St. Francis Institute of Technology, Mumbai-400 103

Department Of Information Technology

A.Y. 2024-2025
Class: TE-ITA/B, Semester: VI

Subject: **Business Intelligence Lab**

Experiment – 7: To implement K-means clustering algorithm using open source tools, WEKA & ORANGE

1. **Aim: :** To implement K-means clustering algorithm using open source tools, WEKA & ORANGE

1. **Objectives:** After study of this experiment, the students will be able to
   Implement K Means

1. **Outcomes:** After study of this experiment, the students will be able to

   CO 4: Design and Implement various clustering data mining techniques such as Partitioning methods, Hierarchical Methods, Density - Based methods along with identification and analysis of outlier.

1. **Prerequisite:** Introduction to all the three clustering algorithms & Problem solving approach.

1. **Requirements:** Personal Computer, Windows XP operating system/Windows 7, Internet Connection, Microsoft Word, WEKA tool, ORANGE tool.

1. **Theory:**
   a. Explain K means (graph) algorithm
   b. Explain K medoids algorithm

1. **Laboratory Exercise:** Implementation of K means clustering algorithm using WEKA & ORANGE along with screenshots.

1. **Post-Experiments Exercise**

a. **Questions:**
   - Explain advantages and disadvantages of K means
   - K means (graph) solved numerical

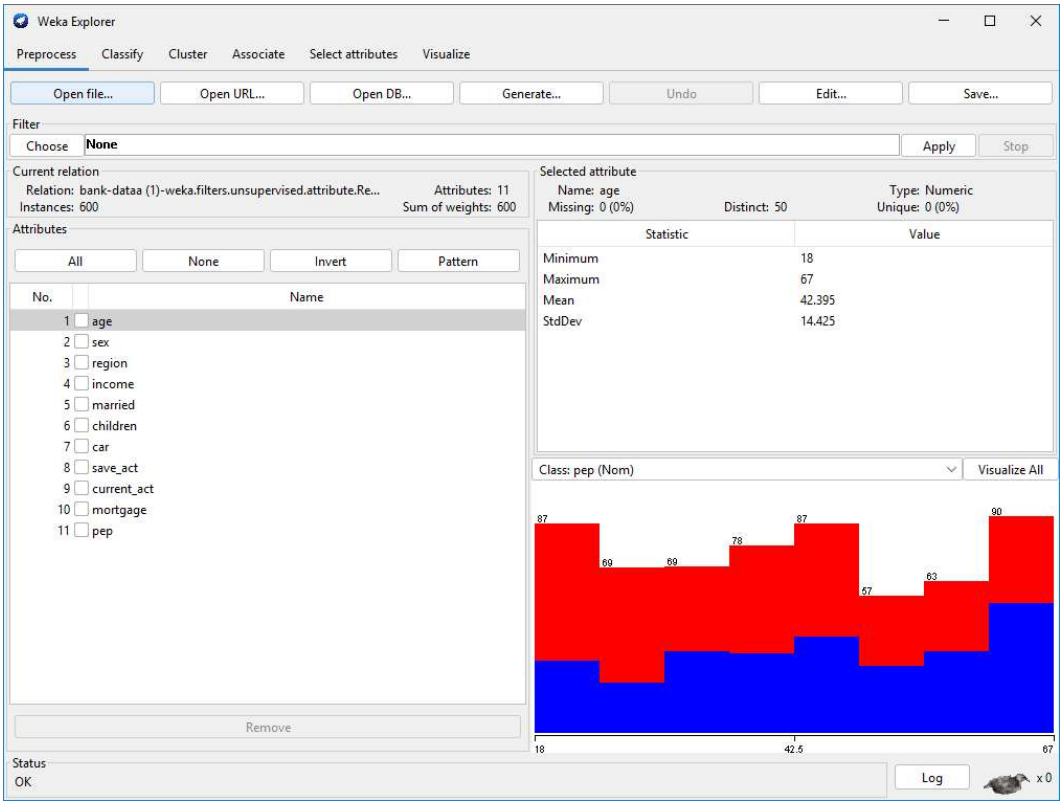
a. **Conclusion:**
   - Summary of Experiment
   - Importance of Experiment
   - Application of Experiment

