Documentation Data Mining

by:

- عبدالرحمن رعدان
- مصعب الحبيشي
- هادي الجمعي
- مالك المصبحي
- عبدالعزيز عبدالغني
- ماجد النائب

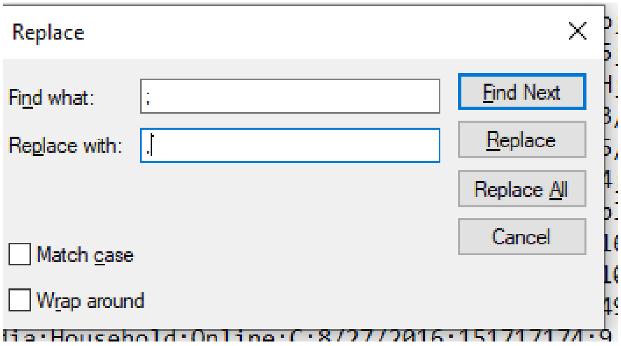
اشراف :

أ. ابراهيم الذارحي

Data Preprocessing Steps

Handling Delimiters

- **Step 1:** Converting Commas to Semicolons in the Dataset
  Objective: Ensure the dataset is properly formatted by replacing all semicolons (;) with commas (,).

  Find what: ;
  Replace with: ,

  lia:Household:Online:C:8/27/2016:151717174:9

- **Step 2:** Clean the "Country" Column Data **Objective**: Ensure data quality by addressing issues with special characters. In our database, we found that the value "Cote d'Ivoire" in the "Country" column contains an apostrophe, which is causing an error in the Wiki.

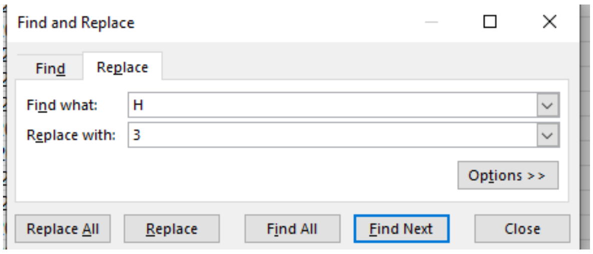
  Problem: The value "Cote d'Ivoire" includes an apostrophe that leads to errors during data analysis or query execution.

- **Step** 3: Re-encoding the "Order Priority" Column We performed re-encoding of the textual values in the "Order Priority" column into numerical values to facilitate processing by machine learning algorithms.

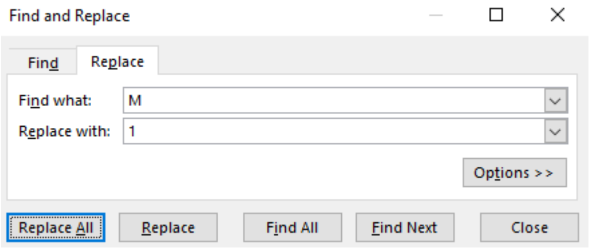
  The values were transformed as follows:

  ```
  Critical → 3
  High → 2
  Medium → 1
  Low → 0
  ```

  This step was necessary because some algorithms, such as Naïve Bayes and K-Means, require numerical data for analysis and processing.

  Find what: H
  Replace with: 3

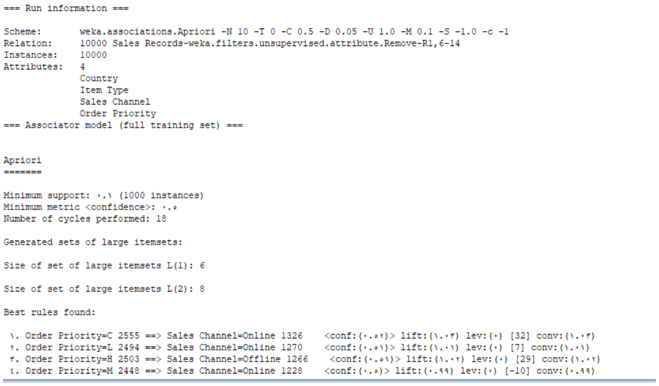
  

  Find what: M
  Replace with: 1

Algorithm Implementation

Apriori Algorithm (Association Rule Mining)

