

Final Project: Bank Customer Information and Marketing Response

Team Members: Puja Shah & Sanjida Chowdhury

Introduction of the Data

Bank Customer Information and Marketing Response

Weka Explorer

Preprocess Cluster Associate Select attributes Visualize

Open file... Open URL... Open DB... Generate... Undo Edit... Save...

Filter
Choose **None** Apply Stop

Current relation
Relation: bank
Instances: 4521 Attributes: 17 Sum of weights: 4521

Attributes
All None Invert Pattern

No.	Name
1	<input type="checkbox"/> age
2	<input type="checkbox"/> job
3	<input type="checkbox"/> marital
4	<input type="checkbox"/> education
5	<input type="checkbox"/> default
6	<input type="checkbox"/> balance
7	<input type="checkbox"/> housing
8	<input type="checkbox"/> loan
9	<input type="checkbox"/> contact
10	<input type="checkbox"/> day
11	<input type="checkbox"/> month
12	<input type="checkbox"/> duration
13	<input type="checkbox"/> campaign
14	<input type="checkbox"/> pdays
15	<input type="checkbox"/> previous
16	<input type="checkbox"/> poutcome
17	<input type="checkbox"/> y

Remove

Selected attribute
Name: age
Missing: 0 (0%) Distinct: 67 Type: Numeric
Unique: 4 (0%)

Statistic	Value
Minimum	19
Maximum	87
Mean	41.17
StdDev	10.576

Class: y (Nom) Visualize All

Age Range	Frequency
14-15	29
16-17	68
18-19	274
20-21	423
22-23	417
24-25	529
26-27	289
28-29	277
30-31	256
32-33	336
34-35	222
36-37	213
38-39	271
40-41	161
42-43	165
44-45	156
46-47	70
48-49	15
50-51	15
52-53	7
54-55	19
56-57	10
58-59	9
60-61	8
62-63	13
64-65	1
66-67	5
68-69	2

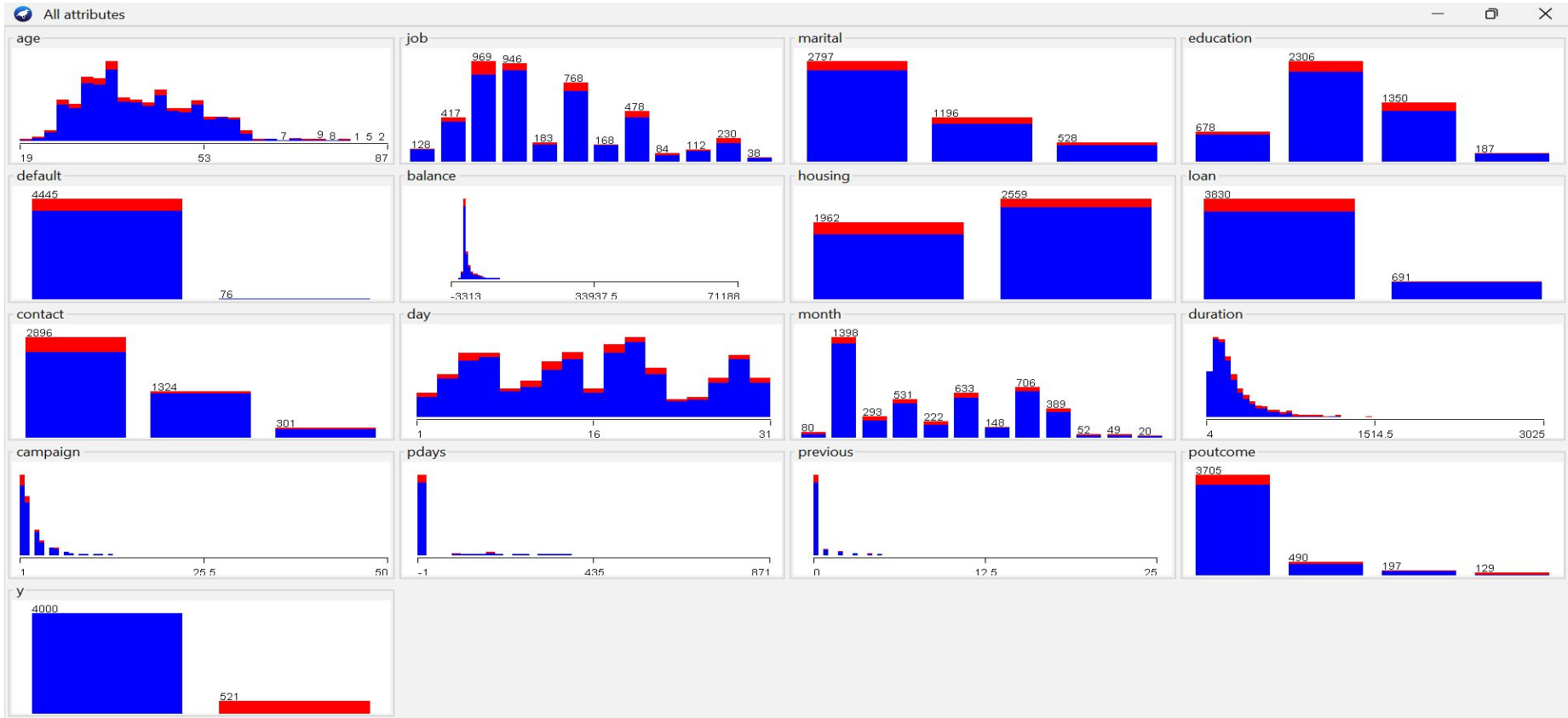
Description of the Data

- The “Bank Customer Information and Marketing Response” dataset sourced from Kaggle. The link to the data set is here:
<https://www.kaggle.com/datasets/zain280/bank-customer-information-and-marketing-response>
- This dataset offers a glimpse into bank customers' behavior and their responses to marketing campaigns.
- It includes various details like customer demographics, financial status, and interactions with marketing efforts.
- The primary aim is to analyze the factors impacting customers' subscription to a term deposit and the success of marketing strategies.
- Number of Attributes: 17
- Number of Objects/Rows: 4521
- Numeric attributes: 'age', 'balance', 'day', 'duration', 'campaign', 'pdays', 'previous'
- Categorical attributes: 'job', 'marital', 'education', 'default', 'housing', 'loan', 'contact', 'month', 'poutcome', 'y'

