Documentation Data Mining

by:

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

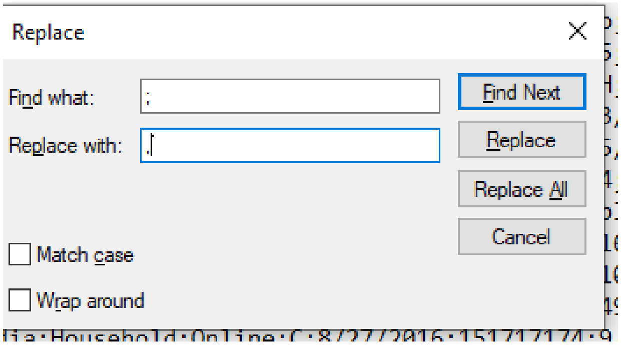
اشراف :

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: ,

- **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.

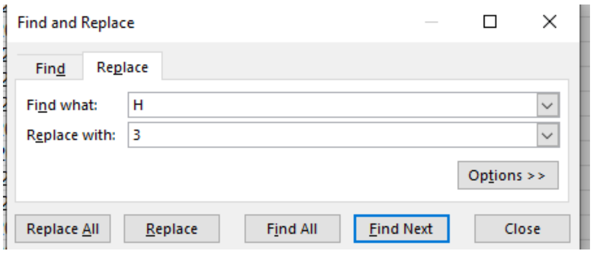
  Problem: The value "Cote d'Ivoire" includes an apostrophe that leads to errors during data analysis or query execution.

- **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.