Name: Vishal Rajesh Mahajan BI Lab EXP 07

Class: TE IT A Roll No: 56

K-Means Clustering in Weka

Loading Pre-processed Dataset in WEKA:



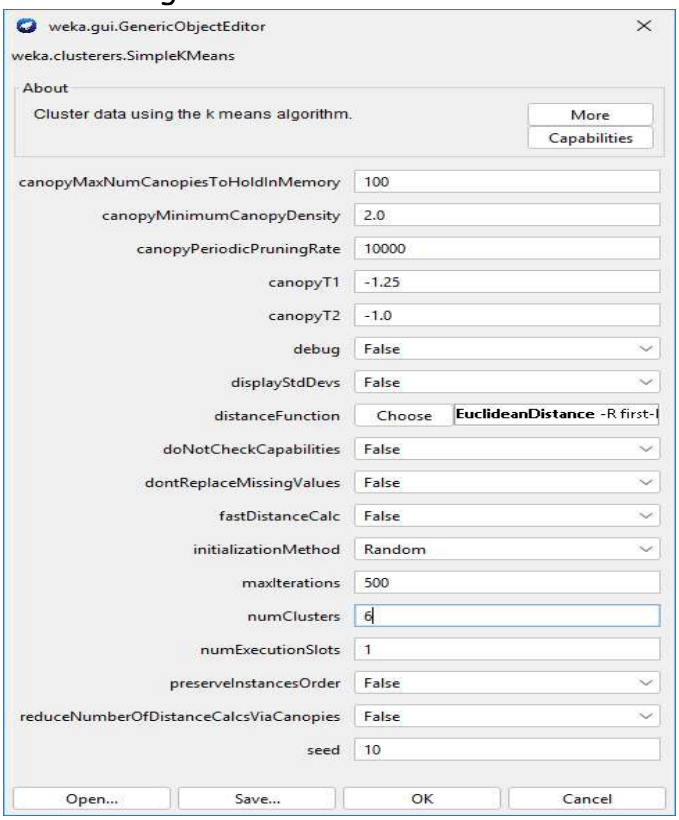
Pre-processed Dataset:

```
@relation 'bank-dataa (1)-
weka.filters.unsupervised.attribute.Remove-R1'

@attribute age numeric
@attribute sex {FEMALE,MALE}
@attribute region {INNER_CITY,TOWN,RURAL,SUBURBAN}
@attribute income numeric
@attribute married {NO,YES}
@attribute children numeric
@attribute car {NO,YES}
@attribute save_act {NO,YES}
@attribute current_act {NO,YES}
@attribute mortgage {NO,YES}
@attribute pep {YES,NO}

@data
48,FEMALE,INNER_CITY,17546,NO,1,NO,NO,NO,NO,YES
40,MALE,TOWN,30085.1,YES,3,YES,NO,YES,YES,NO
51,FEMALE,INNER_CITY,16575.4,YES,0,YES,YES,YES,NO,NO
23,FEMALE,TOWN,20375.4,YES,3,NO,NO,YES,NO,NO
57,FEMALE,RURAL,50576.3,YES,0,NO,YES,NO,NO,NO
57,FEMALE,TOWN,37869.6,YES,2,NO,YES,YES,NO,YES
```

Name: Vishal Rajesh Mahajan BI Lab EXP 07

Class: TE IT A Roll No: 56

Setting the Number of Clusters to 6



K-Means Clustering Result in WEKA:

The configuration for the K-Means clustering scheme is as follows:

"weka.clusterers.SimpleKMeans -init 0 -max-candidates 100 -periodic-pruning 10000 -min-density 2.0 -t1 -1.25 -t2 -1.0 -N 6 -A "weka.core.EuclideanDistance -R first-last" -I 500 -num-slots 1 -S 10"

