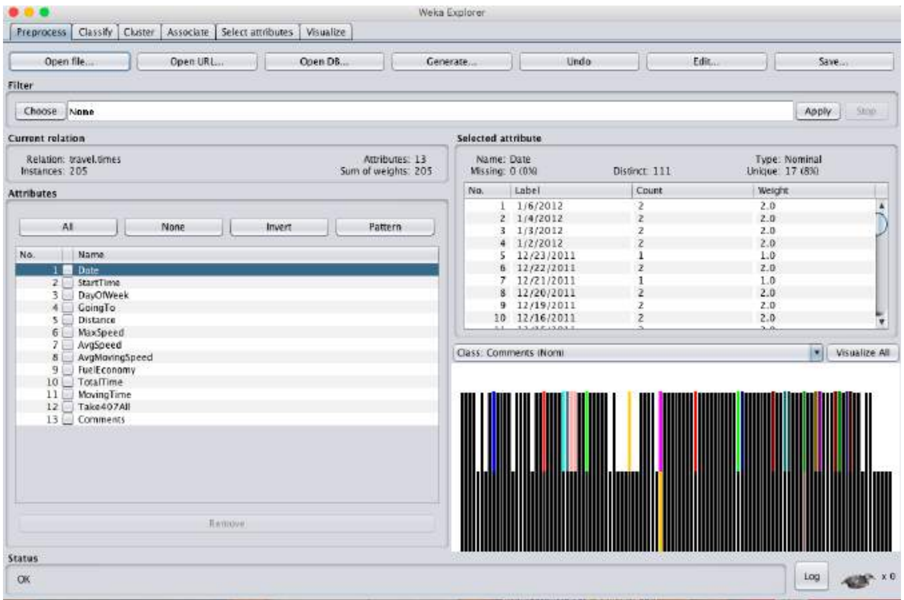
DATA PREPROCESSING AND ANALYSIS FOR DATASET USING WEKA

DESCRIPTION :

Consider a dataset of traveltimes.csv file where it contains the columns of (or) attributes as Date, StartTime, DayOfWeek, GoingTo, Distance, MaxSpeed, AvgSpeed, AvgMovingSpeed, FuelEconomy, TotalTime, MovingTime, Take407All comments.

PREPROCESS :



OBSERVATION :

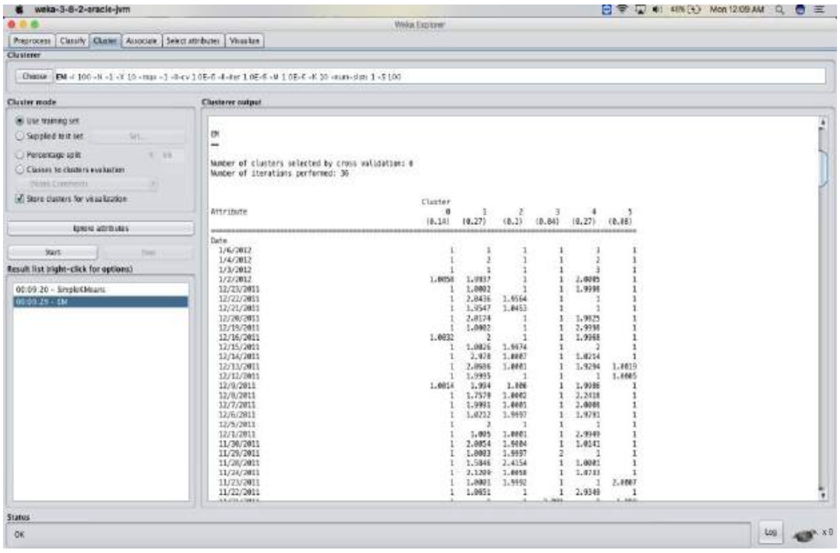
A. ATTRIBUTE TYPE :

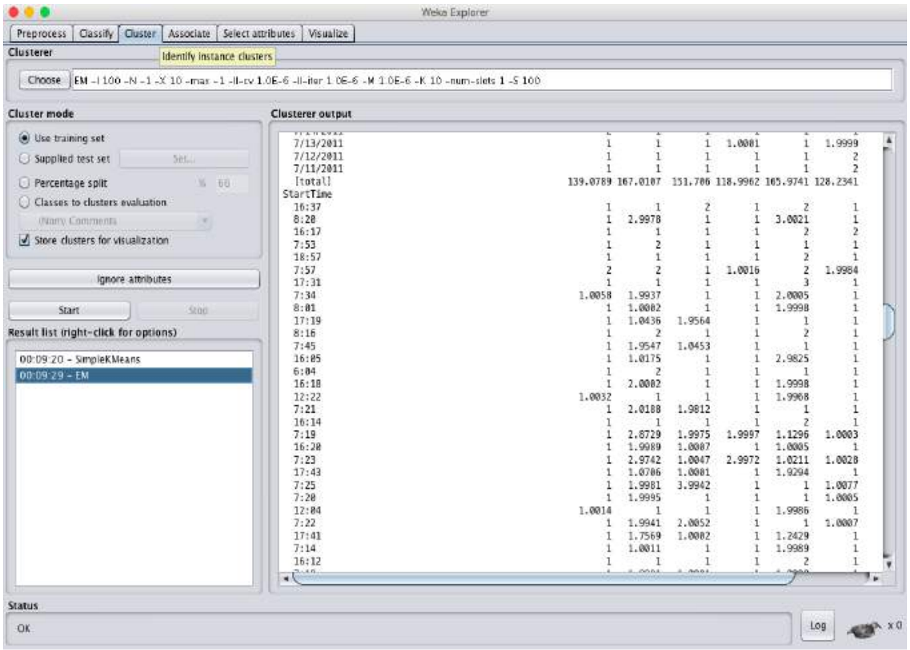
S.NO	ATTRIBUTE	TYPE
1.	Date	Nominal
2.	Start Time	Nominal
3.	Day Of Week	Nominal
4.	Going To	Nominal
5.	Distance	Numeric
6.	Max Speed	Numeric
7.	Avg Speed	Numeric
8.	Avg Moving Speed	Numeric
9.	Fuel Economy	Nominal
10.	Total Time	Numeric
11.	Moving Time	Numeric
12.	Comments	Nominal
13.	Take 407 All	Nominal

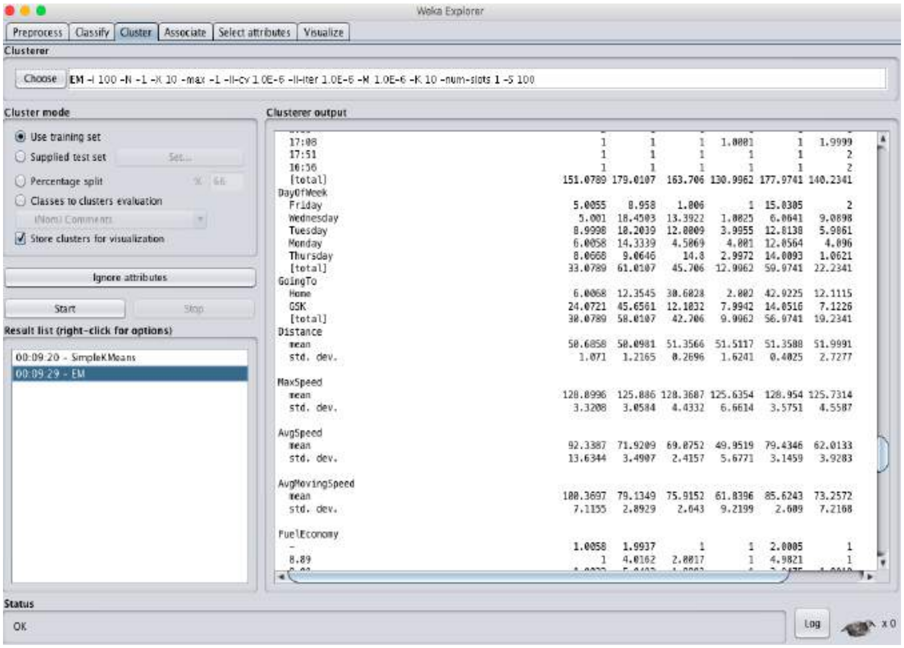
B. PERCENTAGE OF MISSING VALUES :

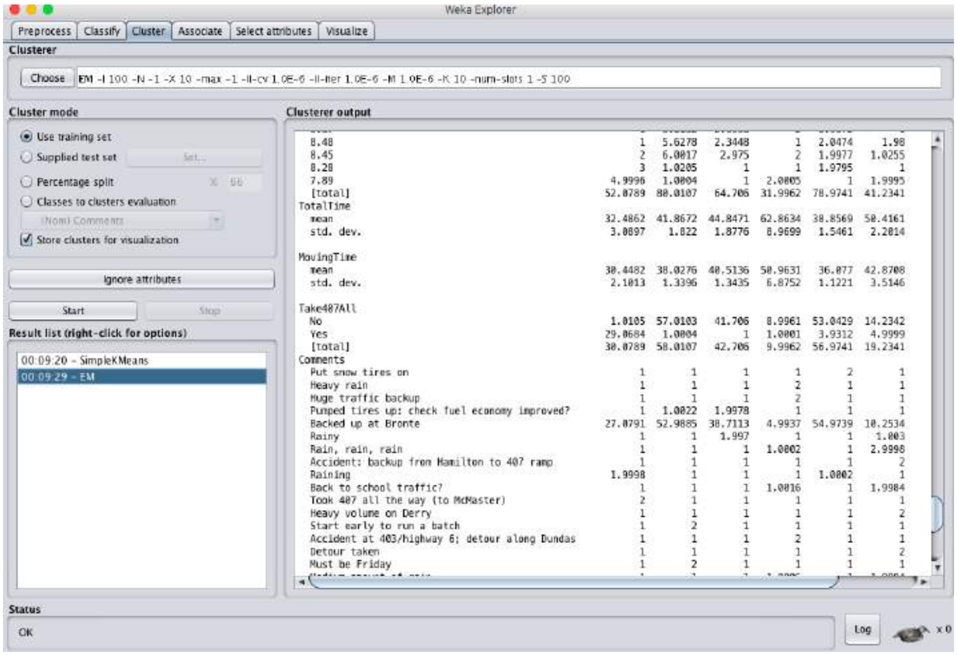
S.NO	ATTRIBUTE	Percentage Of Missing Values
1.	Date	0 %
2.	Start Time	0 %
3.	Day Of Week	0 %
4.	Going To	0 %
5.	Distance	0 %
6.	Max Speed	0 %
7.	Avg Speed	0 %
8.	Avg Moving Speed	0 %
9.	Fuel Economy	8 %
10.	Total Time	0 %
11.	Moving Time	0 %
12.	Comments	88 %
13.	Take 407 All	0 %

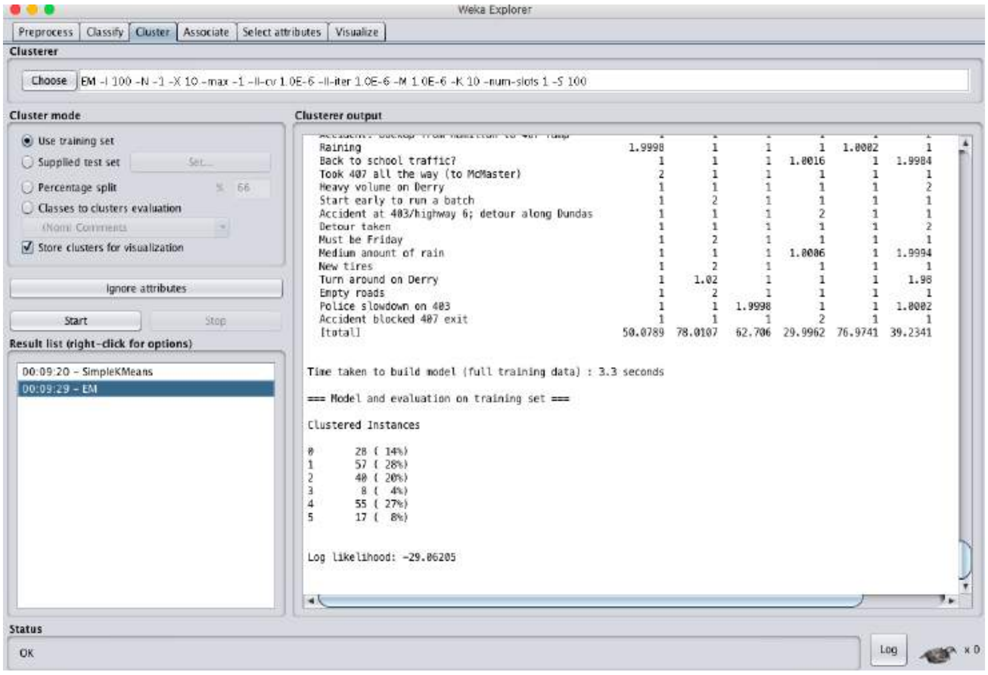
C. MIN, MAX, MEAN, STANDARD DEVIATION :


weka-3-8-2-oracle-jvm
Weka Explorer
Preprocess
Classify
Cluster
Associate
Select attributes
Visualize
Clusterer
Choose
Cluster mode
Use training set
Supplied test set
Percentage split
Classes to clusters evaluation
Store clusters for visualization
Ignore attributes
Start
Stop
Result list (right-click for options)
00:09:20 - SimpleKMeans
00:09:29 - EM
Clusterer output
EM
Number of clusters selected by cross validation: 6
Number of iterations performed: 36
Attribute
Cluster
Date
Status
OK
Log


Weka Explorer
Preprocess
Classify
Cluster
Associate
Select attributes
Visualize
Clusterer
Identify instance clusters
Choose
EM -I 100 -N -1 -X 10 -max -1 -ll-cv 1.0E-6 -ll-iter 1.0E-6 -M 1.0E-6 -K 10 -num-slots 1 -S 100
Cluster mode
Use training set
Supplied test set
Set...
Percentage split
% 66
Classes to clusters evaluation
(Nom) Comments
Store clusters for visualization
Ignore attributes
Start
Stop
Result list (right-click for options)
00:09:20 - SimpleKMeans
00:09:29 - EM
Clusterer output
7/13/2011 1 1 1 1.0001 1 1.9999
7/12/2011 1 1 1 1 1 2
7/11/2011 1 1 1 1 1 2
[total] 139.0789 167.0107 151.706 118.9962 165.9741 128.2341
StartTime
16:37 1 1 2 1 2 1
8:20 1 2.9978 1 1 3.0021 1
16:17 1 1 1 1 2 2
7:53 1 2 1 1 1 1
18:57 1 1 1 1 2 1
7:57 2 2 1 1.0016 2 1.9984
17:31 1 1 1 1 3 1
7:34 1.0058 1.9937 1 1 2.0005 1
8:01 1 1.0002 1 1 1.9998 1
17:19 1 1.0436 1.9564 1 1 1
8:16 1 2 1 1 2 1
7:45 1 1.9547 1.0453 1 1 1
16:05 1 1.0175 1 1 2.9825 1
6:04 1 2 1 1 1 1
16:18 1 2.0002 1 1 1.9998 1
12:22 1.0032 1 1 1 1.9968 1
7:21 1 2.0188 1.9812 1 1 1
16:14 1 1 1 1 2 1
7:19 1 2.8729 1.9975 1.9997 1.1296 1.0003
16:20 1 1.9989 1.0007 1 1.0005 1
7:23 1 2.9742 1.0047 2.9972 1.0211 1.0028
17:43 1 1.0706 1.0001 1 1.9294 1
7:25 1 1.9981 3.9942 1 1 1.0077
7:20 1 1.9995 1 1 1 1.0005
12:04 1.0014 1 1 1 1.9986 1
7:22 1 1.9941 2.0052 1 1 1.0007
17:41 1 1.7569 1.0002 1 1.2429 1
7:14 1 1.0011 1 1 1.9989 1
16:12 1 1 1 1 2 1
Status
OK
Log


Weka Explorer
Preprocess
Classify
Cluster
Associate
Select attributes
Visualize
Clusterer
Choose
EM -I 100 -N -1 -X 10 -max -1 -ll-cv 1.0E-6 -ll-iter 1.0E-6 -M 1.0E-6 -K 10 -num-slots 1 -S 100
Cluster mode
Use training set
Supplied test set
Percentage split
Classes to clusters evaluation
Store clusters for visualization
Ignore attributes
Start
Stop
Result list (right-click for options)
00:09:20 - SimpleKMeans
00:09:29 - EM
Clusterer output
Status
OK
Log


Weka Explorer
Preprocess
Classify
Cluster
Associate
Select attributes
Visualize
Clusterer
Choose
EM -I 100 -N -1 -X 10 -max -1 -ll-cv 1.0E-6 -ll-iter 1.0E-6 -M 1.0E-6 -K 10 -num-slots 1 -S 100
Cluster mode
Use training set
Supplied test set
Percentage split
Classes to clusters evaluation
Store clusters for visualization
Ignore attributes
Start
Stop
Result list (right-click for options)
00:09:20 - SimpleKMeans
00:09:29 - EM
Clusterer output
Status
OK
Log



DATA SEGMENTATION BYK- MEANS CLUSTER USING WEKA AND R-TOOL

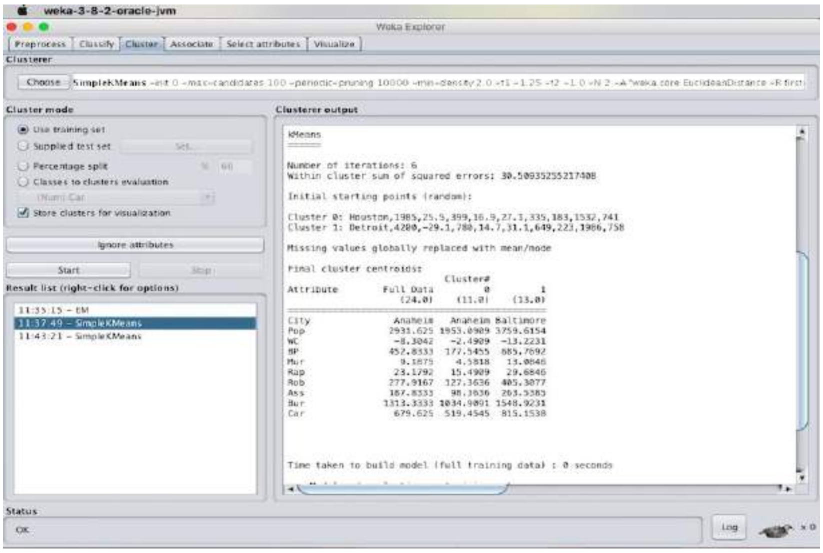
DESCRIPTION :

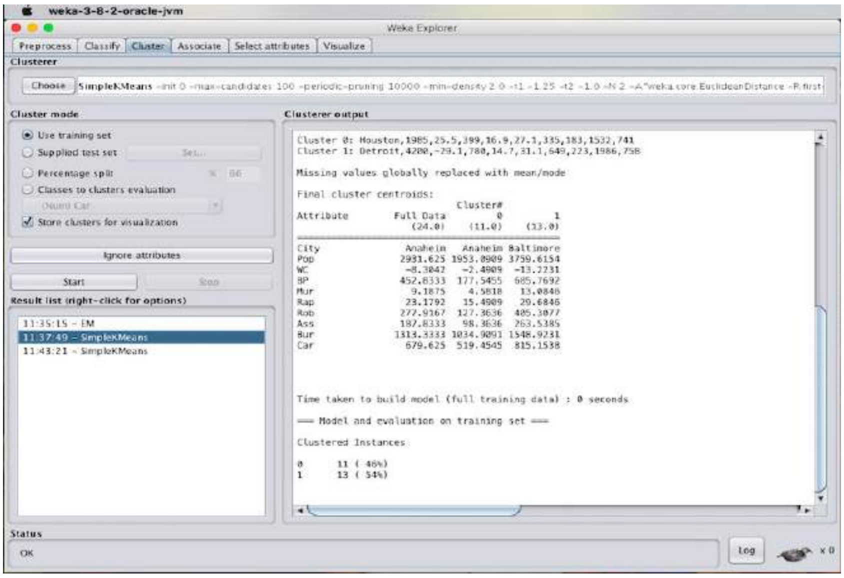
Consider a dataset of citycrimes.csv file of which it contains the attributes are City, Pop, WC, BP, Mur, Rap, Rob, Ass, Bus and car for the performance of the dataset by applying the K-means algorithm in weka and as well using R- tool.

❖ **USING WEKA TOOL :**

STEPS INVOLVED :

- Choose a set of attributes for clustering and for giving a motivation.
- Choose the dataset and import the dataset into Weka tool.
- Cluster the dataset and choose simple K-means algorithm and give the motivation.

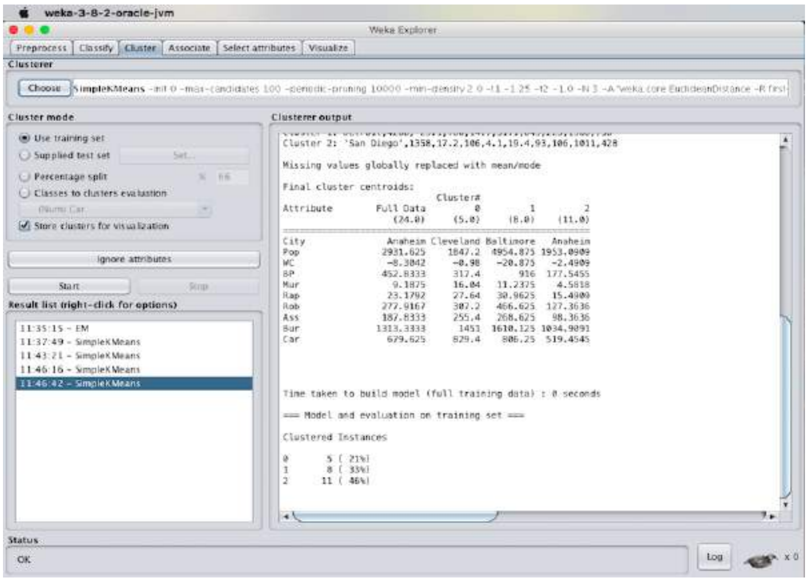


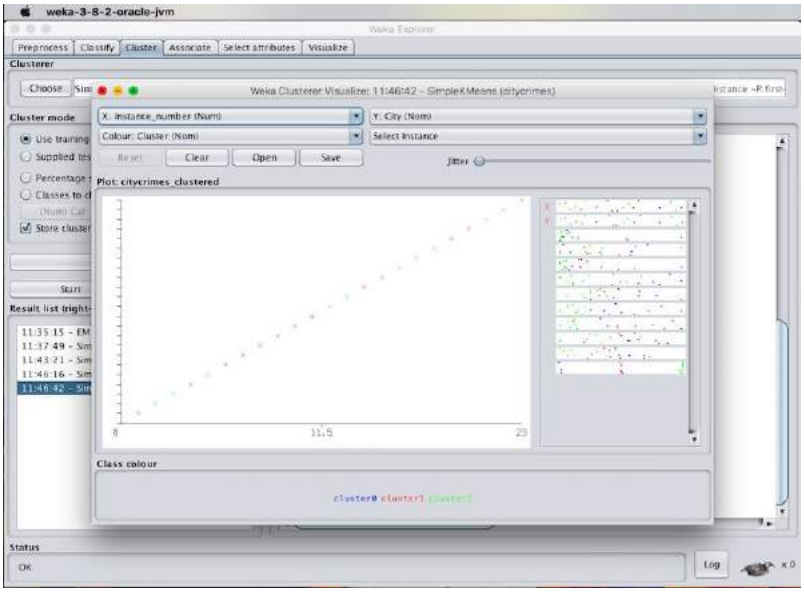


A. Experiment with atleast 2 different number of clusters but with same seed values:

STEPS INVOLVED :

- Compare the two different clusters but with the same seed values.
- Change the number of clusters value and need not to change the seed value.
- Apply the K-means algorithm and start executing the algorithm.





DATASEGMENTATION BY EXPECTATION MAXIMISATION ALGORITHM THROUGH WEKA

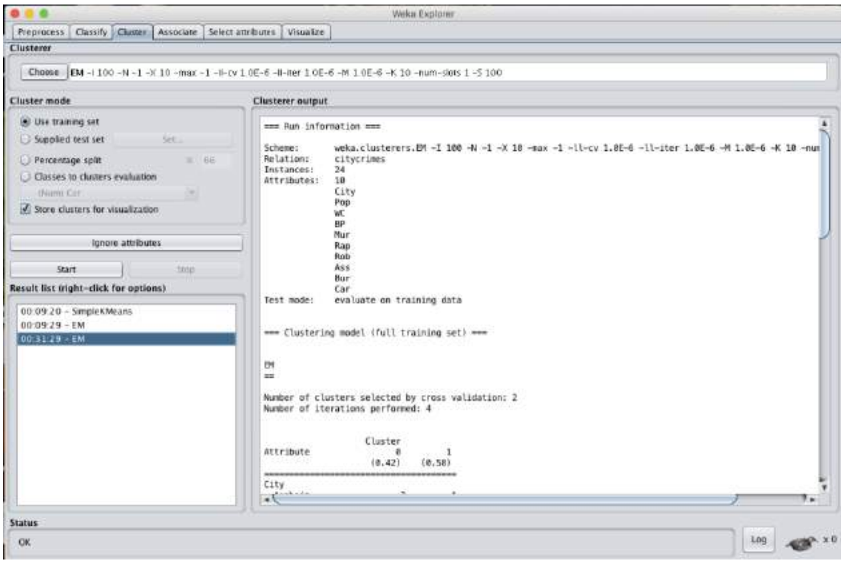
DESCRIPTION :

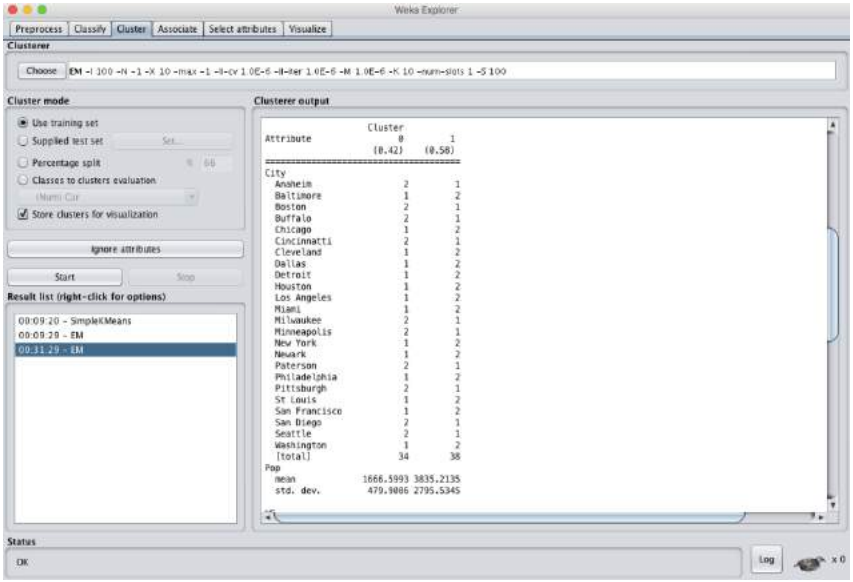
Consider a dataset of citycrimes.csv file of which it contains the attributes are City, Pop, WC, BP, Mur, Rap, Rob, Ass, Bus and car for the performance of the dataset by applying the K-means algorithm in weka and as well using R- tool.

