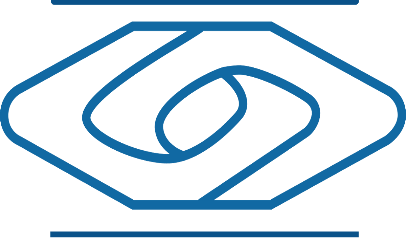
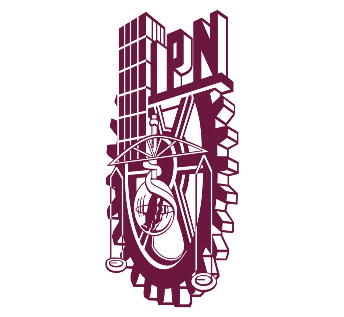
**INSTITUTO POLITÉCNICO NACIONAL**

**ESCUELA SUPERIOR DE CÓMPUTO**

**Data Mining**

**TAREA 5.** Árboles

**Equipo: 4**

**Alumnos:**

* Flores Ponce Alan Marcelo
* García Cruz Octavio Arturo
* Sampayo Hernández Mauro

**Grupo: 3CV15**

**Profesora:** Ocampo Botello Fabiola

**INSTITUTO POLITÉCNCO NACIONAL**

**ESCUELA SUPERIOR DE CÓMPUTO**

**MATERIAL EDUCATIVO PARA LA UNIDAD DE APRENDIZAJE DE DATA MINING 2023–2**

*Grupo 3CV15*

**PRACTICA DE ÁRBOLES DE DECISIÓN**

|  |  |
| --- | --- |
| Equipo No.: | 4 |

|  |
| --- |
| Nombres: |
| Flores Ponce Alan Marcelo |
| García Cruz Octavio Arturo |
| Sampayo Hernández Mauro |

**PRIMERA PARTE. NIVEL DE INGRESO**

1. **Descripción del conjunto de datos.**

**Autores del conjunto de datos:**

**Donante:** Ronny Kohavi and Barry Becker. Data Mining and Visualization. Silicon Graphics. e-mail: ronnyk '@' live.com for questions.

**Enlace de acceso:** <https://archive.ics.uci.edu/ml/datasets/adult>

1. **Objetivo de la práctica.**

Establecer tres objetivos para cada uno de los siguientes tipos de árboles: ID3, C4.5 y CART.

|  |  |
| --- | --- |
| **Tipo de árbol** | **Objetivo** |
| **ID3** | Generar un modelo de clasificación y las reglas de decisión correspondientes que reflejen las características demográficas distintivas de personas adultas que cuenten con un **Ingreso** *MAYOR, MENOR* o *IGUAL* a los 50K; con la intención de identificar muestras futuras.   * **Atributo Objetivo: Ingreso (income)**, el cual tiene dos clases: *<=50k* y *>50k.* * **Atributos Independientes:** workclass, education, marital-status, occupation, race, sex y native-country. |
| **C4.5** | Generar un modelo de clasificación y las reglas de decisión correspondientes que nos permita identificar las características demográficas distintivas de personas adultas dependiendo del **Nivel de Educación** que tengan, con la intención de identificar muestras futuras.   * **Atributo Objetivo: Nivel de educación (education)**, el cual tiene las siguientes clases: *Bachelors, Some-college, 11th, HS-grad, Prof-school, Assoc-acdm, Assoc-voc, 9th, 7th-8th, 12th, Masters, 1st-4th, 10th, Doctorate* y *5th-6th, Preschool.* * **Atributos Independientes:** age, workclass, education, marital-status, occupation, race, sex, hours-per-week, native-country e income. |
| **CART** | Generar un modelo de clasificación para estimar las **Horas de Trabajo Semanales** considerando sus características demográficas, con la intención de estimar muestras futuras.   * **Atributo Objetivo: Horas de trabajo semanales (hours-per-week)** * **Atributos Independientes:** age, workclass, education, education-num, marital-status, occupation, relationship, race, sex, capital-gain, native-country e income. |

1. **Diccionario de datos.**

Construya el diccionario de datos considerando la siguiente estructura.

|  |  |  |  |
| --- | --- | --- | --- |
| **No** | **Nombre** | **Tipo** | **Dominio** |
| **0** | age | Numérico | Números enteros Positivos |
| **1** | workclass | Categórico | Clase de Trabajo |
| **2** | fnlwgt | Numérico | Números enteros Positivos |
| **3** | education | Categórico | Bachelors, Some-college, 11th, HS-grad, Prof-school, Assoc-acdm, Assoc-voc, 9th, 7th-8th, 12th, Masters, 1st-4th, 10th, Doctorate, 5th-6th, Preschool |
| **4** | education-num | Numérico | Números enteros Positivos |
| **5** | marital-status | Categórico | Married-civ-spouse, Divorced, Never-married, Separated, Widowed, Married-spouse-absent, Married-AF-spouse |
| **6** | occupation | Categórico | Ocupación del individuo |
| **7** | relationship | Categórico | Wife, Own-child, Husband, Not-in-family, Other-relative, Unmarried |
| **8** | race | Categórico | White, Asian-Pac-Islander, Amer-Indian-Eskimo, Other, Black |
| **9** | sex | Categórico | Female, Male |
| **10** | capital-gain | Numérico | Números enteros Positivos |
| **11** | capital-loss | Numérico | Números enteros Positivos |
| **12** | hours-per-week | Numérico | Número de horas |
| **13** | native-country | Categórico | Nombre del país nativo del individuo |
| **14** | income | categórico | Nivel de ingresos del individuo |

1. **Resultados**

Presente los resultados considerando lo siguiente:

* Realice y describa los resultados de cinco consultas descriptivas en el conjunto de datos

