Laboratório

Clusterização com WEKA Explorer

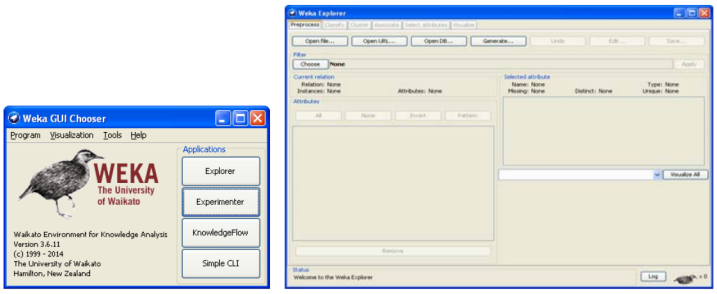
Faça o download dos datasets **car-browsers.arff*** e **iris.arff*** , e execute clusterização conforme as páginas abaixo, onde encontram-se tutoriais.:

*disponível em: www.inf.ufpr.br/menotti/am-201/data.zip

1. Preparando os dados para classificação

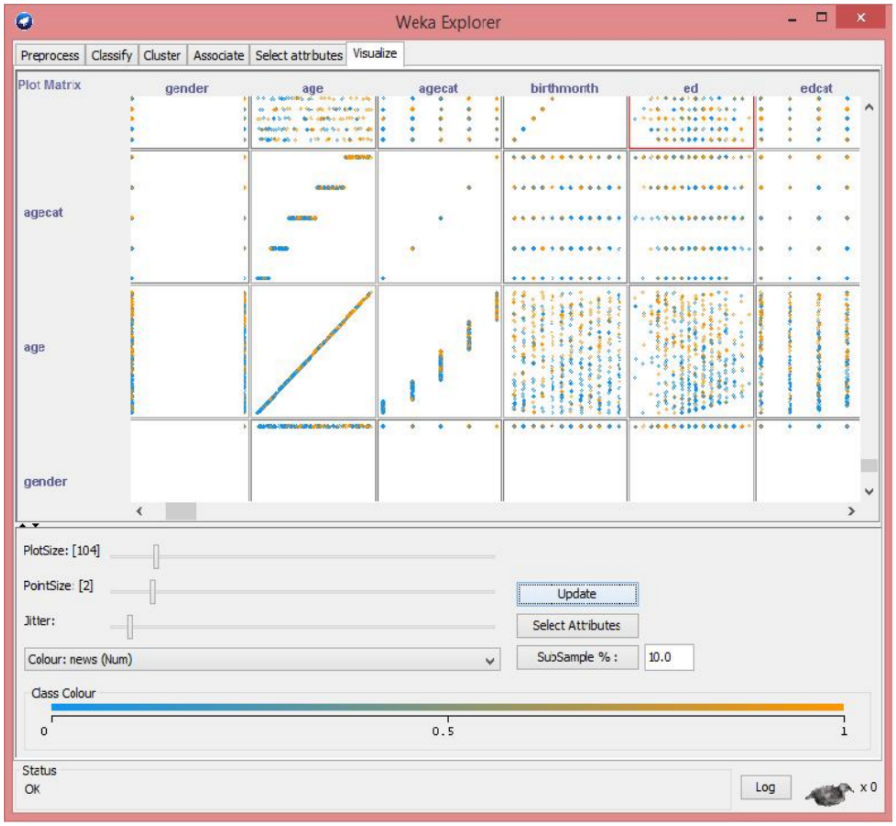
1. Inicie uma sessão do **Weka** ou execute em linha de comando: ***java –jar weka.jar***.

2. Quando a **GUI Chooser** surgir, selecione o **Explorer** a partir das quatro opções do lado direito.



3. Estamos no **Preprocess** agora. Clique no botão **Open** para abrir a caixa de diálogo padrão através da qual você pode selecionar um arquivo. Escolha o arquivo **telco_lab3.csv**.

