STUDY OF THE ASSOCIATION RULES ON SUPERMARKET PRODUCTS AND EMPLOYEE STAFFING

DATA MINING IN BUSINESS

Ágatha del Olmo Tirado | 2nd BIA | 07/12/2023

BUSINESS INTELLIGENCE & ANALYTICS

INDEX

INTRODUCTION

This report presents a detailed analysis of the induction rules based on data from the business and consumer environment collected in the "employees.arff" and "supermarket.arff" files. The aim of this study is to employ association techniques to find interesting patterns in both contexts. Throughout the study we have used the "a priori" technique in the Explorer environment in Weka, both with its predetermined metrics and with certain variations as we were interested.

For the "employees.arff" dataset, the aim is to discover association rules that allow us to understand the relationships between characteristics such as salary, marital status, car ownership, number of children, form of housing, etc. to which each employee belongs. The goal is to obtain information that can be useful in human resource management decision-making.

In the case of "supermercado.arff", we seek to identify purchasing patterns associated with the variable "total" (total amount of purchase) to understand which products tend to be purchased together and how these associations can be used to improve marketing strategies or product availability in the establishment.

GLOSSARY OF CONCEPTS

For a better understanding of the metrics discussed in the study, we have made a glossary of concepts for all readers to follow.

A priori algorithm

An algorithm that iteratively reduces minimum coverage until it finds the number of rules indicated with the specified minimum confidence. You have the option to extract rules with class association (with a class as the predictor of the rule).

Coverage

Percentage of times the antecedent and the consequent are given at the same time in the database.
$Coverage\ (X\text{->}Y) = P(XnY) = P(X) * P(Y/X) = P(Y) * p(X/Y)$

Confidence

Percentage of instances in which the antecedent has led to the consequent.
$Conf\ (X\text{->}Y) = P(Y/X)$

- High Confidence: Values above 80-90% are generally considered high.

- Medium confidence: Values in the range of 60-80% can be considered median.

- Low Confidence: Values below 60% are considered low

Confidence

Percentage of instances in which the antecedent has led to the consequent.

$Conf\ (X\text{->}Y) = P(Y/X)$

- High Confidence: Values above 80-90% are generally considered high.

- Medium confidence: Values in the range of 60-80% can be considered median.

- Low Confidence: Values below 60% are considered low

Lift

The survey measures the level of interest.
$Lift\ (X\text{->}Y) = P(Y/X)\ /\ P(Y)$

- Positive association: If lift>1 indicates a positive dependence. The greater the lift, the stronger the direct association.

- Null association: If lift=1, null dependence. When a rule indicates a null association, this implies that the rule can be discarded.

- Negative association: If lift < 1 indicates a negative dependence. The smaller the lift, the stronger the inverse association.

Levarage

It measures the correlation between antecedent and consequential by comparing the coverage of the latter under the assumption of independence and the actual coverage of the database.
*Levarage (X->Y) = P (XuY) - P(X) * P(Y)*

- <u>High co-occurrence:</u> If Lev > 0, X, and Y are positively correlated, the observed coverage is greater than the expected coverage.

- <u>Null joint occurrence</u>: if Lev=0, X and Y are uncorrelated.

- <u>Low co-occurrence:</u> If Lev < 0, X, and Y are negatively correlated, they tend not to occur in the same rule.

Conviction

It quantifies how independent antecedent and consequential they are. This measure is very useful for identifying meaningful association rules.
Conviction (X->Y) = (P(X). P(Ȳ)) / P(XnȲ)

- <u>Dependence</u>: The higher the value of Conv, the more dependent the consequent is on the antecedent (in a negative sense (inversely) or positively (direct).

- <u>Independence</u>: Conv = 1 implies independence.

INTERPRETATION OF THE RULES

To begin with, let's perform an exhaustive interpretation of the metrics and meaning of each of the ten rules in both databases with the "Apriori" default algorithm. It is worth mentioning that the conviction metric has been interpreted in relation to the rest of the rules of each database.

