Documentation Data Mining

by:

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

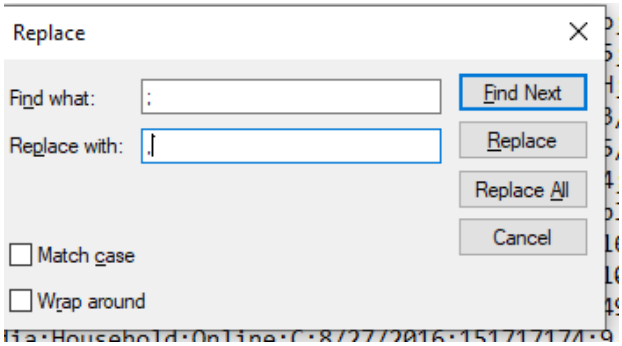
اشراف :

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 (,).



- **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.

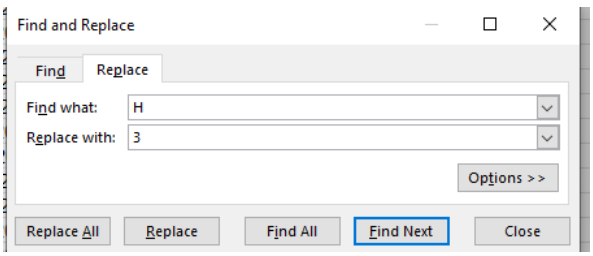
  Problem: The value "Cote d'Ivoire" includes an apostrophe that leads to errors during data analysis or query execution.

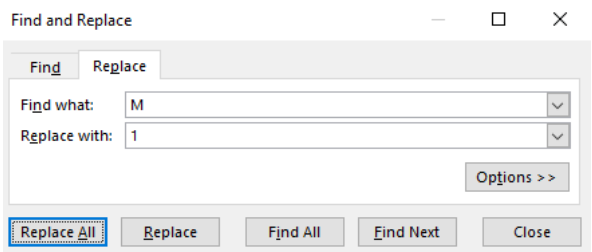
- **Step** 3: Re-encoding the "Order Priority" Column We performed re-encoding of the textual values in the "Order Priority" column into numerical values to facilitate processing by machine learning algorithms.

  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.





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)

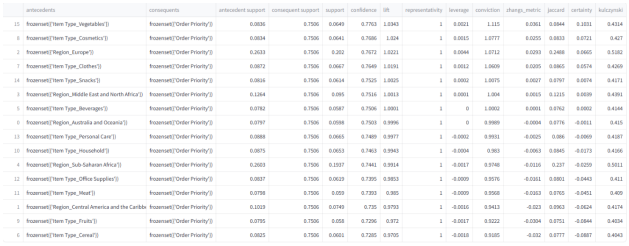
Preprocessing for Apriori

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_Europe
0	1	0	0	0
1	1	0	0	0
2	1	0	0	0
3	0	0	0	0
4	1	0	0	0



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:

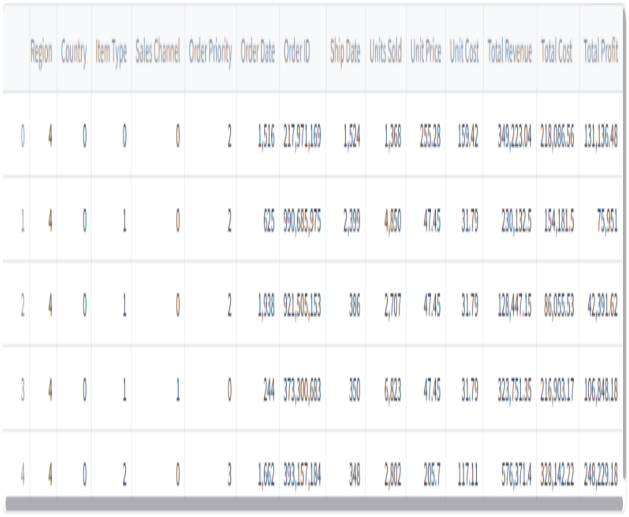
The Apriori algorithm is used to identify frequent itemsets in a dataset and generate association rules. It helps in discovering interesting relationships between variables in large databases.

