TEMA 7. EXTRACCIÓN DE REGLAS DE ASOCIACIÓN

Contenidos

I. Introducción
II. Definiciones
III. Extracción de reglas de asociación
IV. Apriori
V. Reglas de asociación en Orange
VI. Actividad

Introducción

PROBLEMA

Dado un conjunto de eventos (objetos), encontrar reglas que describan relaciones causa-efecto entre ellos

Ejemplo: la cesta de la compra

Detectar cuándo la ocurrencia de un artículo esta asociada a la ocurrencia de otros artículos en la misma compra

Introducción

ANÁLISIS DE LA CESTA DE LA COMPRA

Origen de las reglas de asociación

Aplicaciones:

Diseño de catálogos, distribución de los productos en los supermercados, . . .

Introducción

Colocación de productos en las estanterías de un supermercado

- **Objetivo:** Identificar productos que muchos clientes compran juntos.
- **Solución:** Procesar los datos de los terminales de punto de venta proporcionados por los escáneres de código de barras.
- **Ejemplo:** Si un cliente hombre compra pañales, es muy probable que compre cerveza

Introducción

Promociones y ofertas

Si identificamos una regla del tipo: **{café} → {leche}**

- Aquellos clientes que compran café, tienen tendencia a comprar leche.
- No ocurre lo mismo a la inversa: los que compran leche no muestran una tendencia especial a comprar café.

Introducción

En la terminología del "análisis de la cesta de la compra" los datos consisten en:

- Una serie de registros de transacciones
- Cada registro contiene un conjunto de artículos comprados por un cliente

Cestas de la compra

Cliente	Compras
1	Pan, leche, vino, huevos
2	Pan, pañales, cerveza
3	Leche, pañales, vino, cerveza

Productos

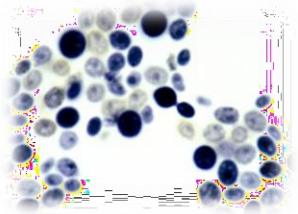
Pan
Leche
Vino
Huevos
Pañales
Cerveza

Los datos se pueden visualizar como una matriz:

Cliente	Pan	Leche	Vino	Huevos	Pañales	Cerveza
1	1	1	1	1	0	0
2	1	0	1	0	1	1
3	0	1	1	0	1	1

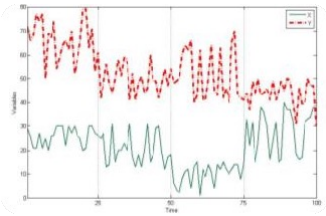
Aplicaciones Reales

Microarray de Spellman y ciclo celular de la levadura



Predicciones agronómicas

Series temporales multidimensionales



Aplicaciones

Microarray de inflamación y respuesta inmune

Datos sísmicos temporales

Enfermedad Alzheimer

Contenido total de ozono en la atmósfera mediante mediciones por satélite

Series temporales del mundo real: Contaminación atmosférica O_3, NO, SO_2

Aplicaciones Reales. Contaminación Atmosférica

Reglas de asociación encontradas por QARGA para condiciones climatológicas con respecto a O_3 (mg / m^3) y SO_2 (mg / m^3)

Rule	Conf. (%)	Lift
Temp. E [38, 42] and Hum. E [25, 33] and Hour E [15, 18] ⇒ O_3 E [140, 206]	90	6.61
Temp. E [16, 22] and Hum. E [75, 90] ⇒ O_3 E [22, 110]	100	1.43
Temp. E [42.9, 45.0] ⇒ SO_2 E [3.7, 7.5]	100	1.72

Reglas de asociación encontradas por Apriori

Rule (%)	Conf.	Lift	#intervals discretization
Temp. E [24, 27] ⇒ O_3 E [90, 115]	33	1.43	10
Hum. E [14, 40] and Dir. E [120, 240] ⇒ O_3 E [99, 183]	73	1.80	3

*** Todas las variables fueron proporcionadas por la estación meteorológica de la ciudad de Sevilla (Spain)**

M. Martínez Ballesteros, F.Martínez-Álvarez, A. Troncoso Lora, J.C. Riquelme Santos. "An Evolutionary Algorithm to Discover Quantitative Association Rules in Multidimensional Time Series". Soft Computing. Vol. 15, No. 10, pp. 2065-2084. 2011. ISSN: 1432-7643. DOI:10.1007/S00500-011-0705-4.