EMPLOYEE DATABASE

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

- **Trust** (Conf): 100%. It indicates that, in 100% of individuals who own property, they also have a car.
- **Lift**: 1.31. It indicates a positive association in which the occurrence of owning a car is 1.31 times more likely when owning a property compared to the overall occurrence of owning a car.
- **Coverage**: 132/318=0.42%. Indicates that this rule is found in 0.42% of cases in the database.
- **Conviction**: 31.13. It indicates a fairly high dependency between antecedent and consequent.

2. Alq/Prop=Low Rent/Year=None 119==> Married=No 119 <conf:(1)> lift:(1.81) lev:(0.17) [53] conv:(53.14)

- **Trust** (Conf): 100%. It indicates that, in 100% of the individuals who have a rental and do not have any sick leave per year, they are not married.
- **Lift**: 1.81. It indicates that the occurrence of being unmarried is 1.81 times more likely when you do not have any sick leave per year and have a rent compared to the general occurrence of not being married.
- **Coverage**: 53/318=0.37%. Indicates that this rule is found in 0.37% of the cases in the database.
- **Conviction**: 53.14. It indicates a very high dependence between antecedent and consequent.

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

- **Trust** (Conf): 100%. It indicates that, in 100% of individuals who have a Less than 12500€, they are men.
- **Lift**: 1.71. It indicates that the occurrence of being male is 1.71 times more likely if you have a Salary less than 12500€ compared to the general occurrence of being male.
- **Coverage**: 111/318=0.35%. Indicates that this rule is found in 0.35% of the cases in the database.
- **Conviction**: 46.08. It indicates a fairly high dependency between antecedent and consequent.

4. Married=Yes Alq/Prop=Prop 104==> Car=Yes 104 <conf:(1)> lift:(1.31) lev:(0.08) [24] conv:(24.53)

- **Trust** (Conf): 100%. It indicates that, in 100% of individuals who are married and own property, they have a car.
- **Lift**: 1.31. It indicates that the occurrence of owning a car is 1.31 times more likely when married and owning property compared to the overall occurrence of owning a car.
- **Coverage**: 104/318=0.33%. Indicates that this rule is found in 0.33% of the cases in the database.
- **Conviction**: 24.53. It indicates a not very high dependency between antecedent and consequent.

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

- **Trust** (Conf): 100%. It indicates that, in 100% of individuals who are men and own property, they are married.

- **Lift**: 0.18. It indicates that the occurrence of being married is 0.18 times more likely when you are a man and have property compared to the overall occurrence of being married.
- **Coverage**: 104/318=0.33%. Indicates that this rule is found in 0.33% of the cases in the database.
- **Conviction**: 57.56. It indicates a very high dependence between antecedent and consequent.

6. Married=Yes alq/prop=prop 104==> sex=H 104 <conf:(1)> lift:(1.71) lev:(0.14) [43] conv:(43.17)

- **Trust** (Conf): 100%. It indicates that 100% of individuals who are married and own property are men.
- **Lift**: 1.71. It indicates that the occurrence of being male is 1.71 times more likely when married and owning property compared to the overall occurrence of being male.
- **Coverage**: 104/318=0.33%. Indicates that this rule is found in 0.33% of the cases in the database.
- **Conviction**: 43.17. It indicates a fairly high dependency between antecedent and consequent.

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

- **Trust** (Conf): 100%. It indicates that, in 100% of individuals who are men and own property, they have a car.
- **Lift**: 1.31. It indicates that the occurrence of owning a car is 1.31 times more likely when you are a man and have property compared to the general occurrence of owning a car.
- **Coverage**: 104/318=0.33%. Indicates that this rule is found in 0.33% of the cases in the database.
- **Conviction**: 24.53. It indicates a not very high dependency between antecedent and consequent.