- **Objective:** Discover frequent itemsets and generate association rules.
- **Selected Columns:** Item Type, Sales Channel.
- **Parameters:**
  - Support Threshold: A reasonable value based on dataset characteristics.
  - Confidence Threshold: Set a meaningful value to filter rules.
  - Lift: Evaluated to assess rule significance.
- **Steps:**
  - Preprocessed data by removing duplicates and inconsistencies.
  - Implemented Apriori using association rule mining tools in Weka.



```
=== Run information ===

Scheme:       weka.associations.Apriori -N 10 -T 0 -C 0.5 -D 0.05 -U 1.0 -M 0.1 -S -1.0 -c -1
Relation:     10000 Sales Records-weka.filters.unsupervised.attribute.Remove-R1,6-14
Instances:    10000
Attributes:   4
              Country
              Item Type
              Sales Channel
              Order Priority
=== Associator model (full training set) ===


Apriori
=======

Minimum support: ٠.١ (1000 instances)
Minimum metric <confidence>: ٠.٥
Number of cycles performed: 18

Generated sets of large itemsets:

Size of set of large itemsets L(1): 6

Size of set of large itemsets L(2): 8

Best rules found:

 ١. Order Priority=C 2555 ==> Sales Channel=Online 1326    <conf:(٠.٥٢)> lift:(١.٠٣) lev:(٠) [32] conv:(١.٠٣)
 ٢. Order Priority=L 2494 ==> Sales Channel=Online 1270    <conf:(٠.٥١)> lift:(١.٠١) lev:(٠) [7] conv:(١.٠١)
 ٣. Order Priority=H 2503 ==> Sales Channel=Offline 1266    <conf:(٠.٥١)> lift:(١.٠٢) lev:(٠) [29] conv:(١.٠٢)
 ٤. Order Priority=M 2448 ==> Sales Channel=Online 1228    <conf:(٠.٥)> lift:(٠.٩٩) lev:(٠) [-10] conv:(٠.٩٩)
```

Algorithm and Reasoning:

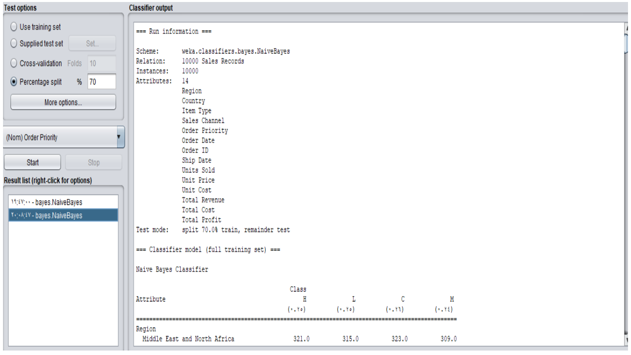
The Apriori algorithm is used to identify frequent itemsets in a dataset and generate association rules. It helps in discovering interesting relationships between variables in large databases.

Parameters:

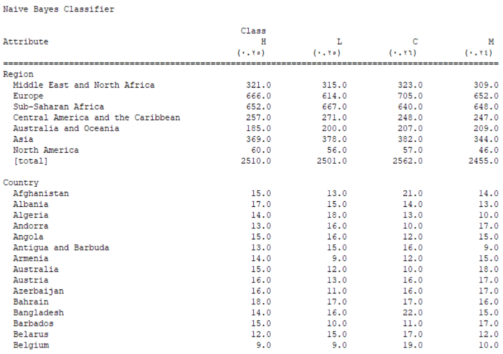
- **Support Threshold:** A minimum percentage of records in the dataset that contain the itemset. In our case, with 10,000 records, a support threshold of 0.05 means any itemset must appear in at least 500 transactions to be considered significant.
- **Confidence Threshold:** The likelihood that a rule is true for the dataset. A threshold of 0.5 means we are interested in rules where the likelihood is at least 50%.
- **Lift:** Measures the importance of a rule. Lift > 1 indicates a strong association.