1. ***Edad promedio* de personas adultas de acuerdo a su *nivel de ingresos (<= 50K o > 50K)*, clasificadas en base a su *clase trabajador*a.**

|  |  |
| --- | --- |
| **Nodo** | GroupBy |
| **Configuración** |  |
| **Resultados** |  |
| **Explicación Resultados** | * Todas las personas adultas que *nunca han trabajado* tienen ingresos MENORES a los 50K, y son por lo general jóvenes que rondan la edad de 20 años. * Todas las personas adultas que *trabajan sin recibir un salario específico* tienen ingresos MENORES a los 50K. * Las edades promedio para la mayoría de las clases trabajadoras, son ligeramente más altas para aquellos adultos que tienen ingresos MAYORES A los 50K, con edades promedio que van de entre los 40 a 56 años; mientras que, para aquellos con ingresos MENORES a los 50K, las edades promedio van desde los 35 a los 42 años. |

1. **Horas de Trabajo Semanales promedio de cada clase trabajadora.**

|  |  |
| --- | --- |
| **Nodo** | GroupBy |
| **Configuración** |  |
|  |  |
| **Resultados** |  |
| **Explicación Resultados** | * Todas las clases trabajadoras, a excepción de que aquellos que *no especificaron* la clase trabajadora a la que pertenecen, o *no pertenecen* a ninguna,trabajan en promedio 40 horas a la semana * Lo adultos que *trabajan por cuenta propia* (clases *Self-enp-inc*  y *Self-emp-not-inc)* son los que más horas trabajan a la semana, trabajando 48 y 44 horas semanales en promedio respectivamente. |

1. **El porcentaje de personas adultas de cierta raza presentes en cada clase trabajadora.**

|  |  |
| --- | --- |
| **Nodo** | CrossTable |
| **Configuración** |  |
| **Resultados** |  |
| **Explicación Resultados** | * En todas las clases trabajadoras predominan las personas de *raza blanca* (siendo un entre un 70% u 80% del total de trabajadores en la mayoría de los casos), seguido de las personas de *raza negra*. |

1. **El porcentaje de personas adultas que cuentan con cierto Nivel Educativo presentes en cada clase trabajadora.**

|  |  |
| --- | --- |
| **Nodo** | CrossTable |
| **Configuración** |  |
| **Resultados** |  |
| **Explicación Resultados** | * En todas las clases trabajadoras predominan las personas que han concluido estudios de nivel *Superior* (**Tertiary Education**), seguidos por las personas que han concluido estudios de nivel *Medio Superior* (**High School**). * En todas las clases trabajadoras, la presencia de las personas que solo cuentan con educación *primaria* (**Elementary School**) es prácticamente inexistente. |

1. **El porcentaje de personas adultas de cierto sexo presentes en cada clase trabajadora.**

|  |  |
| --- | --- |
| **Nodo** | CrossTable |
| **Configuración** |  |
| **Resultados** |  |
| **Explicación Resultados** | * En todas las clases trabajadoras predominan las personas de *sexo masculino*. |

* Presente propiedades estadísticas del conjunto de datos
* Describa las medidas generadas a partir de la matriz de confusión (archivo anexo)
* Analice este comportamiento en función la cantidad de elementos de cada tipo que existen en el conjunto de datos
* Anexe el modelo y las reglas generadas.

**INDICACIONES:**

* Genere un diccionario de datos de todos los atributos del conjunto de datos
* Para cada uno de los tres tipos de árboles, realice lo siguiente:
  + Cree una portada para identificar cada tipo de árbol
  + Genere el diccionario de datos con los datos considerados en cada caso
  + Añada el diagrama general del árbol
  + Añada las pantallas de configuración correspondientes
  + Describa los resultados obtenidos de la matriz de confusión de los árboles ID3 y C4.5. Se anexan medidas a describir.
  + Presente las reglas generadas en cada caso.

**Árbol ID3**

*Predice si un individuo tiene un INGRESO* ***menor o igual a 50k (<= 50k)****, o* ***mayor a 50k (> 50k)****.*

|  |  |
| --- | --- |
|  | **FUNCIóN:**    Mauro Sampayo |

1. **Diccionario de datos.**

|  |  |  |  |
| --- | --- | --- | --- |
| **No** | **Nombre** | **Tipo** | **Dominio** |
| **1** | workclass | Categórico | Clase de Trabajo |
| **2** | education | Categórico | Bachelors, Some-college, 11th, HS-grad, Prof-school, Assoc-acdm, Assoc-voc, 9th, 7th-8th, 12th, Masters, 1st-4th, 10th, Doctorate, 5th-6th, Preschool |
| **3** | marital-status | Categórico | Married-civ-spouse, Divorced, Never-married, Separated, Widowed, Married-spouse-absent, Married-AF-spouse |
| **4** | occupation | Categórico | Ocupación del individuo |
| **5** | race | Categórico | White, Asian-Pac-Islander, Amer-Indian-Eskimo, Other, Black |
| **6** | sex | Categórico | Female, Male |
| **7** | native-country | Categórico | Nombre del país nativo del individuo |
| **8** | income | categórico | Nivel de ingresos del individuo |

1. **Diagrama general del árbol.**

A screenshot of a computer

Description automatically generated with low confidence

1. **Configuración.**

A diagram of a decision tree

Description automatically generated with low confidence

|  |  |
| --- | --- |
| **Nodo** | **Configuracion** |
| Partitioning | Se configuró que se usara el 70% del conjunto de datos como **conjunto de entrenamiento**, y el 30% restante como **conjunto de prueba**.  Se utilizó el *muestreo estratificado* para generar el árbol. |
| Decision Tree Learner |  |
| Decision Tree Predictor |  |

