



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 POLITÉCNICO 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>

2. 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: $\leq 50k$ 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

	<p>demográficas distintivas de personas adultas dependiendo del Nivel de Educación que tengan, con la intención de identificar muestras futuras.</p> <ul style="list-style-type: none"> – Atributo Objetivo: Nivel de educación (education), el cual tiene las siguientes clases: <i>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</i>. – Atributos Independientes: age, workclass, education, marital-status, occupation, race, sex, hours-per-week, native-country e income.
CART	<p>Generar un modelo de clasificación para estimar las Horas de Trabajo Semanales considerando sus características demográficas, con la intención de estimar muestras futuras.</p> <ul style="list-style-type: none"> – 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.

3. Diccionario de datos.

Construya el diccionario de datos considerando la siguiente estructura.

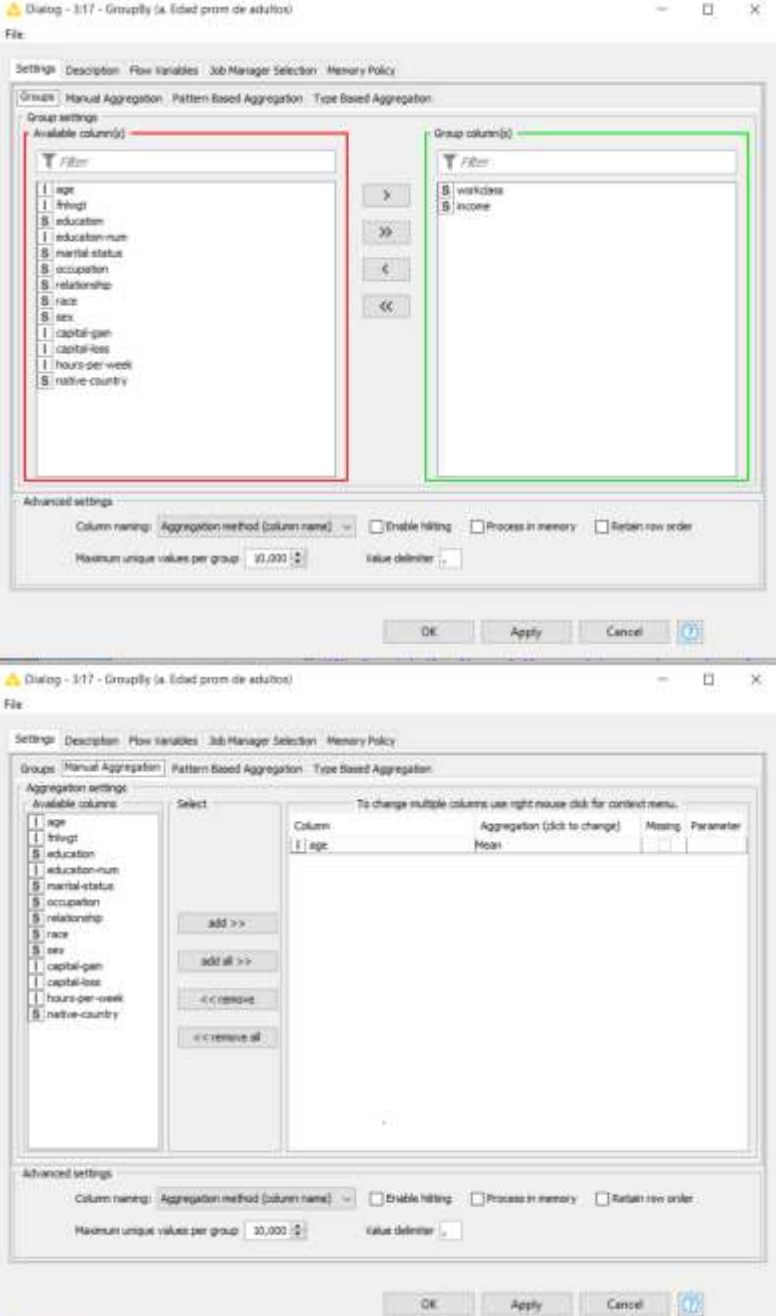
No	Nombre	Tipo	Dominio
0	age	Numérico	Números enteros Positivos
1	workclass	Categórico	Clase de Trabajo
2	fnlwgt	Numérico	Números enteros Positivos
3	education	Categórico	Bachelors, Some-college, 11th, HS-grad, Prof-school, Assoc-acdm, Assoc-voc, 9th, 7th-8th, 12th, Masters, 1st-4th, 10th, Doctorate, 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

4. Resultados

Presente los resultados considerando lo siguiente:

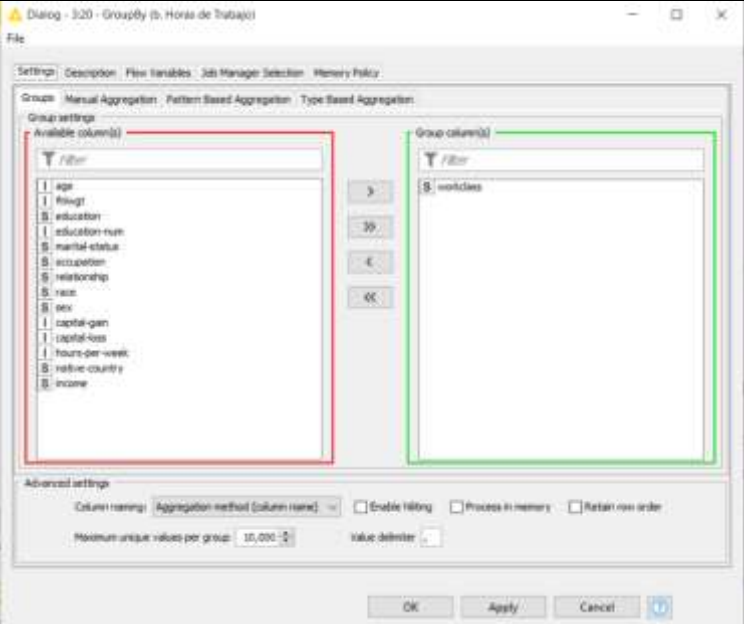
- Realice y describa los resultados de cinco consultas descriptivas en el conjunto de datos

a) **Edad promedio de personas adultas de acuerdo a su nivel de ingresos ($\leq 50K$ o $> 50K$), clasificadas en base a su clase trabajadora.**

Nodo	GroupBy
Configuración	 <p>The image displays two screenshots of the Orange3 GroupBy widget configuration dialog. The top screenshot shows the 'Group settings' tab, where the 'Available column(s)' list on the left contains various features like age, income, education, etc. The 'Group column(s)' list on the right contains 'workclass' and 'income'. The bottom screenshot shows the 'Aggregation settings' tab, where the 'Available columns' list on the left is the same. The 'Select' table on the right shows 'age' selected with the 'Mean' aggregation method. Both screenshots include an 'Advanced settings' section at the bottom with options for column naming, aggregation method, and other parameters.</p>

Resultados	<table><tr><th>Row ID</th><th>S workclass</th><th>S income</th><th>D Mean(age)</th></tr><tr><td>Row0</td><td>Federal-gov</td><td><=50K</td><td>40.625</td></tr><tr><td>Row1</td><td>Federal-gov</td><td>>50K</td><td>45.712</td></tr><tr><td>Row2</td><td>Local-gov</td><td><=50K</td><td>40.705</td></tr><tr><td>Row3</td><td>Local-gov</td><td>>50K</td><td>44.254</td></tr><tr><td>Row4</td><td>Never-worked</td><td><=50K</td><td>20.571</td></tr><tr><td>Row5</td><td>Not specified</td><td><=50K</td><td>39.258</td></tr><tr><td>Row6</td><td>Not specified</td><td>>50K</td><td>55.618</td></tr><tr><td>Row7</td><td>Private</td><td><=50K</td><td>35.113</td></tr><tr><td>Row8</td><td>Private</td><td>>50K</td><td>42.815</td></tr><tr><td>Row9</td><td>Self-emp-inc</td><td><=50K</td><td>43.206</td></tr><tr><td>Row10</td><td>Self-emp-inc</td><td>>50K</td><td>48.249</td></tr><tr><td>Row11</td><td>Self-emp-not-inc</td><td><=50K</td><td>44.389</td></tr><tr><td>Row12</td><td>Self-emp-not-inc</td><td>>50K</td><td>46.428</td></tr><tr><td>Row13</td><td>State-gov</td><td><=50K</td><td>37.279</td></tr><tr><td>Row14</td><td>State-gov</td><td>>50K</td><td>45.21</td></tr><tr><td>Row15</td><td>Without-pay</td><td><=50K</td><td>47.786</td></tr></table>	Row ID	S workclass	S income	D Mean(age)	Row0	Federal-gov	<=50K	40.625	Row1	Federal-gov	>50K	45.712	Row2	Local-gov	<=50K	40.705	Row3	Local-gov	>50K	44.254	Row4	Never-worked	<=50K	20.571	Row5	Not specified	<=50K	39.258	Row6	Not specified	>50K	55.618	Row7	Private	<=50K	35.113	Row8	Private	>50K	42.815	Row9	Self-emp-inc	<=50K	43.206	Row10	Self-emp-inc	>50K	48.249	Row11	Self-emp-not-inc	<=50K	44.389	Row12	Self-emp-not-inc	>50K	46.428	Row13	State-gov	<=50K	37.279	Row14	State-gov	>50K	45.21	Row15	Without-pay	<=50K	47.786
Row ID	S workclass	S income	D Mean(age)																																																																		
Row0	Federal-gov	<=50K	40.625																																																																		
Row1	Federal-gov	>50K	45.712																																																																		
Row2	Local-gov	<=50K	40.705																																																																		
Row3	Local-gov	>50K	44.254																																																																		
Row4	Never-worked	<=50K	20.571																																																																		
Row5	Not specified	<=50K	39.258																																																																		
Row6	Not specified	>50K	55.618																																																																		
Row7	Private	<=50K	35.113																																																																		
Row8	Private	>50K	42.815																																																																		
Row9	Self-emp-inc	<=50K	43.206																																																																		
Row10	Self-emp-inc	>50K	48.249																																																																		
Row11	Self-emp-not-inc	<=50K	44.389																																																																		
Row12	Self-emp-not-inc	>50K	46.428																																																																		
Row13	State-gov	<=50K	37.279																																																																		
Row14	State-gov	>50K	45.21																																																																		
Row15	Without-pay	<=50K	47.786																																																																		
Explicación Resultados	<ul style="list-style-type: none">• Todas las personas adultas que <i>nunca han trabajado</i> tienen ingresos MENORES a los 50K, y son por lo general jóvenes que rondan la edad de 20 años.• Todas las personas adultas que <i>trabajan sin recibir un salario específico</i> 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.																																																																				

b) Horas de Trabajo Semanales promedio de cada clase trabajadora.

Nodo	GroupBy
Configuración	

Resultados

Row ID	S workclass	D Mean(hours-per-week)
Row0	Federal-gov	41.379
Row1	Local-gov	40.983
Row2	Never-worked	28.429
Row3	Not specified	31.919
Row4	Private	40.267
Row5	Self-emp-inc	48.818
Row6	Self-emp-not-inc	44.422
Row7	State-gov	39.032
Row8	Without-pay	32.714

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-emp-inc* y *Self-emp-not-inc*) son los que más horas trabajan a la semana, trabajando 48 y 44 horas semanales en promedio respectivamente.

c) El porcentaje de personas adultas de cierta raza presentes en cada clase trabajadora.