Naïve Bayes (Classification)

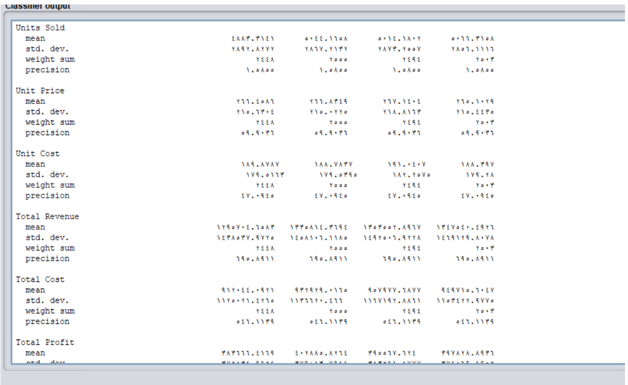
- **Objective:** Build a probabilistic model to classify data into predefined classes.
- **Selected Columns:** Region, Country, Item Type, Sales Channel, Order Priority.
- **Steps:**
  - Split dataset into training (70%) and testing (30%) sets.
  - Assumed feature independence.
  - Evaluated model using accuracy, precision, recall, and F1-score.



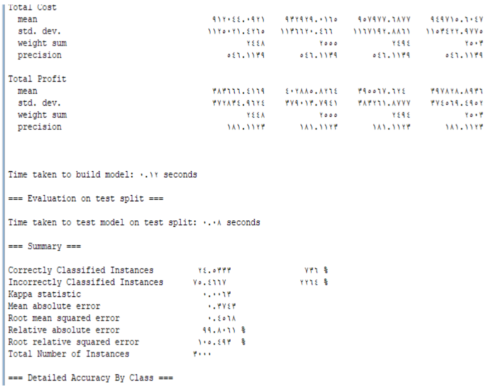
```
=== Run information ===

Scheme:       weka.classifiers.bayes.NaiveBayes
Relation:     10000 Sales Records
Instances:    10000
Attributes:   14
              Region
              Country
              Item Type
              Sales Channel
              Order Priority
              Order Date
              Order ID
              Ship Date
              Units Sold
              Unit Price
              Unit Cost
              Total Revenue
              Total Cost
              Total Profit
Test mode:    split 70.0% train, remainder test

=== Classifier model (full training set) ===

Naive Bayes Classifier

                                     Class
Attribute                                H          L          C          M
                                    (٠.٢٥)     (٠.٢٥)     (٠.٢٦)     (٠.٢٤)
=================================================================================
Region
  Middle East and North Africa        321.0      315.0      323.0      309.0
```



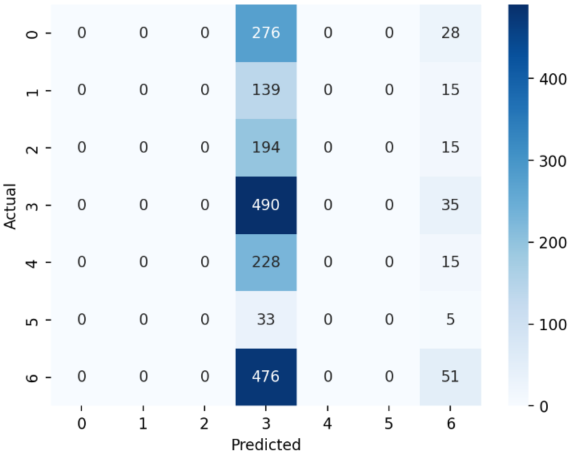
```
Naive Bayes Classifier

                                           Class
Attribute                                      H            L            C            M
                                          (٠.٢٥)       (٠.٢٥)       (٠.٢٦)       (٠.٢٤)
=================================================================================================
Region
  Middle East and North Africa             321.0        315.0        323.0        309.0
  Europe                                   666.0        614.0        705.0        652.0
  Sub-Saharan Africa                       652.0        667.0        640.0        648.0
  Central America and the Caribbean        257.0        271.0        248.0        247.0
  Australia and Oceania                    185.0        200.0        207.0        209.0
  Asia                                     369.0        378.0        382.0        344.0
  North America                             60.0         56.0         57.0         46.0
  [total]                                 2510.0       2501.0       2562.0       2455.0

Country
  Afghanistan                               15.0         13.0         21.0         14.0
  Albania                                   17.0         15.0         14.0         13.0
  Algeria                                   14.0         18.0         13.0         10.0
  Andorra                                   13.0         16.0         10.0         17.0
  Angola                                    15.0         16.0         12.0         15.0
  Antigua and Barbuda                       13.0         15.0         16.0          9.0
  Armenia                                   14.0          9.0         12.0         15.0
  Australia                                 15.0         12.0         10.0         18.0
  Austria                                   16.0         13.0         16.0         17.0
  Azerbaijan                                16.0         11.0         16.0         17.0
  Bahrain                                   18.0         17.0         17.0         16.0
  Bangladesh                                14.0         16.0         22.0         15.0
  Barbados                                  15.0         10.0         11.0         17.0
  Belarus                                   12.0         15.0         17.0         12.0
  Belgium                                    9.0          9.0         19.0         10.0
```