1. **Resultados Obtenidos.**

A screenshot of a computer

Description automatically generated with low confidence

|  |  |  |
| --- | --- | --- |
|  | **Cálculo** | **Explicación** |
| Negativo verdadero | **A** =1506 | El 15.44% de los registros fueron clasificados correctamente como **>50k.** |
| Positivo falso | **B** = 846 | El 8.66% de los registros fueron clasificados incorrectamente como **<=50k**, cuando en realidad eran **>50k.** |
| Negativo falso | **C** = 997 | El 10.2% de los registros fueron clasificados incorrectamente como **>50k**, cuando en realidad eran **<=50k.** |
| Positivo verdadero | **D** = 6420 | El 65.71% de los registros fueron clasificados correctamente como **<=50k.** |
| Tasa de exactitud |  | El 81% de los registros totales, fueron clasificados de forma CORRECTA. |
| Tasa de error |  | El 18% de los registros totales, fueron clasificados de forma EQUIVOCADA. |
| Precisión |  | El 60% de los ejemplos clasificados como clase **positiva**, son realmente **positivos**. |
| Sensibilidad (*Recall*) |  | El clasificador puede reconocer **muestras positivas** en el 64% de los casos. |
| Tasa de positivos falsos |  | La tasa de **positivos falsos** es del 13%. |
| Tasa de negativos falsos |  | La tasa de **negativos falsos** es del 35%. |
| Especificidad |  | El clasificador puede reconocer **muestras negativas** en el 86% de los casos. |

1. **Reglas Obtenidas.**

A screenshot of a computer

Description automatically generated with medium confidence

**Árbol C4.5**

*Predice el NIVEL DE EDUCACIÓN de un individuo.*

|  |  |
| --- | --- |
|  | **Abstract**    Mauro Sampayo |

1. **Diccionario de datos.**

|  |  |  |  |
| --- | --- | --- | --- |
| **No** | **Nombre** | **Tipo** | **Dominio** |
| **0** | age | Numérico | Números enteros Positivos |
| **1** | education | Categórico | Bachelors, Some-college, 11th, HS-grad, Prof-school, Assoc-acdm, Assoc-voc, 9th, 7th-8th, 12th, Masters, 1st-4th, 10th, Doctorate, 5th-6th, Preschool |
| **2** | occupation | Categórico | Ocupación del individuo |
| **3** | race | Categórico | White, Asian-Pac-Islander, Amer-Indian-Eskimo, Other, Black |
| **4** | sex | Categórico | Female, Male |
| **5** | native-country | Categórico | Nombre del país nativo del individuo |
| **6** | income | categórico | Nivel de ingresos del individuo |

1. **Diagrama general del árbol.**

Debido al tamaño del árbol generado, solo se mostrarán los primeros 2 niveles del diagrama. Sin embargo, el diagrama completo puede encontrarse dentro de la carpeta “*arboles generados*”, con el nombre de **nivelIngreso\_C4.5.**

A screenshot of a computer

Description automatically generated with medium confidence

1. **Configuración.**

Con el objetivo de que el predictor pudiese ser mas preciso, se englobaron los 15 valores del dominio de los **niveles de educación**, dentro de 4 grandes grupos, basándonos en los niveles que el sistema educativo de los Estados Unidos considera:

1. **Elementary School:** Engloba prescolar y los 6 primeros niveles de educación.
2. **Middle School:** Engloba el séptimo y octavo nivel de educación.
3. **High School:** Engloba los últimos cuatro niveles de educación (del noveno al doceavo).
4. **Tertiary Education:** Engloba todo tipo de estudios superiores (licenciaturas, maestrías, doctorados y derivados).

El tratamiento sobre estos datos se hizo con ayuda de un código programado en Python. Este código se adjunta a continuación.

**import** knime**.**scripting**.**io **as** knio

# --- Diccionario de Niveles de Educacion

education **=** **{**

"Preschool" **:** 'Elementary School'**,**

"1st-4th" **:** 'Elementary School'**,**

"5th-6th" **:** 'Elementary School'**,**

"7th-8th" **:** 'Middle School'**,**

"9th" **:** 'High School'**,**

"10th" **:** 'High School'**,**

"11th" **:** 'High School'**,**

"12th" **:** 'High School'**,**

"HS-grad" **:** 'High School'**,**

"Bachelors" **:** 'Tertiary Education'**,**

"Some-college" **:** 'Tertiary Education'**,**

"Prof-school" **:** 'Tertiary Education'**,**

"Assoc-acdm" **:** 'Tertiary Education'**,**

"Assoc-voc" **:** 'Tertiary Education'**,**

"Masters" **:** 'Tertiary Education'**,**

"Doctorate" **:** 'Tertiary Education'

**}**

# Convirtiendo la tabla de ENTRADA a dataFrame

df **=** knio**.**input\_tables**[**0**].**to\_pandas**()**

# --------------- TRATAMIENTO DE DATOS ---------------

# Tratamiendo de valores FALTANTES

**for** col **in** df**.**columns**:**

df**[**col**]=**df**[**col**].**replace**(**'?'**,** 'Not specified'**)**

# Tratamiento de valores de EDUCACION, para reducirlos a solo 4 grupos:

df**.**replace**({**"education"**:** education**},**inplace**=True)**

# --------------- TRATAMIENTO DE DATOS ---------------

# Convirtiendo el dataFrame a un dataTable e igualandolo a la tabla de SALIDA

knio**.**output\_tables**[**0**]** **=** knio**.**Table**.**from\_pandas**(**df**)**