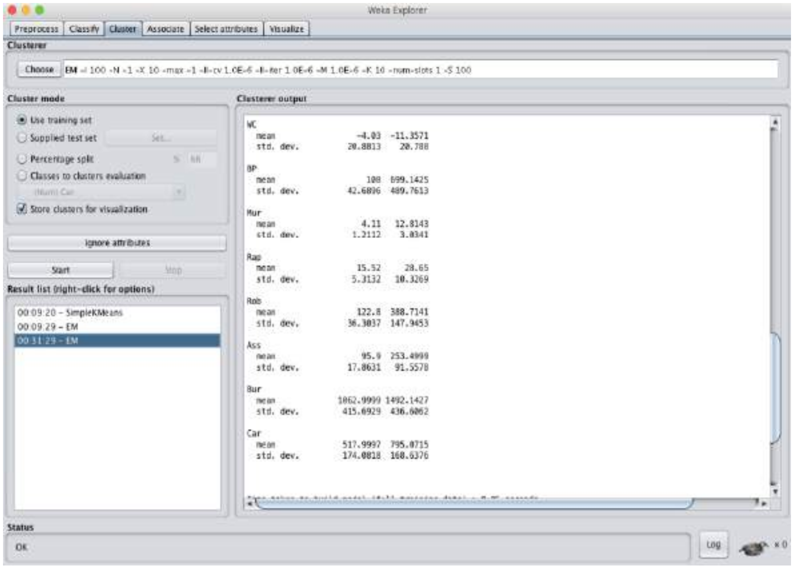
When the clustering is been made through the expectation maximization algorithm by setting minimum standard deviation values then the results will be of the following :

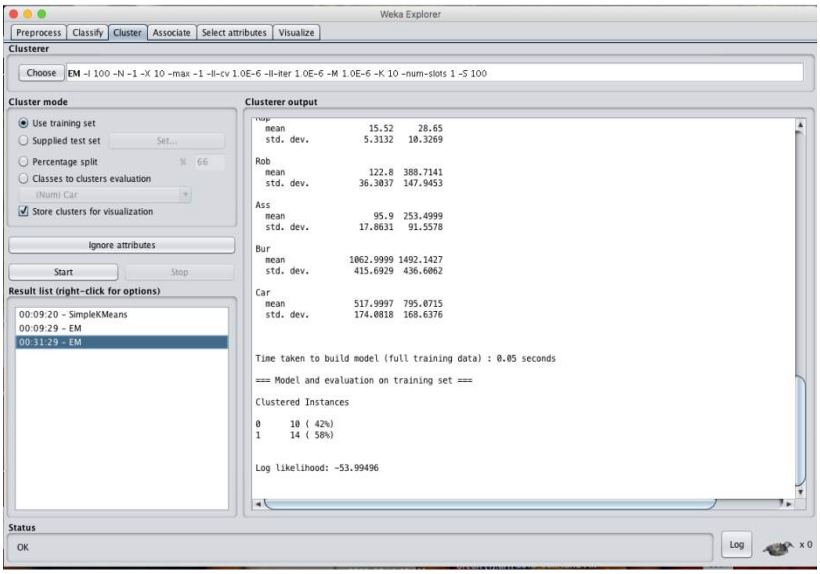
- ❖ **Steps Involved :**

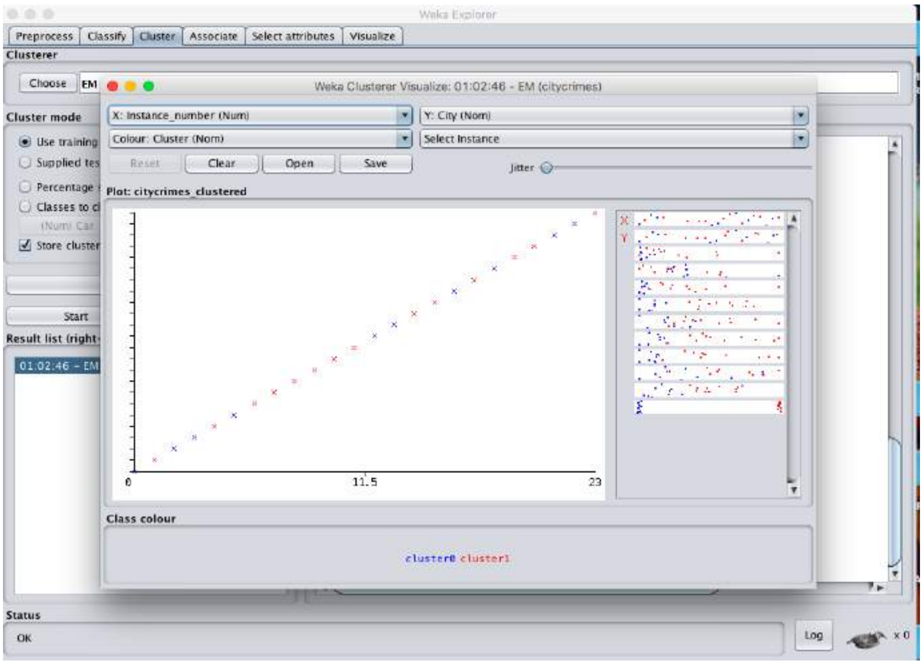
- Initially, load the dataset into the weka tool and check for all the attributes present in the dataset.
- Then move to cluster panel and apply the EM algorithm technique for the datasheet.
- Finally, Observe the results that are obtained.



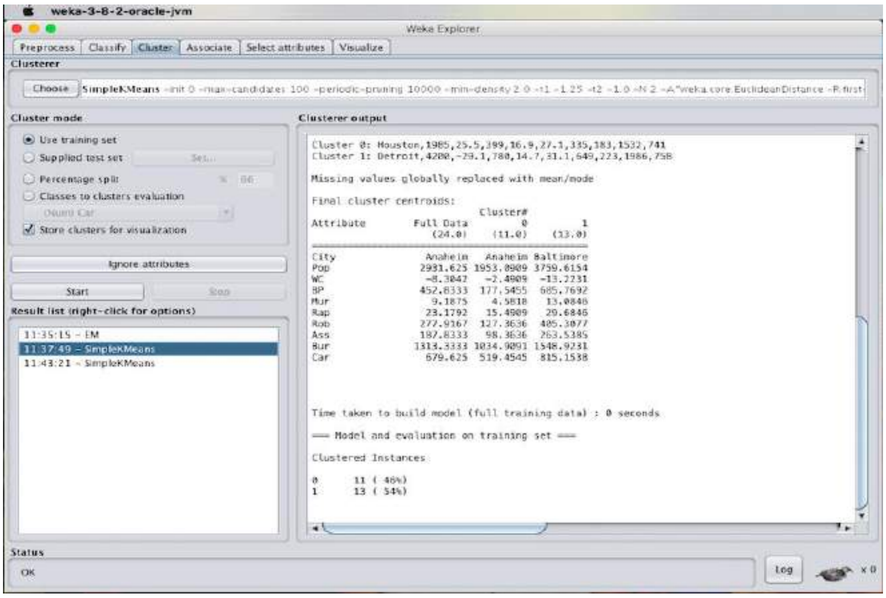



Weka Explorer
Preprocess
Classify
Cluster
Associate
Select attributes
Visualize
Clusterer
Choose
EM -I 100 -N -1 -X 10 -max -1 -ll-cv 1.0E-6 -ll-iter 1.0E-6 -M 1.0E-6 -K 10 -num-slots 1 -S 100
Cluster mode
Use training set
Supplied test set
Set...
Percentage split
% 66
Classes to clusters evaluation
(Num) Car
Store clusters for visualization
Ignore attributes
Start
Stop
Result list (right-click for options)
00:09:20 - SimpleKMeans
00:09:29 - EM
00:31:29 - EM
Clusterer output
WC
mean -4.03 -11.3571
std. dev. 20.8813 20.788
BP
mean 108 699.1425
std. dev. 42.6896 489.7613
Mur
mean 4.11 12.8143
std. dev. 1.2112 3.8341
Rap
mean 15.52 28.65
std. dev. 5.3132 10.3269
Rob
mean 122.8 388.7141
std. dev. 36.3037 147.9453
Ass
mean 95.9 253.4999
std. dev. 17.8631 91.5578
Bur
mean 1062.9999 1492.1427
std. dev. 415.6929 436.6062
Car
mean 517.9997 795.0715
std. dev. 174.0818 168.6376
Status
OK
Log
x 0


Weka Explorer
Preprocess
Classify
Cluster
Associate
Select attributes
Visualize
Clusterer
Choose
EM -I 100 -N -1 -X 10 -max -1 -ll-cv 1.0E-6 -ll-iter 1.0E-6 -M 1.0E-6 -K 10 -num-slots 1 -S 100
Cluster mode
Use training set
Supplied test set
Set...
Percentage split
% 66
Classes to clusters evaluation
(Num) Car
Store clusters for visualization
Ignore attributes
Start
Stop
Result list (right-click for options)
00:09:20 - SimpleKMeans
00:09:29 - EM
00:31:29 - EM
Clusterer output
mean 15.52 28.65
std. dev. 5.3132 10.3269
Rob
mean 122.8 388.7141
std. dev. 36.3037 147.9453
Ass
mean 95.9 253.4999
std. dev. 17.8631 91.5578
Bur
mean 1062.9999 1492.1427
std. dev. 415.6929 436.6062
Car
mean 517.9997 795.0715
std. dev. 174.0818 168.6376
Time taken to build model (full training data) : 0.05 seconds
=== Model and evaluation on training set ===
Clustered Instances
0 10 (42%)
1 14 (58%)
Log likelihood: -53.99496
Status
OK
Log
x 0



❖ K- MEANS ALGORITHM:



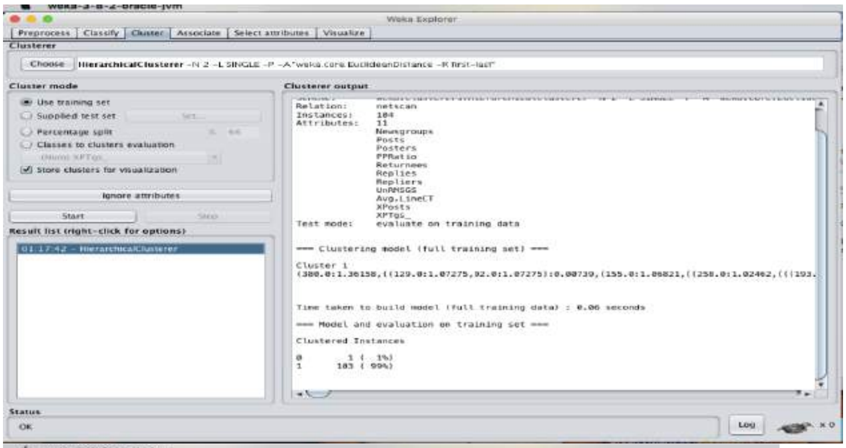
DATA SEGMENTATION BY COBWEB – HIERARCHIAL CLUSTERING ALGORITHM USING WEKA TOOL

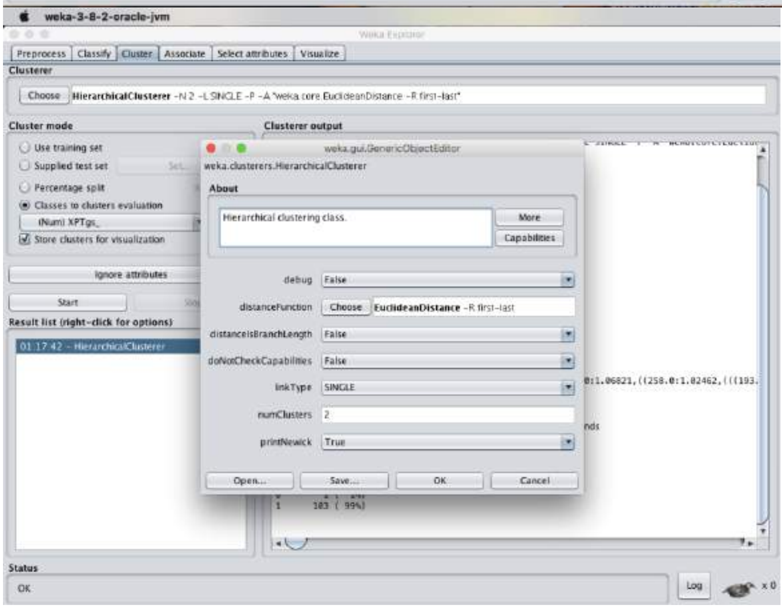
.

DESCRIPTION:

Consider a dataset netscan.csv where it contains the attributes of Newsgroups, posts, posters, PPRatio, Returnees, Replies, Repliers, UnRMsgs, Avg.LineCT, Xports, XPTgs. Each attribute will have different types of the meanings.

HIERARCHIIAL CLUSTERING:





FREQUENT PATTERN MINING USING ASSOCIATION RULE THROUGH
WEKA AND R TOOLS

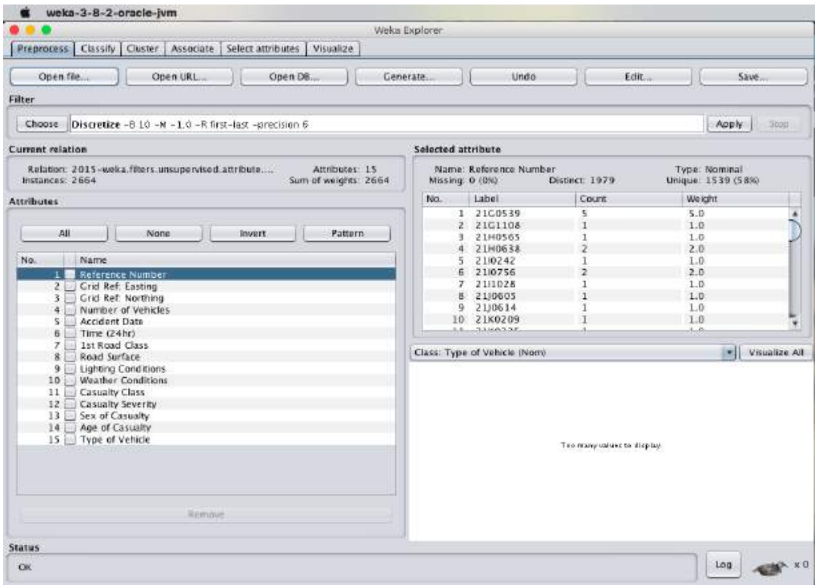
DESCRIPTION :

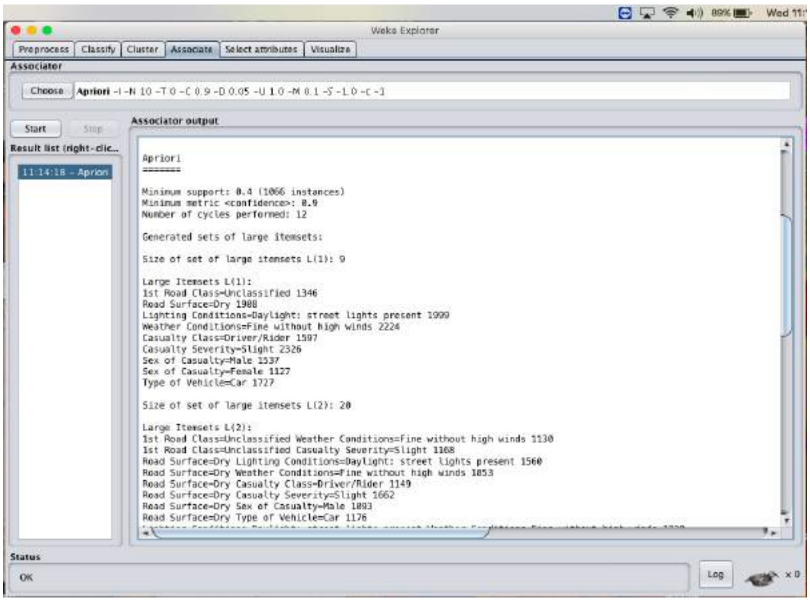
Consider a dataset of 2015.csv file of which it contains the attributes are Reference Number, Grid ref: Easting, Grid Ref: Northing, Number of vehicles, Accident date, Time(24 hr), 1st Road class, Road Surface, Lighting conditions, Weather conditions, casuality class, Sex of casuality, Age of casuality, Type of casuality for the performance of the dataset by applying the Apriori algorithm in weka and as well using R- tool.

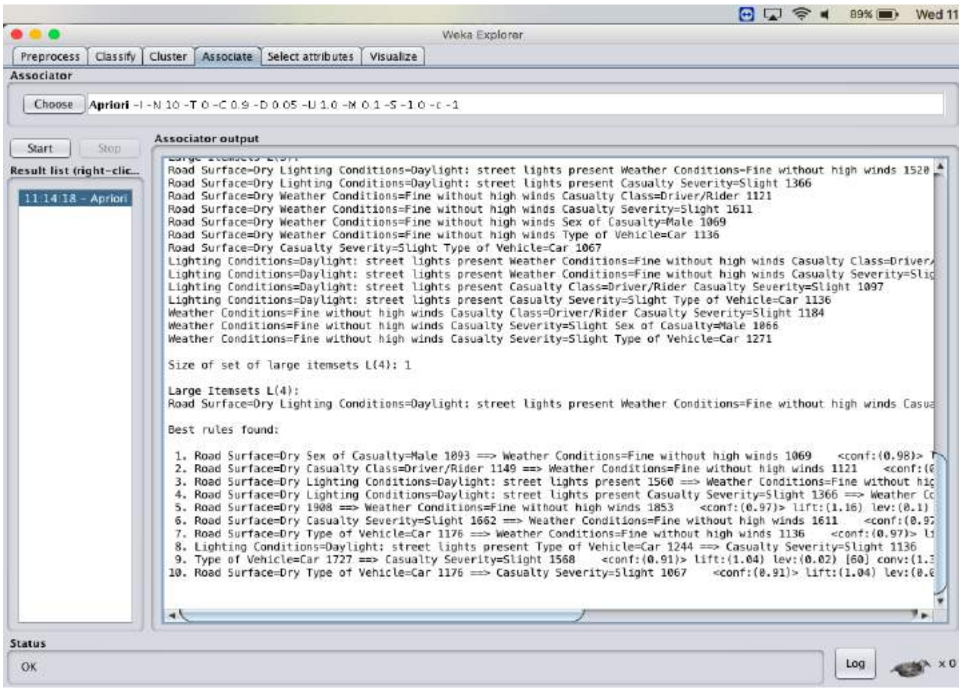
- **USING WEKA TOOL :**

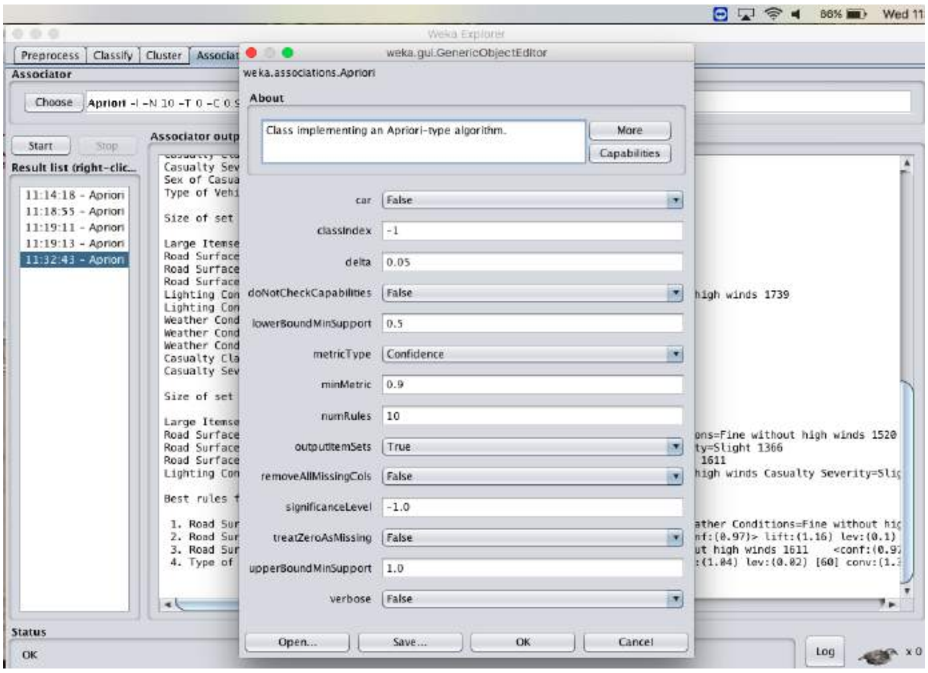
STEPS INVOLVED :

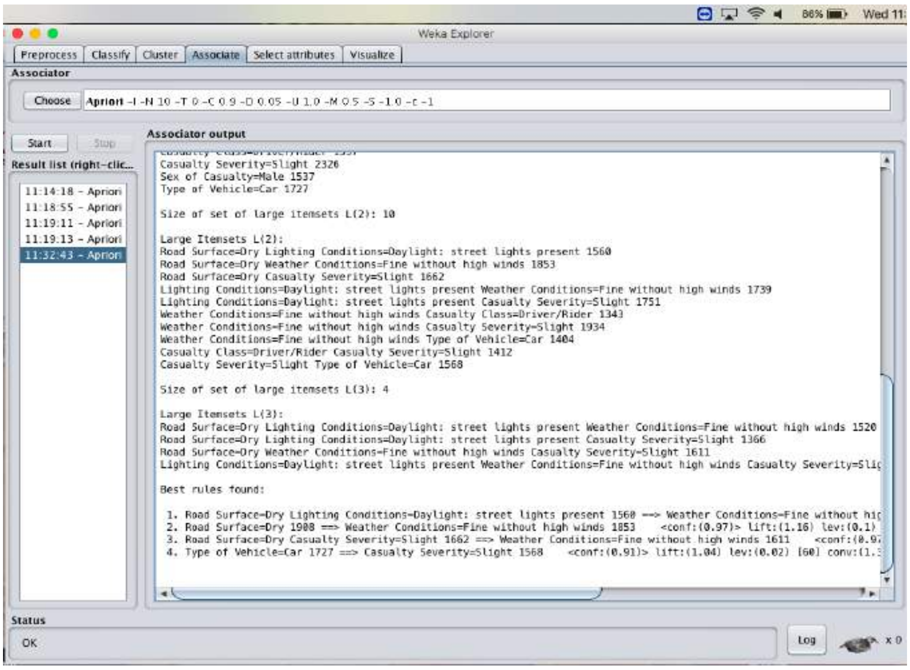
- Choose a set of attributes for clustering and for giving a motivation.
- Choose the dataset and import the dataset into Weka tool.
- Discretize the attributes from numeric to nominal to perform the algorithm.
- Cluster the dataset and choose simple Apriori algorithm.
- Set the Upper bound min_sup and lower bound min_sup values.