8. Car=Yes Alq/Prop=Prop Sex=H 104==> Married=Yes 104 <conf:(1)> lift:(2.24) lev:(0.18) [57] conv:(57.56)

- **Trust** (Conf): 100%. It indicates that, in 100% of the individuals who are self-driving and own property and are men, they are married.
- **Lift**: 2.24. It indicates that the occurrence of being married is 2.24 times more likely when you are a man, have a car and own property compared to the

general occurrence of being married.
- **Coverage**: 104/318=0.33%. Indicates that this rule is found in 0.33% of the cases in the database.
- **Conviction**: 57.56. It indicates a very high dependence between antecedent and consequent.

9. Married=Yes Alq/Prop=Prop Sex=H 104==> Car=Yes 104 <conf:(1)> lift:(1.31) lev:(0.08) [24] conv:(24.53)

- **Trust** (Conf): 100%. It indicates that, in 100% of individuals who are married, own property and are men, they have a car.
- **Lift**: 1.31. It indicates that the occurrence of owning a car is 1.31 times more likely when you are married, a man, and own property compared to the general occurrence of owning a car.
- **Coverage**: 104/318=0.33%. Indicates that this rule is found in 0.33% of the cases in the database.
- **Conviction**: 24.53. It indicates a not very high dependency between antecedent and consequent.

10. Married=Yes Car=Yes Sex=H 104==> Alq/Prop=Prop 104 <conf:(1)> lift:(2.41) lev:(0.19) [60] conv:(60.83)

- **Trust** (Conf): 100%. It indicates that 100% of individuals who are married, are men and have a car, they own property.

- **Lift**: 2.41. It indicates that the occurrence of owning property is 2.41 times more likely when you are married, a man, and have a car compared to the general occurrence of owning property.

- **Coverage**: 104/318=0.33%. Indicates that this rule is found in 0.33% of the cases in the database.

- **Conviction**: 60.83. It indicates a very high dependence between antecedent and consequent, in fact it is the one that presents the most among these ten.

SUPERMARKET DATABASE

1. Biscuits=T Frozen=T Fruits=T Total=High 788==> Bakery=T 723 <conf:(0.92)> lift:(1.27) lev:(0.03) [155] conv:(3.35)

- **Confidence** (Conf): 92%. It indicates that, on 92% of the occasions in which biscuits, frozen foods, fruits were purchased, and the total amount of the purchase was high, bakery was also purchased.
- **Lift**: 1.27. It indicates that bakery occurrence is 1.27 times more likely when buying cookies, frozen, fruits, and the total purchase amount is high compared to the overall bakery occurrence.
- **Coverage**: 723/4627=0.16%. Indicates that this rule is found in 0.16% of cases in the database.
- **Conviction**: 3.35. It indicates an average dependence between antecedent and consequent.

2. Kitchen utensils=t biscuits=t fruits=t total=high 760 ==> bakery=t 696 <conf:(0.92)> lift:(1.27) lev:(0.03) [149] conv:(3.28)

- **Confidence** (Conf): 92%. It indicates that, on 92% of the occasions in which cookies, kitchen utensils, fruits, and the total amount of the purchase was high, bakery was also purchased.
- **Lift**: 1.27. It indicates that bakery occurrence is 1.27 times more likely when cookies, kitchen utensils, fruits are purchased, and the total purchase amount is high compared to the overall bakery occurrence.
- **Coverage**: 696/4627=0.15%. Indicates that this rule is found in 0.15% of cases in the database.
- **Conviction**: 3.28. It indicates an average dependence between antecedent and consequent.

