ESTUDIO DE LAS REGLAS DE ASOCIACIÓN EN PRODUCTOS DE SUPERMERCADOS Y PLANTILLAS DE EMPLEADOS

MINERÍA DE DATOS EN NEGOCIOS

Ágatha del Olmo Tirado | 2ºBIA | 07/12/2023

VNIVERSITAT DE VALÈNCIA

INTELIGENCIA Y ANALÍTICA DE NEGOCIOS

ÍNDICE

INTRODUCCIÓN

Este informe presenta un análisis detallado de las reglas de inducción basado en datos del entorno empresarial y de consumo recopilados en los archivos "empleados.arff" y “supermercado.arff”. El objetivo de este estudio es emplear técnicas de asociación para encontrar patrones interesantes en ambos contextos. Durante todo el estudio hemos empleado la técnica de “apriori” en el entorno de Explorer en Weka, tanto con sus métricas predeterminadas como con ciertas variaciones según nos interesaba.

Para el conjunto de datos "empleados.arff", se pretende descubrir reglas de asociación que permitan entender las relaciones entre características como el sueldo, el estado civil, la posesión de coche, la cantidad de hijos, la forma de vivienda, etc al que pertenece cada empleado. El objetivo es obtener información que pueda ser útil en la toma de decisiones de la gestión de recursos humanos.

En el caso de "supermercado.arff", se busca identificar patrones de compra asociados a la variable "total" (monto total de la compra) para entender qué productos tienden a comprarse juntos y cómo estas asociaciones pueden ser utilizadas para mejorar estrategias de marketing o disposición de productos en el establecimiento.

GLOSARIO DE CONCEPTOS

Para una mejor comprensión de las métricas comentadas en el estudio, hemos realizado un glosario de conceptos para que todo lector pueda seguirlo.

Algoritmo a priori

Algoritmo que reduce iterativamente la cobertura mínima hasta que encuentra el número de reglas indicadas con la confianza mínima especificada. Tiene la opción de extraer reglas con asociación de clase (con una clase como predictor de la regla).

Cobertura

Porcentaje de veces que se dan el antecedente y el consecuente a la vez en la base de datos.
*Cobertura (X->Y) = P(XnY) = P(X) * P(Y/X) = P(Y) * p(X/Y)*

Confianza

Porcentaje de instancias en las que el antecedente ha derivado en el consecuente.
Conf (X->Y) = P(Y/X)

- Alta Confianza: Valores superiores al 80-90% se consideran generalmente altos.

- Confianza Media: Valores en el rango del 60-80% pueden considerarse medianos.

- Baja Confianza: Valores por debajo del 60% se consideran bajos

Confianza

Porcentaje de instancias en las que el antecedente ha derivado en el consecuente.

Conf (X->Y) = P(Y/X)

- Alta Confianza: Valores superiores al 80-90% se consideran generalmente altos.

- Confianza Media: Valores en el rango del 60-80% pueden considerarse medianos.

- Baja Confianza: Valores por debajo del 60% se consideran bajos

Lift

El levantamiento mide el nivel de interés.
Lift (X->Y) = P(Y/X) / P(Y)

- Asociación positiva: Si lift>1 indica una dependencia positiva. Cuanto mayor sea el levantamiento, más fuerte es la asociación directa.

- Asociación nula: Si lift=1, dependencia nula. Cuando una regla indica una asociación nula esto implica que la regla se puede descartarl

- Asociación negativa: Si lift < 1 indica una dependencia negativa. Cuanto menor sea el levantamiento, más fuerte es la asociación inversa.

Levarage

Mide la correlación entre antecedente y consecuente comparando la cobertura de estos bajo la suposición de independencia y la cobertura real de la base de datos.
*Levarage (X->Y) = P (XuY) - P(X) * P(Y)*

- Ocurrencia conjunta alta: si Lev > 0, X e Y están positivamente correlacionados, la cobertura observada es mayor a la cobertura esperada.

- Ocurrencia conjunta nula: si Lev=0, X e Y son incorrelacionados.

- Ocurrencia conjunta baja: si Lev < 0, X e Y están negativamente correlacionados, tienden a no darse en la misma regla.

Convicción

Cuantifica qué tan independientes son antecedente y consecuente. Esta medida es muy útil para identificar reglas de asociación significativas.
Convicción (X->Y) = (P(X).P($\bar{Y}$)) / P(Xn$\bar{Y}$)

- Dependencia: Cuanto mayor sea el valor de Conv, más dependiente es el consecuente del antecedente (en sentido negativo -inverso- o positivo -directo-).

- Independencia: Conv = 1 implica independencia.

