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 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**

1. Describa la minería de datos

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

1. 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. |

1. ¿Cuál es la diferencia entre un árbol de clasificación y un árbol de regresión?

|  |
| --- |
| En el árbol de clasificación las variable 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.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Criterio** | **ID3** | **C4.5** | **CART** | **Random Forest** |
| 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  variable durante un período 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  datos de atributos nominales | Admite tanto continuo como  datos de atributos nominales |

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Criterio** | **ID3** | **C4.5** | **CART** | **Random Forest** |
| Ventajas | * Árbol rápido y conto * Busca en todo el conjunto de datos * Encuentra los nodos hoja permitiendo así que los datos de prueba sean podado y reduciendo el número de pruebas. | * C4.5 es fácil de implementar. * C4.5 crea modelos que se pueden interpretar fácilmente. * Puede manejar valores categóricos y continuos. * Puede lidiar con el ruido y lidiar con el valor faltante   atributos. | * CART puede manejar los valores perdidos automáticamente usando divisiones sustitutas. * Utiliza cualquier combinación de variables continuas / discretas. * CART realiza automáticamente la selección de variables. * CART puede establecer interacciones entre variables. | * Reconoce valores atípicos y anomalías en conocimientos datos. * Es uno de los algoritmos de aprendizaje más precisos. disponible. Para muchos conjuntos de datos, produce una gran clasificación de clasificadores precisos. * Da una estimación de las variables importantes en clasificación. |
| Desventajas | * Para una muestra pequeña, los datos pueden estar sobre ajustados o sobre clasificado. * Para tomar una decisión, solo se prueba un atributo en un instantáneo consumiendo así mucho tiempo. | * Una pequeña variación en los datos puede llevar a diferentes decisiones de árboles cuando se usa C4.5. * Para un conjunto de entrenamiento pequeño, C4.5 no funciona muy bien. | * CART puede tener árboles de decisión inestables. * CART se divide solo por una variable. * No paramétrico. | * A veces la clasificación hecha por Random Forest son difíciles de interpretar por los humanos. * Random Forest a veces se sobreajuste con conjuntos de datos con Tareas ruidosas de clasificación / regresión. |

**Ejercicio 3**

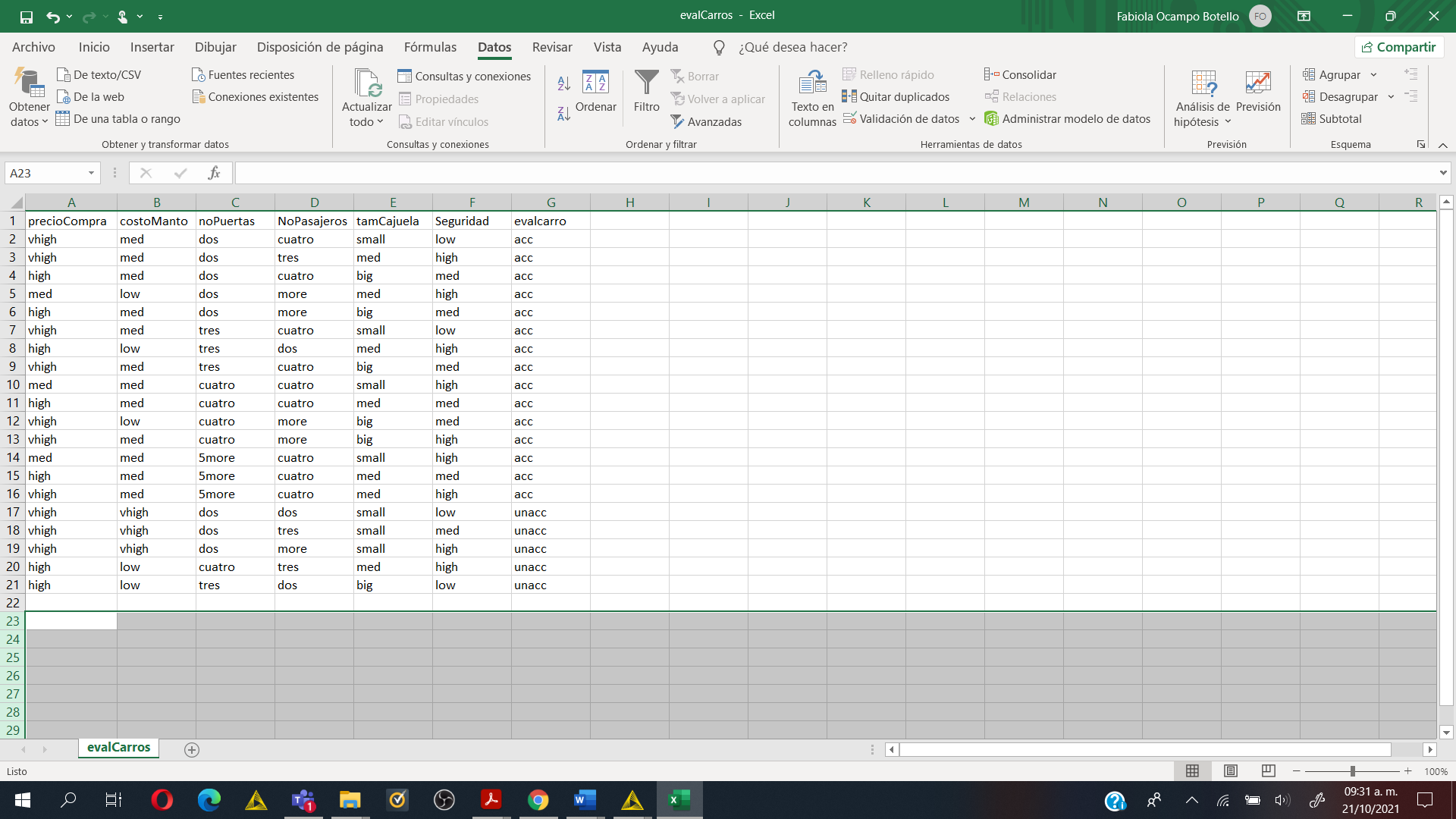
Considerando los siguientes criterios de selección de atributo para particionamiento: *Entropy (Information Gain)*, *Gain Ratio* and *Gini Index*. Realice una descripción con sus propias palabras de cada uno de ellos.