A diagram of a tree learner

Description automatically generated with low confidence

|  |  |
| --- | --- |
| **Nodo** | **Configuracion** |
| X-Partitioner | Se utilizó el *muestreo estratificado* para generar el árbol, con 10 validaciones. |
| Decision Tree Learner |  |
| Decision Tree Predictor |  |
| X-Aggregator | A screenshot of a computer  Description automatically generated |

1. **Resultados Obtenidos.**

A screenshot of a computer

Description automatically generated with medium confidence

|  |  |  |
| --- | --- | --- |
| **Medida** | **Cálculo** | **Explicación** |
| Tasa de exactitud |  | El 66.79% de los registros totales, fueron clasificados de forma CORRECTA. |
| Tasa de error |  | El 33.2% de los registros totales, fueron clasificados de forma EQUIVOCADA. |

1. **Reglas Obtenidas.**

A screenshot of a computer

Description automatically generated with low confidence

**Árbol CART**

*Predice las HORAS DE TRABAJO SEMANALES de un individuo.*

|  |  |
| --- | --- |
|  | **FUNCIÓN:**    Mauro Sampayo |

1. **Diccionario de datos.**

|  |  |  |  |
| --- | --- | --- | --- |
| **No** | **Nombre** | **Tipo** | **Dominio** |
| **0** | age | Numérico | Números enteros Positivos |
| **1** | workclass | Categórico | Clase de Trabajo |
| **2** | education-num | Numérico | Números enteros Positivos |
| **3** | race | Categórico | White, Asian-Pac-Islander, Amer-Indian-Eskimo, Other, Black |
| **4** | sex | Categórico | Female, Male |
| **5** | hours-per-week | Numérico | Número de horas |
| **6** | native-country | Categórico | Nombre del país nativo del individuo |
| **7** | income | categórico | Nivel de ingresos del individuo |

1. **Diagrama general del árbol.**

Debido al gran tamaño del árbol generado, solo se mostrarán los primeros 3 niveles del diagrama. Sin embargo, dentro de la carpeta “*arboles generados*”, puede encontrar este diagrama extendido hasta el 5to nivel, con el nombre de **nivelIngreso\_CART.**

A screenshot of a computer

Description automatically generated with low confidence

1. **Configuración.**

A diagram of a tree learner

Description automatically generated with low confidence

|  |  |
| --- | --- |
| **Nodo** | **Configuracion** |
| X-Partitioner | Se utilizó el *muestreo aleatorio* para generar el árbol, con 10 validaciones. |
| Decision Tree Learner |  |
| Decision Tree Predictor |  |
| X-Aggregator |  |

**SEGUNDA PARTE. VENTAS DE PRODUCTOS**

Considere el conjunto de datos que trabajó para el tablero dinámico.

En este ejercicio se realizó una adaptación del conjunto de datos donado por: Aung Pyae

Intención del conjunto de datos original:

Este conjunto de datos contiene las ventas históricas de una empresa de supermercados registradas en 3 sucursales diferentes durante 3 meses.

Enlace de acceso al conjunto de datos original: <https://www.kaggle.com/aungpyaeap/supermarket-sales>

**INDICACIONES:**

* Genere un diccionario de datos de todos los atributos del conjunto de datos

|  |  |  |  |
| --- | --- | --- | --- |
| **No** | **Nombre** | **Tipo** | **Dominio** |
| **0** | Invoice ID | Numérico | Número de identificación |
| **1** | Branch | Categórico | Letra desde A-C |
| **2** | City | Categorico | Ubicación del Supermercado |
| **3** | Customer type | Categórico | Member o Normal |
| **4** | Gender | Categorico | Female, Male |
| **5** | Product line | Categórico | Categoria de Producto |
| **6** | Unit Price | Numérico | Precio por unidad |
| **7** | Quantity | Numerico | Cantidad de productos comprados |
| **8** | Tax 5% | Numerico | Porcentaje de Impuesto |
| **9** | Total | Numerico | Total pagado |
| **10** | Date | Fecha | Fecha de la compra |
| **11** | Time | Numérico | Hora de la compra: Día XX/Mes XX/Año XXXX |
| **12** | Payment | Numérico | Forma de pago Ewallet, Credit card, Cash |
| **13** | cogs | Numerico | Costo de productos vendidos |
| **14** | gross margin percentage | Numerico | Porcentaje de margen bruto |
| **15** | gross income | Numérico | Ingreso Bruto |
| **16** | Rating | Numérico | Calificación de estratificación del cliente en su experiencia de compra general (en una escala de 1 a 10) |

* Para cada uno de los tres tipos de árboles, realice lo siguiente:
  + Cree una portada para identificar cada tipo de árbol
  + Genere el diccionario de datos con los datos considerados en cada caso
  + Añada el diagrama general del árbol
  + Añada las pantallas de configuración correspondientes
  + Describa los resultados obtenidos de la matriz de confusión de los árboles ID3 y C4.5. Se anexan medidas a describir.
  + Presente las reglas generadas en cada caso.

|  |  |
| --- | --- |
| Use los siguientes nodos: |  |

**Árbol ID3**

*Predice el TIPO DE CLIENTE que realizó una compra, de acuerdo al producto que compró, su género y el tipo de pago que uso.*

|  |  |
| --- | --- |
|  | **FUNCIóN:**    Mauro Sampayo |

1. **Diccionario de datos.**

|  |  |  |  |
| --- | --- | --- | --- |
| **No** | **Nombre** | **Tipo** | **Dominio** |
| **0** | Customer type | Categórico | Member o Normal |
| **1** | Gender | Categorico | Female, Male |
| **2** | Product line | Categórico | Categoria de Producto |
| **3** | Payment | Numérico | Forma de pago Ewallet, Credit card, Cash |