INTERPRETACIÓN DE LAS REGLAS

Para empezar, vamos a realizar una interpretación exhaustiva de las métricas y sentido de cada una de las diez reglas en ambas bases de datos con el algoritmo "Apriori" default. Cabe mencionar que la métrica de convicción se ha interpretado en relación con el resto de reglas de cada base de datos.

BASE DE DATOS DE EMPLEADOS

1. Alq/Prop=Prop 132 ==> Coche=Sí 132 <conf:(1)> lift:(1.31) lev:(0.1) [31] conv:(31.13)

- **Confianza** (Conf): 100%. Indica que, en el 100% de los individuos que tienen propiedad, tienen también coche.
- **Lift**: 1.31. Indica una asociación positiva en la que la ocurrencia de tener coche es 1.31 veces más probable cuando se tiene una propiedad en comparación con la ocurrencia general de tener coche.
- **Cobertura**: 132/318=0.42%. Indica que esta regla se encuentra en el 0.42% de los casos de la base de datos.
- **Convicción**: 31.13. Indica una dependencia bastante alta entre antecedente y consecuente.

2. Alq/Prop=Alquiler Bajas/Año=ninguna 119 ==> Casado=No 119 <conf:(1)> lift:(1.81) lev:(0.17) [53] conv:(53.14)

- **Confianza** (Conf): 100%. Indica que, en el 100% de los individuos que tienen alquiler y no tienen ninguna baja laboral al año, no están casados.
- **Lift**: 1.81. Indica que la ocurrencia de no estar casado es 1.81 veces más probable cuando no se tiene ninguna baja laboral al año y se tiene un alquiler en comparación con la ocurrencia general de no estar casado.
- **Cobertura**: 53/318=0.37%. Indica que esta regla se encuentra en el 0.37% de los casos de la base de datos.
- **Convicción**: 53.14. Indica una dependencia muy alta entre antecedente y consecuente.

3. Sueldo ='(-inf-12500]' 111 ==> Sexo=H 111 <conf:(1)> lift:(1.71) lev:(0.14) [46] conv:(46.08)

- **Confianza** (Conf): 100%. Indica que, en el 100% de los individuos que tienen un sueldo menor que 12500€, son hombres.
- **Lift**: 1.71. Indica que la ocurrencia ser hombre es 1.71 veces más probable si se tiene un sueldo menor que 12500€ en comparación con la ocurrencia general de ser hombre.
- **Cobertura**: 111/318=0.35%. Indica que esta regla se encuentra en el 0.35% de los casos de la base de datos.
- **Convicción**: 46.08. Indica una dependencia bastante alta entre antecedente y consecuente.

4. Casado=Sí Alq/Prop=Prop 104 ==> Coche=Sí 104 <conf:(1)> lift:(1.31) lev:(0.08) [24] conv:(24.53)

- **Confianza** (Conf): 100%. Indica que, en el 100% de los individuos que están casados y tienen propiedad, tienen coche.
- **Lift**: 1.31. Indica que la ocurrencia de tener coche es 1.31 veces más probable cuando se está casado y se tiene una propiedad en comparación con la ocurrencia general de tener coche.
- **Cobertura**: 104/318=0.33%. Indica que esta regla se encuentra en el 0.33% de los casos de la base de datos.
- **Convicción**: 24.53. Indica una dependencia no muy alta entre antecedente y consecuente.

5. Alq/Prop=Prop Sexo=H 104 ==> Casado=Sí 104 <conf:(1)> lift:(2.24) lev:(0.18) [57] conv:(57.56)

- **Confianza** (Conf): 100%. Indica que, en el 100% de los individuos que son hombres y tienen propiedad, están casados.

- **Lift**: 0.18. Indica que la ocurrencia de estar casado es 0.18 veces más probable cuando se es hombre y se tiene una propiedad en comparación con la ocurrencia general de estar casado.
- **Cobertura**: 104/318=0.33%. Indica que esta regla se encuentra en el 0.33% de los casos de la base de datos.
- **Convicción**: 57.56. Indica una dependencia muy alta entre antecedente y consecuente.

6. Casado=Sí Alq/Prop=Prop 104 ==> Sexo=H 104 <conf:(1)> lift:(1.71) lev:(0.14) [43] conv:(43.17)

- **Confianza** (Conf): 100%. Indica que, en el 100% de los individuos que están casados y tienen propiedad, son hombres.
- **Lift**: 1.71. Indica que la ocurrencia de ser hombre es 1.71 veces más probable cuando se está casado y se tiene una propiedad en comparación con la ocurrencia general de ser hombre.
- **Cobertura**: 104/318=0.33%. Indica que esta regla se encuentra en el 0.33% de los casos de la base de datos.
- **Convicción**: 43.17. Indica una dependencia bastante alta entre antecedente y consecuente.

