos nominales
--	--	--	------------------------------	------------------------------

<b>Criterio</b>	<b>ID3</b>	<b>C4.5</b>	<b>CART</b>	<b>Random Forest</b>
Ventajas	<ul style="list-style-type: none"> <li>Árbol rápido y barato</li> <li>Busca en todo el conjunto de datos</li> <li>Encuentra los nodos hoja permitiendo así que los datos de prueba sean podados y reduciendo el número de pruebas.</li> </ul>	<ul style="list-style-type: none"> <li>C4.5 es fácil de implementar.</li> <li>C4.5 crea modelos que se pueden interpretar fácilmente.</li> <li>Puede manejar valores categóricos y continuos.</li> <li>Puede lidiar con el ruido y lidiar con el valor faltante de atributos.</li> </ul>	<ul style="list-style-type: none"> <li>CART puede manejar los valores perdidos automáticamente usando divisiones sustitutas.</li> <li>Utiliza cualquier combinación de variables continuas / discretas.</li> <li>CART realiza automáticamente la selección de variables.</li> <li>CART puede establecer interacciones entre variables.</li> </ul>	<ul style="list-style-type: none"> <li>Reconoce valores atípicos y anomalías en los conocimientos de datos.</li> <li>Es uno de los algoritmos de aprendizaje más precisos disponibles. Para muchos conjuntos de datos, produce una gran clasificación de clasificador es precisos.</li> <li>Da una estimación de las variables importantes en la clasificación.</li> </ul>
Desventajas	<ul style="list-style-type: none"> <li>Para una muestra pequeña, los datos pueden estar sobreajustados o sobre clasificados.</li> <li>Para tomar una decisión, solo se prueba un</li> </ul>	<ul style="list-style-type: none"> <li>Una pequeña variación en los datos puede llevar a diferentes decisiones de árboles cuando</li> </ul>	<ul style="list-style-type: none"> <li>CART puede tener árboles de decisión inestables.</li> <li>CART se divide solo por una variable.</li> </ul>	<ul style="list-style-type: none"> <li>A veces la clasificación hecha por Random Forest son difíciles de interpretar por los humanos.</li> <li>Random Forest a veces se sobreajuste con</li> </ul>

## DATA MINING

	atributo en un instantáneo o consumiendo así mucho tiempo.	se usa C4.5. <ul style="list-style-type: none"><li>• Para un conjunto de entrenamiento pequeño, C4.5 no funciona muy bien.</li></ul>	<ul style="list-style-type: none"><li>• No paramétrico.</li></ul>	conjuntos de datos con Tareas ruidosas de clasificación / regresión.
--	--	--	---	--

**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.

<b>Criterio</b>	<b>Descripción</b>
Entropy (Information Gain)	Valor entre 0-1 que indica la incertidumbre en una variable aleatoria.
Gain Ratio	Medida de ganancia de información, es mejor seleccionar atributos con una gran cantidad de valores
Gini Index	Medida de impureza de un conjunto de datos, son los atributos de división.

## 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							

Aplicación del proceso del cálculo de medidas de evaluación al conjunto de datos del carro (accesible, no accesible)

## Paso 1: Cálculo de la entropía total

Evaluación Carro	
Accesible	No accesible
15	5

$$E(\text{Total}) = -\frac{15}{20} \log_2 \left( \frac{15}{20} \right) - \frac{5}{20} \log_2 \left( \frac{5}{20} \right) = 0.8112$$

## Paso 2: Dividir el conjunto de datos en los diversos atributos

Atributo objetivo	Evaluación Carro
Atributo	Dominio
Precio Compra	Muy alto
	Alto
	Medio
Costo Mantenimiento	Muy alto
	Medio
	Bajo
Número Puertas	Dos
	Tres
	Cuatro
	Cinco o más
Número Pasajeros	Dos
	Tres
	Cuatro
	Más
Tamaño Cajuela	Grande
	Mediana
	Pequeña
Seguridad	Alta
	Mediana
	Baja

## Paso 3: Se calcula la entropía de cada rama y se suman proporcionalmente para calcular la entropía total

Para Precio Compra

Precio Compra	Muy alto	Accesible	7
		No accesible	3
	Alto	Accesible	5
		No accesible	2
	Mediano	Accesible	3
		No accesible	0

$$E(\text{Muy alto}) = -\frac{7}{20} \log_2\left(\frac{7}{20}\right) - \frac{3}{20} \log_2\left(\frac{3}{20}\right) = 0.9406$$



$$E(\text{Alto}) = -\frac{5}{20} \log_2\left(\frac{5}{20}\right) - \frac{2}{20} \log_2\left(\frac{2}{20}\right) = 0.8321$$

$$E(\text{Mediano}) = 0$$

$$E(\text{Precio Compra, Evaluación}) = \frac{10}{20}(0.9406) + \frac{7}{20}(0.8321) + \frac{3}{20}(0.4105) = 0.7615$$

Para Costo Mantenimiento

Costo Mantenimiento	Muy alto	Accesible	0
		No accesible	3
	Medio	Accesible	12
		No accesible	0
	Bajo	Accesible	3
		No accesible	2

$$E(\text{Muy alto}) = 0$$

$$E(\text{Alto}) = 0$$

$$E(\text{Mediano}) = -\frac{3}{20} \log_2\left(\frac{3}{20}\right) - \frac{2}{20} \log_2\left(\frac{2}{20}\right) = 0.7427$$

$$E(\text{Precio Compra, Evaluación}) = \frac{5}{20}(0.7427) = 0.1856$$

Para Número Puertas

Precio Compra	Dos	Accesible	5
		No accesible	3
	Tres	Accesible	3
		No accesible	1
	Cuatro	Accesible	4
		No accesible	1
	Cinco o más	Accesible	3
		No accesible	0

$$E(\text{Dos}) = -\frac{5}{20} \log_2\left(\frac{5}{20}\right) - \frac{3}{20} \log_2\left(\frac{3}{20}\right) = 0.9105$$

$$E(\text{Tres}) = -\frac{3}{20} \log_2\left(\frac{3}{20}\right) - \frac{1}{20} \log_2\left(\frac{1}{20}\right) = 0.6266$$

$$E(\text{Cuatro}) = -\frac{4}{20} \log_2\left(\frac{4}{20}\right) - \frac{1}{20} \log_2\left(\frac{1}{20}\right) = 0.6804$$

$$E(\text{Cinco o más}) = 0$$

$$E(\text{Número Puertas, Evaluación}) = \frac{8}{20}(0.9105) + \frac{4}{20}(0.6266) + \frac{5}{20}(0.6804) + \frac{3}{20}(0.4105) = 0.6595$$

