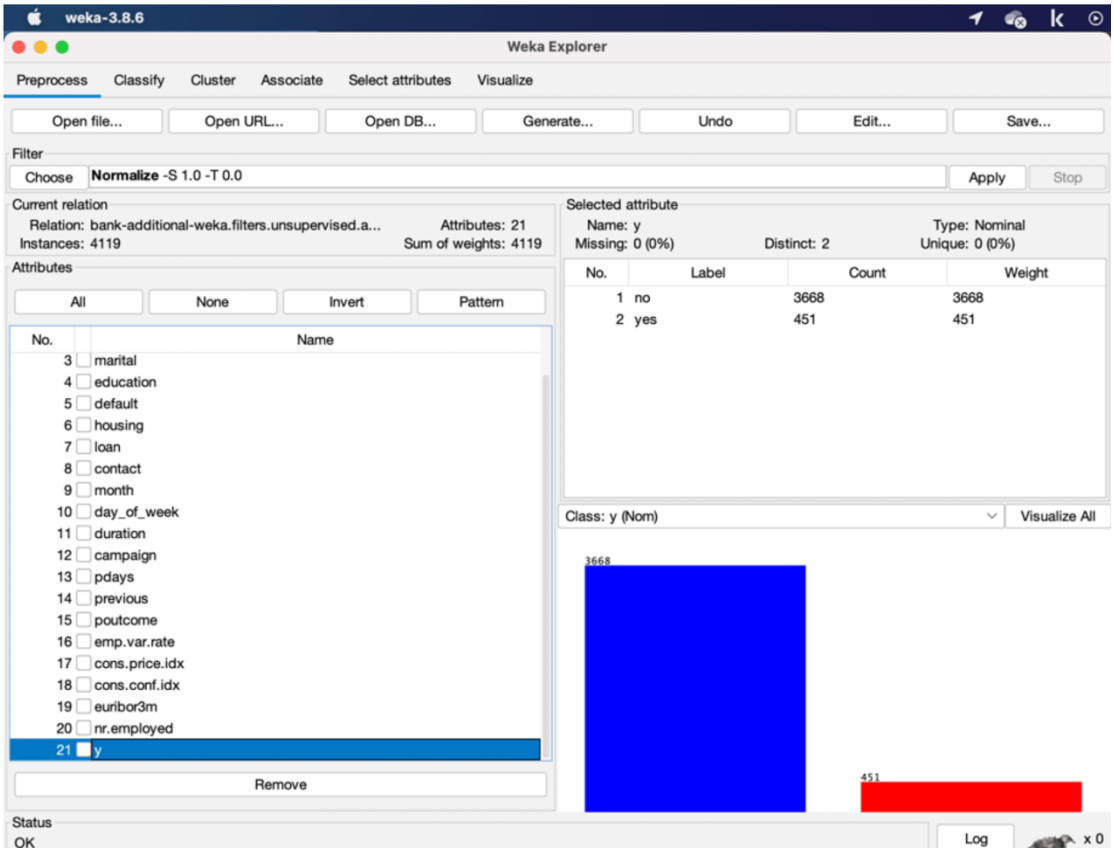
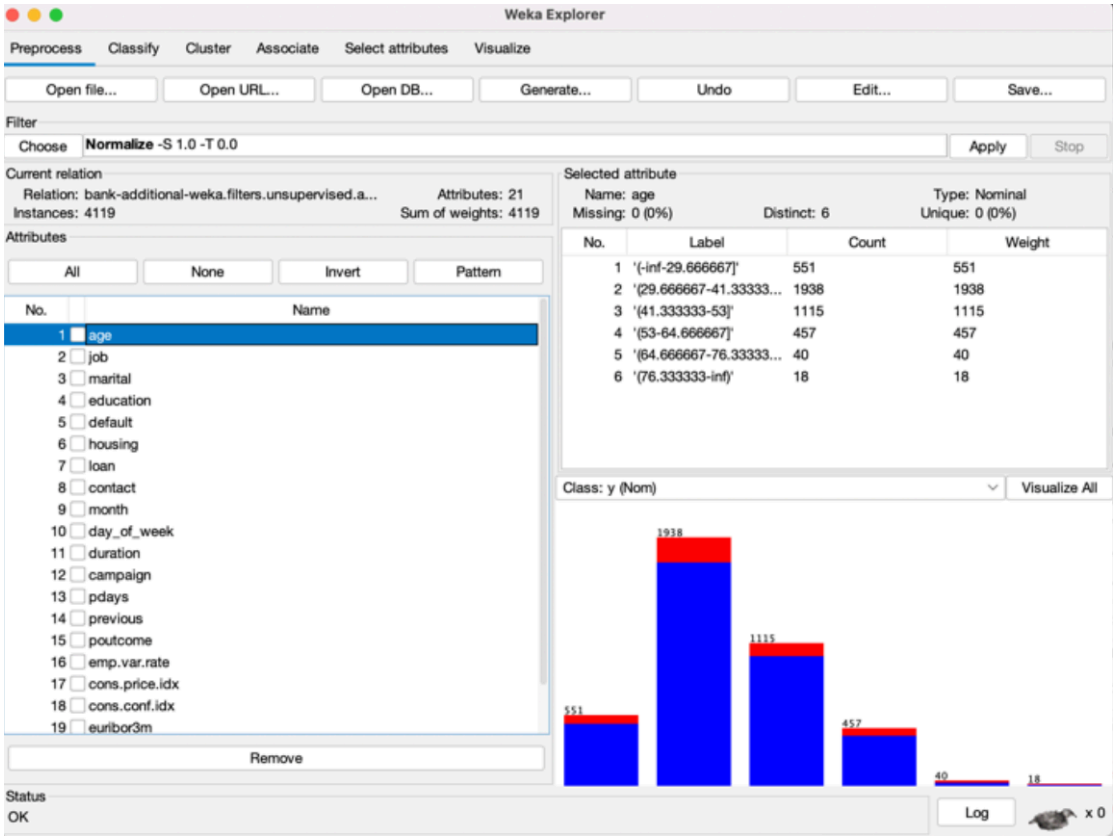
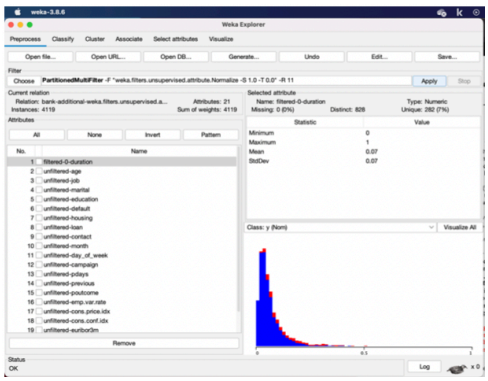
Step1 & 2 : Firstly, we use the "Bank Marketing (with social/economic context)" dataset, specifically the "**bank-additional.csv**" file, to build the model. Moreover, The target variable is *"y"* which indicates whether the client subscribed to a term deposit or not.

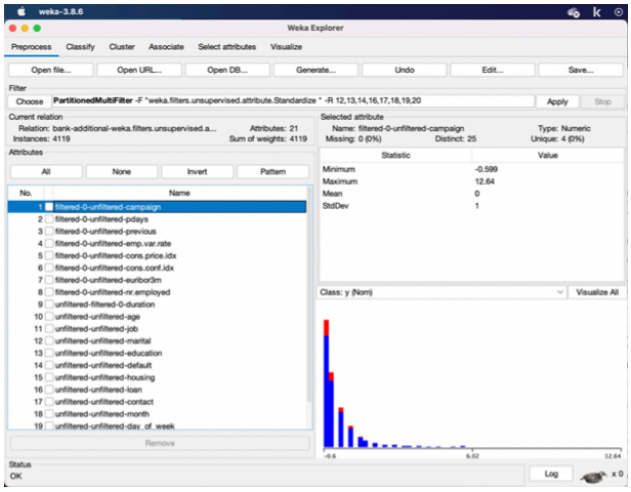


Step 3 : Now we carry out the data pre-processing part . Firstly, I group client into 6 equal width age groups.