7. Alq/Prop=Prop Sexo=H 104 ==> Coche=Sí 104 <conf:(1)> lift:(1.31) lev:(0.08) [24] conv:(24.53)

- **Confianza** (Conf): 100%. Indica que, en el 100% de los individuos que son hombres y tienen propiedad, tienen coche.
- **Lift**: 1.31. Indica que la ocurrencia de tener coche es 1.31 veces más probable cuando se es hombre y se tiene una propiedad en comparación con la ocurrencia general de tener coche.
- **Cobertura**: 104/318=0.33%. Indica que esta regla se encuentra en el 0.33% de los casos de la base de datos.
- **Convicción**: 24.53. Indica una dependencia no muy alta entre antecedente y consecuente.

8. Coche=Sí Alq/Prop=Prop Sexo=H 104 ==> Casado=Sí 104 <conf:(1)> lift:(2.24) lev:(0.18) [57] conv:(57.56)

- **Confianza** (Conf): 100%. Indica que, en el 100% de los individuos que están tienen coche y tienen propiedad y son hombres, están casados.
- **Lift**: 2.24. Indica que la ocurrencia de estar casado es 2.24 veces más probable cuando se es hombre, se tiene coche y se tiene una propiedad en comparación con la

ocurrencia general de estar casado.
- **Cobertura**: 104/318=0.33%. Indica que esta regla se encuentra en el 0.33% de los casos de la base de datos.
- **Convicción**: 57.56. Indica una dependencia muy alta entre antecedente y consecuente.

9. Casado=Sí Alq/Prop=Prop Sexo=H 104 ==> Coche=Sí 104 <conf:(1)> lift:(1.31) lev:(0.08) [24] conv:(24.53)

- **Confianza** (Conf): 100%. Indica que, en el 100% de los individuos que están casados, tienen propiedad y son hombres, tienen coche.
- **Lift**: 1.31. Indica que la ocurrencia de tener coche es 1.31 veces más probable cuando se está casado, se es hombre y se tiene una propiedad en comparación con la ocurrencia general de tener coche.
- **Cobertura**: 104/318=0.33%. Indica que esta regla se encuentra en el 0.33% de los casos de la base de datos.
- **Convicción**: 24.53. Indica una dependencia no muy alta entre antecedente y consecuente.

10. Casado=Sí Coche=Sí Sexo=H 104 ==> Alq/Prop=Prop 104 <conf:(1)> lift:(2.41) lev:(0.19) [60] conv:(60.83)

- **Confianza** (Conf): 100%. Indica que, en el 100% de los individuos que están casados, son hombres y tienen coche, tienen propiedad.

- **Lift**: 2.41. Indica que la ocurrencia de tener propiedad es 2.41 veces más probable cuando se está casado, se es hombre y se tiene coche en comparación con la ocurrencia general de tener propiedad.

- **Cobertura**: 104/318=0.33%. Indica que esta regla se encuentra en el 0.33% de los casos de la base de datos.

- **Convicción**: 60.83. Indica una dependencia muy alta entre antecedente y consecuente, de hecho es la que más presenta entre estas diez.

BASE DE DATOS DE SUPERMERCADO

1. galletas=t congelados=t frutas=t total=alto 788 ==> panaderia=t 723 <conf:(0.92)> lift:(1.27) lev:(0.03) [155] conv:(3.35)

- **Confianza** (Conf): 92%. Indica que, en el 92% de las ocasiones en las que se compraron galletas, congelados, frutas, y el monto total de la compra fue alto, también se compró panadería.
- **Lift**: 1.27. Indica que la ocurrencia de panadería es 1.27 veces más probable cuando se compran galletas, congelados, frutas, y el monto total de la compra es alto en comparación con la ocurrencia general de panadería.
- **Cobertura**: 723/4627=0.16%. Indica que esta regla se encuentra en el 0.16% de los casos de la base de datos.
- **Convicción**: 3.35. Indica una dependencia media entre antecedente y consecuente.

2. utiles cocina=t galletas=t frutas=t total=alto 760 ==> panaderia=t 696 <conf:(0.92)> lift:(1.27) lev:(0.03) [149] conv:(3.28)

- **Confianza** (Conf): 92%. Indica que, en el 92% de las ocasiones en las que se compraron galletas, útiles de cocina, frutas, y el monto total de la compra fue alto, también se compró panadería.
- **Lift**: 1.27. Indica que la ocurrencia de panadería es 1.27 veces más probable cuando se compran galletas, útiles de cocina, frutas, y el monto total de la compra es alto en comparación con la ocurrencia general de panadería.
- **Cobertura**: 696/4627=0.15%. Indica que esta regla se encuentra en el 0.15% de los casos de la base de datos.
- **Convicción**: 3.28. Indica una dependencia media entre antecedente y consecuente.