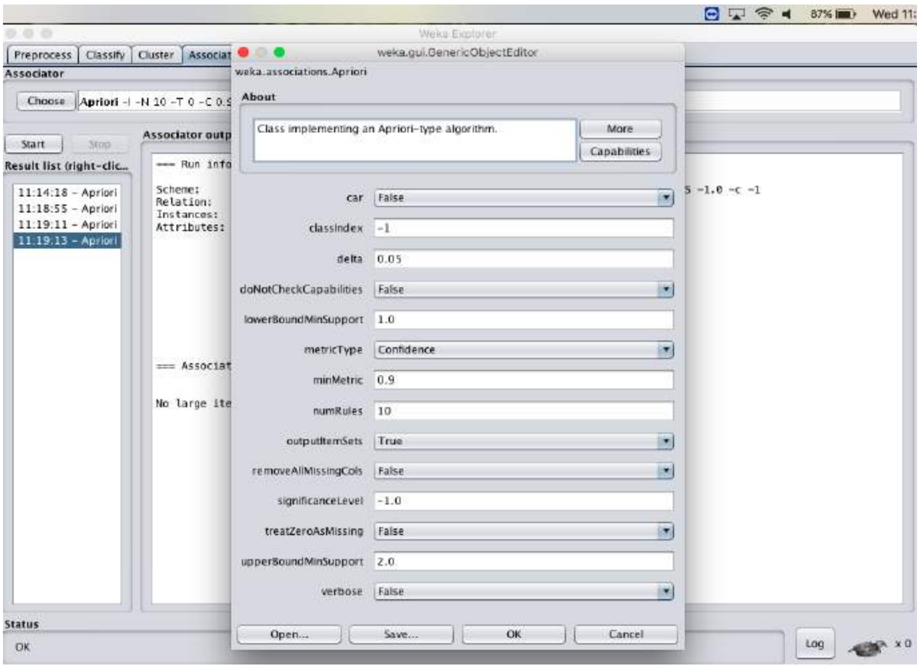

weka-3-8-2-oracle-jvm
Weka Explorer
Preprocess
Classify
Cluster
Associate
Select attributes
Visualize
Open file...
Open URL...
Open DB...
Generate...
Undo
Edit...
Save...
Filter
Choose
Discretize -B 10 -M -1.0 -R first-last -precision 6
Apply
Stop
Current relation
Relation: 2015-weka.filters.unsupervised.attribute....
Attributes: 15
Instances: 2664
Sum of weights: 2664
Selected attribute
Name: Reference Number
Type: Nominal
Missing: 0 (0%)
Distinct: 1979
Unique: 1539 (58%)
No. Label Count Weight
1 21C0539 5 5.0
2 21G1108 1 1.0
3 21H0565 1 1.0
4 21H0638 2 2.0
5 21I0242 1 1.0
6 21I0756 2 2.0
7 21I1028 1 1.0
8 21J0605 1 1.0
9 21J0614 1 1.0
10 21K0209 1 1.0
Attributes
All
None
Invert
Pattern
No. Name
1 Reference Number
2 Grid Ref: Easting
3 Grid Ref: Northing
4 Number of Vehicles
5 Accident Date
6 Time (24hr)
7 1st Road Class
8 Road Surface
9 Lighting Conditions
10 Weather Conditions
11 Casualty Class
12 Casualty Severity
13 Sex of Casualty
14 Age of Casualty
15 Type of Vehicle
Remove
Class: Type of Vehicle (Nom)
Visualize All
Too many values to display
Status
OK
Log
x 0

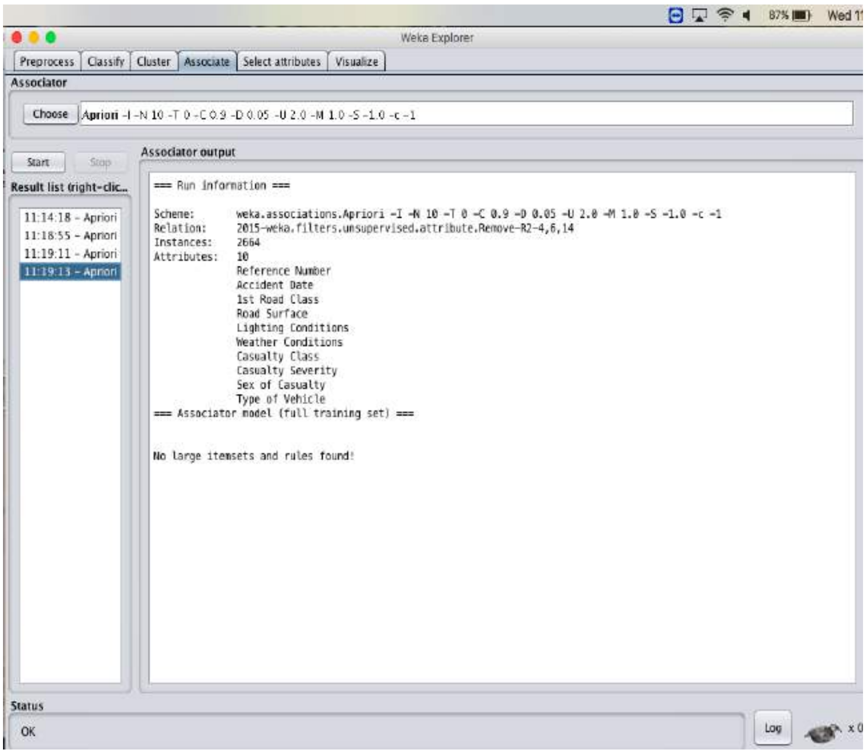

Weka Explorer
Preprocess
Classify
Cluster
Associate
Select attributes
Visualize
Associator
Choose
Apriori -I -N 10 -T 0 -C 0.9 -D 0.05 -U 1.0 -M 0.1 -S -1.0 -c -1
Start
Stop
Associator output
Result list (right-clic...
11:14:18 - Apriori
Apriori
=======
Minimum support: 0.4 (1066 instances)
Minimum metric <confidence>: 0.9
Number of cycles performed: 12
Generated sets of large itemsets:
Size of set of large itemsets L(1): 9
Large Itemsets L(1):
1st Road Class=Unclassified 1346
Road Surface=Dry 1988
Lighting Conditions=Daylight: street lights present 1999
Weather Conditions=Fine without high winds 2224
Casualty Class=Driver/Rider 1597
Casualty Severity=Slight 2326
Sex of Casualty=Male 1537
Sex of Casualty=Female 1127
Type of Vehicle=Car 1727
Size of set of large itemsets L(2): 20
Large Itemsets L(2):
1st Road Class=Unclassified Weather Conditions=Fine without high winds 1130
1st Road Class=Unclassified Casualty Severity=Slight 1108
Road Surface=Dry Lighting Conditions=Daylight: street lights present 1560
Road Surface=Dry Weather Conditions=Fine without high winds 1853
Road Surface=Dry Casualty Class=Driver/Rider 1149
Road Surface=Dry Casualty Severity=Slight 1662
Road Surface=Dry Sex of Casualty=Male 1093
Road Surface=Dry Type of Vehicle=Car 1176
Status
OK
Log
x 0


Weka Explorer
Preprocess
Classify
Cluster
Associate
Select attributes
Visualize
Associator
Choose
Apriori -I -N 10 -T 0 -C 0.9 -D 0.05 -U 1.0 -M 0.1 -S -1.0 -c -1
Start
Stop
Result list (right-clic...
11:14:18 - Apriori
Associator output
Road Surface=Dry Lighting Conditions=Daylight: street lights present Weather Conditions=Fine without high winds 1520
Road Surface=Dry Lighting Conditions=Daylight: street lights present Casualty Severity=Slight 1366
Road Surface=Dry Weather Conditions=Fine without high winds Casualty Class=Driver/Rider 1121
Road Surface=Dry Weather Conditions=Fine without high winds Casualty Severity=Slight 1611
Road Surface=Dry Weather Conditions=Fine without high winds Sex of Casualty=Male 1069
Road Surface=Dry Weather Conditions=Fine without high winds Type of Vehicle=Car 1136
Road Surface=Dry Casualty Severity=Slight Type of Vehicle=Car 1067
Lighting Conditions=Daylight: street lights present Weather Conditions=Fine without high winds 1097
Lighting Conditions=Daylight: street lights present Casualty Severity=Slight Type of Vehicle=Car 1136
Weather Conditions=Fine without high winds Casualty Class=Driver/Rider Casualty Severity=Slight 1184
Weather Conditions=Fine without high winds Casualty Severity=Slight Sex of Casualty=Male 1066
Weather Conditions=Fine without high winds Casualty Severity=Slight Type of Vehicle=Car 1271
Size of set of large itemsets L(4): 1
Large Itemsets L(4):
Best rules found:
1. Road Surface=Dry Sex of Casualty=Male 1093 ==> Weather Conditions=Fine without high winds 1069 <conf:(0.98)>
2. Road Surface=Dry Casualty Class=Driver/Rider 1149 ==> Weather Conditions=Fine without high winds 1121
3. Road Surface=Dry Lighting Conditions=Daylight: street lights present 1560 ==> Weather Conditions=Fine without high winds
4. Road Surface=Dry Lighting Conditions=Daylight: street lights present Casualty Severity=Slight 1366 ==> Weather Co
5. Road Surface=Dry 1908 ==> Weather Conditions=Fine without high winds 1853 <conf:(0.97)> lift:(1.16) lev:(0.1)
6. Road Surface=Dry Casualty Severity=Slight 1662 ==> Weather Conditions=Fine without high winds 1611
7. Road Surface=Dry Type of Vehicle=Car 1176 ==> Weather Conditions=Fine without high winds 1136 <conf:(0.97)>
8. Lighting Conditions=Daylight: street lights present Type of Vehicle=Car 1244 ==> Casualty Severity=Slight 1136
9. Type of Vehicle=Car 1727 ==> Casualty Severity=Slight 1568 <conf:(0.91)> lift:(1.04) lev:(0.02) [60] conv:(1.3
10. Road Surface=Dry Type of Vehicle=Car 1176 ==> Casualty Severity=Slight 1067 <conf:(0.91)> lift:(1.04)
Status
OK
Log


Weka Explorer
weka.gui.GenericObjectEditor
weka.associations.Apriori
About
Class implementing an Apriori-type algorithm.
More
Capabilities
car False
classIndex -1
delta 0.05
doNotCheckCapabilities False
lowerBoundMinSupport 0.5
metricType Confidence
minMetric 0.9
numRules 10
outputItemSets True
removeAllMissingCols False
significanceLevel -1.0
treatZeroAsMissing False
upperBoundMinSupport 1.0
verbose False
Open...
Save...
OK
Cancel
Result list (right-clic...
11:14:18 - Apriori
11:18:55 - Apriori
11:19:11 - Apriori
11:19:13 - Apriori
11:32:43 - Apriori
Status
OK
Log


Weka Explorer
Preprocess
Classify
Cluster
Associate
Select attributes
Visualize
Associator
Choose
Apriori -I -N 10 -T 0 -C 0.9 -D 0.05 -U 1.0 -M 0.5 -S -1.0 -c -1
Start
Stop
Associator output
Result list (right-clic...
11:14:18 - Apriori
11:18:55 - Apriori
11:19:11 - Apriori
11:19:13 - Apriori
11:32:43 - Apriori
Casualty Severity=Slight 2326
Sex of Casualty=Male 1537
Type of Vehicle=Car 1727
Size of set of large itemsets L(2): 10
Large Itemsets L(2):
Road Surface=Dry Lighting Conditions=Daylight: street lights present 1560
Road Surface=Dry Weather Conditions=Fine without high winds 1853
Road Surface=Dry Casualty Severity=Slight 1662
Lighting Conditions=Daylight: street lights present Weather Conditions=Fine without high winds 1739
Lighting Conditions=Daylight: street lights present Casualty Severity=Slight 1751
Weather Conditions=Fine without high winds Casualty Class=Driver/Rider 1343
Weather Conditions=Fine without high winds Casualty Severity=Slight 1934
Weather Conditions=Fine without high winds Type of Vehicle=Car 1404
Casualty Class=Driver/Rider Casualty Severity=Slight 1412
Casualty Severity=Slight Type of Vehicle=Car 1568
Size of set of large itemsets L(3): 4
Large Itemsets L(3):
Road Surface=Dry Lighting Conditions=Daylight: street lights present Weather Conditions=Fine without high winds 1520
Road Surface=Dry Lighting Conditions=Daylight: street lights present Casualty Severity=Slight 1366
Road Surface=Dry Weather Conditions=Fine without high winds Casualty Severity=Slight 1611
Lighting Conditions=Daylight: street lights present Weather Conditions=Fine without high winds Casualty Severity=Slig
Best rules found:
1. Road Surface=Dry Lighting Conditions=Daylight: street lights present 1560 ==> Weather Conditions=Fine without hig
2. Road Surface=Dry 1908 ==> Weather Conditions=Fine without high winds 1853 <conf:(0.97)> lift:(1.16) lev:(0.1)
3. Road Surface=Dry Casualty Severity=Slight 1662 ==> Weather Conditions=Fine without high winds 1611 <conf:(0.97
4. Type of Vehicle=Car 1727 ==> Casualty Severity=Slight 1568 <conf:(0.91)> lift:(1.04) lev:(0.02) [60] conv:(1.1
Status
OK
Log
x 0


Weka Explorer
Preprocess
Classify
Cluster
Associat
Associator
Choose
Apriori -I -N 10 -T 0 -C 0.9
Start
Stop
Associator outp
Result list (right-clic...
11:14:18 - Apriori
11:18:55 - Apriori
11:19:11 - Apriori
11:19:13 - Apriori
=== Run info
Scheme:
Relation:
Instances:
Attributes:
=== Associat
No large ite
S -1.0 -c -1
weka.gui.GenericObjectEditor
weka.associations.Apriori
About
Class implementing an Apriori-type algorithm.
More
Capabilities
car False
classIndex -1
delta 0.05
doNotCheckCapabilities False
lowerBoundMinSupport 1.0
metricType Confidence
minMetric 0.9
numRules 10
outputItemSets True
removeAllMissingCols False
significanceLevel -1.0
treatZeroAsMissing False
upperBoundMinSupport 2.0
verbose False
Open...
Save...
OK
Cancel
Status
OK
Log
x 0



Weka Explorer
Preprocess
Classify
Cluster
Associate
Select attributes
Visualize
Associator
Choose
Apriori -I -N 10 -T 0 -C 0.9 -D 0.05 -U 2.0 -M 1.0 -S -1.0 -c -1
Start
Stop
Result list (right-clic...
11:14:18 - Apriori
11:18:55 - Apriori
11:19:11 - Apriori
11:19:13 - Apriori
Associator output
=== Run information ===
Scheme: weka.associations.Apriori -I -N 10 -T 0 -C 0.9 -D 0.05 -U 2.0 -M 1.0 -S -1.0 -c -1
Relation: 2015-weka.filters.unsupervised.attribute.Remove-R2-4,6,14
Instances: 2664
Attributes: 10
Reference Number
Accident Date
1st Road Class
Road Surface
Lighting Conditions
Weather Conditions
Casualty Class
Casualty Severity
Sex of Casualty
Type of Vehicle
=== Associator model (full training set) ===
No large itemsets and rules found!
Status
OK
Log

FREQUENT PATTERN MINING USING FP GROWTH THROUGH WEKA TOOL

DESCRIPTION :

Consider a dataset of 2015.csv file of which it contains the attributes are Reference Number, Grid ref: Easting, Grid Ref: Northing, Number of vehicles, Accident date, Time(24 hr), 1st Road class, Road Surface, Lighting conditions, Weather conditions, casuality class, Sex of casuality, Age of casuality, Type of casuality for the performance of the dataset by applying the FP algorithm in weka tool.

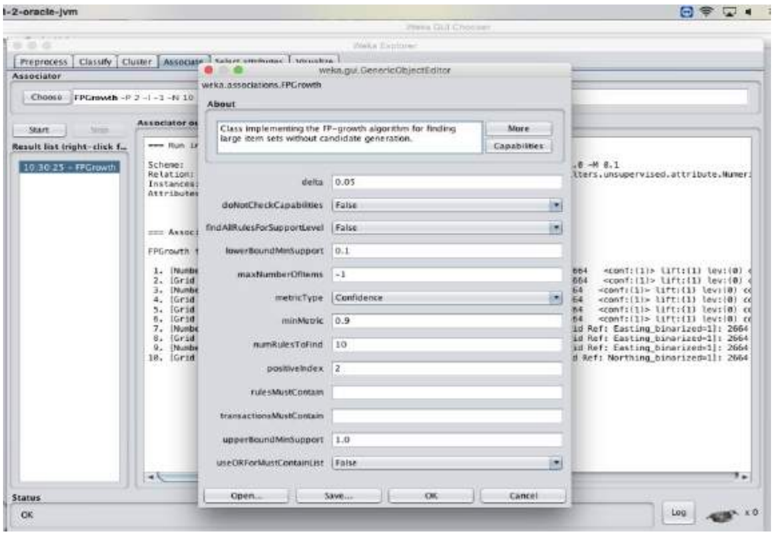
- **USING WEKA TOOL :**

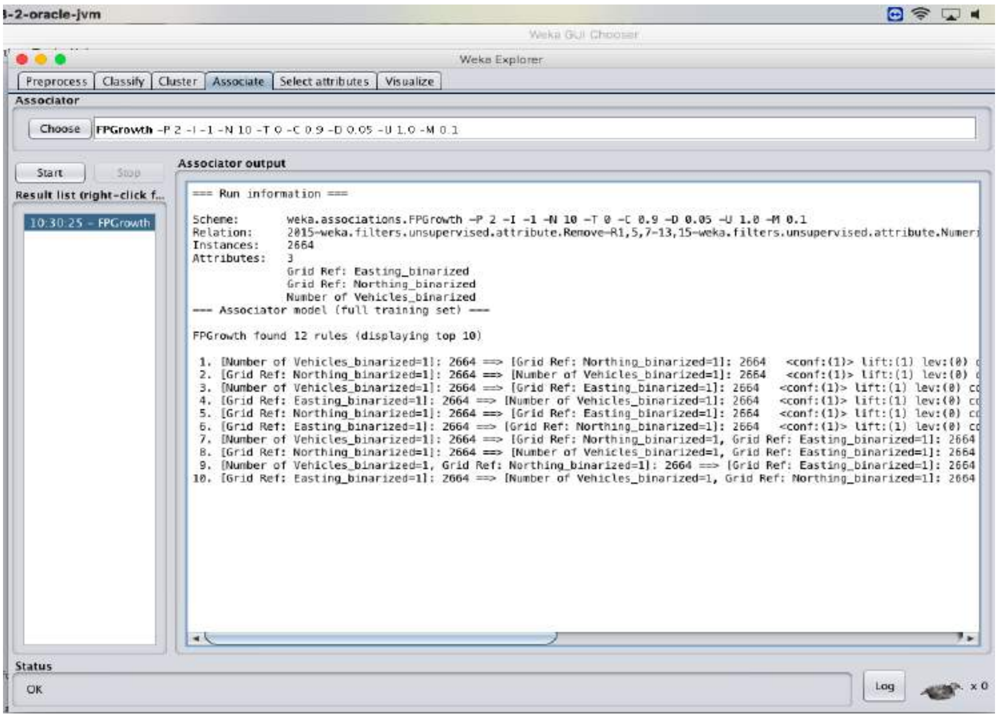
STEPS INVOLVED :

- Choose a set of attributes for clustering and for giving a motivation.
- Choose the dataset and import the dataset into Weka tool.
- Discretize the attributes from all data types to nominal to perform the algorithm.
- Associate the attributes with the FP growth algorithm.
- Set the Upper bound min_sup and lower bound min_sup values.

OBSERVATIONS :

1) When the association rules are of values:
   a) Upper bound min_sup = 1.0
   b) Lower bound min_sup = 0.1
   c) Metric type = confidence.



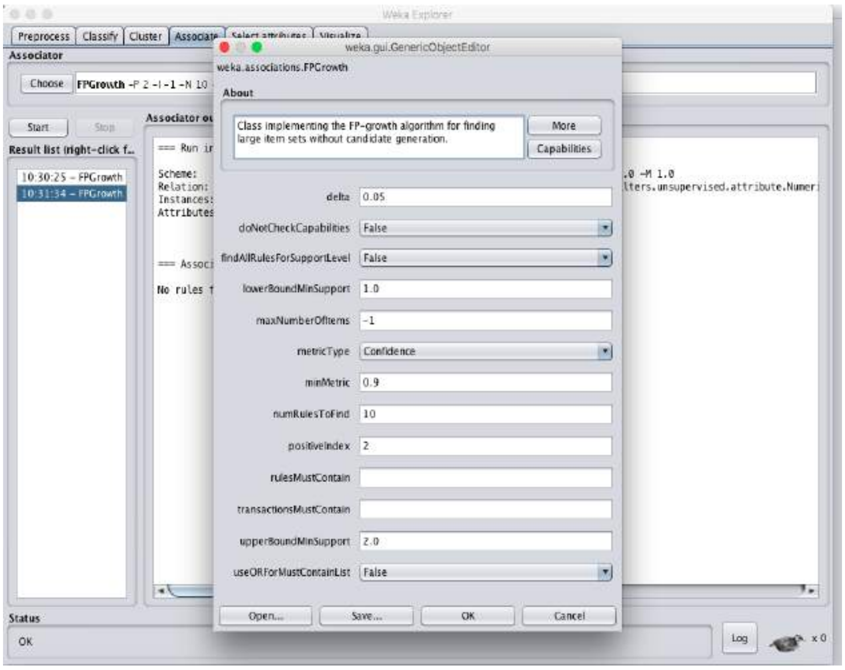


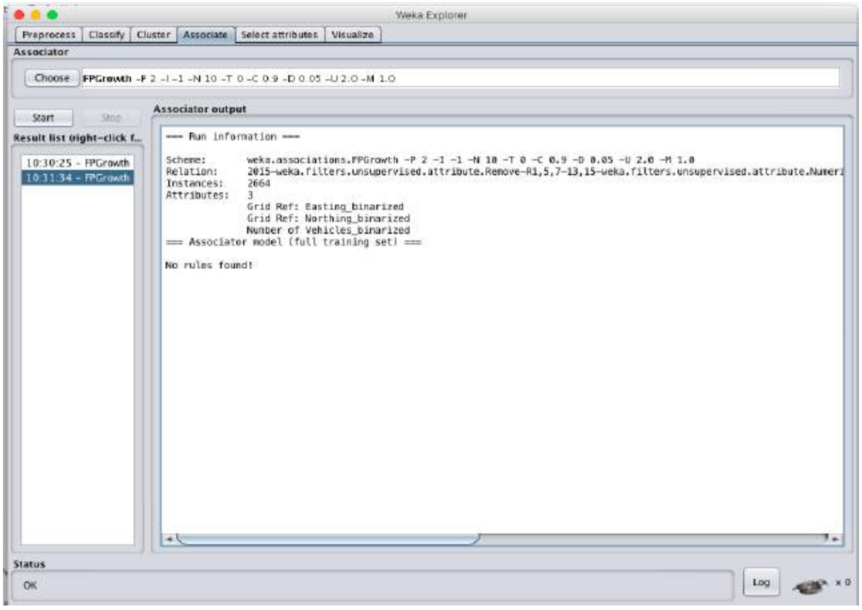
2) When the association rules are of values:

a) Upper bound min_sup = 2.0

b) Lower bound min_sup = 1.0

c) Metric type = confidence.

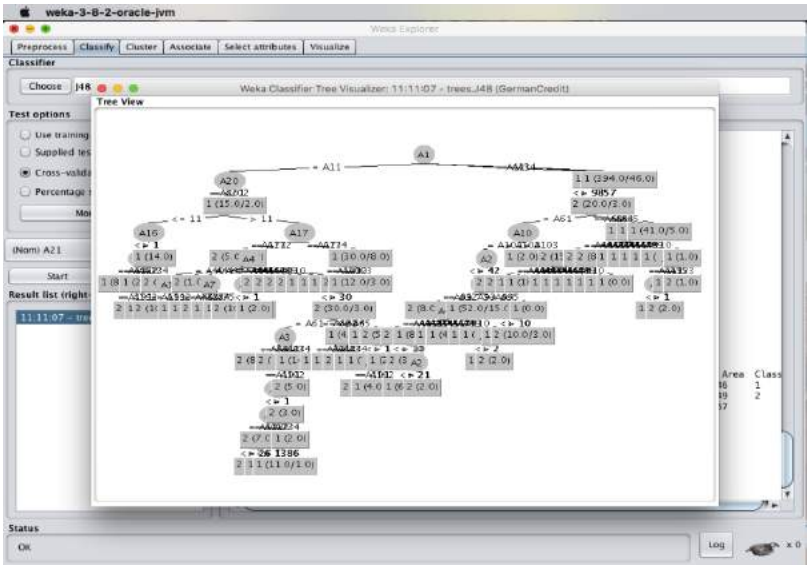

Weka Explorer
Preprocess
Classify
Cluster
Associate
Associator
Choose
FPGrowth -P 2 -I -1 -N 10
Start
Stop
Result list (right-click f...
10:30:25 - FPGrowth
10:31:34 - FPGrowth
Associator output
=== Run in
Scheme:
Relation:
Instances:
Attributes
=== Associ
No rules f
.0 -M 1.0
lters.unsupervised.attribute.Numer
Status
OK
Log
x 0
weka.gui.GenericObjectEditor
weka.associations.FPGrowth
About
Class implementing the FP-growth algorithm for finding large item sets without candidate generation.
More
Capabilities
delta 0.05
doNotCheckCapabilities False
findAllRulesForSupportLevel False
lowerBoundMinSupport 1.0
maxNumberOfItems -1
metricType Confidence
minMetric 0.9
numRulesToFind 10
positiveIndex 2
rulesMustContain
transactionsMustContain
upperBoundMinSupport 2.0
useORForMustContainList False
Open...
Save...
OK
Cancel


Weka Explorer
Preprocess
Classify
Cluster
Associate
Select attributes
Visualize
Associator
Choose
FPGrowth -P 2 -I -1 -N 10 -T 0 -C 0.9 -D 0.05 -U 2.0 -M 1.0
Start
Stop
Associator output
Result list (right-click f...
10:30:25 - FPGrowth
10:31:34 - FPGrowth
=== Run information ===
Scheme: weka.associations.FPGrowth -P 2 -I -1 -N 10 -T 0 -C 0.9 -D 0.05 -U 2.0 -M 1.0
Relation: 2015-weka.filters.unsupervised.attribute.Remove-R1,5,7-13,15-weka.filters.unsupervised.attribute.Numeri
Instances: 2664
Attributes: 3
Grid Ref: Easting_binarized
Grid Ref: Northing_binarized
Number of Vehicles_binarized
=== Associator model (full training set) ===
No rules found!
Status
OK
Log
x 0

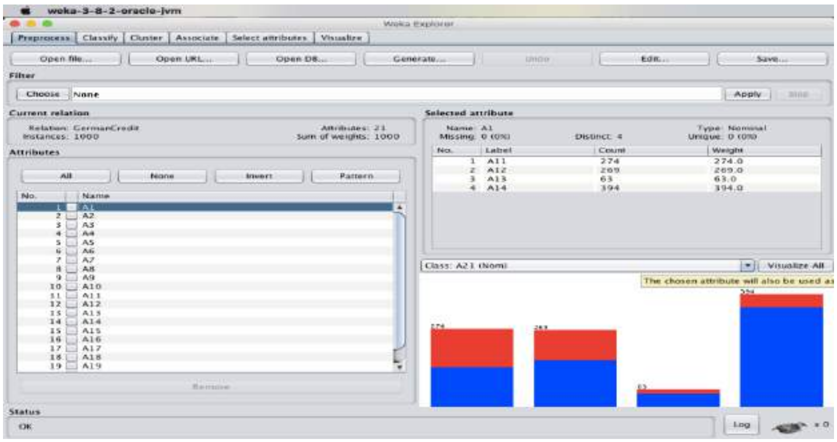
PREDICTION OF CATEGORICAL DATA USING DECISION TREE ALGORTIHM THROUGH WEKA

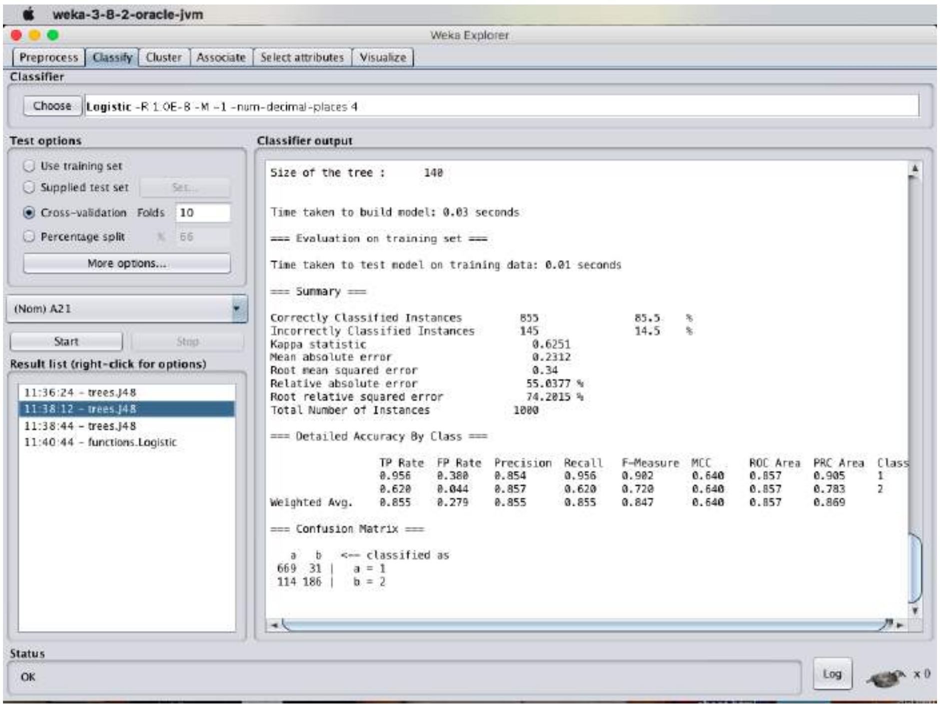
- **Decision Tree :**

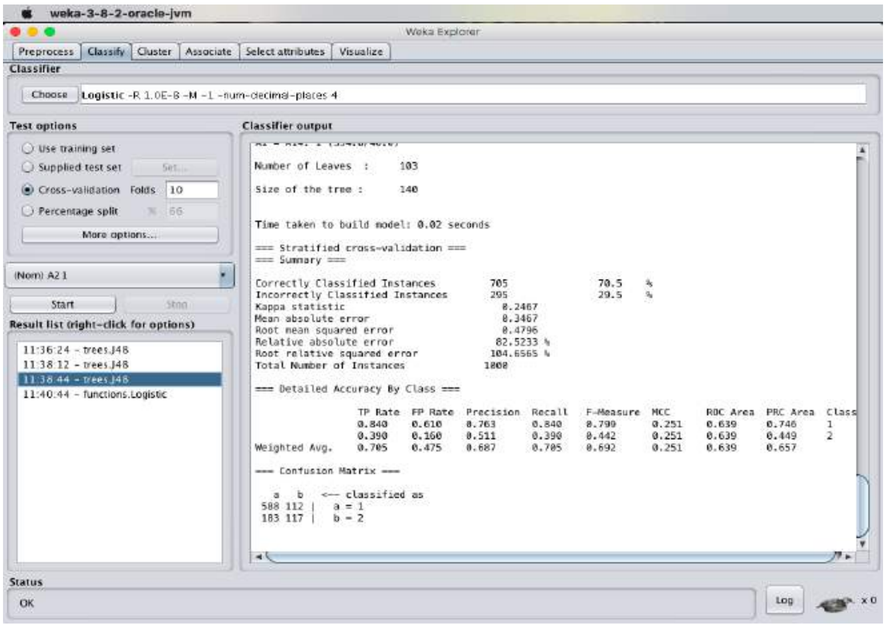
Visualize the decision tree for the given dataset.



or.