Definiciones

- **Itemset:**
  - Conjunto de uno o más items, p.ej. {pan, leche}
  - **K-itemset.** Itemset con k elementos
- **Soporte** de un itemset (support):
  - Frecuencia relativa de aparición del itemset en cuestión dentro de todos los itemsets del conjunto de datos

    p.ej. sup({pañales, cerveza}) = 2/3
- **Itemset frecuente:**
  - Itemset con soporte igual o superior a un umbral de soporte establecido por el usuario (MinSup).

Itemsets del conjunto de datos:

Cliente	Compras
1	Pan, leche, vino, huevos
2	Pan, pañales, cerveza
3	Leche, pañales, vino, cerveza

Definiciones

Descubrimiento de Reglas de Asociación:

Aprendizaje no supervisado para descubrir relaciones entre atributos

Una regla de asociación es una **implicación** de la forma:

$$A \Rightarrow C$$

Antecedente **Consecuente**

A y C están formados por itemsets

Definiciones

- **Reglas de asociación Booleanas:**
  - Asociaciones entre la presencia y ausencia de items. *E.g. compra A o no compra A*.

$$\text{Leche y Huevos} \Rightarrow \text{Pan}$$

- **Reglas de asociación nominales:**
  - Asociaciones entre las propiedades o valores de items (discretos).

$$\text{Temperatura es FRÍA y Humedad es NORMAL} \Rightarrow \text{Jugar es SI}$$

- **Reglas de asociación cuantitativas:**
  - Asociaciones entre items o atributos cuantitativos (continuos).

$$\text{Temperatura E } [38, 42] \text{ y Humedad E } [25, 33] \Rightarrow \text{Ozono Troposférico E } [140,206]$$

Definiciones

- **Regla de Asociación**
  - Implicación de la forma A → C , donde A e C son itemsets
  - Ejemplo: {Leche, Huevos} → {Pan}
- **Métricas de evaluación**
  - **Soporte:** Fracción de ejemplos que contienen tanto A como C:

$$Sup(A \rightarrow C) = P(A \wedge C) = \frac{\#(A \wedge C)}{N}$$

  - **Confianza:** Frecuencia de casos en los que aparece A y C con respecto al total de casos que incluyen A:

$$Conf(A \rightarrow C) = P(A \mid C) = \frac{Sup(A \rightarrow C)}{Sup(A)}$$

  - **Lift:** Indica cuándo una regla es mejor prediciendo el resultado que asumiendo el resultado de forma aleatoria. Si el resultado es mayor que uno, la regla es significativa:

$$Lift(A \rightarrow C) = \frac{Conf(A \rightarrow C)}{Sup(C)}$$

Cliente	Compras
1	Pan, leche, vino, huevos
2	Pan, pañales, cerveza
3	Leche, pañales, vino, cerveza
4	Pan, leche, pañales, cerveza
5	Pan, leche, pañales