4. Você pode ignorar atributos irrelevantes durante o processo de **Clustering**, como **custIds**. Para identificar atributos redundantes, poderíamos verificar a correlação a partir da Visualização do dataset na aba **Visualize**. Pode-se ver que os atributos **age** and **agecat** estão correlacionados. Um deles deve ser ignorado. Nós mantemos o atributo **age** para fins de agrupamento; também mantemos **ed** (removendo **edcat**), então temos 8 atributos restantes para **Clustering** (vamos ignorar **custIds**, **agecat** e **edcat** quando realizamos o **Clustering**).

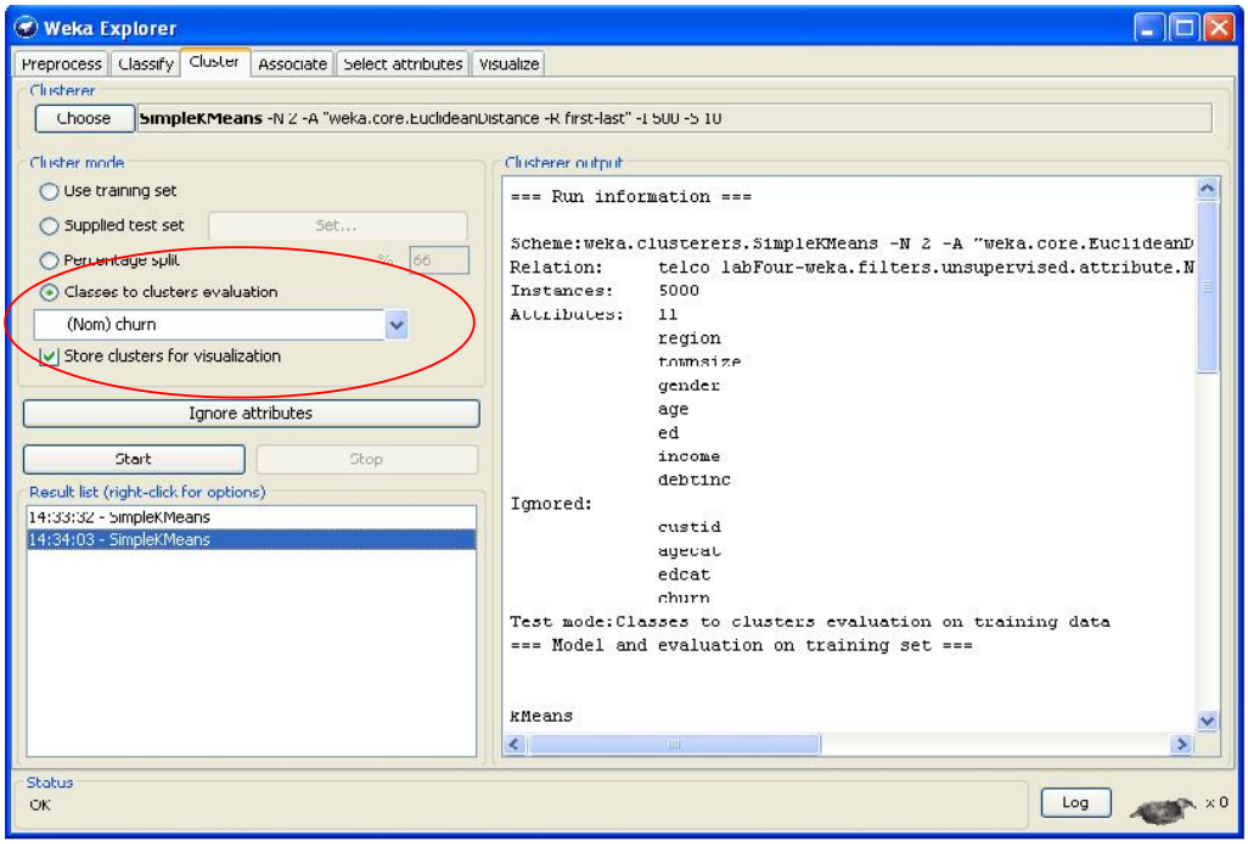


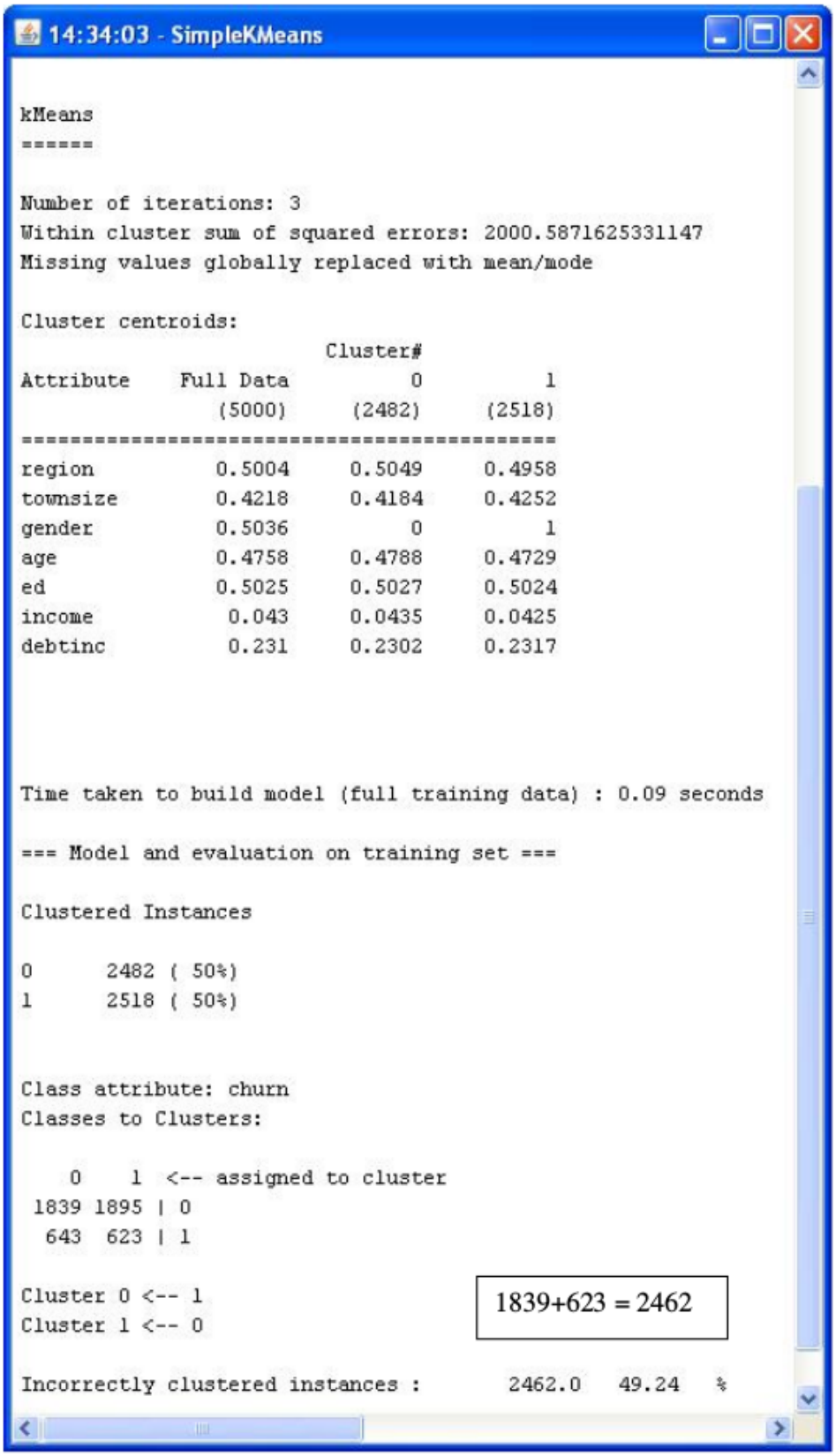
5. Antes de fazer o **Clustering** com Weka, precisamos **normalizar** seus valores de dados

numéricos (use o filtro **Normalize**). Como temos o rótulo de classe, gostaríamos de defini-lo como **nominal** antes da normalização. Esta informação será usada para avaliar o desempenho do clustering.