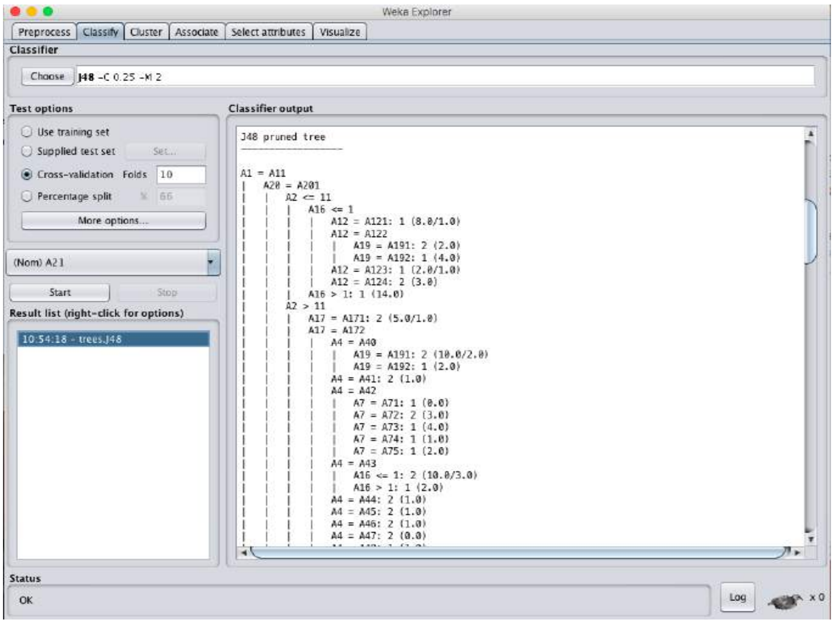




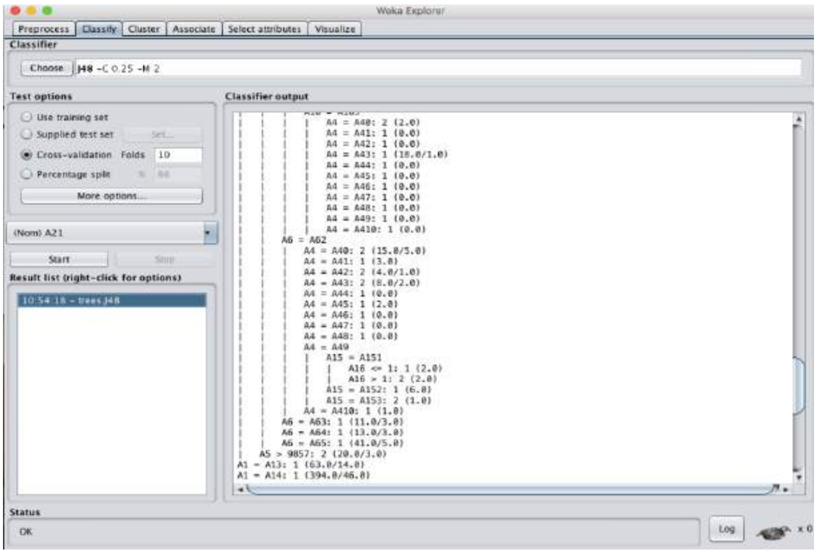


- **CROSS VALIDATION ANALYSIS :**

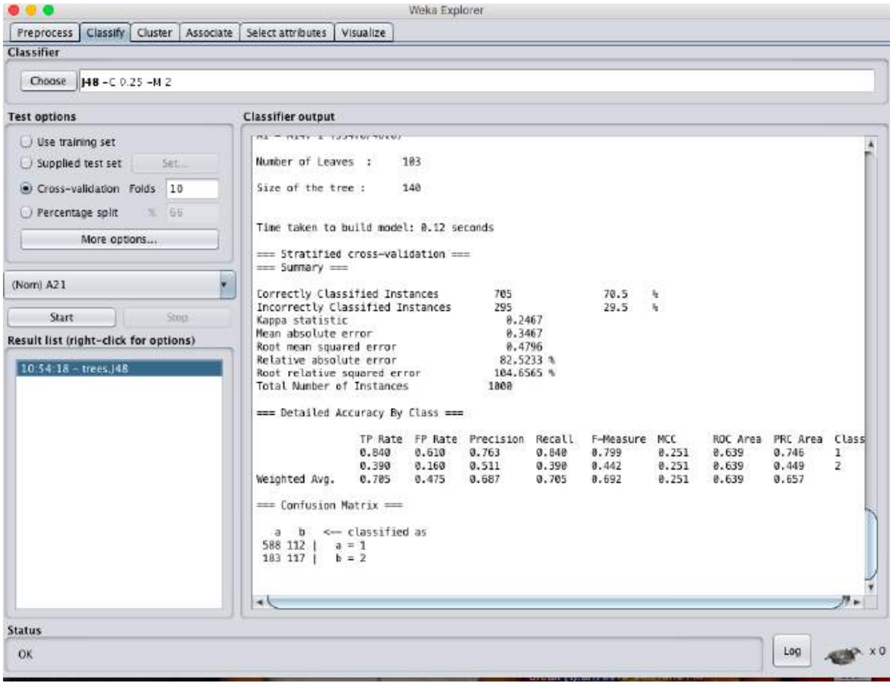
  - **When cross validation folds are 10 :**



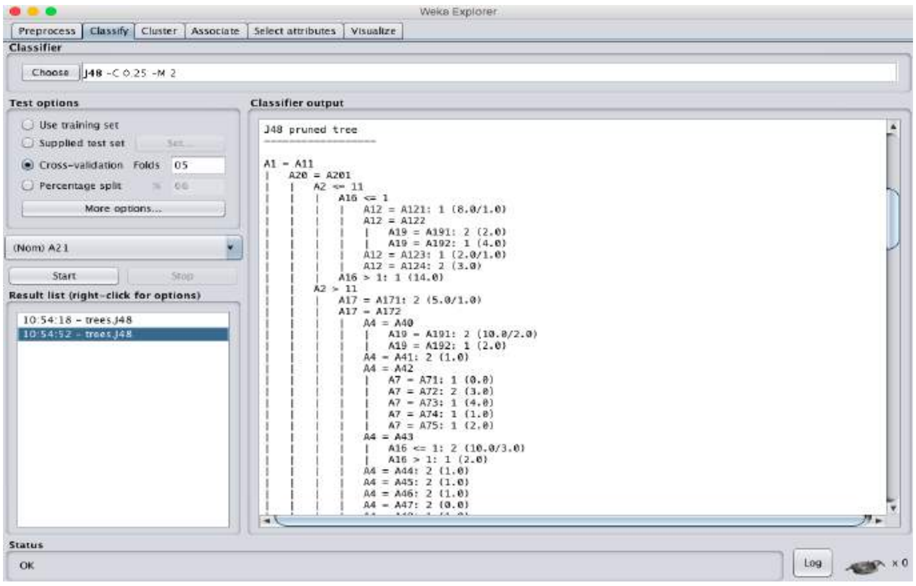
Weka Explorer
Preprocess
Classify
Cluster
Associate
Select attributes
Visualize
Classifier
Choose
J48 -C 0.25 -M 2
Test options
Use training set
Supplied test set
Set...
Cross-validation Folds 10
Percentage split % 66
More options...
(Nom) A21
Start
Stop
Result list (right-click for options)
10:54:18 - trees.J48
Classifier output
J48 pruned tree
A1 = A11
A20 = A201
A2 <= 11
A16 <= 1
A12 = A121: 1 (8.0/1.0)
A12 = A122
A19 = A191: 2 (2.0)
A19 = A192: 1 (4.0)
A12 = A123: 1 (2.0/1.0)
A12 = A124: 2 (3.0)
A16 > 1: 1 (14.0)
A2 > 11
A17 = A171: 2 (5.0/1.0)
A17 = A172
A4 = A40
A19 = A191: 2 (10.0/2.0)
A19 = A192: 1 (2.0)
A4 = A41: 2 (1.0)
A4 = A42
A7 = A71: 1 (0.0)
A7 = A72: 2 (3.0)
A7 = A73: 1 (4.0)
A7 = A74: 1 (1.0)
A7 = A75: 1 (2.0)
A4 = A43
A16 <= 1: 2 (10.0/3.0)
A16 > 1: 1 (2.0)
A4 = A44: 2 (1.0)
A4 = A45: 2 (1.0)
A4 = A46: 2 (1.0)
A4 = A47: 2 (0.0)
Status
OK
Log
x 0

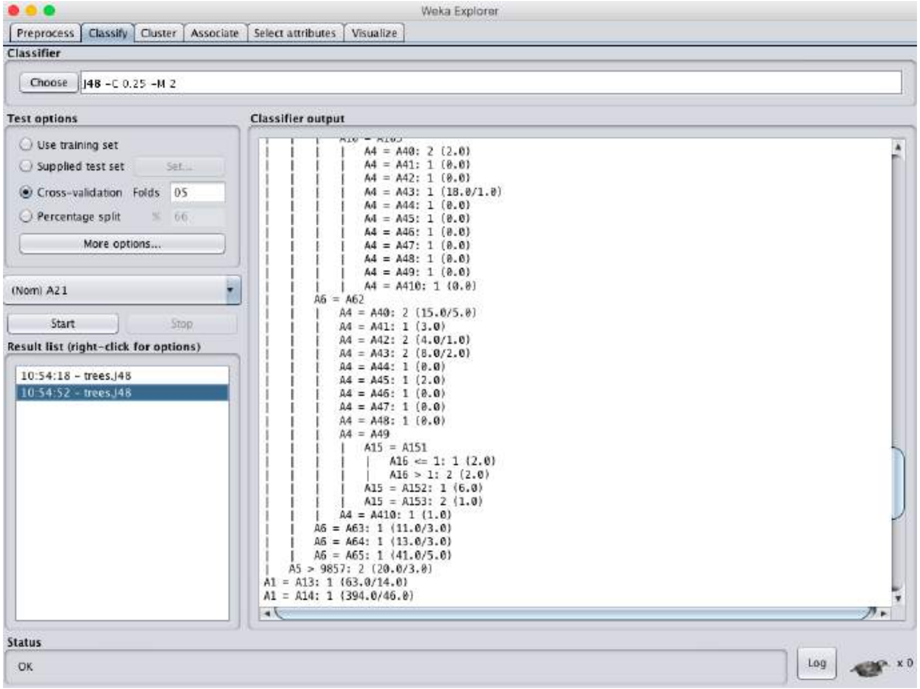


Weka Explorer
Preprocess
Classify
Cluster
Associate
Select attributes
Visualize
Classifier
Choose
J48 -C 0.25 -M 2
Test options
Use training set
Supplied test set
Set...
Cross-validation Folds 10
Percentage split % 66
More options...
(Nom) A21
Start
Stop
Result list (right-click for options)
10:54:18 - trees.J48
Classifier output
A4 = A40: 2 (2.0)
A4 = A41: 1 (0.0)
A4 = A42: 1 (0.0)
A4 = A43: 1 (18.0/1.0)
A4 = A44: 1 (0.0)
A4 = A45: 1 (0.0)
A4 = A46: 1 (0.0)
A4 = A47: 1 (0.0)
A4 = A48: 1 (0.0)
A4 = A49: 1 (0.0)
A4 = A410: 1 (0.0)
A6 = A62
A4 = A40: 2 (15.0/5.0)
A4 = A41: 1 (3.0)
A4 = A42: 2 (4.0/1.0)
A4 = A43: 2 (8.0/2.0)
A4 = A44: 1 (0.0)
A4 = A45: 1 (2.0)
A4 = A46: 1 (0.0)
A4 = A47: 1 (0.0)
A4 = A48: 1 (0.0)
A4 = A49
A15 = A151
A16 <= 1: 1 (2.0)
A16 > 1: 2 (2.0)
A15 = A152: 1 (6.0)
A15 = A153: 2 (1.0)
A4 = A410: 1 (1.0)
A6 = A63: 1 (11.0/3.0)
A6 = A64: 1 (13.0/3.0)
A6 = A65: 1 (41.0/5.0)
A5 > 9857: 2 (20.0/3.0)
A1 = A13: 1 (63.0/14.0)
A1 = A14: 1 (394.0/46.0)
Status
OK
Log
x 0

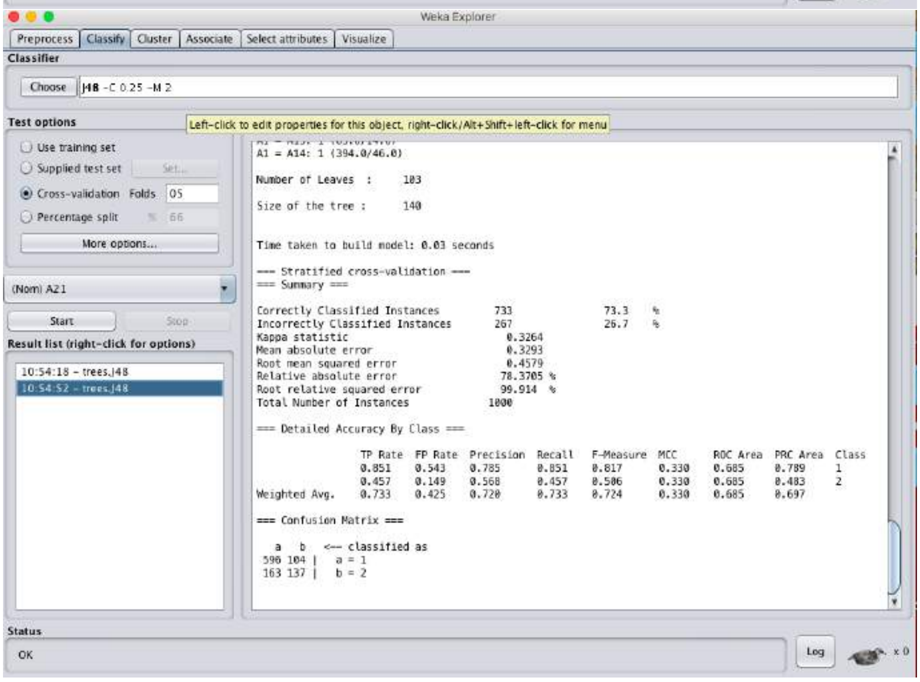


- **When cross validation folds are : 05 :-**





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 05
Percentage split % 66
More options...
(Nom) A21
Start
Stop
Result list (right-click for options)
10:54:18 - trees.J48
10:54:52 - trees.J48
Classifier output
A4 = A40: 2 (2.0)
A4 = A41: 1 (0.0)
A4 = A42: 1 (0.0)
A4 = A43: 1 (18.0/1.0)
A4 = A44: 1 (0.0)
A4 = A45: 1 (0.0)
A4 = A46: 1 (0.0)
A4 = A47: 1 (0.0)
A4 = A48: 1 (0.0)
A4 = A49: 1 (0.0)
A4 = A410: 1 (0.0)
A6 = A62
A4 = A40: 2 (15.0/5.0)
A4 = A41: 1 (3.0)
A4 = A42: 2 (4.0/1.0)
A4 = A43: 2 (8.0/2.0)
A4 = A44: 1 (0.0)
A4 = A45: 1 (2.0)
A4 = A46: 1 (0.0)
A4 = A47: 1 (0.0)
A4 = A48: 1 (0.0)
A4 = A49
A15 = A151
A16 <= 1: 1 (2.0)
A16 > 1: 2 (2.0)
A15 = A152: 1 (6.0)
A15 = A153: 2 (1.0)
A4 = A410: 1 (1.0)
A6 = A63: 1 (11.0/3.0)
A6 = A64: 1 (13.0/3.0)
A6 = A65: 1 (41.0/5.0)
A5 > 9857: 2 (20.0/3.0)
A1 = A13: 1 (63.0/14.0)
A1 = A14: 1 (394.0/46.0)
Status
OK
Log
x 0



Weka Explorer
Preprocess
Classify
Cluster
Associate
Select attributes
Visualize
Classifier
Choose
J48 -C 0.25 -M 2
Test options
Left-click to edit properties for this object, right-click/Alt+Shift+left-click for menu
Use training set
Supplied test set
Set...
Cross-validation Folds 05
Percentage split % 66
More options...
(Nom) A21
Start
Stop
Result list (right-click for options)
10:54:18 - trees.J48
10:54:52 - trees.J48
A1 = A14: 1 (394.0/46.0)
Number of Leaves : 103
Size of the tree : 140
Time taken to build model: 0.03 seconds
=== Stratified cross-validation ===
=== Summary ===
Correctly Classified Instances 733 73.3 %
Incorrectly Classified Instances 267 26.7 %
Kappa statistic 0.3264
Mean absolute error 0.3293
Root mean squared error 0.4579
Relative absolute error 78.3705 %
Root relative squared error 99.914 %
Total Number of Instances 1000
=== Detailed Accuracy By Class ===
TP Rate FP Rate Precision Recall F-Measure MCC ROC Area PRC Area Class
0.851 0.543 0.785 0.851 0.817 0.330 0.685 0.789 1
0.457 0.149 0.568 0.457 0.506 0.330 0.685 0.483 2
Weighted Avg. 0.733 0.425 0.720 0.733 0.724 0.330 0.685 0.697
=== Confusion Matrix ===
a b <-- classified as
596 104 | a = 1
163 137 | b = 2
Status
OK
Log
x 0

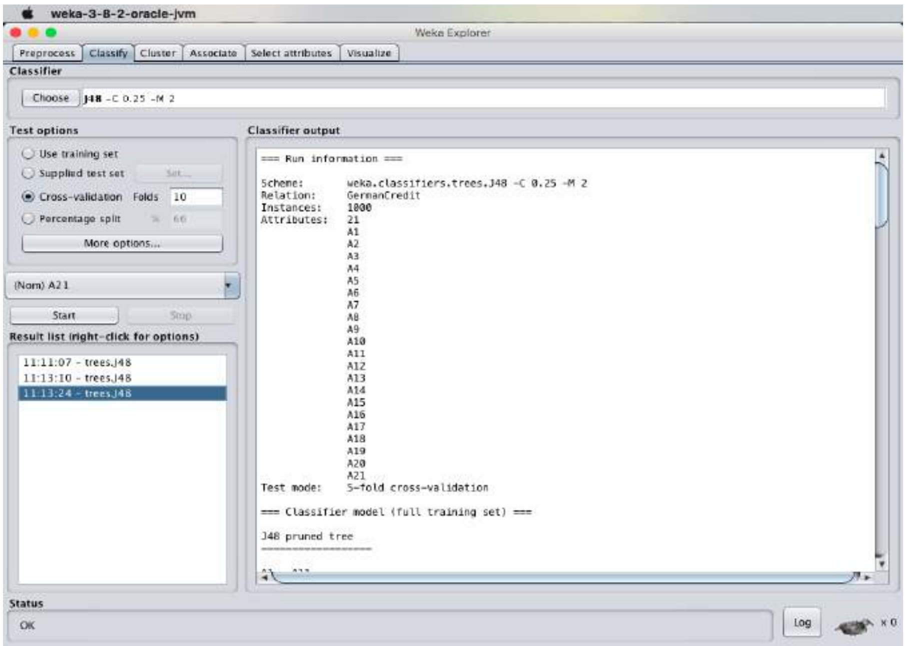
PREDICTION OF CATEGORICAL DATA USING SMO ALGORTIHM THROUGH WEKA

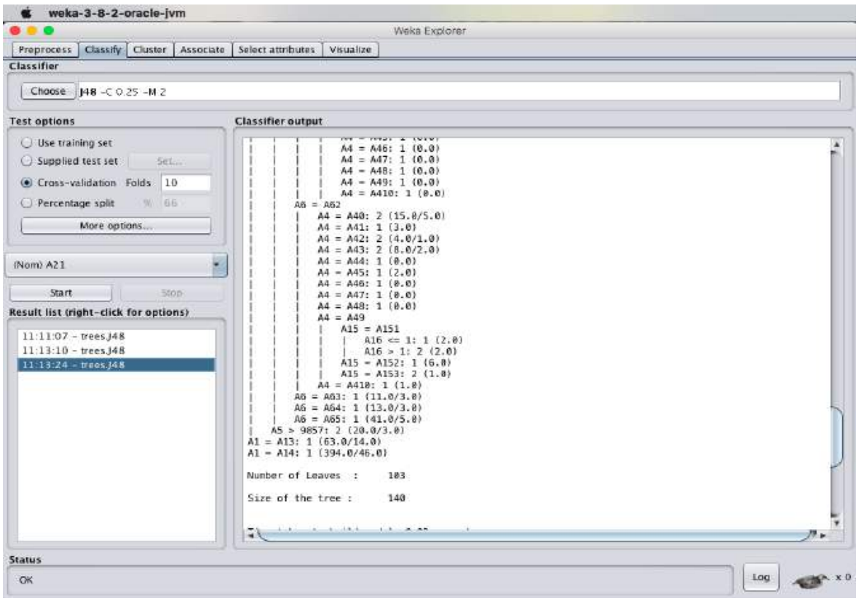
DESCRIPTION :

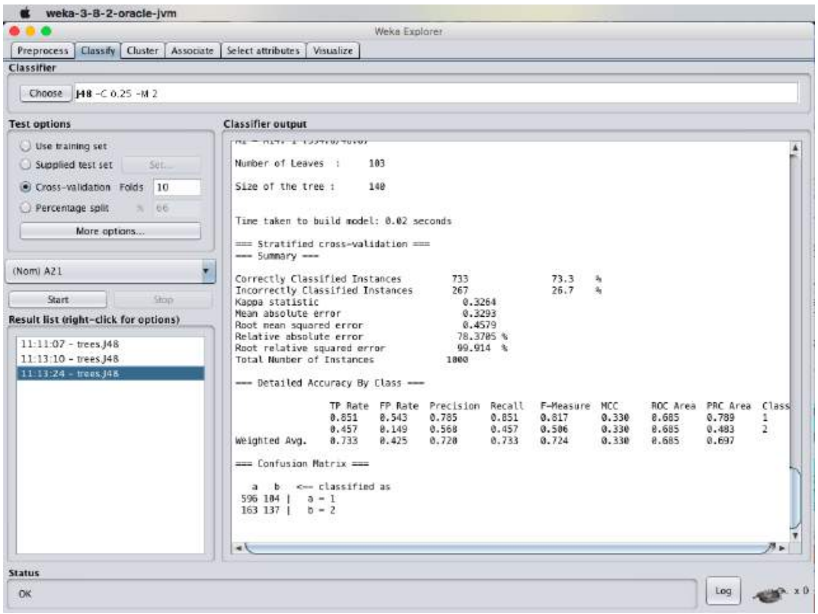
Consider the german credit dataset which can be downloaded from the UCI repository.

❖ **DECISION TREE :**

A tree has many analogies in real life, and turns out that it has influenced a wide area of machine learning, covering both classification and regression. In decision analysis, a decision tree can be used to visually and explicitly represent decisions and decision making. As the name goes, it uses a tree-like model of decisions. Though a commonly used tool in data mining for deriving a strategy to reach a particular goal, its also widely used in machine learning, which will be the main focus of this article.