|  |  |
| --- | --- |
| **Criterio** | **Descripción** |
| Entropy (Information Gain) | Valor entre 0-1 que índica 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.



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(Total) =

# 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(Muy alto) = = 0.9406

E(Alto) = = 0.8321

E(Mediano) =

E(Precio Compra, Evaluación) = = 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(Muy alto) =

E(Alto) =

E(Mediano) = = 0.7427

E(Precio Compra, Evaluación) = = 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(Dos) = = 0.9105

E(Tres) = = 0.6266

E(Cuatro) = = 0.6804

E(Cinco o más) =

E(Número Puertas, Evaluación) = = 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(Dos) = = 0.5482

E(Tres) = = 0.5482

E(Cuatro) =

E(Cinco o más) = = 0.6804

E(Número Pasajeros, Evaluación) = = 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(Grande) = = 0.715

E(Mediana) = = 0.7371

E(Pequeña) = = 0.8749

E(Tamaño Cajuela, Evaluación) = = 0.779

Para Seguridad

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

E(Grande) = = 0.8622

E(Mediana) = = 0.7371

E(Pequeña) = = 0.6643

E(Tamaño Cajuela, Evaluación) = = 0.7788

# Paso 4: Se calcula la ganancia de información

Para Precio Compra: Gain(PrecioCompra, Evaluación) = 0.8112 – 0.7615 = 0.0497

Para Costo Mantenimiento: Gain(Costo Mantenimiento, Evaluación) = 0.8112 – 0.1856 = 0.6256

Para Número Puertas: Gain(Número Puertas, Evaluación) = 0.8112 – 0.6595 = 0.1517

Para Número Pasajeros: Gain(Número Pasajeros, Evaluación) = 0.8112 – 0.3267 = 0.4845

Para Tamaño de Cajuela: Gain(Tamaño Cajuela, Evaluación) = 0.8112 – 0.779 = 0.0322

Para Seguridad: Gain(Seguridad, 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:

SplitInfoMantenimimento = = 1.3527

GainRatio(Costo Mantenimiento) =

Se calcula Gini Index con respecto a la variable Evaluación

Gini(Evaluación) = 1-

Se Calcula

=

Se Calcula

=

Se Calcula

=

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

SplitInfoPrecioCompra = = 1.4406

GainRatio(PrecioCompra) =

Se Calcula

=

Se Calcula

=

Se Calcula

=

SplitInfoNumeroPuertas = = 1.9037

GainRatio(NumeroPuestas) =

Se Calcula

=

Se Calcula

=

Se Calcula

=

Se Calcula

=

Se Calcula

=

Se Calcula

=

SplitInfoNúmeroPasajeros = = 1.8394

GainRatio(NúmeroPasajeros) = 0.2634

Se Calcula

=

Se Calcula

=

Se Calcula

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

Se Calcula

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

SplitInfoTamañoCajuela = = 1.5812

GainRatio(TamañoCajuela) =

Se Calcula

=

Se Calcula

=

Se Calcula

=

SplitInfoSeguridad = = 1.5128

GainRatio(Seguridad) =

Se Calcula

=

Se Calcula

=

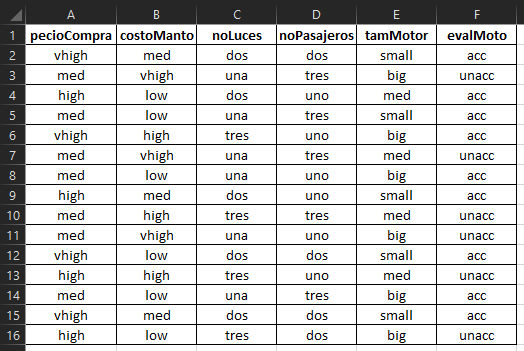
Se Calcula

=

**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:



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(Total) =

# 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(Muy alto) =

E(Alto)=

E(Mediano)

E(Precio Compra, Evaluación) =

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(Muy alto)

E(Alto)

E(Mediano)

E(Bajo)

E(Precio Mantenimiento, Evaluación) =

Para Número Luces

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

E(Uno)

E(Dos)

E(Tres)

E(Número Luces, Evaluación)=

Para Número Pasajeros

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

E(Uno)

E(Dos)

E(Tres)

E(Número Pasajeros, Evaluación) =

Para Tamaño Motor

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

E(Grande)

E(Mediano)

E(Pequeño)

E(Tamaño Motor, Evaluación) =

# Paso 4: Se calcula la ganancia de información

* Para Precio Compra:

Gain(PrecioCompra, Evaluación)

* Para Costo Mantenimiento:

Gain(Costo Mantenimiento, Evaluación)

* Para Número Luces:

Gain(Número Puertas, Evaluación)

* Para Número Pasajeros:

Gain(Número Pasajeros, Evaluación)

* Para Tamaño Motor:

Gain(Tamaño Cajuela, Evaluación)

# 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:

*SplitInfoMantenimimento =*

*GainRatio(Costo Mantenimiento) =*

Se calcula Gini Index con respecto a la variable Evaluación

Gini(Evaluación) =

Se Calcula

=

Se Calcula

=

Se Calcula

=

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

*SplitInfoPrecioCompra = = 1.4406*

*GainRatio(PrecioCompra) =*

*Se Calcula*

*=*

*Se Calcula*

*=*

*Se Calcula*

*=*

*SplitInfoNumeroLuces = = 1.9037*

*GainRatio(NumeroLuces) =*

*Se Calcula*

*=*

*Se Calcula*

*=*

*Se Calcula*

*=*

*Se Calcula*

*=*

*Se Calcula*

*=*

*Se Calcula*

*=*

*SplitInfoNúmeroPasajeros = = 1.8394*

*GainRatio(NúmeroPasajeros) = 0.2634*

*Se Calcula*

*=*

*Se Calcula*

*=*

*Se Calcula*

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

*Se Calcula*

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

*SplitInfoTamañoMotor = = 1.5812*

*GainRatio(TamañoMotor) =*

*Se Calcula*

*=*

*Se Calcula*

*=*

*Se Calcula*

*=*

**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). **Negativo verdadero** |
| Positivo falso | 2 | B son los ejemplares clasificados como positivos de forma incorrecta, ya que son negativos originalmente. **Falso positivo** |
| Negativo falso | 5 | C son los ejemplares clasificados como negativos de forma incorrecta, ya que son positivos originalmente. **Falso negativo** |
| Positivo verdadero | 40 | D son los ejemplares clasificados como positivos de forma adecuada (son positivos originalmente). **Positivo verdadero** |
| Tasa de exactitud |  | Tasa de exactitud  Las clasificaciones que hizo de forma correcta |
| Tasa de error |  | La tasa de error  Las clasificaciones que hizo de forma equivocada |
| Precisión |  | Precisión  La precisión (*Precision*) mide cuántos ejemplos clasificados como clase "positiva" son realmente "positivos". |
| Sensibilidad (*Recall*) |  | Recall  La sensibilidad (*Sensitivity*) (también conocida como recuerdo (*recall*)) evalúa qué tan bien el clasificador puede reconocer muestras positivas |
| Tasa de positivos falsos |  | Nos revela el porcentaje de falsos positivos según la matriz de confusión. |
| Tasa de negativos falsos |  | Nos revela el porcentaje de falsos negativos según la matriz de confusión. |
| Especificidad |  | Especificidad  La medida de especificidad (*specificity*) mide que tan bien el clasificador puede reconocer las muestras negativas |