3. utiles cocina=t congelados=t frutas=t total=alto 770 ==> panaderia=t 705 <conf:(0.92)> lift:(1.27) lev:(0.03) [150] conv:(3.27)

- **Confianza** (Conf): 92%. Indica que, en el 92% de las ocasiones en las que se compraron útiles de cocina, congelados, frutas, y el monto total de la compra fue alto, también se compró panadería.
- **Lift**: 1.27. Indica que la ocurrencia de panadería es 1.27 veces más probable cuando se compran útiles de cocina, congelados, frutas, y el monto total de la compra es alto en comparación con la ocurrencia general de panadería.
- **Cobertura**: 705/4627=0.15%. Indica que esta regla se encuentra en el 0.15% de los casos de la base de datos.
- **Convicción**: 3.27. Indica una dependencia media entre antecedente y consecuente.

4. galletas=t frutas=t verdura=t total=alto 815 ==> panaderia=t 746 <conf:(0.92)> lift:(1.27) lev:(0.03) [159] conv:(3.26)

- **Confianza** (Conf): 92%. Indica que, en el 92% de las ocasiones en las que se compraron galletas, frutas, verduras, y el monto total de la compra fue alto, también se compró panadería.
- **Lift**: 1.27. Indica que la ocurrencia de panadería es 1.27 veces más probable cuando se compran galletas, frutas, verduras, y el monto total de la compra es alto en comparación con la ocurrencia general de panadería.
- **Cobertura**: 746/4627=0.16%. Indica que esta regla se encuentra en el 0.16% de los casos de la base de datos.
- **Convicción**: 3.26. Indica una dependencia media entre antecedente y consecuente.

5. snacks=t frutas=t total=alto 854 ==> panaderia=t 779 <conf:(0.91)> lift:(1.27) lev:(0.04) [164] conv:(3.15)

- **Confianza** (Conf): 91%. Indica que, en el 91% de las ocasiones en las que se compraron snacks, frutas, y el monto total de la compra fue alto, también se compró panadería.
- **Lift**: 1.27. Indica que la ocurrencia de panadería es 1.27 veces más probable cuando se compran snacks, frutas, y el monto total de la compra es alto en comparación con la ocurrencia general de panadería.
- **Cobertura**: 779/4627=0.17%. Indica que esta regla se encuentra en el 0.17% de los casos de la base de datos.
- **Convicción**: 3.15. Indica una dependencia ligeramente inferior a la media entre antecedente y consecuente.

6. galletas=t congelados=t verdura=t total=alto 797 ==> panaderia=t 725 <conf:(0.91)> lift:(1.26) lev:(0.03) [151] conv:(3.06)

- **Confianza** (Conf): 91%. Indica que, en el 91% de las ocasiones en las que se compraron galletas, congelados, verdura, y el monto total de la compra fue alto, también se compró panadería.
- **Lift**: 1.26. Indica que la ocurrencia de panadería es 1.26 veces más probable cuando se compran galletas, congelados, verdura, y el monto total de la compra es alto en comparación con la ocurrencia general de panadería.
- **Cobertura**: 725/4627=0.16%. Indica que esta regla se encuentra en el 0.16% de los casos de la base de datos.
- **Convicción**: 3.06. Indica una dependencia ligeramente inferior a la media entre antecedente y consecuente.

7. utiles cocina=t galletas=t verdura=t total=alto 772 ==> panaderia=t 701 <conf:(0.91)> lift:(1.26) lev:(0.03) [145] conv:(3.01)

- **Confianza** (Conf): 91%. Indica que, en el 91% de las ocasiones en las que se compraron galletas, útiles de cocina, verdura, y el monto total de la compra fue alto,

también se compró panadería.
- **Lift**: 1.26. Indica que la ocurrencia de panadería es 1.26 veces más probable cuando se compran galletas, útiles de cocina, verdura, y el monto total de la compra es alto en comparación con la ocurrencia general de panadería.
- **Cobertura**: 701/4627=0.15%. Indica que esta regla se encuentra en el 0.15% de los casos de la base de datos.
- **Convicción**: 3.01. Indica una dependencia inferior a la media entre antecedente y consecuente.