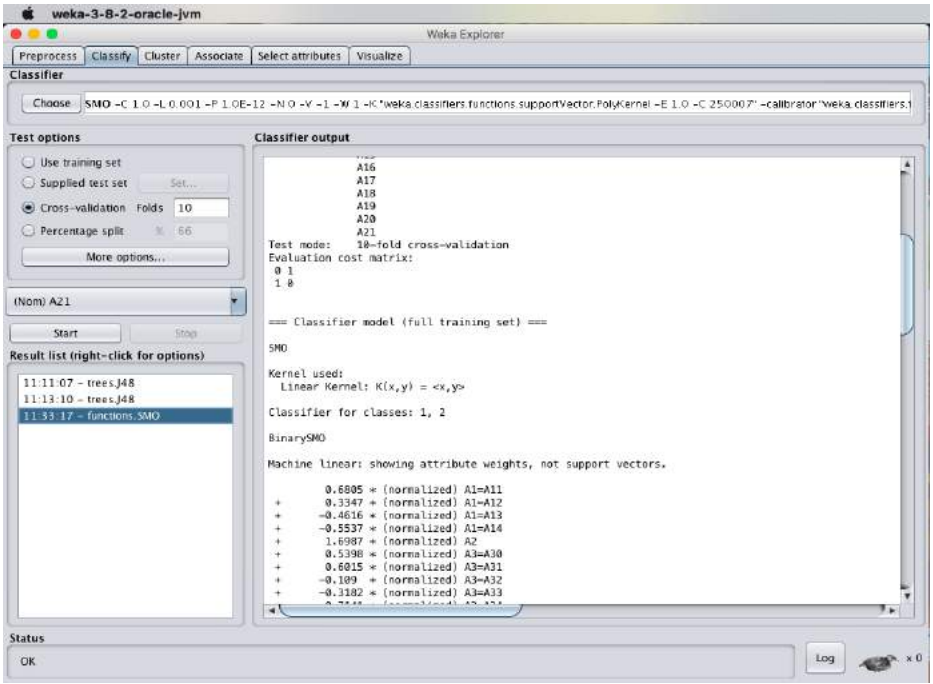


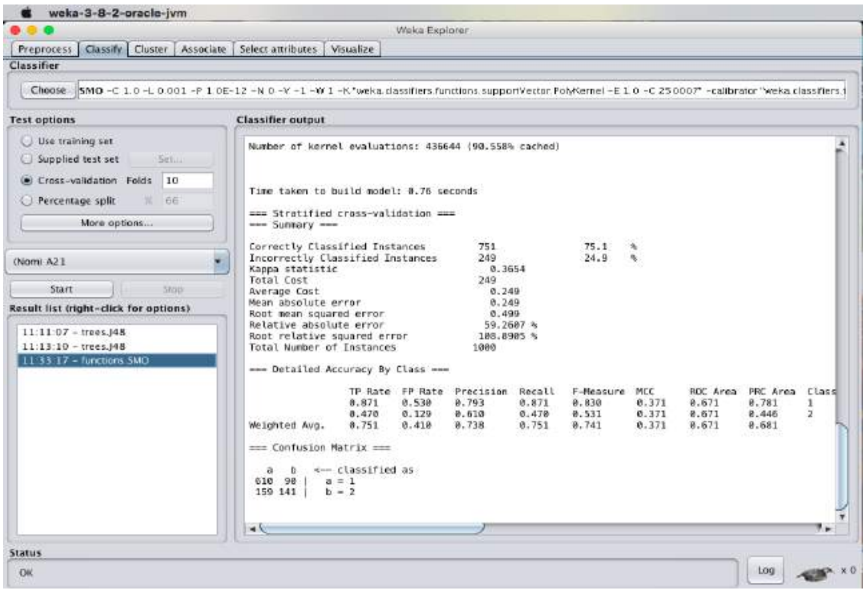




❖ SMO ALGORITHM:

The iterative algorithm Sequential Minimal Optimization (SMO) is used for solving quadratic programming (QP) problems. One example where QP problems are relevant is during the training process of support vector machines (SVM). The SMO algorithm is used to solve in this example a constraint optimization problem. John Platt proposed this algorithm in 1998 and it was successfully used since then. We describe here the basics of the algorithm in the light of big data.


weka-3-8-2-oracle-jvm
Weka Explorer
Preprocess
Classify
Cluster
Associate
Select attributes
Visualize
Classifier
Choose
SMO -C 1.0 -L 0.001 -P 1.0E-12 -N 0 -V -1 -W 1 -K "weka.classifiers.functions.supportVector.PolyKernel -E 1.0 -C 250007" -calibrator "weka.classifiers.
Test options
Use training set
Supplied test set
Set...
Cross-validation Folds 10
Percentage split % 66
More options...
(Nom) A21
Start
Stop
Result list (right-click for options)
11:11:07 - trees.J48
11:13:10 - trees.J48
11:33:17 - functions.SMO
Classifier output
A16
A17
A18
A19
A20
A21
Test mode: 10-fold cross-validation
Evaluation cost matrix:
0 1
1 0
=== Classifier model (full training set) ===
SMO
Kernel used:
Linear Kernel: K(x,y) = <x,y>
Classifier for classes: 1, 2
BinarySMO
Machine linear: showing attribute weights, not support vectors.
0.6805 * (normalized) A1=A11
+ 0.3347 * (normalized) A1=A12
+ -0.4616 * (normalized) A1=A13
+ -0.5537 * (normalized) A1=A14
+ 1.6987 * (normalized) A2
+ 0.5398 * (normalized) A3=A30
+ 0.6015 * (normalized) A3=A31
+ -0.109 * (normalized) A3=A32
+ -0.3182 * (normalized) A3=A33
Status
OK
Log
x 0


weka-3-8-2-oracle-jvm
Weka Explorer
Preprocess
Classify
Cluster
Associate
Select attributes
Visualize
Classifier
Choose
SMO -C 1.0 -L 0.001 -P 1.0E-12 -N 0 -V -1 -W 1 -K "weka.classifiers.functions.supportVector.PolyKernel -E 1.0 -C 250007" -calibrator "weka.classifiers.
Test options
Use training set
Supplied test set
Set...
Cross-validation Folds 10
Percentage split % 66
More options...
(Nom) A21
Start
Stop
Result list (right-click for options)
11:11:07 - trees.J48
11:13:10 - trees.J48
11:33:17 - functions.SMO
Classifier output
Number of kernel evaluations: 436644 (90.558% cached)
Time taken to build model: 0.76 seconds
=== Stratified cross-validation ===
=== Summary ===
Correctly Classified Instances 751 75.1 %
Incorrectly Classified Instances 249 24.9 %
Kappa statistic 0.3654
Total Cost 249
Average Cost 0.249
Mean absolute error 0.249
Root mean squared error 0.499
Relative absolute error 59.2607 %
Root relative squared error 108.8905 %
Total Number of Instances 1000
=== Detailed Accuracy By Class ===
TP Rate FP Rate Precision Recall F-Measure MCC ROC Area PRC Area Class
0.871 0.530 0.793 0.871 0.830 0.371 0.671 0.781 1
0.470 0.129 0.610 0.470 0.531 0.371 0.671 0.446 2
Weighted Avg. 0.751 0.410 0.738 0.751 0.741 0.371 0.671 0.681
=== Confusion Matrix ===
a b <-- classified as
610 90 | a = 1
159 141 | b = 2
Status
OK
Log
x 0

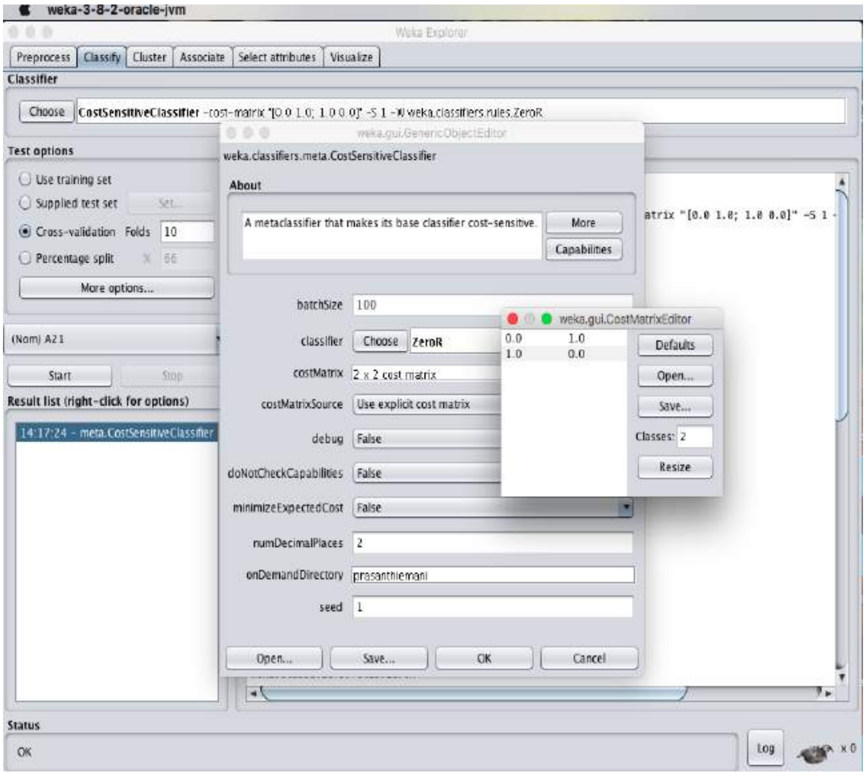
1. Set the cost sensitive evaluation and compare the obtained results.

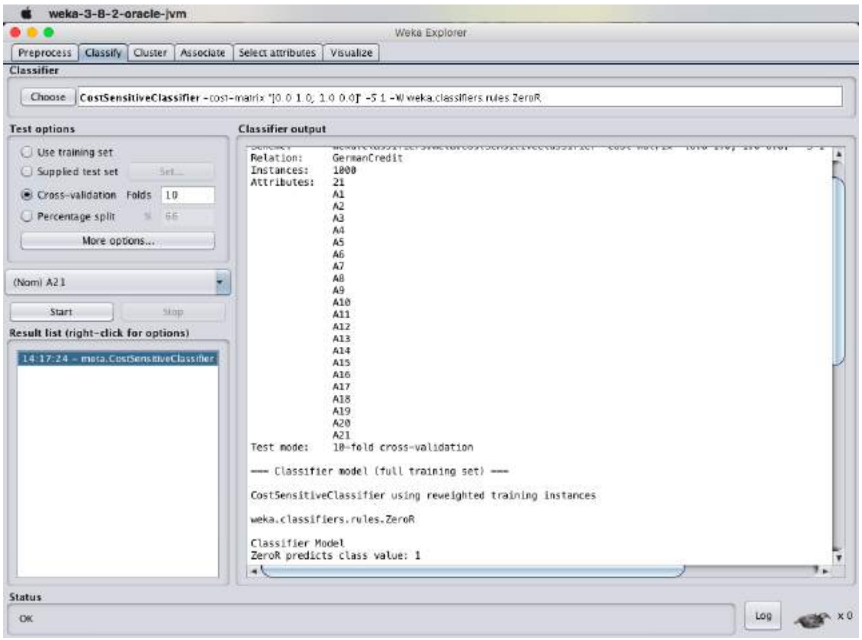
Cost-Sensitive Learning is a type of learning in data mining that takes the misclassification costs (and possibly other types of cost) into consideration. The goal of this type of learning is to minimize the total cost. The key difference between cost-sensitive learning and cost-insensitive learning is that cost-sensitive learning treats the different misclassifications differently. Costinsensitive learning does not take the misclassification costs into consideration. The goal of this type of learning is to pursue a high accuracy of classifying examples into a set of known classes.

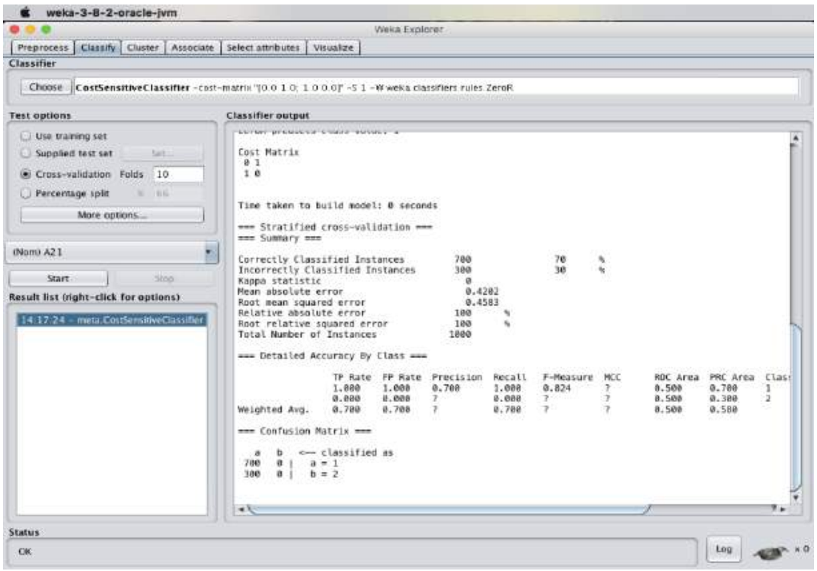
STEPS :

- Classifythe dataset with the cost sensitive classifier technique.
- Change the cost matrix to 2*2 matrix and execute.

ANALYSIS :







2. **What is the significance of the following parameters :**

a) **Mean Absolute Error :**

Mean Absolute Error (MAE) is similar to the Mean Squared Error, but it uses absolute values instead of squaring. This measure is not as popular as MSE, though its meaning is more intuitive (the "average error").

b) **Total Number of Instances :**

The data present consists of various instances of the class. In the case of german_credit dataset, the total number of instances present in the german credit dataset are 1000 instances.

EVALUATING ACCURACY OF THE CLASSIFIERS

DESCRIPTION :

Consider the german credit dataset which can be downloaded from the UCI repository.

ANALYSIS :

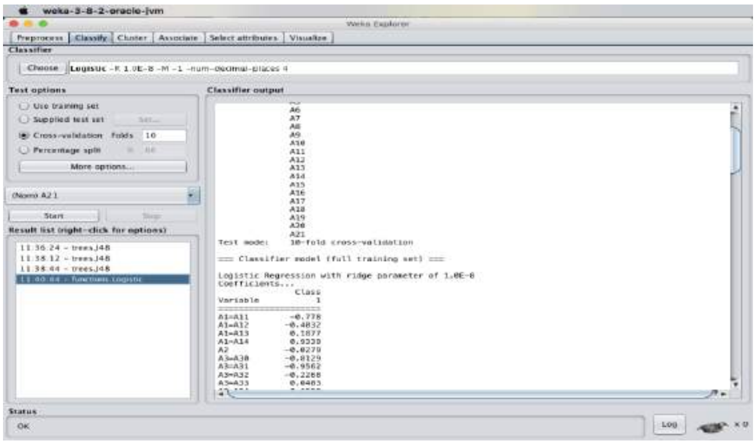
A) **Logistic Regression :**

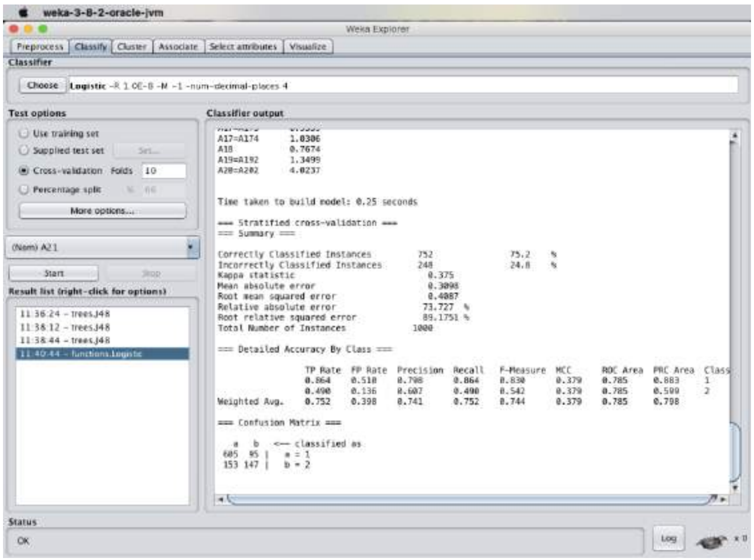
Logistic regression predicts the probability of an outcome that can only have two values (i.e. a dichotomy). The prediction is based on the use of one or several predictors (numerical and categorical).

Steps :

- Load the dataset into the weka tool and preprocess it.
- Apply the classification the logistic regression technique and execute for the result.

Output :





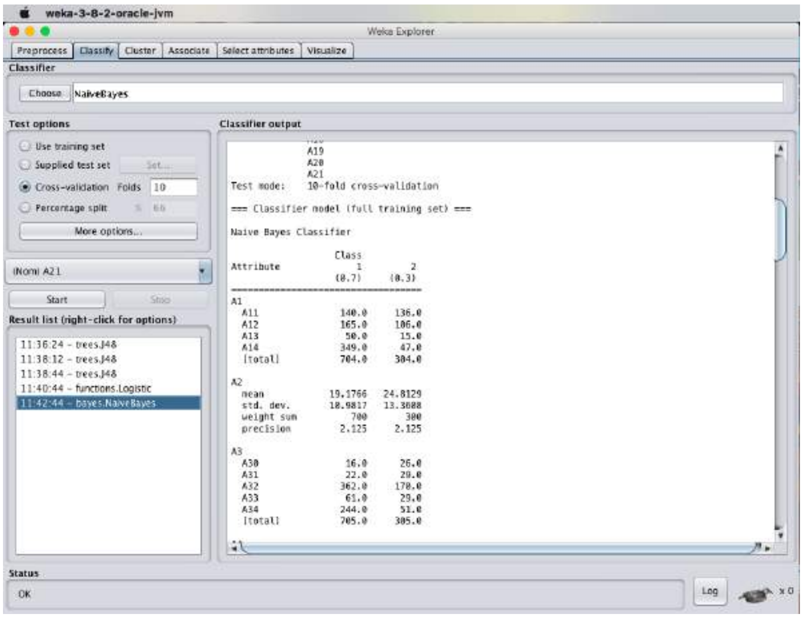
B) Naïve Bayes Algorithm :

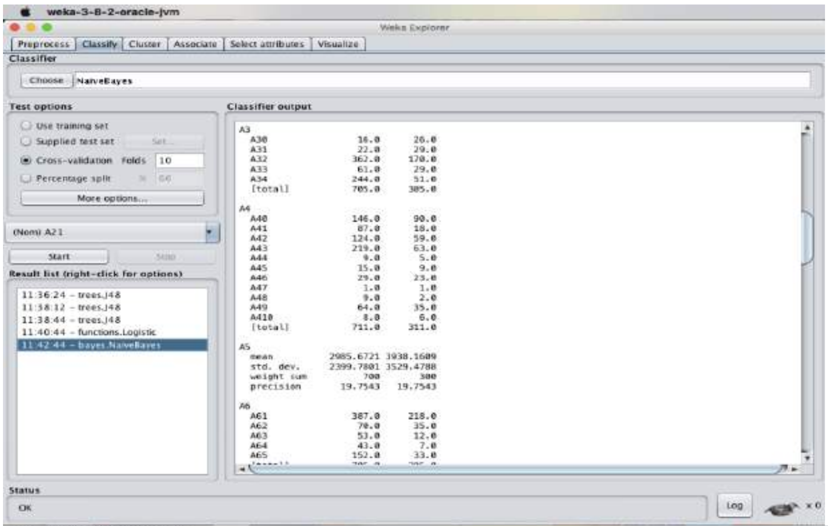
The Naive Bayesian classifier is based on Bayes' theorem with the independence assumptions between predictors. A Naive Bayesian model is easy to build, with no complicated iterative parameter estimation which makes it particularly useful for very large datasets. Despite its simplicity, the Naive Bayesian classifier often does surprisingly well and is widely used because it often outperforms more sophisticated classification methods.

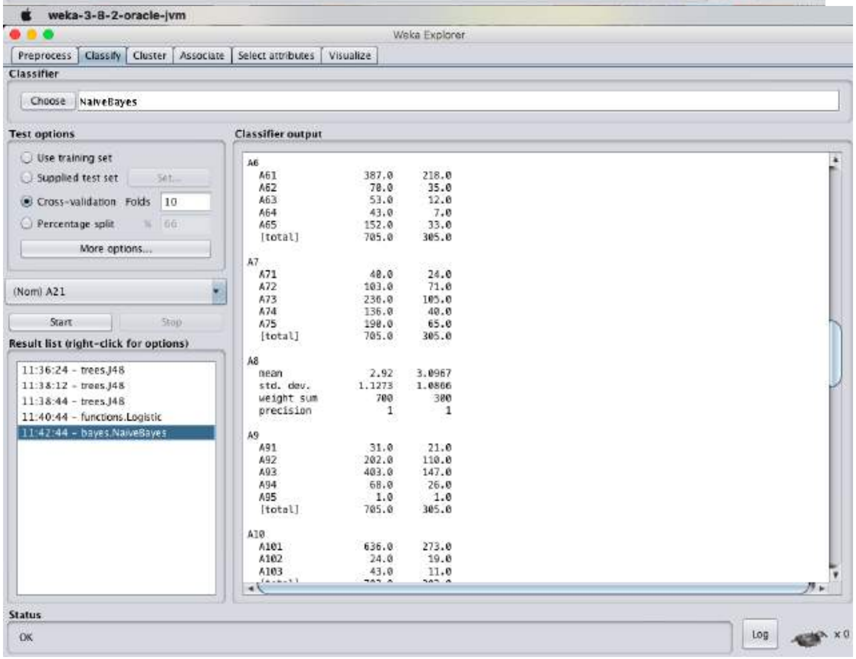
Steps :

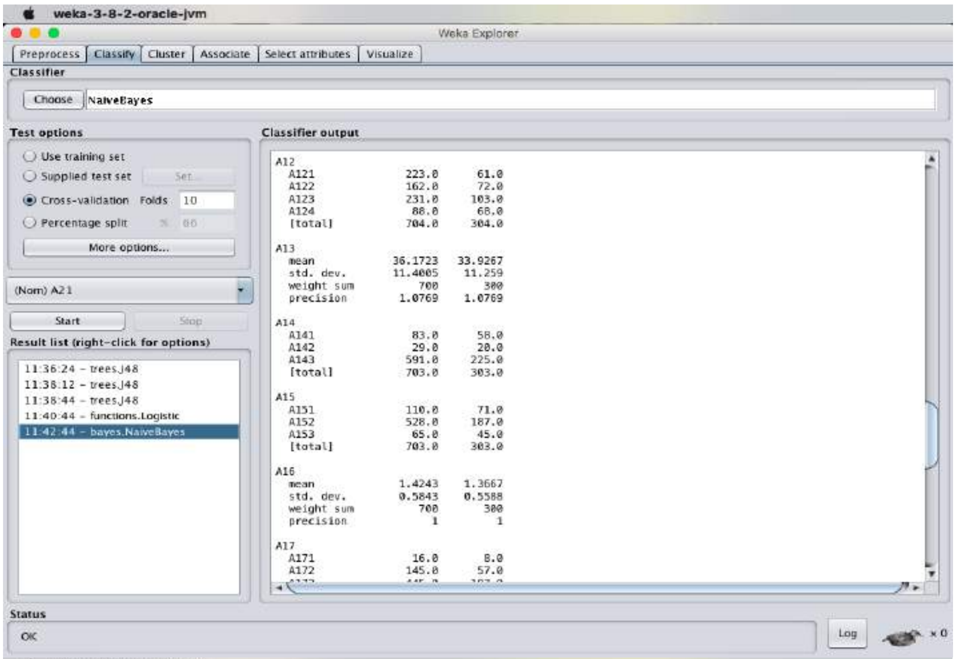
- Load the dataset into the weka tool and preprocess it.
- Apply the classification the Naïve bayes technique and execute for the result.

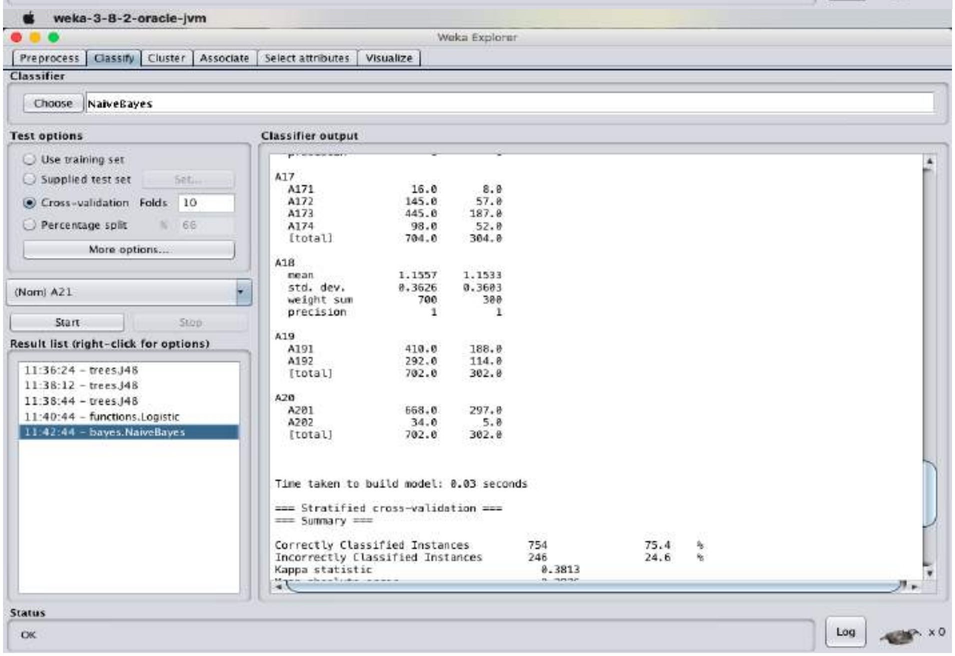
Output :