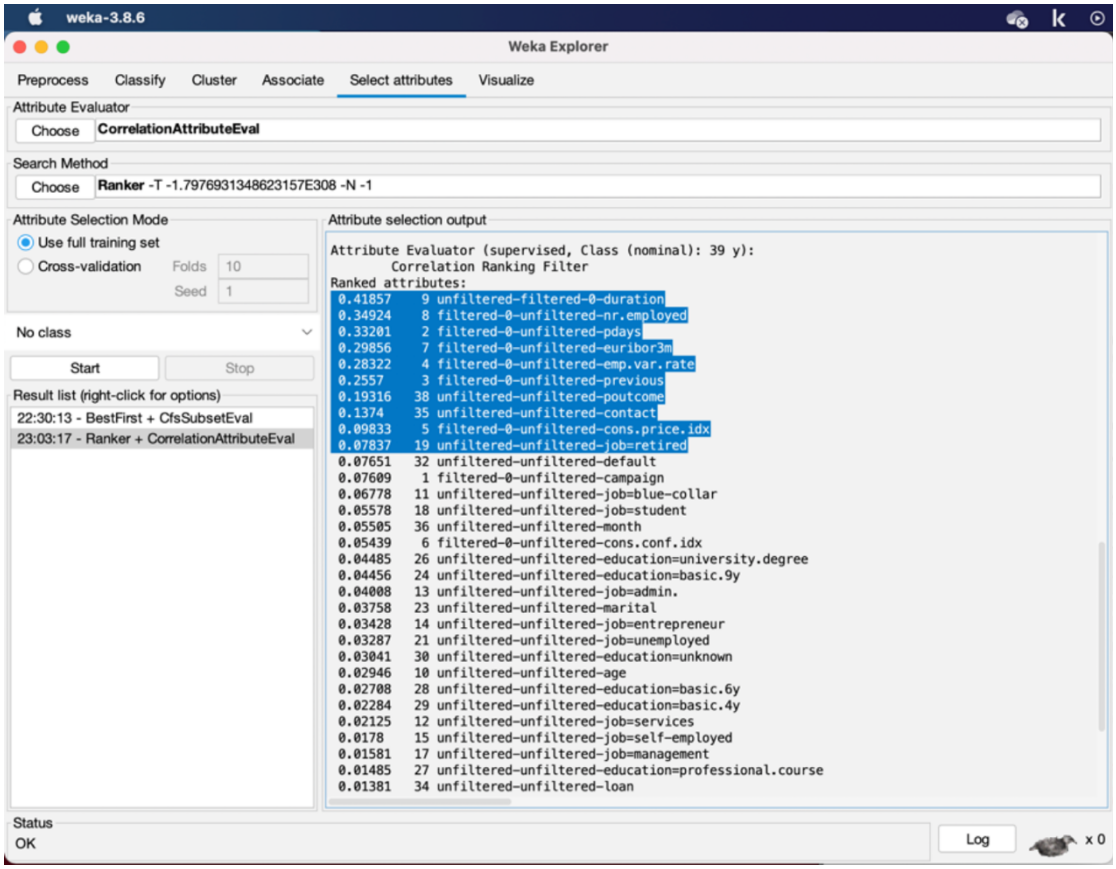


Then, by using normalization and standardization. In normalization, we set the minimum value be 0 while maximum value be 1. In standardization, we set the mean be 0 and standard deviation be 1.

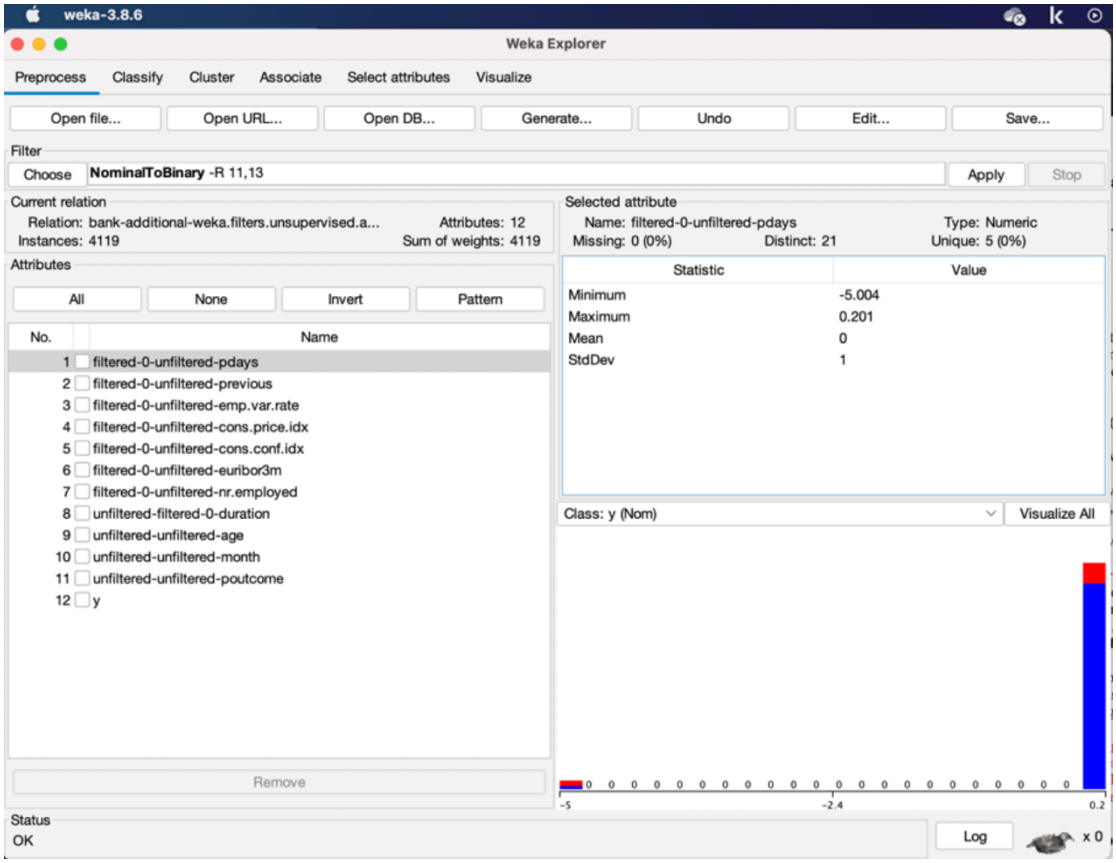




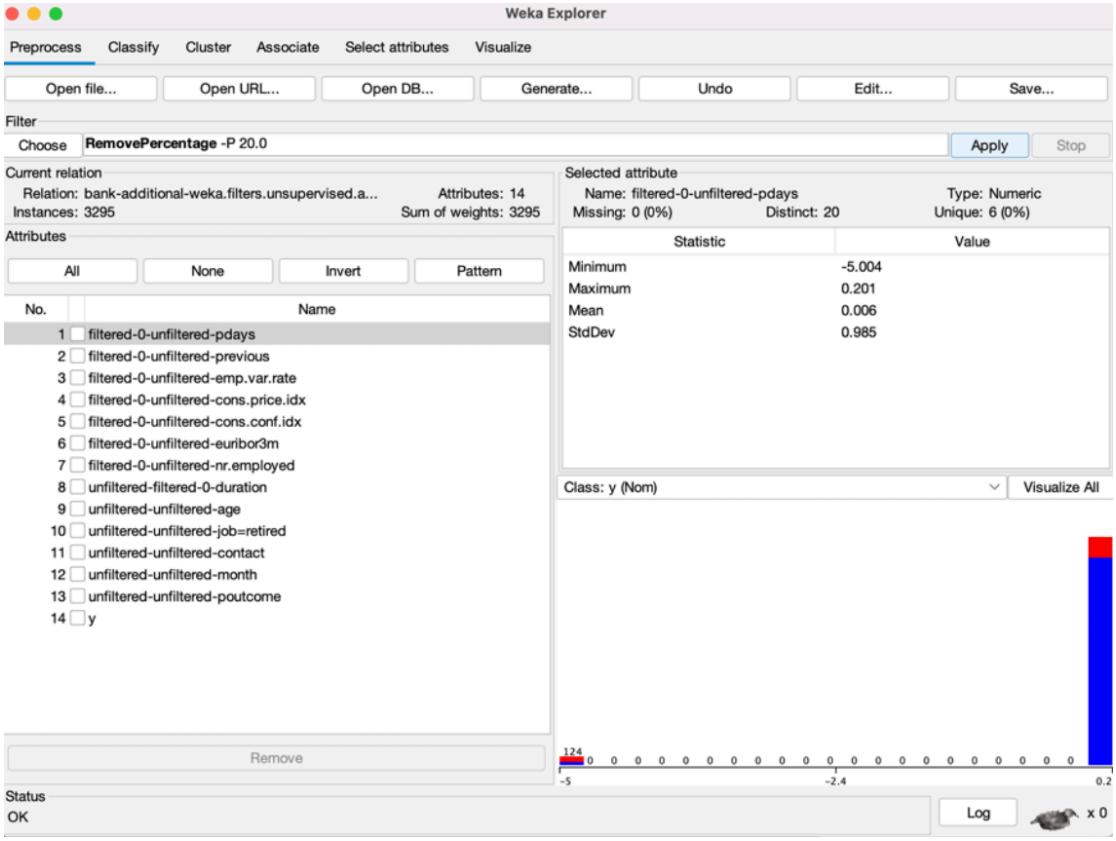
After that, I eliminate variables not related to the target variable by employing correlation-based selection methods and retain the top 10 attributes.