SimpleKMeans

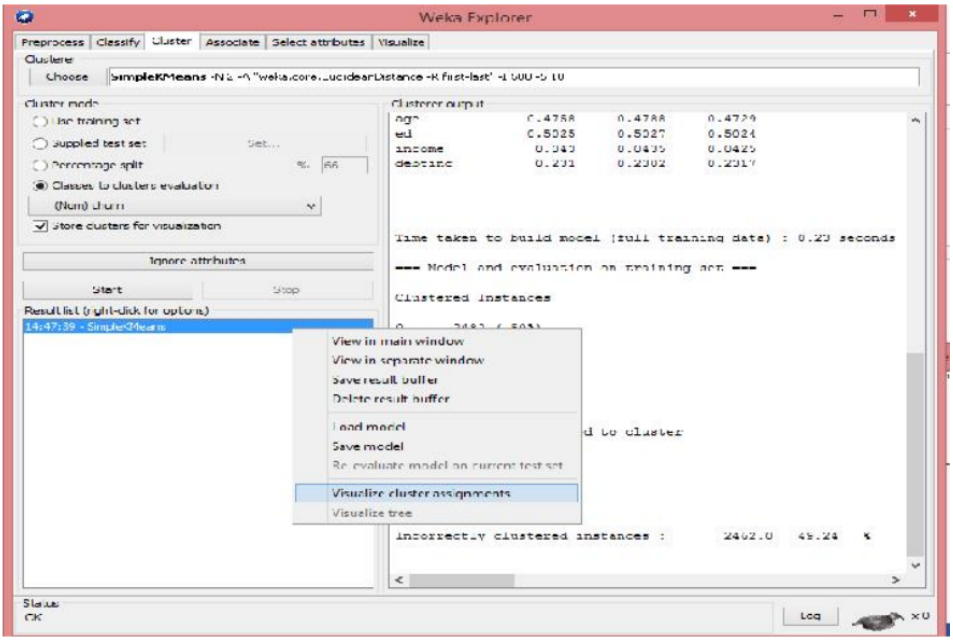
1. Para executar o **Clustering** no dataset dados, clique na guia **Cluster** e escolha o algoritmo **SimpleKMeans**. Definimos ***k = 2*** para este dataset. Clique em ***Classes to clusters evaluation*** e selecione o último atributo (**churn**) como **classe**. Clique em ***Store Clusters for visualization***. Clique em ***Ignore attributes*** e seleciona **custIds**, **agecat**, **edcat** e o último atributo **churn** (classe, que não será clusterizado). Em seguida, clique em ***Start***.