Ejemplo: {cerveza} -> {pañales}
Sup({cerveza}) = 3/5
Sup({pañales}) = 4/5

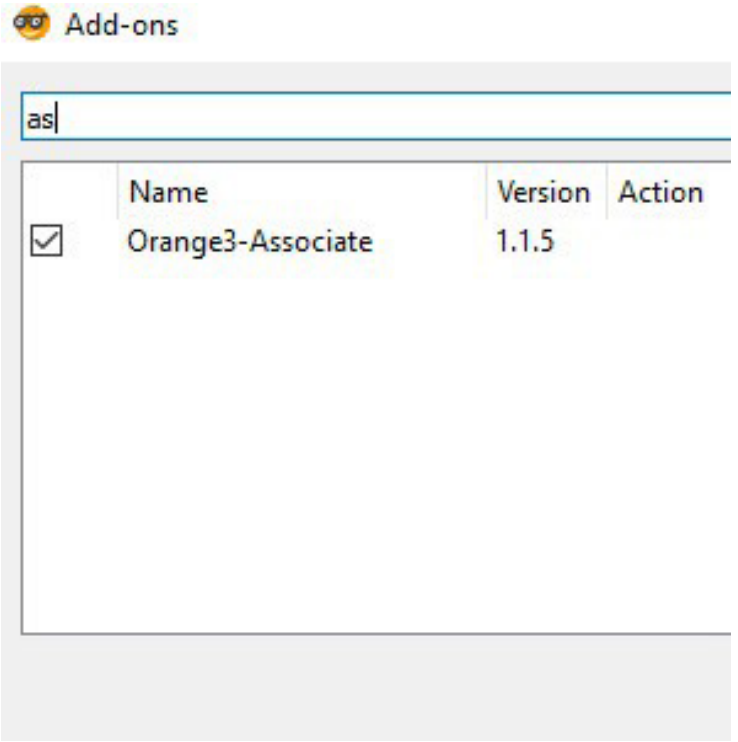
Sup({cerveza}→{pañales}) = Sup({cerveza, pañales}) = 3/5

Conf ({cerveza}→{pañales}) = Sup({cerveza, pañales}) / Sup({cerveza})
= (3/5) / (3/5) = 1

Lift ({cerveza}→{pañales}) = Conf({cerveza}→{pañales}) / Sup({pañales})
= 1 / (4/5) = 5/4 = 1,25

Reglas de asociación en Orange

- El add-on “Associate” contiene los métodos orientados a buscar asociaciones entre datos.
- Es importante reseñar que estos métodos solo funcionan con datos cualitativos.
- Es un add-on sencillo y simple de usar, con pocas opciones, basta con seleccionar un método, configurarlo y probarlo.



Reglas de asociación en Orange

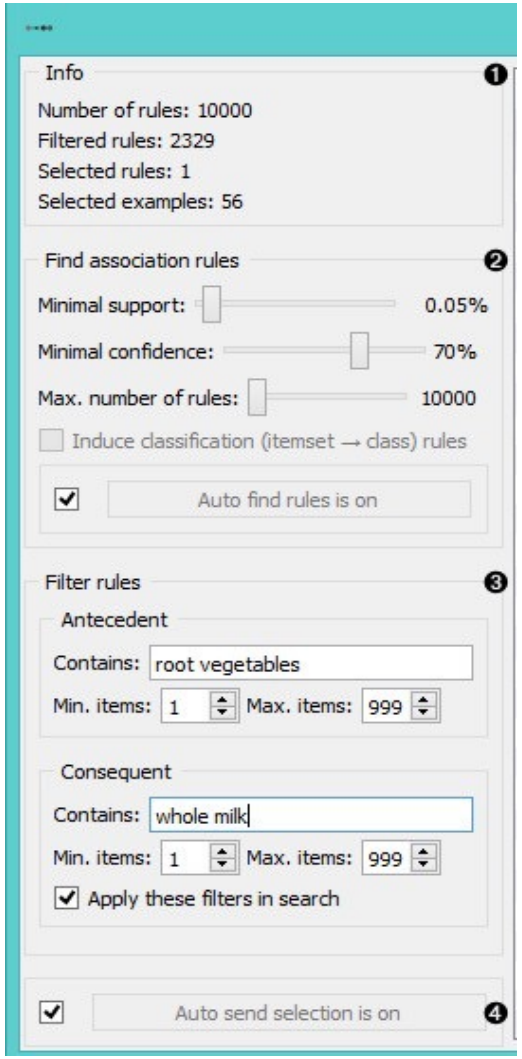
FPGrowth → Usado por Orange

- Presenta un buen rendimiento utilizando la estructura FP-Tree (árbol de patrones frecuentes).
- Haciendo uso de dicha estructura de almacenamiento se reduce el coste de computación del proceso de obtención de reglas de asociación.
- No es necesario generar los conjuntos de ítems candidatos a frecuentes y tampoco comprobar si superan un determinado umbral.
- Reduce iterativamente el soporte mínimo hasta que encuentra el número de reglas requerido con un valor mínimo para la métrica indicada.