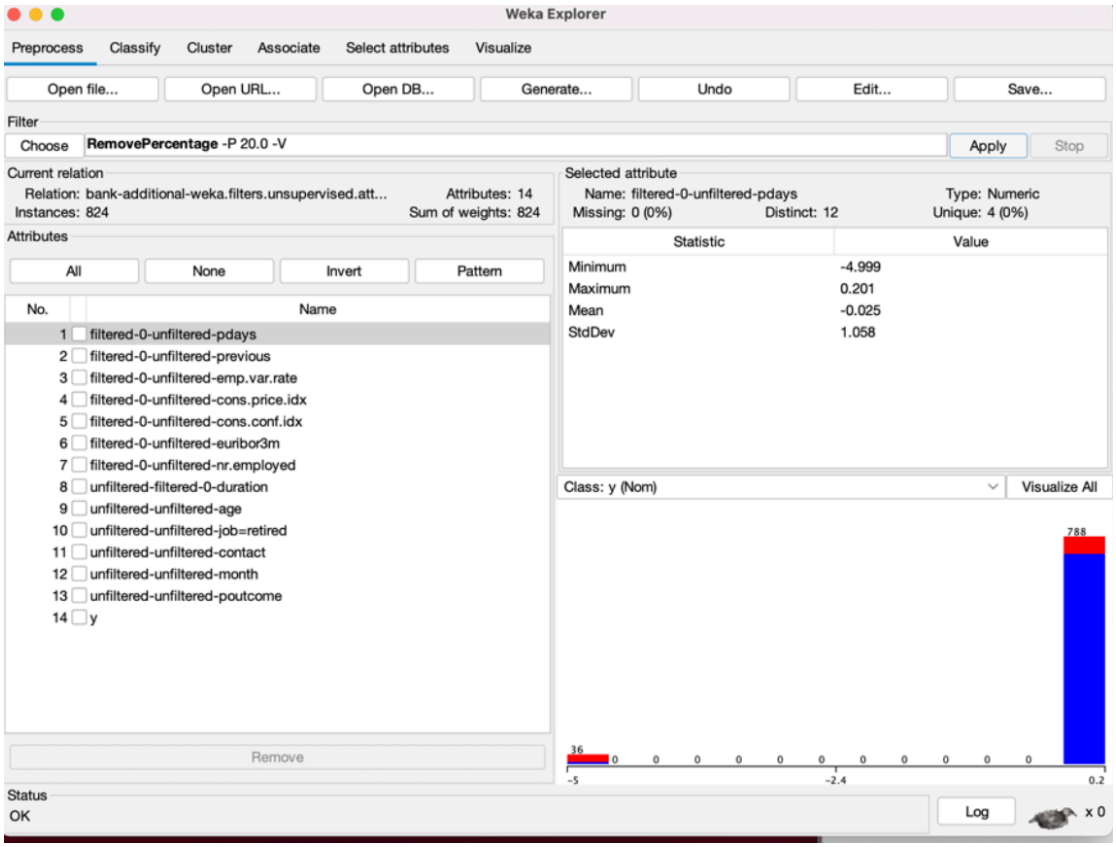


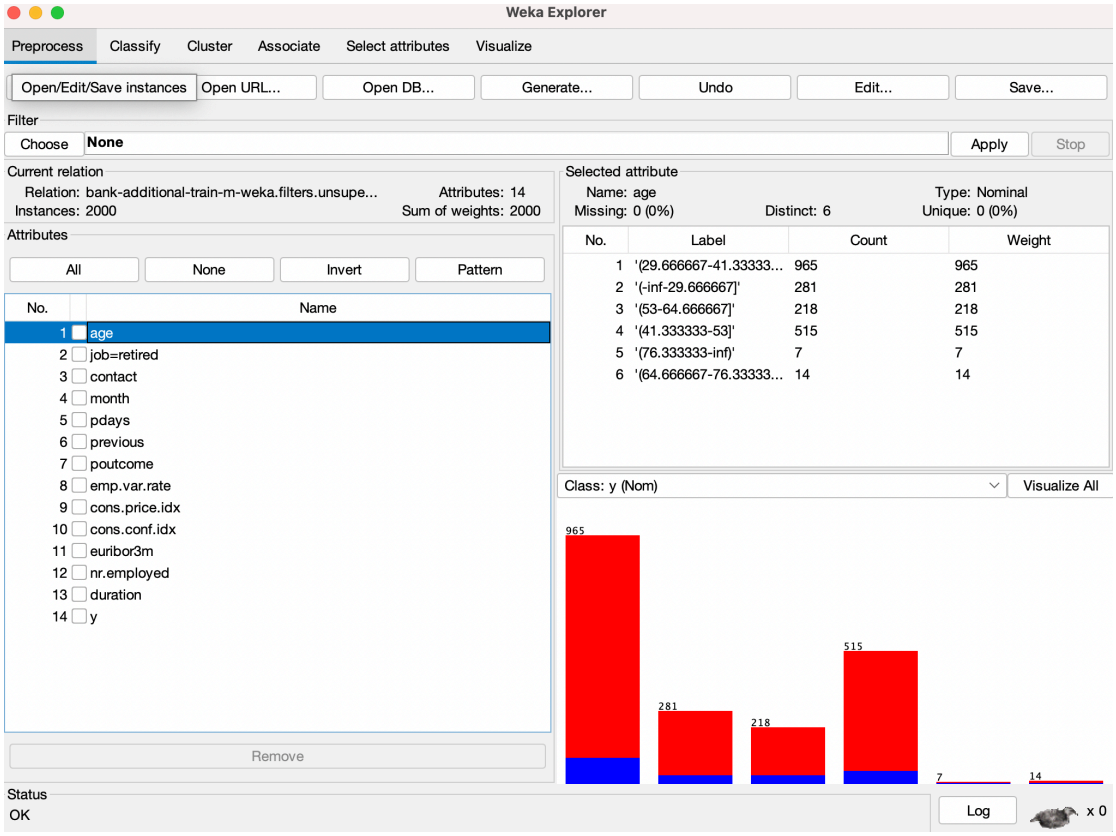
I also eliminate variables not related to the target variable by employing information gain-based selection methods and retain the top 10 attributes.



Step 4&5 : I split the whole dataset into two subsets, namely, a training set and a testing set and I save the training and testing sets as two separate files for subsequent processing.