1. **Diagrama general del árbol.**

Debido al tamaño del árbol generado, solo se mostrarán los primeros 2 niveles del diagrama. Sin embargo, el diagrama completo puede encontrarse dentro de la carpeta “*arboles generados*”, con el nombre de **ventasProductos\_ID3.**

A screenshot of a computer

Description automatically generated with low confidence

1. **Configuración.**

A picture containing text, screenshot, diagram, line

Description automatically generated

|  |  |
| --- | --- |
| **Nodo** | **Configuracion** |
| Partitioning | Se configuró que se usara el 80% del conjunto de datos como **conjunto de entrenamiento**, y el 20% restante como **conjunto de prueba**.  Se utilizó el *muestreo estratificado* para generar el árbol. |
| Decision Tree Predictor |  |
| Decision Tree Learner |  |

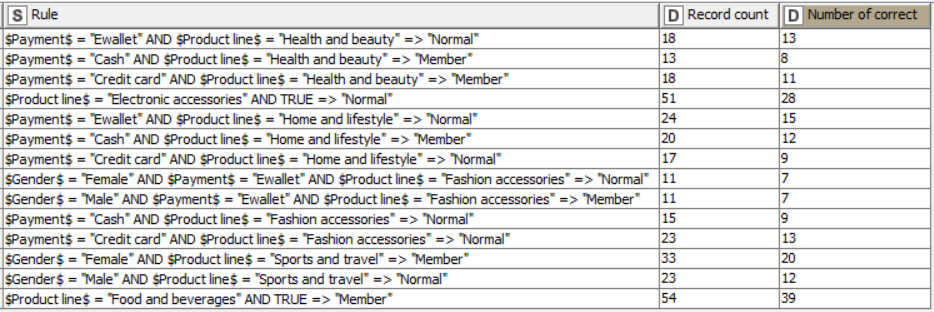
1. **Resultados Obtenidos.**

A screenshot of a member

Description automatically generated with medium confidence

|  |  |  |
| --- | --- | --- |
|  | **Cálculo** | **Explicación** |
| Negativo verdadero | **A** =23 | El 27.77% de los registros fueron clasificados correctamente como **Normal.** |
| Positivo falso | **B** = 17 | El 20.48% de los registros fueron clasificados fueron clasificados incorrectamente como **Member**, cuando en realidad eran **Normal.** |
| Negativo falso | **C** = 22 | El 26.5% de los registros fueron clasificados fueron clasificados incorrectamente como **Normal**, cuando en realidad eran **Member.** |
| Positivo verdadero | **D** = 21 | El 25.3% de los registros fueron clasificados correctamente como **Member.** |
| Tasa de exactitud |  | El 53% de los registros totales, fueron clasificados de forma CORRECTA. |
| Tasa de error |  | El 46% de los registros totales, fueron clasificados de forma EQUIVOCADA. |
| Precisión |  | El 51% de los ejemplos clasificados como clase **positiva**, son realmente **positivos**. |
| Sensibilidad (*Recall*) |  | El clasificador puede reconocer **muestras positivas** en el 57% de los casos. |
| Tasa de positivos falsos |  | La tasa de **positivos falsos** es del 51%. |
| Tasa de negativos falsos |  | La tasa de **negativos falsos** es del 42%. |
| Especificidad |  | El clasificador puede reconocer **muestras negativas** en el 48% de los casos. |

1. **Reglas Obtenidas.**



**Árbol C4.5**

*Predice el METODO DE PAGO que fue utilizado por un individuo.*

|  |  |
| --- | --- |
|  | **Abstract**    Mauro Sampayo |

1. **Diccionario de datos.**

|  |  |  |  |
| --- | --- | --- | --- |
| **No** | **Nombre** | **Tipo** | **Dominio** |
| **0** | Customer type | Categórico | Member o Normal |
| **1** | Gender | Categorico | Female, Male |
| **2** | Payment | Numérico | Forma de pago Ewallet, Credit card, Cash |

1. **Diagrama general del árbol.**

A picture containing text, diagram

Description automatically generated

1. **Configuración.**

A diagram of a tree learner

Description automatically generated with low confidence

|  |  |
| --- | --- |
| **Nodo** | **Configuracion** |
| X-Partitioner | Se utilizó el *muestreo estratificado* para generar el árbol, con 10 validaciones. |
| Decision Tree Predictor |  |
| Decision Tree Learner |  |
| X-Aggregator |  |

1. **Resultados Obtenidos.**

A picture containing text, screenshot, font, number

Description automatically generated

|  |  |  |
| --- | --- | --- |
| **Medida** | **Cálculo** | **Explicación** |
| Tasa de exactitud |  | El 33.57% de los registros totales, fueron clasificados de forma CORRECTA. |
| Tasa de error |  | El 66.42% de los registros totales, fueron clasificados de forma EQUIVOCADA. |

1. **Reglas Obtenidas.**

A screenshot of a computer

Description automatically generated with low confidence

**Árbol CART**

*Predice la CALIFICACIÓN otorgada por el cliente.*

|  |  |
| --- | --- |
|  | **FUNCIÓN:**    Mauro Sampayo |

1. **Diccionario de datos.**

|  |  |  |  |
| --- | --- | --- | --- |
| **No** | **Nombre** | **Tipo** | **Dominio** |
| **0** | Branch | Categórico | Letra desde A-C |
| **1** | Customer type | Categórico | Member o Normal |
| **2** | Gender | Categorico | Female, Male |
| **3** | Product line | Categórico | Categoria de Producto |
| **4** | Rating | Numérico | Calificación de estratificación del cliente en su experiencia de compra general (en una escala de 1 a 10) |

1. **Diagrama general del árbol.**

Debido al tamaño del árbol generado, solo se mostrarán los primeros 3 niveles del diagrama. Sin embargo, el diagrama completo puede encontrarse dentro de la carpeta “*arboles generados*”, con el nombre de **ventasProductos\_CART.**