Reglas de asociación en Orange

Nodo Association Rules

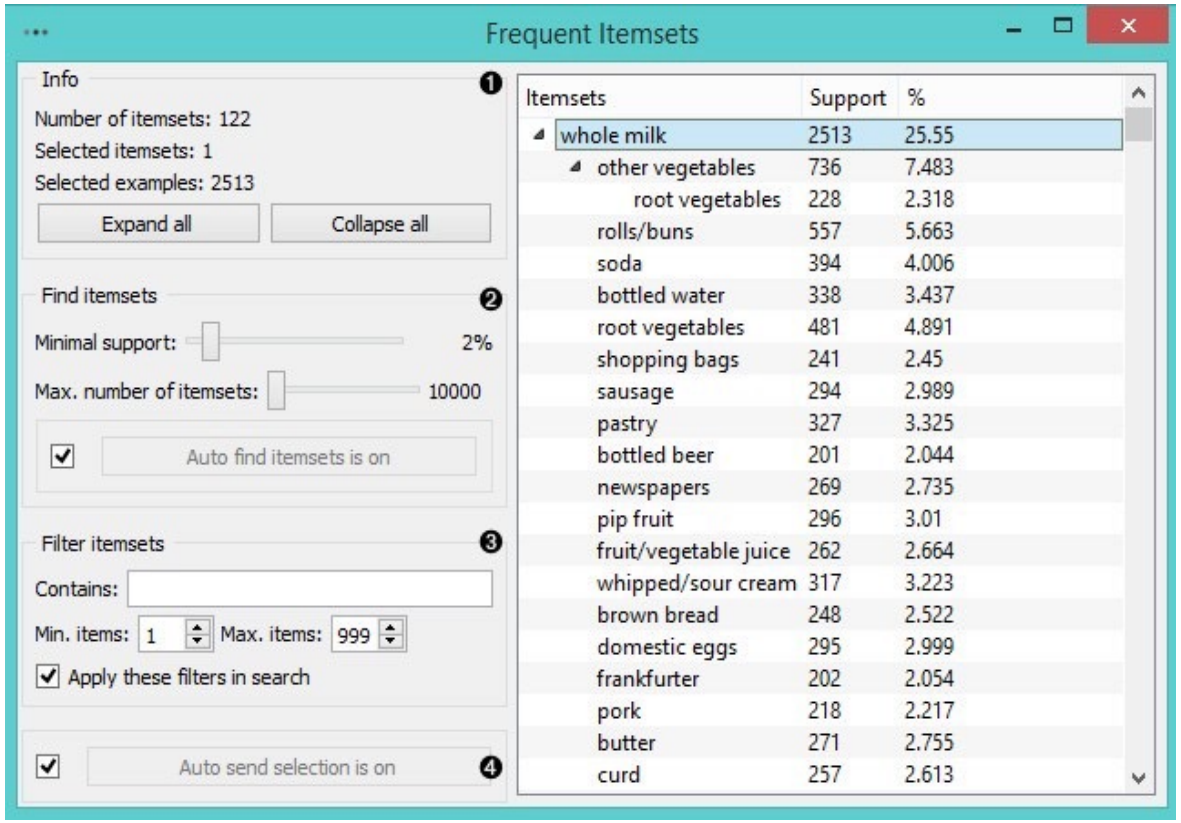
- Find association rules: se puede indicar el criterio a seguir para inducir las reglas:
  - Soporte mínimo
  - Confianza mínima
  - Máx número de reglas
- Filter rules: las reglas generadas se pueden filtrar por varios criterios. Por ejemplo, el antecedente o el consecuente