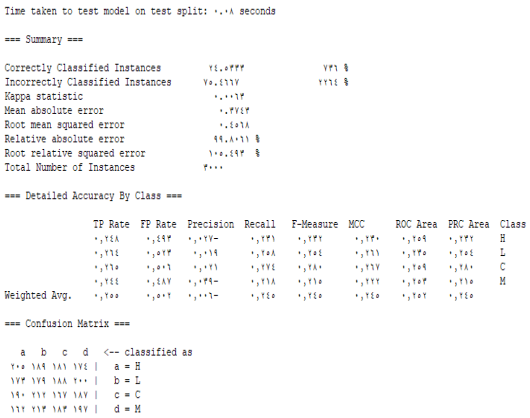


```
Units Sold
  mean                          ٤٨٨٣.٣١٤١      ٥٠٤٤.١١٥٨      ٥٠١٤.١٨٠٢      ٥٠٦٦.٣١٥٨
  std. dev.                     ٢٨٩٢.٨٢٧٢      ٢٨٦٧.٢١٣٢      ٢٨٧٣.٢٥٥٧      ٢٨٥٦.١١١٦
  weight sum                         ٢٤٤٨           ٢٥٥٥           ٢٤٩٤           ٢٥٠٣
  precision                        ١.٥٨٥٥         ١.٥٨٥٥         ١.٥٨٥٥         ١.٥٨٥٥

Unit Price
  mean                           ٢٦٦.٤٥٨٦       ٢٦٦.٨٣٤٩       ٢٦٧.١٤٠٤       ٢٦٥.١٠٢٩
  std. dev.                      ٢١٥.٦٣٠٤       ٢١٥.٠٢٢٥       ٢١٨.٨١٦٣       ٢١٥.٤٤٣٥
  weight sum                         ٢٤٤٨           ٢٥٥٥           ٢٤٩٤           ٢٥٠٣
  precision                       ٥٩.٩٠٣٦        ٥٩.٩٠٣٦        ٥٩.٩٠٣٦        ٥٩.٩٠٣٦

Unit Cost
  mean                           ١٨٩.٨٧٨٧       ١٨٨.٧٨٣٧       ١٩١.٠٤٠٧        ١٨٨.٣٩٧
  std. dev.                      ١٧٩.٥١٦٣       ١٧٩.٥٣٩٥       ١٨٢.٢٥٧٥         ١٧٩.٢٨
  weight sum                         ٢٤٤٨           ٢٥٥٥           ٢٤٩٤           ٢٥٠٣
  precision                       ٤٧.٠٩٤٥        ٤٧.٠٩٤٥        ٤٧.٠٩٤٥        ٤٧.٠٩٤٥

Total Revenue
  mean                       ١٢٩٥٧٠٤.٦٥٨٣   ١٣٣٥٨١٤.٣١٩٤   ١٣٥٣٥٥٢.٨٩٦٧   ١٣٤٧٥٤٠.٤٩٢٦
  std. dev.                  ١٤٣٨٥٣٧.٩٧٢٥   ١٤٥٨١٠٦.١١٨٥   ١٤٩٢٥٠٦.٩٢٢٨   ١٤٦٩١٢٩.٨٠٧٨
  weight sum                         ٢٤٤٨           ٢٥٥٥           ٢٤٩٤           ٢٥٠٣
  precision                      ٦٩٥.٨٩١١       ٦٩٥.٨٩١١       ٦٩٥.٨٩١١       ٦٩٥.٨٩١١

Total Cost
  mean                        ٩١٢٠٤٤.٠٩٢١    ٩٣٢٩٢٩.٠١٦٥    ٩٥٧٩٧٧.٦٨٧٧    ٩٤٩٧١٥.٦٠٤٧
  std. dev.                  ١١٢٥٠٢١.٤٢٦٥    ١١٣٦٦٢٠.٤٦٦   ١١٦٧١٩٢.٨٨٦١   ١١٥٣٤٢٢.٩٧٧٥
  weight sum                         ٢٤٤٨           ٢٥٥٥           ٢٤٩٤           ٢٥٠٣
  precision                      ٥٤٦.١١٣٩       ٥٤٦.١١٣٩       ٥٤٦.١١٣٩       ٥٤٦.١١٣٩

Total Profit
  mean                        ٣٨٣٦٦٦.٤١٦٩    ٤٠٢٨٨٥.٨٢٦٤     ٣٩٥٥٦٧.٦٢٤    ٣٩٧٨٢٨.٨٩٣٦
```



```
Total Cost
  mean                       ٩١٢٠٤٤.٠٩٢١    ٩٣٢٩٢٩.٠١٦٥    ٩٥٧٩٧٧.٦٨٧٧    ٩٤٩٧١٥.٦٠٤٧
  std. dev.                 ١١٢٥٠٢١.٤٢٦٥    ١١٣٦٦٢٠.٤٦٦   ١١٦٧١٩٢.٨٨٦١   ١١٥٣٤٢٢.٩٧٧٥
  weight sum                        ٢٤٤٨           ٢٥٥٥           ٢٤٩٤           ٢٥٠٣
  precision                     ٥٤٦.١١٣٩       ٥٤٦.١١٣٩       ٥٤٦.١١٣٩       ٥٤٦.١١٣٩