Description of Main Attributes/Features

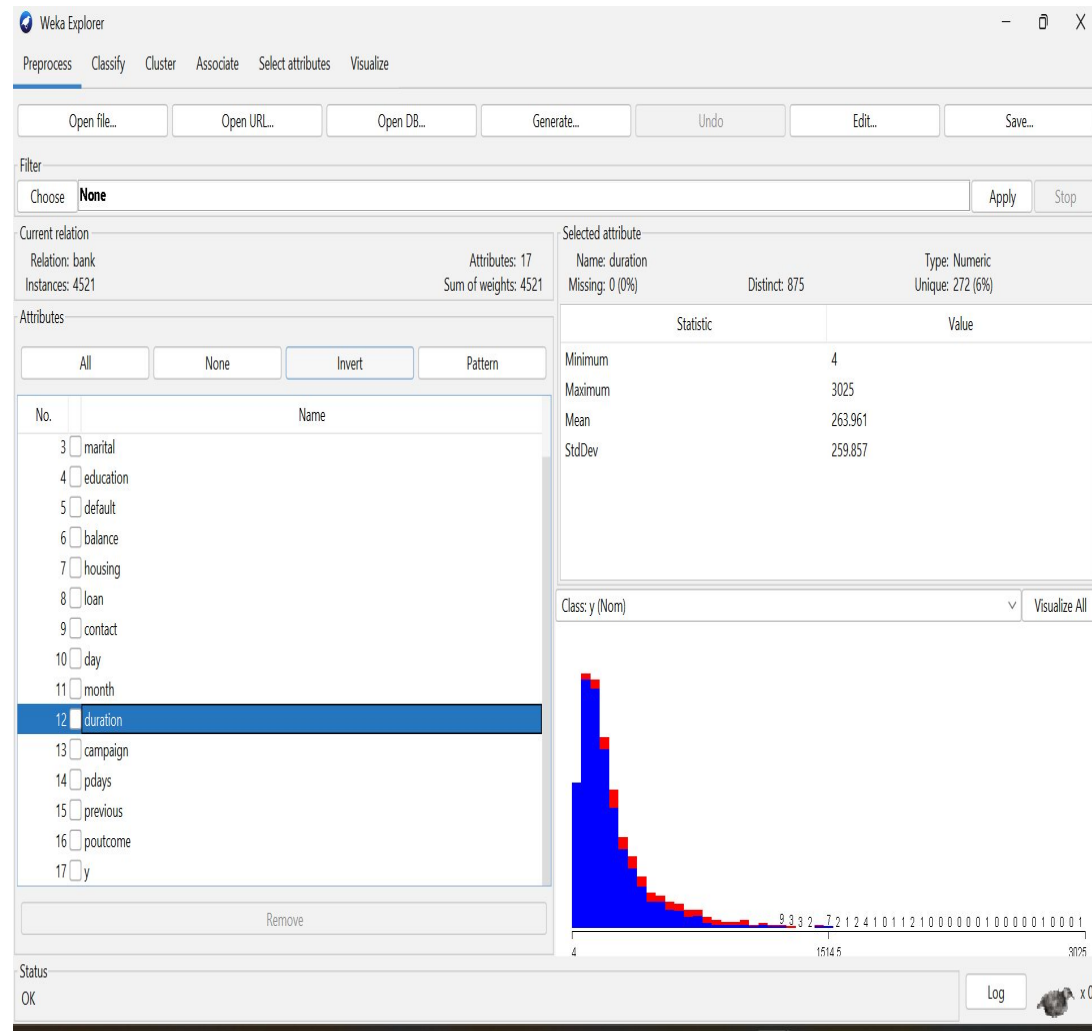
1. Demographic Information:
 - age: Age of the customer (numeric)
 - job: Occupation of the customer (categorical: 'admin.', 'blue-collar', 'entrepreneur', 'housemaid', 'management', 'retired', 'self-employed', 'services', 'student', 'technician', 'unemployed', 'unknown')
 - marital: Marital status of the customer (categorical: 'divorced', 'married', 'single')
 - education: Education level of the customer (categorical: 'primary', 'secondary', 'tertiary', 'unknown')
2. Financial Information:
 - default: Whether the customer has credit in default (binary: 'yes', 'no')
 - balance: Balance in the customer's account (numeric)
 - housing: Whether the customer has a housing loan (binary: 'yes', 'no')
 - loan: Whether the customer has a personal loan (binary: 'yes', 'no')
3. Marketing Campaign Information:
 - contact: Communication contact type (categorical: 'cellular', 'telephone', 'unknown')
 - day: Last contact day of the month (numeric)
 - month: Last contact month of the year (categorical: 'jan', 'feb', 'mar', ..., 'nov', 'dec')
 - duration: Duration of the last contact in seconds (numeric)
 - campaign: Number of contacts performed during this campaign and for this client (numeric)
 - pdays: Number of days that passed by after the client was last contacted from a previous campaign (numeric)
 - previous: Number of contacts performed before this campaign and for this client (numeric)
 - poutcome: Outcome of the previous marketing campaign (categorical: 'failure', 'success', 'unknown', 'other')
4. Response to Marketing Campaign (Target Variable):
 - y: Whether the customer subscribed to a term deposit (binary: 'yes', 'no')

Visualization



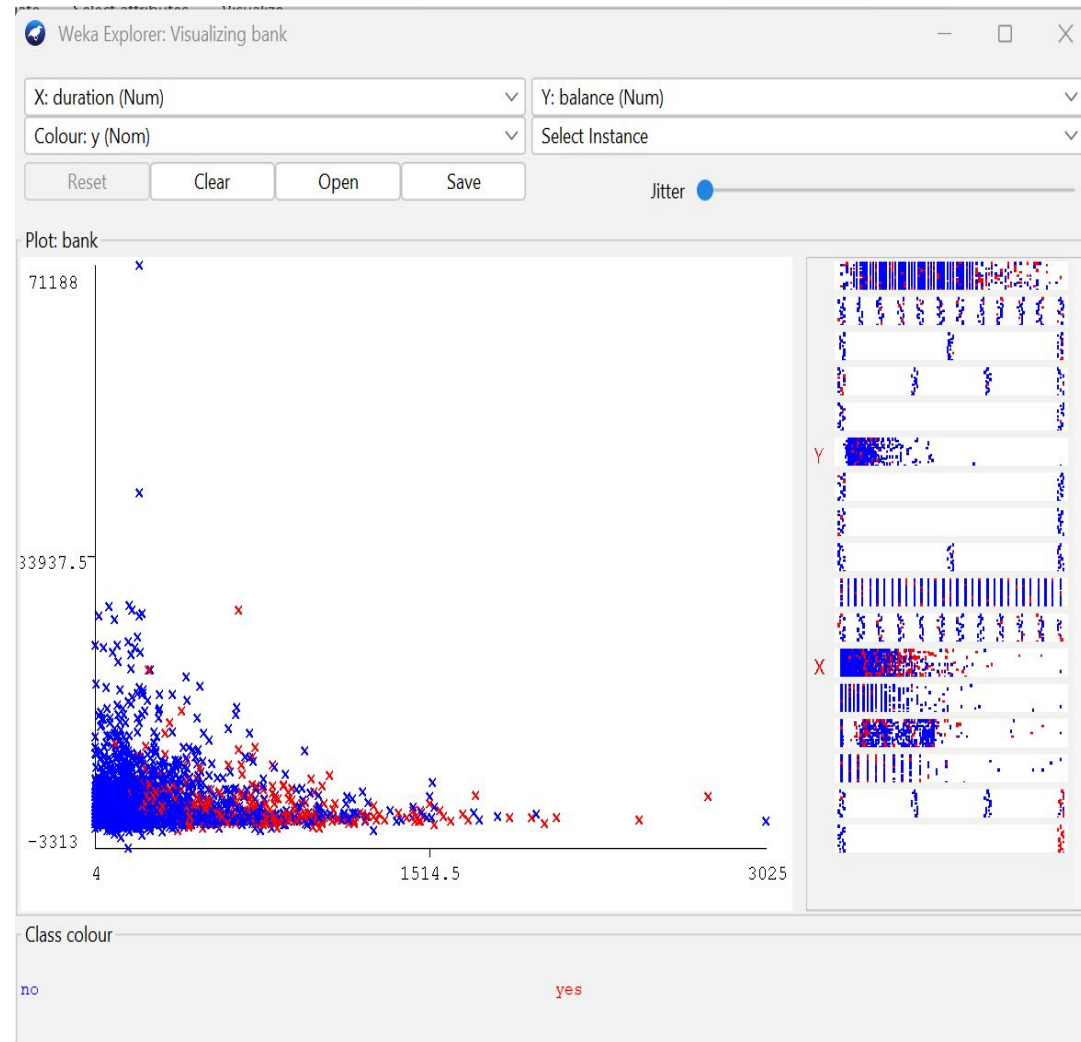
Visualization

- **Range:** 'Duration' ranges from 4 seconds to 3025 seconds, indicating a wide variation in the length of interactions between the bank and its customers during marketing campaigns.
- **Mean:** The average duration of interactions is 263.961 seconds, suggesting that most interactions are of moderate length.
- **Standard Deviation:** With a standard deviation of 259.857 seconds, there's a considerable spread of data around the mean, indicating variability in the duration of interactions.
- **Interpretation:** The distribution of 'Duration' likely represents the different types of interactions customers have with the bank during marketing campaigns. Longer durations may indicate more engaging interactions, potentially leading to higher subscription rates. Analyzing this attribute can provide insights into the effectiveness of marketing strategies and customer engagement levels.