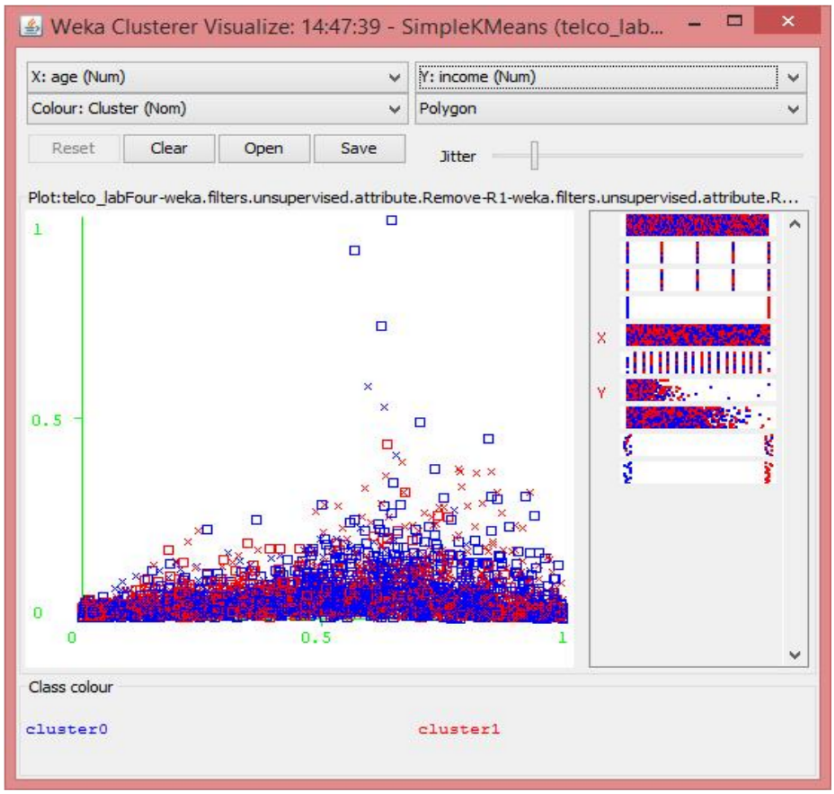


2. Você poderia visualizar os resultados da clusterização clicando com o botão **Direito** na

lista de resultados e escolher **Visualize clusters assignments**. Você pode selecionar uma combinação diferente de dois atributos como ***X*** e ***Y***.

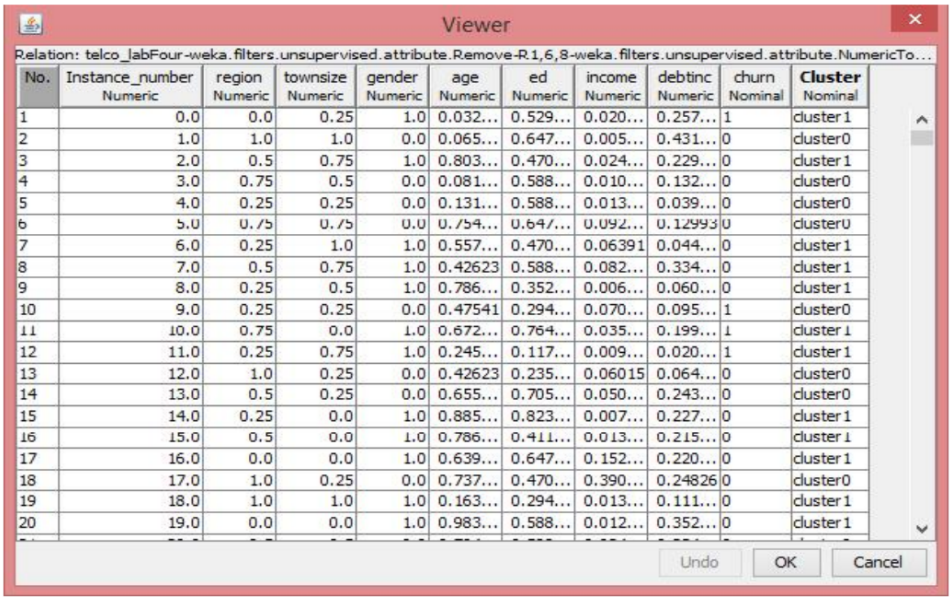


3. Você pode salvar os resultados da clusterização clicando no botão **Save** quando estiver visualizando o resultado da clusterização.