Para Número Pasajeros

Numero Pasajeros	Dos	Accesible	1
		No accesible	2
	Tres	Accesible	1
		No accesible	2
	Cuatro	Accesible	9
		No accesible	0
	Más	Accesible	4
		No accesible	1

$$E(\text{Dos}) = -\frac{1}{20} \log_2\left(\frac{1}{20}\right) - \frac{2}{20} \log_2\left(\frac{2}{20}\right) = 0.5482$$

$$E(\text{Tres}) = -\frac{1}{20} \log_2\left(\frac{1}{20}\right) - \frac{2}{20} \log_2\left(\frac{2}{20}\right) = 0.5482$$

$$E(\text{Cuatro}) = 0$$

$$E(\text{Cinco o más}) = -\frac{4}{20} \log_2\left(\frac{4}{20}\right) - \frac{1}{20} \log_2\left(\frac{1}{20}\right) = 0.6804$$

$$E(\text{Número Pasajeros, Evaluación}) = \frac{3}{20}(0.5482) + \frac{3}{20}(0.5482) + \frac{5}{20}(0.6804) = 0.3267$$

Para Tamaño de Cajuela

Costo Mantenimiento	Grande	Accesible	5
		No accesible	1
	Mediana	Accesible	6
		No accesible	1
	Pequeña	Accesible	4
		No accesible	3

$$E(\text{Grande}) = -\frac{5}{20} \log_2\left(\frac{5}{20}\right) - \frac{1}{20} \log_2\left(\frac{1}{20}\right) = 0.715$$

$$E(\text{Mediana}) = -\frac{6}{20} \log_2\left(\frac{6}{20}\right) - \frac{1}{20} \log_2\left(\frac{1}{20}\right) = 0.7371$$

$$E(\text{Pequeña}) = -\frac{4}{20} \log_2\left(\frac{4}{20}\right) - \frac{3}{20} \log_2\left(\frac{3}{20}\right) = 0.8749$$

$$E(\text{Tamaño Cajuela, Evaluación}) = \frac{6}{20}(0.715) + \frac{7}{20}(0.7371) + \frac{7}{20}(0.8749) = 0.779$$

Para Seguridad

Seguridad	Alta	Accesible	7
		No accesible	2
	Mediana	Accesible	6
		No accesible	1
	Baja	Accesible	2
		No accesible	2

$$E(\text{Grande}) = -\frac{7}{20} \log_2\left(\frac{7}{20}\right) - \frac{2}{20} \log_2\left(\frac{2}{20}\right) = 0.8622$$

$$E(\text{Mediana}) = -\frac{6}{20} \log_2\left(\frac{6}{20}\right) - \frac{1}{20} \log_2\left(\frac{1}{20}\right) = 0.7371$$

$$E(\text{Pequeña}) = -\frac{2}{20} \log_2\left(\frac{2}{20}\right) - \frac{2}{20} \log_2\left(\frac{2}{20}\right) = 0.6643$$

$$E(\text{Tamaño Cajuela, Evaluación}) = \frac{9}{20}(0.8622) + \frac{7}{20}(0.7371) + \frac{4}{20}(0.6643) = 0.7788$$

## Paso 4: Se calcula la ganancia de información

Para Precio Compra:  $\text{Gain}(\text{PrecioCompra, Evaluación}) = 0.8112 - 0.7615 = 0.0497$

Para Costo Mantenimiento:  $\text{Gain}(\text{Costo Mantenimiento, Evaluación}) = 0.8112 - 0.1856 = 0.6256$

Para Número Puertas:  $\text{Gain}(\text{Número Puertas, Evaluación}) = 0.8112 - 0.6595 = 0.1517$

Para Número Pasajeros:  $\text{Gain}(\text{Número Pasajeros}, \text{Evaluación}) = 0.8112 - 0.3267 = 0.4845$

Para Tamaño de Cajuela:  $\text{Gain}(\text{Tamaño Cajuela}, \text{Evaluación}) = 0.8112 - 0.779 = 0.0322$

Para Seguridad:  $\text{Gain}(\text{Seguridad}, \text{Evaluación}) = 0.8112 - 0.7788 = 0.0324$

## Paso 5: Elección del nodo de Decisión.

Variable	Ganancia	(Accesible, No Accesible)
Precio de Compra	0.0497	Muy Alto = (7, 3) Alto = (5, 2) Medio = (3, 0)
Costo Mantenimiento	0.6256	Muy Alto = (0, 3) Medio = (12, 0) Bajo = (3, 2)
Número Puertas	0.1517	Dos = (5, 3) Tres = (3, 1) Cuatro = (4, 1) Cinco o Más = (3, 0)
Número Pasajeros	0.4845	Dos = (1, 2) Tres = (1, 2) Cuatro = (9, 0) Más = (4, 1)
Tamaño Cajuela	0.0322	Grande = (5, 1) Mediana = (6, 1) Pequeña = (4, 3)
Seguridad	0.0324	Alta = (7, 2) Mediana = (6, 1) Baja = (2, 2)

Costo Mantenimiento es la variable que da una mayor ganancia

Se calcula SplitInfo de Mantenimiento:

$$\text{SplitInfoMantenimiento} = -\frac{3}{20} \log_2\left(\frac{3}{20}\right) - \frac{12}{20} \log_2\left(\frac{12}{20}\right) - \frac{5}{20} \log_2\left(\frac{5}{20}\right) = 1.3527$$

$$\text{GainRatio}(\text{Costo Mantenimiento}) = \frac{0.6256}{1.3527} = 0.4624$$

Se calcula Gini Index con respecto a la variable Evaluación

$$\text{Gini}(\text{Evaluación}) = 1 - \left(\frac{15}{20}\right)^2 - \left(\frac{5}{20}\right)^2 = 0.375$$