3. Kitchen utensils=T frozen=T fruits=T total=High 770==> Bakery=T 705 <conf:(0.92)> lift:(1.27) lev:(0.03) [150] conv:(3.27)

- **Confidence** (Conf): 92%. It indicates that, on 92% of the occasions in which kitchen utensils, frozen foods, fruits were purchased, and the total amount of the purchase was high, bakery was also purchased.
- **Lift**: 1.27. It indicates that the occurrence of bakery is 1.27 times more likely when buying kitchen utensils, frozen, fruits, and the total amount of the purchase is high compared to the general occurrence of bakery.
- **Coverage**: 705/4627=0.15%. Indicates that this rule is found in 0.15% of cases in the database.
- **Conviction**: 3.27. It indicates an average dependence between antecedent and consequent.

4. Biscuits=T Fruits=T Vegetables=Total T=High 815==> Bakery=T 746

<conf:(0.92)> lift:(1.27) lev:(0.03) [159] conv:(3.26)

- **Confidence** (Conf): 92%. It indicates that, on 92% of the occasions in which cookies, fruits, vegetables were purchased, and the total amount of the purchase was high, bakery was also purchased.
- **Lift**: 1.27. It indicates that bakery occurrence is 1.27 times more likely when cookies, fruits, vegetables are purchased, and the total purchase amount is high compared to the overall bakery occurrence.
- **Coverage**: 746/4627=0.16%. Indicates that this rule is found in 0.16% of cases in the database.
- **Conviction**: 3.26. It indicates an average dependence between antecedent and consequent.

5. Snacks=T Fruits=T Total=High 854==> Bakery=T 779 <conf:(0.91)> lift:(1.27) lev:(0.04) [164] conv:(3.15)

- **Confidence** (Conf): 91%. It indicates that, on 91% of the occasions in which snacks and fruits were purchased, and the total amount of the purchase was high, bakery was also purchased.
- **Lift**: 1.27. It indicates that the bakery occurrence is 1.27 times more likely when snacks, fruits, and the total purchase amount is high compared to the general bakery occurrence.
- **Coverage**: 779/4627=0.17%. Indicates that this rule is found in 0.17% of cases in the database.
- **Conviction**: 3.15. It indicates a slightly lower-than-average dependence between antecedent and consequent.

6. biscuits=t frozen=t vegetables=t total=high 797 ==> bakery=t 725 <conf:(0.91)> lift:(1.26) lev:(0.03) [151] conv:(3.06)

- **Confidence** (Conf): 91%. It indicates that, on 91% of the occasions in which biscuits, frozen foods, vegetables were purchased, and the total amount of the purchase was high, bakery was also purchased.
- **Lift**: 1.26. It indicates that bakery occurrence is 1.26 times more likely when buying biscuits, frozen, vegetables, and the total purchase amount is high compared to the overall bakery occurrence.
- **Coverage**: 725/4627=0.16%. Indicates that this rule is found in 0.16% of cases in the database.
- **Conviction**: 3.06. It indicates a slightly lower-than-average dependence between antecedent and consequent.

7. Kitchen utensils=T biscuits=T vegetables=T total=High 772==> Bakery=T 701 <conf:(0.91)> lift:(1.26) lev:(0.03) [145] conv:(3.01)

- **Confidence** (Conf): 91%. It indicates that, on 91% of the occasions in which biscuits, kitchen utensils, vegetables were purchased, and the total amount of the purchase was high,

Bakery was also bought.
- **Lift**: 1.26. It indicates that bakery occurrence is 1.26 times more likely when biscuits, kitchen utensils, vegetables are purchased, and the total purchase amount is high compared to the overall bakery occurrence.
- **Coverage**: 701/4627=0.15%. Indicates that this rule is found in 0.15% of cases in the database.
- **Conviction**: 3.01. It indicates a lower-than-average dependence between antecedent and consequent.

8. Biscuits=T Fruits=Total T=High 954==> Bakery=T 866 <conf:(0.91)> lift:(1.26) lev:(0.04) [179] conv:(3)

- **Confidence** (Conf): 91%. It indicates that, on 91% of the occasions in which cookies and fruits were purchased, and the total amount of the purchase was high, bakery was also purchased.
- **Lift**: 1.26. It indicates that bakery occurrence is 1.26 times more likely when buying cookies, fruits, and the total purchase amount is high compared to the overall bakery occurrence.
- **Coverage**: 866/4627=0.19%. Indicates that this rule is found in 0.19% of the cases in the database.
- **Conviction**: 3. Indicates a below-average dependence between antecedent and consequent.