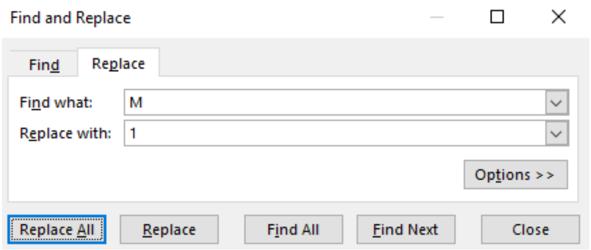
  The values were transformed as follows:

```
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

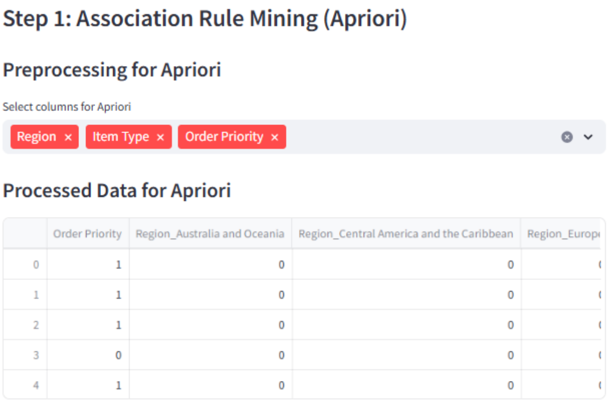


Find what: M
Replace with: 1

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.



Step 1: Association Rule Mining (Apriori)

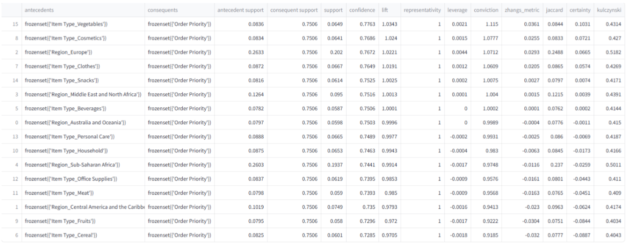
Preprocessing for Apriori

Select columns for Apriori

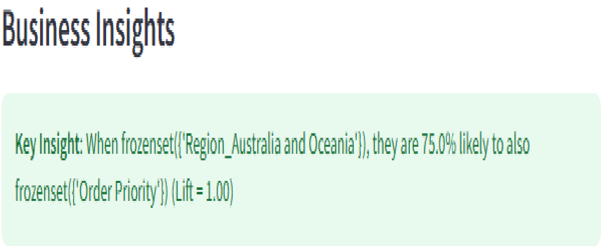
Region × Item Type × Order Priority ×

Processed Data for Apriori

	Order Priority	Region_Australia and Oceania	Region_Central America and the Caribbean	Region_Europ
0	1	0	0	(
1	1	0	0	(
2	1	0	0	(
3	0	0	0	(
4	1	0	0	(



	antecedents	consequents	antecedent support	consequent support	support	confidence	lift	representativity	leverage	conviction	zhangs_metric	jaccard	certainty	kulczynski
15	frozenset({'Item Type_Vegetables'})	frozenset({'Order Priority'})	0.0836	0.7506	0.0649	0.7763	1.0343	1	0.0021	1.115	0.0361	0.0844	0.1031	0.4314
8	frozenset({'Item Type_Cosmetics'})	frozenset({'Order Priority'})	0.0834	0.7506	0.0641	0.7686	1.024	1	0.0015	1.0777	0.0255	0.0833	0.0721	0.427
2	frozenset({'Region_Europe'})	frozenset({'Order Priority'})	0.2633	0.7506	0.202	0.7672	1.0221	1	0.0044	1.0712	0.0293	0.2488	0.0665	0.5182
7	frozenset({'Item Type_Clothes'})	frozenset({'Order Priority'})	0.0872	0.7506	0.0667	0.7649	1.0191	1	0.0012	1.0609	0.0205	0.0865	0.0574	0.4269
14	frozenset({'Item Type_Snacks'})	frozenset({'Order Priority'})	0.0816	0.7506	0.0614	0.7525	1.0025	1	0.0002	1.0075	0.0027	0.0797	0.0074	0.4171
3	frozenset({'Region_Middle East and North Africa'})	frozenset({'Order Priority'})	0.1264	0.7506	0.095	0.7516	1.0013	1	0.0001	1.004	0.0015	0.1215	0.0039	0.4391
5	frozenset({'Item Type_Beverages'})	frozenset({'Order Priority'})	0.0782	0.7506	0.0587	0.7506	1.0001	1	0	1.0002	0.0001	0.0762	0.0002	0.4144
0	frozenset({'Region_Australia and Oceania'})	frozenset({'Order Priority'})	0.0797	0.7506	0.0598	0.7503	0.9996	1	0	0.9989	-0.0004	0.0776	-0.0011	0.415
13	frozenset({'Item Type_Personal Care'})	frozenset({'Order Priority'})	0.0888	0.7506	0.0665	0.7489	0.9977	1	-0.0002	0.9931	-0.0025	0.086	-0.0069	0.4187
10	frozenset({'Item Type_Household'})	frozenset({'Order Priority'})	0.0875	0.7506	0.0653	0.7463	0.9943	1	-0.0004	0.983	-0.0063	0.0845	-0.0173	0.4166
4	frozenset({'Region_Sub-Saharan Africa'})	frozenset({'Order Priority'})	0.2603	0.7506	0.1937	0.7441	0.9914	1	-0.0017	0.9748	-0.0116	0.237	-0.0259	0.5011
12	frozenset({'Item Type_Office Supplies'})	frozenset({'Order Priority'})	0.0837	0.7506	0.0619	0.7395	0.9853	1	-0.0009	0.9576	-0.0161	0.0801	-0.0443	0.411
11	frozenset({'Item Type_Meat'})	frozenset({'Order Priority'})	0.0798	0.7506	0.059	0.7393	0.985	1	-0.0009	0.9568	-0.0163	0.0765	-0.0451	0.409