Se Calcula  $\text{Gini}_{\text{CostoMantenimiento} \in (\text{Bajo}, \text{Medio})} = \frac{17}{20} \text{Gini}(\text{Evaluación1}) + \frac{3}{20} \text{Gini}(\text{Evaluación2})$

$$= \frac{17}{20} \left(1 - \left(\frac{15}{17}\right)^2 - \left(\frac{2}{17}\right)^2\right) + \frac{3}{20} \left(1 - \left(\frac{3}{3}\right)^2 - \left(\frac{0}{3}\right)^2\right) = 0.1764$$

Se Calcula  $\text{Gini}_{\text{CostoMantenimiento} \in (\text{Bajo}, \text{Alto})} = \frac{8}{20} \text{Gini}(\text{Evaluación1}) + \frac{12}{20} \text{Gini}(\text{Evaluación2})$

$$= \frac{8}{20} \left(1 - \left(\frac{3}{8}\right)^2 - \left(\frac{5}{8}\right)^2\right) + \frac{12}{20} \left(1 - \left(\frac{12}{12}\right)^2 - \left(\frac{0}{12}\right)^2\right) = 0.1875$$

Se Calcula  $\text{Gini}_{\text{CostoMantenimiento} \in (\text{Medio}, \text{Muy alto})} = \frac{15}{20} \text{Gini}(\text{Evaluación1}) + \frac{5}{20} \text{Gini}(\text{Evaluación2})$

$$= \frac{15}{20} \left(1 - \left(\frac{12}{15}\right)^2 - \left(\frac{3}{15}\right)^2\right) + \frac{5}{20} \left(1 - \left(\frac{3}{5}\right)^2 - \left(\frac{2}{5}\right)^2\right) = 0.36$$

Proponemos realizar este proceso para los demás atributos, para ver si obtenemos valores menores de Gini Index

$$\text{SplitInfoPrecioCompra} = -\frac{10}{20} \log_2\left(\frac{10}{20}\right) - \frac{7}{20} \log_2\left(\frac{7}{20}\right) - \frac{3}{20} \log_2\left(\frac{3}{20}\right) = 1.4406$$

$$\text{GainRatio(PrecioCompra)} = \frac{0.0497}{1.4406} = 0.0344$$

$$\begin{aligned} \text{Se Calcula } Gini_{\text{PrecioCompra} \in (\text{Medio}, \text{Alto})} &= \frac{10}{20} Gini(\text{Evaluación1}) + \frac{10}{20} Gini(\text{Evaluación2}) \\ &= \frac{10}{20} \left(1 - \left(\frac{8}{10}\right)^2 - \left(\frac{2}{10}\right)^2\right) + \frac{10}{20} \left(1 - \left(\frac{7}{10}\right)^2 - \left(\frac{3}{10}\right)^2\right) = 0.37 \end{aligned}$$

$$\begin{aligned} \text{Se Calcula } Gini_{\text{PrecioCompra} \in (\text{Medio}, \text{Muy alto})} &= \frac{13}{20} Gini(\text{Evaluación1}) + \\ &\frac{7}{20} Gini(\text{Evaluación2}) \\ &= \frac{13}{20} \left(1 - \left(\frac{10}{13}\right)^2 - \left(\frac{3}{13}\right)^2\right) + \frac{7}{20} \left(1 - \left(\frac{5}{7}\right)^2 - \left(\frac{2}{7}\right)^2\right) = 0.3736 \end{aligned}$$

$$\begin{aligned} \text{Se Calcula } Gini_{\text{PrecioCompra} \in (\text{Alto}, \text{Muy alto})} &= \frac{17}{20} Gini(\text{Evaluación1}) + \frac{3}{20} Gini(\text{Evaluación2}) \\ &= \frac{17}{20} \left(1 - \left(\frac{12}{17}\right)^2 - \left(\frac{5}{17}\right)^2\right) + \frac{3}{20} \left(1 - \left(\frac{3}{3}\right)^2 - \left(\frac{0}{3}\right)^2\right) = 0.3529 \end{aligned}$$

$$\text{SplitInfoNumeroPuertas} = -\frac{8}{20} \log_2\left(\frac{8}{20}\right) - \frac{4}{20} \log_2\left(\frac{4}{20}\right) - \frac{5}{20} \log_2\left(\frac{5}{20}\right) - \frac{3}{20} \log_2\left(\frac{3}{20}\right) = 1.9037$$

$$\text{GainRatio(NumeroPuestas)} = \frac{0.1517}{1.9037} = 0.0796$$

$$\begin{aligned} \text{Se Calcula } Gini_{\text{NúmeroPuertas} \in (\text{Dos}, \text{Tres})} &= \frac{12}{20} Gini(\text{Evaluación1}) + \frac{8}{20} Gini(\text{Evaluación2}) \\ &= \frac{12}{20} \left(1 - \left(\frac{8}{12}\right)^2 - \left(\frac{4}{12}\right)^2\right) + \frac{8}{20} \left(1 - \left(\frac{7}{8}\right)^2 - \left(\frac{1}{8}\right)^2\right) = 0.3541 \end{aligned}$$

$$\begin{aligned} \text{Se Calcula } Gini_{\text{NúmeroPuertas} \in (\text{Tres}, \text{Cuatro})} &= \frac{9}{20} Gini(\text{Evaluación1}) + \frac{11}{20} Gini(\text{Evaluación2}) \\ &= \frac{9}{20} \left(1 - \left(\frac{7}{9}\right)^2 - \left(\frac{2}{9}\right)^2\right) + \frac{11}{20} \left(1 - \left(\frac{8}{11}\right)^2 - \left(\frac{3}{11}\right)^2\right) = 0.3737 \end{aligned}$$

$$\begin{aligned} \text{Se Calcula } Gini_{\text{NúmeroPuertas} \in (\text{Cuatro}, \text{Cinco o más})} &= \frac{8}{20} Gini(\text{Evaluación1}) + \\ &\frac{12}{20} Gini(\text{Evaluación2}) \\ &= \frac{8}{20} \left(1 - \left(\frac{7}{8}\right)^2 - \left(\frac{1}{8}\right)^2\right) + \frac{12}{20} \left(1 - \left(\frac{8}{12}\right)^2 - \left(\frac{4}{12}\right)^2\right) = 0.3541 \end{aligned}$$