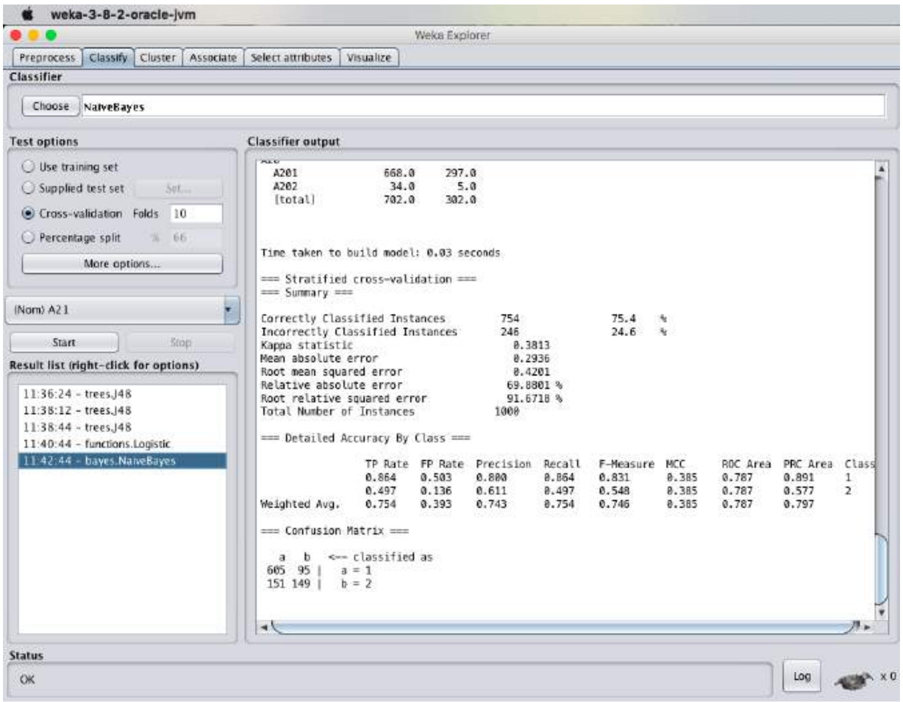

weka-3-8-2-oracle-jvm
Weka Explorer
Preprocess
Classify
Cluster
Associate
Select attributes
Visualize
Classifier
Choose
NaiveBayes
Test options
Use training set
Supplied test set
Cross-validation Folds 10
Percentage split % 66
More options...
(Nom) A21
Start
Stop
Result list (right-click for options)
11:36:24 - trees.J48
11:38:12 - trees.J48
11:38:44 - trees.J48
11:40:44 - functions.Logistic
11:42:44 - bayes.NaiveBayes
Classifier output
A19
A20
A21
Test mode: 10-fold cross-validation
=== Classifier model (full training set) ===
Naive Bayes Classifier
Class
Attribute 1 2
(0.7) (0.3)
A1
A11 140.0 136.0
A12 165.0 106.0
A13 50.0 15.0
A14 349.0 47.0
[total] 704.0 304.0
A2
mean 19.1766 24.8129
weight sum 700 300
precision 2.125 2.125
A3
A30 16.0 26.0
A31 22.0 29.0
A32 362.0 170.0
A33 61.0 29.0
A34 244.0 51.0
[total] 705.0 305.0
Status
OK
Log


weka-3-8-2-oracle-jvm
Weka Explorer
Preprocess
Classify
Cluster
Associate
Select attributes
Visualize
Classifier
Choose
NaiveBayes
Test options
Use training set
Supplied test set
Cross-validation Folds 10
Percentage split
More options...
(Nom) A21
Start
Stop
Result list (right-click for options)
11:36:24 - trees.J48
11:38:12 - trees.J48
11:38:44 - trees.J48
11:40:44 - functions.Logistic
11:42:44 - bayes.NaiveBayes
Classifier output
A3
A30 16.0 26.0
A31 22.0 29.0
A32 362.0 170.0
A33 61.0 29.0
A34 244.0 51.0
[total] 705.0 305.0
A4
A40 146.0 90.0
A41 87.0 18.0
A42 124.0 59.0
A43 219.0 63.0
A44 9.0 5.0
A45 15.0 9.0
A46 29.0 23.0
A47 1.0 1.0
A48 9.0 2.0
A49 64.0 35.0
A410 8.0 6.0
[total] 711.0 311.0
A5
weight sum 700 300
precision 19.7543 19.7543
A6
A61 387.0 218.0
A62 70.0 35.0
A63 53.0 12.0
A64 43.0 7.0
A65 152.0 33.0
Status
OK
Log


weka-3-8-2-oracle-jvm
Weka Explorer
Preprocess
Classify
Cluster
Associate
Select attributes
Visualize
Classifier
Choose
NaiveBayes
Test options
Use training set
Supplied test set
Cross-validation Folds 10
Percentage split % 66
More options...
(Nom) A21
Start
Stop
Result list (right-click for options)
11:36:24 - trees.J48
11:38:12 - trees.J48
11:38:44 - trees.J48
11:40:44 - functions.Logistic
11:42:44 - bayes.NaiveBayes
Classifier output
A6
A61 387.0 218.0
A62 70.0 35.0
A63 53.0 12.0
A64 43.0 7.0
A65 152.0 33.0
[total] 705.0 305.0
A7
A71 40.0 24.0
A72 103.0 71.0
A73 236.0 105.0
A74 136.0 40.0
A75 190.0 65.0
[total] 705.0 305.0
A8
mean 2.92 3.0967
std. dev. 1.1273 1.0866
weight sum 700 300
precision 1 1
A9
A91 31.0 21.0
A92 202.0 110.0
A93 403.0 147.0
A94 68.0 26.0
A95 1.0 1.0
[total] 705.0 305.0
A10
A101 636.0 273.0
A102 24.0 19.0
A103 43.0 11.0
Status
OK
Log


weka-3-8-2-oracle-jvm
Weka Explorer
Preprocess | Classify | Cluster | Associate | Select attributes | Visualize
Classifier
Choose NaiveBayes
Test options
Use training set
Supplied test set Set...
Cross-validation Folds 10
Percentage split % 66
More options...
(Nom) A21
Start Stop
Result list (right-click for options)
11:36:24 - trees.J48
11:38:12 - trees.J48
11:38:44 - trees.J48
11:40:44 - functions.Logistic
11:42:44 - bayes.NaiveBayes
Classifier output
A12
A121 223.0 61.0
A122 162.0 72.0
A123 231.0 103.0
A124 88.0 68.0
[total] 704.0 304.0
A13
mean 36.1723 33.9267
std. dev. 11.4005 11.259
weight sum 700 300
precision 1.0769 1.0769
A14
A141 83.0 58.0
A142 29.0 20.0
A143 591.0 225.0
[total] 703.0 303.0
A15
A151 110.0 71.0
A152 528.0 187.0
A153 65.0 45.0
[total] 703.0 303.0
A16
mean 1.4243 1.3667
std. dev. 0.5843 0.5588
weight sum 700 300
precision 1 1
A17
A171 16.0 8.0
A172 145.0 57.0
Status
OK
Log
x 0


weka-3-8-2-oracle-jvm
Weka Explorer
Preprocess | Classify | Cluster | Associate | Select attributes | Visualize
Classifier
Choose NaiveBayes
Test options
Use training set
Supplied test set Set...
Cross-validation Folds 10
Percentage split % 66
More options...
(Nom) A21
Start Stop
Result list (right-click for options)
11:36:24 - trees.J48
11:38:12 - trees.J48
11:38:44 - trees.J48
11:40:44 - functions.Logistic
11:42:44 - bayes.NaiveBayes
Classifier output
A17
A171 16.0 8.0
A172 145.0 57.0
A173 445.0 187.0
A174 98.0 52.0
[total] 704.0 304.0
A18
mean 1.1557 1.1533
std. dev. 0.3626 0.3603
weight sum 700 300
precision 1 1
A19
A191 410.0 188.0
A192 292.0 114.0
[total] 702.0 302.0
A20
A201 668.0 297.0
A202 34.0 5.0
[total] 702.0 302.0
Time taken to build model: 0.03 seconds
=== Stratified cross-validation ===
=== Summary ===
Correctly Classified Instances 754 75.4 %
Incorrectly Classified Instances 246 24.6 %
Kappa statistic 0.3813
Status
OK
Log
x 0



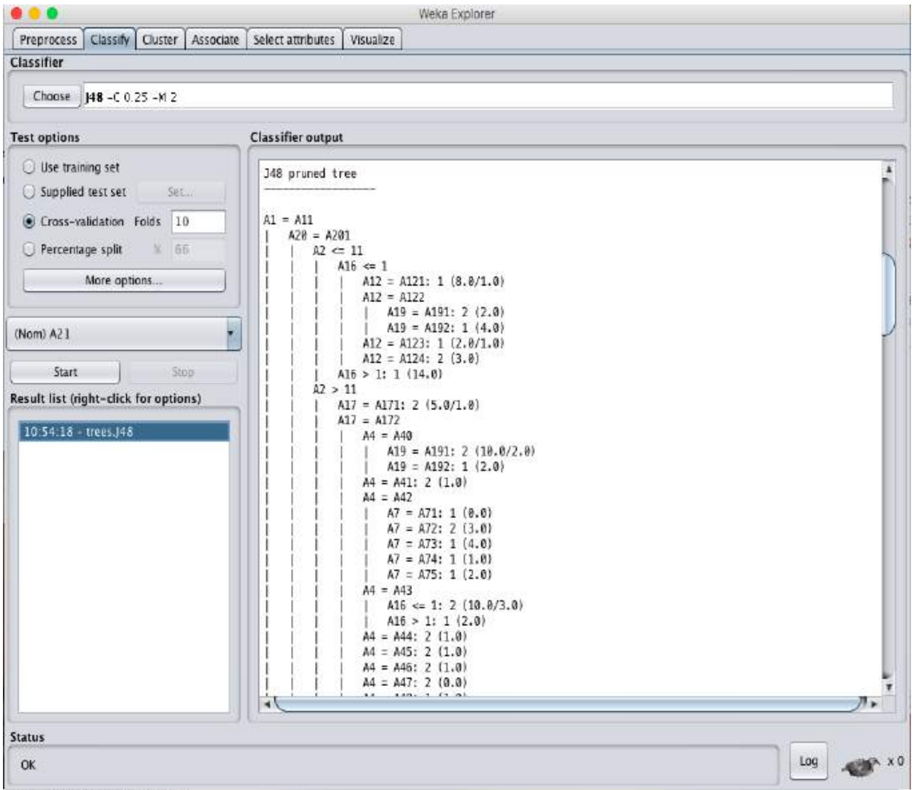
C) J48 Algorithm :

Classification is the process of building a model of classes from a set of records that contain class labels. Decision Tree Algorithm is to find out the way the attributes-vector behaves for a number of instances. Also on the bases of the training instances the classes for the newly generated instances are being found. This algorithm generates the rules for the prediction of the target variable. With the help of tree classification algorithm the critical distribution of the data is easily understandable.

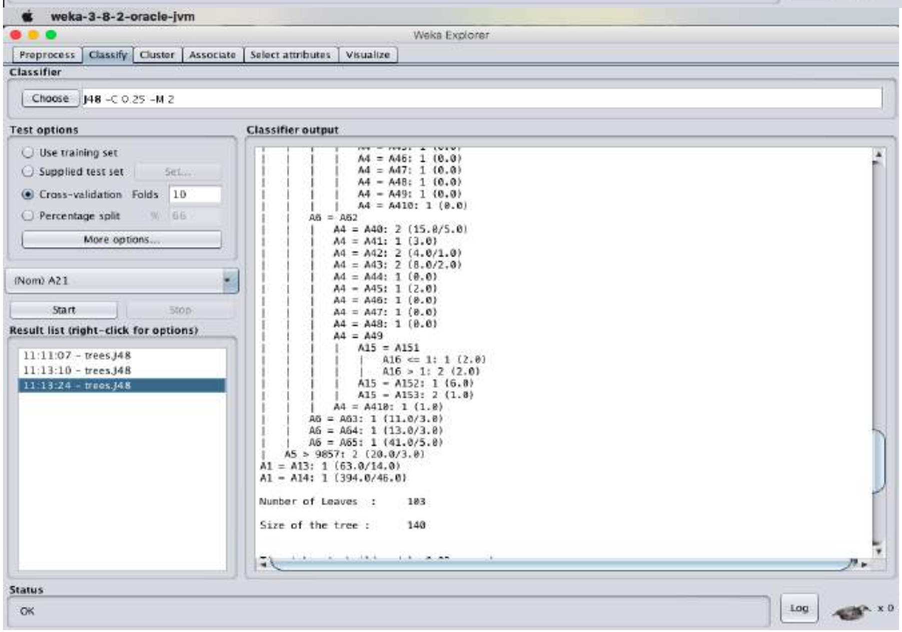
Steps :

- Load the dataset into the weka tool and preprocess it.
- Apply the classification the J48 technique and execute for the result.

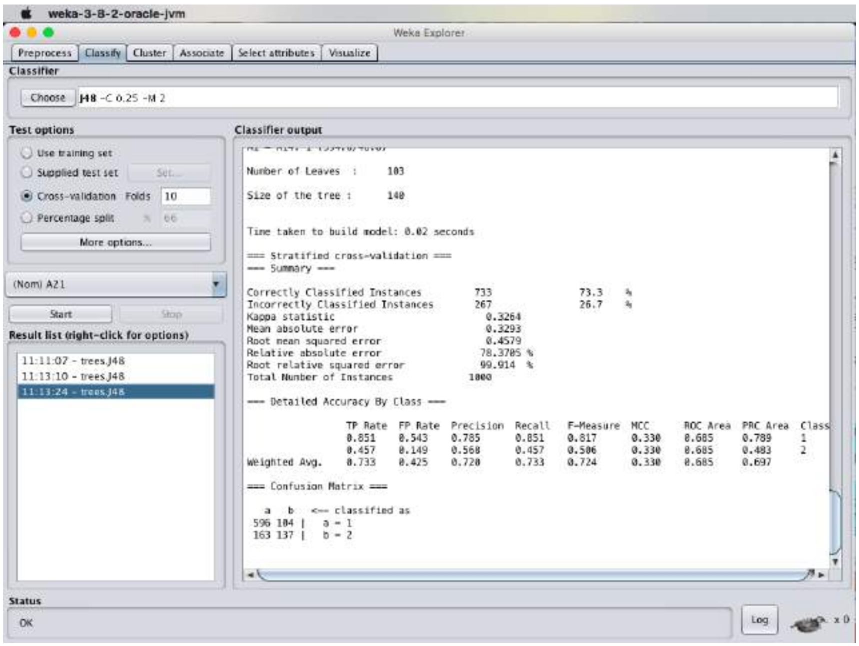
Output :

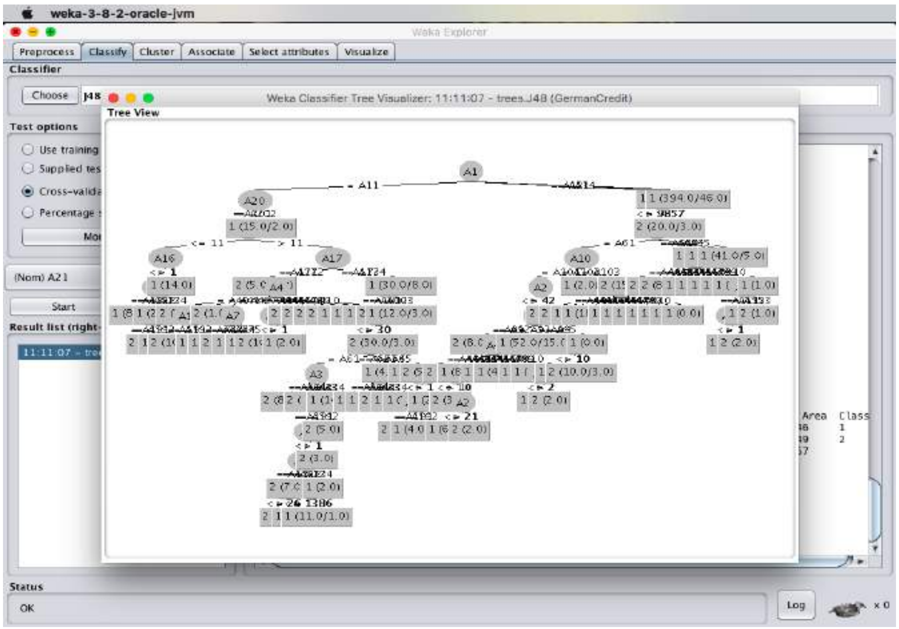


Weka Explorer
Preprocess
Classify
Cluster
Associate
Select attributes
Visualize
Classifier
Choose
J48 -C 0.25 -M 2
Test options
Use training set
Supplied test set
Set...
Cross-validation Folds 10
Percentage split % 66
More options...
(Nom) A21
Start
Stop
Result list (right-click for options)
10:54:18 - trees.J48
Classifier output
J48 pruned tree
A1 = A11
A20 = A201
A2 <= 11
A16 <= 1
A12 = A121: 1 (8.0/1.0)
A12 = A122
A19 = A191: 2 (2.0)
A19 = A192: 1 (4.0)
A12 = A123: 1 (2.0/1.0)
A12 = A124: 2 (3.0)
A16 > 1: 1 (14.0)
A2 > 11
A17 = A171: 2 (5.0/1.0)
A17 = A172
A4 = A40
A19 = A191: 2 (10.0/2.0)
A19 = A192: 1 (2.0)
A4 = A41: 2 (1.0)
A4 = A42
A7 = A71: 1 (0.0)
A7 = A72: 2 (3.0)
A7 = A73: 1 (4.0)
A7 = A74: 1 (1.0)
A7 = A75: 1 (2.0)
A4 = A43
A16 <= 1: 2 (10.0/3.0)
A16 > 1: 1 (2.0)
A4 = A44: 2 (1.0)
A4 = A45: 2 (1.0)
A4 = A46: 2 (1.0)
A4 = A47: 2 (0.0)
Status
OK
Log
x 0



weka-3-8-2-oracle-jvm
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) A21
Start
Stop
Result list (right-click for options)
11:11:07 - trees.J48
11:13:10 - trees.J48
11:13:24 - trees.J48
Classifier output
A4 = A46: 1 (0.0)
A4 = A47: 1 (0.0)
A4 = A48: 1 (0.0)
A4 = A49: 1 (0.0)
A4 = A410: 1 (0.0)
A6 = A62
A4 = A40: 2 (15.0/5.0)
A4 = A41: 1 (3.0)
A4 = A42: 2 (4.0/1.0)
A4 = A43: 2 (8.0/2.0)
A4 = A44: 1 (0.0)
A4 = A45: 1 (2.0)
A4 = A46: 1 (0.0)
A4 = A47: 1 (0.0)
A4 = A48: 1 (0.0)
A4 = A49
A15 = A151
A16 <= 1: 1 (2.0)
A16 > 1: 2 (2.0)
A15 = A152: 1 (6.0)
A15 = A153: 2 (1.0)
A4 = A410: 1 (1.0)
A6 = A63: 1 (11.0/3.0)
A6 = A64: 1 (13.0/3.0)
A6 = A65: 1 (41.0/5.0)
A5 > 9857: 2 (20.0/3.0)
A1 = A13: 1 (63.0/14.0)
A1 = A14: 1 (394.0/46.0)
Number of Leaves : 103
Size of the tree : 140
Status
OK
Log
x 0





D) K-Nearest Neighbor :

K-Nearest Neighbors is one of the most basic yet essential classification algorithms in Machine Learning. It belongs to the supervised learning domain and finds intense application in pattern recognition, data mining and intrusion detection.

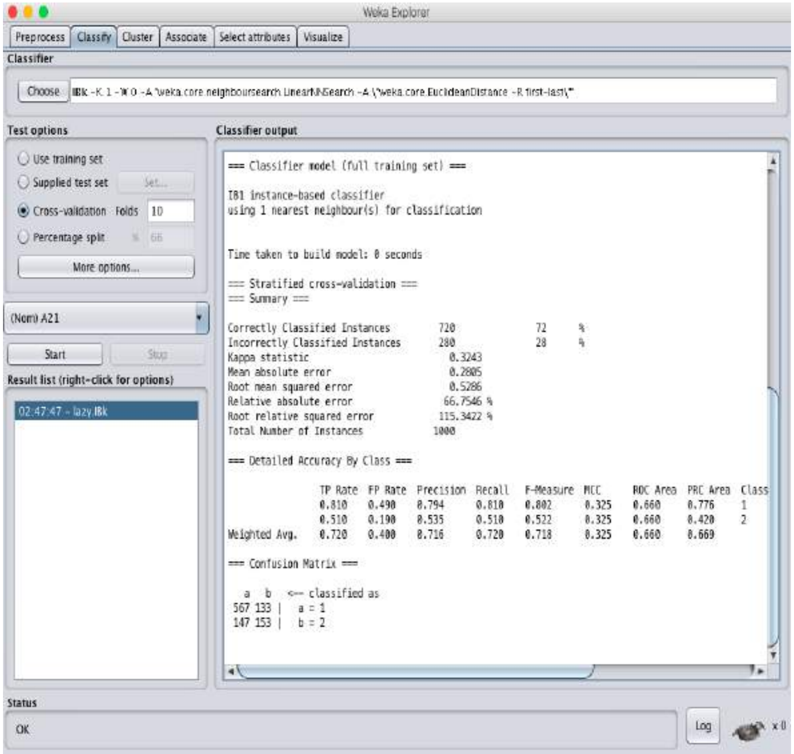
It is widely disposable in real-life scenarios since it is non-parametric, meaning, it does not make any underlying assumptions about the distribution of data (as opposed to other algorithms such as GMM, which assume a Gaussian distribution of the given data).

We are given some prior data (also called training data), which classifies coordinates into groups identified by an attribute.

Steps :

- Load the dataset into the weka tool and preprocess it.
- Apply the classification the K- Nearest Neighbor technique and execute for the result.

Output :

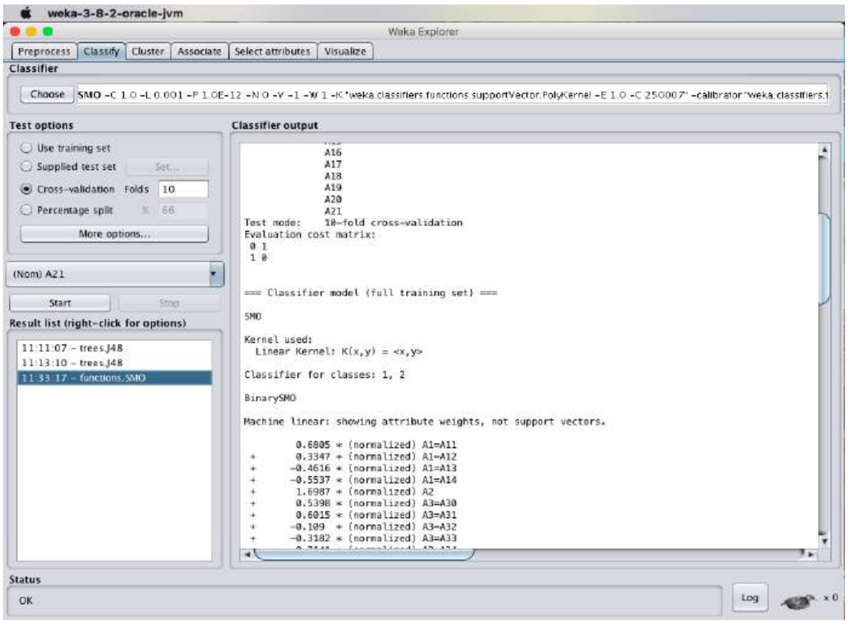


E) SMO Algorithm :

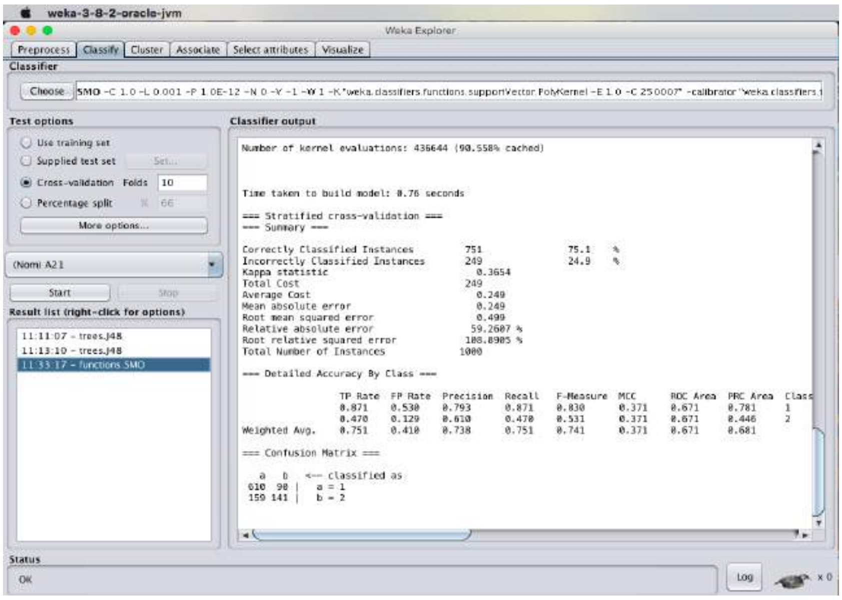
The iterative algorithm Sequential Minimal Optimization (SMO) is used for solving quadratic programming (QP) problems. One example where QP problems are relevant is during the training process of support vector machines (SVM). The SMO algorithm is used to solve in this example a constraint optimization problem. John Platt proposed this algorithm in 1998 and it was successfully used since then. We describe here the basics of the algorithm in the light of big data.

Steps :

- Load the dataset into the weka tool and preprocess it.
- Apply the classification the Sequential Minimal Optimization (SMO)technique and execute for the result.



weka-3-8-2-oracle-jvm
Weka Explorer
Preprocess
Classify
Cluster
Associate
Select attributes
Visualize
Classifier
Choose
SMO -C 1.0 -L 0.001 -P 1.0E-12 -N 0 -V -1 -W 1 -K "weka.classifiers.functions.supportVector.PolyKernel -E 1.0 -C 250007" -calibrator "weka.classifiers.
Test options
Use training set
Supplied test set
Set...
Cross-validation Folds 10
Percentage split % 66
More options...
(Nom) A21
Start
Stop
Result list (right-click for options)
11:11:07 - trees.J48
11:13:10 - trees.J48
11:33:17 - functions.SMO
Classifier output
A16
A17
A18
A19
A20
A21
Test mode: 10-fold cross-validation
Evaluation cost matrix:
0 1
1 0
=== Classifier model (full training set) ===
SMO
Kernel used:
Linear Kernel: K(x,y) = <x,y>
Classifier for classes: 1, 2
BinarySMO
Machine linear: showing attribute weights, not support vectors.
0.6805 * (normalized) A1=A11
+ 0.3347 * (normalized) A1=A12
+ -0.4616 * (normalized) A1=A13
+ -0.5537 * (normalized) A1=A14
+ 1.6987 * (normalized) A2
+ 0.5398 * (normalized) A3=A30
+ 0.6015 * (normalized) A3=A31
+ -0.109 * (normalized) A3=A32
+ -0.3182 * (normalized) A3=A33
Status
OK
Log
x 0



weka-3-8-2-oracle-jvm
Weka Explorer
Preprocess
Classify
Cluster
Associate
Select attributes
Visualize
Classifier
Choose
SMO -C 1.0 -L 0.001 -P 1.0E-12 -N 0 -V -1 -W 1 -K "weka.classifiers.functions.supportVector.PolyKernel -E 1.0 -C 250007" -calibrator "weka.classifiers.
Test options
Use training set
Supplied test set
Set...
Cross-validation Folds 10
Percentage split % 66
More options...
(Nom) A21
Start
Stop
Result list (right-click for options)
11:11:07 - trees.J48
11:13:10 - trees.J48
11:33:17 - functions.SMO
Classifier output
Number of kernel evaluations: 436644 (90.558% cached)
Time taken to build model: 0.76 seconds
=== Stratified cross-validation ===
=== Summary ===
Correctly Classified Instances 751 75.1 %
Incorrectly Classified Instances 249 24.9 %
Kappa statistic 0.3654
Total Cost 249
Average Cost 0.249
Mean absolute error 0.249
Root mean squared error 0.499
Relative absolute error 59.2607 %
Root relative squared error 108.8985 %
Total Number of Instances 1000
=== Detailed Accuracy By Class ===
TP Rate FP Rate Precision Recall F-Measure MCC ROC Area PRC Area Class
0.871 0.530 0.793 0.871 0.830 0.371 0.671 0.781 1
0.470 0.129 0.610 0.470 0.531 0.371 0.671 0.446 2
Weighted Avg. 0.751 0.410 0.738 0.751 0.741 0.371 0.671 0.681
=== Confusion Matrix ===
a b <-- classified as
610 90 | a = 1
159 141 | b = 2
Status
OK
Log
x 0

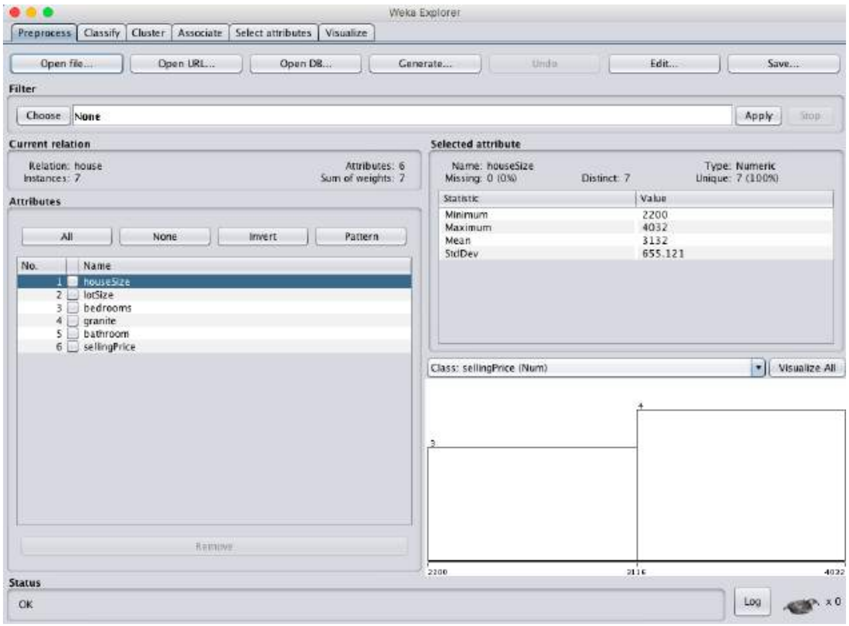
NUMERICAL PREDICTION ANALYSIS USING LINEAR REGRESSION THROUGH WEKA

DESCRIPTION :

Consider a dataset of house.arff where it contains the attributes as house size, lot size, bedrooms, granite, bathroom and the selling price.