9. Frozen=T Fruits=T Vegetables=T Total=High 834==> Bakery=T 757 <conf:(0.91)> lift:(1.26) lev:(0.03) [156] conv:(3)

- **Confidence** (Conf): 91%. It indicates that, on 91% of the occasions in which frozen fruits, fruits and vegetables were purchased, and the total amount of the purchase was high, bakery was also purchased.
- **Lift**: 1.26. It indicates that the bakery occurrence is 1.26 times more likely when frozen foods, fruits, and vegetables are purchased, and the total purchase amount is high compared to the overall bakery occurrence.
- **Coverage**: 757/4627=0.16%. Indicates that this rule is found in 0.16% of cases in the database.
- **Conviction**: 3. Indicates a below-average dependence between antecedent and consequent.

10. Frozen=T Fruits=T Total=High 969==> Bakery=T 877 <conf:(0.91)> lift:(1.26) lev:(0.04) [179] conv:(2.92)

- **Confidence** (Conf): 91%. It indicates that, on 91% of the occasions in which frozen foods and fruits were purchased, and the total amount of the purchase was high, bakery was also purchased.

- **Lift**: 1.26. It indicates that bakery occurrence is 1.26 times more likely when frozen and fruits are purchased, and the total purchase amount is high compared to the overall bakery occurrence.
- **Coverage**: 877/4627=0.19%. Indicates that this rule is found in 0.19% of the cases in the database.
- **Conviction**: 2.92. It indicates the lowest dependence between antecedent and consequent among these ten rules.

BILL AMOUNT AS A PREDICTOR

If we change to True car in 'show properties', that is, if we use reference class as a predictor, which by default will be the last variable, which in this case is "total", no rule with a confidence level higher than 0.9 is generated. However, if we lower this threshold to 0.8, we get the following rules:

1. Kitchen utensils=t biscuits=t sauces=t frozen=t tissues=t 574 ==> total=high 470 conf:(0.82)

2. bakery=t biscuits=t sauces=t frozen=t tissues=t 600 ==> total=high 491 conf:(0.82)

3. bakery=t kitchen utensils=t sauces=t frozen=t tissues=t 620 ==> total=high 506 conf:(0.82)

4. bakery=t kitchen utensils=t biscuits=t sauces=t tissues=t 595 ==> total=high 483 conf:(0.81)

5. bakery=t biscuits=t sauces=t tissues=t vegetables=t 583 ==> total=high 469 conf:(0.8)

6. Bakery=t sauces=t frozen=t tissues=t vegetables=t 610 ==> total=high 490 conf:(0.8)

The most prevalent food combination that precedes in a high-value total includes bakery, kitchen utensils, sauces, frozen and tissues.

The most common combination of products is the third with bakery, kitchen utensils, sauces, frozen foods, tissues. This combination of items has a total of 620 appearances, and 506 of these have resulted in a high-value purchase. This represents a coverage of 506/4627=0.11%, i.e. very close to 0, so we can say that it is not too relevant.

COVERAGE GREATER THAN 50%

By setting a minimum coverage of 0.5, it is necessary to reduce trust to 0.78 to be able to get two rules:

1. Cream=T 2939==> Bakery=T 2337   <conf:(0.8)> lift:(1.1) lev:(0.05) [221] conv:(1.37)

2. Fruits=T 2962==> Bakery=T 2325   <conf:(0.78)> lift:(1.09) lev:(0.04) [193] conv:(1.3)

Initially, you might suggest arranging the cream and fruits next to the bakery. However, as we know that it is not unusual, but rather the opposite, that bread is bought together with any combination of products, this recommendation and any other that includes bakery loses relevance.