A screenshot of a computer

Description automatically generated with low confidence

1. **Configuración.**

A diagram of a tree learner

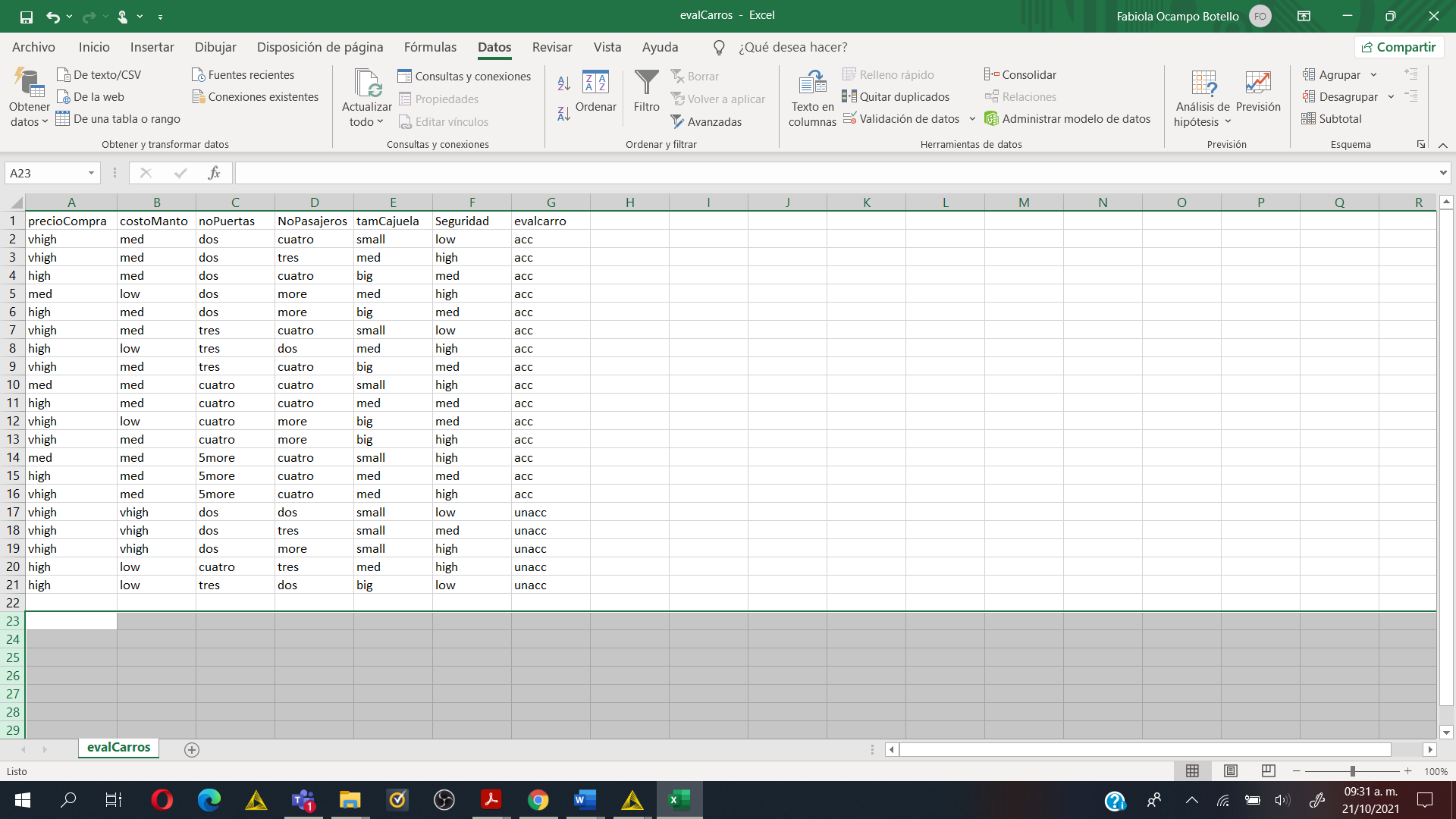
Description automatically generated with medium confidence

|  |  |
| --- | --- |
| **Nodo** | **Configuracion** |
| X-Partitioner | Se utilizó el *muestreo aleatorio* para generar el árbol, con 10 validaciones. |
| Decision Tree Learner |  |
| Decision Tree Predictor |  |
| X-Aggregator |  |

**TERCERA PARTE. CONSTRUCCIÓN DE ÁRBOL**

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.



1. **Se calcula la Entropía total**

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| |  |  | | --- | --- | | **evalcarro** | | | acc | unacc | | 15 | 5 | | No. total de Registros = 20 |

1. **Se divide el conjunto de datos en los diversos atributos**

|  |  |
| --- | --- |
| **Atributo Objetivo** | evalcarro |
|  |  |
| **Atributo** | **Dominio** |
| **precioCompra** | vhigh med high |
| **costoManto** | med  low  vhigh |
| **noPuertas** | dos tres cuatro 5more |
| **NoPasajeros** | cuatro tres more dos |
| **tamCajuela** | Small big  med |
| **Seguridad** | Low  High  med |

1. **Se calcula la Entropía en cada rama, y se suman proporcionalmente para calcular la Entropía total**
2. **Se calcula la Ganancia de Información de cada rama**

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | | | | **Count** |
| **precioCompra** | *vhigh* | **evalcarro** | ***acc*** | 7 |
| ***unacc*** | 3 |
| *Med* | **evalcarro** | ***acc*** | 3 |
| ***unacc*** | 0 |
| *high* | **evalcarro** | ***acc*** | 5 |
| ***unacc*** | 2 |

