**Problem 1**: Consider the 10 retail transactions of cosmetics. For example, the first transaction includes a lipstick, mascara, and an eye liner, while the second transaction includes an eye shadow and mascara

|  |  |
| --- | --- |
| **Transaction** | **Cosmetics Purchased** |
| 1 | Lipstick, mascara, eye liner |
| 2 | Eye shadow, mascara |
| 3 | Primer, concealer, lip gloss |
| 4 | Face brush, concealer, primer |
| 5 | Mascara, lipstick, eye liner |
| 6 | Mascara, eye liner, eye shadow |
| 7 | Lip gloss, eye liner, mascara |
| 8 | Mascara, eye shadow, lip gloss |
| 9 | Face brush, mascara, eye liner, lipstick |
| 10 | Eye liner, lipstick, primer, mascara |

1. Create a frequency distribution that summarizes the number of purchases for each cosmetic product.

Ans:

|  |  |
| --- | --- |
| Item | Count |
| Lipstick | 4 |
| Mascara | 8 |
| Eye liner | 6 |
| Eye Shadow | 3 |
| Primer | 3 |
| Concealer | 2 |
| Lip gloss | 3 |
| Face brush | 2 |

1. For the transactions database above, how many total possible positive association rules are there?

Ans:

No of association rules =

=

= 6050

1. If the minimum threshold of any item-sets is 4, create valid 1-item-sets, 2-item-sets, and 3-item-sets lists.

Ans: 1-Item Sets:

|  |  |
| --- | --- |
| **Item** | **Count** |
| Lip Stick | 4 |
| Mascara | 8 |
| Eye Liner | 6 |

2- Item Sets:

|  |  |
| --- | --- |
| **Item** | **Count** |
| Lip Stick, Mascara | 4 |
| Lip Stick, Eye Liner | 4 |
| Mascara, Eye Liner | 6 |

3-Item Sets:

|  |  |
| --- | --- |
| **Item** | **Count** |
| Lip Stick, Mascara, Eye Liner | 4 |

1. From the previous item-sets, create all possible association rules.

Ans:

From 2-Item Set:

|  |  |
| --- | --- |
| **Antecedent** | **Consequent** |
| If Lip Stick | Then Mascara |
| If Lip Stick | Then Eye Liner |
| If Mascara | Then Lip Stick |
| If Mascara | Then Eye Liner |
| If Eye Liner | Then Mascara |
| If Eye Liner | Then Lip Stick |

From 3-Item Set:

|  |  |
| --- | --- |
| **Antecedent** | **Consequent** |
| If Lip Stick | Then Mascara & Eye Liner |
| If Lip Stick & Mascara | Then Eye Liner |
| If Lip Stick & Eye Liner | Then Mascara |
| If Mascara | Then Lip Stick & Eye Liner |
| If Mascara & Eye Liner | Then Lip Stick |
| If Eye Liner | Then Mascara & Lip Stick |

1. Compute the rule support of the two association rules:

Ans:

1. If mascara then eye liner

Support = (6/10) \* 100

= 60%

1. If mascara and eye liner then lipstick

Support = (4/10)\*100

= 40%

1. Compute the antecedent support of the two association rules:

Ans:

* 1. If mascara then eye liner

Support = (8/10)\*100

= 80%

* 1. If mascara and eye liner then lipstick

Support = (6/10)\*100

= 60%

1. Compute the rule confidence/accuracy of the two association rules:

Ans:

* 1. If mascara then eye liner

Confidence = (6/8)\*100

= 75%

* 1. If mascara and eye liner then lipstick

Confidence = (4/6)\*100

= 66.67%

1. Compute the Lift of the two association rules:

Ans:

* 1. If mascara then eye liner

Lift = 0.75/ 0.6

= 1.25

* 1. If mascara and eye liner then lipstick

Lift = 0.667/ 0.4

= 1.6675

1. Compute the deployability of the two association rules:

Ans:

* 1. If mascara then eye liner

Deployability = P(A) – P(A and B)

= Antecedent Support – Rule Support

= 80 - 60

= 20 %

* 1. If mascara and eye liner then lipstick

Deployability = P(A) – P(A and B)

= Antecedent Support – Rule Support

= 60 – 40

= 20%

1. Compute the J-measure for the association rule: if mascara then eye liner.

Ans: P (X) = 0.8

P (Y) = 0.6

P (Y | X) = 0.75

J-measure = 0.8[(0.75) \* ln (0.75/0.6) + (1-0.75) \* ln (1-0.75/1-0.6)]

= 0.8[(0.75) \* ln (0.75/0.6) + (0.25) \* ln (0.25/0.4)]

= 0.8[(0.75) \* ln (1.25) + (0.25) \* ln (0.625)]

= 0.8[(0.75) \* (0.22314) + (0.25) \* (-0.47)]

= 0.8[0.167355 – 0.1175]

= 0.8 [0.049855]

= 0.039884

1. Compute the confidence ratio of the association rule: if mascara then eye liner.

Ans: Confidence ratio = 1 – min [(P (Y|X)/ P(Y)), (P(Y)/P(Y|X))]

= 1 – min [(0.75/0.6), (0.6/0.75)]

= 1 – min [0.6/0.75]

= 0.2

**Problem 2:**

Suppose that the data mining task is to cluster a large data set of candy bars into two clusters. The input **Calories** is the number of calories per serving and the input **Carb** is the amount of carbohydrates in grams. In this table, only a small sample is extracted from the data set.