8. galletas=t frutas=t total=alto 954 ==> panaderia=t 866 <conf:(0.91)> lift:(1.26) lev:(0.04) [179] conv:(3)

- **Confianza** (Conf): 91%. Indica que, en el 91% de las ocasiones en las que se compraron galletas, frutas, y el monto total de la compra fue alto, también se compró panadería.
- **Lift**: 1.26. Indica que la ocurrencia de panadería es 1.26 veces más probable cuando se compran galletas, frutas, y el monto total de la compra es alto en comparación con la ocurrencia general de panadería.
- **Cobertura**: 866/4627=0.19%. Indica que esta regla se encuentra en el 0.19% de los casos de la base de datos.
- **Convicción**: 3. Indica una dependencia inferior a la media entre antecedente y consecuente.

9. congelados=t frutas=t verdura=t total=alto 834 ==> panaderia=t 757 <conf:(0.91)> lift:(1.26) lev:(0.03) [156] conv:(3)

- **Confianza** (Conf): 91%. Indica que, en el 91% de las ocasiones en las que se compraron congelados, frutas y verduras, y el monto total de la compra fue alto, también se compró panadería.
- **Lift**: 1.26. Indica que la ocurrencia de panadería es 1.26 veces más probable cuando se compran congelados, frutas y verduras, y el monto total de la compra es alto en comparación con la ocurrencia general de panadería.
- **Cobertura**: 757/4627=0.16%. Indica que esta regla se encuentra en el 0.16% de los casos de la base de datos.
- **Convicción**: 3. Indica una dependencia inferior a la media entre antecedente y consecuente.

10. congelados=t frutas=t total=alto 969 ==> panaderia=t 877 <conf:(0.91)> lift:(1.26) lev:(0.04) [179] conv:(2.92)

- **Confianza** (Conf): 91%. Indica que, en el 91% de las ocasiones en las que se compraron congelados y frutas, y el monto total de la compra fue alto, también se compró panadería.

- **Lift**: 1.26. Indica que la ocurrencia de panadería es 1.26 veces más probable cuando se compran congelados y frutas, y el monto total de la compra es alto en comparación con la ocurrencia general de panadería.
- **Cobertura**: 877/4627=0.19%. Indica que esta regla se encuentra en el 0.19% de los casos de la base de datos.
- **Convicción**: 2.92. Indica una la dependencia más inferior entre antecedente y consecuente de entre estas diez reglas.

MONTO DE LA FACTURA COMO PREDICTOR

Si cambiamos a True car en 'show properties', es decir, si usamos clase de referencia como predictor, que por defecto será la última variable, que en este caso es "total", no se genera ninguna regla con nivel de confianza superior a 0.9. Sin embargo, si reducimos este umbral a 0.8, obtenemos las siguientes reglas:

1. utiles cocina=t galletas=t salsas=t congelados=t pañuelos papel=t 574 ==> total=alto 470 conf:(0.82)

2. panaderia=t galletas=t salsas=t congelados=t pañuelos papel=t 600 ==> total=alto 491 conf:(0.82)

3. panaderia=t utiles cocina=t salsas=t congelados=t pañuelos papel=t 620 ==> total=alto 506 conf:(0.82)

4. panaderia=t utiles cocina=t galletas=t salsas=t pañuelos papel=t 595 ==> total=alto 483 conf:(0.81)

5. panaderia=t galletas=t salsas=t pañuelos papel=t verdura=t 583 ==> total=alto 469 conf:(0.8)

6. panaderia=t salsas=t congelados=t pañuelos papel=t verdura=t 610 ==> total=alto 490 conf:(0.8)

La combinación de alimentos más frecuente que precede en un total de importe alto incluye panadería, útiles de cocina, salsas, congelado y pañuelos de papel.

La combinación de productos que se da con más frecuencia es la tercera con panadería, útiles de cocina, salsas, congelados, pañuelos de papel. Esta combinación de ítems tiene un total de 620 apariciones, y 506 de estas ha desembocado en una compra de importe alto. Esto representa una cobertura de 506/4627=0.11%, es decir, muy cercano a 0, por lo que podemos decir que no es demasiado relevante.

COBERTURA SUPERIOR AL 50%

Estableciendo una cobertura mínima del 0.5, es necesario reducir la confianza hasta 0.78 para poder obtener dos reglas:

1. nata=t 2939 ==> panaderia=t 2337 <conf:(0.8)> lift:(1.1) lev:(0.05) [221] conv:(1.37)

2. frutas=t 2962 ==> panaderia=t 2325 <conf:(0.78)> lift:(1.09) lev:(0.04) [193] conv:(1.3)

Inicialmente, se podría sugerir la disposición de la nata y las frutas junto a la panadería. No obstante, como sabemos que no es algo inusual, sino más bien al contrario, que se compre pan junto a cualquier combinación de productos, esta recomendación y cualquier otra que incluya panadería pierde relevancia.