Total Profit
  mean                       ٣٨٣٦٦٦.٤١٦٩    ٤٠٢٨٨٥.٨٢٦٤     ٣٩٥٥٦٧.٦٢٤    ٣٩٧٨٢٨.٨٩٣٦
  std. dev.                  ٣٧٢٨٣٤.٩٦٢٤    ٣٧٩٠١٣.٧٩٤١    ٣٨٣٢٦١.٨٧٧٧    ٣٧٤٥٦٩.٤٩٥٢
  weight sum                        ٢٤٤٨           ٢٥٥٥           ٢٤٩٤           ٢٥٠٣
  precision                     ١٨١.١١٢٣       ١٨١.١١٢٣       ١٨١.١١٢٣       ١٨١.١١٢٣



Time taken to build model: ٠.١٢ seconds

=== Evaluation on test split ===

Time taken to test model on test split: ٠.٠٨ seconds

=== Summary ===

Correctly Classified Instances         ٢٤.٥٣٣              ٧٣٦ %
Incorrectly Classified Instances     ٧٥.٤٦٦٧              ٢٢٦٤ %
Kappa statistic                        ٠.٠٠٦٣
Mean absolute error                     ٠.٣٧٤٣
Root mean squared error                 ٠.٤٥٦٨
Relative absolute error               ٩٩.٨٠٦١ %
Root relative squared error          ١٠٥.٤٩٣  %
Total Number of Instances            ٣٠٠٠

=== Detailed Accuracy By Class ===
```





```
Time taken to test model on test split: ٠.٠٨ seconds

=== Summary ===

Correctly Classified Instances         ٢٤.٥٣٣              ٧٣٦ %
Incorrectly Classified Instances     ٧٥.٤٦٦٧              ٢٢٦٤ %
Kappa statistic                        ٠.٠٠٦٣
Mean absolute error                     ٠.٣٧٤٣
Root mean squared error                 ٠.٤٥٦٨
Relative absolute error               ٩٩.٨٠٦١ %
Root relative squared error          ١٠٥.٤٩٣  %
Total Number of Instances            ٣٠٠٠

=== Detailed Accuracy By Class ===

                 TP Rate  FP Rate  Precision  Recall   F-Measure  MCC      ROC Area  PRC Area  Class
                 ٠,٢٤٨     ٠,٤٩٣   ٠,٠٢٧-      ٠,٢٣١    ٠,٢٣٢      ٠,٢٣٠    ٠,٢٥٩     ٠,٢٣٢     H
                 ٠,٢٦٤     ٠,٥٢٣    ٠,٠١٩      ٠,٢٥٨    ٠,٢٥٤      ٠,٢٦١    ٠,٢٣٥     ٠,٢٥٤     L
                 ٠,٢٦٥     ٠,٥٠٦    ٠,٠٢١      ٠,٢٧٤    ٠,٢٨٠      ٠,٢٦٧    ٠,٢٥٩     ٠,٢٨٠     C
                 ٠,٢٤٤     ٠,٤٨٧   ٠,٠٣٩-      ٠,٢١٨    ٠,٢١٥      ٠,٢٢٢    ٠,٢٥٣     ٠,٢١٥     M
Weighted Avg.    ٠,٢٥٥     ٠,٥٠٢   ٠,٠٠٦-      ٠,٢٤٥    ٠,٢٤٥      ٠,٢٤٥    ٠,٢٥٢     ٠,٢٤٥

=== Confusion Matrix ===

   a   b   c   d   <-- classified as
 ٢٠٥ ١٨٩ ١٨١ ١٧٤ |   a = H
 ١٧٣ ١٧٩ ١٨٨ ٢٠٠ |   b = L
 ١٩٠ ٢١٢ ١٦٧ ١٨٧ |   c = C
 ١٦٢ ٢١٣ ١٨٣ ١٩٧ |   d = M
```