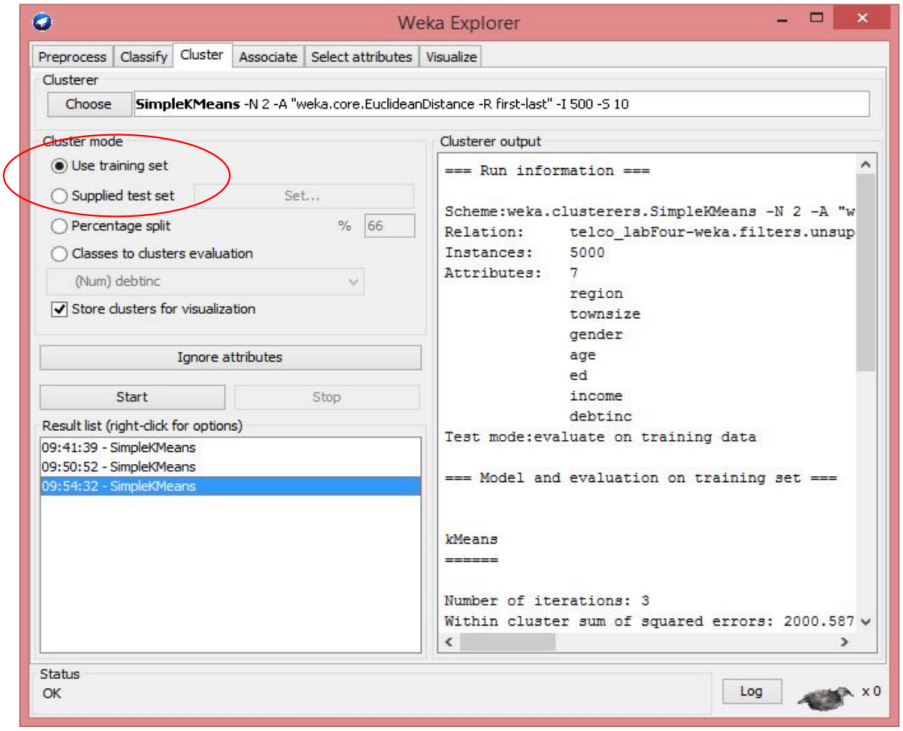


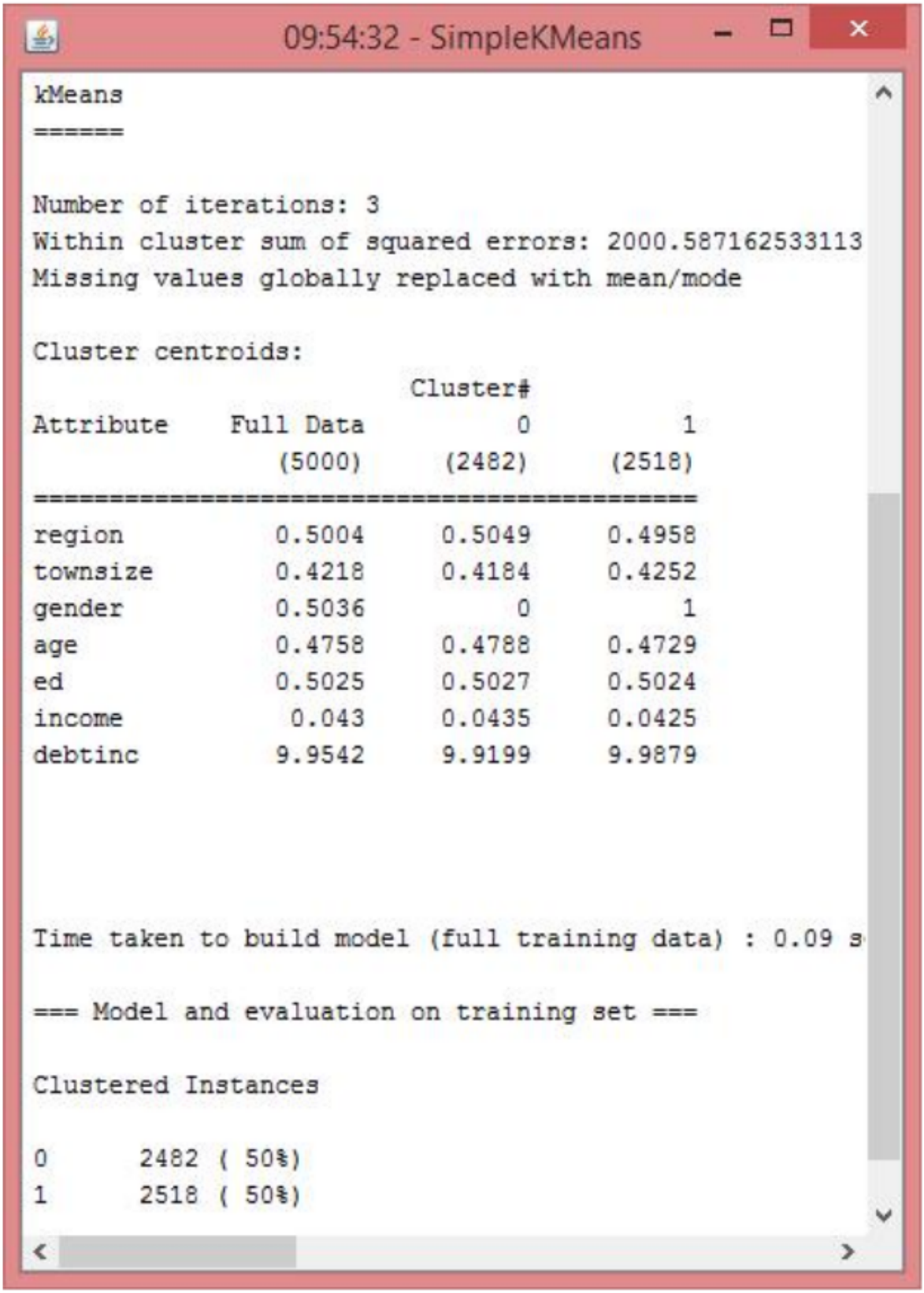
4. Os resultados são salvos em um arquivo **.arff**. Você pode usar o próprio Weka para abrí-lo