|  |  |  |
| --- | --- | --- |
| **Brand** | **Calories** | **Carb** |
| 1. Peanut Butter Twix | 311.0 | 31.4 |
| 2. Baby Ruth | 275.0 | 39.0 |
| 3. Caramel Twix | 284.5 | 37.5 |
| 4. 5th Avenue | 279.5 | 41.0 |
| 5. Snickers | 273.0 | 34.0 |
| 6. Twizzlers | 262.5 | 66.0 |
| 7. Reese's Pieces | 258.0 | 34.0 |
| 8. Mr. Goodbar | 257.0 | 25.5 |
| 9. Whatchamacallit | 256.5 | 30.0 |
| 10. Oh Henry! | 245.5 | 37.0 |
| 11. M&M's Peanut | 242.5 | 28.5 |
| 12. Krackel | 236.0 | 29.0 |

In the K-Means clustering algorithm, the Euclidean distance is used to cluster the records. For the data above, use the K-Means clustering algorithm to perform the computations to cluster the candy bars into two groups, using the two inputs **Calories** and **Carb**. Suppose that initially, the candy bars 4 and 8 were randomly chosen as the two centroids of the two clusters. Using Excel, perform the computations of each phase to reach the solution.

* + - 1. Create a scatter plot of the above data. Label the two axes and give the graph a title.

Ans:

A graph with blue dots

Description automatically generated

A graph of brands data

Description automatically generated

* + - 1. In phase 1, what candy bars constitute clusters 1 and 2 and what are their centroids. Compute the sum of square error (SSE).

Ans: Cluster-1 = (1,2,3,4,5,6)

Cluster-2 = (7,8,9,10,11,12)

Sum of Square Error(SSE) = 3181.91

Centroids = (279.5,41) (257, 25.5)

* + - 1. In the final phase, what candy bars constitute clusters 1 and 2 and what are their centroids? Compute the sum of square error (SSE).

Ans: Cluster-1 = (1,2,3,4,5,6)

Cluster-2 = (7,8,9,10,11,12)

Sum of Square Error (SSE) = 2647.9252

Centroids = (280.91, 41.483) (249.25, 30.667)

**Problem 3**: Randy Johnson is an insurance adjustor for a national auto insurance company. Using historical insurance claim data, Randy built an insurance fraud detection model with the help of a data scientist. After applying the model on the test data set, it generated the following confusion matrix. A fraudulent insurance claim is labeled as Class 1, while a nonfraudulent claim is labeled as Class 0.

|  |  |  |
| --- | --- | --- |
|  | Predicted Class 1 | Predicted Class 0 |
| Actual Class 1 | 130(TP) | 170(FN) |
| Actual Class 0 | 2,402(FP) | 27,298(TN) |

1. Compute the lift.

Ans: Lift = (N\*TP) / ((TP+FP)\*(TP+FN))

Where N= TP + FN + FP + TN

Lift = (30000\*130) / ((130+2402)\*(130+170))

= (3900000)/((2532)\*(300))

= 3900000/759600

= 5.1342

1. Compute the accuracy.

Ans: Accuracy = (TP + TN)/(P + N)

= 0.91426

Accuracy = 91.426 %

1. Compute the false positive rate.

Ans: P = TP + FN

N = FP + TN

False Positive Rate = FP/N = 0.08087

1. Compute the false negative rate.

Ans: P = TP + FN

N = FP + TN

False Negative Rate = FN/P = 0.5666

1. Compute the specificity.

Ans: Specificity = TrueNegatives / (TrueNegatives + FalsePositives)

Specificity = 0.9191

1. Compute the precision.

Ans: Precision = TruePositives / (TruePositives + FalsePositives)

Precision = 0.05134

1. Compute the sensitivity/recall.

Ans: Recall = TruePositives / (TruePositives + FalseNegatives)

Recall = 0.4333

1. Compute the F-1 Score.

Ans: F1 Score = 2(Precision x Sensitivity)/(Precision + Sensitivity)

F1 Score = 0.0918

**Problem 4**: This is a small data set extracted from a social media data set, where **1** indicates **Yes** (the user uses the social media platform) and **0** indicates **No** (the user doesn’t use the social media platform).

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| **User ID** | **Facebook** | **Snapchat** | **Instagram** | **Pinterest** | **Tumblr** | **LinkedIn** | **Twitter** |
| 1 | 1 | 1 | 0 | 0 | 0 | 1 | 0 |
| 2 | 1 | 1 | 1 | 1 | 0 | 1 | 1 |
| 3 | 1 | 0 | 0 | 0 | 0 | 1 | 1 |
| 4 | 0 | 0 | 1 | 1 | 0 | 1 | 0 |
| 5 | 0 | 1 | 0 | 0 | 1 | 1 | 1 |

1. Compute the matching coefficient between user 1 and the others.

Matching coefficient between (User 1 and User 2) = 4/7 = 0.5714

Matching coefficient between (User 1 and User 3) = 5/7 = 0.7142

Matching coefficient between (User 1 and User 4) = 3/7 = 0.4285

Matching coefficient between (User 1 and User 5) = 4/7 = 0.5714

2. Compute the Jaccard’s coefficient between user 1 and the others.

Jaccard’s coefficient between (User 1 and User 2) = 3/(7-1) = 0.5

Jaccard’s coefficient between (User 1 and User 3) = 2/(7-3) = 0.5

Jaccard’s coefficient between (User 1 and User 4) = 1/(7-2) = 0.2

Jaccard’s coefficient between (User 1 and User 5) = 2/(7-2) = 0.4

**Problem 5**: Suppose that we have the following data for just one input. Using the single-linkage clustering technique, identify the clusters.

0 0 1 3 3 6 7 9 10 10

Ans:

Step-1: 0 0 3 3 10 10

Step-2: 0 0 1 3 3 6 7 9 10 10

Step-3: 0 0 1 3 3 6 7 9 10 10

Step-4: 0 0 1 3 3 6 7 9 10 10