$$\begin{aligned} \text{Se Calcula } Gini_{\text{NúmeroPuertas} \in (\text{Dos}, \text{Cinco o más})} &= \frac{11}{20} Gini(\text{Evaluación1}) + \\ &\frac{9}{20} Gini(\text{Evaluación2}) \\ &= \frac{11}{20} \left(1 - \left(\frac{8}{11}\right)^2 - \left(\frac{3}{11}\right)^2\right) + \frac{9}{20} \left(1 - \left(\frac{7}{9}\right)^2 - \left(\frac{2}{9}\right)^2\right) = 0.3737 \end{aligned}$$

$$\begin{aligned} \text{Se Calcula } Gini_{\text{NúmeroPuertas} \in (\text{Dos}, \text{Cuatro})} &= \frac{13}{20} Gini(\text{Evaluación1}) + \frac{7}{20} Gini(\text{Evaluación2}) \\ &= \frac{13}{20} \left(1 - \left(\frac{9}{13}\right)^2 - \left(\frac{4}{13}\right)^2\right) + \frac{7}{20} \left(1 - \left(\frac{6}{7}\right)^2 - \left(\frac{1}{7}\right)^2\right) = 0.3626 \end{aligned}$$

$$\begin{aligned} \text{Se Calcula } Gini_{\text{NúmeroPuertas} \in (\text{Tres}, \text{Cinco})} &= \frac{7}{20} Gini(\text{Evaluación1}) + \frac{13}{20} Gini(\text{Evaluación2}) \\ &= \frac{7}{20} \left(1 - \left(\frac{6}{7}\right)^2 - \left(\frac{1}{7}\right)^2\right) + \frac{13}{20} \left(1 - \left(\frac{9}{13}\right)^2 - \left(\frac{4}{13}\right)^2\right) = 0.3626 \end{aligned}$$

$$\text{SplitInfoNúmeroPasajeros} = -\frac{3}{20} \log_2\left(\frac{3}{20}\right) - \frac{3}{20} \log_2\left(\frac{3}{20}\right) - \frac{9}{20} \log_2\left(\frac{9}{20}\right) - \frac{5}{20} \log_2\left(\frac{5}{20}\right) = 1.8394$$

$$\text{GainRatio(NúmeroPasajeros)} = \frac{0.4845}{1.8394} = 0.2634$$

$$\begin{aligned} \text{Se Calcula } Gini_{\text{NúmeroPasajeros} \in (\text{Dos}, \text{Tres})} &= \frac{6}{20} Gini(\text{Evaluación1}) + \frac{14}{20} Gini(\text{Evaluación2}) \\ &= \frac{6}{20} \left(1 - \left(\frac{2}{6}\right)^2 - \left(\frac{4}{6}\right)^2\right) + \frac{14}{20} \left(1 - \left(\frac{13}{14}\right)^2 - \left(\frac{1}{14}\right)^2\right) = 0.2261 \end{aligned}$$

Se Calcula  $Gini_{NúmeroPasajeros \in (Tres, Cuatro)} = \frac{12}{20} Gini(Evaluación1) + \frac{8}{20} Gini(Evaluación2)$

$$= \frac{12}{20} \left(1 - \left(\frac{10}{12}\right)^2 - \left(\frac{2}{12}\right)^2\right) + \frac{8}{20} \left(1 - \left(\frac{5}{8}\right)^2 - \left(\frac{3}{8}\right)^2\right) = 0.3541$$

Se Calcula  $Gini_{NúmeroPasajeros \in (Dos, Cuatro)} = \frac{12}{20} Gini(Evaluación1) + \frac{8}{20} Gini(Evaluación2)$

Como son el número de elementos el número de pasajeros cuando son dos y tres, el resultado es el mismo, o sea

Se Calcula  $Gini_{NúmeroPasajeros \in (Tres, Más)} = \frac{8}{20} Gini(Evaluación1) + \frac{13}{20} Gini(Evaluación2)$

Como son el número de elementos el número de pasajeros cuando son dos y tres, el resultado es el mismo

$$SplitInfoTamañoCajuela = -\frac{6}{20} \log_2\left(\frac{6}{20}\right) - \frac{7}{20} \log_2\left(\frac{7}{20}\right) - \frac{7}{20} \log_2\left(\frac{7}{20}\right) = 1.5812$$

$$GainRatio(TamañoCajuela) = \frac{0.0322}{1.5812} = 0.0203$$

Se Calcula  $Gini_{TamañoCajuela \in (Pequeña, Mediana)} = \frac{14}{20} Gini(Evaluación1) + \frac{6}{20} Gini(Evaluación2)$

$$= \frac{14}{20} \left(1 - \left(\frac{10}{14}\right)^2 - \left(\frac{4}{14}\right)^2\right) + \frac{6}{20} \left(1 - \left(\frac{5}{6}\right)^2 - \left(\frac{1}{6}\right)^2\right) = 0.3736$$

Se Calcula  $Gini_{TamañoCajuela \in (Mediana, Grande)} = \frac{13}{20} Gini(Evaluación1) + \frac{7}{20} Gini(Evaluación2)$

$$= \frac{13}{20} \left(1 - \left(\frac{11}{13}\right)^2 - \left(\frac{2}{13}\right)^2\right) + \frac{7}{20} \left(1 - \left(\frac{4}{7}\right)^2 - \left(\frac{3}{7}\right)^2\right) = 0.3406$$

Se Calcula  $Gini_{TamañoCajuela \in (Pequeña, Grande)} = \frac{13}{20} Gini(Evaluación1) + \frac{7}{20} Gini(Evaluación2)$

$$= \frac{13}{20} \left(1 - \left(\frac{9}{13}\right)^2 - \left(\frac{4}{13}\right)^2\right) + \frac{7}{20} \left(1 - \left(\frac{6}{7}\right)^2 - \left(\frac{1}{7}\right)^2\right) = 0.3626$$

$$SplitInfoSeguridad = -\frac{9}{20} \log_2\left(\frac{9}{20}\right) - \frac{7}{20} \log_2\left(\frac{7}{20}\right) - \frac{4}{20} \log_2\left(\frac{4}{20}\right) = 1.5128$$

$$GainRatio(Seguridad) = \frac{0.0324}{1.5128} = 0.0214$$

Se Calcula  $Gini_{Seguridad \in (Baja, Mediana)} = \frac{11}{20} Gini(Evaluación1) + \frac{9}{20} Gini(Evaluación2)$

$$= \frac{11}{20} \left(1 - \left(\frac{8}{11}\right)^2 - \left(\frac{3}{11}\right)^2\right) + \frac{9}{20} \left(1 - \left(\frac{7}{9}\right)^2 - \left(\frac{2}{9}\right)^2\right) = 0.3737$$