Reglas de asociación en Orange

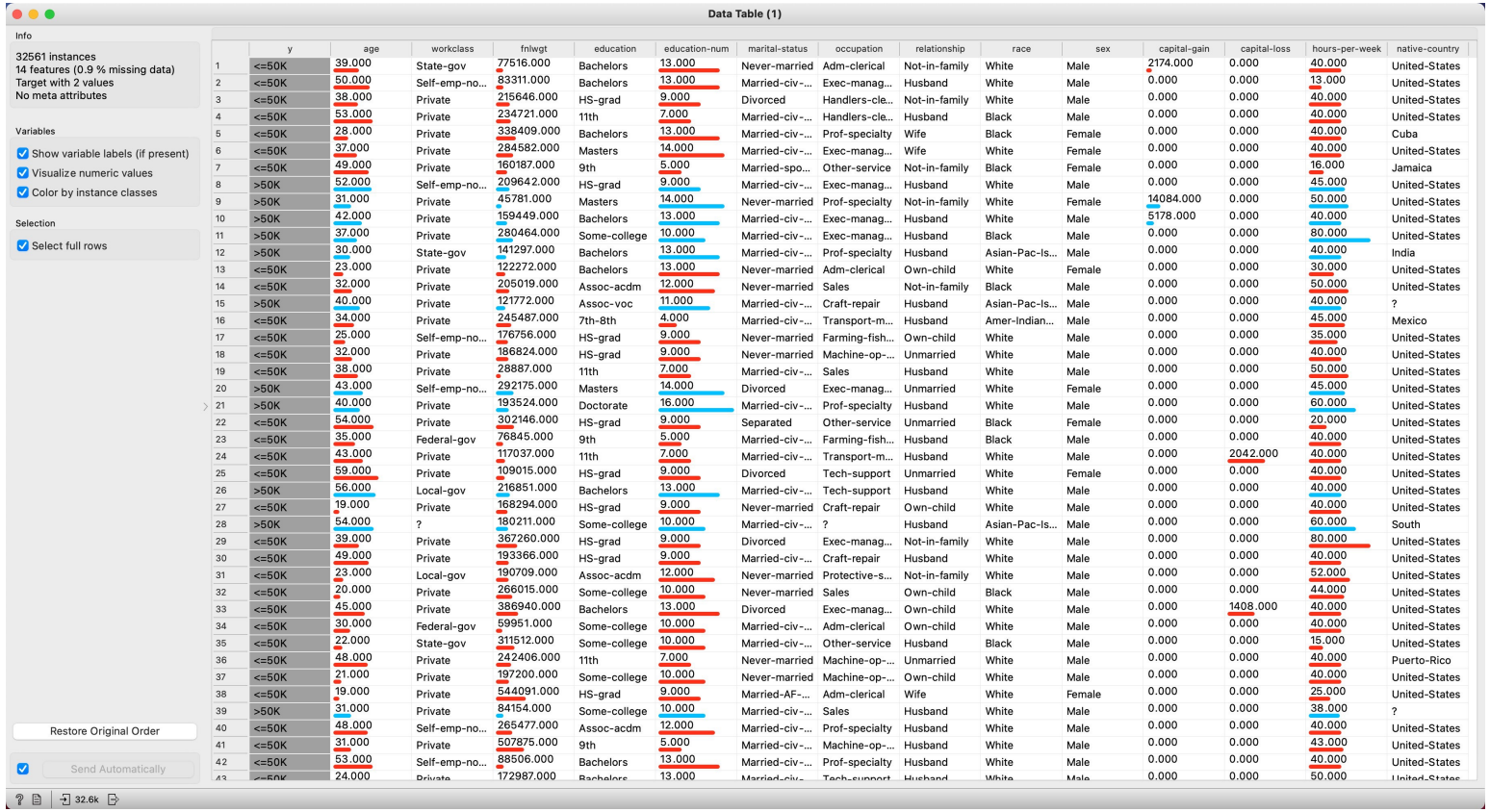
Nodo Frequent Itemsets

- Este widget encuentra itemsets frecuentes según unos parámetros a elegir:
  - Soporte mínimo
  - Máximo número de itemsets
  - Filtro