Nodo	CrossTable																																																																													
Configuración	<div><div>Dialog - 3:19 - Crosstab (local) (c. Porcentaje de adultos)</div><div>File</div><div><div>Settings</div><div>Flow Variables</div><div>Job Manager Selection</div><div>Memory Policy</div></div><div><div>Row variable:</div><div>S</div><div>workclass</div></div><div><div>Column variable:</div><div>S</div><div>race</div></div><div><div>Weight column:</div><div>?</div><div><none></div></div><div><input type="checkbox"/> Enable hilling</div><div><div>OK</div><div>Apply</div><div>Cancel</div><div>?</div></div></div>																																																																													
Resultados	<table><tr><th>Frequency Row Percent</th><th>Amer-Indian-Eskimo</th><th>Asian-Pac-Islander</th><th>Black</th><th>Other</th><th>White</th><th>Total</th></tr><tr><td>Federal-gov</td><td>19 1.9792%</td><td>44 4.5833%</td><td>169 17.6042%</td><td>7 0.7292%</td><td>721 75.1042%</td><td>960</td></tr><tr><td>Local-gov</td><td>36 1.72%</td><td>39 1.8634%</td><td>288 13.7602%</td><td>10 0.4778%</td><td>1,720 82.1787%</td><td>2,093</td></tr><tr><td>Never-worked</td><td></td><td></td><td>2 28.5714%</td><td></td><td>5 71.4286%</td><td>7</td></tr><tr><td>Not specified</td><td>25 1.3617%</td><td>65 3.3403%</td><td>213 11.6013%</td><td>23 1.2527%</td><td>1,510 82.2444%</td><td>1,836</td></tr><tr><td>Private</td><td>190 0.8372%</td><td>713 3.1415%</td><td>2,176 9.5876%</td><td>213 0.9385%</td><td>19,404 85.4952%</td><td>22,696</td></tr><tr><td>Self-emp-inc</td><td>2 0.1792%</td><td>46 4.1219%</td><td>23 2.0609%</td><td>5 0.448%</td><td>1,040 93.19%</td><td>1,116</td></tr><tr><td>Self-emp-not-inc</td><td>24 0.9445%</td><td>73 2.8729%</td><td>93 3.66%</td><td>9 0.3542%</td><td>2,342 92.1684%</td><td>2,541</td></tr><tr><td>State-gov</td><td>15 1.1556%</td><td>58 4.4684%</td><td>159 12.2496%</td><td>4 0.3082%</td><td>1,062 81.8182%</td><td>1,298</td></tr><tr><td>Without-pay</td><td></td><td>1 7.1429%</td><td>1 7.1429%</td><td></td><td>12 85.7143%</td><td>14</td></tr><tr><td>Total</td><td>311</td><td>1,039</td><td>3,124</td><td>271</td><td>27,816</td><td>32,561</td></tr></table>	Frequency Row Percent	Amer-Indian-Eskimo	Asian-Pac-Islander	Black	Other	White	Total	Federal-gov	19 1.9792%	44 4.5833%	169 17.6042%	7 0.7292%	721 75.1042%	960	Local-gov	36 1.72%	39 1.8634%	288 13.7602%	10 0.4778%	1,720 82.1787%	2,093	Never-worked			2 28.5714%		5 71.4286%	7	Not specified	25 1.3617%	65 3.3403%	213 11.6013%	23 1.2527%	1,510 82.2444%	1,836	Private	190 0.8372%	713 3.1415%	2,176 9.5876%	213 0.9385%	19,404 85.4952%	22,696	Self-emp-inc	2 0.1792%	46 4.1219%	23 2.0609%	5 0.448%	1,040 93.19%	1,116	Self-emp-not-inc	24 0.9445%	73 2.8729%	93 3.66%	9 0.3542%	2,342 92.1684%	2,541	State-gov	15 1.1556%	58 4.4684%	159 12.2496%	4 0.3082%	1,062 81.8182%	1,298	Without-pay		1 7.1429%	1 7.1429%		12 85.7143%	14	Total	311	1,039	3,124	271	27,816	32,561
Frequency Row Percent	Amer-Indian-Eskimo	Asian-Pac-Islander	Black	Other	White	Total																																																																								
Federal-gov	19 1.9792%	44 4.5833%	169 17.6042%	7 0.7292%	721 75.1042%	960																																																																								
Local-gov	36 1.72%	39 1.8634%	288 13.7602%	10 0.4778%	1,720 82.1787%	2,093																																																																								
Never-worked			2 28.5714%		5 71.4286%	7																																																																								
Not specified	25 1.3617%	65 3.3403%	213 11.6013%	23 1.2527%	1,510 82.2444%	1,836																																																																								
Private	190 0.8372%	713 3.1415%	2,176 9.5876%	213 0.9385%	19,404 85.4952%	22,696																																																																								
Self-emp-inc	2 0.1792%	46 4.1219%	23 2.0609%	5 0.448%	1,040 93.19%	1,116																																																																								
Self-emp-not-inc	24 0.9445%	73 2.8729%	93 3.66%	9 0.3542%	2,342 92.1684%	2,541																																																																								
State-gov	15 1.1556%	58 4.4684%	159 12.2496%	4 0.3082%	1,062 81.8182%	1,298																																																																								
Without-pay		1 7.1429%	1 7.1429%		12 85.7143%	14																																																																								
Total	311	1,039	3,124	271	27,816	32,561																																																																								
Explicación Resultados	<ul style="list-style-type: none">En todas las clases trabajadoras predominan las personas de <i>raza blanca</i> (siendo un entre un 70% u 80% del total de trabajadores en la mayoría de los casos), seguido de las personas de <i>raza negra</i>.																																																																													

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

Nodo	CrossTable																																																																		
Configuración	<div><div>Dialog - 3:21 - Crosstab (local) (d. Porcentaje de adultos)</div><div>File</div><div><div>Settings</div><div>Flow Variables</div><div>Job Manager Selection</div><div>Memory Policy</div></div><div><div>Row variable:</div><div>S</div><div>workclass</div></div><div><div>Column variable:</div><div>S</div><div>education</div></div><div><div>Weight column:</div><div>?</div><div><none></div></div><div><input type="checkbox"/> Enable highting</div><div><div>OK</div><div>Apply</div><div>Cancel</div><div></div></div></div>																																																																		
Resultados	<table><tr><th>Frequency Row Percent</th><th>Elementary School</th><th>High School</th><th>Middle School</th><th>Tertiary Education</th><th>Total</th></tr><tr><td>Federal-gov</td><td>1 0.1042%</td><td>286 29.7917%</td><td>2 0.2083%</td><td>671 69.8958%</td><td>960</td></tr><tr><td>Local-gov</td><td>17 0.8122%</td><td>612 29.2403%</td><td>28 1.3378%</td><td>1,436 68.6097%</td><td>2,093</td></tr><tr><td>Never-worked</td><td></td><td>4 57.1429%</td><td>1 14.2857%</td><td>2 28.5714%</td><td>7</td></tr><tr><td>Not specified</td><td>47 2.5599%</td><td>841 45.8061%</td><td>72 3.9216%</td><td>876 47.7124%</td><td>1,836</td></tr><tr><td>Private</td><td>443 1.9519%</td><td>10,118 44.5805%</td><td>424 1.8682%</td><td>11,711 51.5994%</td><td>22,696</td></tr><tr><td>Self-emp-inc</td><td>6 0.5376%</td><td>329 29.4803%</td><td>14 1.2545%</td><td>767 68.7276%</td><td>1,116</td></tr><tr><td>Self-emp-not-inc</td><td>32 1.2593%</td><td>1,046 41.1649%</td><td>94 3.6993%</td><td>1,369 53.8764%</td><td>2,541</td></tr><tr><td>State-gov</td><td>6 0.4622%</td><td>311 23.9599%</td><td>10 0.7704%</td><td>971 74.8074%</td><td>1,298</td></tr><tr><td>Without-pay</td><td></td><td>9 64.2857%</td><td>1 7.1429%</td><td>4 28.5714%</td><td>14</td></tr><tr><td>Total</td><td>552</td><td>13,556</td><td>646</td><td>17,807</td><td>32,561</td></tr></table>	Frequency Row Percent	Elementary School	High School	Middle School	Tertiary Education	Total	Federal-gov	1 0.1042%	286 29.7917%	2 0.2083%	671 69.8958%	960	Local-gov	17 0.8122%	612 29.2403%	28 1.3378%	1,436 68.6097%	2,093	Never-worked		4 57.1429%	1 14.2857%	2 28.5714%	7	Not specified	47 2.5599%	841 45.8061%	72 3.9216%	876 47.7124%	1,836	Private	443 1.9519%	10,118 44.5805%	424 1.8682%	11,711 51.5994%	22,696	Self-emp-inc	6 0.5376%	329 29.4803%	14 1.2545%	767 68.7276%	1,116	Self-emp-not-inc	32 1.2593%	1,046 41.1649%	94 3.6993%	1,369 53.8764%	2,541	State-gov	6 0.4622%	311 23.9599%	10 0.7704%	971 74.8074%	1,298	Without-pay		9 64.2857%	1 7.1429%	4 28.5714%	14	Total	552	13,556	646	17,807	32,561
Frequency Row Percent	Elementary School	High School	Middle School	Tertiary Education	Total																																																														
Federal-gov	1 0.1042%	286 29.7917%	2 0.2083%	671 69.8958%	960																																																														
Local-gov	17 0.8122%	612 29.2403%	28 1.3378%	1,436 68.6097%	2,093																																																														
Never-worked		4 57.1429%	1 14.2857%	2 28.5714%	7																																																														
Not specified	47 2.5599%	841 45.8061%	72 3.9216%	876 47.7124%	1,836																																																														
Private	443 1.9519%	10,118 44.5805%	424 1.8682%	11,711 51.5994%	22,696																																																														
Self-emp-inc	6 0.5376%	329 29.4803%	14 1.2545%	767 68.7276%	1,116																																																														
Self-emp-not-inc	32 1.2593%	1,046 41.1649%	94 3.6993%	1,369 53.8764%	2,541																																																														
State-gov	6 0.4622%	311 23.9599%	10 0.7704%	971 74.8074%	1,298																																																														
Without-pay		9 64.2857%	1 7.1429%	4 28.5714%	14																																																														
Total	552	13,556	646	17,807	32,561																																																														
Explicación Resultados	<ul style="list-style-type: none">En todas las clases trabajadoras predominan las personas que han concluido estudios de nivel <i>Superior</i> (Tertiary Education), seguidos por las personas que han concluido estudios de nivel <i>Medio Superior</i> (High School).En todas las clases trabajadoras, la presencia de las personas que solo cuentan con educación <i>primaria</i> (Elementary School) es prácticamente inexistente.																																																																		

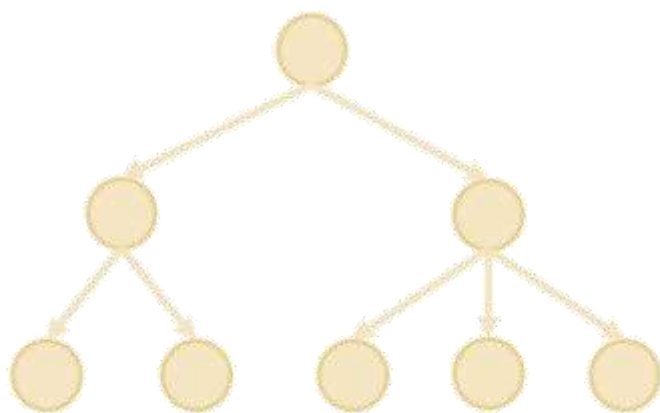
- e) El porcentaje de personas adultas de cierto sexo presentes en cada clase trabajadora.

Nodo	CrossTable																																												
Configuración	<div><div>Dialog - 3:22 - Crosstab (local) (e. Porcentaje de adultos)</div><div><div>File</div><div>SettingsFlow VariablesJob Manager SelectionMemory Policy</div><div><div>Row variable:workclass</div><div>Column variable:sex</div><div>Weight column: <none></div><div><input type="checkbox"/> Enable hilling</div></div><div><div>OK</div><div>Apply</div><div>Cancel</div><div>?</div></div></div></div>																																												
Resultados	<table><thead><tr><th>Frequency Row Percent</th><th>Female</th><th>Male</th><th>Total</th></tr></thead><tbody><tr><td>Federal-gov</td><td>315 32.8125%</td><td>645 67.1875%</td><td>960</td></tr><tr><td>Local-gov</td><td>835 39.8949%</td><td>1,258 60.1051%</td><td>2,093</td></tr><tr><td>Never-worked</td><td>2 28.5714%</td><td>5 71.4286%</td><td>7</td></tr><tr><td>Not specified</td><td>839 45.6972%</td><td>997 54.3028%</td><td>1,836</td></tr><tr><td>Private</td><td>7,752 34.1558%</td><td>14,944 65.8442%</td><td>22,696</td></tr><tr><td>Self-emp-inc</td><td>135 12.0968%</td><td>981 87.9032%</td><td>1,116</td></tr><tr><td>Self-emp-not-inc</td><td>399 15.7025%</td><td>2,142 84.2975%</td><td>2,541</td></tr><tr><td>State-gov</td><td>489 37.6733%</td><td>809 62.3267%</td><td>1,298</td></tr><tr><td>Without-pay</td><td>5 35.7143%</td><td>9 64.2857%</td><td>14</td></tr><tr><td>Total</td><td>10,771</td><td>21,790</td><td>32,561</td></tr></tbody></table>	Frequency Row Percent	Female	Male	Total	Federal-gov	315 32.8125%	645 67.1875%	960	Local-gov	835 39.8949%	1,258 60.1051%	2,093	Never-worked	2 28.5714%	5 71.4286%	7	Not specified	839 45.6972%	997 54.3028%	1,836	Private	7,752 34.1558%	14,944 65.8442%	22,696	Self-emp-inc	135 12.0968%	981 87.9032%	1,116	Self-emp-not-inc	399 15.7025%	2,142 84.2975%	2,541	State-gov	489 37.6733%	809 62.3267%	1,298	Without-pay	5 35.7143%	9 64.2857%	14	Total	10,771	21,790	32,561
Frequency Row Percent	Female	Male	Total																																										
Federal-gov	315 32.8125%	645 67.1875%	960																																										
Local-gov	835 39.8949%	1,258 60.1051%	2,093																																										
Never-worked	2 28.5714%	5 71.4286%	7																																										
Not specified	839 45.6972%	997 54.3028%	1,836																																										
Private	7,752 34.1558%	14,944 65.8442%	22,696																																										
Self-emp-inc	135 12.0968%	981 87.9032%	1,116																																										
Self-emp-not-inc	399 15.7025%	2,142 84.2975%	2,541																																										
State-gov	489 37.6733%	809 62.3267%	1,298																																										
Without-pay	5 35.7143%	9 64.2857%	14																																										
Total	10,771	21,790	32,561																																										
Explicación Resultados	<ul style="list-style-type: none">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:
 - o Cree una portada para identificar cada tipo de árbol
 - o Genere el diccionario de datos con los datos considerados en cada caso
 - o Añada el diagrama general del árbol
 - o Añada las pantallas de configuración correspondientes
 - o Describa los resultados obtenidos de la matriz de confusión de los árboles ID3 y C4.5. Se anexan medidas a describir.
 - o Presente las reglas generadas en cada caso.



