Documentation

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:** Checking for Outliers, Missing Values, and Special Characters
  Objective: Ensure the data quality by identifying and handling the following issues:

    - **Outliers:** Perform statistical analysis (such as interquartile range or z-score) to detect extreme data points that may affect model performance.
    - **Missing Values:** Identify missing entries and handle them through appropriate methods, such as imputation (mean, median, or mode) or row removal when necessary.
    - **Special Characters:** Detect and clean special characters such as the apostrophe in entries like "Cote d'Ivoire" to avoid parsing errors. These can be replaced by suitable alternatives (e.g., removing the apostrophe or using escape sequences).

  Steps Taken:

    - Applied functions to identify and visualize outliers.
    - Used conditional checks to replace or remove missing values.
    - Implemented a find-and-replace mechanism for problematic characters.

Algorithm Implementation

Apriori Algorithm (Association Rule Mining)

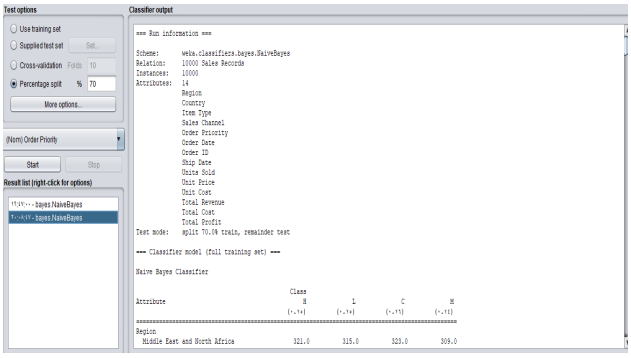
- **Objective:** Discover frequent itemsets and generate association rules.
- **Selected Columns:** Item Type, Sales Channel.
- **Parameters:**
    - Support Threshold: A reasonable value based on dataset characteristics.
    - Confidence Threshold: Set a meaningful value to filter rules.
    - Lift: 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: ..\ (1000 instances)
Minimum metric <confidence>: ..*
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:

 \. Order Priority=C 2555 ==> Sales Channel=Online 1326    <conf:(..*\)> lift:(\..*) lev:(.) [32] conv:(\..*)
 *. Order Priority=L 2494 ==> Sales Channel=Online 1270    <conf:(..*\)> lift:(\..*) lev:(.) [7] conv:(\..*)
```

```
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)
```

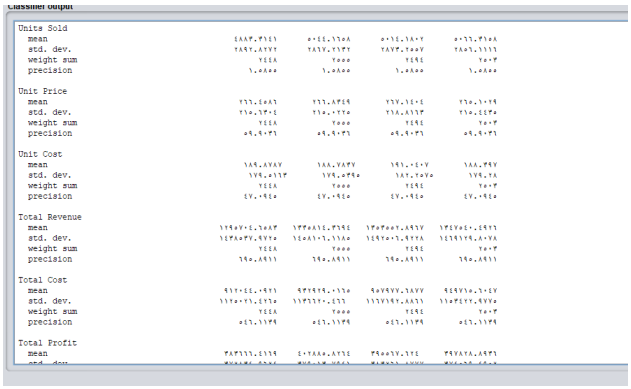
Naïve Bayes (Classification)

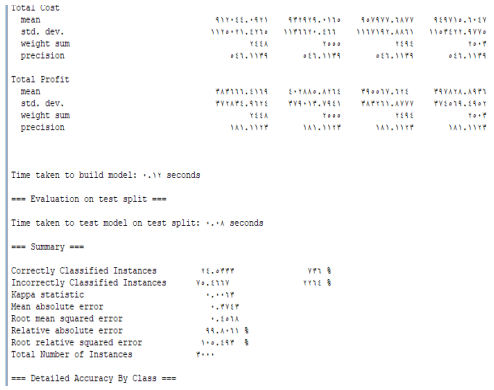
- **Objective:** Build a probabilistic model to classify data into predefined classes.
- **Selected Columns:** Region, Country, Item Type, Sales Channel, Order Priority.
- **Steps:**
  - Split dataset into training (70%) and testing (30%) sets.
  - Assumed feature independence.
  - Evaluated model using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



Naive Bayes Classifier

Attribute	Class H (٠.٢٥)	L (٠.٢٥)	C (٠.٢٦)	M (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






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

=== Summary ===

Correctly Classified Instances         ٢٤.٥٣٣٣               ٧٣١ %
Incorrectly Classified Instances      ٧٥.٤٦٧٧               ٢٦٩٤ %
Kappa statistic                          ٠.٠٠٦٣
Mean absolute error                      ٠.٣٧٤٣
Root mean squared error                  ٠.٤٠٦٨
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     ٠,٢٧٠       ٠,٢٣١    ٠,٢٣٧     ٠,٢٣٠    ٠,٢٠٩     ٠,٢٣٧     H
               ٠,٢٦٤     ٠,٥٠٣     ٠,٢٦٣       ٠,٢٥٨    ٠,٢٥٤     ٠,٢٦١    ٠,٢٠٥     ٠,٢٤٤     L
               ٠,٢٦٥     ٠,٥٠٦     ٠,٢٦١       ٠,٢٧٤    ٠,٢٦٨     ٠,٢٦٧    ٠,٢٠٩     ٠,٢٤٨     C
               ٠,٢٤٤     ٠,٤٨٧     ٠,٢٣٩       ٠,٢٦٨    ٠,٢٦٠     ٠,٢٦٠    ٠,٢٠٥     ٠,٢٤٠     M
Weighted Avg.  ٠,٢٥٥     ٠,٥٠٢     ٠,٢٥٠٦       ٠,٢٥٥    ٠,٢٥٥     ٠,٢٥٥    ٠,٢٠٧     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
```

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

- Used metrics such as accuracy, precision, recall, F1-score, Silhouette Score, and Inertia.

Results and Insights

- Visualizations provided insights into sales trends, classification outcomes, and cluster formations.
- Performance scores highlighted the effectiveness of different algorithms for the dataset.