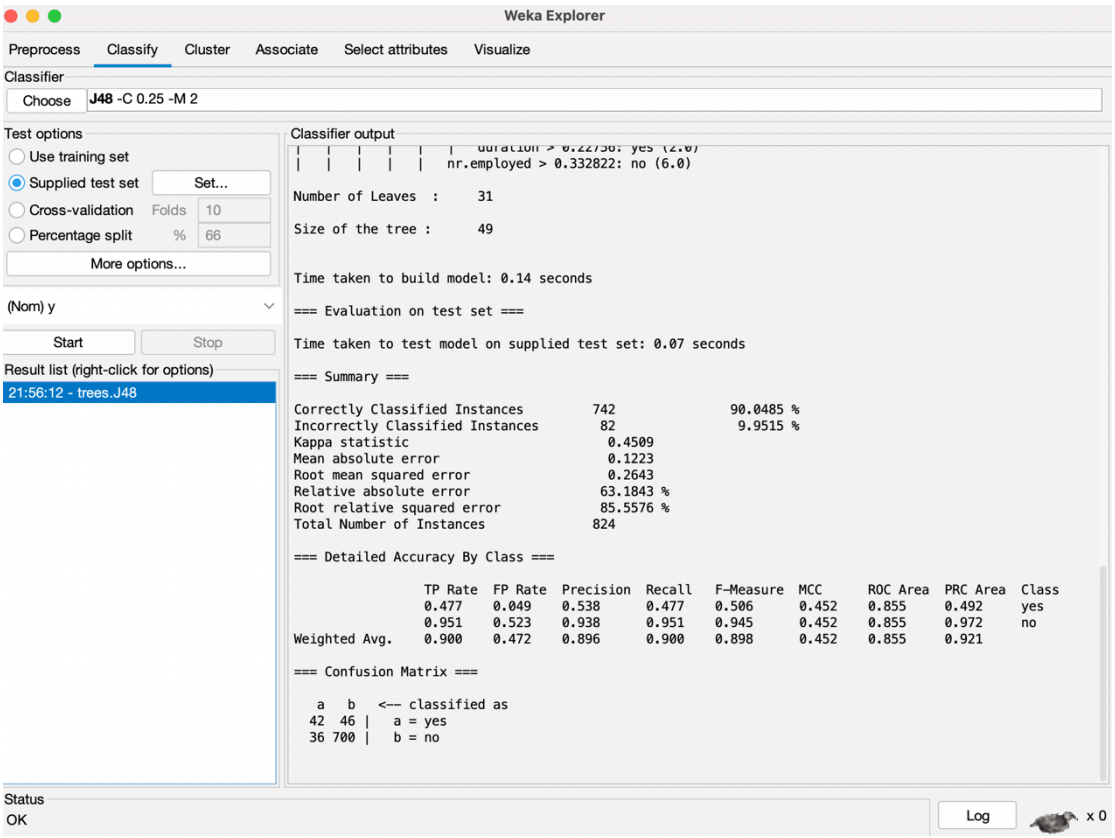

Weka Explorer
Preprocess
Classify
Cluster
Associate
Select attributes
Visualize
Open file...
Open URL...
Open DB...
Generate...
Undo
Edit...
Save...
Filter
Choose
RemovePercentage -P 20.0
Apply
Stop
Current relation
Relation: bank-additional-weka.filters.unsupervised.a...
Attributes: 14
Instances: 3295
Sum of weights: 3295
Selected attribute
Name: filtered-0-unfiltered-pdays
Type: Numeric
Missing: 0 (0%)
Distinct: 20
Unique: 6 (0%)
Attributes
All
None
Invert
Pattern
No.
Name
1 filtered-0-unfiltered-pdays
2 filtered-0-unfiltered-previous
3 filtered-0-unfiltered-emp.var.rate
4 filtered-0-unfiltered-cons.price.idx
5 filtered-0-unfiltered-cons.conf.idx
6 filtered-0-unfiltered-euribor3m
7 filtered-0-unfiltered-nr.employed
8 unfiltered-filtered-0-duration
9 unfiltered-unfiltered-age
10 unfiltered-unfiltered-job=retired
11 unfiltered-unfiltered-contact
12 unfiltered-unfiltered-month
13 unfiltered-unfiltered-poutcome
14 y
Statistic
Value
Minimum
-5.004
Maximum
0.201
Mean
0.006
StdDev
0.985
Class: y (Nom)
Visualize All
124
-5
-2.4
0.2
Remove
Status
OK
Log
x 0


Weka Explorer
Preprocess
Classify
Cluster
Associate
Select attributes
Visualize
Open file...
Open URL...
Open DB...
Generate...
Undo
Edit...
Save...
Filter
Choose
RemovePercentage -P 20.0 -V
Apply
Stop
Current relation
Relation: bank-additional-weka.filters.unsupervised.att...
Attributes: 14
Instances: 824
Sum of weights: 824
Selected attribute
Name: filtered-0-unfiltered-pdays
Type: Numeric
Missing: 0 (0%)
Distinct: 12
Unique: 4 (0%)
Attributes
All
None
Invert
Pattern
No.
Name
1 filtered-0-unfiltered-pdays
2 filtered-0-unfiltered-previous
3 filtered-0-unfiltered-emp.var.rate
4 filtered-0-unfiltered-cons.price.idx
5 filtered-0-unfiltered-cons.conf.idx
6 filtered-0-unfiltered-euribor3m
7 filtered-0-unfiltered-nr.employed
8 unfiltered-filtered-0-duration
9 unfiltered-unfiltered-age
10 unfiltered-unfiltered-job=retired
11 unfiltered-unfiltered-contact
12 unfiltered-unfiltered-month
13 unfiltered-unfiltered-poutcome
14 y
Statistic
Value
Minimum
-4.999
Maximum
0.201
Mean
-0.025
StdDev
1.058
Class: y (Nom)
Visualize All
788
36
-5
-2.4
0.2
Remove
Status
OK
Log
x 0

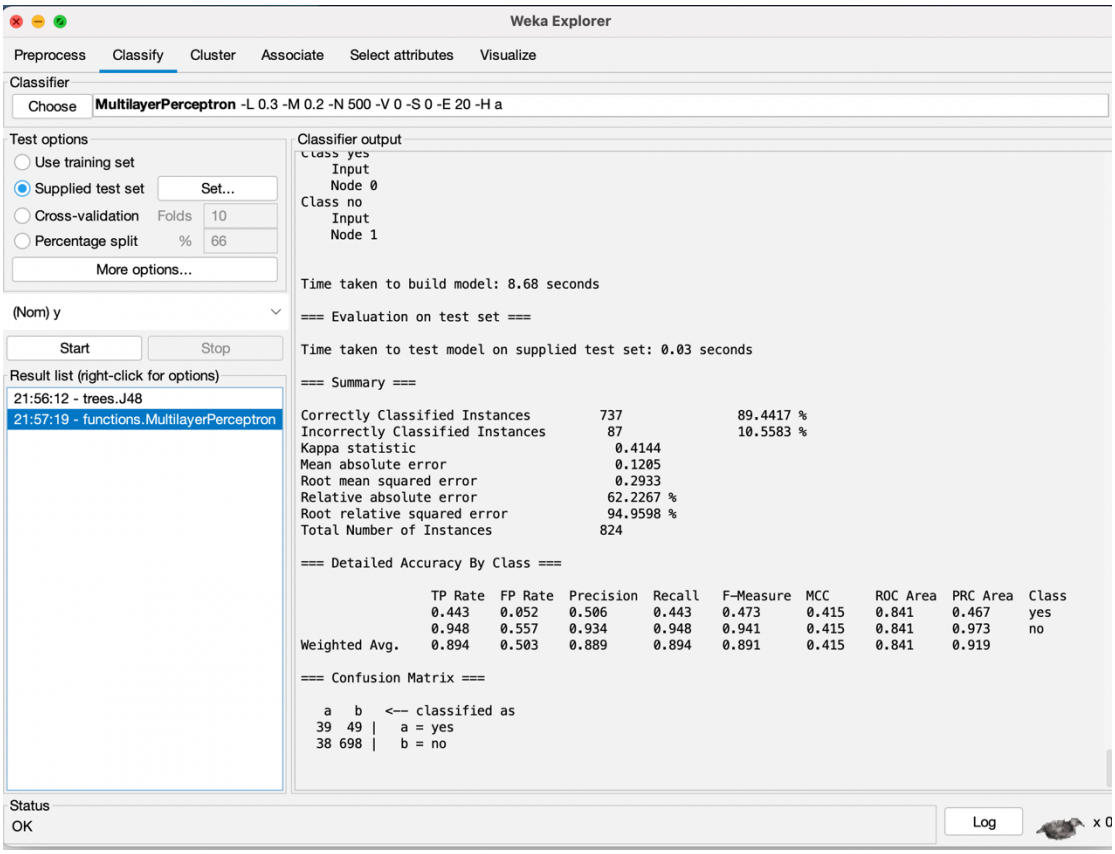


(Data-prepocessing part is completed)