Árbol ID3

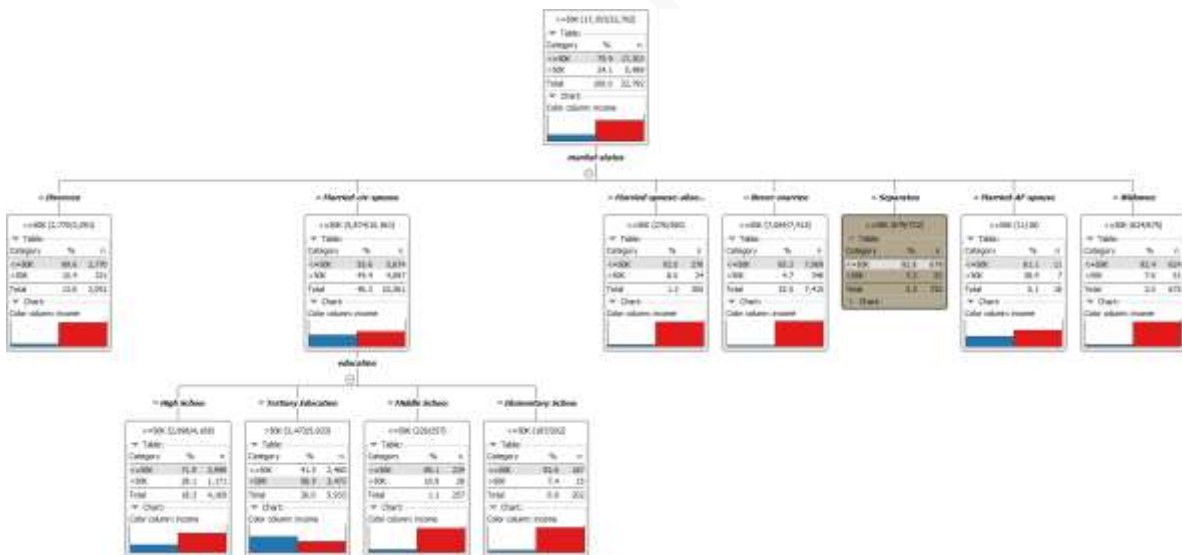
FUNCIÓN:

*Predice si un individuo tiene un INGRESO **menor** o igual a 50k ($\leq 50k$), o mayor a 50k ($> 50k$).*

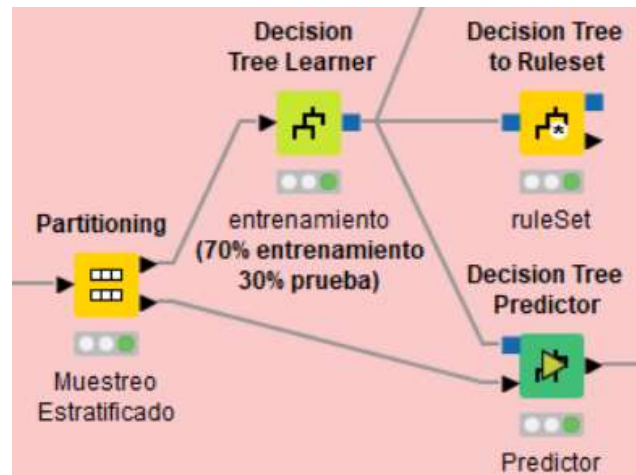
1. Diccionario de datos.

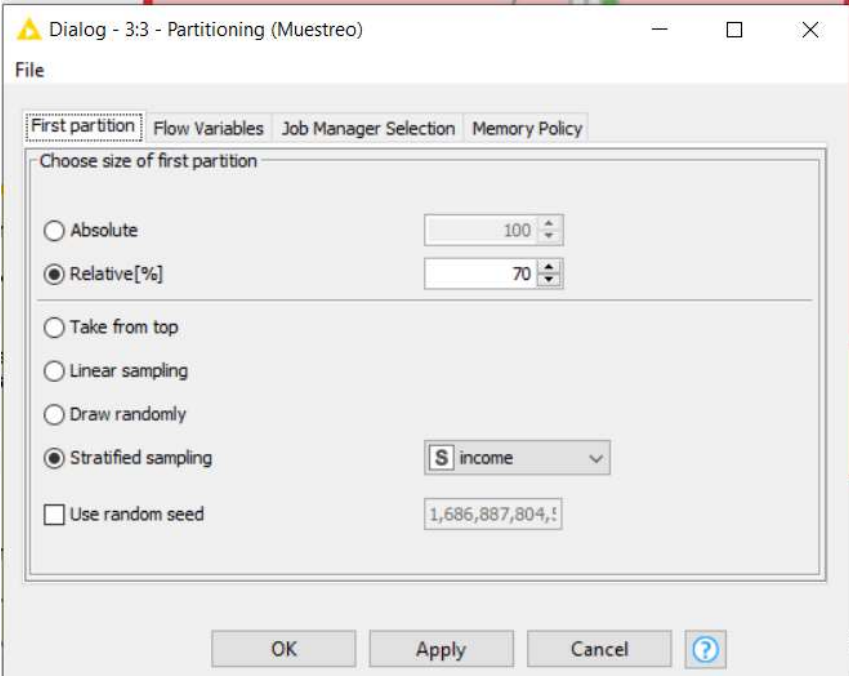
No	Nombre	Tipo	Dominio
1	workclass	Categorico	Clase de Trabajo
2	education	Categorico	Bachelors, Some-college, 11th, HS-grad, Prof-school, Assoc-acdm, Assoc-voc, 9th, 7th-8th, 12th, Masters, 1st-4th, 10th, Doctorate, 5th-6th, Preschool
3	marital-status	Categorico	Married-civ-spouse, Divorced, Never-married,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

2. Diagrama general del árbol.



3. Configuración.



Nodo	Configuración
Partitioning	<p>Se configuró que se usara el 70% del conjunto de datos como conjunto de entrenamiento, y el 30% restante como conjunto de prueba.</p> <p>Se utilizó el <i>muestreo estratificado</i> para generar el árbol.</p> 

Decision Tree Learner

Dialog - 3:4 - Decision Tree Learner (entrenamiento)

File

Options PMMLSettings Flow Variables Job Manager Selection

General

Class column: S income

Quality measure: Gain ratio

Pruning method: No pruning

☒ Reduced Error Pruning

Min number records per node: 1,600

Number records to store for view: 10,000

☒ Average split point

Number threads: 8

☐ Skip nominal columns without domain information

Root split

☐ Force root split column

Root split column: S native-country

Binary nominal splits

☐ Binary nominal splits

Max #nominal: 10

☐ Filter invalid attribute values in child nodes

OK Apply Cancel ?

Decision Tree Predictor

Dialog - 3:6 - Decision Tree Predictor (Predictor)

File

Options Flow Variables Job Manager Selection Memory Policy

Maximum number of stored patterns for HLLite-ing: 10,000

☐ Change prediction column name

Prediction (income)

☐ Append columns with normalized class distribution

Suffix for probability columns:

OK Apply Cancel ?

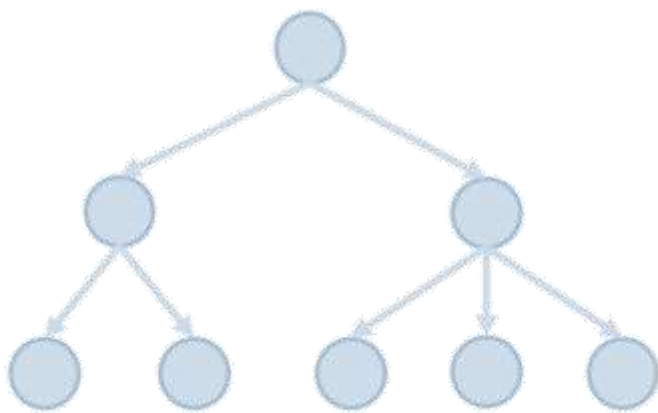
4. Resultados Obtenidos.

Row ID	I <=50K	I >50K
<=50K	6420	997
>50K	846	1506

	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	$\frac{a+d}{a+b+c+d} = \frac{1506+6420}{1506+846+997+6420} = \frac{7926}{9769} = 0.81$	El 81% de los registros totales, fueron clasificados de forma CORRECTA. Accuracy: 81.134%
Tasa de error	$\frac{b+c}{a+b+c+d} = \frac{846+997}{1506+846+997+6420} = 0.18$	El 18% de los registros totales, fueron clasificados de forma EQUIVOCADA. Error: 18.866%
Precisión	$\frac{d}{b+d} = \frac{6420}{846+6420} = \frac{6420}{7266} = 0.60$	El 60% de los ejemplos clasificados como clase positiva , son realmente positivos .
Sensibilidad (<i>Recall</i>)	$\frac{d}{c+d} = \frac{6420}{997+6420} = \frac{6420}{7417} = 0.64$	El clasificador puede reconocer muestras positivas en el 64% de los casos.
Tasa de positivos falsos	$\frac{b}{a+b} = \frac{846}{1506+846} = \frac{846}{2352} = 0.13$	La tasa de positivos falsos es del 13%.
Tasa de negativos falsos	$\frac{c}{c+d} = \frac{997}{997+6420} = \frac{997}{7417} = 0.35$	La tasa de negativos falsos es del 35%.
Especificidad	$\frac{a}{a+b} = \frac{1506}{1506+846} = \frac{1506}{2352} = 0.86$	El clasificador puede reconocer muestras negativas en el 86% de los casos.

5. Reglas Obtenidas.

S Rule	D Record count	D Number of correct
\$marital-status\$ = "Divorced" AND TRUE => "<=50K"	3,091	2,770
\$education\$ = "High School" AND \$marital-status\$ = "Married-civ-spouse" => "<=50K"	4,169	2,998
\$education\$ = "Tertiary Education" AND \$marital-status\$ = "Married-civ-spouse" => ">50K"	5,933	3,473
\$education\$ = "Middle School" AND \$marital-status\$ = "Married-civ-spouse" => "<=50K"	257	229
\$education\$ = "Elementary School" AND \$marital-status\$ = "Married-civ-spouse" => "<=50K"	202	187
\$marital-status\$ = "Married-spouse-absent" AND TRUE => "<=50K"	300	276
\$marital-status\$ = "Never-married" AND TRUE => "<=50K"	7,415	7,069
\$marital-status\$ = "Separated" AND TRUE => "<=50K"	732	679
\$marital-status\$ = "Married-AF-spouse" AND TRUE => "<=50K"	18	11
\$marital-status\$ = "Widowed" AND TRUE => "<=50K"	675	624



Árbol C4.5

ABSTRACT

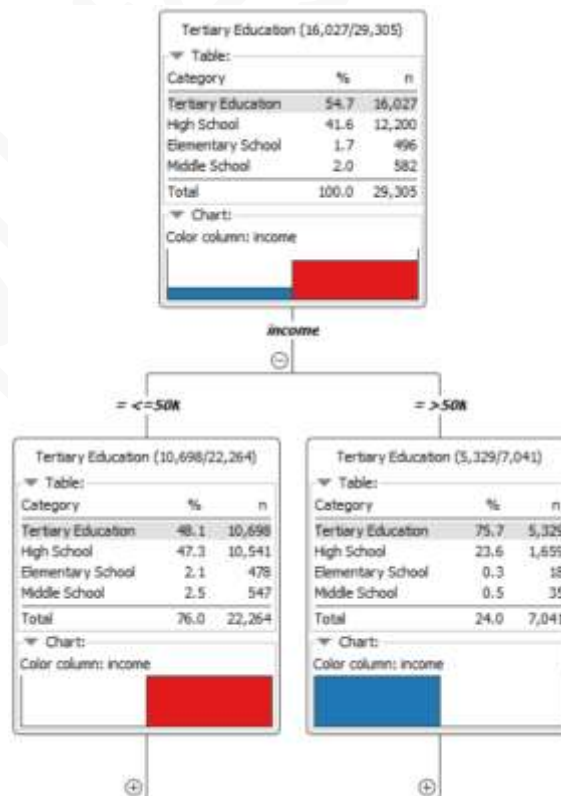
Predice el NIVEL DE EDUCACIÓN de un individuo.

1. Diccionario de datos.

No	Nombre	Tipo	Dominio
0	age	Numérico	Números enteros Positivos
1	education	Categorico	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	Categorico	Ocupación del individuo
3	race	Categorico	White, Asian-Pac-Islander, Amer-Indian-Eskimo, Other, Black
4	sex	Categorico	Female, Male
5	native-country	Categorico	Nombre del país nativo del individuo
6	income	categorico	Nivel de ingresos del individuo

2. 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**.



3. 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()