Actividad – Reglas asociativas en salarios

- Utilizaremos el conjunto de datos **Adult** que viene integrado en Orange

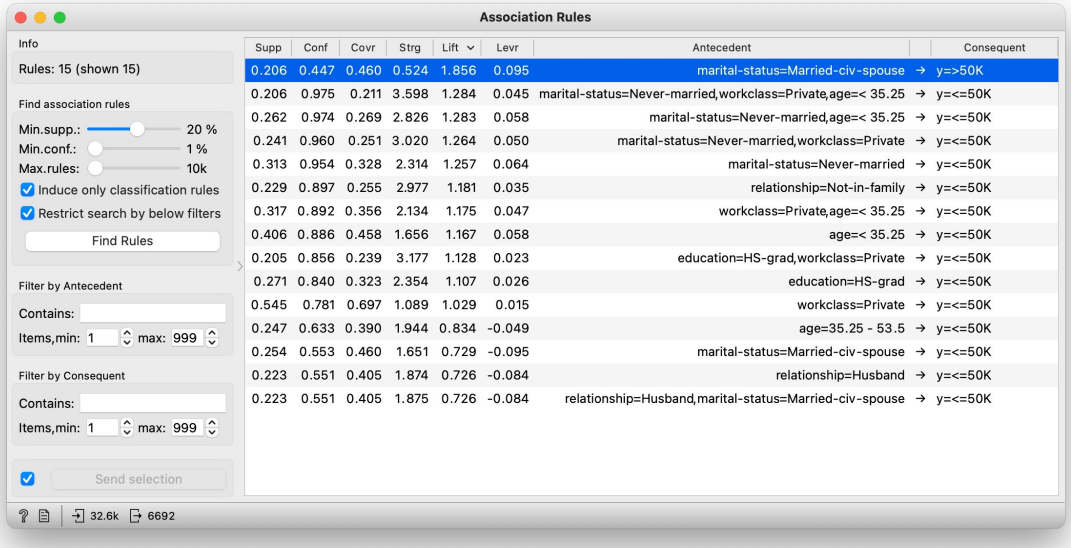


	y	age	workclass	fnlwgt	education	education-num	marital-status	occupation	relationship	race	sex	capital-gain	capital-loss	hours-per-week	native-country
1	<=50K	39.000	State-gov	77516.000	Bachelors	13.000	Never-married	Adm-clerical	Not-in-family	White	Male	2174.000	0.000	40.000	United-States
2	<=50K	50.000	Self-emp-no...	83311.000	Bachelors	13.000	Married-civ-...	Exec-manag...	Husband	White	Male	0.000	0.000	13.000	United-States
3	<=50K	38.000	Private	215646.000	HS-grad	9.000	Divorced	Handlers-cle...	Not-in-family	White	Male	0.000	0.000	40.000	United-States
4	<=50K	53.000	Private	234721.000	11th	7.000	Married-civ-...	Handlers-cle...	Husband	Black	Male	0.000	0.000	40.000	United-States
5	<=50K	28.000	Private	338409.000	Bachelors	13.000	Married-civ-...	Prof-specialty	Wife	Black	Female	0.000	0.000	40.000	Cuba
6	<=50K	37.000	Private	284582.000	Masters	14.000	Married-civ-...	Exec-manag...	Wife	White	Female	0.000	0.000	40.000	United-States
7	<=50K	49.000	Private	160187.000	9th	5.000	Married-spo...	Other-service	Not-in-family	Black	Female	0.000	0.000	16.000	Jamaica
8	>50K	52.000	Self-emp-no...	209642.000	HS-grad	9.000	Married-civ-...	Exec-manag...	Husband	White	Male	0.000	0.000	45.000	United-States
9	>50K	31.000	Private	45781.000	Masters	14.000	Never-married	Prof-specialty	Not-in-family	White	Female	14084.000	0.000	50.000	United-States
10	>50K	42.000	Private	159449.000	Bachelors	13.000	Married-civ-...	Exec-manag...	Husband	White	Male	5178.000	0.000	40.000	United-States
11	>50K	37.000	Private	280464.000	Some-college	10.000	Married-civ-...	Exec-manag...	Husband	Black	Male	0.000	0.000	80.000	United-States
12	>50K	30.000	State-gov	141297.000	Bachelors	13.000	Married-civ-...	Prof-specialty	Husband	Asian-Pac-Is...	Male	0.000	0.000	40.000	India
13	<=50K	23.000	Private	122272.000	Bachelors	13.000	Never-married	Adm-clerical	Own-child	White	Female	0.000	0.000	30.000	United-States
14	<=50K	32.000	Private	205019.000	Assoc-acdm	12.000	Never-married	Sales	Not-in-family	Black	Male	0.000	0.000	50.000	United-States
15	>50K	40.000	Private	121772.000	Assoc-voc	11.000	Married-civ-...	Craft-repair	Husband	Asian-Pac-Is...	Male	0.000	0.000	40.000	?
16	<=50K	34.000	Private	245487.000	7th-8th	4.000	Married-civ-...	Transport-m...	Husband	Amer-Indian...	Male	0.000	0.000	45.000	Mexico
17	<=50K	25.000	Self-emp-no...	176756.000	HS-grad	9.000	Never-married	Farming-fish...	Own-child	White	Male	0.000	0.000	35.000	United-States
18	<=50K	32.000	Private	186824.000	HS-grad	9.000	Never-married	Machine-op-...	Unmarried	White	Male	0.000	0.000	40.000	United-States
19	<=50K	38.000	Private	28887.000	11th	7.000	Married-civ-...	Sales	Husband	White	Male	0.000	0.000	50.000	United-States
20	>50K	43.000	Self-emp-no...	292175.000	Masters	14.000	Divorced	Exec-manag...	Unmarried	White	Female	0.000	0.000	45.000	United-States