DATABASE LIMITED TO A HIGH TOTAL AMOUNT

If we select only the cases that involve a high bill with the *command weka.filters.unsupervised.instance.RemoveWithValues -S 0.0 -C last -L 1*, and remove the "Total" attribute, which no longer makes sense because there is only a high total, the first four rules that we get with the "a priori" algorithm are the following:

1. Biscuits=T Frozen=T Fruits=T 788==> Bakery=T 723   <conf:(0.92)> lift:(1.09) lev:(0.04) [59] conv:(1.89)

2. Kitchen utensils=t biscuits=t fruit=T 760==> bakery=T 696   <conf:(0.92)> lift:(1.09) lev:(0.03) [56] conv:(1.85)

3. Kitchen utensils=t frozen=t fruits=t 770 ==> bakery=t 705   <conf:(0.92)> lift:(1.09) lev:(0.03) [56] conv:(1.85)

4. biscuits=t fruits=t vegetables=t 815 ==> bakery=t 746   <conf:(0.92)> lift:(1.09) lev:(0.04) [60] conv:(1.84)

We cannot conclude that these are the most frequent products among high total purchases, since in these rules confidence is maximized, which has nothing to do with the totality of the cases, but with the bakery total. On the other hand, it would be more appropriate to maximize the coverage, since it would show us, from the database of high total amount, how many occurrences the rules have, which is the definition of frequency.

BAKERY AS A CONSEQUENCE

If we ask the algorithm to extract 30 rules, we get the following:

1. Biscuits=T Frozen=T Snacks=T Fruits=T Vegetables=T Total=High 510 ==> Bakery=T 478<conf:(0.94)> lift:(1.3) lev:(0.02) [110] conv:(4.33)
2. Biscuits=T Frozen=T Cheeses=T Fruits=T Total=High 495==> Bakery=T 463 <conf:(0.94)> lift:(1.3) lev:(0.02) [106] conv:(4.2)
3. biscuits=t cheeses=t fruits=t vegetables=t total=high 513 ==> bakery=t 479 <conf:(0.93)> lift:(1.3) lev:(0.02) [109] conv:(4.11)
4. Kitchen utensils=t biscuits=t snacks=t fruits=t total=high 557 ==> bakery=t 520 <conf:(0.93)> lift:(1.3) lev:(0.03) [119] conv:(4.11)
5. Kitchen utensils=t cheeses=t fruits=t vegetables=t total=high 519 ==> bakery=t 483 <conf:(0.93)> lift:(1.29) lev:(0.02) [109] conv:(3.93)
6. Frozen=T Snacks=T Tissues=T Fruits=T Total=High 518==> Bakery=T 482 <conf:(0.93)> lift:(1.29) lev:(0.02) [109] conv:(3.92)
7. Juices=T Biscuits=T Snacks=T Fruits=T Total=High 529 ==> Bakery=T 492 <conf:(0.93)> lift:(1.29) lev:(0.02) [111] conv:(3.9)
8. Biscuits=T Cheeses=T Fruits=T Total=High 584==> Bakery=T 543<conf:(0.93)> lift:(1.29) lev:(0.03) [122] conv:(3.9)
9. biscuits=t snacks=t fruits=t vegetables=t total=high 596 ==> bakery=t 554 <conf:(0.93)> lift:(1.29) lev:(0.03) [125] conv:(3.89)
10. Kitchen utensils=t biscuits=t frozen=t fruits=t vegetables=t total=high 561 ==> bakery=t 521 <conf:(0.93)> lift:(1.29) lev:(0.03) [117] conv:(3.84)
11. Biscuits=T Frozen=T Snacks=T Fruits=T Total=High 589==> Bakery=T 547 <conf:(0.93)> lift:(1.29) lev:(0.03) [123] conv:(3.84)
12. Kitchen utensils=t frozen=t snacks=t fruits=t total=high 558 ==> bakery=t 518 <conf:(0.93)> lift:(1.29) lev:(0.03) [116] conv:(3.81)
13. Biscuits=T Snacks=T Tissues=T Fruits=T Total=High 515 ==> Bakery=T 478 <conf:(0.93)> lift:(1.29) lev:(0.02) [107] conv:(3.8)
14. Kitchen utensils=t Cheeses=T Fruits=T total=High 584 ==> Bakery=T 542 <conf:(0.93)> lift:(1.29) lev:(0.03) [121] conv:(3.81)
15. Kitchen utensils=t frozen=t tissues=t fruits=t vegetables=t total=high 513 ==> bakery=t 476 <conf:(0.93)> lift:(1.29) lev:(0.02) [106] conv:(3.78)
16. Biscuits=T canned vegetables=T fruits=T total=High 523 ==> Bakery=T 485 <conf:(0.93)> lift:(1.29) lev:(0.02) [108] conv:(3.76)
17. Snacks=T Cheeses=T Fruits=T Total=High 535==> Bakery=T 496<conf:(0.93)> lift:(1.29) lev:(0.02) [110] conv:(3.75)
18. biscuits=t cream=t margarine=t fruits=t total=high 506 ==> bakery=t 469 <conf:(0.93)> lift:(1.29) lev:(0.02) [104] conv:(3.73)
19. snacks=t tissues=t fruits=t vegetables=t total=tall 530 ==> bakery=t 491 <conf:(0.93)> lift:(1.29) lev:(0.02) [109] conv:(3.71)
20. Frozen=T Snacks=T Cream=T Fruits=T Total=High 528 ==> Bakery=T 489 <conf:(0.93)> lift:(1.29) lev:(0.02) [109] conv:(3.7)
21. Kitchen utensils=t frozen=t tissues=t fruits=t total=high 581 ==> bakery=t 538 <conf:(0.93)> lift:(1.29) lev:(0.03) [119] conv:(3.7)

```
22. Biscuits=T Frozen=T Tissues=T Fruits=T Vegetables=T Total=High 513==>
Bakery=T 475        <conf:(0.93)> lift:(1.29) lev:(0.02) [105] conv:(3.69)
23. Kitchen utensils=t frozen=t margarine=t fruits=t total=high 553 ==> bakery=t 512
    <conf:(0.93)> lift:(1.29) lev:(0.02) [114] conv:(3.69)
24. Frozen=T Snacks=T Fruits=T Vegetables=T Total=High 593 ==> Bakery=T 549
<conf:(0.93)> lift:(1.29) lev:(0.03) [122] conv:(3.69)
25. Frozen=t Cheeses=T Fruits=T Total=High 579==> Bakery=T 536        <conf:(0.93)>
lift:(1.29) lev:(0.03) [119] conv:(3.69)
26. biscuits=t frozen=t cream=t margarine=t total=high 537 ==> bakery=t 497
<conf:(0.93)> lift:(1.29) lev:(0.02) [110] conv:(3.67)
27. Biscuits=T Cheeses=T Cream=T Total=High 548==> Bakery=T 507
    <conf:(0.93)> lift:(1.29) lev:(0.02) [112] conv:(3.66)
28. Canned vegetables=t frozen=t fruits=t total=high 521 ==> bakery=t 482
<conf:(0.93)> lift:(1.29) lev:(0.02) [107] conv:(3.65)
29. biscuits=t cream=t margarine=t vegetables=t total=tall 507 ==> bakery=t 469
<conf:(0.93)> lift:(1.29) lev:(0.02) [104] conv:(3.64)
30. Biscuits=T Frozen=T Margarine=T Fruits=T Total=High 560==> Bakery=T 518
<conf:(0.93)> lift:(1.29) lev:(0.02) [114] conv:(3.65)
```

As we can see, the minimum confidence set is 0.93 in the thirtieth rule. And it is interesting to note that in all the rules obtained, the consequent is bakery. This repetition of the consequent may be due to the nature of the consequent, because "regardless" of what type of purchase is made, bread is always bought, therefore, these rules are impractical in reality. To avoid this phenomenon, we could simply reduce the maximum confidence, which could lead to more varied rules, as they would be "less obvious".