Se Calcula  $Gini_{Seguridad \in (Mediana, Alta)} = \frac{16}{20} Gini(Evaluación1) + \frac{4}{20} Gini(Evaluación2)$

$$= \frac{16}{20} \left(1 - \left(\frac{13}{16}\right)^2 - \left(\frac{3}{16}\right)^2\right) + \frac{4}{20} \left(1 - \left(\frac{2}{4}\right)^2 - \left(\frac{2}{4}\right)^2\right) = 0.3437$$

Se Calcula  $Gini_{Seguridad \in (Baja, Alta)} = \frac{13}{20} Gini(Evaluación1) + \frac{7}{20} Gini(Evaluación2)$

$$= \frac{13}{20} \left(1 - \left(\frac{9}{13}\right)^2 - \left(\frac{4}{13}\right)^2\right) + \frac{7}{20} \left(1 - \left(\frac{6}{7}\right)^2 - \left(\frac{1}{7}\right)^2\right) = 0.3626$$

## 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.

El conjunto de datos que planteamos es el siguiente:

	A	B	C	D	E	F
1	precioCompra	costoManto	noLuces	noPasajeros	tamMotor	evalMoto
2	vhigh	med	dos	dos	small	acc
3	med	vhigh	una	tres	big	unacc
4	high	low	dos	uno	med	acc
5	med	low	una	tres	small	acc
6	vhigh	high	tres	uno	big	acc
7	med	vhigh	una	tres	med	unacc
8	med	low	una	uno	big	acc
9	high	med	dos	uno	small	acc
10	med	high	tres	tres	med	unacc
11	med	vhigh	una	uno	big	unacc
12	vhigh	low	dos	dos	small	acc
13	high	high	tres	uno	med	unacc
14	med	low	una	tres	big	acc
15	vhigh	med	dos	dos	small	acc
16	high	low	tres	dos	big	unacc

Aplicación del proceso del cálculo de medidas de evaluación al conjunto de datos de motos (accesible, no accesible)

## Paso 1: Cálculo de la entropía total

Evaluación Moto	
Accesible	No accesible
9	6

$$E(\text{Total}) = -\frac{9}{15} \log_2 \left( \frac{9}{15} \right) - \frac{6}{15} \log_2 \left( \frac{6}{15} \right) = 0.9709$$

## Paso 2: Dividir el conjunto de datos en los diversos atributos

Atributo objetivo	Evaluación Motos
Atributo	Dominio
Precio Compra	Muy alto
	Alto
	Medio
Costo Mantenimiento	Muy alto
	Alto
	Medio
	Bajo
Número Luces	Uno
	Dos
	Tres
Número Pasajeros	Uno
	Dos
	Tres
Tamaño Motor	Pequeña
	Mediana
	Grande

## Paso 3: Se calcula la entropía de cada rama y se suman proporcionalmente para calcular la entropía total

Para Precio Compra

Precio Compra	Muy alto	Accesible	4
		No accesible	0
	Alto	Accesible	2
		No accesible	2
	Mediano	Accesible	3
		No accesible	4

$$E(\text{Muy alto}) = -\frac{4}{15} \log_2 \left( \frac{4}{15} \right) - \frac{0}{15} \log_2 \left( \frac{0}{15} \right) = 0.5085$$

$$E(\text{Alto}) = -\frac{2}{15} \log_2 \left( \frac{2}{15} \right) - \frac{2}{15} \log_2 \left( \frac{2}{15} \right) = 0.7751$$

$$E(\text{Mediano}) = -\frac{3}{15} \log_2 \left( \frac{3}{15} \right) - \frac{4}{15} \log_2 \left( \frac{4}{15} \right) = 0.9728$$

$$E(\text{Precio Compra, Evaluación}) = \frac{4}{15}(0.5085) + \frac{4}{15}(0.7751) + \frac{7}{15}(0.9728) = 0.7962$$

Para Costo Mantenimiento

Costo Mantenimiento	Muy alto	Accesible	0
		No accesible	3
	Alto	Accesible	1
		No accesible	2
	Medio	Accesible	3
		No accesible	0
	Bajo	Accesible	5
		No accesible	1

$$E(\text{Muy alto}) = -\frac{0}{15} \log_2 \left( \frac{0}{15} \right) - \frac{3}{15} \log_2 \left( \frac{3}{15} \right) = 0.4643$$

$$E(\text{Alto}) = -\frac{1}{15} \log_2 \left( \frac{1}{15} \right) - \frac{2}{15} \log_2 \left( \frac{2}{15} \right) = 0.6480$$

$$E(\text{Mediano}) = -\frac{3}{15} \log_2 \left( \frac{3}{15} \right) - \frac{4}{15} \log_2 \left( \frac{4}{15} \right) = 0.9728$$

$$E(\text{Bajo}) = -\frac{3}{15} \log_2 \left( \frac{3}{15} \right) - \frac{0}{15} \log_2 \left( \frac{0}{15} \right) = 0.4643$$

$$E(\text{Precio Mantenimiento, Evaluación}) = \frac{3}{15}(0.4643) + \frac{3}{15}(0.6480) + \frac{3}{15}(0.9728) + \frac{6}{15}(0.4643) = 0.6027$$

Para Número Luces

Número Luces	Una	Accesible	4
		No accesible	2
	Dos	Accesible	5
		No accesible	0
	Tres	Accesible	1
		No accesible	3

$$E(\text{Uno}) = -\frac{4}{15} \log_2 \left( \frac{4}{15} \right) - \frac{2}{15} \log_2 \left( \frac{2}{15} \right) = 0.8960$$

$$E(\text{Dos}) = -\frac{5}{15} \log_2 \left( \frac{5}{15} \right) - \frac{0}{15} \log_2 \left( \frac{0}{15} \right) = 0.5283$$

$$E(\text{Tres}) = -\frac{1}{15} \log_2 \left( \frac{1}{15} \right) - \frac{3}{15} \log_2 \left( \frac{3}{15} \right) = 0.7248$$

$$E(\text{Número Luces, Evaluación}) = \frac{6}{15}(0.8960) + \frac{5}{15}(0.5283) + \frac{4}{15}(0.7248) = 0.7277$$