E(*vhigh*)= E(ACC, UNACC)= 0.94

E(*high*)= E(ACC, UNACC)= 0.83

E(*Med*)= E(ACC, UNACC)= 0.41=0

*Probabilidades*

*Vhigh: P(10/20)= 0.5 Med: P(3/20)=0.15 high: P(7/20)=0.35*

*E(*evalcarro, *precioCompra)= 0.5 \* 0.94 + 0.83 \* 0.35 + 0.15 \* 0 = 0.76*

*E(*evalcarro, *precioCompra)= 0.76*

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | | | | **Count** |
| **costoManto** | *med* | **evalcarro** | ***acc*** | 12 |
| ***unacc*** | 0 |
| *low* | **evalcarro** | ***acc*** | 3 |
| ***unacc*** | 2 |
| *vhigh* | **evalcarro** | ***acc*** | 0 |
| ***unacc*** | 3 |

E(*med*)= E(ACC, UNACC)= 0.44

E(*low*)= E(ACC, UNACC)= 0.74

E(*vhigh*)= E(ACC, UNACC)= 0.41

*Probabilidades*

*med: P(12/20)= 0.6 low: P(5/20)= 0.25 vhigh: P(3/20)= 0.15*

*E(*evalcarro, *costoManto)= 0.6 \* 0 + 0.25 \* 0.74 + 0.15 \* 0 = 0.185*

*E(*evalcarro, *costoManto) = 0.185*

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | | | | **Count** |
| **noPuertas** | *dos* | **evalcarro** | ***acc*** | 5 |
| ***unacc*** | 3 |
| *tres* | **evalcarro** | ***acc*** | 3 |
| ***unacc*** | 1 |
| *cuatro* | **evalcarro** | ***acc*** | 4 |
| ***unacc*** | 1 |
| *5more* | **evalcarro** | ***acc*** | 3 |
| ***unacc*** | 0 |

E(*dos*)= E(ACC, UNACC)= 0.91

E(*tres*)= E(ACC, UNACC)=0.62

E(*cuatro*)= E(ACC, UNACC)= 0.67

E(*5more*)= E(ACC, UNACC)= 0.41

*Probabilidades*

*dos: P(8/20)=0.4 tres: P(4/20)= 0.2*

*cuatro: P(5/20)=0.25 5moreP(3/20)= 0.15*

*E(*evalcarro, *noPuertas)= 0.4 \* 0.91 + 0.2 \* 0.62 + 0.25 \* 0.67 + 0.15 \* 0 = 0.65*

*E(*evalcarro, *noPuertas) = 0.65*

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | | | | **Count** |
| **noPasajeros** | *cuatro* | **evalcarro** | ***acc*** | 9 |
| ***unacc*** | 0 |
| *tres* | **evalcarro** | ***acc*** | 1 |
| ***unacc*** | 2 |
| *more* | **evalcarro** | ***acc*** | 4 |
| ***unacc*** | 1 |
| *dos* | **evalcarro** | ***acc*** | 1 |
| ***unacc*** | 2 |

E(Cuatro)= E(ACC, UNACC)= 0.51

E(tres)= E(ACC, UNACC)= 0.54

E(more)= E(ACC, UNACC)= 0.67

E(dos)= E(ACC, UNACC)= 0.54

*Probabilidades*

*cuatro: P(9/20)= 0.45 tres: P(3/20)= 0.15*

*more: P(5/20)=0.25 dos: P(3/20)= 0.15*

*E(*evalcarro, *noPasajeros)= 0.45 \* 0 + 0.15 \* 0.54 + 0.25 \* 0.67 + 0.15 \* 0.54= 0.32*

*E(*evalcarro, *noPasajeros) = 0.32*

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | | | | **Count** |
| **tamCajuela** | *small* | **evalcarro** | ***acc*** | 4 |
| ***unacc*** | 3 |
| *big* | **evalcarro** | ***acc*** | 5 |
| ***unacc*** | 1 |
| *med* | **evalcarro** | ***acc*** | 6 |
| ***unacc*** | 1 |

E(small)= E(ACC, UNACC)= 0.87

E(big)= E(ACC, UNACC)= 0.71

E(med)= E(ACC, UNACC)= 0.73

*Probabilidades*

*small: P(7/20)=0.35 big: P(6/20)=0.3 med: P(7/20)=0.35*

*E(*evalcarro, *tamCajuela)= 0.35 \* 0.87 + 0.3 \* 0.71 + 0.35 \* 0.73 =0.77*

*E(*evalcarro tamCajuela*) = 0.77*

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | | | | **Count** |
| **Seguridas** | *low* | **evalcarro** | ***acc*** | 2 |
| ***unacc*** | 2 |
| *high* | **evalcarro** | ***acc*** | 7 |
| ***unacc*** | 2 |
| *med* | **evalcarro** | ***acc*** | 6 |
| ***unacc*** | 1 |

E(low)= E(ACC, UNACC)= 0.66

E(high)= E(ACC, UNACC)= 0.86

E(med)= E(ACC, UNACC)= 0.73

*Probabilidades*

*low: P(4/20)=0.2 high: P(9/20)=0.45 med: P(7/20)=0.35*

*E(*evalcarro, *tamCajuela)= 0.2 \* 0.66 + 0.45 \* 0.86 + 0.35 \* 0.73=0.77*

*E(*evalcarro tamCajuela*) = 0.77*

1. **Elegir el Atributo con mayor ganancia de Informacion**

|  |  |
| --- | --- |
| ***Variable*** | ***Ganancia*** |
| precioCompra | *0.05* |
| costoManto | *0.625* |
| noPuertas | *0.16* |
| NoPasajeros | *0.49* |
| tamCajuela | *0.04* |
| Seguridad | *0.04* |

**CUARTA PARTE. CONSTRUCCIÓN DE ÁRBOL**

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.