21	>50K	40.000	Private	193524.000	Doctorate	16.000	Married-civ-...	Prof-specialty	Husband	White	Male	0.000	0.000	60.000	United-States
22	<=50K	54.000	Private	302146.000	HS-grad	9.000	Separated	Other-service	Unmarried	Black	Female	0.000	0.000	20.000	United-States
23	<=50K	35.000	Federal-gov	76845.000	9th	5.000	Married-civ-...	Farming-fish...	Husband	Black	Male	0.000	0.000	40.000	United-States
24	<=50K	43.000	Private	117037.000	11th	7.000	Married-civ-...	Transport-m...	Husband	White	Male	0.000	2042.000	40.000	United-States
25	<=50K	59.000	Private	109015.000	HS-grad	9.000	Divorced	Tech-support	Unmarried	White	Female	0.000	0.000	40.000	United-States
26	>50K	56.000	Local-gov	216851.000	Bachelors	13.000	Married-civ-...	Tech-support	Husband	White	Male	0.000	0.000	40.000	United-States
27	<=50K	19.000	Private	168294.000	HS-grad	9.000	Never-married	Craft-repair	Own-child	White	Male	0.000	0.000	40.000	United-States
28	>50K	54.000	?	180211.000	Some-college	10.000	Married-civ-...	?	Husband	Asian-Pac-Is...	Male	0.000	0.000	60.000	South
29	<=50K	39.000	Private	367260.000	HS-grad	9.000	Divorced	Exec-manag...	Not-in-family	White	Male	0.000	0.000	80.000	United-States
30	<=50K	49.000	Private	193366.000	HS-grad	9.000	Married-civ-...	Craft-repair	Husband	White	Male	0.000	0.000	40.000	United-States
31	<=50K	23.000	Local-gov	190709.000	Assoc-acdm	12.000	Never-married	Protective-s...	Not-in-family	White	Male	0.000	0.000	52.000	United-States
32	<=50K	20.000	Private	266015.000	Some-college	10.000	Never-married	Sales	Own-child	Black	Male	0.000	0.000	44.000	United-States
33	<=50K	45.000	Private	386940.000	Bachelors	13.000	Divorced	Exec-manag...	Own-child	White	Male	0.000	1408.000	40.000	United-States
34	<=50K	30.000	Federal-gov	59951.000	Some-college	10.000	Married-civ-...	Adm-clerical	Own-child	White	Male	0.000	0.000	40.000	United-States
35	<=50K	22.000	State-gov	311512.000	Some-college	10.000	Married-civ-...	Other-service	Husband	Black	Male	0.000	0.000	15.000	United-States
36	<=50K	48.000	Private	242406.000	11th	7.000	Never-married	Machine-op-...	Unmarried	White	Male	0.000	0.000	40.000	Puerto-Rico
37	<=50K	21.000	Private	197200.000	Some-college	10.000	Never-married	Machine-op-...	Own-child	White	Male	0.000	0.000	40.000	United-States
38	<=50K	19.000	Private	544091.000	HS-grad	9.000	Married-AF-...	Adm-clerical	Wife	White	Female	0.000	0.000	25.000	United-States
39	>50K	31.000	Private	84154.000	Some-college	10.000	Married-civ-...	Sales	Husband	White	Male	0.000	0.000	38.000	?
40	<=50K	48.000	Self-emp-no...	265477.000	Assoc-acdm	12.000	Married-civ-...	Prof-specialty	Husband	White	Male	0.000	0.000	40.000	United-States
41	<=50K	31.000	Private	507875.000	9th	5.000	Married-civ-...	Machine-op-...	Husband	White	Male	0.000	0.000	43.000	United-States
42	<=50K	53.000	Self-emp-no...	88506.000	Bachelors	13.000	Married-civ-...	Prof-specialty	Husband	White	Male	0.000	0.000	40.000	United-States
43	<=50K	24.000	Private	172987.000	Bachelors	13.000	Married-civ-...	Tech-support	Husband	White	Male	0.000	0.000	50.000	United-States