Name: Vishal Rajesh Mahajan BI Lab EXP 07

Class: TE IT A Roll No: 56

11:08:23 - SimpleKMeans

```
=== Run information ===

Scheme:       weka.clusterers.SimpleKMeans -init 0 -max-candidates 100 -periodic-pruning 10000 -min-dens
Relation:     bank-dataa (1)-weka.filters.unsupervised.attribute.Remove-R1
Instances:    600
Attributes:   11
              age
              sex
              region
              income
              married
              children
              car
              save_act
              current_act
              mortgage
              pep
Test mode:    evaluate on training data


=== Clustering model (full training set) ===


kMeans
======

Number of iterations: 16
Within cluster sum of squared errors: 1360.8718391528164

Initial starting points (random):

Cluster 0: 25,FEMALE,RURAL,14505.3,NO,3,NO,YES,YES,NO,NO
Cluster 1: 61,FEMALE,RURAL,22942.9,YES,2,NO,YES,YES,NO,NO
Cluster 2: 54,FEMALE,INNER_CITY,31095.6,YES,2,NO,NO,YES,NO,YES
Cluster 3: 36,FEMALE,TOWN,26920.8,YES,0,NO,NO,YES,NO,NO
Cluster 4: 42,MALE,INNER_CITY,15499.9,YES,0,YES,NO,YES,YES,YES
Cluster 5: 50,MALE,TOWN,40972.9,NO,2,YES,YES,YES,YES,YES
```

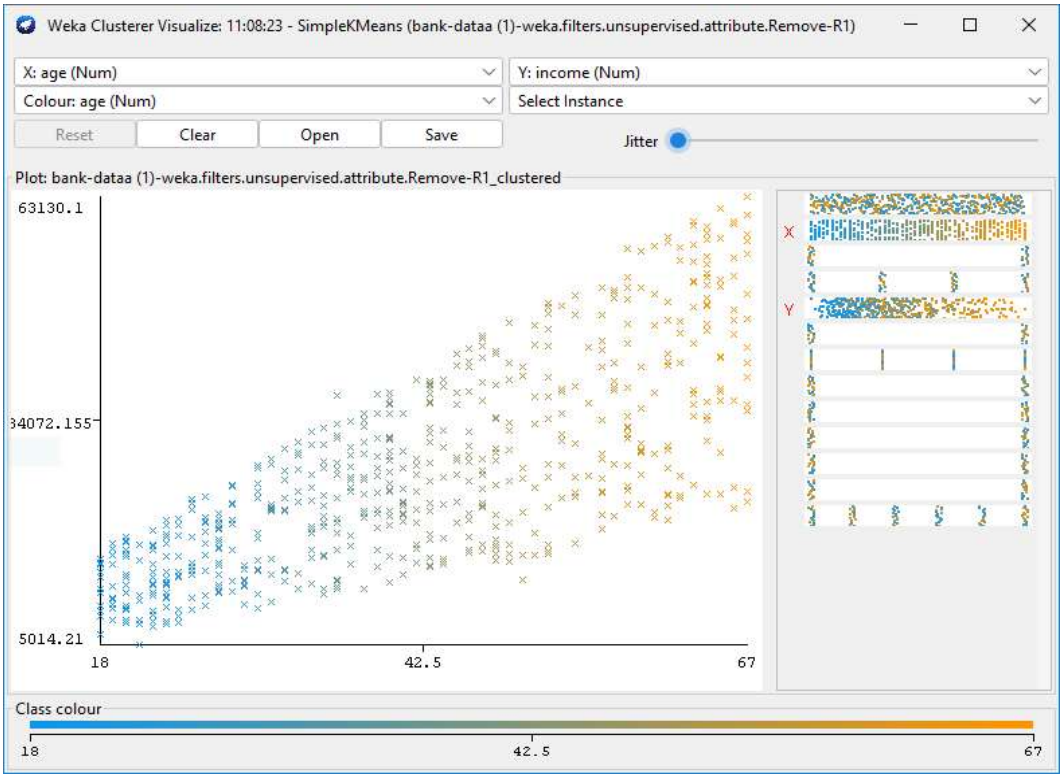
11:08:23 - SimpleKMeans

```
Cluster 5: 50,MALE,TOWN,40972.9,NO,2,YES,YES,YES,YES,YES

Missing values globally replaced with mean/mode

Final cluster centroids:
                        Cluster#
Attribute      Full Data          0          1          2          3          4          5
                 (600.0)     (72.0)    (166.0)     (71.0)     (58.0)     (99.0)    (134.0)
=============================================================================================
age               42.395    43.4444    43.7952    38.7465    37.3103     38.404    47.1791
sex               FEMALE     FEMALE     FEMALE     FEMALE     FEMALE       MALE       MALE
region        INNER_CITY      RURAL INNER_CITY INNER_CITY       TOWN INNER_CITY       TOWN
income        27524.0312  29322.789   28672.09 20239.3776 20600.8528  25720.037 33324.4929
married              YES         NO        YES        YES        YES        YES         NO
children          1.0117     2.0139     0.6265     0.6761     1.6207      0.899     0.9478
car                  NO         NO         NO         NO         NO        YES        YES
save_act            YES        YES        YES         NO         NO         NO        YES
current_act         YES        YES        YES        YES        YES        YES        YES
mortgage             NO         NO         NO         NO         NO        YES         NO
pep                  NO         NO         NO        YES         NO        YES        YES



Time taken to build model (full training data) : 0.05 seconds

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

Clustered Instances

0       72 ( 12%)
1      166 ( 28%)
2       71 ( 12%)
3       58 ( 10%)
4       99 ( 17%)
5      134 ( 22%)
```