Visualization

- Due to the high standard deviation, we can expect a considerable spread of points in the scatter plot. This spread could signify variability in the duration of interactions across customers with different balance levels.
- The presence of outliers, especially customers with extremely high or low balances, may impact the scatter plot shape and the perceived relationship between 'Balance' and 'Duration'

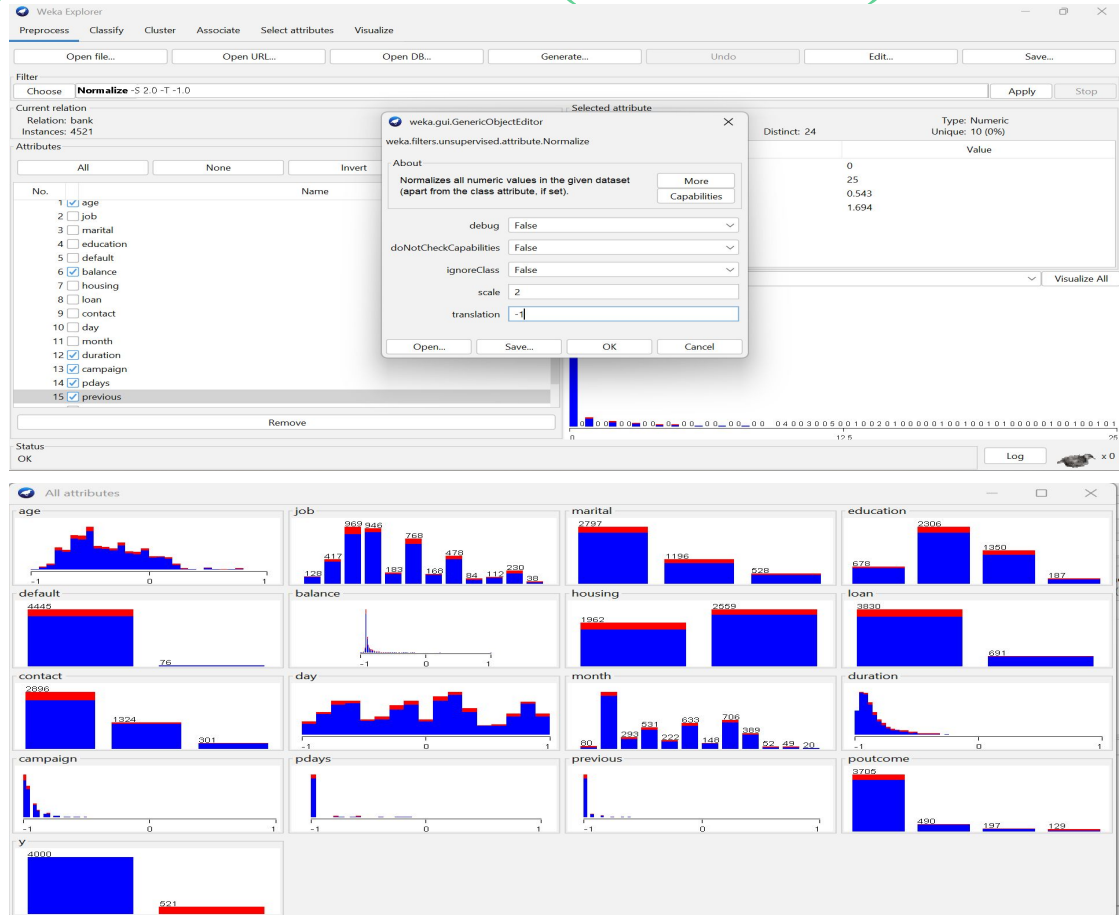


Goal

- The goal of the analysis is to perform classification to predict whether bank customers will subscribe to a term deposit or not, based on their demographic, financial, and campaign interaction data.
- This task is essential for banks to optimize marketing strategies, improve subscription rates, and allocate resources effectively.
- Additionally, cluster analysis will be conducted to identify distinct customer segments, allowing for personalized marketing communication and better targeting of specific customer groups.
- The analysis aims to enhance subscription outcomes and customer engagement through data-driven insights.

Pre-processing - Normalization (min-max)

- Normalization Applied: Numerical features ('age', 'balance', 'day', 'duration', 'campaign', 'pdays', and 'previous') have been normalized using Min-Max scaling.
- Range Standardization: Feature values are now uniformly scaled within the range of -1 to 2, ensuring consistent magnitudes across all attributes.



Classification: J48 Decision Tree

We chose 'Age', 'Balance', 'Duration', and 'Poutcome' as attributes for the classification model because they provide critical insights into financial behavior, customer engagement, and past campaign responses, which are key determinants of term deposit subscriptions.

Age: Reflects life stage and financial behavior.

Balance: Indicates financial capacity and stability.

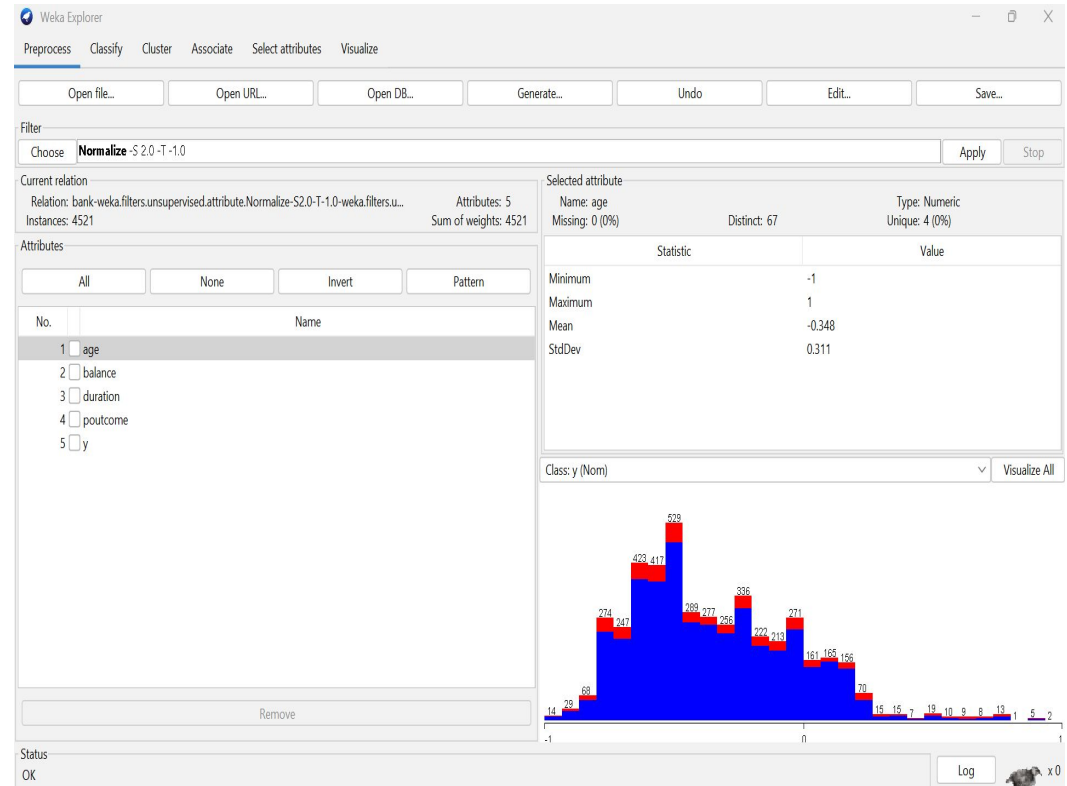
Duration: Shows engagement level during the marketing campaign.