1	frozenset({'Region_Central America and the Caribbe	frozenset({'Order Priority'})	0.1019	0.7506	0.0749	0.735	0.9793	1	-0.0016	0.9413	-0.023	0.0963	-0.0624	0.4174
9	frozenset({'Item Type_Fruits'})	frozenset({'Order Priority'})	0.0795	0.7506	0.058	0.7296	0.972	1	-0.0017	0.9222	-0.0304	0.0751	-0.0844	0.4034
6	frozenset({'Item Type_Cereal'})	frozenset({'Order Priority'})	0.0825	0.7506	0.0601	0.7285	0.9705	1	-0.0018	0.9185	-0.032	0.0777	-0.0887	0.4043



Business Insights

Key Insight: When frozenset({'Region_Australia and Oceania'}), they are 75.0% likely to also frozenset({'Order Priority'}) (Lift = 1.00)

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, Item Type, country.
- **Steps:**
  - Split dataset into training (80%) and testing (20%) sets.
  - Assumed feature independence.
  - Evaluated model using accuracy, precision, recall, and F1-score.

Algorithm and Reasoning:

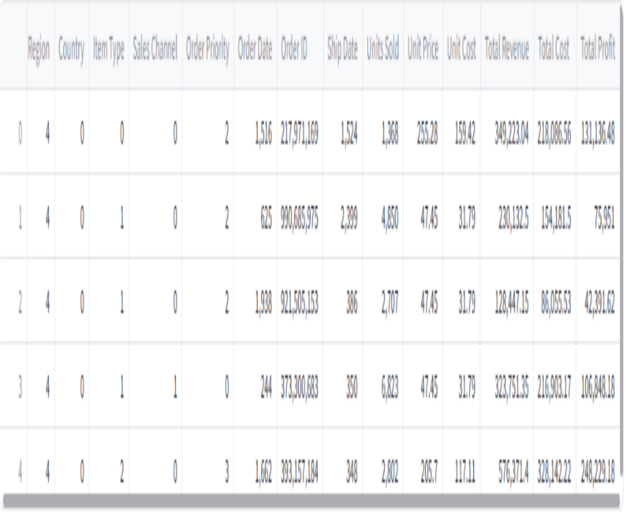
Naïve Bayes is a probabilistic classification algorithm based on Bayes' Theorem. It assumes that the presence of a particular feature in a class is unrelated to the presence of any other feature.

Relevant Columns:

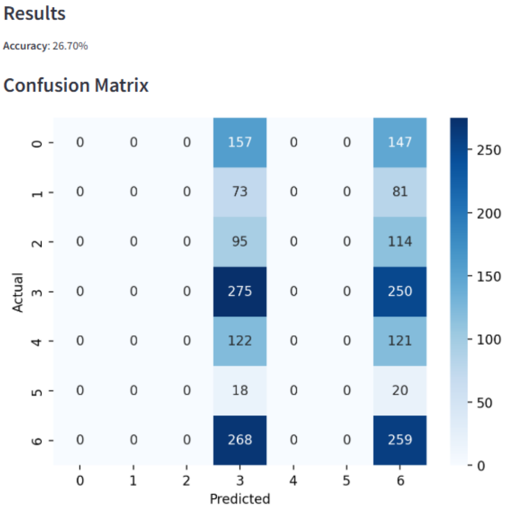
- **Region:** Useful for understanding geographic influence on purchasing behavior.
- **Item Type:** To know the types of products that are bought together.

Results:

- **Accuracy:** 36.25%.
- **Confusion Matrix:** Shows the number of correct and incorrect classifications.



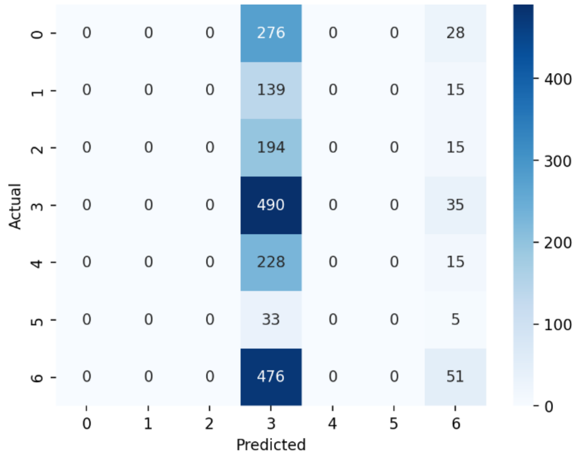
	Region	Country	Item Type	Sales Channel	Order Priority	Order Date	Order ID	Ship Date	Units Sold	Unit Price	Unit Cost	Total Revenue	Total Cost	Total Profit
0	4	0	0	0	2	1,516	217,971,169	1,524	1,368	255.28	159.42	349,223.04	218,086.56	131,136.48
1	4	0	1	0	2	625	990,685,975	2,399	4,850	47.45	31.79	230,132.5	154,181.5	75,951
2	4	0	1	0	2	1,938	921,505,153	386	2,707	47.45	31.79	128,447.15	86,055.53	42,391.62
3	4	0	1	1	0	244	373,300,683	350	6,823	47.45	31.79	323,751.35	216,903.17	106,848.18
4	4	0	2	0	3	1,662	393,157,184	348	2,802	205.7	117.11	576,371.4	328,142.22	248,229.18

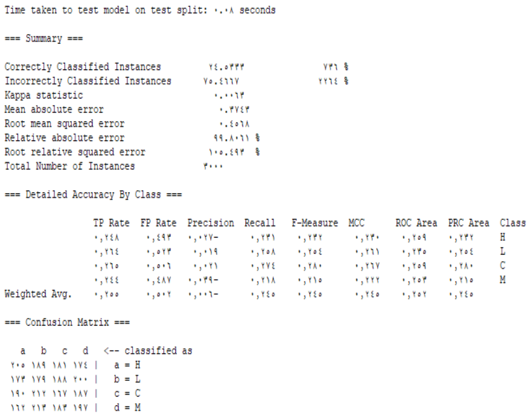


Results

Accuracy: 26.70%

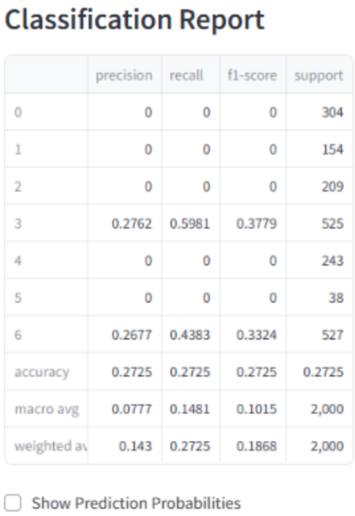
Confusion Matrix





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

=== Summary ===

Correctly Classified Instances        ٢٤.٥٣٣٣               ٧٣٦ %
Incorrectly Classified Instances     ٧٥.٤٦٦٧              ٢٢٦٤ %
Kappa statistic                        ٠.٠٠١٣
Mean absolute error                     ٠.٣٧٤٣
Root mean squared error                 ٠.٤٥١٨
Relative absolute error               ٩٩.٨٠١١ %
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
```