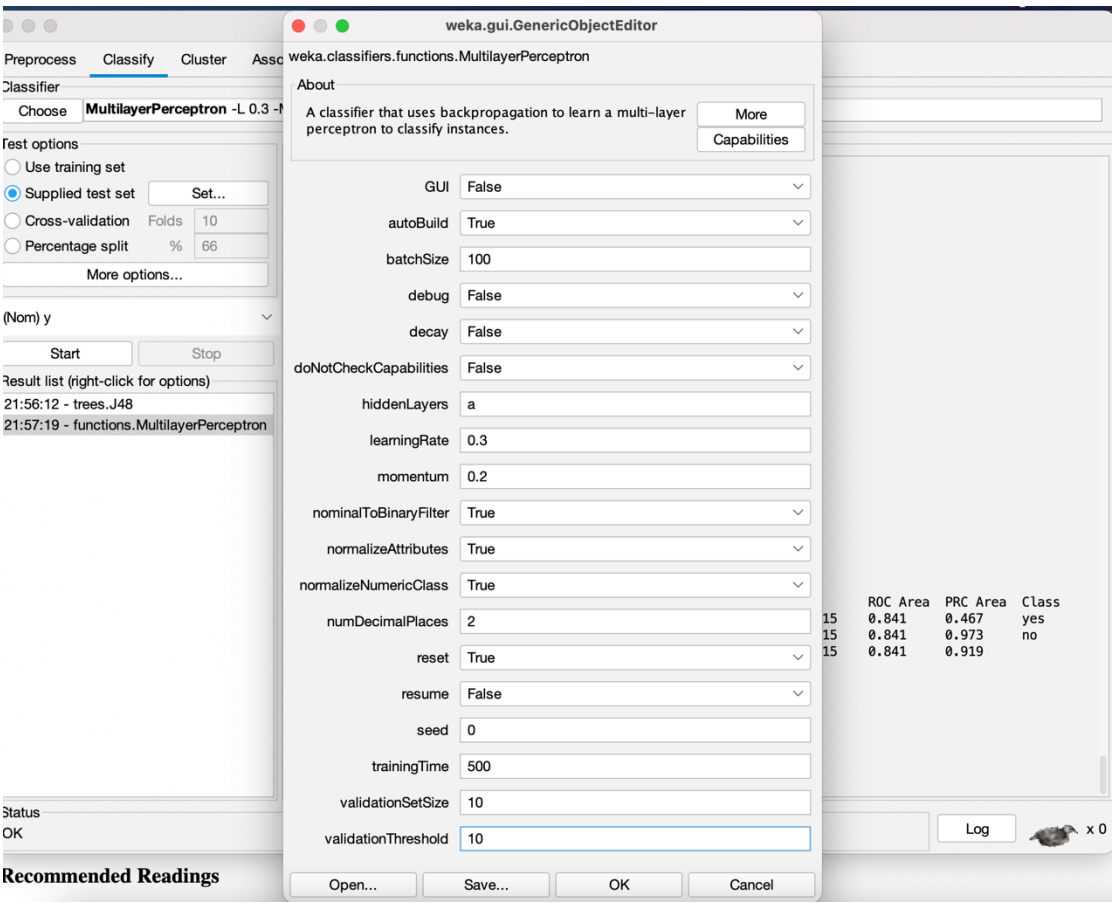
Step 6&7: Similar to the spirit of Assignment 2, I use the training set to train a model and use the testing set to measure the performance of the model. I then carry out decision tree model, neural network model, and logistic regression model respectively by trying different parameters in these models.



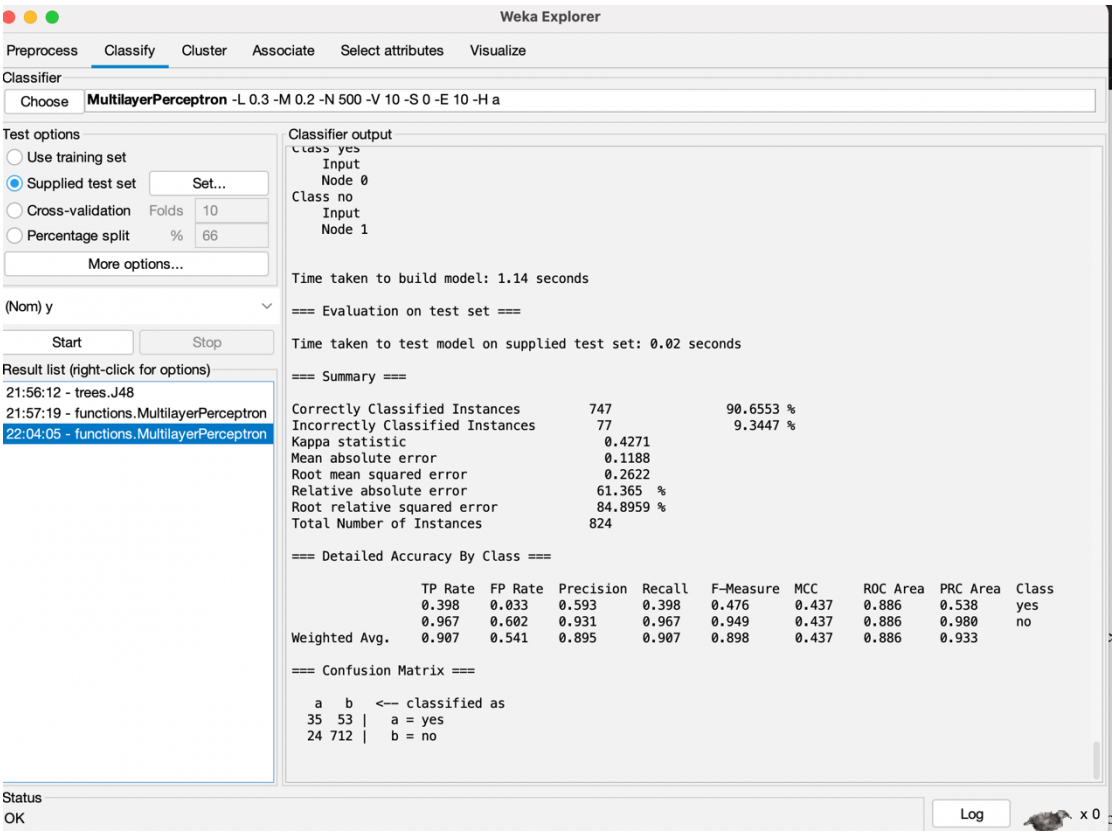
(I employ Decision tree J48 model first – with 90.0485% accuracy)

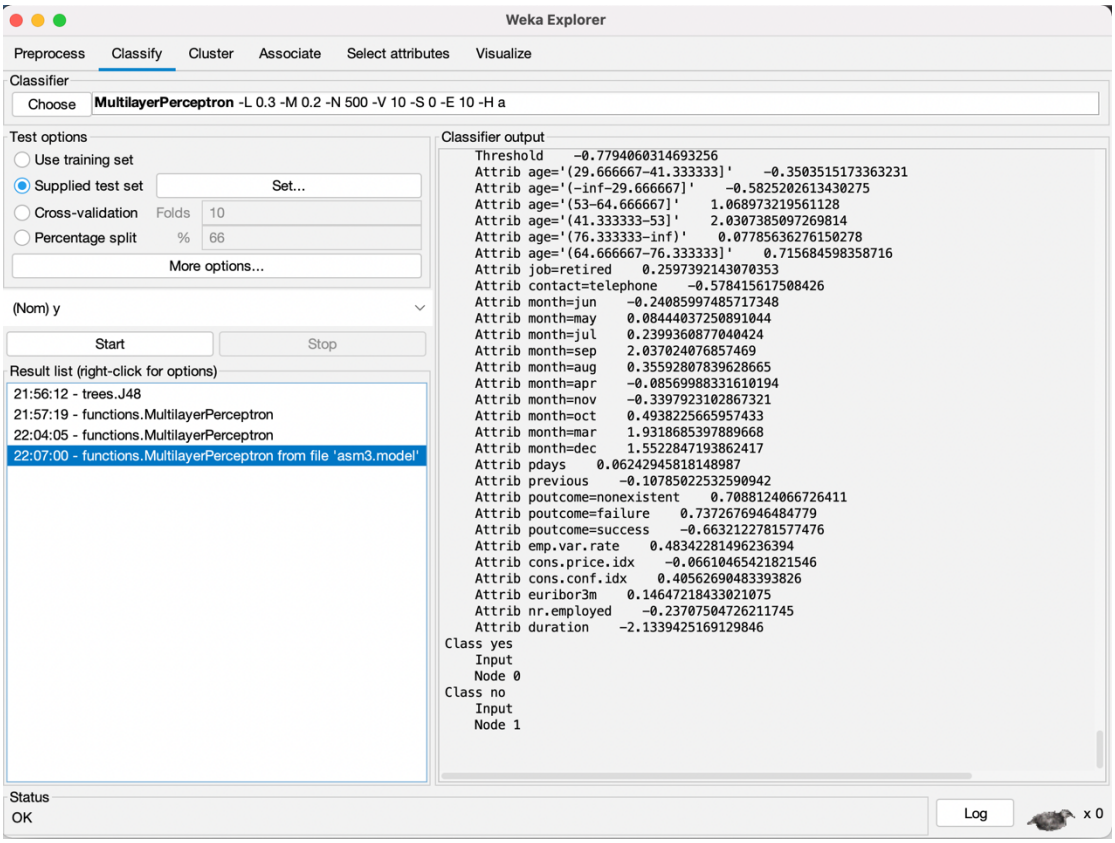


(Then I employ Neural network MultilayerPerceptron model with 89.4417% accuracy)