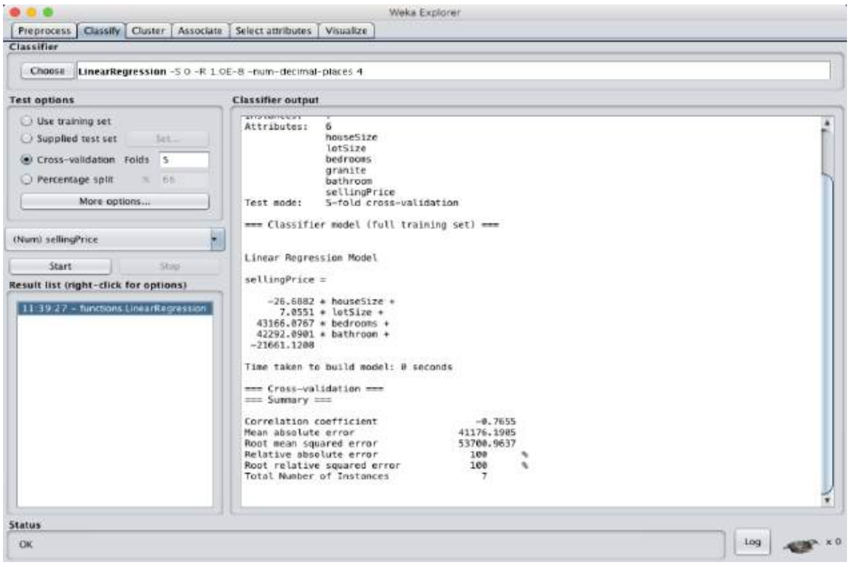
Steps :

- Load the dataset into the weka tool and check for the attributes.
- Classify the data using linear regression analysis method (or) technique.
- Check for the cross-validation folds where the value of the folds should be less than the value of the instances present in the dataset.
- Observe the cross validation summary after applying the linear regression technique for the price of the house.

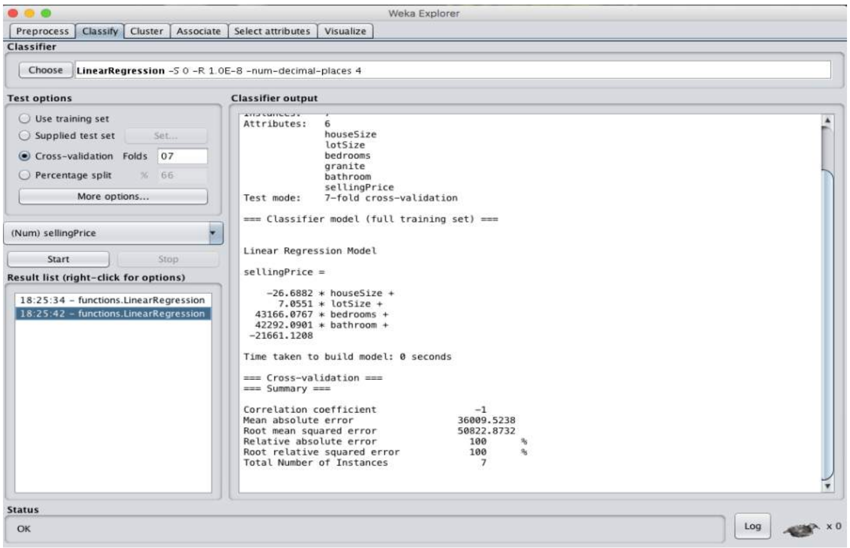


OBSERVATION :

- ❖ **When cross validation folds = 05 :**



❖ **When cross validation folds = 10 :**



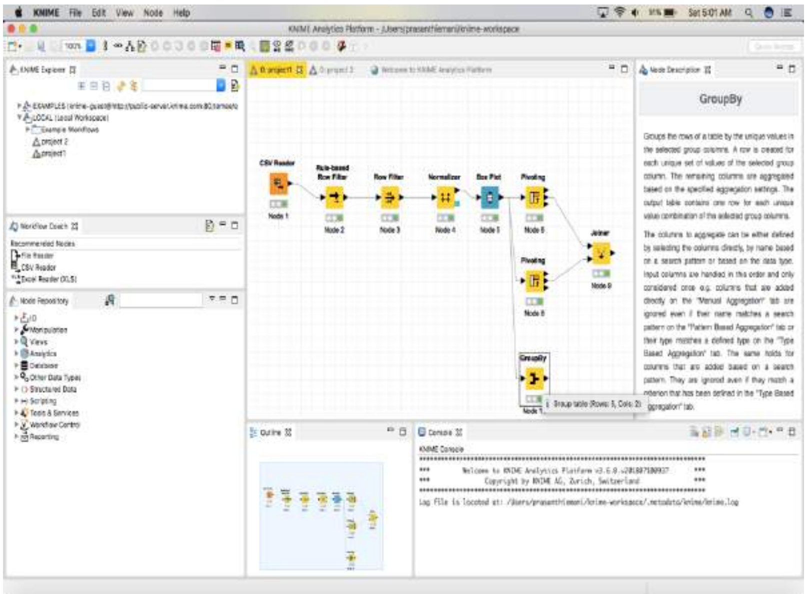
EXTRACT TRANSFORM LOAD (ETL) AND OLAP OPERATION USING KNIME TOOL

DESCRIPTION :

Consider a dataset movies.csv where it contains the attributes ad title, genre, director, year, duration, gross, budget, cast_facebook_likes, votes, reviews, rating.

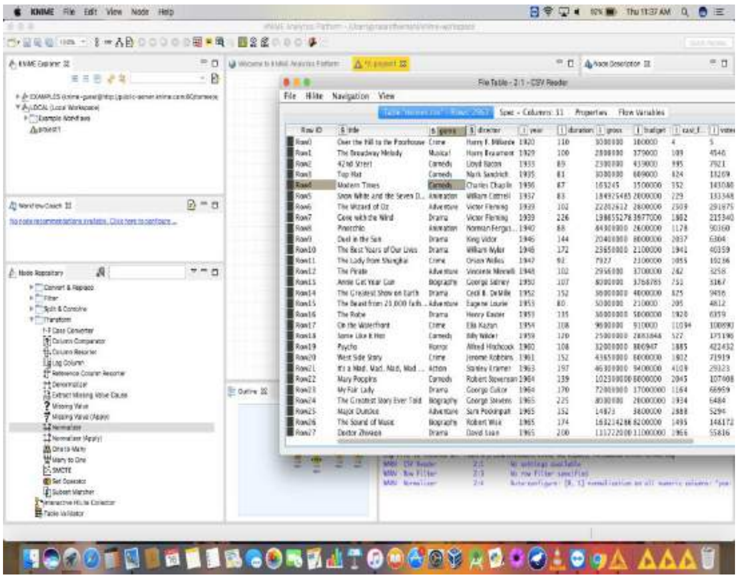
STEPS :

➢ Import the dataset into the knime tool using csv reader.



❖ **CSV Reader :**

Execute the CSV reader,look at the table of the loaded dataset.



❖ **Rule-Based Row Filter :**

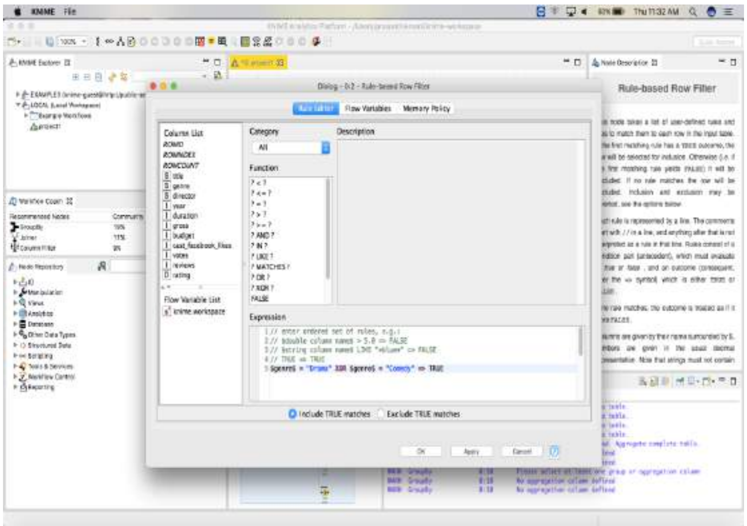
This node takes a list of user-defined rules and tries to match them to each row in the input table. If the first matching rule has a TRUE outcome, the row will be selected for inclusion. Otherwise (i.e. if the first matching rule yields FALSE) it will be excluded. If no rule matches the row will be excluded.

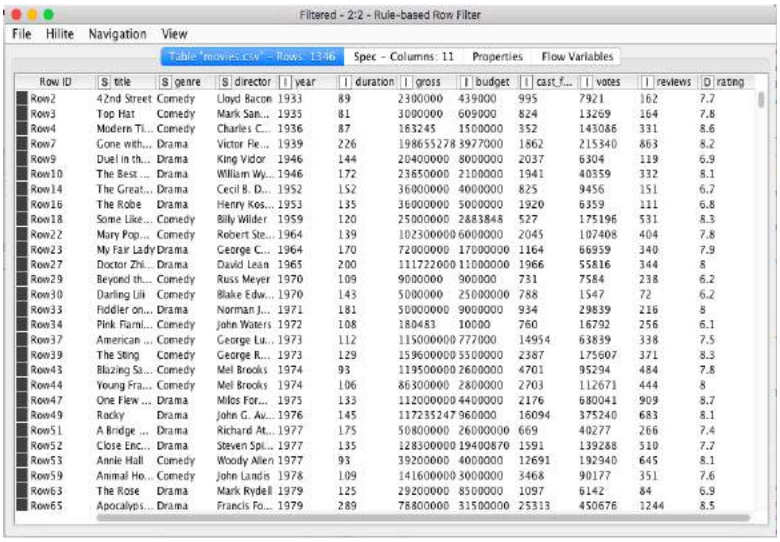
Steps :

- Make a connection between CSV reader and rule based row filter.
- Configure rule based row filter.
- Execute and Check out for the table after applying rule based row filter.

Code :

$genre$ = “Drama” XOR $genre$ = “Comedy” => TRUE




Filtered - 2:2 - Rule-based Row Filter

File Hilite Navigation View

Table "movies.csv" - Rows: 1346 | Spec - Columns: 11 | Properties | Flow Variables

Row ID	S title	S genre	S director	I year	I duration	I gross	I budget	I cast_f...	I votes	I reviews	D rating
Row2	42nd Street	Comedy	Lloyd Bacon	1933	89	2300000	439000	995	7921	162	7.7
Row3	Top Hat	Comedy	Mark San...	1935	81	3000000	609000	824	13269	164	7.8
Row4	Modern Ti...	Comedy	Charles C...	1936	87	163245	1500000	352	143086	331	8.6
Row7	Gone with...	Drama	Victor Fle...	1939	226	198655278	3977000	1862	215340	863	8.2
Row9	Duel in th...	Drama	King Vidor	1946	144	20400000	8000000	2037	6304	119	6.9
Row10	The Best ...	Drama	William Wy...	1946	172	23650000	2100000	1941	40359	332	8.1
Row14	The Great...	Drama	Cecil B. D...	1952	152	36000000	4000000	825	9456	151	6.7
Row16	The Robe	Drama	Henry Kos...	1953	135	36000000	5000000	1920	6359	111	6.8
Row18	Some Like...	Comedy	Billy Wilder	1959	120	25000000	2883848	527	175196	531	8.3
Row22	Mary Pop...	Comedy	Robert Ste...	1964	139	102300000	6000000	2045	107408	404	7.8
Row23	My Fair Lady	Drama	George C...	1964	170	72000000	17000000	1164	66959	340	7.9
Row27	Doctor Zhi...	Drama	David Lean	1965	200	111722000	11000000	1966	55816	344	8
Row29	Beyond th...	Comedy	Russ Meyer	1970	109	9000000	900000	731	7584	238	6.2
Row30	Darling Lili	Comedy	Blake Edw...	1970	143	5000000	25000000	788	1547	72	6.2
Row33	Fiddler on...	Drama	Norman J...	1971	181	50000000	9000000	934	29839	216	8
Row34	Pink Flami...	Comedy	John Waters	1972	108	180483	10000	760	16792	256	6.1
Row37	American ...	Comedy	George Lu...	1973	112	115000000	777000	14954	63839	338	7.5
Row39	The Sting	Comedy	George R...	1973	129	159600000	5500000	2387	175607	371	8.3
Row43	Blazing Sa...	Comedy	Mel Brooks	1974	93	119500000	2600000	4701	95294	484	7.8
Row44	Young Fra...	Comedy	Mel Brooks	1974	106	86300000	2800000	2703	112671	444	8
Row47	One Flew ...	Drama	Milos For...	1975	133	112000000	4400000	2176	680041	909	8.7
Row49	Rocky	Drama	John G. Av...	1976	145	117235247	960000	16094	375240	683	8.1
Row51	A Bridge ...	Drama	Richard At...	1977	175	50800000	26000000	669	40277	266	7.4
Row52	Close Enc...	Drama	Steven Spi...	1977	135	128300000	19400870	1591	139288	510	7.7
Row53	Annie Hall	Comedy	Woody Allen	1977	93	39200000	4000000	12691	192940	645	8.1
Row59	Animal Ho...	Comedy	John Landis	1978	109	141600000	3000000	3468	90177	351	7.6
Row63	The Rose	Drama	Mark Rydell	1979	125	29200000	8500000	1097	6142	84	6.9
Row65	Apocalyps...	Drama	Francis Fo...	1979	289	78800000	31500000	25313	450676	1244	8.5

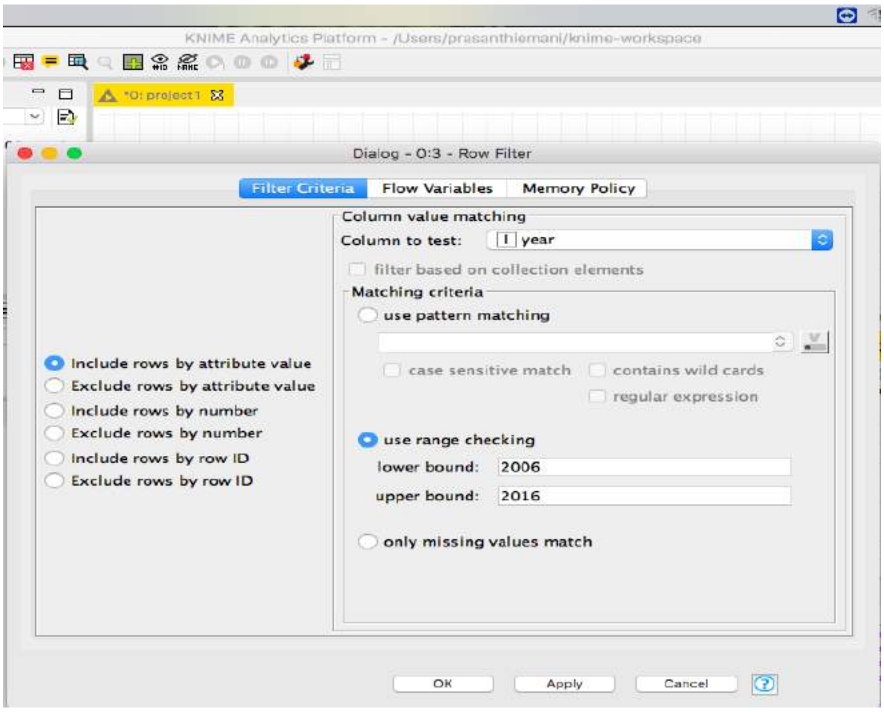
❖ **Row Filter :**

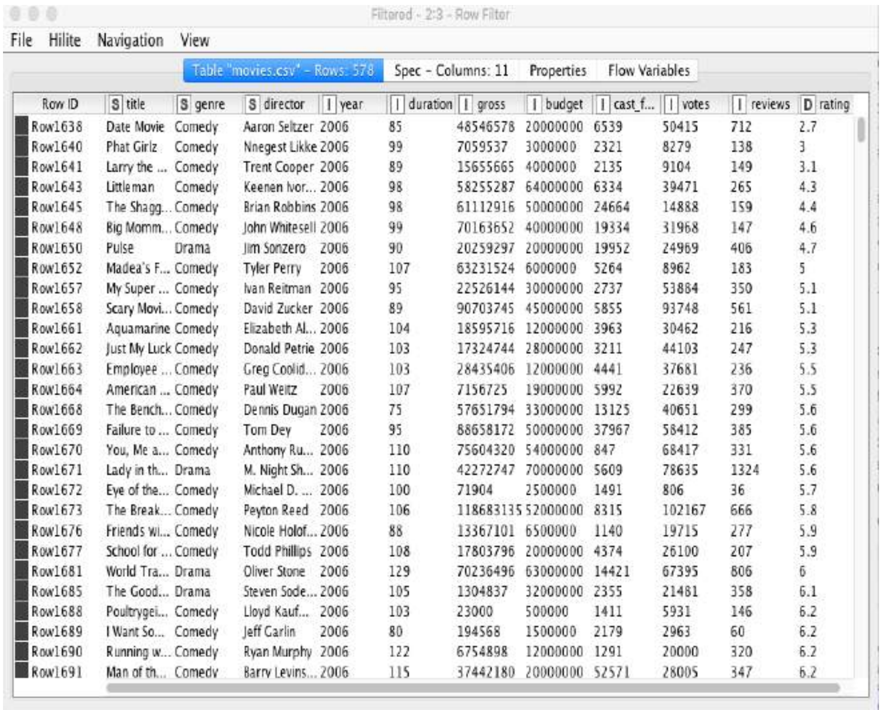
3 matching criteria on data colums: on String by full or partial pattern matching, on numbers by range, on missing values, all of them also on collection columns. 1 matching criterion on row numbers: from row number to row number. 1 matching criterion on RowID: full and partial patterm matching. Partial pattern matching is obtained through wild cards and RegEx. All matching criteria can be used in Include or Exclude mode. Include keeps the match results. Exclude excludes it.

Steps :

- Make a connection between rule based row filter and row filter.
- Configure row filter.
- Execute and Check out for the table after applying row filter.

➢ Use range checking :
Lower Bound : 2006
Upper Bound : 2016



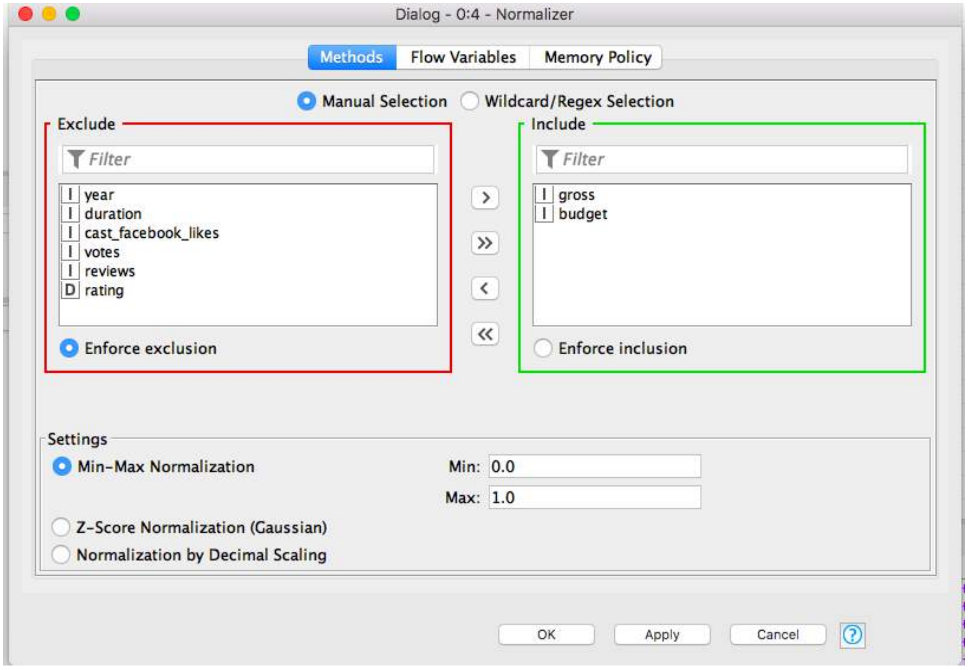


Row ID	S title	S genre	S director	I year	I duration	I gross	I budget	I cast_f...	I votes	I reviews	D rating
Row1638	Date Movie	Comedy	Aaron Seltzer	2006	85	48546578	20000000	6539	50415	712	2.7
Row1640	Phat Girlz	Comedy	Nnegest Likke	2006	99	7059537	3000000	2321	8279	138	3
Row1641	Larry the ...	Comedy	Trent Cooper	2006	89	15655665	4000000	2135	9104	149	3.1
Row1643	Littleman	Comedy	Keenen Ivor...	2006	98	58255287	64000000	6334	39471	265	4.3
Row1645	The Shagg...	Comedy	Brian Robbins	2006	98	61112916	50000000	24664	14888	159	4.4
Row1648	Big Momm...	Comedy	John Whitesell	2006	99	70163652	40000000	19334	31968	147	4.6
Row1650	Pulse	Drama	Jim Sonzero	2006	90	20259297	20000000	19952	24969	406	4.7
Row1652	Madea's F...	Comedy	Tyler Perry	2006	107	63231524	6000000	5264	8962	183	5
Row1657	My Super ...	Comedy	Ivan Reitman	2006	95	22526144	30000000	2737	53884	350	5.1
Row1658	Scary Movi...	Comedy	David Zucker	2006	89	90703745	45000000	5855	93748	561	5.1
Row1661	Aquamarine	Comedy	Elizabeth Al...	2006	104	18595716	12000000	3963	30462	216	5.3
Row1662	Just My Luck	Comedy	Donald Petrie	2006	103	17324744	28000000	3211	44103	247	5.3
Row1663	Employee ...	Comedy	Greg Coolid...	2006	103	28435406	12000000	4441	37681	236	5.5
Row1664	American ...	Comedy	Paul Weitz	2006	107	7156725	19000000	5992	22639	370	5.5
Row1668	The Bench...	Comedy	Dennis Dugan	2006	75	57651794	33000000	13125	40651	299	5.6
Row1669	Failure to ...	Comedy	Tom Dey	2006	95	88658172	50000000	37967	58412	385	5.6
Row1670	You, Me a...	Comedy	Anthony Ru...	2006	110	75604320	54000000	847	68417	331	5.6
Row1671	Lady in th...	Drama	M. Night Sh...	2006	110	42272747	70000000	5609	78635	1324	5.6
Row1672	Eye of the...	Comedy	Michael D. ...	2006	100	71904	2500000	1491	806	36	5.7
Row1673	The Break...	Comedy	Peyton Reed	2006	106	118683135	52000000	8315	102167	666	5.8
Row1676	Friends wi...	Comedy	Nicole Holof...	2006	88	13367101	6500000	1140	19715	277	5.9
Row1677	School for ...	Comedy	Todd Phillips	2006	108	17803796	20000000	4374	26100	207	5.9
Row1681	World Tra...	Drama	Oliver Stone	2006	129	70236496	63000000	14421	67395	806	6
Row1685	The Good...	Drama	Steven Sode...	2006	105	1304837	32000000	2355	21481	358	6.1
Row1688	Poultrygei...	Comedy	Lloyd Kauf...	2006	103	23000	500000	1411	5931	146	6.2
Row1689	I Want So...	Comedy	Jeff Carlin	2006	80	194568	1500000	2179	2963	60	6.2
Row1690	Running w...	Comedy	Ryan Murphy	2006	122	6754898	12000000	1291	20000	320	6.2
Row1691	Man of th...	Comedy	Barry Levins...	2006	115	37442180	20000000	52571	28005	347	6.2

❖ Normalizer :

Steps :

- Connect normalizer with the row filter.
- Configure normalizer as methods which are to be included for normalization technique and set min and max values.
- Include:
  a) Gross
  b) Budget
  c) Min : 0.0
  d) Max : 0.1
- Execute the normalizer and check for the values in the table where you will find the normalized values of the table.



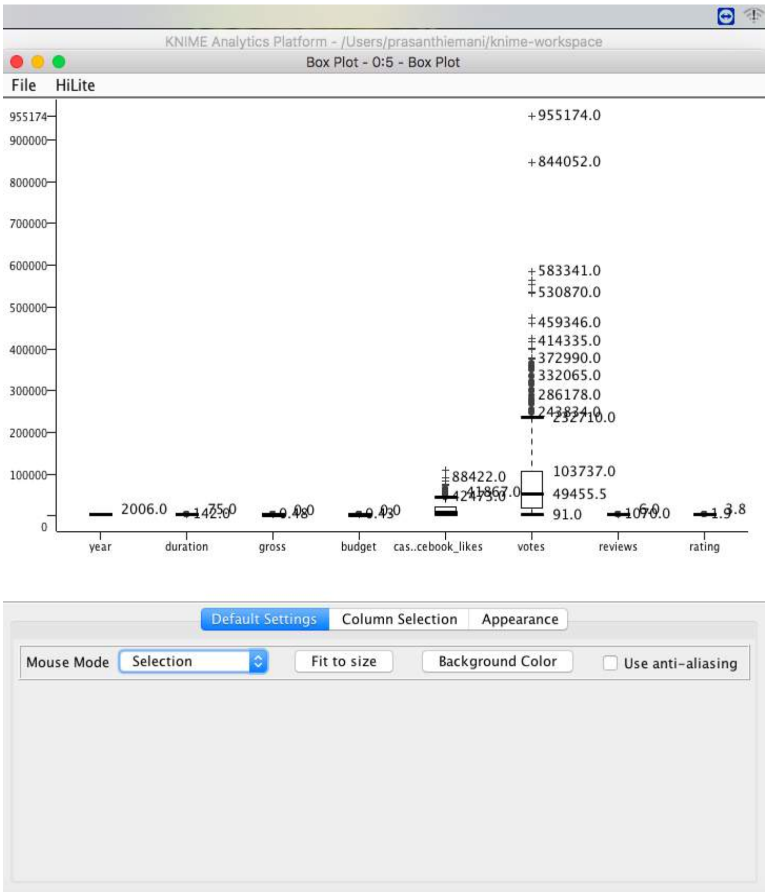
❖ Boxplot :

A box plot displays robust statistical parameters: minimum, lower quartile, median, upper quartile, and maximum. A box plot for one numerical attribute is constructed in the following way: The box itself goes from lower quartile (Q1) to upper quartile (Q3). Median is drawn as horizontal bar inside box. Distance between Q1 and Q3 is called interquartile range (IQR). Above and below box are so-called whiskers. They are drawn at minimum and maximum value as horizontal bars and are connected with the box by a dotted line.

