

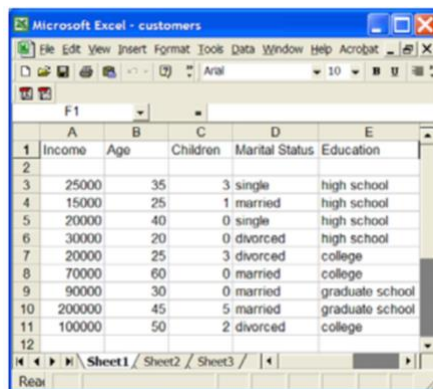
Predictive Analytics BUS2004

Week 4 Workshop

Data Clustering

In this week, we are to work with data Clustering in Weka. Please follow the steps to perform the following tasks:

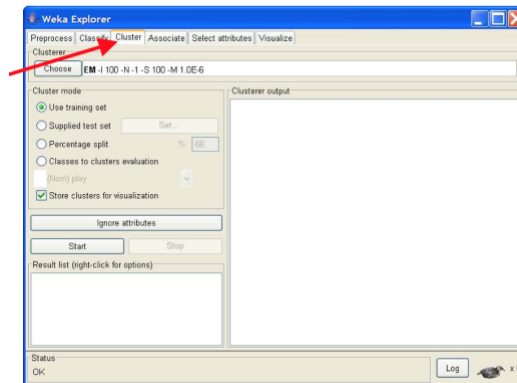
PART I: Clustering data:



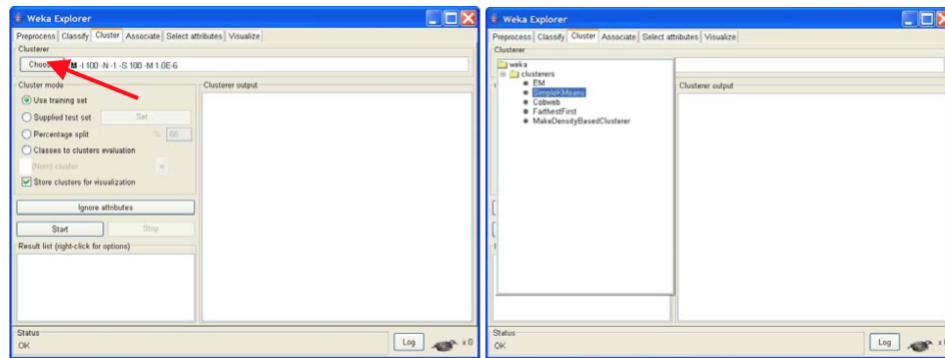
A screenshot of a Microsoft Excel spreadsheet titled 'customers'. The spreadsheet contains a table with 5 columns: Income, Age, Children, Marital Status, and Education. The data is as follows:

	A	B	C	D	E
1	Income	Age	Children	Marital Status	Education
2					
3	25000	35	3	single	high school
4	15000	25	1	married	high school
5	20000	40	0	single	high school
6	30000	20	0	divorced	high school
7	20000	25	3	divorced	college
8	70000	60	0	married	college
9	90000	30	0	married	graduate school
10	200000	45	5	married	graduate school
11	100000	50	2	divorced	college
12					

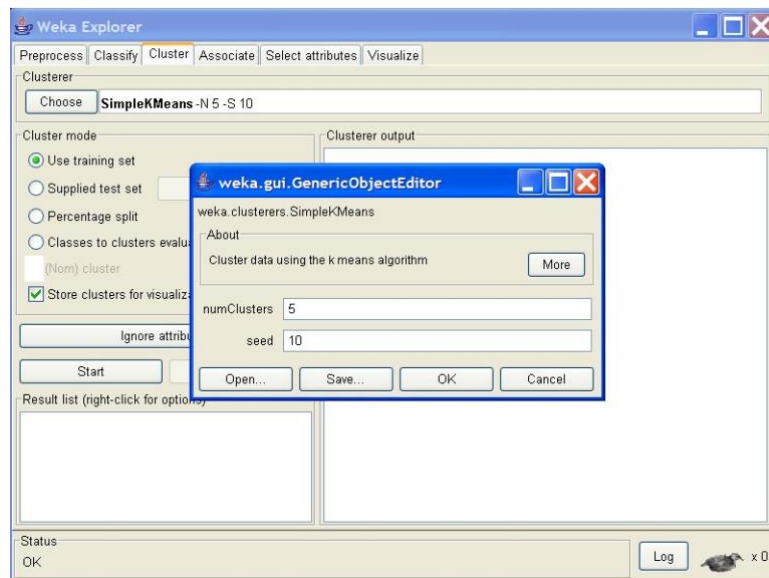
A company X is an international company that wants to group the customer based on similarity. The data has no predefined labels



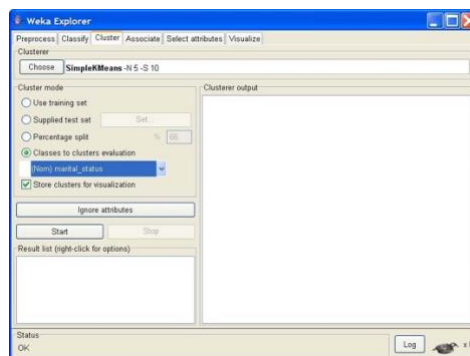
1. Open the file **customers.arff** and load it in Weka: From the main interface, click on the open file button to load the file.