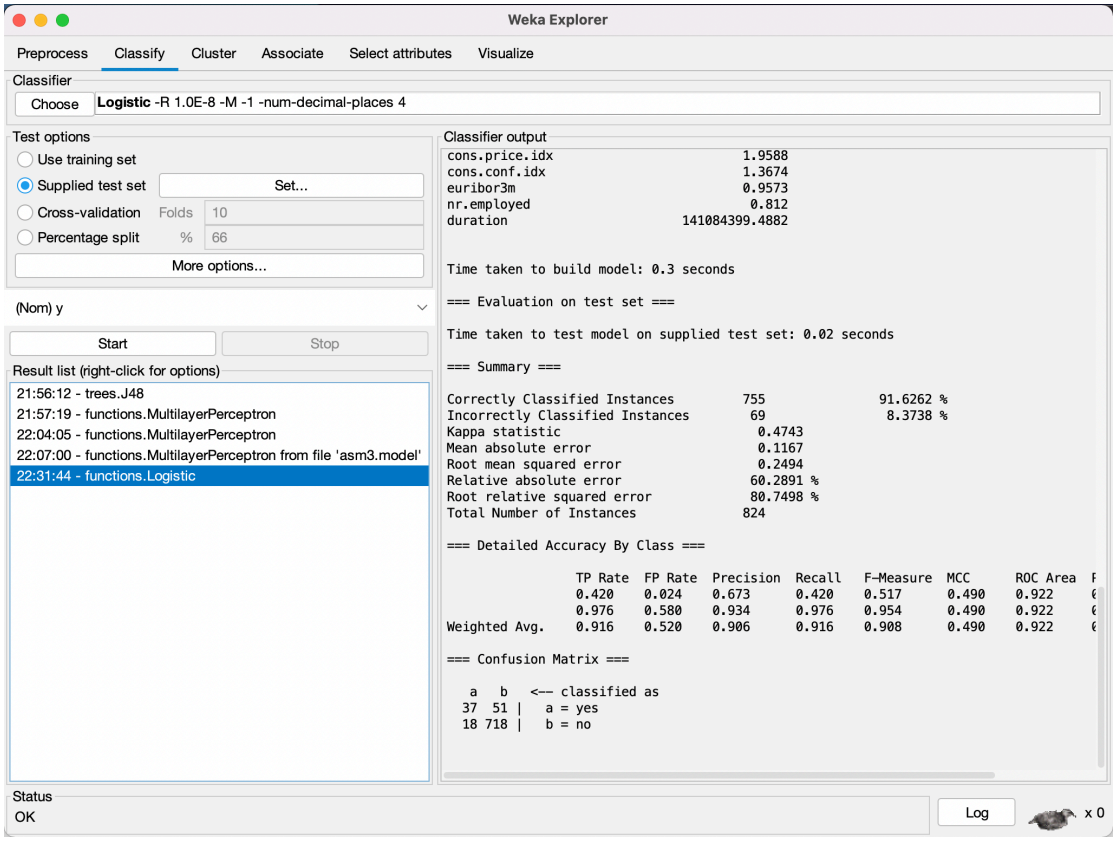
(In order to increase the accuracy of the current Neural Network model, I change the parameter of validationThreshold to 10)


Weka Explorer
Preprocess
Classify
Cluster
Associate
Select attributes
Visualize
Classifier
Choose
MultilayerPerceptron -L 0.3 -M 0.2 -N 500 -V 10 -S 0 -E 10 -H a
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)
21:56:12 - trees.J48
21:57:19 - functions.MultilayerPerceptron
22:04:05 - functions.MultilayerPerceptron
Classifier output
Input
Node 0
Class no
Input
Node 1
Time taken to build model: 1.14 seconds
=== Evaluation on test set ===
Time taken to test model on supplied test set: 0.02 seconds
=== Summary ===
Correctly Classified Instances 747 90.6553 %
Incorrectly Classified Instances 77 9.3447 %
Kappa statistic 0.4271
Mean absolute error 0.1188
Root mean squared error 0.2622
Relative absolute error 61.365 %
Root relative squared error 84.8959 %
Total Number of Instances 824
=== Detailed Accuracy By Class ===
TP Rate FP Rate Precision Recall F-Measure MCC ROC Area PRC Area Class
0.398 0.033 0.593 0.398 0.476 0.437 0.886 0.538 yes
0.967 0.602 0.931 0.967 0.949 0.437 0.886 0.980 no
Weighted Avg. 0.907 0.541 0.895 0.907 0.898 0.437 0.886 0.933
=== Confusion Matrix ===
a b <-- classified as
35 53 | a = yes
24 712 | b = no
Status
OK
Log
x 0

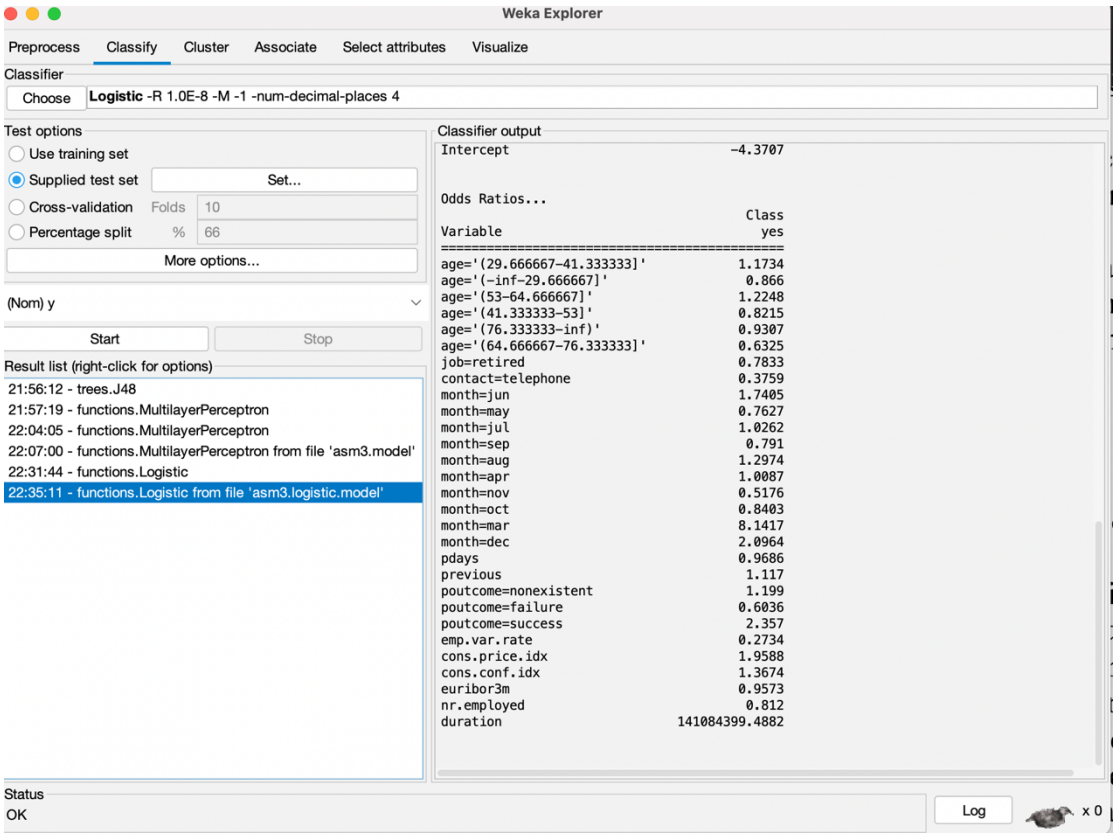

Weka Explorer
Preprocess
Classify
Cluster
Associate
Select attributes
Visualize
Classifier
Choose
MultilayerPerceptron -L 0.3 -M 0.2 -N 500 -V 10 -S 0 -E 10 -H a
Test options
Use training set
Supplied test set
Set...
Cross-validation
Folds
10
Percentage split
%
66
More options...
(Nom) y
Start
Stop
Result list (right-click for options)
21:56:12 - trees.J48
21:57:19 - functions.MultilayerPerceptron
22:04:05 - functions.MultilayerPerceptron
22:07:00 - functions.MultilayerPerceptron from file 'asm3.model'
Classifier output
Threshold -0.7794060314693256
Attrib age='(29.666667-41.333333]' -0.3503515173363231
Attrib age='(-inf-29.666667]' -0.5825202613430275
Attrib age='(53-64.666667]' 1.068973219561128
Attrib age='(41.333333-53]' 2.0307385097269814
Attrib age='(76.333333-inf)' 0.07785636276150278
Attrib age='(64.666667-76.333333]' 0.715684598358716
Attrib job=retired 0.2597392143070353
Attrib contact=telephone -0.578415617508426
Attrib month=jun -0.24085997485717348
Attrib month=may 0.08444037250891044
Attrib month=jul 0.2399360877040424
Attrib month=sep 2.037024076857469
Attrib month=aug 0.35592807839628665
Attrib month=apr -0.08569988331610194
Attrib month=nov -0.3397923102867321
Attrib month=oct 0.4938225665957433
Attrib month=mar 1.9318685397889668
Attrib month=dec 1.5522847193862417
Attrib pdays 0.06242945818148987
Attrib previous -0.10785022532590942
Attrib poutcome=nonexistent 0.7088124066726411
Attrib poutcome=failure 0.7372676946484779
Attrib poutcome=success -0.6632122781577476
Attrib emp.var.rate 0.48342281496236394
Attrib cons.price.idx -0.06610465421821546
Attrib cons.conf.idx 0.40562690483393826
Attrib euribor3m 0.14647218433021075
Attrib nr.employed -0.23707504726211745
Attrib duration -2.1339425169129846
Class yes
Input
Node 0
Class no
Input
Node 1
Status
OK
Log
x 0