Para Número Pasajeros

Número Pasajeros	Uno	Accesible	4
		No accesible	2
	Dos	Accesible	3
		No accesible	1
	Tres	Accesible	2
		No accesible	3

$$E(\text{Uno}) = -\frac{4}{15} \log_2 \left( \frac{4}{15} \right) - \frac{2}{15} \log_2 \left( \frac{2}{15} \right) = 0.8960$$

$$E(\text{Dos}) = -\frac{3}{15} \log_2 \left( \frac{3}{15} \right) - \frac{1}{15} \log_2 \left( \frac{1}{15} \right) = 0.7248$$



$$E(\text{Tres}) = -\frac{2}{15} \log_2 \left( \frac{2}{15} \right) - \frac{3}{15} \log_2 \left( \frac{3}{15} \right) = 0.8519$$

$$E(\text{Número Pasajeros, Evaluación}) = \frac{6}{15} (0.8960) + \frac{4}{15} (0.7248) + \frac{5}{15} (0.8519) = 0.8356$$

Para Tamaño Motor

Tamaño Motor	Grande	Accesible	3
		No accesible	3
	Mediana	Accesible	1
		No accesible	3
	Pequeña	Accesible	5
		No accesible	0

$$E(\text{Grande}) = -\frac{3}{15} \log_2 \left( \frac{3}{15} \right) - \frac{3}{15} \log_2 \left( \frac{3}{15} \right) = 0.9287$$

$$E(\text{Mediano}) = -\frac{1}{15} \log_2 \left( \frac{1}{15} \right) - \frac{3}{15} \log_2 \left( \frac{3}{15} \right) = 0.7248$$

$$E(\text{Pequeño}) = -\frac{5}{15} \log_2 \left( \frac{5}{15} \right) - \frac{0}{15} \log_2 \left( \frac{0}{15} \right) = 0.5283$$

$$E(\text{Tamaño Motor, Evaluación}) = \frac{6}{15} (0.9287) + \frac{4}{15} (0.7248) + \frac{5}{15} (0.5283) = 0.7408$$

## Paso 4: Se calcula la ganancia de información

- Para Precio Compra:  
Gain(Precio Compra, Evaluación) =  $0.9709 - 0.7962 = 0.1747$
- Para Costo Mantenimiento:  
Gain(Costo Mantenimiento, Evaluación) =  $0.9709 - 0.6027 = 0.3682$
- Para Número Luces:  
Gain(Número Puertas, Evaluación) =  $0.9709 - 0.7277 = 0.2432$
- Para Número Pasajeros:  
Gain(Número Pasajeros, Evaluación) =  $0.9709 - 0.8356 = 0.1353$
- Para Tamaño Motor:  
Gain(Tamaño Cajuela, Evaluación) =  $0.9709 - 0.7408 = 0.2301$

## Paso 5: Elección del nodo de Decisión.

Variable	Ganancia	(Accesible, No Accesible)
Precio de Compra	0.1747	Muy Alto = (4, 0) Alto = (2, 2) Medio = (3, 4)
Costo Mantenimiento	0.3682	Muy Alto = (0, 3) Alto = {1, 2} Medio = (3, 0) Bajo = (5, 1)
Número Luces	0.2432	Uno = (4, 2) Dos = (5, 0) Tres = (1, 3)
Número Pasajeros	0.1353	Uno = (4, 2) Dos = (3, 1) Tres = (2, 3)

Tamaño Motor	0.2301	Grande = (3, 3) Mediana = (1, 3) Pequeña = (5, 0)
--------------	--------	---

Costo Mantenimiento es la variable que da una mayor ganancia

Se calcula SplitInfo de Mantenimiento:

$$\text{SplitInfoMantenimiento} = -\frac{3}{15} \log_2 \left( \frac{3}{15} \right) - \frac{3}{15} \log_2 \left( \frac{3}{15} \right) - \frac{3}{15} \log_2 \left( \frac{3}{15} \right) - \frac{6}{15} \log_2 \left( \frac{6}{15} \right) = 1.9219$$

$$\text{GainRatio(Costo Mantenimiento)} = \frac{0.3682}{1.9219} = 0.3993$$

Se calcula Gini Index con respecto a la variable Evaluación

$$\text{Gini(Evaluación)} = 1 - \left( \frac{9}{15} \right)^2 - \left( \frac{6}{15} \right)^2 = 0.48$$

$$\text{Se Calcula } \text{Gini}_{\text{CostoMantenimiento} \in (\text{Bajo, Medio})} = \frac{17}{15} \text{Gini(Evaluación1)} + \frac{3}{15} \text{Gini(Evaluación2)}$$

$$= \frac{17}{20} \left( 1 - \left( \frac{15}{17} \right)^2 - \left( \frac{2}{17} \right)^2 \right) + \frac{3}{20} \left( 1 - \left( \frac{3}{3} \right)^2 - \left( \frac{0}{3} \right)^2 \right) = 0.1764$$

$$\text{Se Calcula } \text{Gini}_{\text{CostoMantenimiento} \in (\text{Bajo, Alto})} = \frac{8}{20} \text{Gini(Evaluación1)} + \frac{12}{20} \text{Gini(Evaluación2)}$$

$$= \frac{8}{20} \left( 1 - \left( \frac{3}{8} \right)^2 - \left( \frac{5}{8} \right)^2 \right) + \frac{12}{20} \left( 1 - \left( \frac{12}{12} \right)^2 - \left( \frac{0}{12} \right)^2 \right) = 0.1875$$

$$\text{Se Calcula } \text{Gini}_{\text{CostoMantenimiento} \in (\text{Medio, Muy alto})} = \frac{15}{20} \text{Gini(Evaluación1)} + \frac{5}{20} \text{Gini(Evaluación2)}$$

$$= \frac{15}{20} \left( 1 - \left( \frac{12}{15} \right)^2 - \left( \frac{3}{15} \right)^2 \right) + \frac{5}{20} \left( 1 - \left( \frac{3}{5} \right)^2 - \left( \frac{2}{5} \right)^2 \right) = 0.36$$