Conjunto de Datos Propuesto Información de estudiantes de Estados Unidos.

A picture containing text, screenshot, number, font

Description automatically generated

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| |  |  | | --- | --- | | **Gender** | | | male | female | | 8 | 7 | | No. total de Registros = 20 |

**Se calcula la Ganancia de Información de cada rama**

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | | | | **Count** |
| **EthnicGroup** | *group A* | **Gender** | ***female*** | 1 |
| ***male*** | 2 |
| *group B* | **Gender** | ***female*** | 5 |
| ***male*** | 1 |
| *group C* | **Gender** | ***female*** | 1 |
| ***male*** | 3 |

E(*group A*)= E(female, male)= 0.64

E(*group B*)= E(female, male)= 0.78

E(*Med*)= E(ACC, UNACC)= 0.72

*Probabilidades*

*group A: P(3/15)= 0.2 group B: P(6/15)=0.4 group C: P(4/15)=0.26*

*E(*evalcarro, *precioCompra)= 0.2 \* 0.64 + 0.4 \* 0.78 + 0.26 \* 0.72 = 0.62*

*E(*evalcarro, *precioCompra)= 0.62*

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | | | | **Count** |
| **ParentEduc** | *bachelor's degree* | **Gender** | ***female*** | 2 |
| ***male*** | 1 |
| *some college* | **Gender** | ***female*** | 2 |
| ***male*** | 3 |
| *master's degree* | **Gender** | ***female*** | 1 |
| ***male*** | 0 |
|  | associate's degree | **Gender** | ***Female*** | 2 |
|  |  |  | ***male*** | 4 |

E(*bachelor's degree*)= E(female, male)= 0.64

E(*some college*)= E(female, male)= 0.85

E(*master's degree*)= E(female, male)= 0

E(associate's degree)= E(female, male)= 0.89

*Probabilidades*

*bachelor's degree: P(3/15)= 0.2*

*some college: P(5/15)= 0.33*

*master's degree: P(1/15)=0.06*

associate's degree*: P(6/15)=0.4*

*E(*evalcarro, *precioCompra)= 0.2 \* 0.64 + 0.33\*0.85 +0.06\*0 +0.4\*0.89=0.76*

*E(*evalcarro, *precioCompra)= 0.76*

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | | | | **Count** |
| **ParentMaritalStatus** | *married* | **Gender** | ***female*** | 5 |
| ***male*** | 5 |
| *single* | **Gender** | ***female*** | 1 |
| ***male*** | 2 |
| *widowed* | **Gender** | ***female*** | 1 |
| ***male*** | 0 |
|  | divorced | **Gender** | ***female*** | 0 |
|  |  |  | ***male*** | 1 |

E(*married*)= E(female, male)= 1.05

1.05

E(*single*)= E(female, male)= 0.64

E(*widowed*)= E(female, male)= 0

E(divorced)= E(female, male)= 0

*Probabilidades*

*married: P(10/15)= 0.66*

*single: P(3/15)= 0.2*

*master's degree: P(1/15)=0.06*

associate's degree*: P(1/15)=0.06*

*E(*evalcarro, *precioCompra)= 0.66 \* 1.05 + 0.2\*0.64 +0.06\*0 +0.06\*0 =0.82*

*E(*evalcarro, *precioCompra)= 0.82*

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | | | | **Count** |
| **TestPrep** | *none* | **Gender** | ***female*** | 6 |
| ***male*** | 5 |
| *completed* | **Gender** | ***female*** | 1 |
| ***male*** | 3 |

E(*none*)= E(female, male)=

E(*group B*)= E(female, male)= 0.78

*Probabilidades*

*married: P(11/15)=0.773*

*single: P(4/15)= 0.26*

|  |  |
| --- | --- |
| ***Variable*** | ***Ganancia*** |
| **EthnicGroup** | *0.37* |
| **ParentEduc** | *0.23* |
| **ParentMaritalStatus** | *0.17* |
| **TestPrep** | *0.26* |

**Elegir el Atributo con mayor ganancia de información**

**QUINTA PARTE. CONSTRUCCIÓN DE ÁRBOL**

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 | **A** = 3 | El 6% de los carros (3 de 50 totales) fueron clasificados correctamente como **No Accesibles.** |
| Positivo falso | **B** = 2 | El 4% de los carros (4 de 50 totales) fueron clasificados incorrectamente como **Accesibles**, cuando en realidad eran **No** **Accesibles.** |
| Negativo falso | **C** = 5 | El 10% de los carros (5 de 50 totales) fueron clasificados incorrectamente como **No Accesibles**, cuando en realidad eran **Accesibles.** |
| Positivo verdadero | **D** = 40 | El 80% de los carros (40 de 50 totales) fueron clasificados correctamente como **Accesibles.** |
| Tasa de exactitud |  | El 86% de los registros totales, fueron clasificados de forma CORRECTA. |
| Tasa de error |  | El 14% de los registros totales, fueron clasificados de forma EQUIVOCADA. |
| Precisión |  | El 95.23% de los ejemplos clasificados como clase **positiva**, son realmente **positivos**. |
| Sensibilidad (*Recall*) |  | El clasificador puede reconocer **muestras positivas** en el 88.88% de los casos. |
| Tasa de positivos falsos |  | La tasa de **positivos falsos** es del 40%. |
| Tasa de negativos falsos |  | La tasa de **negativos falsos** es del 11.11%. |
| Especificidad |  | El clasificador puede reconocer **muestras negativas** en el 60% de los casos. |