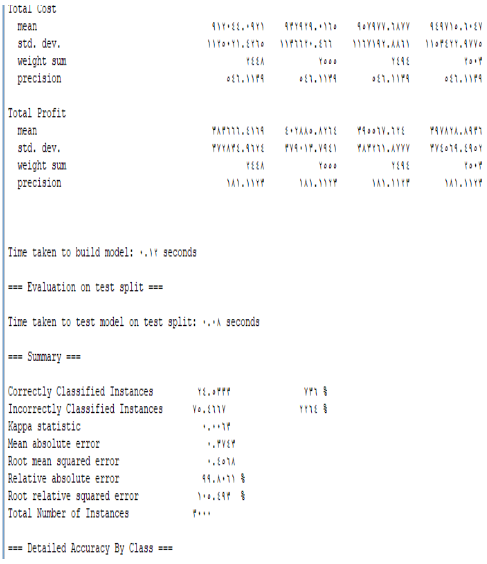
- **Classification Report:** Includes Precision, Recall, F1-Score, and Support.



Classification Report

	precision	recall	f1-score	support
0	0	0	0	304
1	0	0	0	154
2	0	0	0	209
3	0.2762	0.5981	0.3779	525
4	0	0	0	243
5	0	0	0	38
6	0.2677	0.4383	0.3324	527
accuracy	0.2725	0.2725	0.2725	0.2725
macro avg	0.0777	0.1481	0.1015	2,000
weighted a	0.143	0.2725	0.1868	2,000