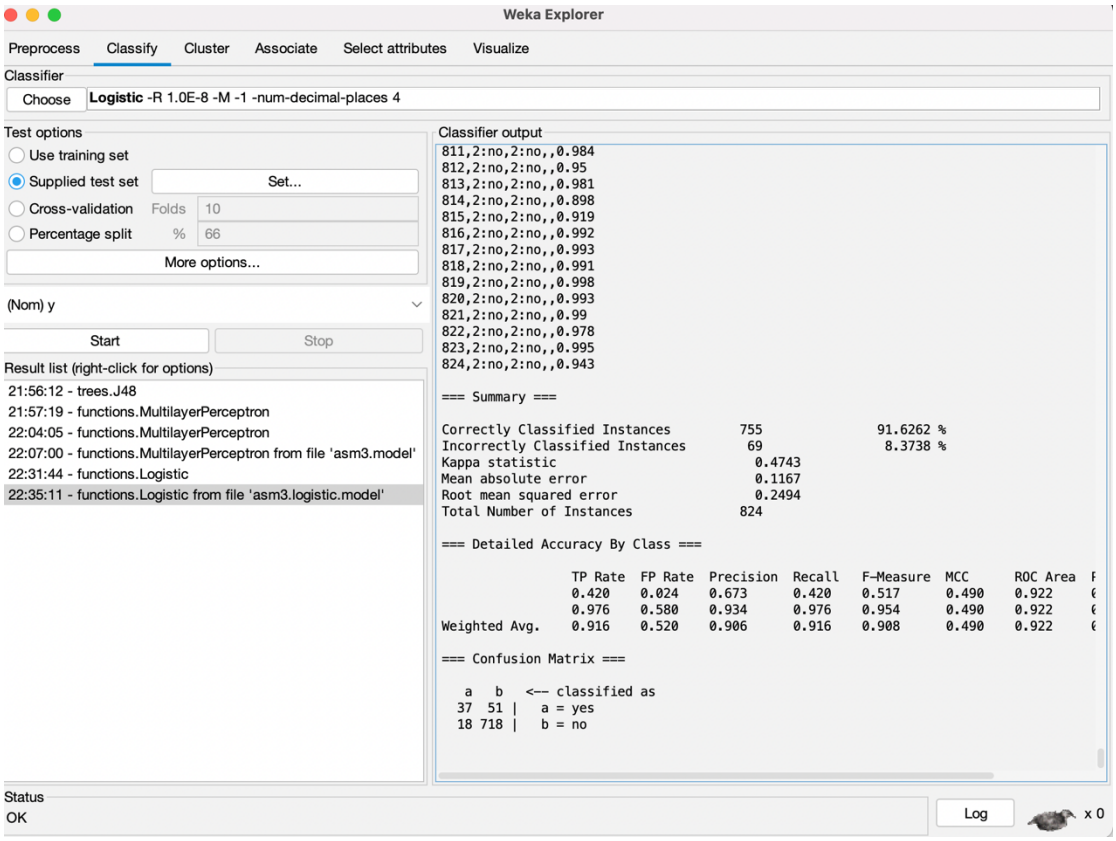
(After changing the parameter, the accuracy of the Neural network MultilayerPerceptron model increased from 88.4417% to 90.6553%)

Step8: After some trials, I decide and save the most desirable model (i.e.: logistic model with the most accuracy among the 3 models) for delivering to the bank for operation.

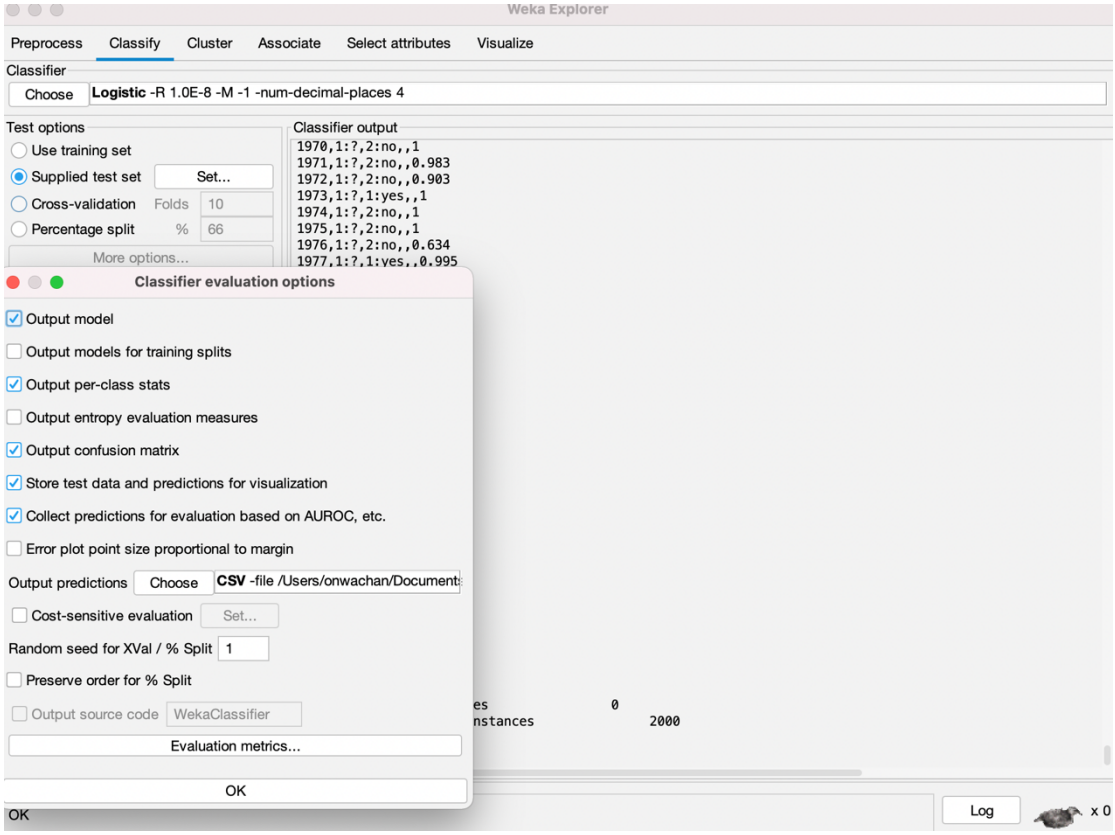


(Logistic model with 91.6262% accuracy)


Weka Explorer
Preprocess
Classify
Cluster
Associate
Select attributes
Visualize
Classifier
Choose
Logistic -R 1.0E-8 -M -1 -num-decimal-places 4
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)
21:56:12 - trees.J48
21:57:19 - functions.MultilayerPerceptron
22:04:05 - functions.MultilayerPerceptron
22:07:00 - functions.MultilayerPerceptron from file 'asm3.model'
22:31:44 - functions.Logistic
22:35:11 - functions.Logistic from file 'asm3.logistic.model'
Classifier output
Intercept -4.3707
Odds Ratios...
Variable Class yes
age='(29.666667-41.333333]' 1.1734
age='(-inf-29.666667]' 0.866
age='(53-64.666667]' 1.2248
age='(41.333333-53]' 0.8215
age='(76.333333-inf)' 0.9307
age='(64.666667-76.333333]' 0.6325
job=retired 0.7833
contact=telephone 0.3759
month=jun 1.7405
month=may 0.7627
month=jul 1.0262
month=sep 0.791
month=aug 1.2974
month=apr 1.0087
month=nov 0.5176
month=oct 0.8403
month=mar 8.1417
month=dec 2.0964
pdays 0.9686
previous 1.117
poutcome=nonexistent 1.199
poutcome=failure 0.6036
poutcome=success 2.357
emp.var.rate 0.2734
cons.price.idx 1.9588
cons.conf.idx 1.3674
euribor3m 0.9573
nr.employed 0.812
duration 141084399.4882
Status
OK
Log
x 0