Parameters:

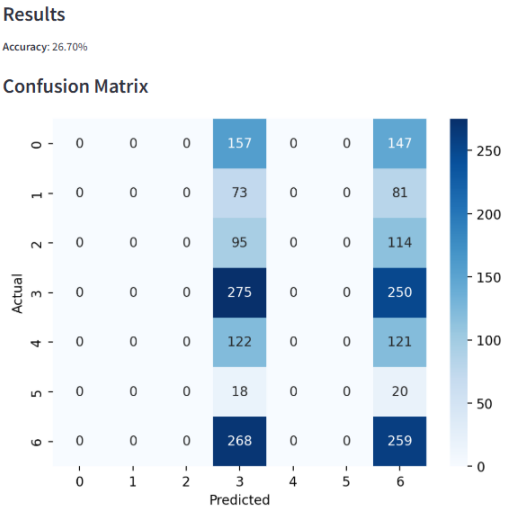
- **Support Threshold:** A minimum percentage of records in the dataset that contain the itemset. In our case, with 10,000 records, a support threshold of 0.05 means any itemset must appear in at least 500 transactions to be considered significant.
- **Confidence Threshold:** The likelihood that a rule 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, 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.



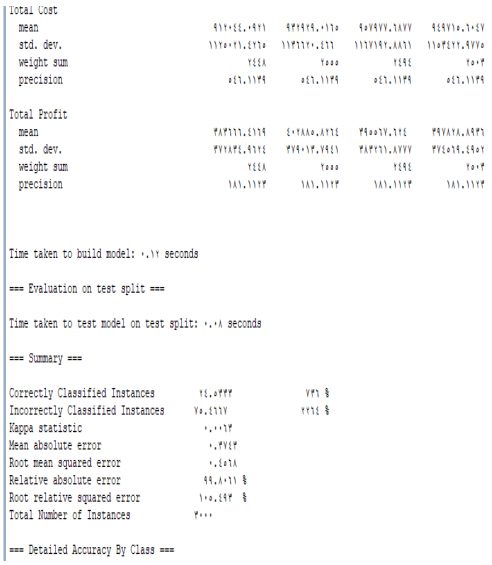
	Region	Country	Item Type	Sales Channel	Order Priority	Order Date	Order ID	Ship Date	Units Sold	Unit Price	Unit Cost	Total Revenue	Total Cost	Total Profit
0	4	0	0	0	2	1,516	217,971,169	1,524	1,366	255.28	159.42	348,723.04	218,006.56	130,716.48
1	4	0	1	0	2	635	990,665,975	2,199	4,850	47.45	31.79	230,132.5	154,181.5	75,951
2	4	0	1	0	2	1,938	921,505,153	186	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	346	2,802	205.7	117.11	576,371.4	328,142.22	248,229.18



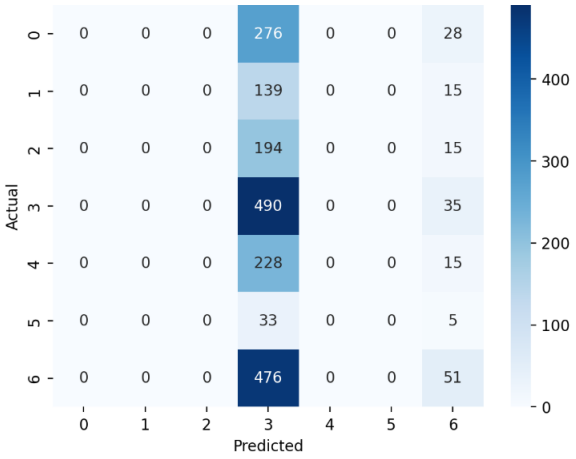
Classification Report

	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 av	0.143	0.2725	0.1868	2,000