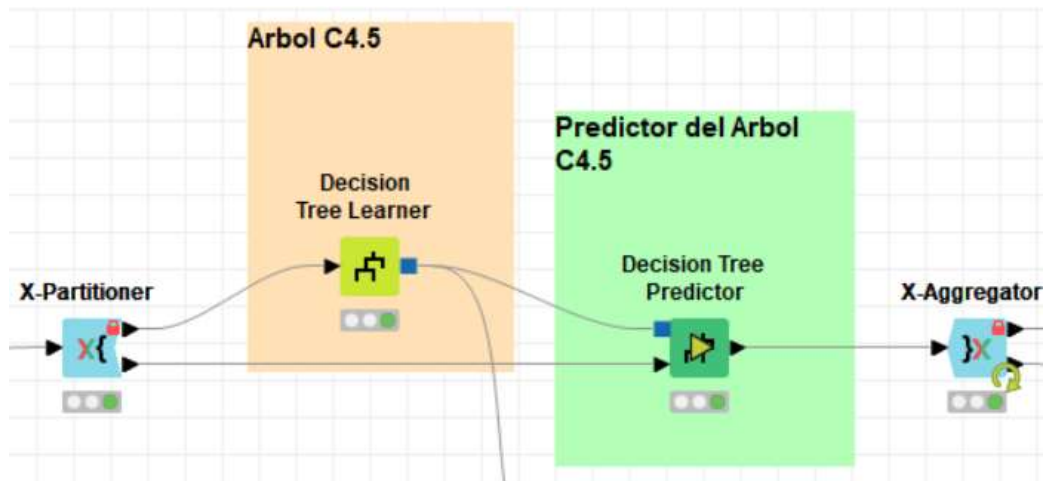
# ----- TRATAMIENTO DE DATOS -----

# 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)
```



Nodo	Configuracion
X-Partitioner	<p>Se utilizó el <i>muestreo estratificado</i> para generar el árbol, con 10 validaciones.</p>

Decision Tree Learner

Dialog - 3:8:16 - Decision Tree Learner

File

Options PMMLSettings Flow Variables Job Manager Selection

General

Class column **S** education

Quality measure Gain ratio

Pruning method MDL

☒ Reduced Error Pruning

Min number records per node 1,000

Number records to store for view 10,000

☒ Average split point

Number threads 4

☐ Skip nominal columns without domain information

Root split

☐ Force root split column

Root split column **S** native-country

Binary nominal splits

☐ Binary nominal splits

Max #nominal 10

☐ Filter invalid attribute values in child nodes

OK Apply Cancel ?

Decision Tree Predictor

Dialog - 3:8:15 - Decision Tree Predictor

File

Options Flow Variables Job Manager Selection Memory Policy

Maximum number of stored patterns for HLite-ing: 10,000

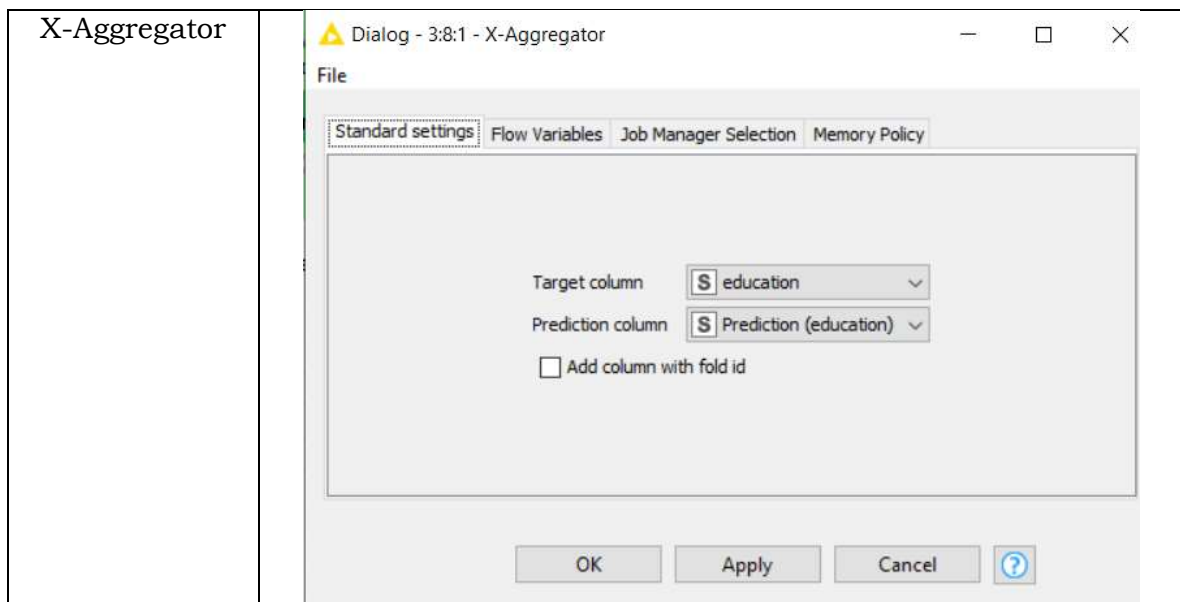
☐ Change prediction column name

Prediction (education)

☐ Append columns with normalized class distribution

Suffix for probability columns

OK Apply Cancel ?



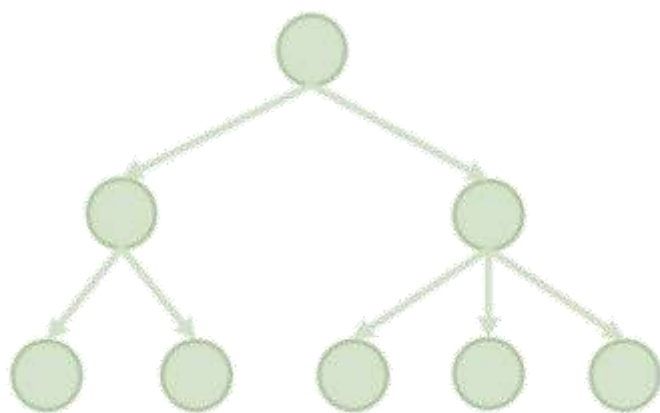
4. Resultados Obtenidos.

Row ID	Tertiar...	High Sc...	Middle ...	Elemen...
Tertiary Educ...	13079	4728	0	0
High School	4886	8670	0	0
Middle School	95	551	0	0
Elementary S...	48	504	0	0

Medida	Cálculo	Explicación
Tasa de exactitud	$\frac{21749}{32561} = 0.6679$	El 66.79% de los registros totales, fueron clasificados de forma CORRECTA. Accuracy: 66.795%
Tasa de error	$\frac{10812}{32561} = 0.332$	El 33.2% de los registros totales, fueron clasificados de forma EQUIVOCADA. Error: 33.205%

5. Reglas Obtenidas.

S Rule	D Record count	D Number of correct
\$occupation\$ = "Adm-clerical" AND \$income\$ = "<=50K" => "Tertiary Education"	2,975	1,729
\$occupation\$ = "Exec-managerial" AND \$income\$ = "<=50K" => "Tertiary Edu..."	1,893	1,337
\$occupation\$ = "Handlers-cleaners" AND \$income\$ = "<=50K" => "High School"	1,150	763
\$occupation\$ = "Prof-specialty" AND \$income\$ = "<=50K" => "Tertiary Educat..."	2,039	1,839
\$occupation\$ = "Other-service" AND \$income\$ = "<=50K" => "High School"	2,829	1,651
\$occupation\$ = "Sales" AND \$income\$ = "<=50K" => "Tertiary Education"	2,376	1,307
\$occupation\$ = "Craft-repair" AND \$income\$ = "<=50K" => "High School"	2,851	1,769
\$occupation\$ = "Farming-fishing" AND \$income\$ = "<=50K" => "High School"	779	426
\$occupation\$ = "Machine-op-inspct" AND \$income\$ = "<=50K" => "High School"	1,595	1,078
\$occupation\$ = "Transport-moving" AND \$income\$ = "<=50K" => "High School"	1,145	782
\$occupation\$ = "Tech-support" AND \$income\$ = "<=50K" => "Tertiary Educati..."	593	469
\$occupation\$ = "Not specified" AND \$income\$ = "<=50K" => "High School"	1,497	716
\$occupation\$ = "Protective-serv" AND \$income\$ = "<=50K" => "Tertiary Educat..."	405	218
\$occupation\$ = "Armed-Forces" AND \$income\$ = "<=50K" => "High School"	8	5
\$occupation\$ = "Priv-house-serv" AND \$income\$ = "<=50K" => "High School"	129	70
\$occupation\$ = "Adm-clerical" AND \$income\$ = ">50K" => "Tertiary Education"	466	307
\$occupation\$ = "Exec-managerial" AND \$income\$ = ">50K" => "Tertiary Educa..."	1,773	1,523
\$occupation\$ = "Handlers-cleaners" AND \$income\$ = ">50K" => "High School"	74	41
\$occupation\$ = "Prof-specialty" AND \$income\$ = ">50K" => "Tertiary Education"	1,664	1,601
\$occupation\$ = "Other-service" AND \$income\$ = ">50K" => "Tertiary Education"	120	68
\$occupation\$ = "Sales" AND \$income\$ = ">50K" => "Tertiary Education"	892	689
\$occupation\$ = "Craft-repair" AND \$income\$ = ">50K" => "Tertiary Education"	836	412
\$occupation\$ = "Farming-fishing" AND \$income\$ = ">50K" => "Tertiary Educati..."	105	51
\$occupation\$ = "Machine-op-inspct" AND \$income\$ = ">50K" => "High School"	225	139
\$occupation\$ = "Transport-moving" AND \$income\$ = ">50K" => "High School"	277	175
\$occupation\$ = "Tech-support" AND \$income\$ = ">50K" => "Tertiary Education"	253	209
\$occupation\$ = "Not specified" AND \$income\$ = ">50K" => "Tertiary Education"	162	117
\$occupation\$ = "Protective-serv" AND \$income\$ = ">50K" => "Tertiary Educat..."	192	147
\$occupation\$ = "Armed-Forces" AND \$income\$ = ">50K" => "Tertiary Education"	1	1
\$occupation\$ = "Priv-house-serv" AND \$income\$ = ">50K" => "Tertiary Educa..."	1	1



Árbol CART

FUNCIÓN:

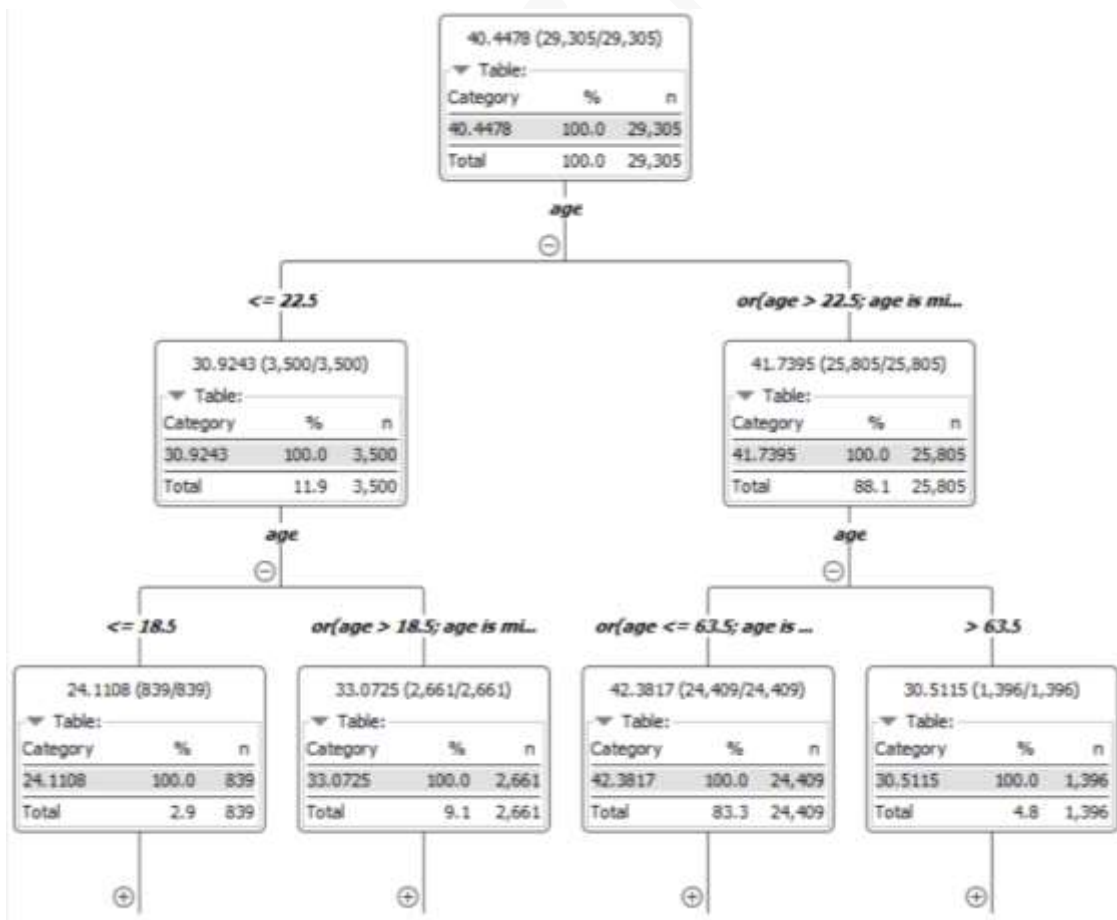
Predice las HORAS DE TRABAJO SEMANALES de un individuo.

1. Diccionario de datos.

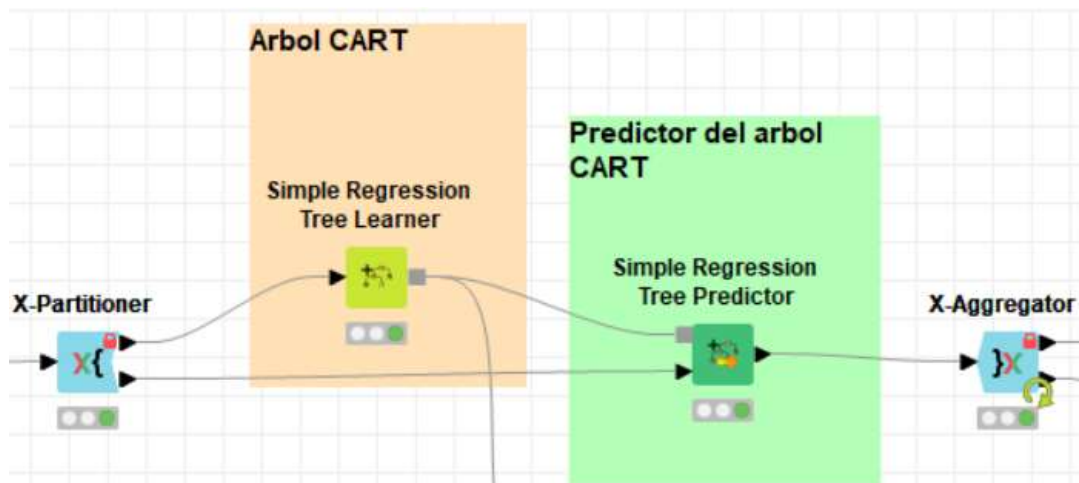
No	Nombre	Tipo	Dominio
0	age	Numérico	Números enteros Positivos
1	workclass	Categorico	Clase de Trabajo
2	education-num	Numérico	Números enteros Positivos
3	race	Categorico	White, Asian-Pac-Islander, Amer-Indian-Eskimo, Other, Black
4	sex	Categorico	Female, Male
5	hours-per-week	Numérico	Número de horas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

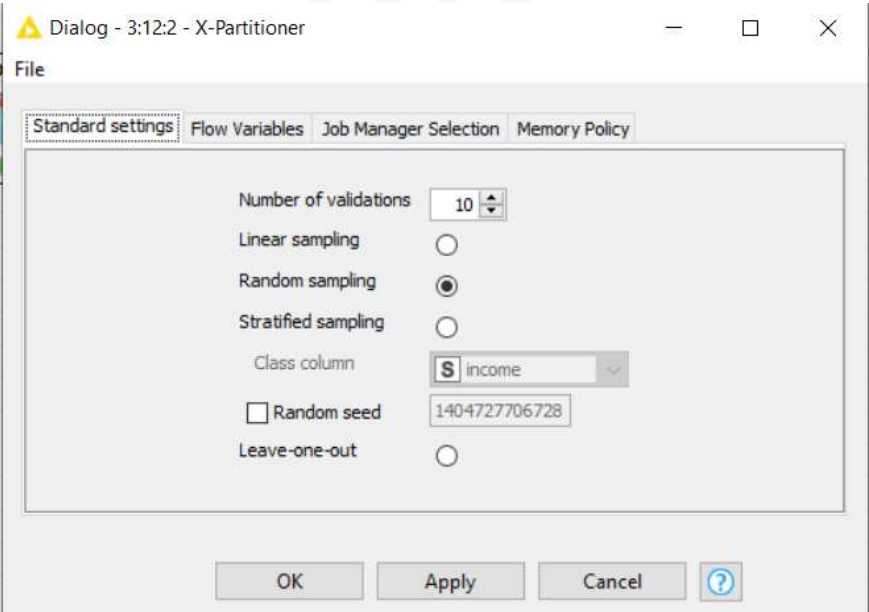
2. Diagrama general del árbol.

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**.



3. Configuración.



Nodo	Configuración
X-Partitioner	<p>Se utilizó el <i>muestreo aleatorio</i> para generar el árbol, con 10 validaciones.</p> 

Decision Tree Learner

Dialog - 3:12:17 - Simple Regression Tree Learner

File

Options | Flow Variables | Job Manager Selection

Target Column: hours-per-week

Attribute Selection

☐ Use fingerprint attribute: [no valid fingerprint input]

☒ Use column attributes

☒ Manual Selection ☐ Wildcard/Regex Selection

Exclude

Filter

- I fringht
- S education
- S marital-status
- S occupation
- S relationship
- I capital-gain
- I capital-loss

☐ Enforce exclusion

Include

Filter

- I age
- S workclass
- I education-num
- S race
- S sex
- S native-country
- S income

☒ Enforce inclusion

Mac Options

☐ Ignore columns without domain information

☐ Enable Highlighting (if patterns to store): 2,000

Tree Options

☒ Use binary splits for nominal attributes

Missing value handling: ignore

☐ Limit number of levels (tree depth): 10

☐ Minimum split node size: 1

OK Apply Cancel ?

Decision Tree Predictor

Dialog - 3:12:18 - Simple Regression Tree Predictor

File