e ver os resultados no novo dataset.



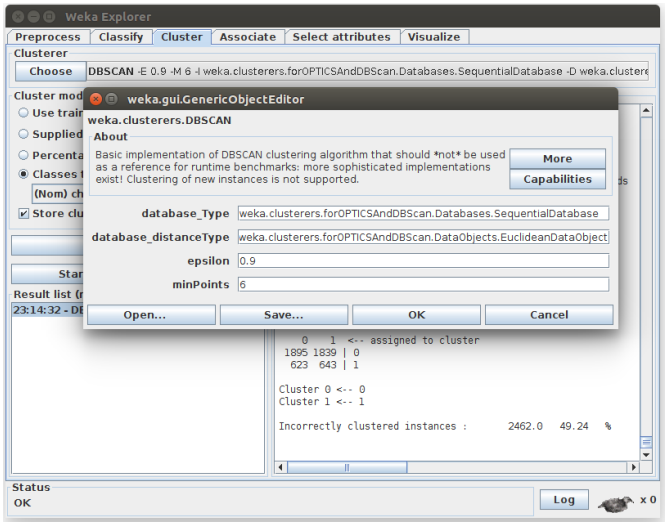
5. Se o dataset não tiver classe definida, quando você realizar a clusterização no dataset, escolha **Use Training Dataset** como Cluster mode.




09:54:32 - SimpleKMeans
kMeans
======
Number of iterations: 3
Within cluster sum of squared errors: 2000.587162533113
Missing values globally replaced with mean/mode
Cluster centroids:
Cluster#
Attribute Full Data 0 1
(5000) (2482) (2518)
region 0.5004 0.5049 0.4958
townsize 0.4218 0.4184 0.4252
gender 0.5036 0 1
age 0.4758 0.4788 0.4729
ed 0.5025 0.5027 0.5024
income 0.043 0.0435 0.0425
debtinc 9.9542 9.9199 9.9879
Time taken to build model (full training data) : 0.09 s
=== Model and evaluation on training set ===
Clustered Instances
0 2482 (50%)
1 2518 (50%)

DBScan

1. Agora vamos abordar o algoritmo **DBScan**. Da mesma forma que fizemos para **SimpleKMeans**, clique em ***Classes to clusters evaluation*** e selecione o último atributo (**churn**) como **classe**. Clique em ***Store Clusters for visualization***. Clique em ***Ignore attributes*** e seleciona **custIds**, **agecat**, **edcat** e o último atributo **churn** (classe, que não será clusterizado). Em seguida, clique em ***Start***. Inicialmente vamos usar os valores default para **epsilon** e **minPoints**.



2. Experimente alterar o valor de **epsilon** para 0.3 e de **minPoints** para 100 e verifique o novo resultado da **Clusterização**.