Poutcome: Previous campaign outcome influences current response.

The target attribute 'Y' (Subscription to Term Deposit) is essential as it represents the outcome we aim to predict.

We removed other attributes to simplify the model and focus on the most influential predictors, reducing the risk of overfitting and ensuring a more targeted analysis.

Attributes to Train the Model:



Classification: J48 Decision Tree

Weka Explorer

Preprocess **Classify** Cluster Associate Select attributes Visualize

Classifier
Choose **J48 -C 0.25 -M 2**

Test options
☐ Use training set
☐ Supplied test set Set...
☒ Cross-validation Folds **10**
☐ Percentage split % **66**
More options...

(Nom) y

Start Stop

Result list (right-click for options)
21:14:02 - trees.J48

Classifier output

=== Run information ===

Scheme: weka.classifiers.trees.J48 -C 0.25 -M 2
Relation: bank-weka.filters.unsupervised.attribute.Normalize-S2.0-T-1.0-weka.filters.unsupervised.attribute.Remove-S2-3,5,8-11,14-weka
Instances: 4521
Attributes: 5
age
balance
duration
poutcome
y

Test mode: 10-fold cross-validation

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

J48 pruned tree

duration <= -0.862959: no (2548.0/73.0)
duration > -0.862959
| duration <= -0.575637
| | poutcome = unknown
| | | age <= 0.205882: no (1253.0/131.0)
| | | age > 0.205882: yes (37.0/17.0)
| | poutcome = failure: no (174.0/35.0)
| | poutcome = other: no (74.0/22.0)
| | poutcome = success: yes (76.0/16.0)
| duration > -0.575637
| | poutcome = unknown

Status
OK

Log x0

Weka Explorer

Preprocess **Classify** Cluster Associate Select attributes Visualize

Classifier
Choose **J48 -C 0.25 -M 2**

Test options
☐ Use training set
☐ Supplied test set Set...
☒ Cross-validation Folds **10**
☐ Percentage split % **66**
More options...

(Nom) y

Start Stop

Result list (right-click for options)
21:14:02 - trees.J48

Classifier output

size of the tree : 25

Time taken to build model: 0 seconds

=== Stratified cross-validation ===
=== Summary ===

Correctly Classified Instances	4016	88.8299 %
Incorrectly Classified Instances	505	11.1701 %
Kappa statistic	0.3297	
Mean absolute error	0.1563	
Root mean squared error	0.2872	
Relative absolute error	76.6024 %	
Root relative squared error	89.9339 %	
Total Number of Instances	4521	

=== Detailed Accuracy By Class ===

	TP Rate	FP Rate	Precision	Recall	F-Measure	MCC	ROC Area	PRC Area	Class
	0.964	0.695	0.914	0.964	0.939	0.345	0.792	0.953	no
	0.305	0.036	0.526	0.305	0.386	0.345	0.792	0.400	yes
Weighted Avg.	0.888	0.619	0.870	0.888	0.875	0.345	0.792	0.889	

=== Confusion Matrix ===

a b <-- classified as

3857	143	a = no
362	159	b = yes

Status
OK

Log x0

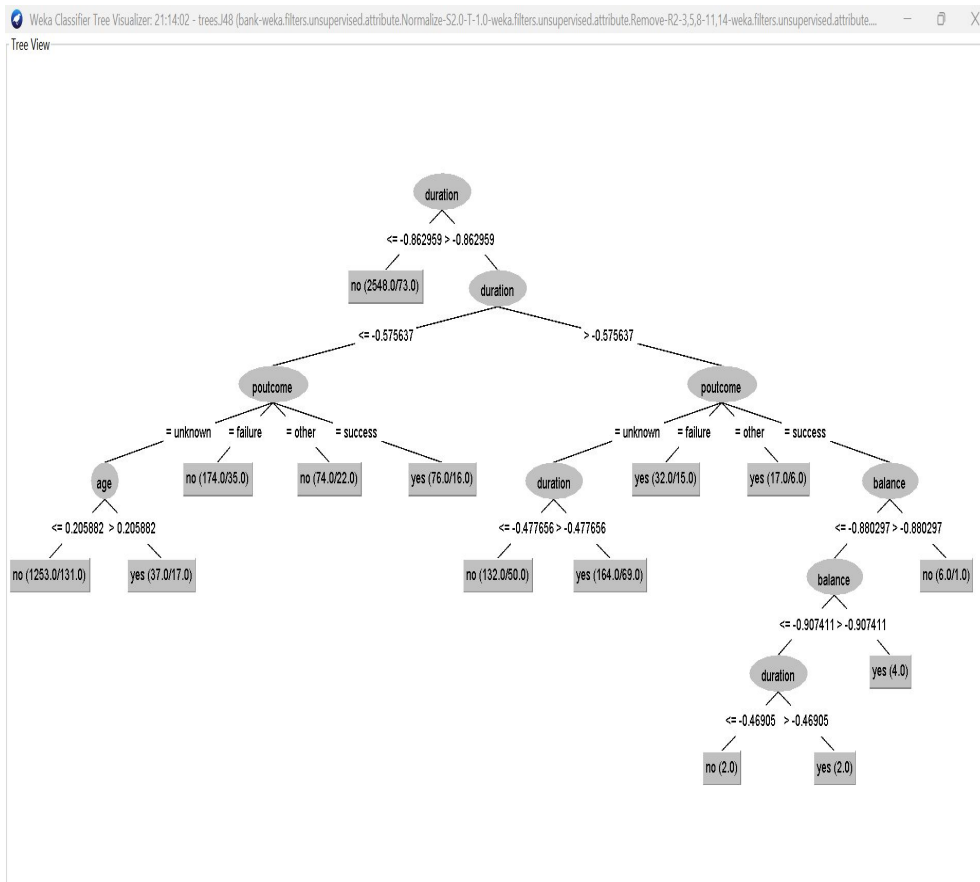
Classification: J48 Decision Tree

Interpreting the model:

- The tree consists of 14 leaves and 23 nodes, indicating a relatively simple structure.
- The root node splits the data based on the duration of the last contact.
- If the duration is less than or equal to -0.862959 (normalized value), the model predicts 'no' subscription.
- If the duration is greater than -0.862959, further splits occur based on 'poutcome', 'age', 'balance', and 'duration'.

Analysis of the goal:

- The model achieved an accuracy of 88.83%, with a higher true positive rate for 'no' subscriptions compared to 'yes'.
- Precision, recall, and F-measure are higher for 'no' subscriptions, indicating better model performance in predicting negative instances.