☐ Show Prediction Probabilities



```
Total Cost
  mean                    ٩١٦٠٤٤.٠٩٢١    ٩٣٩٩٢٩.٠١١٥    ٩٥٧٩٧٧.١٨٧٧    ٩٤٩٧١٥.١٠٤٧
  std. dev.               ١١٦٥٠٣١.٤٦٦٥   ١١٣٩٦٢٠.٤٦٦    ١١٦٧١٩٢.٨٨٦١   ١١٥٣٤٢٢.٩٧٧٥
  weight sum                      ٢٤٨             ٢٥٥٥             ٢٤٩٤             ٢٥٠٣
  precision               ٥٤٦.١١٣٩       ٥٤٦.١١٣٩       ٥٤٦.١١٣٩       ٥٤٦.١١٣٩

Total Profit
  mean                    ٣٨٣١٦١.٤١٦٩    ٤٠٢٨٨٥.٨٢٦٤    ٣٩٥٥٦٧.٢٦٤     ٣٩٧٨٢٨.٨٩٣٦
  std. dev.               ٣٧٢٨٤.٩٦٢٤     ٣٧٩٠٧٣.٧٩٤١    ٣٨٣٢٦١.٨٢٧٧    ٣٧٤٥٦٩.٤٩٥٢
  weight sum                      ٢٤٨             ٢٥٥٥             ٢٤٩٤             ٢٥٠٣
  precision               ١٨٦.١١٣٣       ١٨٦.١١٣٣       ١٨٦.١١٣٣       ١٨٦.١١٣٣


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               ٢٦٦٤ %
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         ٢٤.٥٣٣٣               ٧٣٦ %
Incorrectly Classified Instances       ٧٥.٤٦٦٧               ٢٦٦٤ %
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    ٠,٢٧٧-     ٠,٢٣١    ٠,٢٥١     ٠,٢٢٠-    ٠,٣٠٩     ٠,٢٢٠     H
               ٠,٢٦٤    ٠,٥٢٣    ٠,٠٩٩      ٠,٢٠٨    ٠,٢٠٤     ٠,٢٦١     ٠,٣٣٥     ٠,٢٠٤     L
               ٠,٢٦٥    ٠,٥٠٦    ٠,٠٢٦      ٠,٢٧٤    ٠,٢٨٠     ٠,٢٦٧     ٠,٢٥٩     ٠,٢٨٠     C
               ٠,٢٤٤    ٠,٤٨٧    ٠,٠٣٤-     ٠,٢٦٨    ٠,٢٦٥     ٠,٢٢٢     ٠,٢٥٣     ٠,٢٦٥     M
Weighted Avg.  ٠,٢٥٥    ٠,٥٠٢    ٠,٠٠٦-     ٠,٢٤٥    ٠,٢٤٥     ٠,٢٥٠     ٠,٢٥٢     ٠,٢٤٥

=== Confusion Matrix ===

   a   b   c   d   <-- classified as
٢٠٥ ١٨٩ ١٨١ ١٢٤ |   a = H
١٧٢ ١٧٩ ١٨٨ ٢٠٠ |   b = L
١٩٠ ٢١٢ ١٦٦ ١٨٧ |   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:

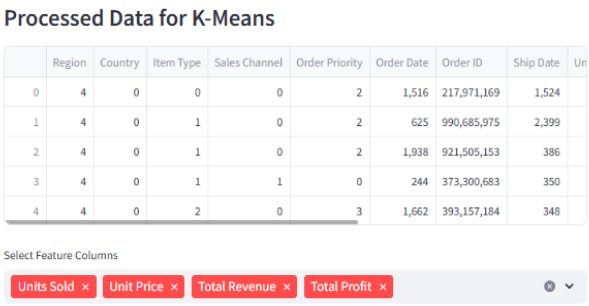
- **Region:** Useful for understanding geographic influence on purchasing behavior.
- **Country:** To identify differences between countries.
- **Item Type:** To know the types of products that are bought together.
- **Sales Channel:** To understand the differences between various sales channels (online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

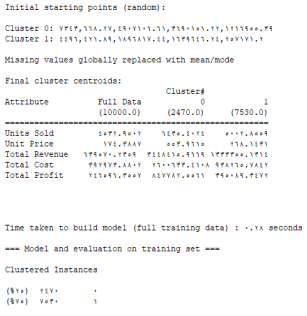
ID3 Algorithm (Decision Trees)

- **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)

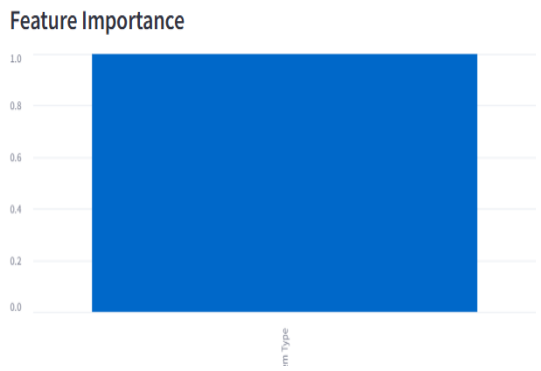
- **Objective:** Partition the data into clusters based on similarity.
- **Selected Columns:** Units Sold, Unit Price, Total Revenue, Total Profit.
- **Parameters:**
  - Number of Clusters (K): Determined using the elbow method.
  - Initialization: Used k-means++ to enhance convergence.
- **Steps:**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

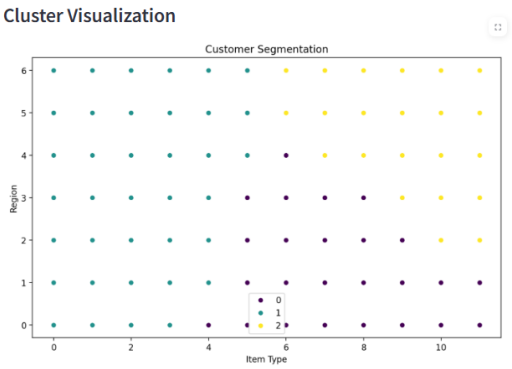




Silhouette Score: 0.60

Evaluation Metrics





Results and Insights