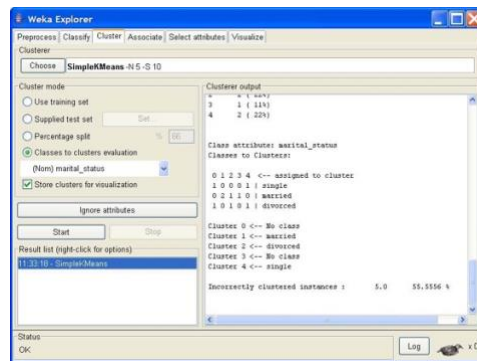


2. Choose Clustering Algorithm:



- 2.1. Under Clustering tab, click the “Choose” button.
- 2.2. From the drop-down list, choose Weka then Clusters.
- 2.3. Select **SimpleKMeans**
- 2.4. right-click on the algorithm, “**weka.gui.GenericObjectEditor**”
- 2.5. Set the value in “numClusters” box to 5
3. Setting Test Options
 - 3.1. It is important to choose the “Cluster mode” before the next step.
 - 3.2. Click on “Classes to cluster evaluation” under cluster mode.
 - 3.3. Select the “marital_status” from the list.
 - 3.4. Click on the ‘Start’ button to execute the algorithm.





PART II: Analysing Results

Fill in the following tables with the results from your clustering:

Run Information gives you the following information:

- The clustering scheme used: _____ with _____ clusters
- The relation's name _____
- Number of instances in the relation. _____
- Number of attributes in the relation _____
- List of attributes used in clustering _____

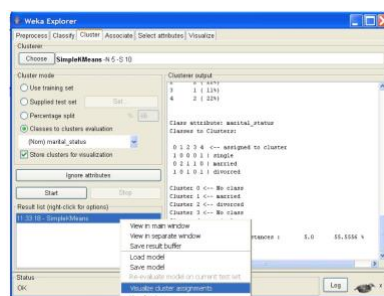
WEKA generated clusters are:

Cluster	Age groups	Marital status	Income	Children	Instances

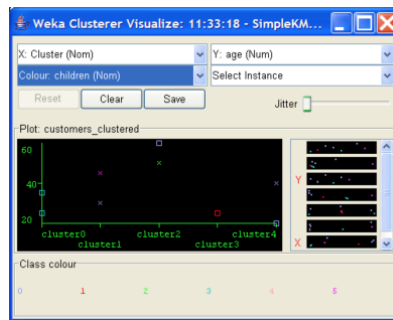
We have _____ incorrectly classified instances, which is _____ %.

PART III: Visualization of Results

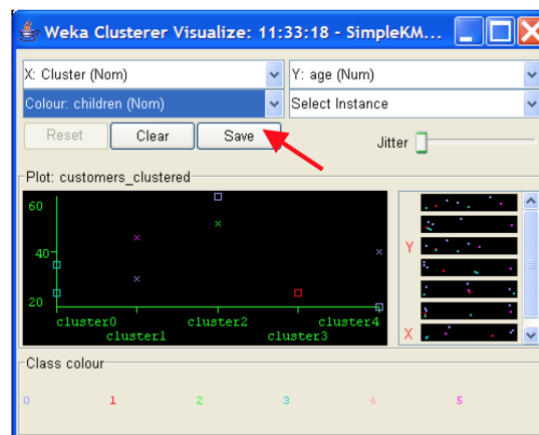
Visualising results and outcomes of a clustering process might have more meaning added to the data and could lead to a better is extraction of insights. In Weka, it is possible to add visualisations of the data and the model as well.



Next window will display 'Weka Clusterer Visualize' window.



1. On the 'Weka Clusterer Visualize' window, from drop-down list, choose the colour scheme.
2. Click on '3' in the 'Class colour' box and select lighter colour from the colour palette.
3. Set X - axis to 'Cluster' attribute, Y - axis to 'Age'.
From the set of horizontal stripes on the right side, you can choose what axes are used in the main graph by clicking on these stripes (click for X-axis, right-click for Y-axis).
4. Select 'Children' as the colour dimension.
The initially correctly clustered instances are represented by crosses, incorrectly clustered once represented as squares. By changing the colour dimension to other attributes, you can see their distribution within each of the clusters.
5. To do so, click 'Save' button in the visualization window and save the result as the file "customers_kmeans.arff".



6. Please note that a new attribute, named – 'cluster', is added to the file. This was added by WEKA and represents the clustering performed by WEKA.
7. Repeat Parts I, II and III using 2 additional clustering algorithms.