Conclusion of the Results: J48 Decision Tree

- The J48 decision tree model achieved an accuracy of 88.83%, with 505 instances incorrectly classified.
- Precision, measuring the model's ability to correctly identify positive predictions, was 52.6% for 'yes' subscriptions and 91.4% for 'no' subscriptions.
- Recall, indicating the model's ability to capture all positive instances, was 30.5% for 'yes' subscriptions and 96.4% for 'no' subscriptions

Classification: Naive Bayes

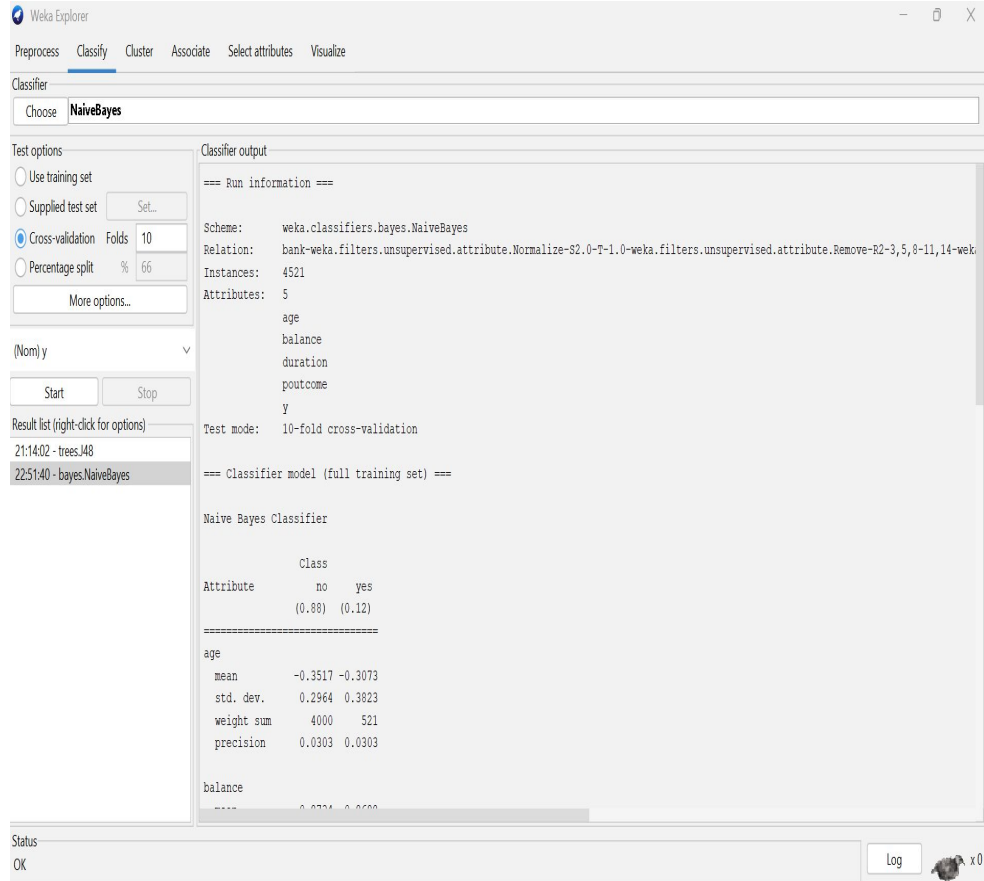
Using the same set of attributes with Naive Bayes as with J48 decision tree.

Interpreting the model:

- Naive Bayes calculates the likelihood of each class (subscribing or not subscribing) given the attribute values.
- For each attribute, the model provides the mean and standard deviation for both classes (yes and no).
- The precision metric measures the proportion of correctly predicted instances for each class.

Analysis of the goal:

- The model's precision and recall for the "yes" class are lower compared to the "no" class, indicating a higher false positive rate.



The screenshot shows the Weka Explorer application window. The 'Classifier' tab is selected, and 'NaiveBayes' is chosen. The 'Test options' section shows 'Cross-validation' with 'Folds' set to 10. The 'Classifier output' section displays the following information:

```
=== Run information ===

Scheme:      weka.classifiers.bayes.NaiveBayes
Relation:    bank-weka.filters.unsupervised.attribute.Normalize-S2.0-T-1.0-weka.filters.unsupervised.attribute.Remove-R2-3,5,8-11,14-weka
Instances:   4521
Attributes:  5
  age
  balance
  duration
  poutcome
  y
Test mode:   10-fold cross-validation

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

Naive Bayes Classifier

      Class
Attribute  no    yes
(0.88) (0.12)
=====
age
mean      -0.3517 -0.3073
std. dev.  0.2964  0.3823
weight sum 4000    521
precision  0.0303  0.0303

balance
mean      0.8734  0.8600
```

The 'Result list' at the bottom shows two entries: '21:14:02 - trees.J48' and '22:51:40 - bayes.NaiveBayes', with the latter being selected.

Classification: Naive Bayes

- The Naive Bayes classifier achieved an accuracy of 89.38%, with 4041 instances correctly classified out of 4521. However, it misclassified 480 instances, leading to an error rate of 10.62%. The Kappa statistic, measuring agreement between observed and expected classifications, indicates moderate agreement at 0.3775.
- Analyzing detailed accuracy by class, the model demonstrates higher precision and recall for the 'no' class compared to the 'yes' class. Specifically, the precision for 'no' is 91.9% and recall is 96.5%, while for 'yes' it is 56.3% and 35.1% respectively.

The screenshot shows the Weka Explorer interface with the Naive Bayes classifier selected. The 'Classifier' dropdown is set to 'NaiveBayes'. The 'Test options' section shows 'Cross-validation' selected with 'Folds' set to 10. The 'Result list' on the left shows the selected model: '22:51:40 - bayes.NaiveBayes'. The 'Classifier output' pane displays the following results:

Time taken to build model: 0.03 seconds

=== Stratified cross-validation ===

=== Summary ===

Correctly Classified Instances	4041	89.3829 %
Incorrectly Classified Instances	480	10.6171 %
Kappa statistic	0.3775	
Mean absolute error	0.1461	
Root mean squared error	0.2914	
Relative absolute error	71.6142 %	
Root relative squared error	91.2684 %	
Total Number of Instances	4521	

=== Detailed Accuracy By Class ===

	TP Rate	FP Rate	Precision	Recall	F-Measure	MCC	ROC Area	PRC Area	Class
	0.965	0.649	0.919	0.965	0.941	0.390	0.836	0.964	no
	0.351	0.036	0.563	0.351	0.433	0.390	0.836	0.448	yes
Weighted Avg.	0.894	0.578	0.878	0.894	0.883	0.390	0.836	0.905	

=== Confusion Matrix ===

a	b	-- Classified as
3858	142	a = no
338	183	b = yes

Status: OK

Log x0

Classification: Conclusion

In conclusion, the analysis of classification models, specifically J48 decision tree and Naive Bayes, provides valuable insights into predicting whether bank customers will subscribe to a term deposit or not.