BASE DE DATOS LIMITADA A UN MONTO TOTAL ELEVADO

Si seleccionamos únicamente los casos que suponen una alta factura con el comando *weka.filters.unsupervised.instance.RemoveWithValues -S 0.0 -C last -L 1*, y eliminamos el atributo "Total", que deja de tener sentidoal haber solo total elevado, las cuatro primeras reglas que obtenemos con el algoritmo "apriori" son las siguientes:

1. galletas=t congelados=t frutas=t 788 ==> panaderia=t 723 <conf:(0.92)> lift:(1.09) lev:(0.04) [59] conv:(1.89)

2. utiles cocina=t galletas=t frutas=t 760 ==> panaderia=t 696 <conf:(0.92)> lift:(1.09) lev:(0.03) [56] conv:(1.85)

3. utiles cocina=t congelados=t frutas=t 770 ==> panaderia=t 705 <conf:(0.92)> lift:(1.09) lev:(0.03) [56] conv:(1.85)

4. galletas=t frutas=t verdura=t 815 ==> panaderia=t 746 <conf:(0.92)> lift:(1.09) lev:(0.04) [60] conv:(1.84)

No podemos concluir que estos sean los productos más frecuentes entre las compras de elevado total, ya que en estas reglas se ve maximizada la confianza, que no tiene que ver con la totalidad de los casos, sino con el total de panadería. En cambio, sería más apropiado maximizar la cobertura, ya que nos mostraría, de la base de datos de monto total elevado, cuántas ocurrencias tienen las reglas, que es la definición de frecuencia.

PANADERÍA COMO CONSECUENTE

Si le pedimos al algoritmo que nos extraiga 30 reglas, obtenemos las siguientes:

1. galletas=t congelados=t snacks=t frutas=t verdura=t total=alto 510 ==> panaderia=t 478 <conf:(0.94)> lift:(1.3) lev:(0.02) [110] conv:(4.33)
2. galletas=t congelados=t quesos=t frutas=t total=alto 495 ==> panaderia=t 463 <conf:(0.94)> lift:(1.3) lev:(0.02) [106] conv:(4.2)
3. galletas=t quesos=t frutas=t verdura=t total=alto 513 ==> panaderia=t 479 <conf:(0.93)> lift:(1.3) lev:(0.02) [109] conv:(4.11)
4. utiles cocina=t galletas=t snacks=t frutas=t total=alto 557 ==> panaderia=t 520 <conf:(0.93)> lift:(1.3) lev:(0.03) [119] conv:(4.11)
5. utiles cocina=t quesos=t frutas=t verdura=t total=alto 519 ==> panaderia=t 483 <conf:(0.93)> lift:(1.29) lev:(0.02) [109] conv:(3.93)
6. congelados=t snacks=t pañuelos papel=t frutas=t total=alto 518 ==> panaderia=t 482 <conf:(0.93)> lift:(1.29) lev:(0.02) [109] conv:(3.92)
7. zumos=t galletas=t snacks=t frutas=t total=alto 529 ==> panaderia=t 492 <conf:(0.93)> lift:(1.29) lev:(0.02) [111] conv:(3.9)
8. galletas=t quesos=t frutas=t total=alto 584 ==> panaderia=t 543 <conf:(0.93)> lift:(1.29) lev:(0.03) [122] conv:(3.9)
9. galletas=t snacks=t frutas=t verdura=t total=alto 596 ==> panaderia=t 554 <conf:(0.93)> lift:(1.29) lev:(0.03) [125] conv:(3.89)
10. utiles cocina=t galletas=t congelados=t frutas=t verdura=t total=alto 561 ==> panaderia=t 521 <conf:(0.93)> lift:(1.29) lev:(0.03) [117] conv:(3.84)
11. galletas=t congelados=t snacks=t frutas=t total=alto 589 ==> panaderia=t 547 <conf:(0.93)> lift:(1.29) lev:(0.03) [123] conv:(3.84)
12. utiles cocina=t congelados=t snacks=t frutas=t total=alto 558 ==> panaderia=t 518 <conf:(0.93)> lift:(1.29) lev:(0.03) [116] conv:(3.81)
13. galletas=t snacks=t pañuelos papel=t frutas=t total=alto 515 ==> panaderia=t 478 <conf:(0.93)> lift:(1.29) lev:(0.02) [107] conv:(3.8)
14. utiles cocina=t quesos=t frutas=t total=alto 584 ==> panaderia=t 542 <conf:(0.93)> lift:(1.29) lev:(0.03) [121] conv:(3.81)
15. utiles cocina=t congelados=t pañuelos papel=t frutas=t verdura=t total=alto 513 ==> panaderia=t 476 <conf:(0.93)> lift:(1.29) lev:(0.02) [106] conv:(3.78)
16. galletas=t conservas verduras=t frutas=t total=alto 523 ==> panaderia=t 485 <conf:(0.93)> lift:(1.29) lev:(0.02) [108] conv:(3.76)
17. snacks=t quesos=t frutas=t total=alto 535 ==> panaderia=t 496 <conf:(0.93)> lift:(1.29) lev:(0.02) [110] conv:(3.75)
18. galletas=t nata=t margarina=t frutas=t total=alto 506 ==> panaderia=t 469 <conf:(0.93)> lift:(1.29) lev:(0.02) [104] conv:(3.73)
19. snacks=t pañuelos papel=t frutas=t verdura=t total=alto 530 ==> panaderia=t 491 <conf:(0.93)> lift:(1.29) lev:(0.02) [109] conv:(3.71)
20. congelados=t snacks=t nata=t frutas=t total=alto 528 ==> panaderia=t 489 <conf:(0.93)> lift:(1.29) lev:(0.02) [109] conv:(3.7)
21. utiles cocina=t congelados=t pañuelos papel=t frutas=t total=alto 581 ==> panaderia=t 538 <conf:(0.93)> lift:(1.29) lev:(0.03) [119] conv:(3.7)