Prediction Settings | Flow Variables | Job Manager Selection | Memory Policy

☐ Change prediction column name

Prediction column name: Prediction (hours-per-week)

OK Apply Cancel ?

X-Aggregator

Dialog - 3:12:1 - X-Aggregator

File

Standard settings | Flow Variables | Job Manager Selection | Memory Policy

Target column: I hours-per-week

Prediction column: D Prediction (hours-per-week)

☐ Add column with fold id

OK Apply Cancel ?

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:

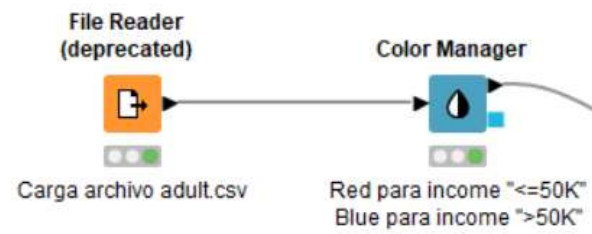
- Genere un diccionario de datos de todos los atributos del conjunto de datos

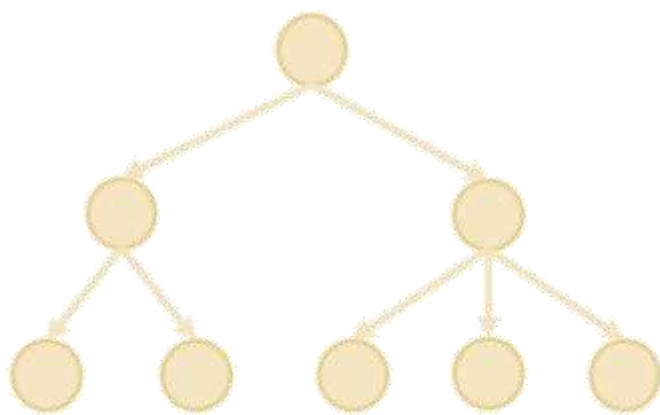
No	Nombre	Tipo	Dominio
0	Invoice ID	Numérico	Número de identificación
1	Branch	Categorico	Letra desde A-C
2	City	Categorico	Ubicación del Supermercado
3	Customer type	Categorico	Member o Normal
4	Gender	Categorico	Female, Male
5	Product line	Categorico	Categoría de Producto
6	Unit Price	Numérico	Precio 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:
 - o Cree una portada para identificar cada tipo de árbol
 - o Genere el diccionario de datos con los datos considerados en cada caso
 - o Añada el diagrama general del árbol
 - o 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

FUNCIÓN:

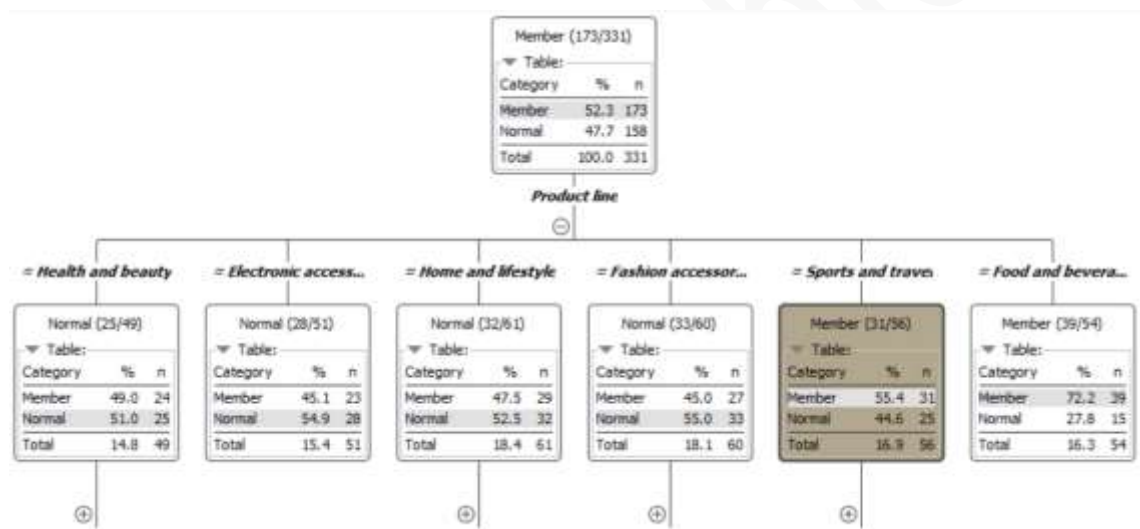
Predice el TIPO DE CLIENTE que realizó una compra, de acuerdo al producto que compró, su género y el tipo de pago que uso.

1. Diccionario de datos.

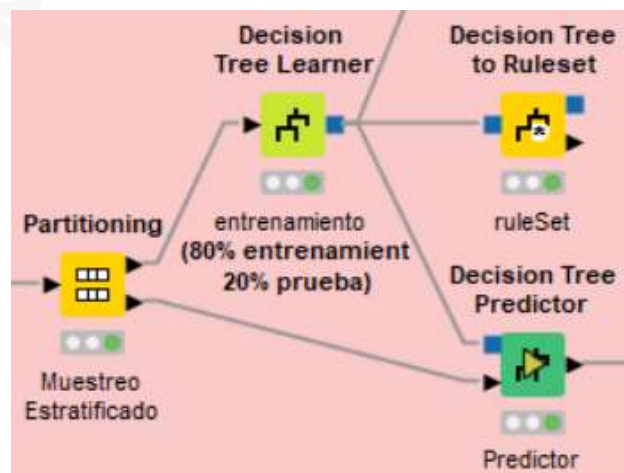
No	Nombre	Tipo	Dominio
0	Customer type	Catagórico	Member o Normal
1	Gender	Categorico	Female, Male
2	Product line	Catagórico	Categoria de Producto
3	Payment	Numérico	Forma de pago Ewallet, Credit card, Cash

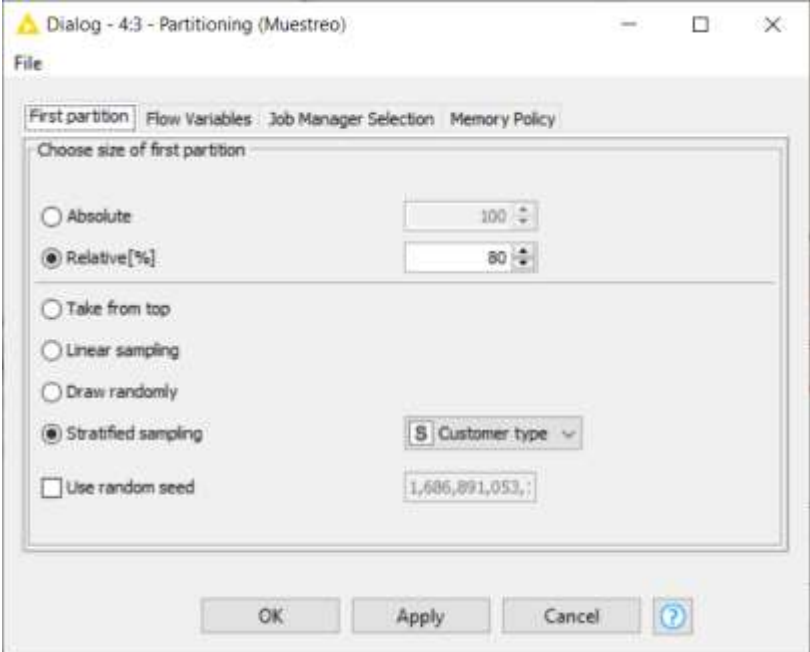
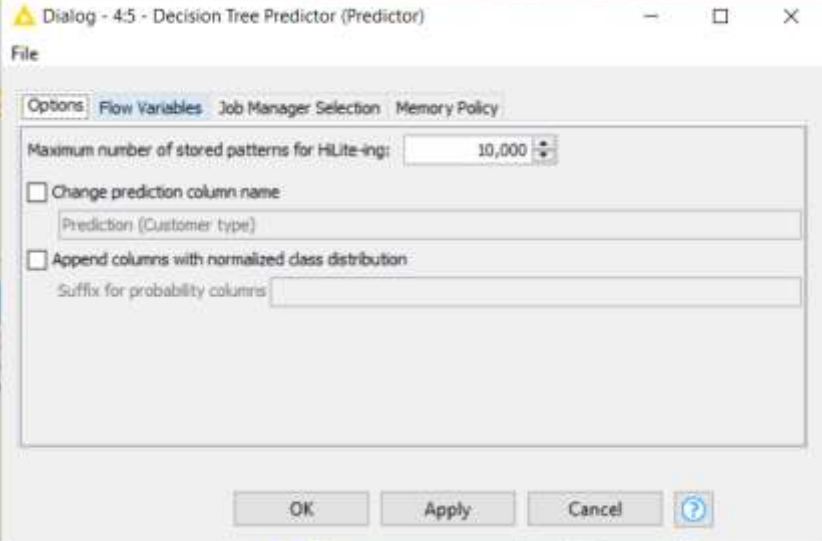
2. Diagrama general del árbol.

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**.



3. Configuración.



Nodo	Configuración
Partitioning	<p>Se configuró que se usara el 80% del conjunto de datos como conjunto de entrenamiento, y el 20% restante como conjunto de prueba.</p> <p>Se utilizó el <i>muestreo estratificado</i> para generar el árbol.</p> 
Decision Tree Predictor	

Decision Tree Learner

Dialog - 4:4 - Decision Tree Learner (entrenamiento)

File

Options PMMLSettings Flow Variables Job Manager Selection

General

Class column **S** Customer type

Quality measure Gain ratio

Pruning method No pruning

☒ Reduced Error Pruning

Min number records per node 10

Number records to store for view 10,000

☒ Average split point

Number threads 8

☐ Skip nominal columns without domain information

Root split

☐ Force root split column

Root split column **S** Gender

Binary nominal splits

☐ Binary nominal splits

Max #nominal 10

☐ Filter invalid attribute values in child nodes

OK Apply Cancel ?

4. Resultados Obtenidos.

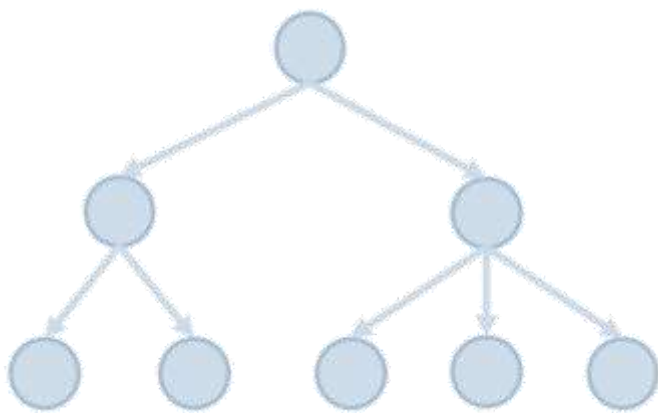
Customer t...	Member	Normal
Member	21	22
Normal	17	23

	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 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	$\frac{a+d}{a+b+c+d} = \frac{23+21}{23+17+22+21} = \frac{44}{83} = 0.53$	El 53% de los registros totales, fueron clasificados de forma CORRECTA. Accuracy: 53.012%
Tasa de error	$\frac{b+c}{a+b+c+d} = \frac{17+22}{23+17+22+21} = \frac{39}{83} = 0.46$	El 46% de los registros totales, fueron clasificados de forma EQUIVOCADA. Error: 46.988%
Precisión	$\frac{d}{b+d} = \frac{21}{17+21} = \frac{21}{38} = 0.51$	El 51% de los ejemplos clasificados como clase positiva , son realmente positivos .
Sensibilidad (<i>Recall</i>)	$\frac{d}{c+d} = \frac{21}{22+21} = \frac{21}{43} = 0.57$	El clasificador puede reconocer muestras positivas en el 57% de los casos.
Tasa de positivos falsos	$\frac{b}{a+b} = \frac{17}{23+17} = \frac{17}{40} = 0.51$	La tasa de positivos falsos es del 51%.
Tasa de negativos falsos	$\frac{c}{c+d} = \frac{22}{22+21} = \frac{22}{43} = 0.42$	La tasa de negativos falsos es del 42%.
Especificidad	$\frac{a}{a+b} = \frac{23}{23+17} = \frac{23}{40} = 0.48$	El clasificador puede reconocer muestras negativas en el 48% de los casos.

5. Reglas Obtenidas.

S	Rule	D	Record count	D	Number of correct
	\$Payment\$ = "Ewallet" AND \$Product line\$ = "Health and beauty" => "Normal"	18	13		
	\$Payment\$ = "Cash" AND \$Product line\$ = "Health and beauty" => "Member"	13	8		
	\$Payment\$ = "Credit card" AND \$Product line\$ = "Health and beauty" => "Member"	18	11		
	\$Product line\$ = "Electronic accessories" AND TRUE => "Normal"	51	28		
	\$Payment\$ = "Ewallet" AND \$Product line\$ = "Home and lifestyle" => "Normal"	24	15		
	\$Payment\$ = "Cash" AND \$Product line\$ = "Home and lifestyle" => "Member"	20	12		
	\$Payment\$ = "Credit card" AND \$Product line\$ = "Home and lifestyle" => "Normal"	17	9		
	\$Gender\$ = "Female" AND \$Payment\$ = "Ewallet" AND \$Product line\$ = "Fashion accessories" => "Normal"	11	7		
	\$Gender\$ = "Male" AND \$Payment\$ = "Ewallet" AND \$Product line\$ = "Fashion accessories" => "Member"	11	7		
	\$Payment\$ = "Cash" AND \$Product line\$ = "Fashion accessories" => "Normal"	15	9		
	\$Payment\$ = "Credit card" AND \$Product line\$ = "Fashion accessories" => "Normal"	23	13		
	\$Gender\$ = "Female" AND \$Product line\$ = "Sports and travel" => "Member"	33	20		
	\$Gender\$ = "Male" AND \$Product line\$ = "Sports and travel" => "Normal"	23	12		
	\$Product line\$ = "Food and beverages" AND TRUE => "Member"	54	39		



Árbol C4.5

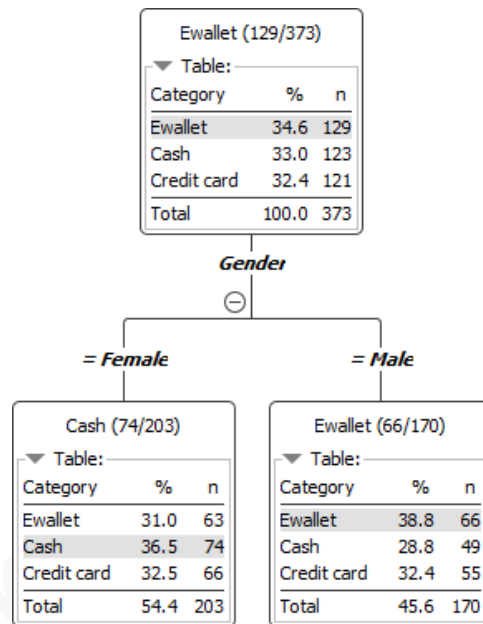
ABSTRACT

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

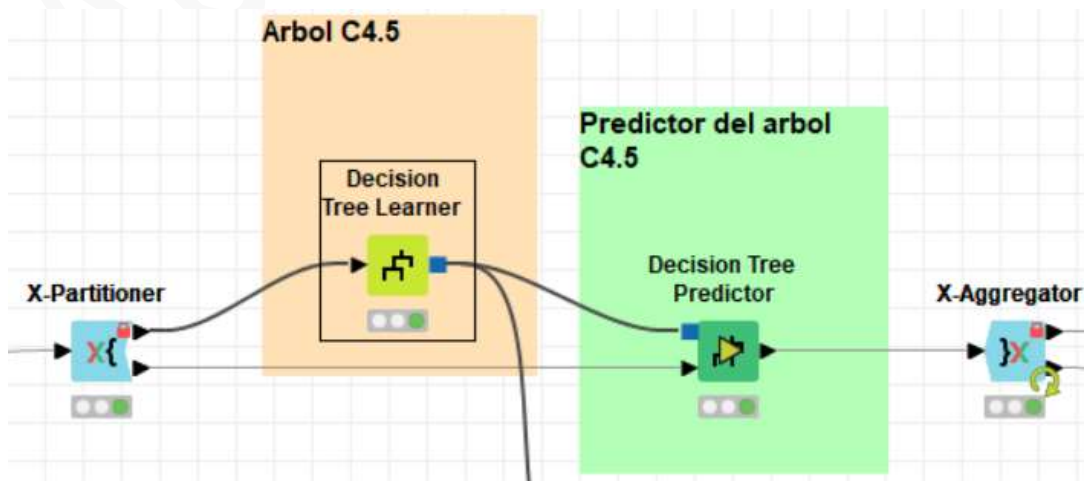
1. Diccionario de datos.

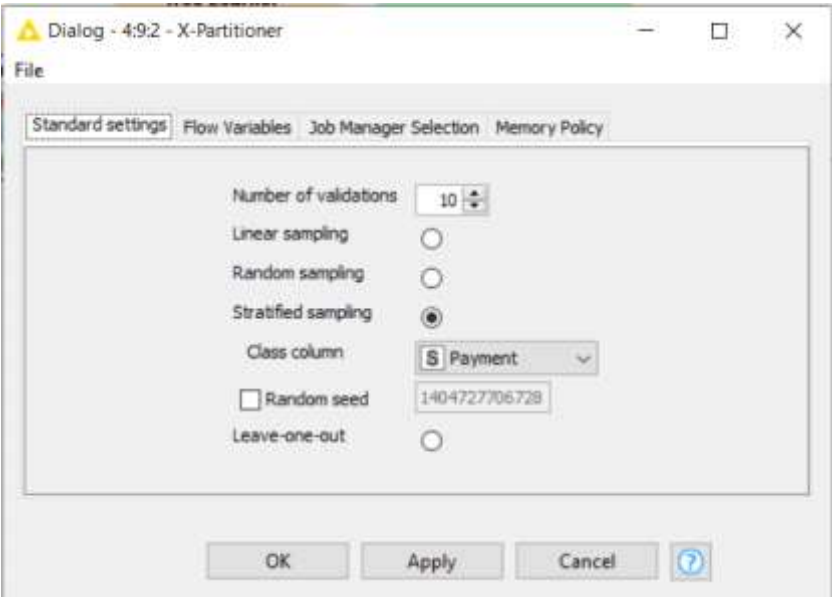
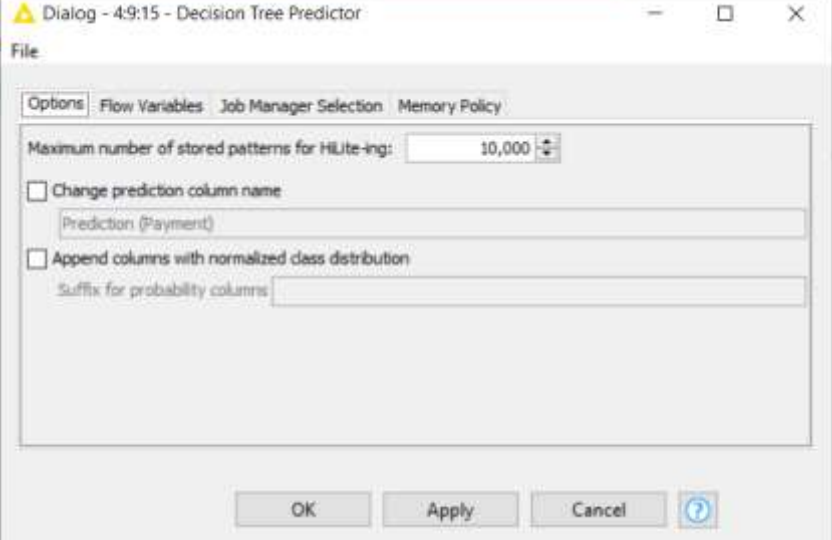
No	Nombre	Tipo	Dominio
0	Customer type	Categorico	Member o Normal
1	Gender	Categorico	Female, Male
2	Payment	Numérico	Forma de pago Ewallet, Credit card, Cash

2. Diagrama general del árbol.



3. Configuración.



Nodo	Configuración
X-Partitioner	<p>Se utilizó el <i>muestreo estratificado</i> para generar el árbol, con 10 validaciones.</p> 
Decision Tree Predictor	

Decision Tree Learner

Dialog - 4:9:16 - Decision Tree Learner

File

Options | PMMLSettings | Flow Variables | Job Manager Selection

General

Class column: Payment

Quality measure:

Pruning method:

☒ Reduced Error Pruning

Min number records per node:

Number records to store for view:

☒ Average split point

Number threads:

☒ Skip nominal columns without domain information

Root split

☐ Force root split column

Root split column: Gender

Binary nominal splits

☐ Binary nominal splits

Max #nominal:

☐ Filter invalid attribute values in child nodes

OK Apply Cancel ?

X-Aggregator

Dialog - 4:9:1 - X-Aggregator

File

Standard settings | Flow Variables | Job Manager Selection | Memory Policy

Target column: Payment

Prediction column: Prediction (Payment)

☐ Add column with fold id

OK Apply Cancel ?

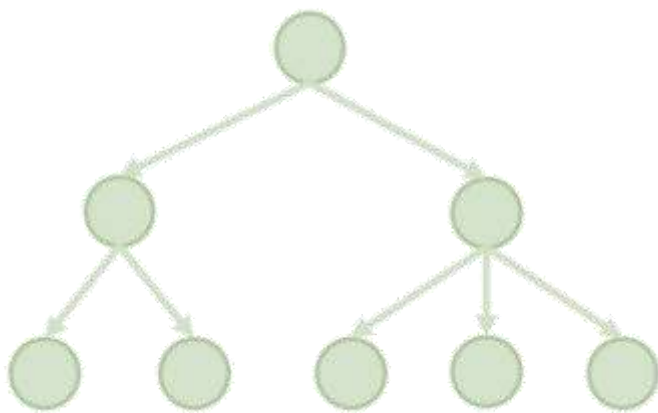
4. Resultados Obtenidos.

Payment \ ...	Ewallet	Cash	Credit card
Ewallet	79	55	9
Cash	73	55	8
Credit card	69	61	5

Medida	Cálculo	Explicación
Tasa de exactitud	$\frac{139}{414} = 0.3357$	El 33.57% de los registros totales, fueron clasificados de forma CORRECTA. Accuracy: 33.575%
Tasa de error	$\frac{275}{414} = 0.6642$	El 66.42% de los registros totales, fueron clasificados de forma EQUIVOCADA. Error: 66.425%

5. Reglas Obtenidas.

S Rule	D Record count	D Number of correct
\$Gender\$ = "Female" AND TRUE => "Cash"	203	74
\$Gender\$ = "Male" AND TRUE => "Ewallet"	170	66



Árbol CART

FUNCIÓN:

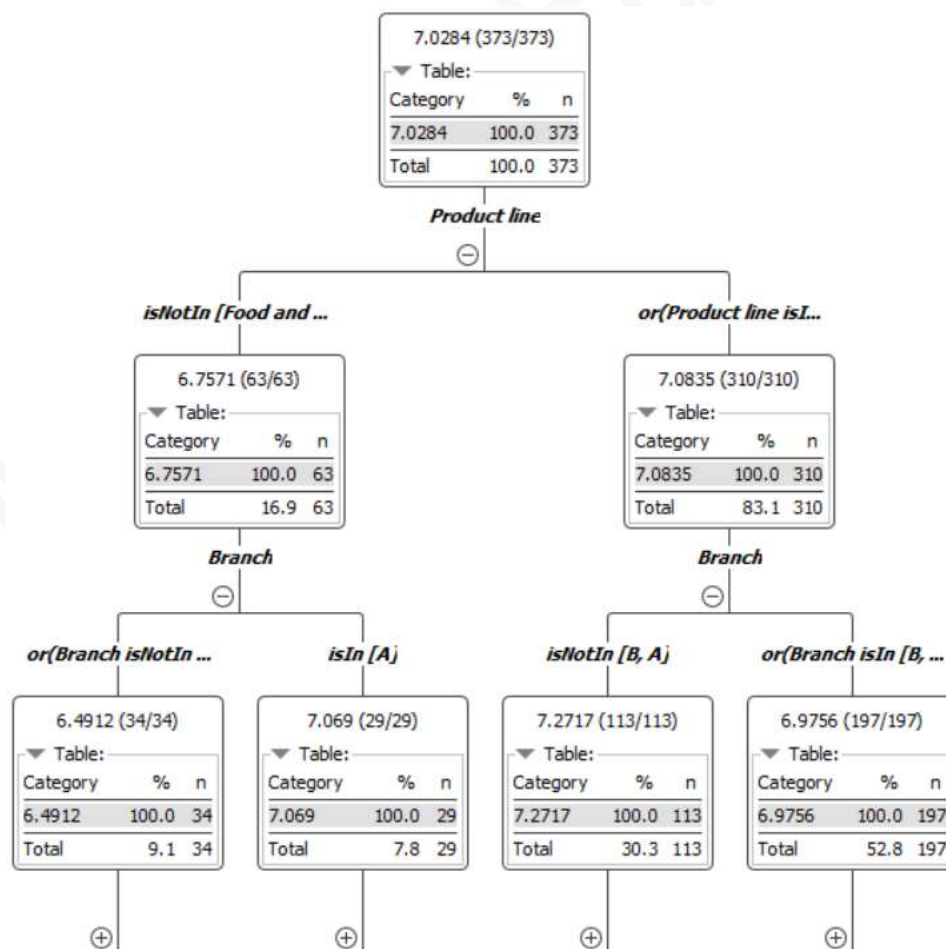
Predice la CALIFICACIÓN otorgada por el cliente.

4. Diccionario de datos.

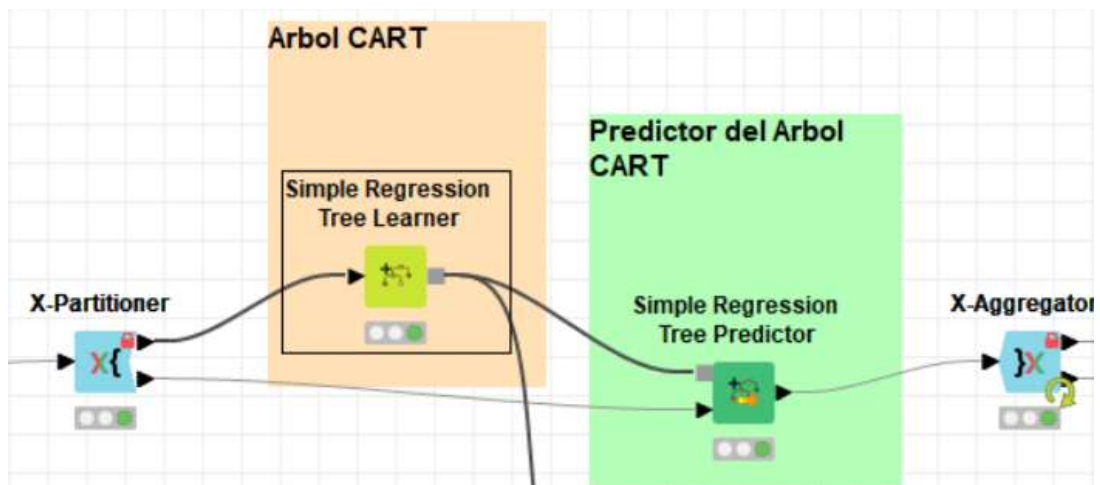
No	Nombre	Tipo	Dominio
0	Branch	Categórico	Letra desde A-C
1	Customer type	Categórico	Member o Normal
2	Gender	Categorico	Female, Male
3	Product line	Categórico	Categoría de Producto
4	Rating	Numérico	Calificación de estratificación del cliente en su experiencia de compra general (en una escala de 1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

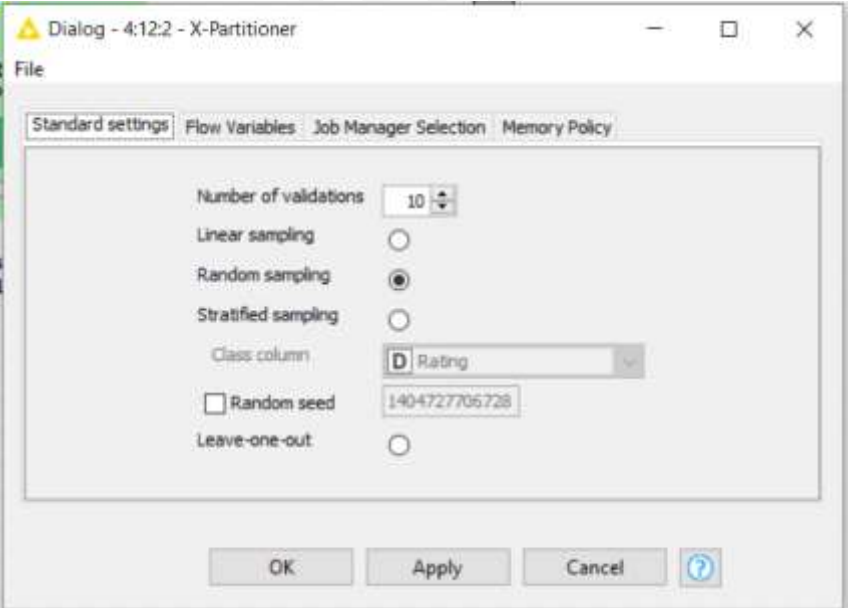
5. Diagrama general del árbol.

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**.



6. Configuración.



Nodo	Configuración
X-Partitioner	<p>Se utilizó el <i>muestreo aleatorio</i> para generar el árbol, con 10 validaciones.</p>  <p>Dialog - 4:12:2 - X-Partitioner</p> <p>File</p> <p>Standard settings Flow Variables Job Manager Selection Memory Policy</p> <p>Number of validations: 10</p> <p>Linear sampling: <input type="radio"/></p> <p>Random sampling: <input checked="" type="radio"/></p> <p>Stratified sampling: <input type="radio"/></p> <p>Class column: D Rating</p> <p><input type="checkbox"/> Random seed: 1404727706728</p> <p>Leave-one-out: <input type="radio"/></p> <p>OK Apply Cancel ?</p>

Decision Tree Learner

Dialog - 4:12:17 - Simple Regression Tree Learner

File

Options | Flow Variables | Job Manager Selection

Target Column: D Rating

Attribute Selection

☐ Use fingerprint attribute: [no valid fingerprint inputs]

☒ Use column attributes

☒ Manual Selection ☐ Wildcard/Regex Selection

Exclude

Filter

- City
- Unit price
- Quantity
- Tax 5%
- Total
- Payment
- costs
- gross margin percentage

☒ Enforce exclusion

Include

Filter

- Brand
- Customer type
- Gender
- Product line

☐ Enforce inclusion

Misc Options

☐ Ignore columns without domain information

☐ Enable Highlighting (#patterns to store): 3,000

Tree Options

☒ Use binary splits for nominal attributes

Missing value handling: XGBoost

☐ Limit number of levels (tree depth): 10

☐ Minimum split node size: 1

OK Apply Cancel ?

Decision Tree Predictor

Dialog - 4:12:18 - Simple Regression Tree Predictor

File

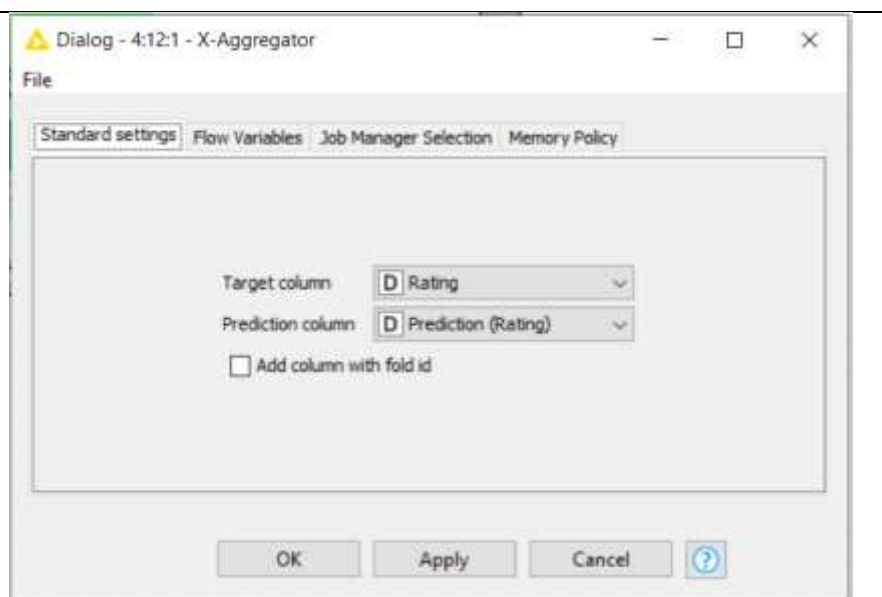
Prediction Settings | Flow Variables | Job Manager Selection | Memory Policy

☐ Change prediction column name

Prediction column name: Prediction (Rating)

OK Apply Cancel ?

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.

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

1. Se calcula la Entropía total

<table border="1"> <tr> <th colspan="2">evalcarro</th></tr> <tr> <td>acc</td><td>unacc</td></tr> <tr> <td>15</td><td>5</td></tr> </table>	evalcarro		acc	unacc	15	5	<p>No. total de Registros = 20</p> $E(s) = \sum_{i=1}^c -p_i \log_2(p_i)$ $E(s) = \left(-\frac{15}{20} \log_2\left(\frac{15}{20}\right)\right) + \left(-\frac{5}{20} \log_2\left(\frac{5}{20}\right)\right)$ $E(s) = (0.31127812) + (0.5)$ $E(s) = 0.81127812$
evalcarro							
acc	unacc						
15	5						

2. 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

3. Se calcula la Entropía en cada rama, y se suman proporcionalmente para calcular la Entropía total
4. Se calcula la Ganancia de Información de cada rama

				Count
precioCompra	<i>vhigh</i>	evalcarro	acc	7
			unacc	3
	<i>Med</i>	evalcarro	acc	3
			unacc	0
	<i>high</i>	evalcarro	acc	5
			unacc	2

$$E(T, X) = \sum_{C \in X} p(C) E(C)$$

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

$$\sum_{i=1}^c -p_i \log_2(p_i) = \left(-\frac{7}{20} \log_2\left(\frac{7}{20}\right)\right) + \left(-\frac{3}{20} \log_2\left(\frac{3}{20}\right)\right) = 0.53 + 0.41 = 0.94$$

$$E(\text{high}) = E(\text{ACC}, \text{UNACC}) = 0.83$$

$$\sum_{i=1}^c -p_i \log_2(p_i) = \left(-\frac{5}{20} \log_2\left(\frac{5}{20}\right)\right) + \left(-\frac{2}{20} \log_2\left(\frac{2}{20}\right)\right) = 0.5 + 0.33 = 0.83$$

$$E(\text{Med}) = E(\text{ACC}, \text{UNACC}) = 0.41 = 0$$

$$\sum_{i=1}^c -p_i \log_2(p_i) = \left(-\frac{3}{20} \log_2\left(\frac{3}{20}\right)\right) = 0.41$$

Probabilidades

$$V_{\text{high}}: P(10/20) = 0.5$$

$$Med: P(3/20) = 0.15$$

$$high: P(7/20) = 0.35$$

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

$$E(\text{evalcarro}, \text{precioCompra}) = 0.76$$

$$GAIN(T, X) = E(T) - E(T, X)$$

$$GAIN(T, X) = 0.81 - 0.76 = 0.05$$

$$GAIN(T, X) = 0.05$$

				Count
costoManto	<i>med</i>	evalcarro	acc	12
			unacc	0
	<i>low</i>	evalcarro	acc	3
			unacc	2
	<i>vhigh</i>	evalcarro	acc	0
			unacc	3

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

$$E(\text{med}) = E(\text{ACC}, \text{UNACC}) = 0.44$$

$$\sum_{i=1}^c -p_i \log_2(p_i) = \left(-\frac{12}{20} \log_2\left(\frac{12}{20}\right)\right) = 0.44$$

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

$$\sum_{i=1}^c -p_i \log_2(p_i) = \left(-\frac{3}{20} \log_2\left(\frac{3}{20}\right)\right) + \left(-\frac{2}{20} \log_2\left(\frac{2}{20}\right)\right) = 0.41 + 0.33 = 0.74$$

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

$$\sum_{i=1}^c -p_i \log_2(p_i) = \left(-\frac{3}{20} \log_2\left(\frac{3}{20}\right)\right) = 0.41$$

Probabilidades

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