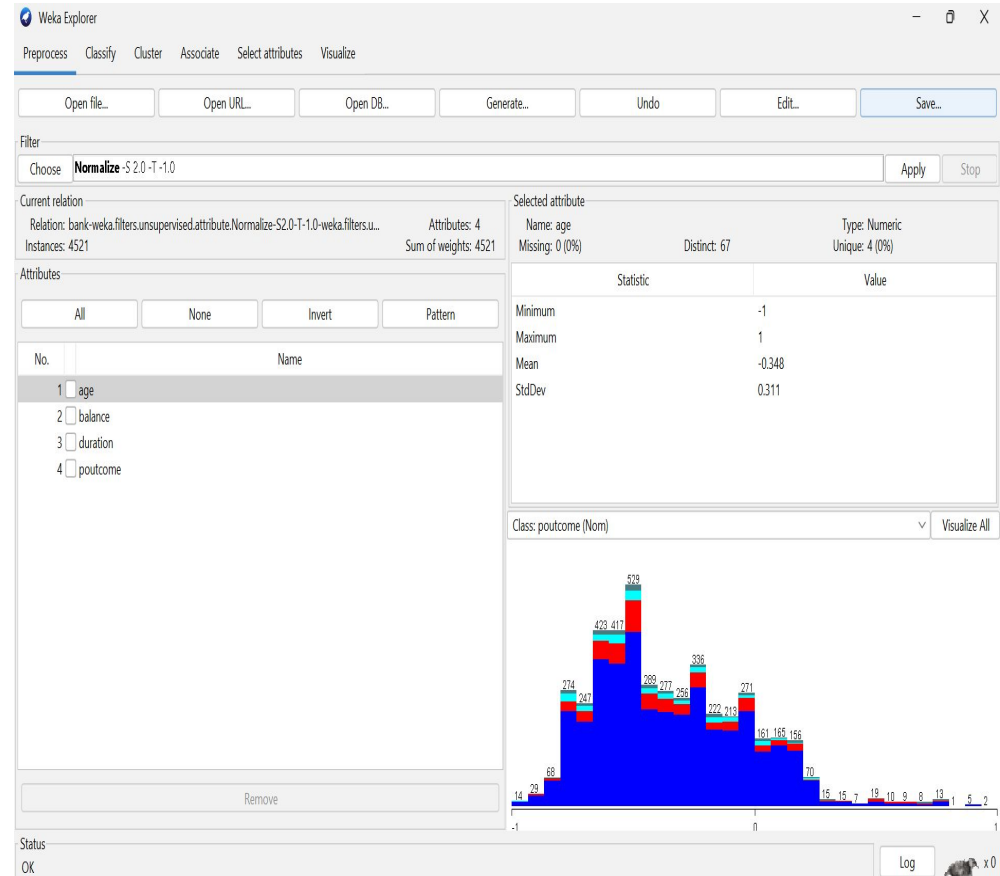
The selection of these models is worth analyzing due to several reasons:

- Decision trees, J48, offer intuitive insights into the decision-making process, making it easier to understand the factors influencing subscription outcomes. Naive Bayes, on the other hand, provides probabilistic reasoning behind classifications, aiding in understanding the likelihood of subscription based on attribute values.
- Both models achieved respectable accuracies, with J48 achieving an accuracy of 88.83% and Naive Bayes achieving 89.38%. Despite slight differences in performance metrics, both models offer reliable predictions.
- Through attribute analysis, we identified key factors influencing subscription decisions. Attributes like 'duration' and 'poutcome' emerged as significant predictors in both models.
- By comparing the performance of different models, such as J48 and Naive Bayes, we gain insights into their strengths and weaknesses.

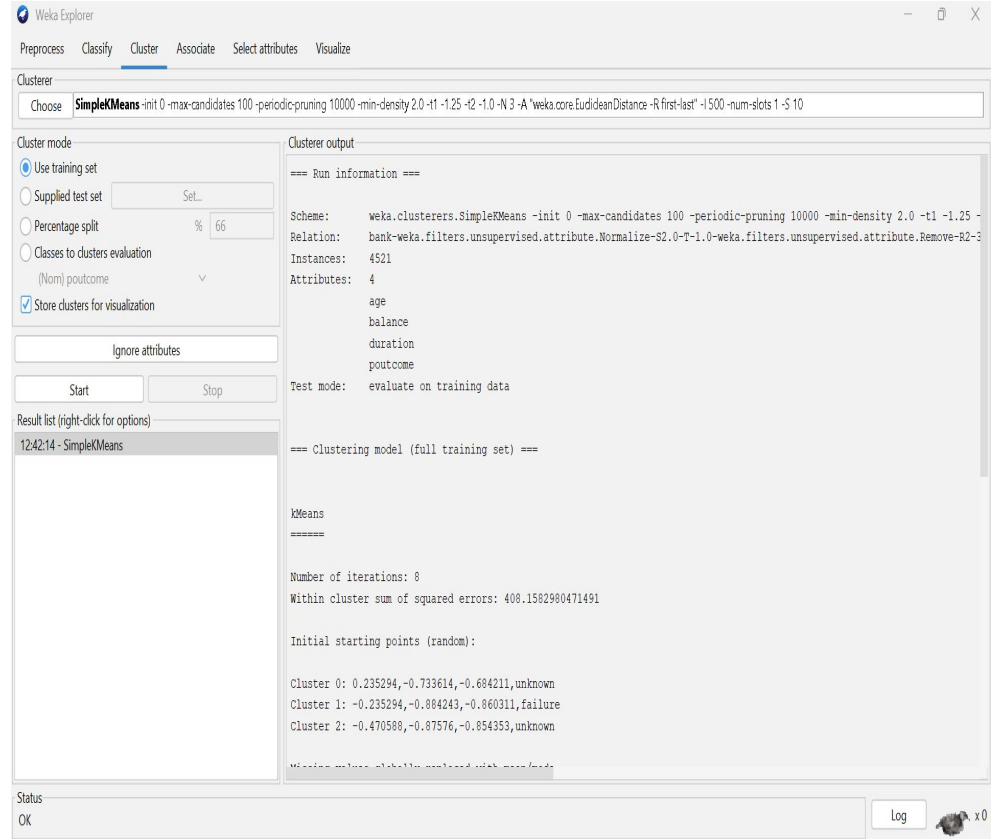
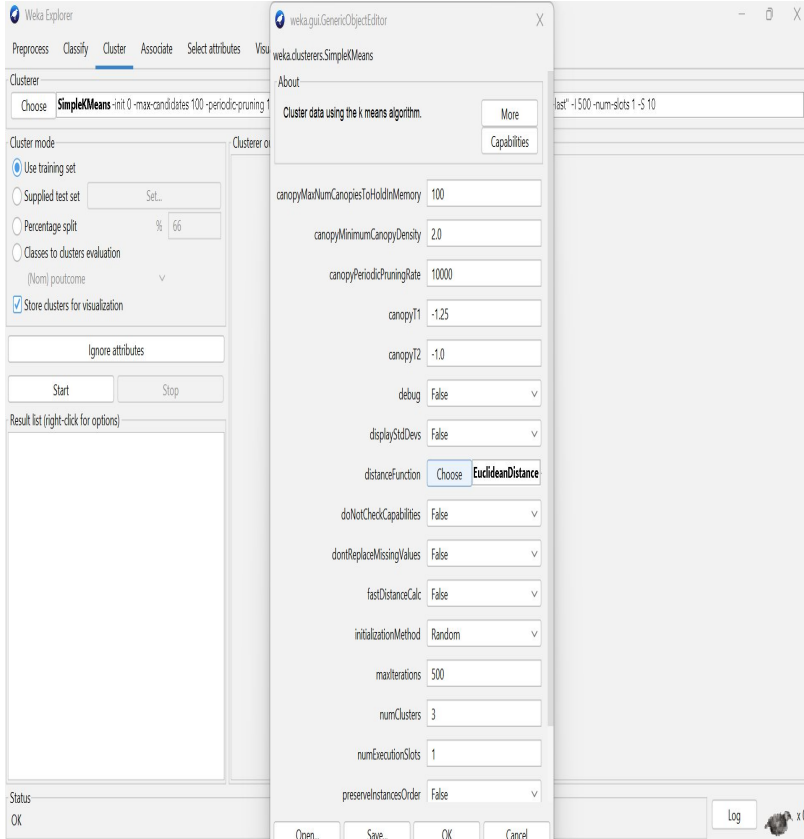
Clustering - K-means

Attribute chosen:

- Age provides a broad view of customer demographics, enabling segmentation based on different life stages or age groups.
- Balance offers information about customers' financial health, guiding strategies for targeting high-value.
- Duration indicates the depth of customer engagement during marketing interactions, helping identify active or responsive segments.
- Poutcome provides insights into previous campaign outcomes, aiding in understanding customer response patterns and preferences.
- The goal is to conduct cluster analysis to identify distinct customer segments based on similarities in demographic characteristics, financial status, campaign engagement, and past behavior. To enhance subscription outcomes and customer engagement through data-driven insights gained from clustering analysis.