Algorithm and Reasoning:

Naïve Bayes is a probabilistic classification algorithm based on Bayes' Theorem. It assumes that the presence of a particular feature in a class is unrelated to the presence of any other feature.

Relevant Columns:

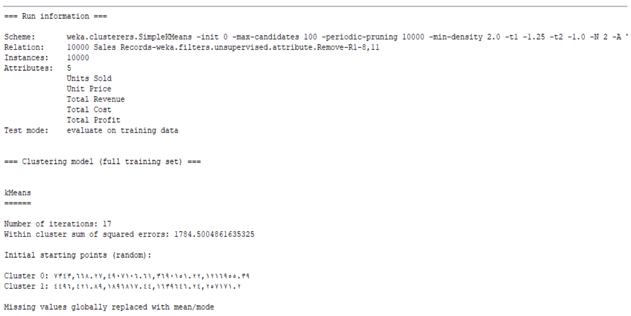
- **Region:** Useful for understanding geographic influence on purchasing behavior.
- **Country:** To identify differences between countries.
- **Item Type:** To know the types of products that are bought together.
- **Sales Channel:** To understand the differences between various sales channels (online and offline).
- **Order Priority:** To understand how order priority affects product relationships.

ID3 Algorithm (Decision Trees)

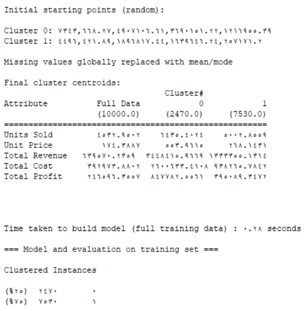
- **Objective:** Create decision trees based on information gain.
- **Selected Columns:** Region, Item Type, Sales Channel, Order Priority, Order Date.
- **Steps:**
  - Used entropy and information gain to construct the tree.
  - Visualized decision tree structure.
  - Evaluated accuracy using cross-validation.

K-Means Algorithm (Clustering)

- **Objective:** Partition the data into clusters based on similarity.
- **Selected Columns:** Units Sold, Unit Price, Total Revenue, Total Profit.
- **Parameters:**
  - Number of Clusters (K): Determined using the elbow method.
  - Initialization: Used k-means++ to enhance convergence.
- **Steps:**
  - Standardized the data.
  - Applied K-Means clustering algorithm.
  - Visualized clusters and centroids.

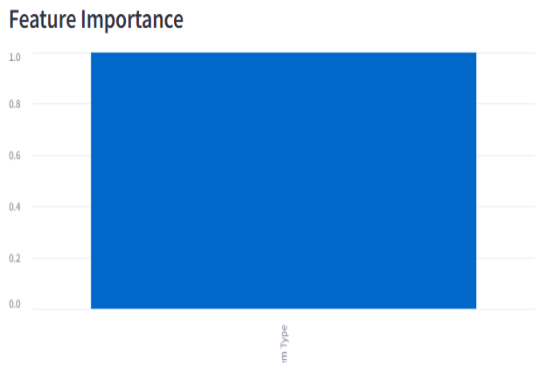


```
=== Run information ===

Scheme:       weka.clusterers.SimpleKMeans -init 0 -max-candidates 100 -periodic-pruning 10000 -min-density 2.0 -t1 -1.25 -t2 -1.0 -N 2 -A '
Relation:     10000 Sales Records-weka.filters.unsupervised.attribute.Remove-R1-8,11
Instances:    10000
Attributes:   5
              Units Sold
              Unit Price
              Total Revenue
              Total Cost
              Total Profit
Test mode:    evaluate on training data


=== Clustering model (full training set) ===


kMeans
======

Number of iterations: 17
Within cluster sum of squared errors: 1784.5004861635325

Initial starting points (random):

Cluster 0: ٧٣٤٣,٦٦٨.٢٧,٤٩٠٧١٠٦.٦١,٣٦٩٠١٥١.٢٢,١٢١٦٩٥٥.٣٩
Cluster 1: ٤٤٩٦,٤٢١.٨٩,١٨٩٦٨١٧.٤٤,١٦٣٩٦٤٦.٢٤,٢٥٧١٧١.٢

Missing values globally replaced with mean/mode
```