The scatter plot represents the relationship between age (on the X-axis) and income (on the Y-axis), with colors differentiating various age groups. In the right panel, additional visualizations display the clustered data. The color bar at the bottom shows the different cluster assignments. The Jitter slider is used to adjust the plot by adding slight randomness, which helps to resolve overlapping data points.

Name: Vishal Rajesh Mahajan BI Lab EXP 07

Class: TE IT A Roll No: 56



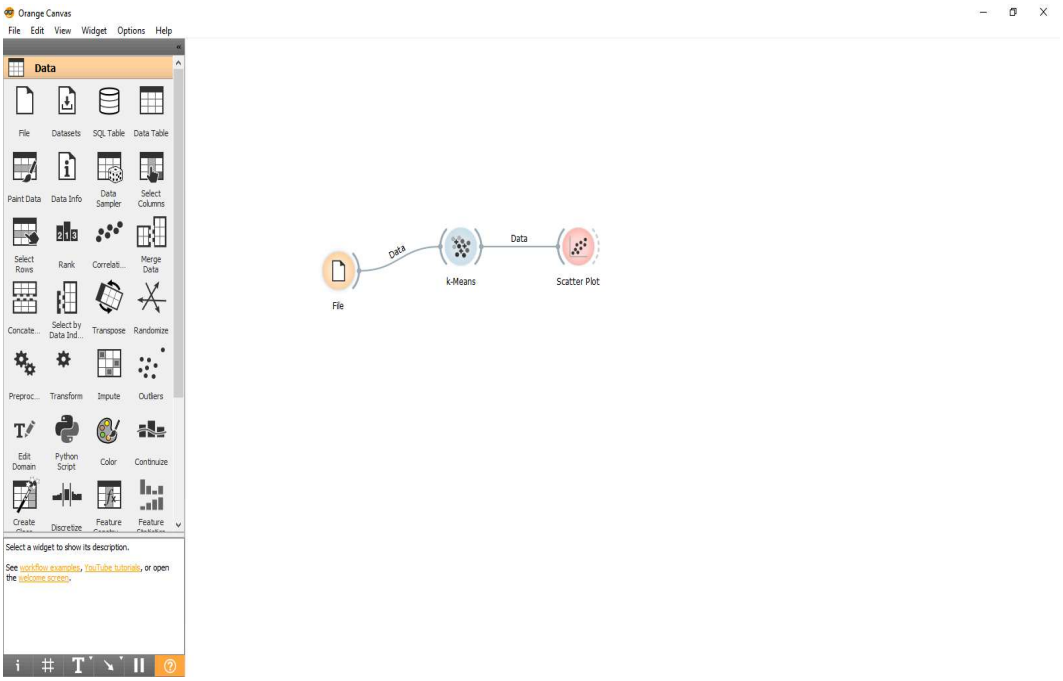
Clustered Dataset:

```
@relation 'bank-dataa (1)-
weka.filters.unsupervised.attribute.Remove-R1_clustered'

@attribute Instance_number numeric
@attribute age numeric
@attribute sex {FEMALE,MALE}
@attribute region {INNER_CITY,TOWN,RURAL,SUBURBAN}
@attribute income numeric
@attribute married {NO,YES}
@attribute children numeric
@attribute car {NO,YES}
@attribute save_act {NO,YES}
@attribute current_act {NO,YES}
@attribute mortgage {NO,YES}
@attribute pep {YES,NO}
@attribute Cluster
{cluster0,cluster1,cluster2,cluster3,cluster4,cluster5}

@data
0,48,FEMALE,INNER_CITY,17546,NO,1,NO,NO,NO,NO,YES,cluster2
1,40,MALE,TOWN,30085.1,YES,3,YES,NO,YES,YES,NO,cluster4
2,51,FEMALE,INNER_CITY,16575.4,YES,0,YES,YES,YES,NO,NO,cluster1
3,23,FEMALE,TOWN,20375.4,YES,3,NO,NO,YES,NO,NO,cluster3
4,57,FEMALE,RURAL,50576.3,YES,0,NO,YES,NO,NO,NO,cluster1
5,57,FEMALE,TOWN,37869.6,YES,2,NO,YES,YES,NO,YES,cluster3
6,22,MALE,RURAL,8877.07,NO,0,NO,NO,YES,NO,YES,cluster2
7,58,MALE,TOWN,24946.6,YES,0,YES,YES,YES,NO,NO,cluster5
8,37,FEMALE,SUBURBAN,25304.3,YES,2,YES,NO,NO,NO,NO,cluster3
9,54,MALE,TOWN,24212.1,YES,2,YES,YES,YES,NO,NO,cluster5
10,66,FEMALE,TOWN,59803.9,YES,0,NO,YES,YES,NO,NO,cluster1
11,52,FEMALE,INNER_CITY,26658.8,NO,0,YES,YES,YES,YES,NO,cluster1
12,44,FEMALE,TOWN,15735.8,YES,1,NO,YES,YES,YES,YES,cluster2
13,66,FEMALE,TOWN,55204.7,YES,1,YES,YES,YES,YES,YES,cluster5
14,36,MALE,RURAL,19474.6,YES,0,NO,YES,YES,YES,NO,cluster1
15,38,FEMALE,INNER_CITY,22342.1,YES,0,YES,YES,YES,YES,NO,cluster
1
```