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

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

$$GAIN(T, X) = E(T) - E(T, X)$$

$$GAIN(T, X) = 0.81 - 0.185 = 0.625$$

$$GAIN(T, X) = 0.625$$

				Count
noPuertas	<i>dos</i>	evalcarro	acc	5
			unacc	3
	<i>tres</i>	evalcarro	acc	3
			unacc	1
	<i>cuatro</i>	evalcarro	acc	4
			unacc	1
	<i>5more</i>	evalcarro	acc	3
			unacc	0

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

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

$$\sum_{i=1}^c -p_i \log_2(p_i) = \left(-\frac{5}{20} \log_2\left(\frac{5}{20}\right)\right) + \left(-\frac{3}{20} \log_2\left(\frac{3}{20}\right)\right) = 0.5 + 0.41 = 0.91$$

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

$$\sum_{i=1}^c -p_i \log_2(p_i) = \left(-\frac{3}{20} \log_2\left(\frac{3}{20}\right)\right) + \left(-\frac{1}{20} \log_2\left(\frac{1}{20}\right)\right) = 0.41 + 0.21 = 0.62$$

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

$$\sum_{i=1}^c -p_i \log_2(p_i) = \left(-\frac{4}{20} \log_2\left(\frac{4}{20}\right)\right) + \left(-\frac{1}{20} \log_2\left(\frac{1}{20}\right)\right) = 0.46 + 0.21 = 0.67$$

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

$$\sum_{i=1}^c -p_i \log_2(p_i) = \left(-\frac{3}{20} \log_2\left(\frac{3}{20}\right)\right) = 0.41$$

Probabilidades

$$dos: P(8/20) = 0.4$$

$$tres: P(4/20) = 0.2$$

$$cuatro: P(5/20) = 0.25$$

$$5more: P(3/20) = 0.15$$

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

$$E(\text{evalcarro}, \text{noPuertas}) = 0.65$$

$$GAIN(T, X) = E(T) - E(T, X)$$

$$GAIN(T, X) = 0.81 - 0.65 = 0.16$$

$$GAIN(T, X) = 0.16$$

				Count
noPasajeros	<i>cuatro</i>	evalcarro	acc	9
			unacc	0
	<i>tres</i>	evalcarro	acc	1
			unacc	2
	<i>more</i>	evalcarro	acc	4
			unacc	1
	<i>dos</i>	evalcarro	acc	1
			unacc	2

$$E(T, X) = \sum_{C \in X} p(C)E(C)$$

$$E(\text{Cuatro}) = E(\text{ACC}, \text{UNACC}) = 0.51$$

$$\sum_{i=1}^c -p_i \log_2(p_i) = \left(-\frac{9}{20} \log_2 \left(\frac{9}{20} \right) \right) = 0.51$$

$$E(\text{tres}) = E(\text{ACC}, \text{UNACC}) = 0.54$$

$$\sum_{i=1}^c -p_i \log_2(p_i) = \left(-\frac{1}{20} \log_2 \left(\frac{1}{20} \right) \right) + \left(-\frac{2}{20} \log_2 \left(\frac{2}{20} \right) \right) = 0.21 + 0.33 = 0.54$$

$$E(\text{more}) = E(\text{ACC}, \text{UNACC}) = 0.67$$

$$\sum_{i=1}^c -p_i \log_2(p_i) = \left(-\frac{4}{20} \log_2 \left(\frac{4}{20} \right) \right) + \left(-\frac{1}{20} \log_2 \left(\frac{1}{20} \right) \right) = 0.46 + 0.21 = 0.67$$

$$E(\text{dos}) = E(\text{ACC}, \text{UNACC}) = 0.54$$

$$\sum_{i=1}^c -p_i \log_2(p_i) = \left(-\frac{1}{20} \log_2 \left(\frac{1}{20} \right) \right) + \left(-\frac{2}{20} \log_2 \left(\frac{2}{20} \right) \right) = 0.21 + 0.33 = 0.54$$

Probabilidades

$$\text{cuatro: } P(9/20) = 0.45$$

$$\text{tres: } P(3/20) = 0.15$$

$$\text{more: } P(5/20) = 0.25$$

$$\text{dos: } P(3/20) = 0.15$$

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

$$E(\text{evalcarro}, \text{noPasajeros}) = 0.32$$

$$GAIN(T, X) = E(T) - E(T, X)$$

$$GAIN(T, X) = 0.81 - 0.32 = 0.49$$

$$GAIN(T, X) = 0.49$$

				Count
tamCajuela	<i>small</i>	evalcarro	acc	4
			unacc	3
	<i>big</i>	evalcarro	acc	5
			unacc	1
	<i>med</i>	evalcarro	acc	6
			unacc	1

$$E(T, X) = \sum_{C \in X} p(C)E(C)$$

$$E(\text{small}) = E(\text{ACC}, \text{UNACC}) = 0.87$$

$$\sum_{i=1}^c -p_i \log_2(p_i) = \left(-\frac{4}{20} \log_2\left(\frac{4}{20}\right)\right) + \left(-\frac{3}{20} \log_2\left(\frac{3}{20}\right)\right) = 0.46 + 0.41 = 0.87$$

$$E(\text{big}) = E(\text{ACC}, \text{UNACC}) = 0.71$$

$$\sum_{i=1}^c -p_i \log_2(p_i) = \left(-\frac{5}{20} \log_2\left(\frac{5}{20}\right)\right) + \left(-\frac{1}{20} \log_2\left(\frac{1}{20}\right)\right) = 0.5 + 0.21 = 0.71$$

$$E(\text{med}) = E(\text{ACC}, \text{UNACC}) = 0.73$$

$$\sum_{i=1}^c -p_i \log_2(p_i) = \left(-\frac{6}{20} \log_2\left(\frac{6}{20}\right)\right) + \left(-\frac{1}{20} \log_2\left(\frac{1}{20}\right)\right) = 0.52 + 0.21 = 0.73$$

Probabilidades

$$\text{small: } P(7/20) = 0.35$$

$$\text{big: } P(6/20) = 0.3$$

$$\text{med: } P(7/20) = 0.35$$

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

$$E(\text{evalcarro tamCajuela}) = 0.77$$

$$GAIN(T, X) = E(T) - E(T, X)$$

$$GAIN(T, X) = 0.81 - 0.77 = 0.04$$

$$GAIN(T, X) = 0.04$$

				Count
Seguridas	<i>low</i>	evalcarro	acc	2
			unacc	2
	<i>high</i>	evalcarro	acc	7
			unacc	2
	<i>med</i>	evalcarro	acc	6
			unacc	1

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

$$E(\text{low}) = E(\text{ACC}, \text{UNACC}) = 0.66$$

$$\sum_{i=1}^c -p_i \log_2(p_i) = \left(-\frac{2}{20} \log_2\left(\frac{2}{20}\right)\right) + \left(-\frac{2}{20} \log_2\left(\frac{2}{20}\right)\right) = 0.33 + 0.33 = 0.66$$

$$E(\text{high}) = E(\text{ACC}, \text{UNACC}) = 0.86$$

$$\sum_{i=1}^c -p_i \log_2(p_i) = \left(-\frac{7}{20} \log_2\left(\frac{7}{20}\right)\right) + \left(-\frac{2}{20} \log_2\left(\frac{2}{20}\right)\right) = 0.53 + 0.33 = 0.86$$

$$E(\text{med}) = E(\text{ACC}, \text{UNACC}) = 0.73$$

$$\sum_{i=1}^c -p_i \log_2(p_i) = \left(-\frac{6}{20} \log_2\left(\frac{6}{20}\right)\right) + \left(-\frac{1}{20} \log_2\left(\frac{1}{20}\right)\right) = 0.52 + 0.21 = 0.73$$

Probabilidades

$$\text{low: } P(4/20) = 0.2$$

$$\text{high: } P(9/20) = 0.45$$

$$\text{med: } P(7/20) = 0.35$$

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

$$E(\text{evalcarro tamCajuela}) = 0.77$$

$$GAIN(T, X) = E(T) - E(T, X)$$

$$GAIN(T, X) = 0.81 - 0.77 = 0.04$$

$$GAIN(T, X) = 0.04$$

5. Elegir el Atributo con mayor ganancia de Informacion

<i><u>Variable</u></i>	<i><u>Ganancia</u></i>
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.

1	Gender	EthnicGroup	ParentEduc	ParentMaritalStatus	TestPrep
2	female	group A	bachelor's degree	married	none
3	female	group C	some college	married	none
4	female	group B	master's degree	single	none
5	male	group A	associate's degree	married	none
6	male	group C	some college	married	none
7	female	group B	associate's degree	married	none
8	female	group B	some college	widowed	completed
9	male	group B	some college	married	none
10	male	group D	associate's degree	single	completed
11	female	group B	associate's degree	married	none
12	male	group C	associate's degree	married	none
13	male	group D	associate's degree	divorced	none
14	female	group B	bachelor's degree	married	none
15	male	group A	some college	single	completed