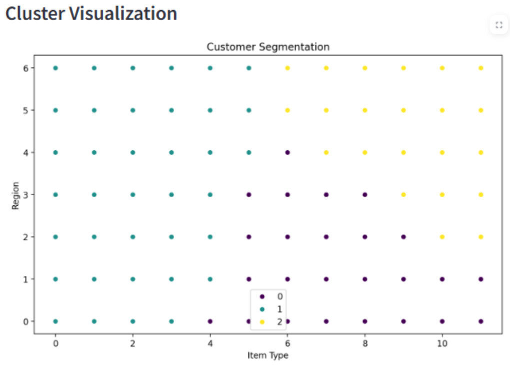


```
Initial starting points (random):

Cluster 0: ٧٣٤٣,٦٦٨.٢٧,٤٩٠٧١٠٦.٦١,٣٦٩٠١٥١.٢٢,١٢١٦٩٥٥.٣٩
Cluster 1: ٤٤٩٦,٤٢١.٨٩,١٨٩٦٨١٧.٤٤,١٦٣٩٦٤٦.٢٤,٢٥٧١٧١.٢

Missing values globally replaced with mean/mode

Final cluster centroids:
                            Cluster#
Attribute       Full Data          0           1
                (10000.0)   (2470.0)    (7530.0)
================================================
Units Sold      ٤٥٣٢.٩٥٠٢   ٦٤٣٥.٤٠٢٤   ٥٠٠٢.٨٥٥٩
Unit Price       ١٧٤.٣٨٨٧    ٥٥٣.٩٦١٥    ٢٦٨.١٤٣١
Total Revenue  ٦٣٩٥٧٠.٢٣٥٩ ٣٤٤٨٤١٥.٩٦٦٩ ١٣٣٣٣٥٥.١٣١٤
Total Cost     ٣٩٢٩٧٣.٨٨٠٢ ٢٦٠٠٦٣٣.٤١٠٨ ٩٣٨٢٦٥.٧٨٤٢
Total Profit   ٢٤٦٥٩٦.٣٥٥٧ ٨٤٧٧٨٢.٥٥٦١ ٣٩٥٠٨٩.٣٤٧٢




Time taken to build model (full training data) : ٠.٢٨ seconds

=== Model and evaluation on training set ===

Clustered Instances

(%٢٥)  ٢٤٧٠       ٠
(%٧٥)  ٧٥٣٠       ١
```

Evaluation Metrics

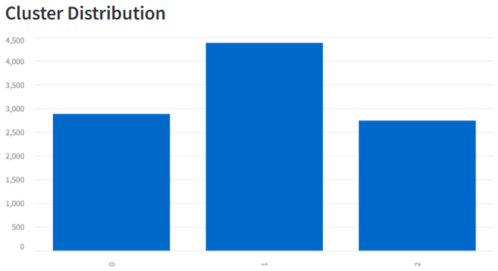


Feature Importance

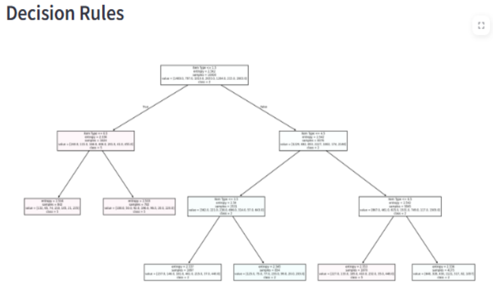


Cluster Visualization

Results and Insights



Cluster Distribution



Decision Rules