22. galletas=t congelados=t pañuelos papel=t frutas=t verdura=t total=alto 513 ==> panaderia=t 475 <conf:(0.93)> lift:(1.29) lev:(0.02) [105] conv:(3.69)
23. utiles cocina=t congelados=t margarina=t frutas=t total=alto 553 ==> panaderia=t 512 <conf:(0.93)> lift:(1.29) lev:(0.02) [114] conv:(3.69)
24. congelados=t snacks=t frutas=t verdura=t total=alto 593 ==> panaderia=t 549 <conf:(0.93)> lift:(1.29) lev:(0.03) [122] conv:(3.69)
25. congelados=t quesos=t frutas=t total=alto 579 ==> panaderia=t 536 <conf:(0.93)> lift:(1.29) lev:(0.03) [119] conv:(3.69)
26. galletas=t congelados=t nata=t margarina=t total=alto 537 ==> panaderia=t 497 <conf:(0.93)> lift:(1.29) lev:(0.02) [110] conv:(3.67)
27. galletas=t quesos=t nata=t total=alto 548 ==> panaderia=t 507 <conf:(0.93)> lift:(1.29) lev:(0.02) [112] conv:(3.66)
28. conservas verduras=t congelados=t frutas=t total=alto 521 ==> panaderia=t 482 <conf:(0.93)> lift:(1.29) lev:(0.02) [107] conv:(3.65)
29. galletas=t nata=t margarina=t verdura=t total=alto 507 ==> panaderia=t 469 <conf:(0.93)> lift:(1.29) lev:(0.02) [104] conv:(3.64)
30. galletas=t congelados=t margarina=t frutas=t total=alto 560 ==> panaderia=t 518 <conf:(0.93)> lift:(1.29) lev:(0.02) [114] conv:(3.65)

Como podemos apreciar, el mínimo de confianza establecido es de 0.93 en la trigésima regla. Y es de interés comentar que en todas las reglas obtenidas, el consecuente es panadería. Esta repetición del consecuente puede deberse a la naturaleza del mismo, pues "independientemente" de qué tipo de compra se haga, siempre se compra pan, por lo tanto, estas reglas son poco prácticas en la realidad. Para evitar este fenómeno, podríamos simplemente reducir la confianza máxima, lo que nos podría generar reglas más variadas, ya que serían "menos evidentes".

EL SENTIDO DE LAS REGLAS: PROPIEDAD Y TENER COCHE

Cuando usamos el algoritmo "Apriori" con las opciones por defecto, observamos que la primera es la siguiente:

1. Alq/Prop=Prop 132 ==> Coche=Sí 132 <conf:(1)> lift:(1.31) lev:(0.1) [31] conv:(31.13)

Puede ser confuso el hecho de que, aunque veamos que la confianza sea del 100%, haya 243 empleados con coches de los cuales sólo 132 tienen la vivienda en propiedad (aproximadamente el 54%).

Esto sucede porque la dirección de la regla lo cambia todo, para entenderlo podemos mirar un ejemplo pequeño y sencillo.