Proponemos realizar este proceso para los demás atributos, para ver si obtenemos valores menores de Gini Index

$$\text{SplitInfoPrecioCompra} = -\frac{10}{20} \log_2 \left( \frac{10}{20} \right) - \frac{7}{20} \log_2 \left( \frac{7}{20} \right) - \frac{3}{20} \log_2 \left( \frac{3}{20} \right) = 1.4406$$

$$\text{GainRatio(PrecioCompra)} = \frac{0.0497}{1.4406} = 0.0344$$

$$\text{Se Calcula } \text{Gini}_{\text{PrecioCompra} \in (\text{Medio, Alto})} = \frac{10}{20} \text{Gini(Evaluación1)} + \frac{10}{20} \text{Gini(Evaluación2)}$$

$$= \frac{10}{20} \left( 1 - \left( \frac{8}{10} \right)^2 - \left( \frac{2}{10} \right)^2 \right) + \frac{10}{20} \left( 1 - \left( \frac{7}{10} \right)^2 - \left( \frac{3}{10} \right)^2 \right) = 0.37$$

$$\text{Se Calcula } \text{Gini}_{\text{PrecioCompra} \in (\text{Medio, Muy alto})} = \frac{13}{20} \text{Gini(Evaluación1)} + \frac{7}{20} \text{Gini(Evaluación2)}$$

$$= \frac{13}{20} \left( 1 - \left( \frac{10}{13} \right)^2 - \left( \frac{3}{13} \right)^2 \right) + \frac{7}{20} \left( 1 - \left( \frac{5}{7} \right)^2 - \left( \frac{2}{7} \right)^2 \right) = 0.3736$$

$$\text{Se Calcula } \text{Gini}_{\text{PrecioCompra} \in (\text{Alto, Muy alto})} = \frac{17}{20} \text{Gini(Evaluación1)} + \frac{3}{20} \text{Gini(Evaluación2)}$$

$$= \frac{17}{20} \left( 1 - \left( \frac{12}{17} \right)^2 - \left( \frac{5}{17} \right)^2 \right) + \frac{3}{20} \left( 1 - \left( \frac{3}{3} \right)^2 - \left( \frac{0}{3} \right)^2 \right) = 0.3529$$

$$\text{SplitInfoNumeroLuces} = -\frac{8}{20} \log_2 \left( \frac{8}{20} \right) - \frac{4}{20} \log_2 \left( \frac{4}{20} \right) - \frac{5}{20} \log_2 \left( \frac{5}{20} \right) - \frac{3}{20} \log_2 \left( \frac{3}{20} \right) = 1.9037$$

$$\text{GainRatio(NumeroLuces)} = \frac{0.1517}{1.9037} = 0.0796$$

$$\text{Se Calcula } \text{Gini}_{\text{NúmeroPuertas} \in (\text{Dos, Tres})} = \frac{12}{20} \text{Gini(Evaluación1)} + \frac{8}{20} \text{Gini(Evaluación2)}$$

$$= \frac{12}{20} \left(1 - \left(\frac{8}{12}\right)^2 - \left(\frac{4}{12}\right)^2\right) + \frac{8}{20} \left(1 - \left(\frac{7}{8}\right)^2 - \left(\frac{1}{8}\right)^2\right) = 0.3541$$

Se Calcula  $Gini_{\text{NúmeroPuertas} \in (\text{Tres}, \text{Cuatro})} = \frac{9}{20} Gini(\text{Evaluación1}) + \frac{11}{20} Gini(\text{Evaluación2})$

$$= \frac{9}{20} \left(1 - \left(\frac{7}{9}\right)^2 - \left(\frac{2}{9}\right)^2\right) + \frac{11}{20} \left(1 - \left(\frac{8}{11}\right)^2 - \left(\frac{3}{11}\right)^2\right) = 0.3737$$

Se Calcula  $Gini_{\text{NúmeroPuertas} \in (\text{Cuatro}, \text{Cinco o más})} = \frac{8}{20} Gini(\text{Evaluación1}) + \frac{12}{20} Gini(\text{Evaluación2})$

$$= \frac{8}{20} \left(1 - \left(\frac{7}{8}\right)^2 - \left(\frac{1}{8}\right)^2\right) + \frac{12}{20} \left(1 - \left(\frac{8}{12}\right)^2 - \left(\frac{4}{12}\right)^2\right) = 0.3541$$

Se Calcula  $Gini_{\text{NúmeroPuertas} \in (\text{Dos}, \text{Cinco o más})} = \frac{11}{20} Gini(\text{Evaluación1}) + \frac{9}{20} Gini(\text{Evaluación2})$

$$= \frac{11}{20} \left(1 - \left(\frac{8}{11}\right)^2 - \left(\frac{3}{11}\right)^2\right) + \frac{9}{20} \left(1 - \left(\frac{7}{9}\right)^2 - \left(\frac{2}{9}\right)^2\right) = 0.3737$$

Se Calcula  $Gini_{\text{NúmeroPuertas} \in (\text{Dos}, \text{Cuatro})} = \frac{13}{20} Gini(\text{Evaluación1}) + \frac{7}{20} Gini(\text{Evaluación2})$

$$= \frac{13}{20} \left(1 - \left(\frac{9}{13}\right)^2 - \left(\frac{4}{14}\right)^2\right) + \frac{7}{20} \left(1 - \left(\frac{6}{7}\right)^2 - \left(\frac{1}{7}\right)^2\right) = 0.3626$$

Se Calcula  $Gini_{\text{NúmeroPuertas} \in (\text{Tres}, \text{Cinco})} = \frac{7}{20} Gini(\text{Evaluación1}) + \frac{13}{20} Gini(\text{Evaluación2})$

$$= \frac{7}{20} \left(1 - \left(\frac{6}{7}\right)^2 - \left(\frac{1}{7}\right)^2\right) + \frac{13}{20} \left(1 - \left(\frac{9}{13}\right)^2 - \left(\frac{4}{13}\right)^2\right) = 0.3626$$

$$\text{SplitInfoNúmeroPasajeros} = -\frac{3}{20} \log_2\left(\frac{3}{20}\right) - \frac{3}{20} \log_2\left(\frac{3}{20}\right) - \frac{9}{20} \log_2\left(\frac{9}{20}\right) - \frac{5}{20} \log_2\left(\frac{5}{20}\right) = 1.8394$$

$$\text{GainRatio(NúmeroPasajeros)} = \frac{0.4845}{1.8394} = 0.2634$$

Se Calcula  $Gini_{\text{NúmeroPasajeros} \in (\text{Dos}, \text{Tres})} = \frac{6}{20} Gini(\text{Evaluación1}) + \frac{14}{20} Gini(\text{Evaluación2})$

$$= \frac{6}{20} \left(1 - \left(\frac{2}{6}\right)^2 - \left(\frac{4}{6}\right)^2\right) + \frac{14}{20} \left(1 - \left(\frac{13}{14}\right)^2 - \left(\frac{1}{14}\right)^2\right) = 0.2261$$