Clustering - K-means



Clustering - K-means

Cluster Centroids:

- Cluster 0: Customers with a moderate age, slightly lower balance, shorter duration, and previous outcome unknown.
- Cluster 1: Customers with a slightly younger age, similar balance, shorter duration, and previous outcome failure.
- Cluster 2: Customers with a higher age, slightly higher balance, longer duration, and previous outcome unknown.

Goal Analysis:

- Clustering has identified three distinct customer segments based on demographic, financial, and campaign interaction attributes.
- By understanding these segments, personalized marketing communication strategies can be developed.
- Better targeting of specific customer groups can improve subscription outcomes and customer engagement.

The screenshot shows the Weka Explorer interface with the 'Cluster' tab selected. The 'SimpleKMeans' algorithm is chosen, and the 'Clusterer output' panel displays the following information:

Clusterer output

Cluster 0: 0.235294, -0.733614, -0.684211, unknown
Cluster 1: -0.235294, -0.884243, -0.860311, failure
Cluster 2: -0.470588, -0.87576, -0.854353, unknown

Missing values globally replaced with mean/mode

Final cluster centroids:

Attribute	Full Data	0	1	2
	(4521.0)	(1492.0)	(579.0)	(2450.0)
age	-0.3479	-0.0139	-0.332	-0.5551
balance	-0.8729	-0.8666	-0.8666	-0.8782
duration	-0.8279	-0.8301	-0.8312	-0.8257
poutcome	unknown	unknown	failure	unknown

Time taken to build model (full training data) : 0.08 seconds

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

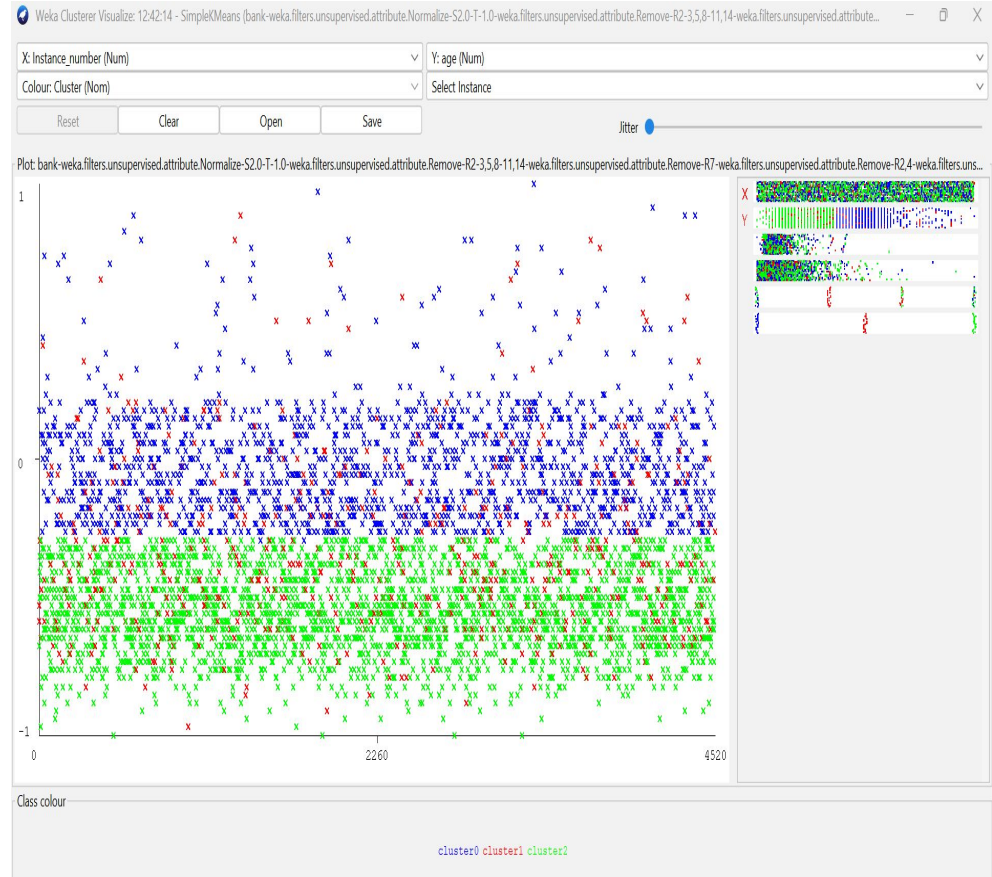
Clustered Instances

0	1492 (33%)
1	579 (13%)
2	2450 (54%)

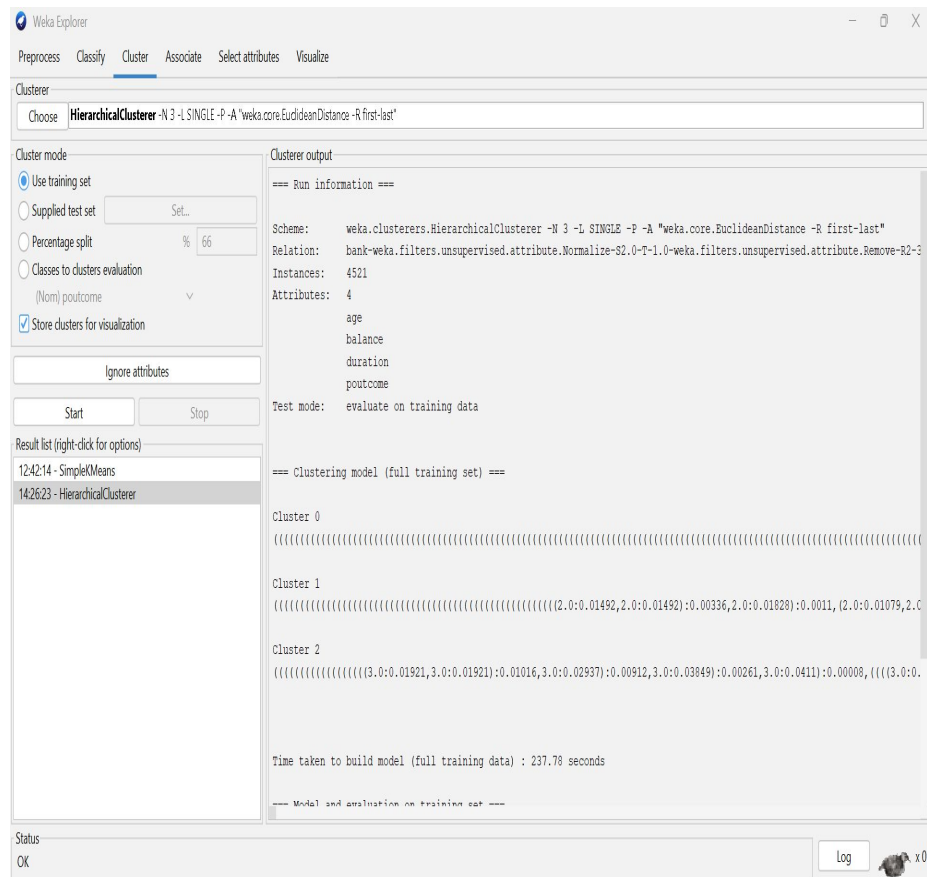
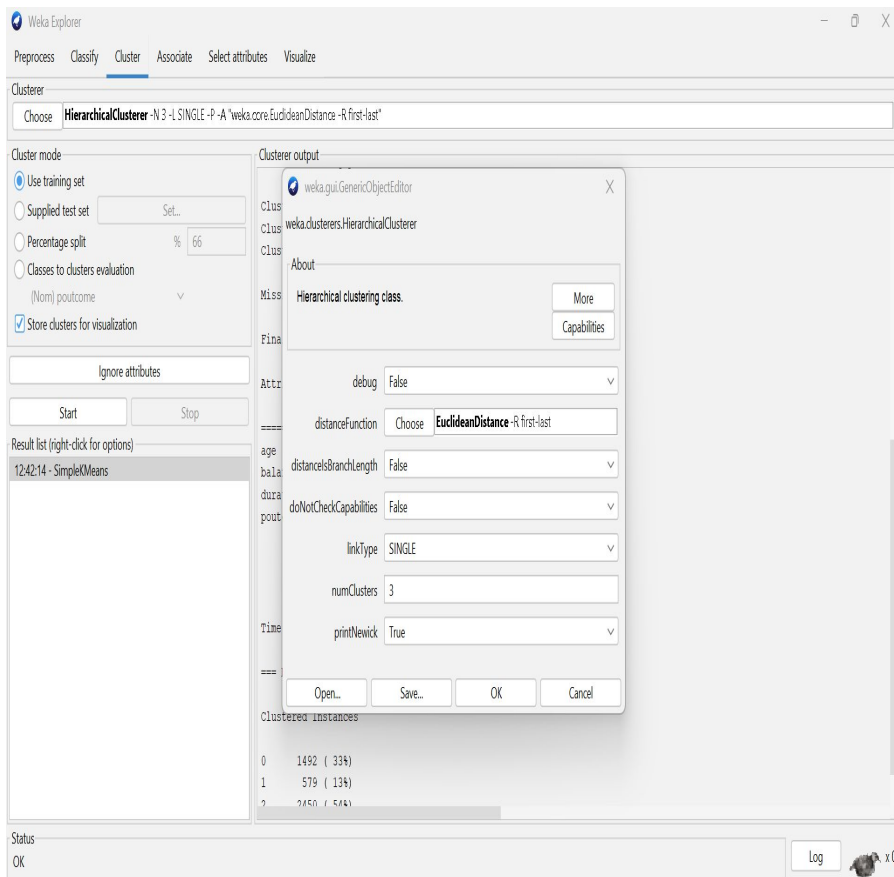
Clustering - K-means