Weka Explorer
Preprocess
Classify
Cluster
Associate
Select attributes
Visualize
Classifier
Choose
Logistic -R 1.0E-8 -M -1 -num-decimal-places 4
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)
21:56:12 - trees.J48
21:57:19 - functions.MultilayerPerceptron
22:04:05 - functions.MultilayerPerceptron
22:07:00 - functions.MultilayerPerceptron from file 'asm3.model'
22:31:44 - functions.Logistic
22:35:11 - functions.Logistic from file 'asm3.logistic.model'
Classifier output
811,2:no,2:no,,0.984
812,2:no,2:no,,0.95
813,2:no,2:no,,0.981
814,2:no,2:no,,0.898
815,2:no,2:no,,0.919
816,2:no,2:no,,0.992
817,2:no,2:no,,0.993
818,2:no,2:no,,0.991
819,2:no,2:no,,0.998
820,2:no,2:no,,0.993
821,2:no,2:no,,0.99
822,2:no,2:no,,0.978
823,2:no,2:no,,0.995
824,2:no,2:no,,0.943
=== Summary ===
Correctly Classified Instances 755 91.6262 %
Incorrectly Classified Instances 69 8.3738 %
Kappa statistic 0.4743
Mean absolute error 0.1167
Root mean squared error 0.2494
Total Number of Instances 824
=== Detailed Accuracy By Class ===
TP Rate FP Rate Precision Recall F-Measure MCC ROC Area
0.420 0.024 0.673 0.420 0.517 0.490 0.922
0.976 0.580 0.934 0.976 0.954 0.490 0.922
Weighted Avg. 0.916 0.520 0.906 0.916 0.908 0.490 0.922
=== Confusion Matrix ===
a b <-- classified as
37 51 | a = yes
18 718 | b = no
Status
OK
Log
x 0

Step 9&10 : I Load the logistic model obtained . Then use it to conduct prediction on the pre-processed file (new client data set, namely, **bank-new- client.csv**) provided. And output the prediction results to a CSV file as follows by re-evaluate the current test set.



Additionally, I assess the quality of new client prediction by yourself, i.e., the prediction quality on **bank-new-client.csv**. To conduct such assessment, I upload my CSV file for the prediction results on our course website under the section "Evaluation". Then I get a score indicating the accuracy of your prediction of new clients.

Assignments and Weka Data Mining Package

- [Weka] Please follow the instruction on the website to install the stable version (3.8) of Weka. It provides different links to suit different OS. Please select the one you are using (if possible, download it before the lab class).
- The bank marketing data set used in Weka notes: bank.csv
- Assignment 1 (Due: Jan 31 (Wed) 17:00): Data Preprocessing Using Weka Specification bank-additional.csv
- Assignment 2 (Due: Mar 05 (Tue) 12:00 Noon): Decision Tree Specification dataset.zip
- Assignment 3 (Due: Apr 18 (Thu) 17:00): Client Subscription Prediction Specification dataset.zip

Required Readings

- " Data Mining - Concepts and Techniques " , Jiawei Han, Micheline Kamber, and Jian Pei, 3rd edition, Elsevier Science, 2011. (e-book can be accessed online via CUHK library)

Recommended Readings

- " Data Mining : Practical Machine Learning Tools and Techniques " , Ian Witten, Eibe Frank, Mark A. Hall, Christopher J. Pal, Fourth edition. Amsterdam, : Elsevier 2017. (e-book can be accessed online via CUHK library)
- " Data Science for Business: what you need to know about data mining and data-analytic thinking " , F. Provost and T. Fawcett, O'Reilly, 2013. (e-book can be accessed online via CUHK library)

Final Exam

- The scope of the normal final exam covers the topics "model evaluation & selection", "practical considerations for classification learning", "classification - neural networks", "classification - logistic regression", and "clustering".
- The scope of the make-up final exam covers **ALL** lecture topics.
- Calculators are **allowed**.
- You cannot bring anything except calculators. Therefore, books, notes, cheat sheets are **NOT** allowed.

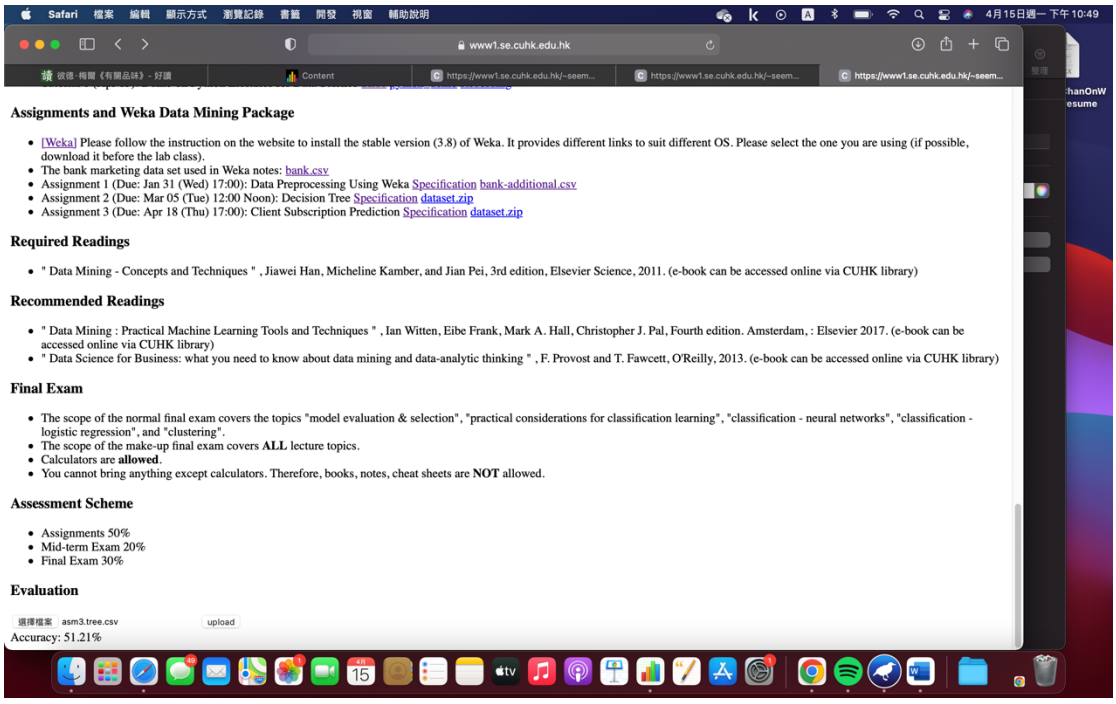
Assessment Scheme

- Assignments 50%
- Mid-term Exam 20%
- Final Exam 30%

Evaluation

選擇檔案 asm3.multilayer.csv upload
Accuracy: 50.85%

(Neural Network model with 50.85% predicted accuracy)



(Decision tree model with 51.21% predicted accuracy)

Assignments and Weka Data Mining Package

- [Weka] Please follow the instruction on the website to install the stable version (3.8) of Weka. It provides different links to suit different OS. Please select the one you are using (if possible, download it before the lab class).
- The bank marketing data set used in Weka notes: bank.csv
- Assignment 1 (Due: Jan 31 (Wed) 17:00): Data Preprocessing Using Weka Specification bank-additional.csv
- Assignment 2 (Due: Mar 05 (Tue) 12:00 Noon): Decision Tree Specification dataset.zip
- Assignment 3 (Due: Apr 18 (Thu) 17:00): Client Subscription Prediction Specification dataset.zip

Required Readings

- " Data Mining - Concepts and Techniques " , Jiawei Han, Micheline Kamber, and Jian Pei, 3rd edition, Elsevier Science, 2011. (e-book can be accessed online via CUHK library)

Recommended Readings

- " Data Mining : Practical Machine Learning Tools and Techniques " , Ian Witten, Eibe Frank, Mark A. Hall, Christopher J. Pal, Fourth edition. Amsterdam, : Elsevier 2017. (e-book can be accessed online via CUHK library)
- " Data Science for Business: what you need to know about data mining and data-analytic thinking " , F. Provost and T. Fawcett, O'Reilly, 2013. (e-book can be accessed online via CUHK library)

Final Exam

- The scope of the normal final exam covers the topics "model evaluation & selection", "practical considerations for classification learning", "classification - neural networks", "classification - logistic regression", and "clustering".
- The scope of the make-up final exam covers **ALL** lecture topics.
- Calculators are **allowed**.
- You cannot bring anything except calculators. Therefore, books, notes, cheat sheets are **NOT** allowed.

Assessment Scheme

- Assignments 50%
- Mid-term Exam 20%
- Final Exam 30%

Evaluation

Accuracy: 70.58%

(Logistic model that employed for delivering the bank for operation with 70.58% of predicted accuracy)