Imaginamos que estos son todos los individuos de una pequeña base de datos:

1. Vivienda = Propiedad, Coche = Sí

2. Vivienda = Alquiler, Coche = Sí

3. Vivienda = Alquiler, Coche = No

4. Vivienda = Propiedad, Coche = Sí

5. Vivienda = Propiedad, Coche = Sí

Hay 4 empleados con coches, y de estos sólo 3 tienen la vivienda en propiedad (el 75%). En cambio, hay 3 empleados con vivienda en propiedad, y de estos todos tienen coche (100%). Es importante ver la dirección de la regla porque el denominador de la probabilidad condicionada cambia (siendo X=Vivienda e Y=Coche, el primer caso sería Conf (Y->X) = P(X/Y) y el segundo caso sería Conf (X->Y) = P(Y/X)).

EL SENTIDO DE LAS REGLAS: BAJO SUELDO Y SER HOMBRE

De entre las mismas reglas del anterior apartado, obtenemos esta:

3. Sueldo ='(-inf-12500]' 111 ==> Sexo=H 111 <conf:(1)> lift:(1.71) lev:(0.14) [46] conv:(46.08)

Esta regla en concreto significa que en nuestra base de datos, dado que el empleado tenga un bajo sueldo, este será hombre el 100% de las veces. De nuevo, puede crear cierta confusión el sentido de la regla, esto no quiere decir que si eres hombre tienes un bajo sueldo, sino precisamente al contrario, no todo hombre tiene un bajo sueldo sino que toda persona que tiene bajo sueldo es hombre.

EL SENTIDO DE LAS REGLAS EN UN PERFIL DEL ÁREA COMERCIAL

Volvemos a usar a opción car en “true”, y por defecto la variable de “Departamento” se usa como predictor de las reglas. Las reglas que afectan al departamento comercial de las que obtenemos son las siguientes:

1. Casado=No Sindic.=Sí 95 ==> departamento=comercial 95 conf:(1)

2. Coche=Sí Sexo=M 90 ==> departamento=comercial 90 conf:(1)

3. Casado=No Hijos=0 Sindic.=Sí 78 ==> departamento=comercial 78 conf:(1)

4. Casado=No Sindic.=Sí Bajas/Año=ninguna 78 ==> departamento=comercial 78 conf:(1)

5. Casado=No Coche=Sí Sindic.=Sí 77 ==> departamento=comercial 77 conf:(1)

El perfil no sería necesariamente un buen predictor del departamento comercial. Esto se da de nuevo por las mismas razones, la confianza es *P(Y/X)*, no al contrario, lo cual, como hemos visto, lo cambia todo.

REGLAS DE POCO INTERÉS

Por último, eliminamos los casos en los que el individuo no pertenezca al departamento comercial y luego eliminamos la propia variable de "Departamento" para que no afecte al estudio. Las reglas que obtenemos y que tienen como predictor ser hombre son las siguientes:

6. Casado=Sí 83 ==> Sexo=H 83 <conf:(1)> lift:(1) lev:(0) [0] conv:(0)

10. Coche=Sí 83 ==> Sexo=H 83 <conf:(1)> lift:(1) lev:(0) [0] conv:(0)

Estas reglas no tienen interés porque el lift=1, esto, como hemos mencionado al inicio, significa que hay una dependencia o asociación nula entre el antecedente y el consecuente, es decir, hay independencia entre tener coche o estar casado y ser hombre.

El hecho de tener un lift alto (positiva o negativamente) no tiene por qué significar que la regla sea de calidad, pero el hecho de que sea 1 nos lleva descartar directamente la regla sin que nos importe que la confianza sea del 100%.

Una solución podría ser usar otra métrica con la que ordenamos las reglas, si usamos lift, nos lo ordena de mayor a menor, y aunque solo obtengamos reglas con dependencias positivas o directas altas, no son =1.

BIBLIOGRAFÍA

Wikipedia contributors. (2023, October 10). *Association rule learning*. Wikipedia. https://en.wikipedia.org/wiki/Association_rule_learning

Wikipedia contributors. (2022, June 24). *Lift (data mining)*. Wikipedia. https://en.wikipedia.org/wiki/Lift_%28data_mining%29

R. Agrawal, R. Srikant: Fast Algorithms for Mining Association Rules in Large Databases. In: 20th International Conference on Very Large Data Bases, 478-499, 1994.

IndexDataMine. (n.d.). https://www.uv.es/mlejarza/datamine/

How do I know the “support” of each association rules in Weka? (n.d.). Stack Overflow. https://stackoverflow.com/questions/30722889/how-do-i-know-the-support-of-each-association-rules-in-weka

El Baúl del Programador. (2018, April 3). *Aprendizaje no Supervisado y Detección de Anomalías: Reglas de Asociación Avanzadas*. Aprendizaje No Supervisado Y Detección De Anomalías: Reglas De Asociación Avanzadas. https://elbauldelprogramador.com/aprendizaje-nosupervisado-reglas-avanzadas/