Actividad – Reglas asociativas en salarios

- Veamos con el nodo **Frequent Itemsets** cuáles son las reglas más comunes:
  1. Si pulsamos sobre Collapse All, sólo se mostrarán los atributos/valores "TOP".
  2. Si se expande cada uno de los ítems, se pueden observar cuáles son los atributos/valores conectados a él.
  3. Intente extraer alguna regla de asociación a partir del primer ítem.

- Para poder extraer las reglas de asociación usaremos el nodo **Association Rules**
  - Observemos la medida **lift**: Si el valor de lift es 1, entonces el antecedente y el consecuente son independientes. Cuanto más alto sea este valor, mayor será la probabilidad de que la existencia del antecedente y el consecuente juntos en una instancia no es sólo una ocurrencia aleatoria, sino debido a una cierta relación entre ellos.



Supp	Conf	Covr	Strg	Lift	Levr	Antecedent		Consequent
0.206	0.447	0.460	0.524	1.856	0.095	marital-status=Married-civ-spouse	→	y=>50K
0.206	0.975	0.211	3.598	1.284	0.045	marital-status=Never-married,workclass=Private,age=< 35.25	→	y=<=50K
0.262	0.974	0.269	2.826	1.283	0.058	marital-status=Never-married,age=< 35.25	→	y=<=50K
0.241	0.960	0.251	3.020	1.264	0.050	marital-status=Never-married,workclass=Private	→	y=<=50K
0.313	0.954	0.328	2.314	1.257	0.064	marital-status=Never-married	→	y=<=50K
0.229	0.897	0.255	2.977	1.181	0.035	relationship=Not-in-family	→	y=<=50K
0.317	0.892	0.356	2.134	1.175	0.047	workclass=Private,age=< 35.25	→	y=<=50K
0.406	0.886	0.458	1.656	1.167	0.058	age=< 35.25	→	y=<=50K
0.205	0.856	0.239	3.177	1.128	0.023	education=HS-grad,workclass=Private	→	y=<=50K
0.271	0.840	0.323	2.354	1.107	0.026	education=HS-grad	→	y=<=50K
0.545	0.781	0.697	1.089	1.029	0.015	workclass=Private	→	y=<=50K
0.247	0.633	0.390	1.944	0.834	-0.049	age=35.25 - 53.5	→	y=<=50K
0.254	0.553	0.460	1.651	0.729	-0.095	marital-status=Married-civ-spouse	→	y=<=50K
0.223	0.551	0.405	1.874	0.726	-0.084	relationship=Husband	→	y=<=50K
0.223	0.551	0.405	1.875	0.726	-0.084	relationship=Husband,marital-status=Married-civ-spouse	→	y=<=50K