☐ Show Prediction Probabilities



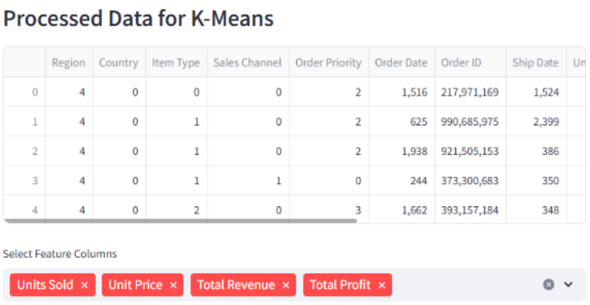
```
Total Cost
  mean                          ٩١٢٠٤٤.٠٩٢١    ٩٣٢٩٢٩.٠١١٥    ٩٥٧٩٧٧.٦٨٧٧    ٩٤٩٧١٥.٦٠٤٧
  std. dev.                    ١١٢٥٠٢١.٤٢٦٥    ١١٣٦٦٢٠.٤٦٦   ١١٦٧١٩٢.٨٨٦١   ١١٥٣٤٢٢.٩٧٧٥
  weight sum                           ٢٤٤٨           ٢٥٥٥           ٢٤٩٤           ٢٥٠٣
  precision                       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        ٢٤.٥٣٣٣               ٧٣٦ %
Incorrectly Classified Instances     ٧٥.٤٦٦٧              ٢٢٦٤ %
Kappa statistic                        ٠.٠٠١٣
Mean absolute error                     ٠.٣٧٤٣
Root mean squared error                 ٠.٤٥١٨
Relative absolute error               ٩٩.٨٠١١ %
Root relative squared error          ١٠٥.٤٩٣  %
Total Number of Instances            ٣٠٠٠

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

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.

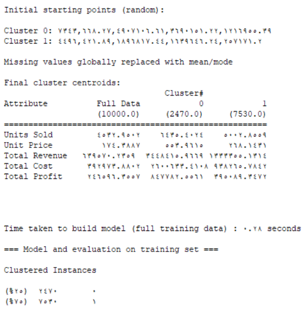


Processed Data for K-Means

	Region	Country	Item Type	Sales Channel	Order Priority	Order Date	Order ID	Ship Date	Un
0	4	0	0	0	2	1,516	217,971,169	1,524	
1	4	0	1	0	2	625	990,685,975	2,399	
2	4	0	1	0	2	1,938	921,505,153	386	
3	4	0	1	1	0	244	373,300,683	350	
4	4	0	2	0	3	1,662	393,157,184	348	

Select Feature Columns

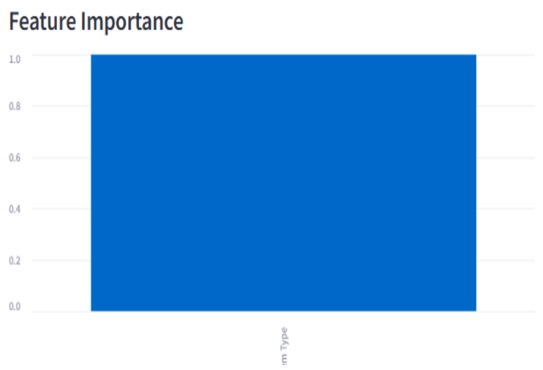
Units Sold × Unit Price × Total Revenue × Total Profit ×

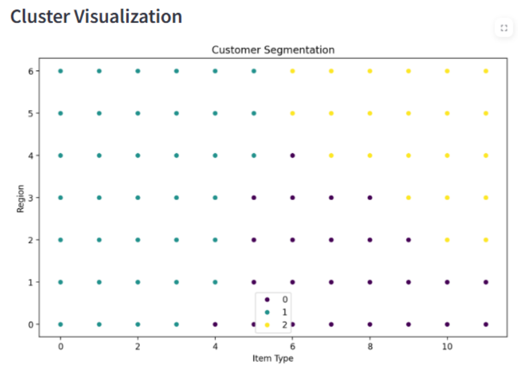


```
Initial starting points (random):

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
```

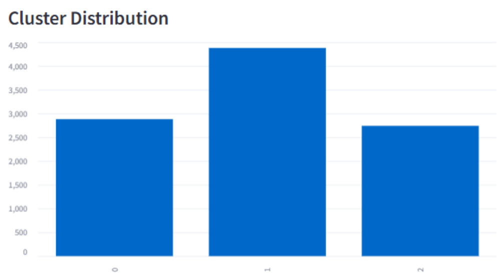
Silhouette Score: 0.60

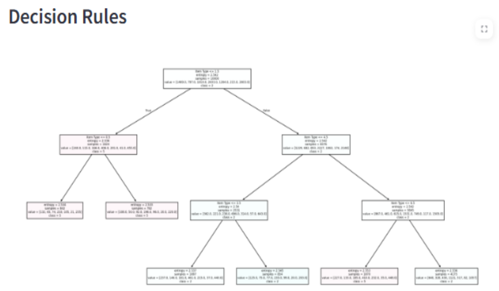
Evaluation Metrics





Results and Insights


Cluster Distribution


Decision Rules
True
False