THE MEANING OF THE RULES: OWNERSHIP AND OWNING A CAR

When we use the "Apriori" algorithm with the default options, we see that the
The first is as follows:

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

It can be confusing that, although we see 100% trust, there are 243 employees with cars, of which only 132 own a home (approximately 54%).

This happens because the direction of the rule changes everything, to understand it we can look at a small and simple example.

We imagine that these are all individuals in a small database:

1. Housing = Ownership, Car = Yes

2. Housing = Rent, Car = Yes

3. Housing = Rent, Car = No

4. Housing = Ownership, Car = Yes

5. Housing = Ownership, Car = Yes

There are 4 employees with cars, and of these only 3 own the home (75%). On the other hand, there are 3 employees who own a home, and of these all have a car (100%). It is important to see the direction of the rule because the denominator of the conditioned probability changes (where X=Housing and Y=Car, the first case would be Conf (Y->X) = P(X/Y) and the second case would be Conf (X->Y) = P(Y/X)).

THE MEANING OF THE RULES: LOW PAY AND BEING A MAN

From among the same rules of the previous section, we obtain this one:

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

This particular rule means that in our database, as long as the employee has a low salary, he will be male 100% of the time. Again, the meaning of the rule can create some confusion, this does not mean that if you are a man you have a low salary, but precisely on the contrary, not every man has a low salary but every person who has a low salary is a man.

THE MEANING OF THE RULES IN A BUSINESS PROFILE

We revert to the "true" car option, and by default the "Department" variable is used as a predictor of the rules. The rules that affect the sales department that we get from are as follows:

1. Married=No Trustee=Yes 95==> department=commercial 95conf:(1)

2. Car=Yes Gender=M 90==> apartment=commercial 90 conf:(1)

3. Married=No Children=0 Trustee=Yes 78==> department=Commercial 78 conf:(1)

4. Married=No Trustee=Yes Departures/Year=none 78 ==> department=commercial 78 conf:(1)

5. Married=No Car=Yes Syndic.=Yes 77==> apartment=Commercial 77 conf:(1)

The profile would not necessarily be a good predictor of the sales department. This is again for the same reasons, trust is *P(Y/X)*, not the other way around, which, as we have seen, changes everything.

LOW-INTEREST RULES

Finally, we remove cases where the individual does not belong to the sales department and then remove the "Department" variable itself so that it does not affect the study. The rules that we obtain and that have as a predictor being male are the following:

6. Married=Yes 83==> Sex=H 83 <conf:(1)> lift:(1) lev:(0) [0] conv:(0)

10. Car=Yes 83==> Sex=H 83 <conf:(1)> lift:(1) lev:(0) [0] conv:(0)

These rules are of no interest because the lift=1, this, as we mentioned at the beginning, means that there is a dependence or null association between the antecedent and the consequent, that is, there is independence between having a car or being married and being a man.

The fact of having a high lift (positively or negatively) does not have to mean that the ruler is of quality, but the fact that it is 1 leads us to directly discard the rule without caring that the confidence is 100%.

A solution could be to use another metric with which we order the rules, if we use lift, it orders it from highest to lowest, and even if we only get rules with high positive or direct dependencies, they are not =1.

BIBLIOGRAPHY

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

The Programmer's Trunk. (2018, April 3). *Unsupervised Learning and Anomaly Detection: Advanced Association Rules*. Unsupervised Learning and Anomaly Detection: Advanced Association Rules. https://elbauldelprogramador.com/aprendizaje-nosupervisado-reglas-Advanced/