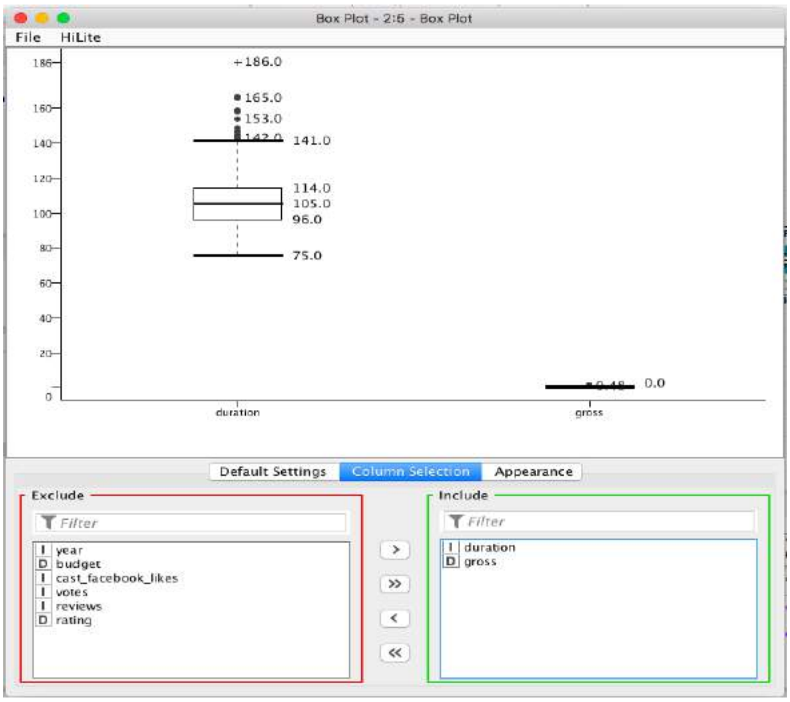
Steps :

- Make connection between normalizer and boxplot.
- View for the boxplot directly.
- We can select the specific columns for the individual boxplot through column selection.

➢ **Boxplot for the whole of dataset :**



➢ **Boxplot for the two of the attributes : gross and duration :**

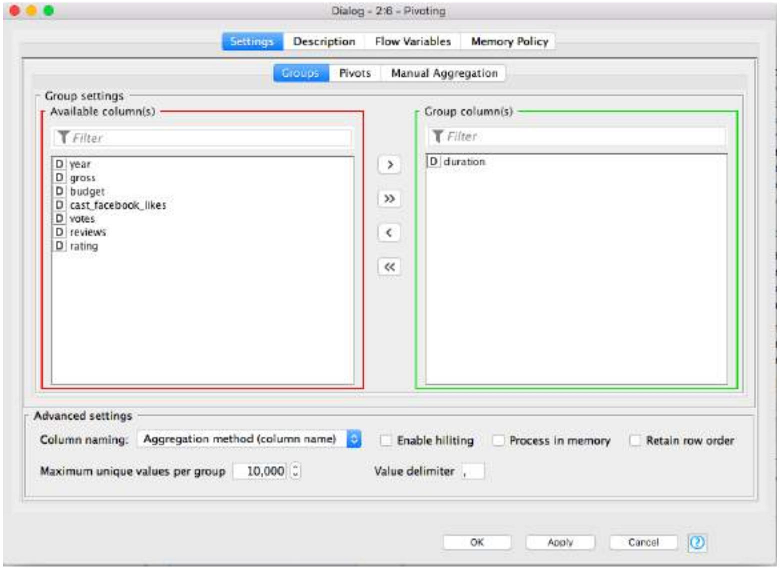


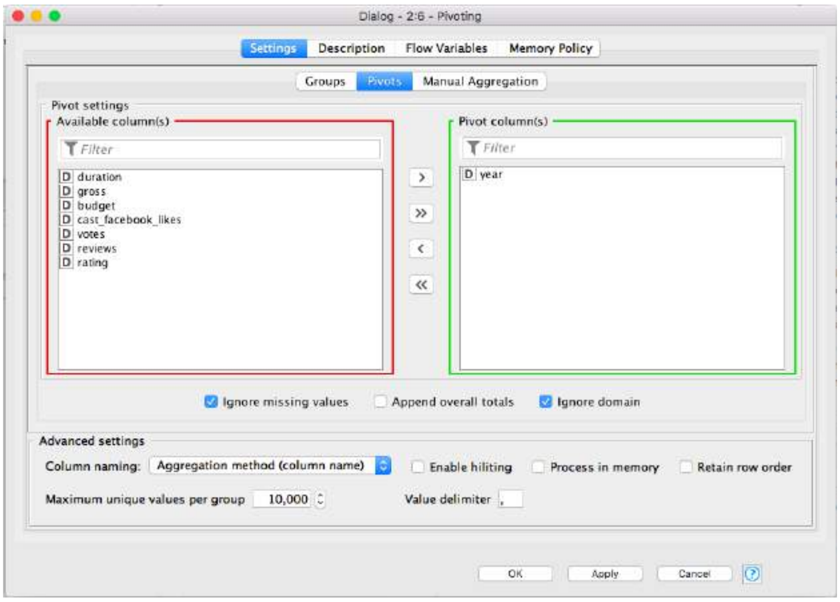
- **Pivoting :**

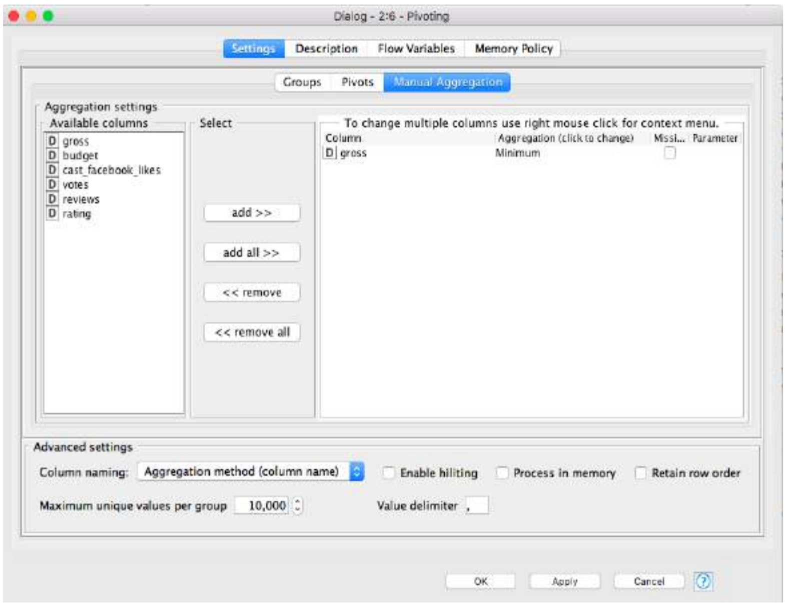
Performs a pivoting on the given input table using a selected number of columns for grouping and pivoting. The group columns will result into unique rows, whereby the pivot values turned into columns for each set of column combinations together with each aggregation method. In addition, the node returns the total aggregation (a) based on only the group columns and (b) based on only the pivoted columns resulting in a single row; optionally, with the total aggregation without pivoting.

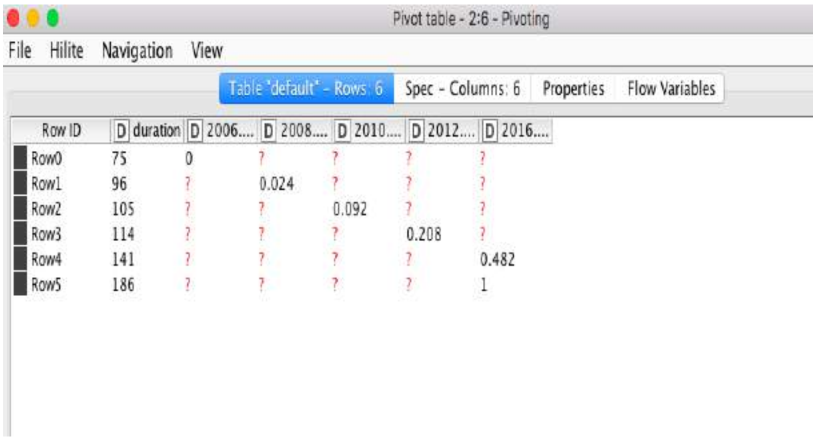
Steps :

- Connect pivot with boxplot and have the connection between them.
- Configure pivoting with 3 different columns for same data type for :
  Groups – duration
  Pivots - year
  Manual Aggregation – gross
- Execute pivoting and checkout for the changes in the table.


Dialog - 2:6 - Pivoting
Settings
Description
Flow Variables
Memory Policy
Groups
Pivots
Manual Aggregation
Group settings
Available column(s)
Filter
D year
D gross
D budget
D cast_facebook_likes
D votes
D reviews
D rating
Group column(s)
Filter
D duration
Advanced settings
Column naming:
Aggregation method (column name)
Enable hiliting
Process in memory
Retain row order
Maximum unique values per group
10,000
Value delimiter
OK
Apply
Cancel


Dialog - 2:6 - Pivoting
Settings
Description
Flow Variables
Memory Policy
Groups
Pivots
Manual Aggregation
Pivot settings
Available column(s)
Filter
D duration
D gross
D budget
D cast_facebook_likes
D votes
D reviews
D rating
Pivot column(s)
Filter
D year
Ignore missing values
Append overall totals
Ignore domain
Advanced settings
Column naming:
Aggregation method (column name)
Enable hiliting
Process in memory
Retain row order
Maximum unique values per group
10,000
Value delimiter
OK
Apply
Cancel


Dialog - 2:6 - Pivoting
Settings
Description
Flow Variables
Memory Policy
Groups
Pivots
Manual Aggregation
Aggregation settings
Available columns
D gross
D budget
D cast_facebook_likes
D votes
D reviews
D rating
Select
add >>
add all >>
<< remove
<< remove all
To change multiple columns use right mouse click for context menu.
Column
Aggregation (click to change)
Missi...
Parameter
D gross
Minimum
Advanced settings
Column naming:
Aggregation method (column name)
Enable hiliting
Process in memory
Retain row order
Maximum unique values per group
10,000
Value delimiter
OK
Apply
Cancel



Row ID	D duration	D 2006....	D 2008....	D 2010....	D 2012....	D 2016....
Row0	75	0	?	?	?	?
Row1	96	?	0.024	?	?	?
Row2	105	?	?	0.092	?	?
Row3	114	?	?	?	0.208	?
Row4	141	?	?	?	?	0.482
Row5	186	?	?	?	?	1

- **Joiner :**

A Joiner node joins two tables together on one or more common key values. Possible join modes: inner join, left outer join, right outer join, full outer join. Two tabs: "Joiner Settings" and "Column Selection". "Joiner Settings" defines the parameters for the join operation: join mode and column keys. "Column Selection" sets which columns to keep and/or drop and strategies to deal with duplicate columns.

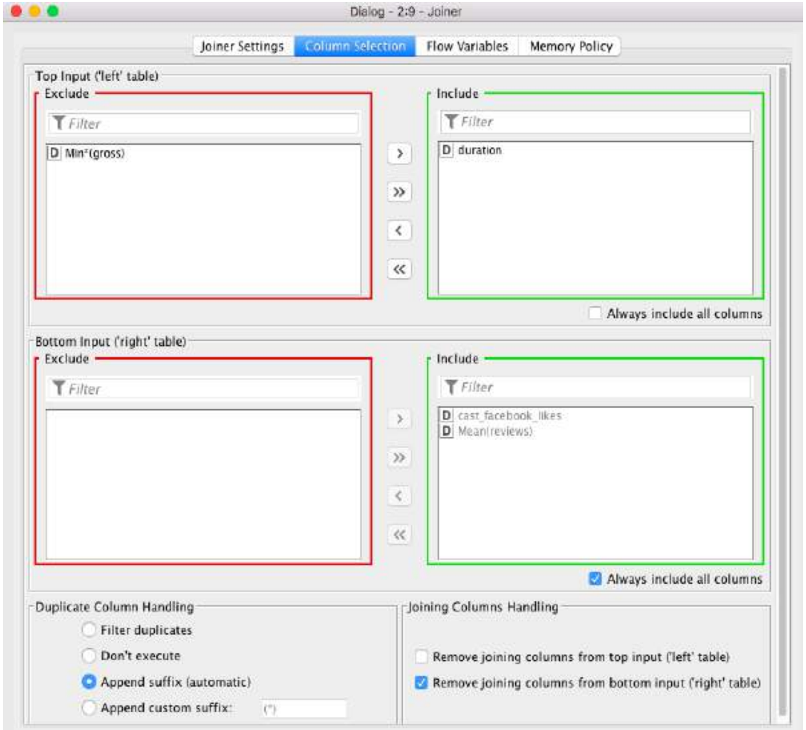
Steps :

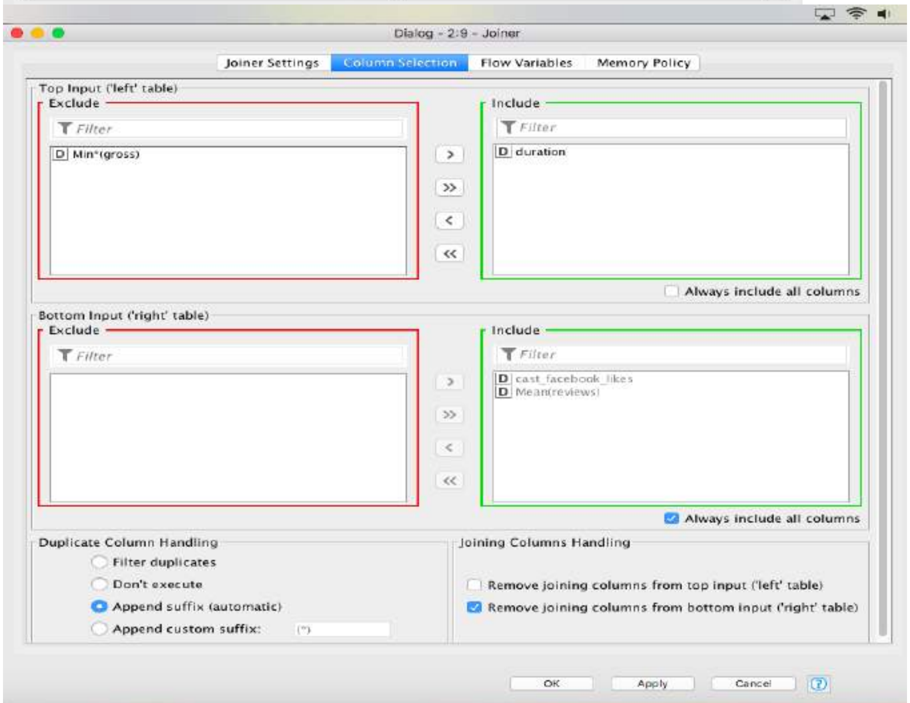
- Have the connection between joiner and pivot so that it is easy to analyse.
- Configure joiner with columns if necessary as :

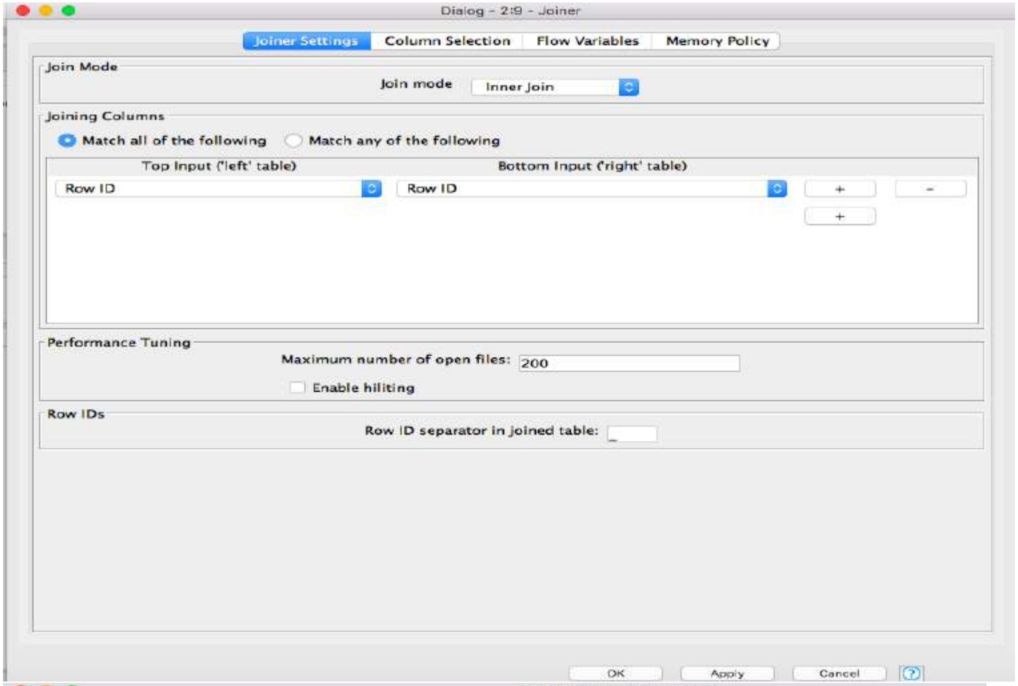
➢ Top Input (left table):
Include : duration

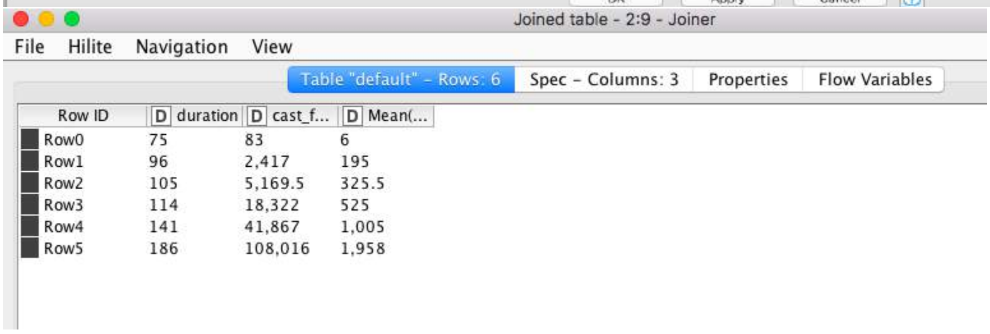
➢ Bottom Input(right table):
Include :
→cast_facebook_likes
→ Mean

- Execute the joiner and check for the final table.


Dialog - 2:9 - Joiner
Joiner Settings
Column Selection
Flow Variables
Memory Policy
Top Input ('left' table)
Exclude
Filter
D Min*(gross)
Include
Filter
D duration
Always include all columns
Bottom Input ('right' table)
Exclude
Filter
Include
Filter
D cast_facebook_likes
D Mean(reviews)
Always include all columns
Duplicate Column Handling
Filter duplicates
Don't execute
Append suffix (automatic)
Append custom suffix:
(*)
Joining Columns Handling
Remove joining columns from top input ('left' table)
Remove joining columns from bottom input ('right' table)


Dialog - 2:9 - Joiner
Joiner Settings
Column Selection
Flow Variables
Memory Policy
Top Input ('left' table)
Exclude
Filter
D Min*(gross)
Include
Filter
D duration
Always include all columns
Bottom Input ('right' table)
Exclude
Filter
Include
Filter
D cast_facebook_likes
D Mean(reviews)
Always include all columns
Duplicate Column Handling
Filter duplicates
Don't execute
Append suffix (automatic)
Append custom suffix:
(*)
Joining Columns Handling
Remove joining columns from top input ('left' table)
Remove joining columns from bottom input ('right' table)
OK
Apply
Cancel





❖ **Group By :**

Groups the rows of a table by the unique values in the selected group columns. A row is created for each unique set of values of the selected group column. The remaining columns are aggregated based on the specified aggregation settings. The output table contains one row for each unique value combination of the selected group columns.

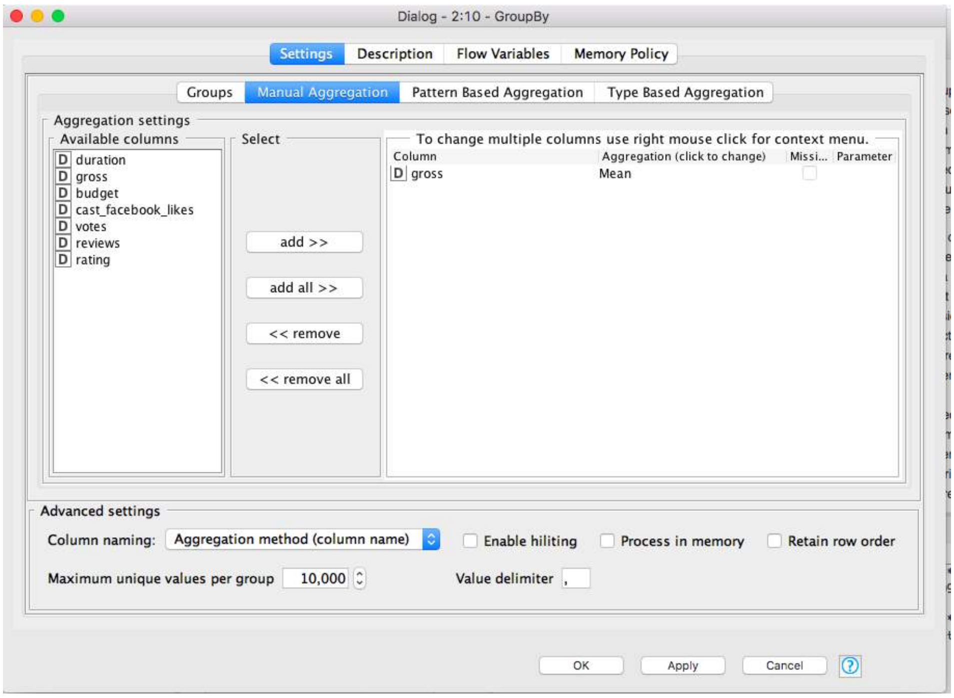
Steps :

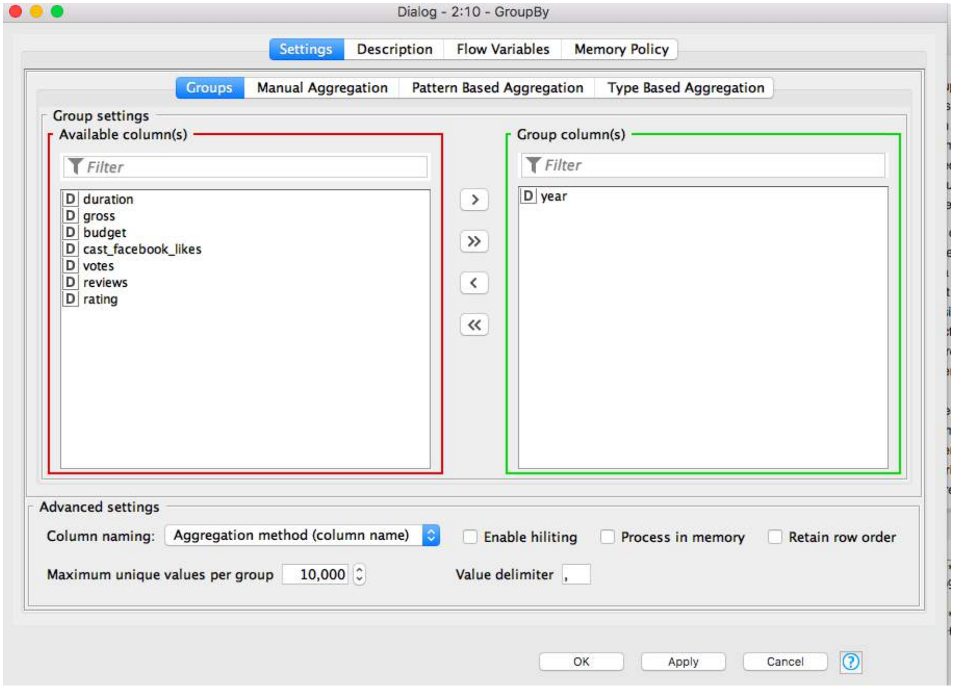
- Make the connection to group by with the boxplot directly.
- Configure group by as the following :

➢ Groups : year

➢ Manual Aggregation : gross

- Execute the group by and check for the analysed table.


Dialog - 2:10 - GroupBy
Settings
Description
Flow Variables
Memory Policy
Groups
Manual Aggregation
Pattern Based Aggregation
Type Based Aggregation
Aggregation settings
Available columns
duration
gross
budget
cast_facebook_likes
votes
reviews
rating
Select
add >>
add all >>
<< remove
<< remove all
To change multiple columns use right mouse click for context menu.
Column
Aggregation (click to change)
Missi...
Parameter
gross
Mean
Advanced settings
Column naming:
Aggregation method (column name)
Enable hiliting
Process in memory
Retain row order
Maximum unique values per group
10,000
Value delimiter
OK
Apply
Cancel


Dialog - 2:10 - GroupBy
Settings
Description
Flow Variables
Memory Policy
Groups
Manual Aggregation
Pattern Based Aggregation
Type Based Aggregation
Group settings
Available column(s)
Filter
duration
gross
budget
cast_facebook_likes
votes
reviews
rating
Group column(s)
Filter
year
Advanced settings
Column naming:
Aggregation method (column name)
Enable hiliting
Process in memory
Retain row order
Maximum unique values per group
10,000
Value delimiter
OK
Apply
Cancel


Group table - 2:10 - GroupBy
File
Hilite
Navigation
View
Table "default" - Rows: 5
Spec - Columns: 2
Properties
Flow Variables
Row ID
year
Mean(...
Row0 2,006 0
Row1 2,008 0.024
Row2 2,010 0.092
Row3 2,012 0.208
Row4 2,016 0.741