Se Calcula  $Gini_{\text{NúmeroPasajeros} \in (\text{Tres}, \text{Cuatro})} = \frac{12}{20} Gini(\text{Evaluación1}) + \frac{8}{20} Gini(\text{Evaluación2})$

$$= \frac{12}{20} \left(1 - \left(\frac{10}{12}\right)^2 - \left(\frac{2}{12}\right)^2\right) + \frac{8}{20} \left(1 - \left(\frac{5}{8}\right)^2 - \left(\frac{3}{8}\right)^2\right) = 0.3541$$

Se Calcula  $Gini_{\text{NúmeroPasajeros} \in (\text{Dos}, \text{Cuatro})} = \frac{12}{20} Gini(\text{Evaluación1}) + \frac{8}{20} Gini(\text{Evaluación2})$

Como son el número de elementos el número de pasajeros cuando son dos y tres, el resultado es el mismo, o sea

Se Calcula  $Gini_{\text{NúmeroPasajeros} \in (\text{Tres}, \text{Más})} = \frac{8}{20} Gini(\text{Evaluación1}) + \frac{13}{20} Gini(\text{Evaluación2})$

Como son el número de elementos el número de pasajeros cuando son dos y tres, el resultado es el mismo

$$\text{SplitInfoTamañoMotor} = -\frac{6}{20} \log_2\left(\frac{6}{20}\right) - \frac{7}{20} \log_2\left(\frac{7}{20}\right) - \frac{7}{20} \log_2\left(\frac{7}{20}\right) = 1.5812$$

$$\text{GainRatio(TamañoMotor)} = \frac{0.0322}{1.5812} = 0.0203$$

Se Calcula  $Gini_{\text{TamañoCajuela} \in (\text{Pequeña}, \text{Mediana})} = \frac{14}{20} Gini(\text{Evaluación1}) + \frac{6}{20} Gini(\text{Evaluación2})$

$$= \frac{14}{20} \left(1 - \left(\frac{10}{14}\right)^2 - \left(\frac{4}{14}\right)^2\right) + \frac{6}{20} \left(1 - \left(\frac{5}{6}\right)^2 - \left(\frac{1}{6}\right)^2\right) = 0.3736$$

$$\begin{array}{l} \text{Se} \quad \quad \text{Calcula} \quad \quad Gini_{\text{TamañoCajuela} \in (\text{Mediana}, \text{Grande})} = \frac{13}{20} Gini(\text{Evaluación1}) + \\ \frac{7}{20} Gini(\text{Evaluación2}) \end{array}$$

$$= \frac{13}{20} \left( 1 - \left( \frac{11}{13} \right)^2 - \left( \frac{2}{13} \right)^2 \right) + \frac{7}{20} \left( 1 - \left( \frac{4}{7} \right)^2 - \left( \frac{3}{7} \right)^2 \right) = 0.3406$$

$$\begin{array}{l} \text{Se} \quad \quad \text{Calcula} \quad \quad Gini_{\text{TamañoCajuela} \in (\text{Pequeña}, \text{Grande})} = \frac{13}{20} Gini(\text{Evaluación1}) + \\ \frac{7}{20} Gini(\text{Evaluación2}) \end{array}$$

$$= \frac{13}{20} \left( 1 - \left( \frac{9}{13} \right)^2 - \left( \frac{4}{13} \right)^2 \right) + \frac{7}{20} \left( 1 - \left( \frac{6}{7} \right)^2 - \left( \frac{1}{7} \right)^2 \right) = 0.3626$$

## 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

Realizamos una modificación al orden de los datos, para una mejor visualización.

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

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

Medida	Cálculo	Explicación
Negativo verdadero	3	A son los ejemplares clasificados como negativos de forma adecuada (son negativos originalmente). <b>Negativo verdadero</b>
Positivo falso	2	B son los ejemplares clasificados como positivos de forma incorrecta, ya que son negativos originalmente. <b>Falso positivo</b>
Negativo falso	5	C son los ejemplares clasificados como negativos de forma incorrecta, ya que son positivos originalmente. <b>Falso negativo</b>
Positivo verdadero	40	D son los ejemplares clasificados como positivos de forma adecuada (son positivos originalmente). <b>Positivo verdadero</b>

Tasa de exactitud	$\frac{(\square + \square)}{(\square + \square + \square + \square)}$ $= \frac{(3 + 40)}{(3 + 2 + 5 + 40)}$ $= 86\%$	<p>Tasa de exactitud</p> <p>Las clasificaciones que hizo de forma correcta</p>
Tasa de error	$\frac{(\square + \square)}{(\square + \square + \square + \square)}$ $= \frac{(2 + 5)}{(3 + 2 + 5 + 40)}$ $= 14\%$	<p>La tasa de error</p> <p>Las clasificaciones que hizo de forma equivocada</p>
Precisión	$\frac{d}{b + d} = \frac{40}{2 + 40} = 95\%$	<p>Precisión</p> <p>La precisión (<i>Precision</i>) mide cuántos ejemplos clasificados como clase "positiva" son realmente "positivos".</p>
Sensibilidad ( <i>Recall</i> )	$\frac{d}{c + d} = \frac{40}{5 + 40} = 88\%$	<p>Recall</p> <p>La sensibilidad (<i>Sensitivity</i>) (también conocida como recuerdo (<i>recall</i>)) evalúa qué tan bien el clasificador puede reconocer muestras positivas</p>
Tasa de positivos falsos	$\frac{b}{a + b} = \frac{2}{3 + 2} = 40\%$	<p>Nos revela el porcentaje de falsos positivos según la matriz de confusión.</p>
Tasa de negativos falsos	$\frac{c}{c + d} = \frac{5}{5 + 40} = 11\%$	<p>Nos revela el porcentaje de falsos negativos según la matriz de confusión.</p>
Especificidad	$\frac{a}{a + b} = \frac{3}{3 + 2} = 60\%$	<p>Especificidad</p> <p>La medida de especificidad (<i>specificity</i>) mide que tan bien el clasificador puede reconocer las muestras negativas</p>