Conclusion:

- Cluster Distribution: Cluster 0 comprises 33% of customers, Cluster 1 comprises 13%, and Cluster 2 comprises 54%.
- The clustering model has successfully grouped customers into distinct segments based on similarities in attributes.
- Insights from clustering can guide marketing strategies for better targeting and engagement, ultimately improving subscription outcomes.
- Within Cluster Sum of Squared Errors: 408.1583
- Cluster Distribution: Cluster 0 (33%), Cluster 1 (13%), Cluster 2 (54%)



Clustering - Hierarchical



Clustering - Hierarchical

Using the same set of attributes with Hierarchical Clustering as with K-means Clustering.

Interpreting the hierarchical clustering model:

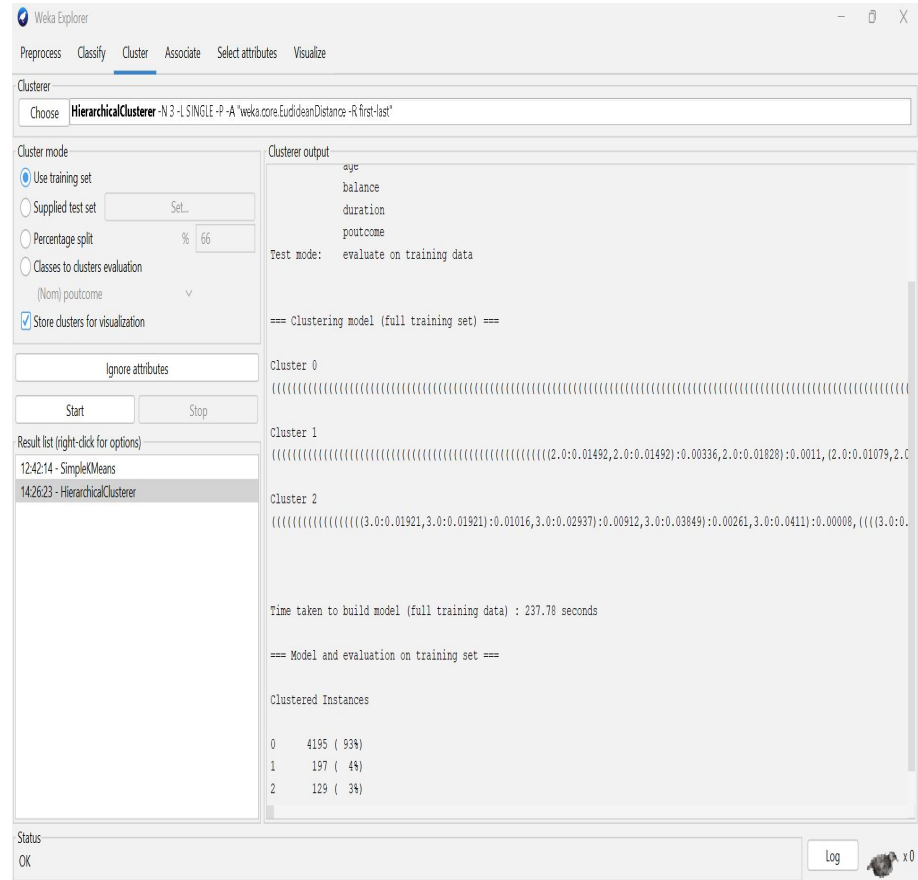
The hierarchical clustering model segmented the dataset into three clusters based on the selected attributes. However, it took significantly longer to build compared to the K-means model, indicating higher computational complexity.

Analysis of the goal:

The goal of hierarchical clustering was to identify distinct customer segments using a subset of attributes, allowing for comparison with the K-means clustering results. This analysis aimed to provide insights into potential differences in cluster structures and their implications for marketing campaign targeting.

Conclusion of the results:

The hierarchical clustering model produced three clusters, with the majority of instances (93%) falling into Cluster 0. This indicates a highly imbalanced distribution compared to the K-means results, where Cluster 0 represented only 33% of instances.



Clustering: Conclusion

- K-means clustering identified three clusters with more balanced distribution among them. It provided a clear segmentation of the dataset. The clusters were based on demographic, financial, and campaign interaction attributes, allowing for personalized communication with customers.
- Hierarchical clustering, on the other hand, yielded three clusters as well but with a highly imbalanced distribution, with the majority of instances falling into one cluster. While it also provided segmentation, the computational complexity was significantly higher compared to K-means.
- Analyzing both clustering methods was worthwhile as it allowed for comparison of cluster structures and their implications for marketing campaign targeting. K-means offered a more straightforward approach with balanced clusters, facilitating clearer insights into customer segments. However, hierarchical clustering provided additional perspective, albeit with higher computational costs.

Conclusion of Overall Data

- The overall data analysis involved both classification and clustering tasks.
- For classification, J48 decision tree and Naive Bayes models were applied, achieving moderate accuracy rates around 89%.
- K-means and hierarchical clustering were employed for clustering, revealing distinct customer segments with varying degrees of balance and computational complexity.
- From classification analysis, we gained insights into customer behavior regarding subscription to term deposits based on demographic and campaign interaction attributes.
- Clustering analysis identified customer segments, offering opportunities for personalized marketing communication and better targeting of specific groups.
- Further analysis could explore ensemble methods or more advanced clustering algorithms to improve classification and clustering accuracy.
- Additionally, integrating external datasets or conducting sentiment analysis on customer feedback could enrich insights and refine marketing strategies.

Thank You