16	male	group C	bachelor's degree	married	completed

<table border="1"> <tr> <th colspan="2">Gender</th></tr> <tr> <td>male</td><td>female</td></tr> <tr> <td>8</td><td>7</td></tr> </table>	Gender		male	female	8	7	<p>No. total de Registros = 20</p> $E(s) = \sum_{i=1}^c -p_i \log_2(p_i)$ $E(s) = \left(-\frac{8}{15} \log_2\left(\frac{8}{15}\right)\right) + \left(-\frac{7}{15} \log_2\left(\frac{7}{15}\right)\right)$ $E(s) = (0.31127812) + (0.5)$ $E(s) = 0.9967$
Gender							
male	female						
8	7						

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

$$E(T, X) = \sum_{C \in X} p(C)E(C)$$

$$E(\text{group A}) = E(\text{female, male}) = 0.64$$

$$\sum_{i=1}^c -p_i \log_2(p_i) = \left(-\frac{1}{15} \log_2\left(\frac{1}{15}\right)\right) + \left(-\frac{2}{15} \log_2\left(\frac{2}{15}\right)\right) = 0.64$$

$$E(\text{group B}) = E(\text{female, male}) = 0.78$$

$$\sum_{i=1}^c -p_i \log_2(p_i) = \left(-\frac{5}{15} \log_2\left(\frac{5}{15}\right)\right) + \left(-\frac{1}{15} \log_2\left(\frac{1}{15}\right)\right) = 0.78$$

$$E(\text{Med}) = E(\text{ACC, UNACC}) = 0.72$$

$$\sum_{i=1}^c -p_i \log_2(p_i) = \left(-\frac{1}{15} \log_2\left(\frac{1}{15}\right)\right) + \left(-\frac{3}{15} \log_2\left(\frac{3}{15}\right)\right) = 0.72$$

Probabilidades

$$\text{group A: } P(3/15) = 0.2 \quad \text{group B: } P(6/15) = 0.4 \quad \text{group C: } P(4/15) = 0.26$$

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

$$E(\text{evalcarro, precioCompra}) = 0.62$$

$$GAIN(T, X) = E(T) - E(T, X)$$

$$GAIN(T, X) = 0.99 - 0.62 = 0.37$$

$$GAIN(T, X) = 0.37$$

				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

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

$$E(\text{bachelor's degree}) = E(\text{female, male}) = 0.64$$

$$\sum_{i=1}^c -p_i \log_2(p_i) = \left(-\frac{2}{15} \log_2\left(\frac{2}{15}\right)\right) + \left(-\frac{1}{15} \log_2\left(\frac{1}{15}\right)\right) = 0.64$$

$$E(\text{some college}) = E(\text{female, male}) = 0.85$$

$$\sum_{i=1}^c -p_i \log_2(p_i) = \left(-\frac{2}{15} \log_2\left(\frac{2}{15}\right)\right) + \left(-\frac{3}{15} \log_2\left(\frac{3}{15}\right)\right) = 0.85$$

$$E(\text{master's degree}) = E(\text{female, male}) = 0$$

$$\sum_{i=1}^c -p_i \log_2(p_i) = \left(-\frac{1}{15} \log_2\left(\frac{1}{15}\right)\right) + \left(-\frac{0}{15} \log_2\left(\frac{0}{15}\right)\right) = 0.$$

$$E(\text{associate's degree}) = E(\text{female, male}) = 0.89$$

$$\sum_{i=1}^c -p_i \log_2(p_i) = \left(-\frac{2}{15} \log_2\left(\frac{2}{15}\right)\right) + \left(-\frac{4}{15} \log_2\left(\frac{4}{15}\right)\right) = 0.89$$

Probabilidades

$$\text{bachelor's degree: } P(3/15) = 0.2$$

$$\text{some college: } P(5/15) = 0.33$$

$$\text{master's degree: } P(1/15) = 0.06$$

$$\text{associate's degree: } P(6/15) = 0.4$$

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

$$E(\text{evalcarro}, \text{precioCompra}) = 0.76$$

$$GAIN(T, X) = E(T) - E(T, X)$$

$$GAIN(T, X) = 0.99 - 0.76 = 0.23$$

$$GAIN(T, X) = 0.23$$

				Count
ParentMarital Status	<i>married</i>	Gender	<i>female</i>	5
			<i>male</i>	5
	<i>single</i>	Gender	<i>female</i>	1
			<i>male</i>	2
	<i>widowed</i>	Gender	<i>female</i>	1
			<i>male</i>	0
	<i>divorced</i>	Gender	<i>female</i>	0
			<i>male</i>	1

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

$$E(\text{married}) = E(\text{female, male}) = 1.05$$

$$\sum_{i=1}^c -p_i \log_2(p_i) = \left(-\frac{5}{15} \log_2\left(\frac{5}{15}\right)\right) + \left(-\frac{5}{15} \log_2\left(\frac{5}{15}\right)\right) = 1.05$$

$$E(\text{single}) = E(\text{female, male}) = 0.64$$

$$\sum_{i=1}^c -p_i \log_2(p_i) = \left(-\frac{1}{15} \log_2\left(\frac{1}{15}\right)\right) + \left(-\frac{2}{15} \log_2\left(\frac{2}{15}\right)\right) = 0.64$$

$$E(\text{widowed}) = E(\text{female, male}) = 0$$

$$\sum_{i=1}^c -p_i \log_2(p_i) = \left(-\frac{1}{15} \log_2\left(\frac{1}{15}\right)\right) + \left(-\frac{0}{15} \log_2\left(\frac{0}{15}\right)\right) = 0.$$

$$E(\text{divorced}) = E(\text{female, male}) = 0$$

$$\sum_{i=1}^c -p_i \log_2(p_i) = \left(-\frac{0}{15} \log_2\left(\frac{0}{15}\right)\right) + \left(-\frac{1}{15} \log_2\left(\frac{1}{15}\right)\right) = 0$$

Probabilidades

$$\text{married: } P(10/15) = 0.66$$

$$\text{single: } P(3/15) = 0.2$$

$$\text{master's degree: } P(1/15) = 0.06$$

$$\text{associate's degree: } P(1/15) = 0.06$$

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

$$E(\text{evalcarro}, \text{precioCompra}) = 0.82$$

$$GAIN(T, X) = E(T) - E(T, X)$$

$$GAIN(T, X) = 0.99 - 0.82 = 0.17$$

$$GAIN(T, X) = 0.17$$

				Count
TestPrep	none	Gender	female	6
			male	5
	completed	Gender	female	1
			male	3

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

$$E(\text{none}) = E(\text{female}, \text{male}) =$$

$$\sum_{i=1}^c -p_i \log_2(p_i) = \left(-\frac{6}{15} \log_2\left(\frac{6}{15}\right)\right) + \left(-\frac{5}{15} \log_2\left(\frac{5}{15}\right)\right) = 1.05$$

$$E(\text{group B}) = E(\text{female}, \text{male}) = 0.78$$

$$\sum_{i=1}^c -p_i \log_2(p_i) = \left(-\frac{1}{15} \log_2\left(\frac{1}{15}\right)\right) + \left(-\frac{3}{15} \log_2\left(\frac{3}{15}\right)\right) = 0.72$$

Probabilidades

$$\text{married: } P(11/15) = 0.773$$

$$\text{single: } P(4/15) = 0.26$$

<u>Variable</u>	<u>Ganancia</u>
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	$\frac{a+d}{a+b+c+d} = \frac{3+40}{40+5+2+3} = \frac{43}{50} = 0.86$	El 86% de los registros totales, fueron clasificados de forma CORRECTA.
Tasa de error	$\frac{b+c}{a+b+c+d} = \frac{2+5}{40+5+2+3} = \frac{7}{50} = 0.14$	El 14% de los registros totales, fueron clasificados de forma EQUIVOCADA.
Precisión	$\frac{d}{b+d} = \frac{40}{2+40} = \frac{20}{21} = 0.9523$	El 95.23% de los ejemplos clasificados como clase positiva , son realmente positivos .
Sensibilidad (<i>Recall</i>)	$\frac{d}{c+d} = \frac{40}{5+40} = \frac{8}{9} = 0.8888$	El clasificador puede reconocer muestras positivas en el 88.88% de los casos.
Tasa de positivos falsos	$\frac{b}{a+b} = \frac{2}{3+2} = \frac{2}{5} = 0.4$	La tasa de positivos falsos es del 40%.
Tasa de negativos falsos	$\frac{c}{c+d} = \frac{5}{5+40} = \frac{1}{9} = 0.1111$	La tasa de negativos falsos es del 11.11%.
Especificidad	$\frac{a}{a+b} = \frac{3}{3+2} = \frac{3}{5} = 0.6$	El clasificador puede reconocer muestras negativas en el 60% de los casos.