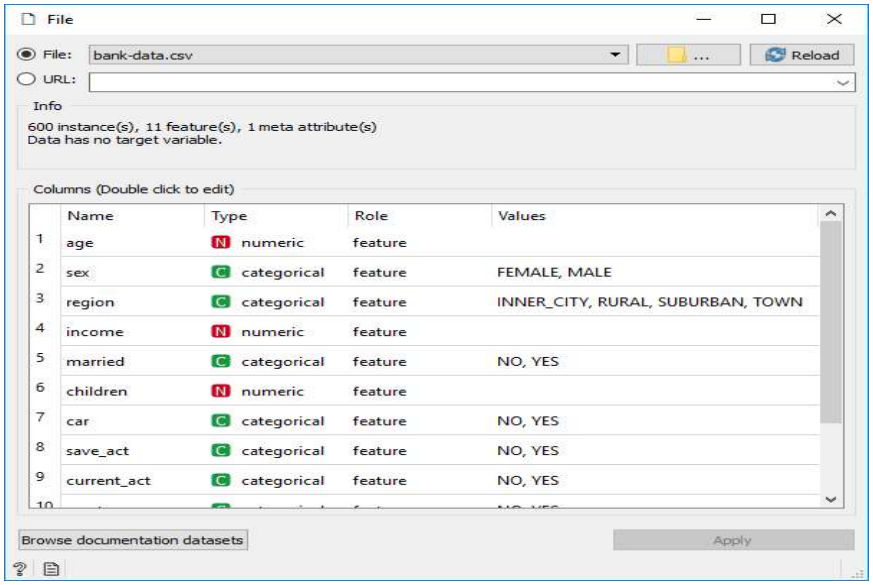
Name: Vishal Rajesh Mahajan BI Lab EXP 07

Class: TE IT A Roll No: 56

K-Means Clustering in Orange

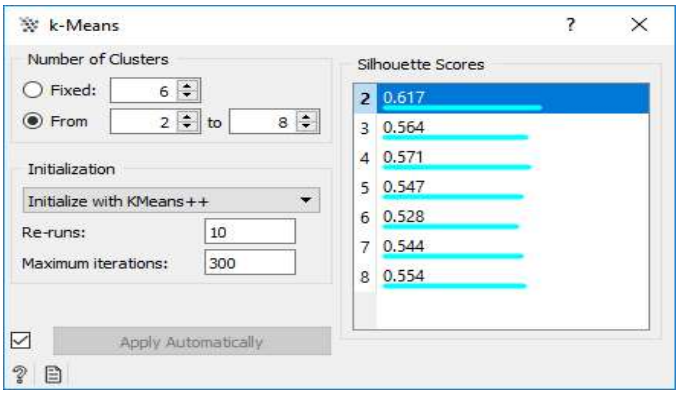


Loading Dataset in ORANGE:

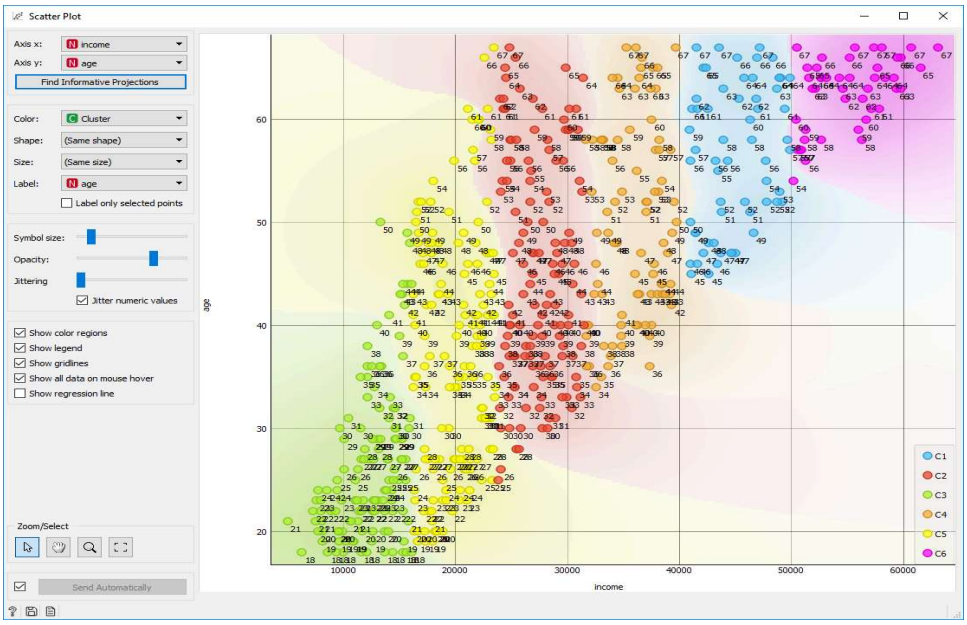


Name: Vishal Rajesh Mahajan BI Lab EXP 07

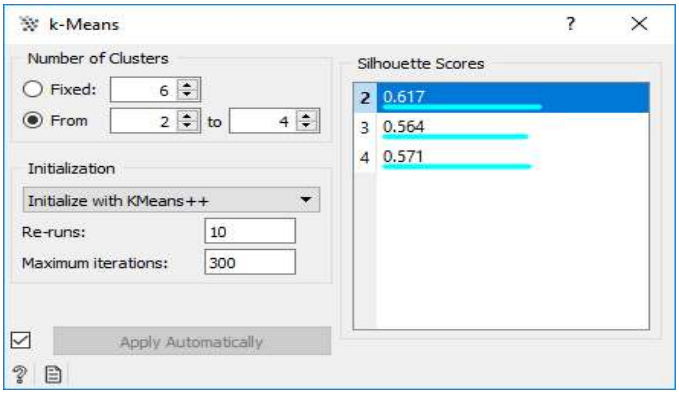
Class: TE IT A Roll No: 56



Graph of Scatter Plot in WEKA:



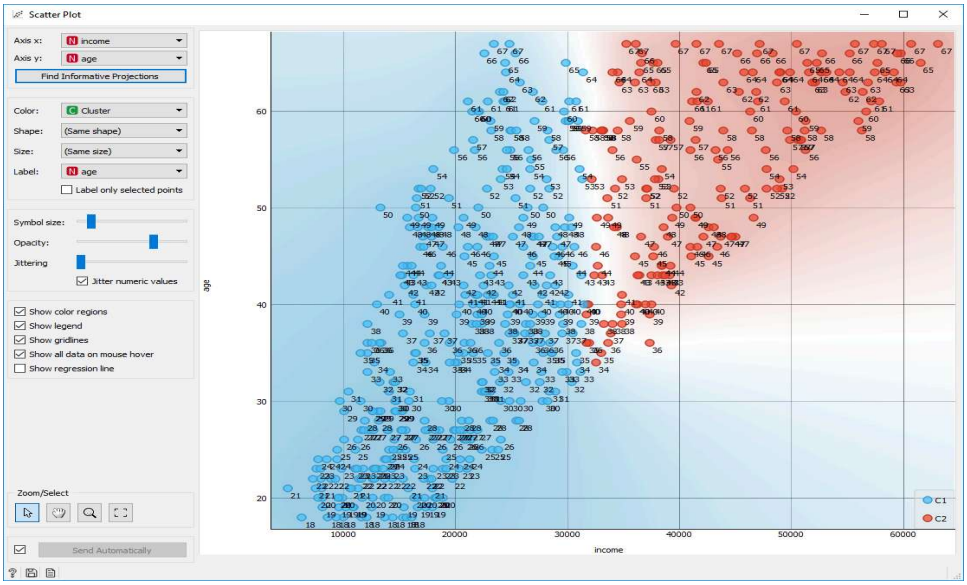
Determining the Number of Clusters Based on Silhouette Score:



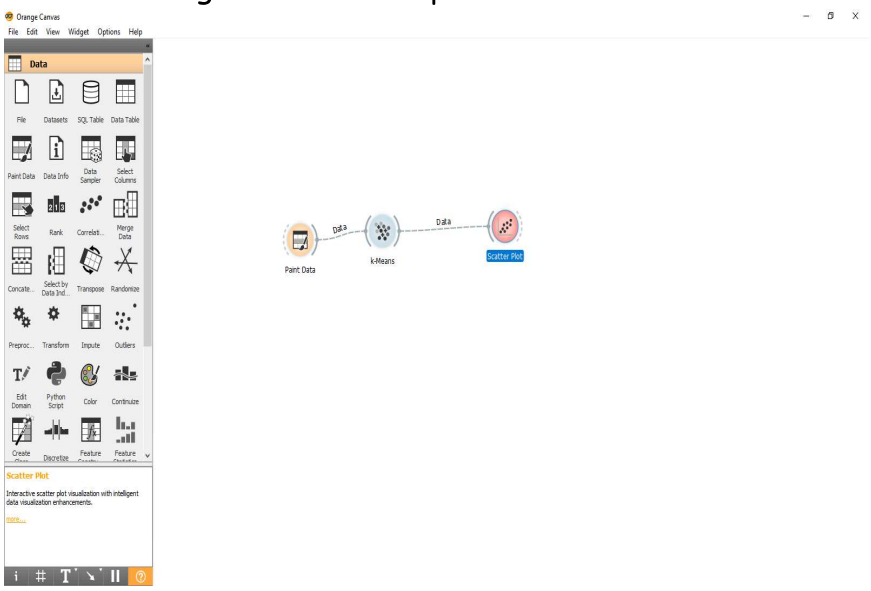
Name: Vishal Rajesh Mahajan BI Lab EXP 07

Class: TE IT A Roll No: 56

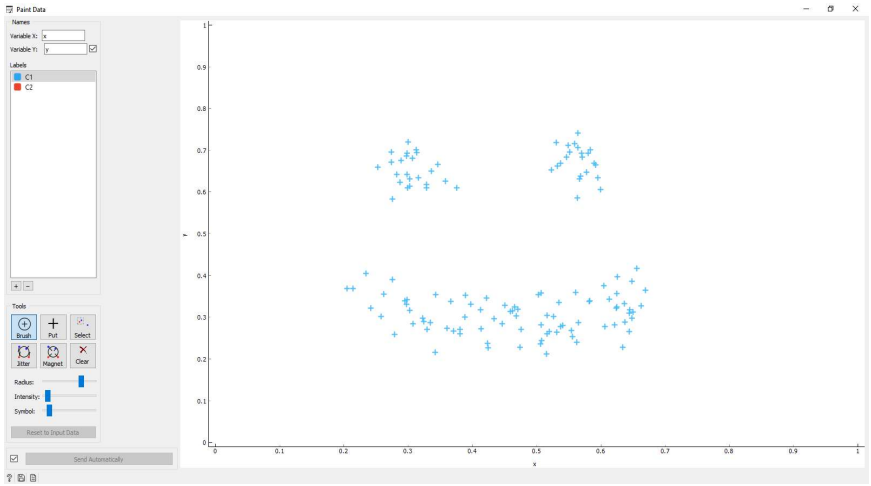
Observing Two Clusters in the Scatter Plot:



Creating a New Data Pipeline for Painted Data:



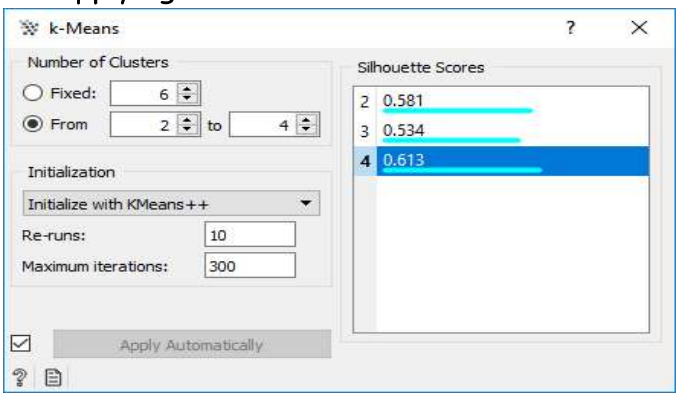
Designing Smiley Shapes in the Painted Data:



Name: Vishal Rajesh Mahajan BI Lab EXP 07

Class: TE IT A Roll No: 56

Applying K-Means on the Painted Data:



Output of Scatter Plot with Clusters: