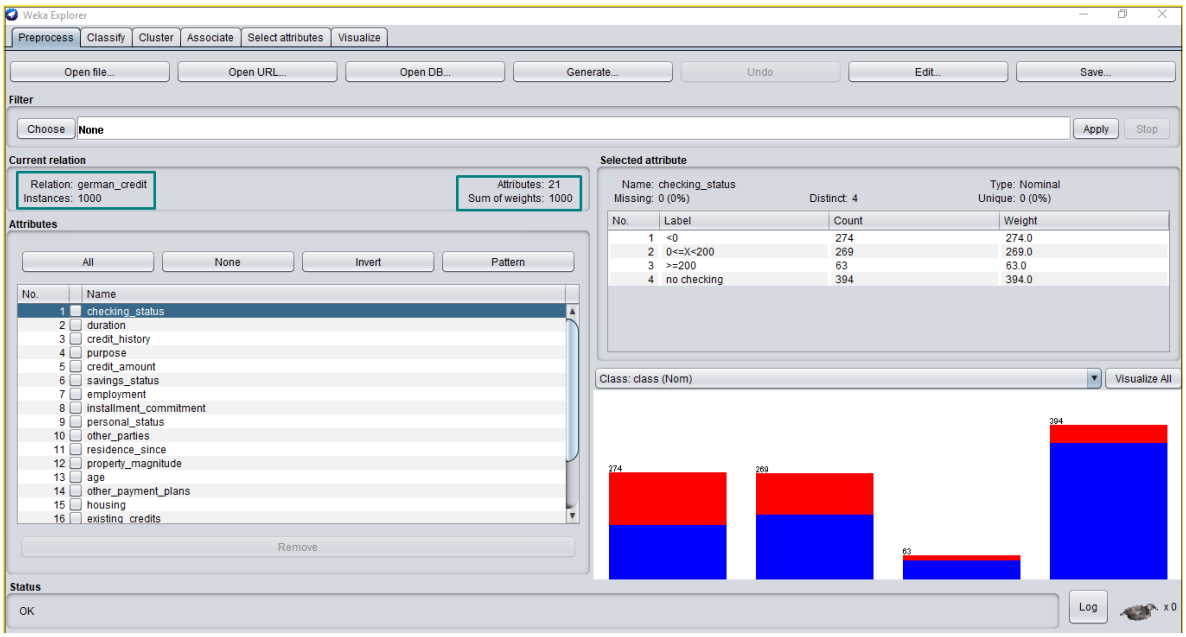
Laura Cristina López Bedoya

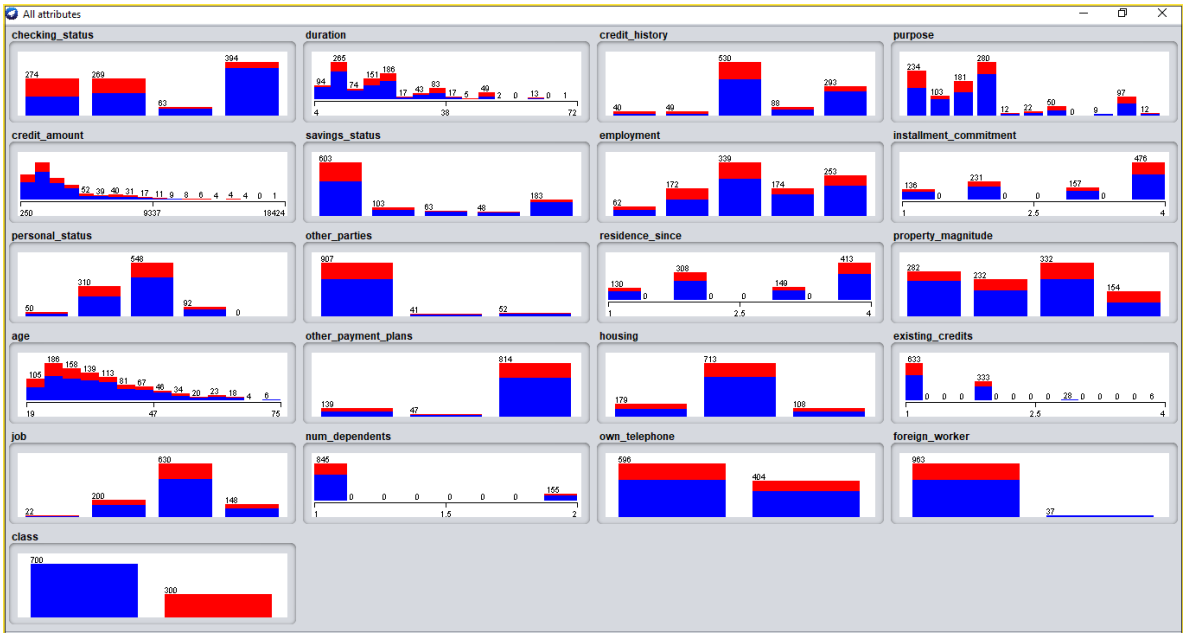
WEKA

PREPARACIÓN DE DATOS

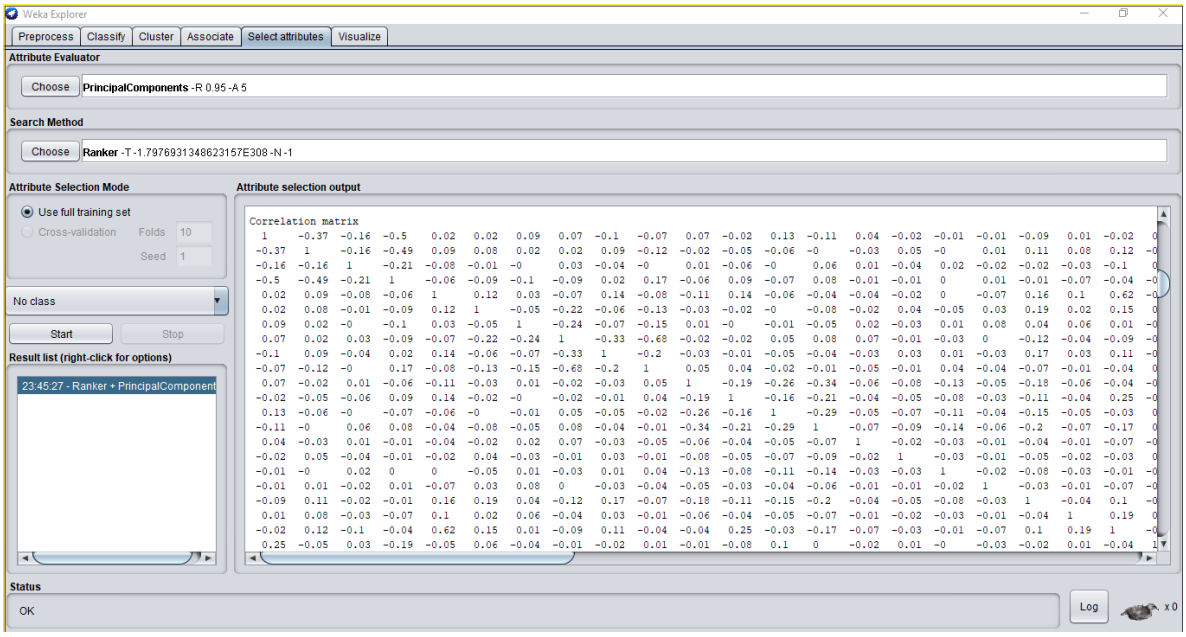
Al cargar el archivo credit-g.arff en Weka se inicia con la preparación y visualización de datos en el programa.

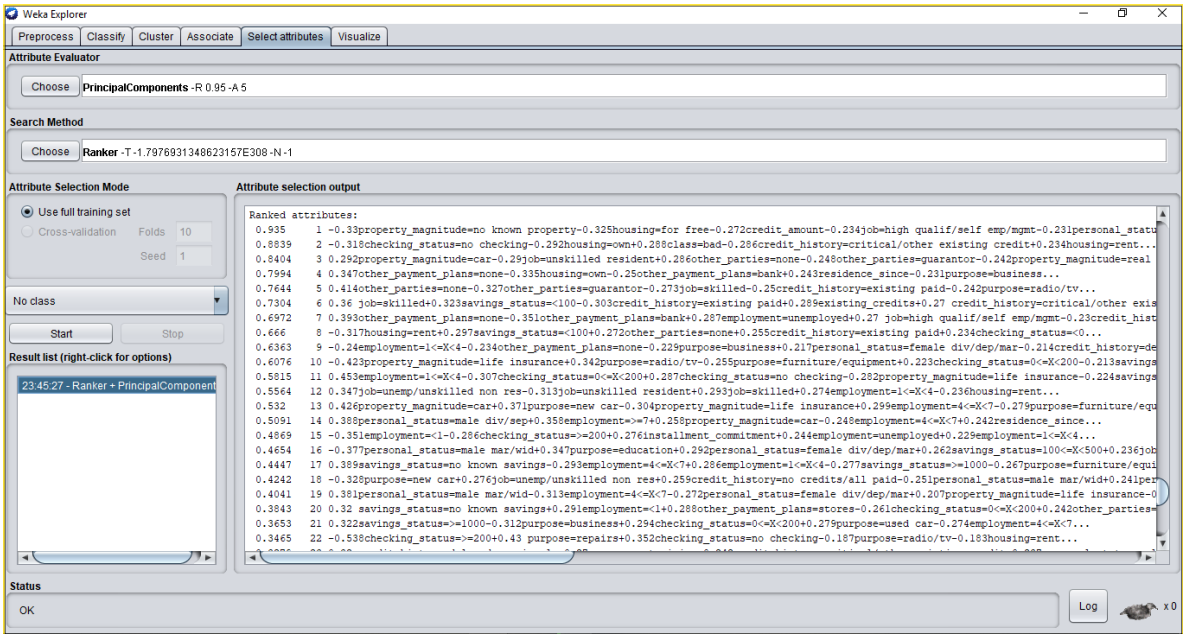
- Visualización de los datos:



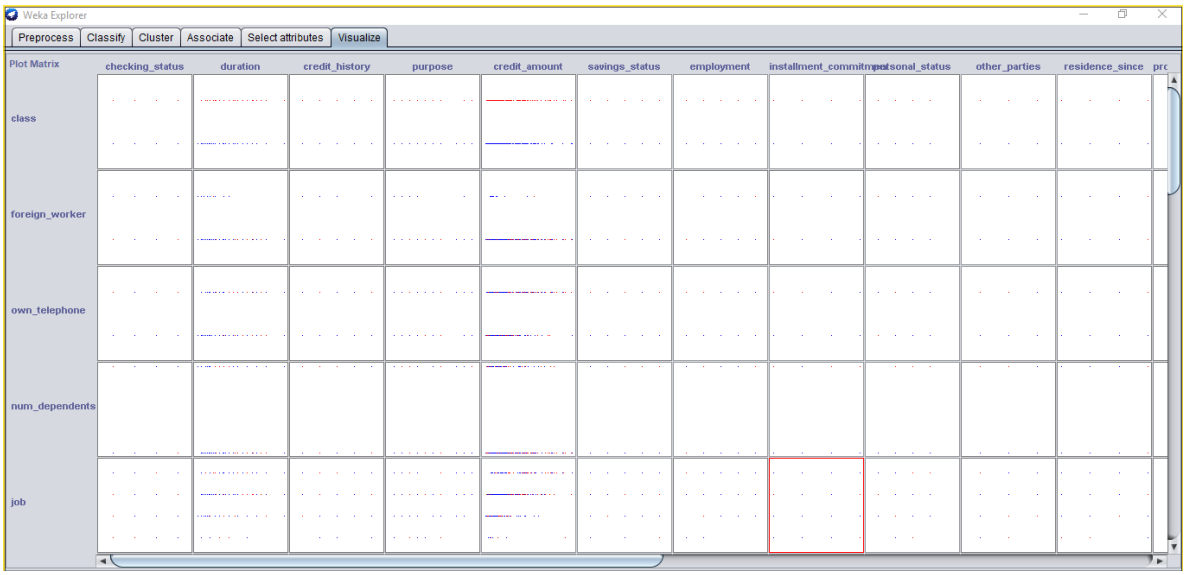



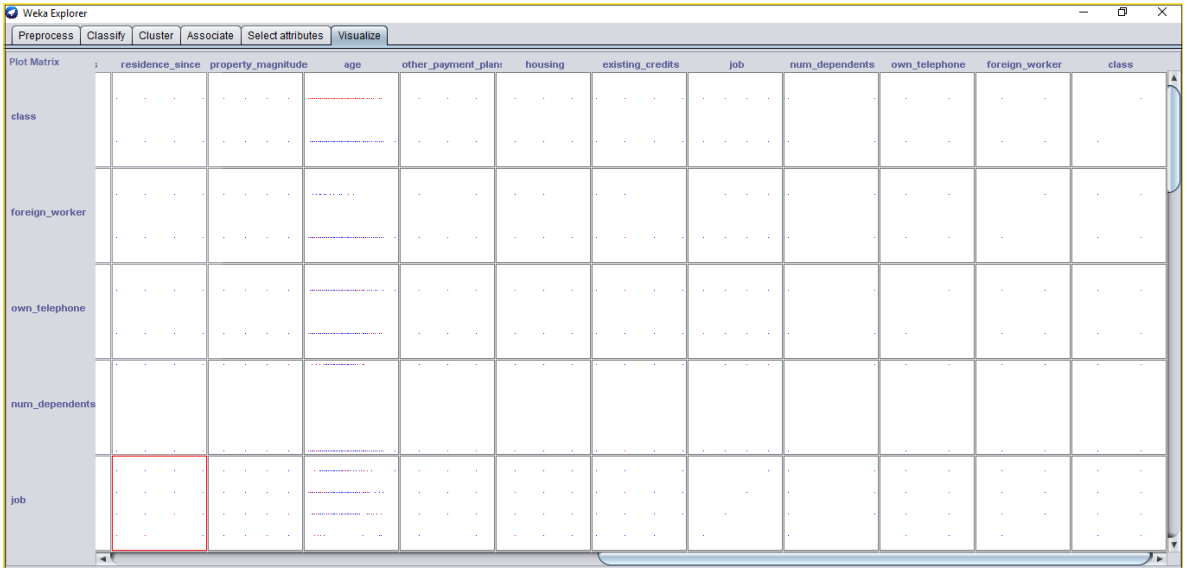
- Matriz de correlación:

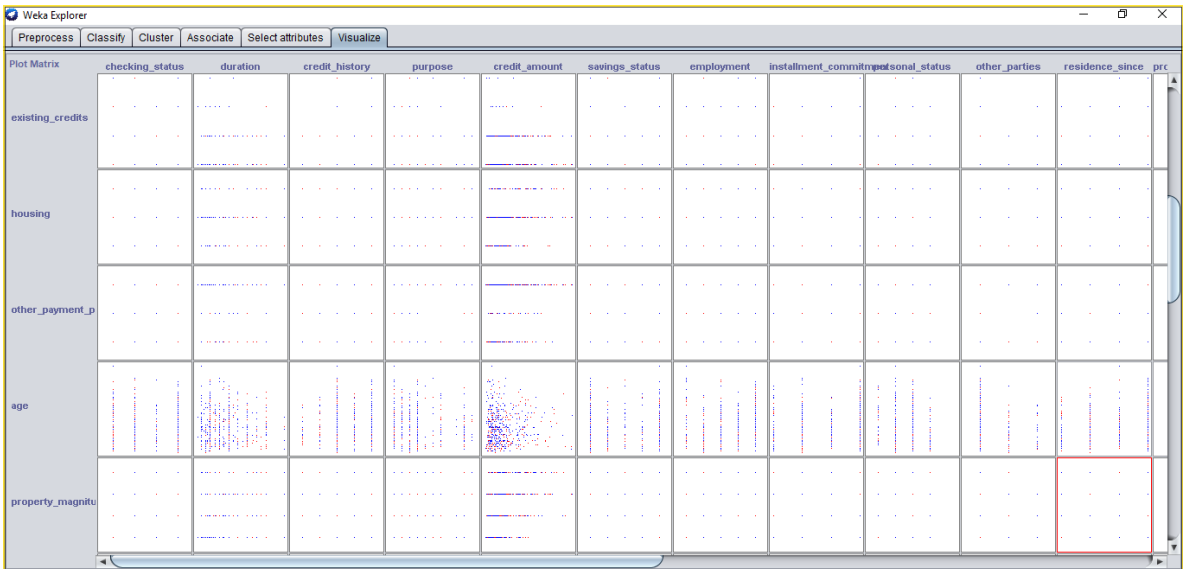


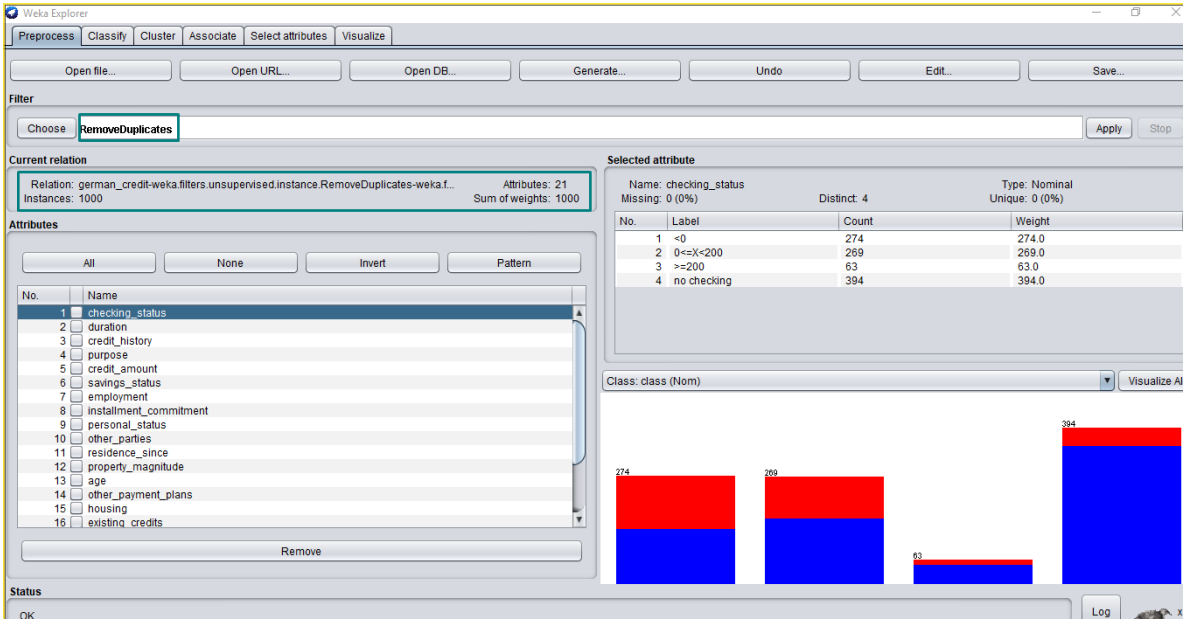
- Visualizar el scatter:



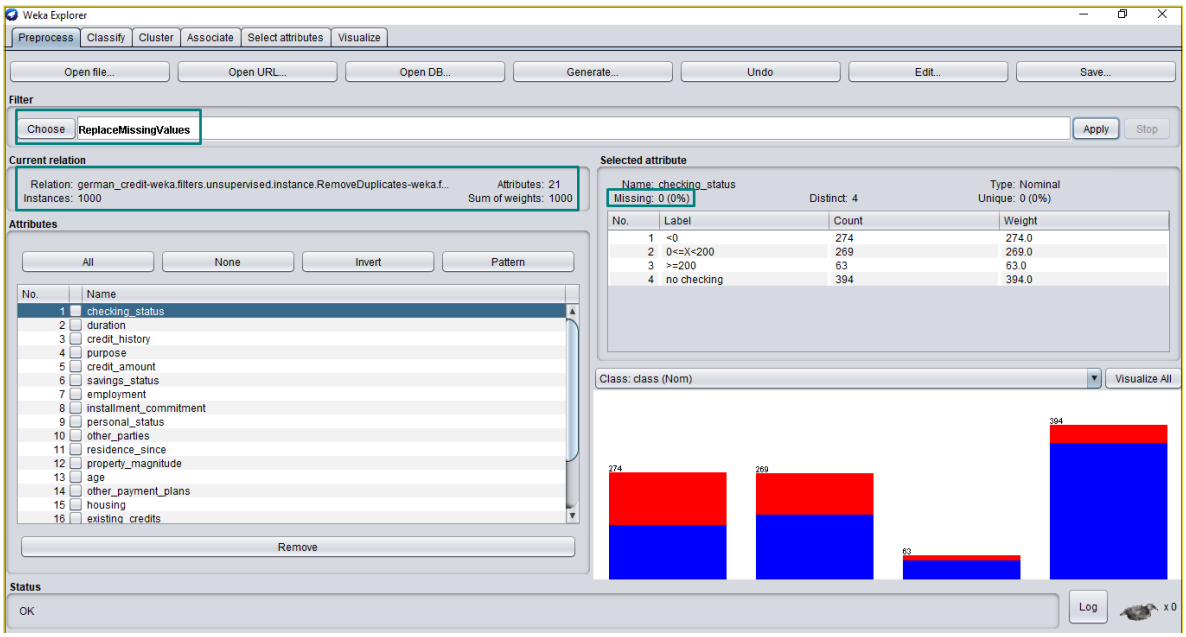





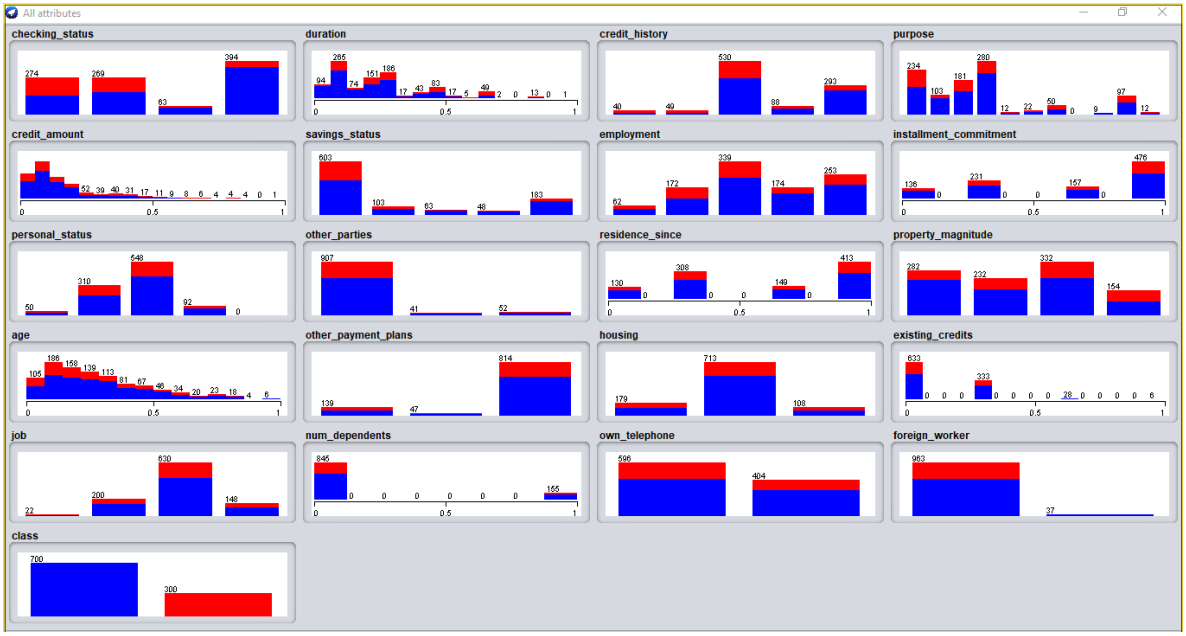

- Remover variables duplicadas: No hay variables duplicadas.



- Remover valores faltantes: El dataset no tiene valores nulos.




- Normalizar las variables para eliminar datos atípicos:

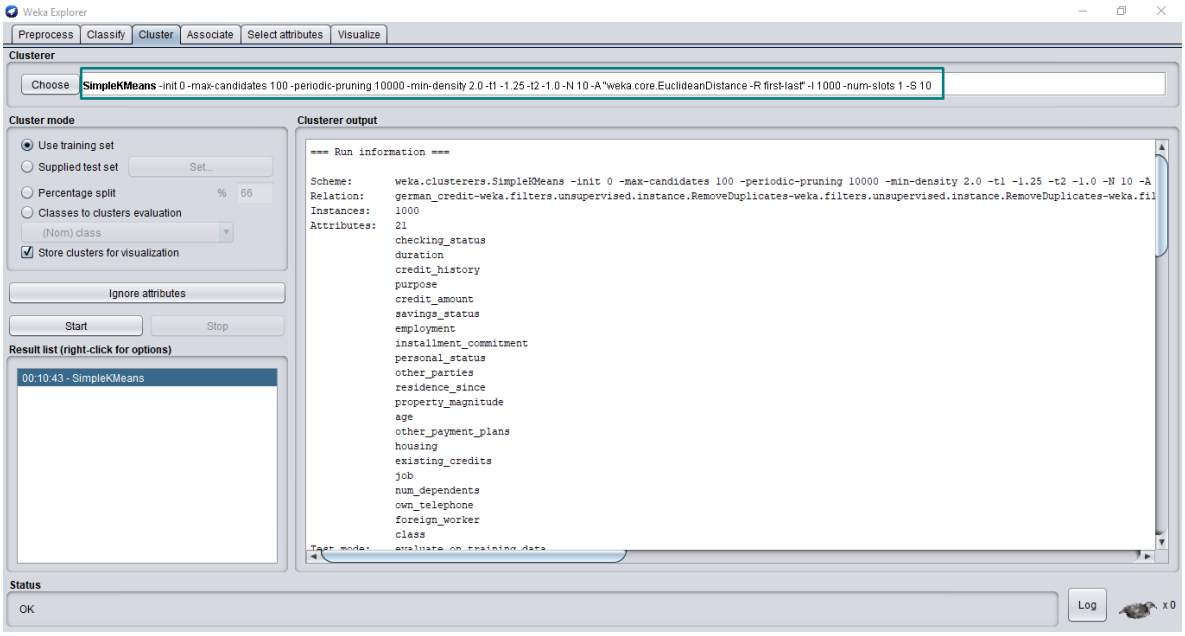


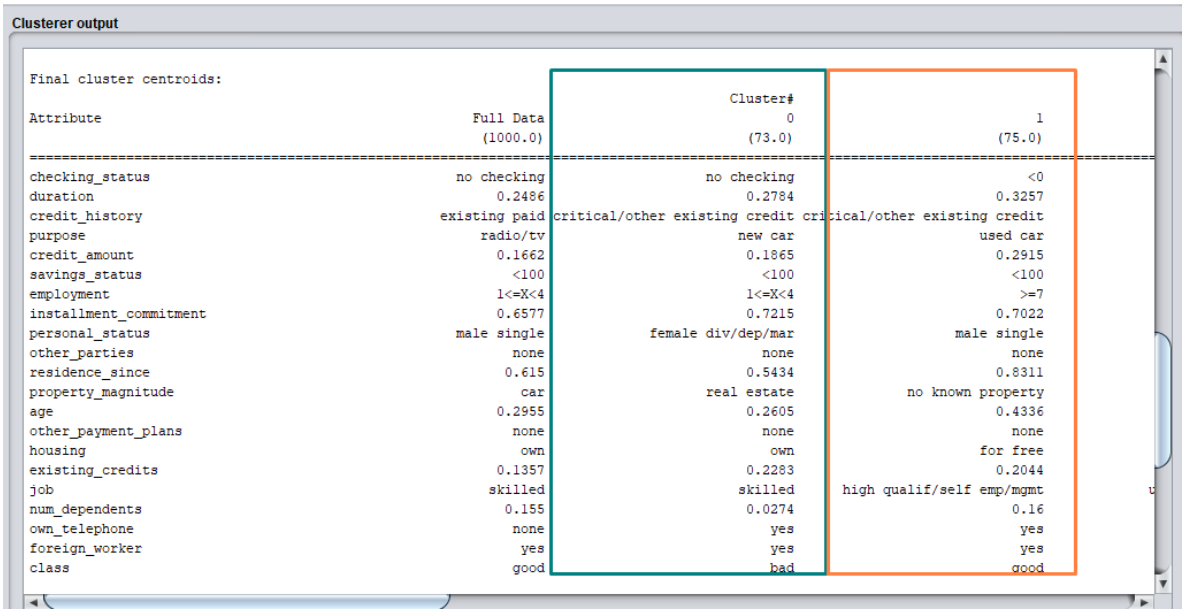
MÉTODOS NO SUPERVISADOS

Se eligen 10 clusters teniendo en cuenta los propósitos por los cuales el usuario solicita el crédito.

- Método Particional:

  K-Means:



Clusterer output

| 2 | 3 | 4 | 5 | |
|---|---|---|---|---|
| (38.0) | (92.0) | (121.0) | (154.0) | |
| 0<=X<200 | 0<=X<200 | <0 | 0<=X<200 | no |
| 0.18 | 0.2206 | 0.2737 | 0.2161 | |
| existing paid | existing paid | existing paid | existing paid | exist |
| radio/tv | radio/tv | new car | radio/tv | furniture/e |
| 0.0751 | 0.1348 | 0.1559 | 0.1182 | |
| <100 | no known savings | <100 | <100 | |
| >=7 | >=7 | <1 | 1<=X<4 | |
| 0.8246 | 0.7717 | 0.6474 | 0.6017 | |
| male mar/wid | male single | female div/dep/mar | female div/dep/mar | mal |
| none | none | none | none | |
| 0.6754 | 0.7645 | 0.7355 | 0.3701 | |
| real estate | life insurance | car | real estate | life i |
| 0.3186 | 0.3651 | 0.1964 | 0.2043 | |
| bank | none | none | none | |
| own | own | rent | own | |
| 0.0877 | 0.1522 | 0.0964 | 0.0563 | |
| unskilled resident | skilled | skilled | skilled | |
| 0.0263 | 0.1196 | 0.0579 | 0.039 | |
| none | none | none | none | |
| yes | yes | yes | yes | |
| good | good | bad | good | |

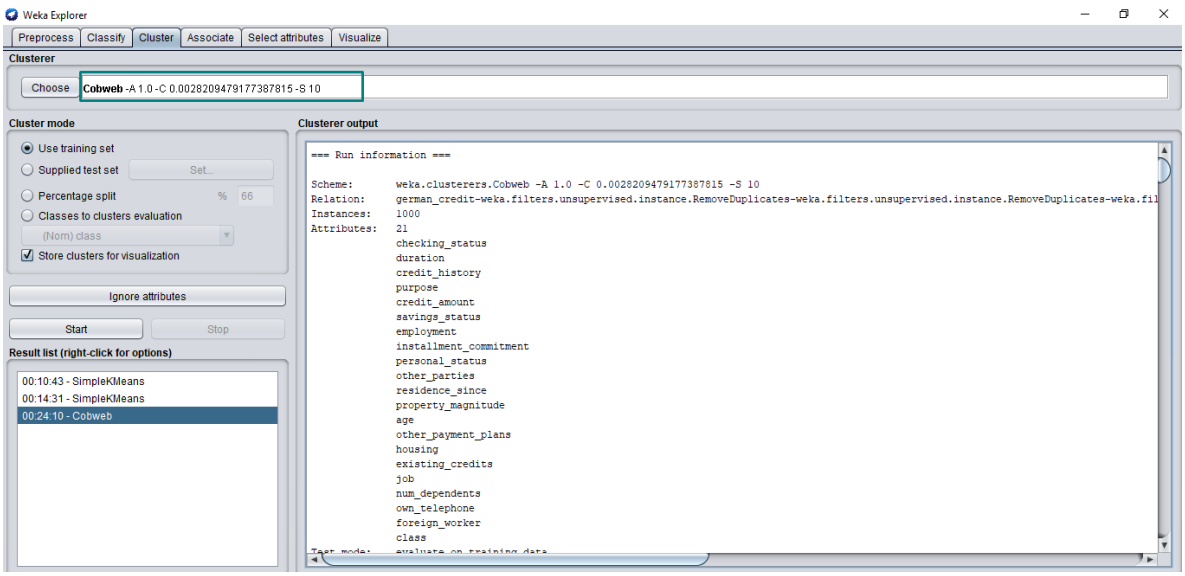
Clusterer output

| 5 | 6 | 7 | 8 | 9 |
|---|---|---|---|---|
| (154.0) | (134.0) | (98.0) | (170.0) | (45.0) |
| 0<=X<200 | no checking | no checking | no checking | no checking |
| 0.2161 | 0.2194 | 0.1567 | 0.3012 | 0.3186 |
| existing paid | existing paid | critical/other existing credit | existing paid | existing paid |
| radio/tv | furniture/equipment | new car | radio/tv | new car |
| 0.1182 | 0.1631 | 0.1272 | 0.1935 | 0.2497 |
| <100 | <100 | <100 | <100 | <100 |
| 1<=X<4 | 4<=X<7 | 1<=X<4 | >=7 | >=7 |
| 0.6017 | 0.5249 | 0.5476 | 0.7392 | 0.6519 |
| male div/dep/mar | male single | male single | male single | male single |
| none | none | none | none | none |
| 0.3701 | 0.4975 | 0.5374 | 0.6471 | 0.9259 |
| real estate | life insurance | real estate | car | no known property |
| 0.2043 | 0.2601 | 0.3519 | 0.3097 | 0.4671 |
| none | none | none | none | bank |
| own | own | own | own | for free |
| 0.0563 | 0.0995 | 0.2177 | 0.149 | 0.1333 |
| skilled | skilled | unskilled resident | skilled | skilled |
| 0.039 | 0.1119 | 0.5408 | 0.1706 | 0.4222 |
| none | none | none | yes | yes |
| yes | yes | yes | yes | yes |
| good | good | good | good | good |

```
Time taken to build model (full training data) : 0.1 seconds

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

Clustered Instances

0       73 (  7%)
1       75 (  8%)
2       38 (  4%)
3       92 (  9%)
4      121 ( 12%)
5      154 ( 15%)
6      134 ( 13%)
7       98 ( 10%)
8      170 ( 17%)
9       45 (  5%)
```


- Métodos Jerárquicos:

COBWEB

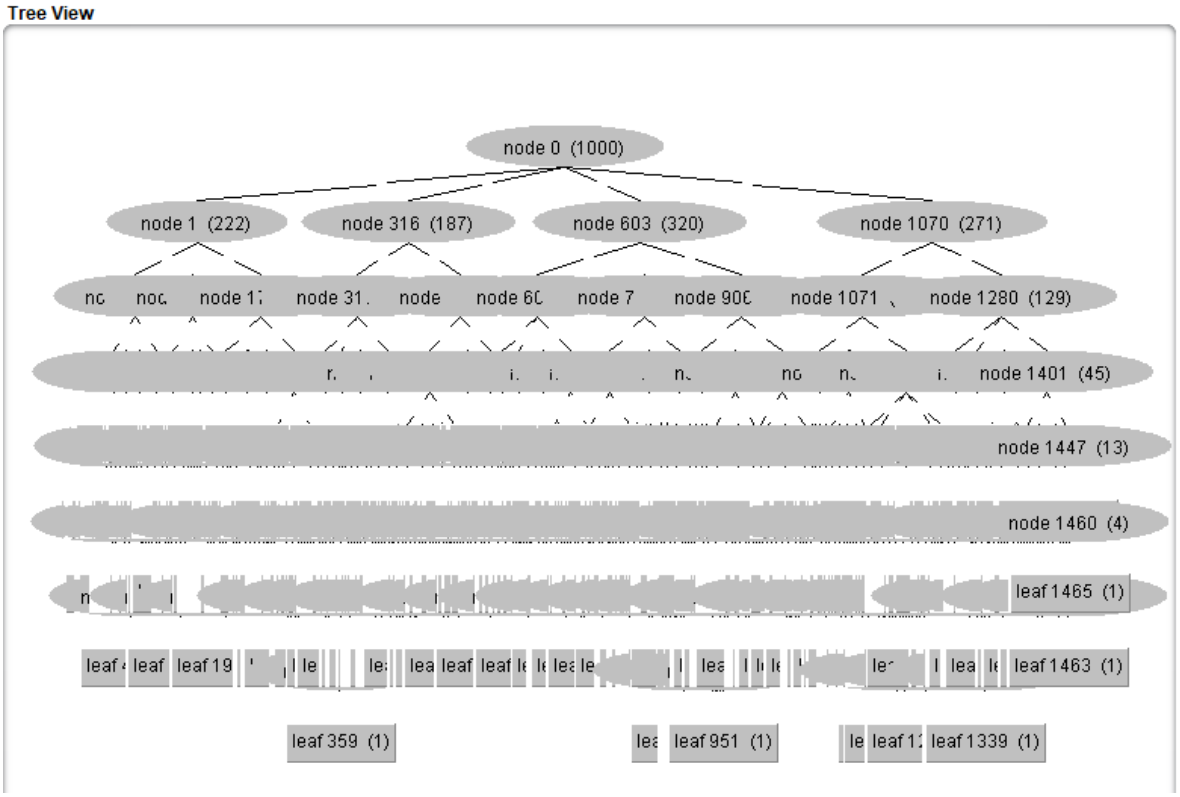


**Clusterer output**

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

Number of merges: 348
Number of splits: 272
Number of clusters: 1466

node 0 [1000]
|   node 1 [222]
|   |   node 2 [61]
|   |   |   node 3 [15]
|   |   |   |   node 4 [2]
|   |   |   |   |   leaf 5 [1]
|   |   |   |   node 4 [2]
|   |   |   |   |   leaf 6 [1]
|   |   |   node 3 [15]
|   |   |   |   node 7 [6]
|   |   |   |   |   leaf 8 [1]
|   |   |   |   node 7 [6]
|   |   |   |   |   leaf 9 [1]
|   |   |   |   node 7 [6]
|   |   |   |   |   leaf 10 [1]
|   |   |   |   node 7 [6]
|   |   |   |   |   node 11 [2]
|   |   |   |   |   |   leaf 12 [1]
|   |   |   |   |   node 11 [2]
|   |   |   |   |   |   leaf 13 [1]
|   |   |   |   node 7 [6]
```

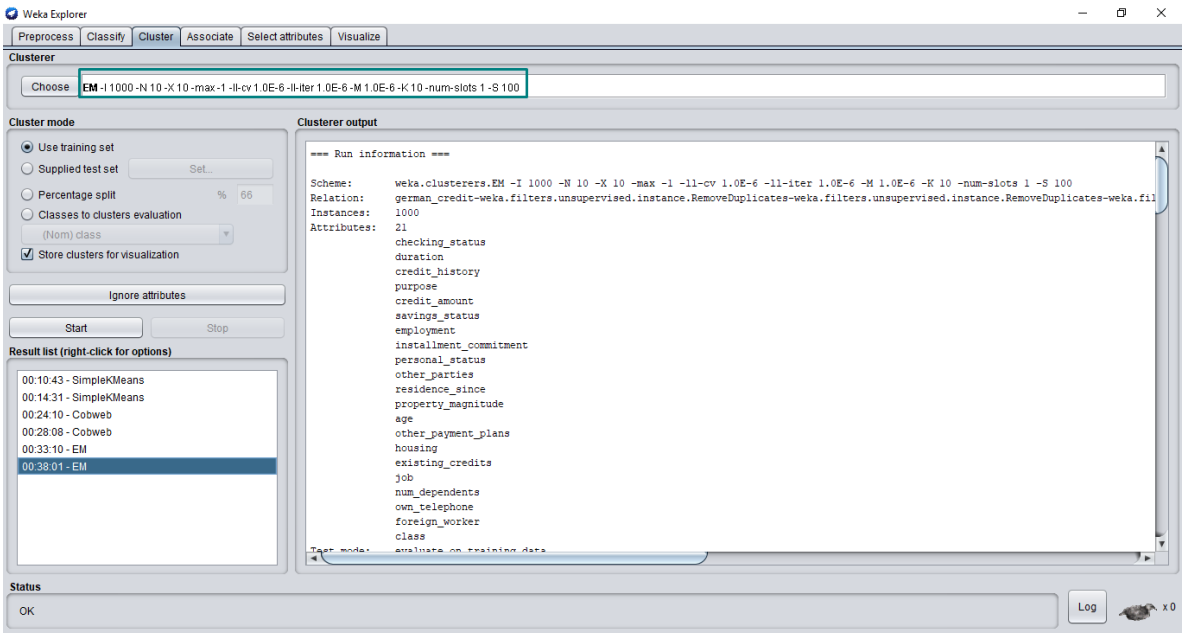
**Clusterer output**

```
Time taken to build model (full training data) : 0.52 seconds

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

Clustered Instances

   4         1 (  0%)
   6         1 (  0%)
  10         1 (  0%)
  11         1 (  0%)
  12         1 (  0%)
  13         1 (  0%)
  16         1 (  0%)
  17         1 (  0%)
  18         1 (  0%)
  20         1 (  0%)
  21         2 (  0%)
  22         1 (  0%)
  24         1 (  0%)
  26         1 (  0%)
  27         3 (  0%)
  29         2 (  0%)
  30         1 (  0%)
  33         1 (  0%)
  35         2 (  0%)
  37         1 (  0%)
  38         1 (  0%)
```




- Método probabilístico:

EM



Clusterer output

```
Number of clusters: 10
Number of iterations performed: 4


                                  Cluster
Attribute                              0        1        2        3        4        5        6        7        8        9
                                  (0.12)    (0.1)   (0.06)   (0.04)   (0.09)   (0.04)   (0.05)    (0.3)   (0.08)    (0.1)
=============================================================================================================================
checking_status
  <0                               8.6472  22.1285  18.3224  15.6245  41.9421   7.6323    3.072   93.619  41.1783  31.8339
  0<=X<200                        22.1919  20.7996  34.6178   5.0409  22.0223  18.9024  12.5008  84.7057  18.0826  40.1358
  >=200                            7.9662   7.3463   1.3021   5.4971   9.5287   1.6201   1.5723  27.9934   6.4408    3.733
  no checking                     89.9761  53.2622  12.5589   22.057  23.4958  20.3138  39.1081  95.3499  15.3516  32.5266
  [total]                        128.7813 103.5365  66.8012  48.2196  96.9889  48.4687  56.2532  301.668  81.0533 108.2294
duration
  mean                             0.2434   0.1683   0.5145   0.0796   0.2586   0.3781   0.3723   0.1513   0.3645   0.3088
  std. dev.                        0.1197   0.0971   0.1879    0.065   0.1443   0.2284   0.1754   0.0863   0.2229   0.1467

credit_history
  no credits/all paid              2.8854   2.0257   7.1165   1.2978   2.2863   1.3632   2.6905   4.9434    8.226  17.1651
  all paid                         1.1225   1.8993   2.5471   5.0062   5.1398   4.4907   2.4908  13.5506   9.2691  13.4839
  existing paid                   49.6467  17.4057  36.5293   7.7913  83.0332  29.4472  13.0324 253.4463  32.0009   17.667
  delayed previously              11.6817   5.5178   7.6255   4.0282   3.3809   5.6176  13.7585   9.6916   8.9887  27.7093
  critical/other existing credit   64.445  77.6879  13.9828  31.0961   4.1486   8.5499  25.2809  21.0361  23.5685  33.2041
  [total]                        129.7813 104.5365  67.8012  49.2196  97.9889  49.4687  57.2532  302.668  82.0533 109.2294
purpose
  new car                         18.4621  26.5781   14.228   29.577  24.4609  16.7872   8.1512  67.3539  18.1014  20.3002
  used car                        22.0358    3.898   9.4013   1.1547   5.6329  13.3362  17.5575  11.3856  20.3317   8.2662
```

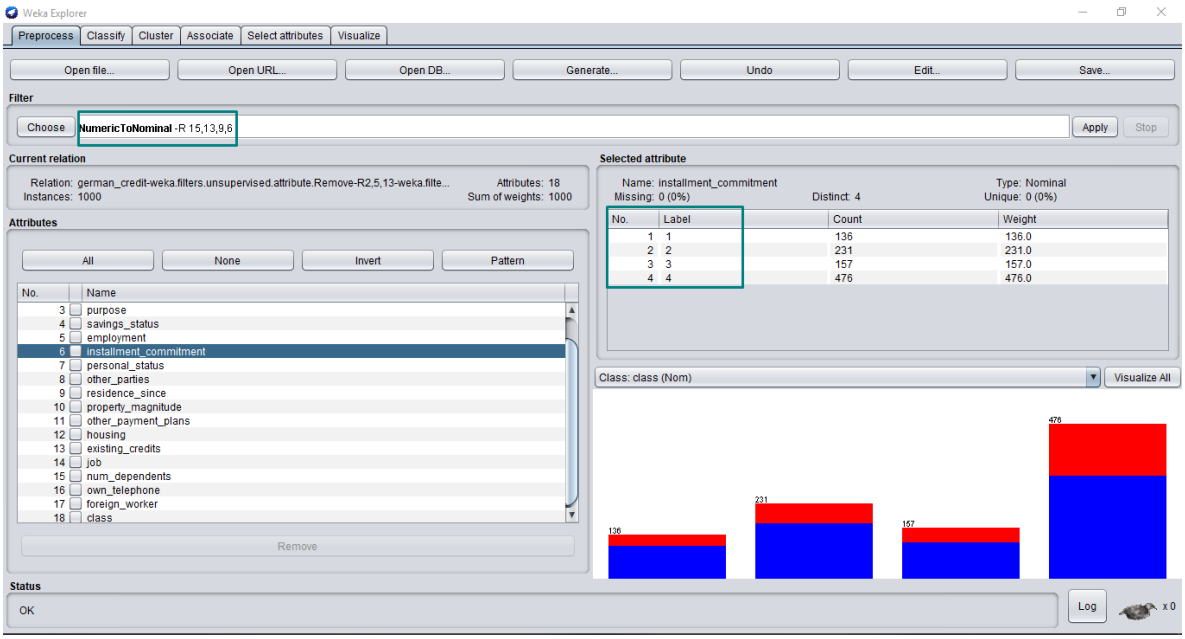
Clusterer output

```
  [total]                        126.7813 101.5365  64.8012  46.2196  94.9889  46.4687  54.2532  299.668  79.0533 106.2294
class
  good                           113.4945  93.7015   21.476  40.2539  56.7012  33.3122  49.7715 210.2252  34.5484  56.5158
  bad                             13.2868    7.835  43.3252   5.9657  38.2877  13.1565   4.4817  89.4428  44.5049  49.7136
  [total]                        126.7813 101.5365  64.8012  46.2196  94.9889  46.4687  54.2532  299.668  79.0533 106.2294


Time taken to build model (full training data) : 0.41 seconds

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

Clustered Instances

0      468 ( 47%)
1       85 (  9%)
2       19 (  2%)
3       42 (  4%)
4       38 (  4%)
5      217 ( 22%)
6       38 (  4%)
7       38 (  4%)
8       46 (  5%)
9        9 (  1%)


Log likelihood: -7.32939
```

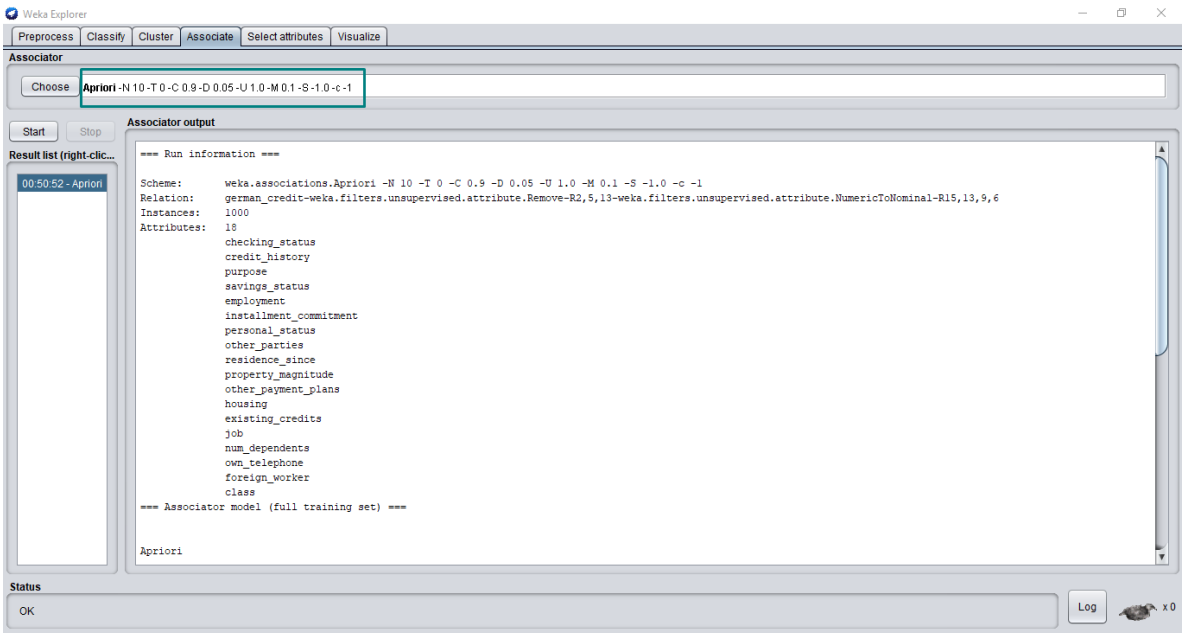
- Reglas de asociación:

  Apriori:

  - Convertir variables numéricas a categóricas:




- Aplicar Apriori:



**Associator output**

```
Apriori
=======

Minimum support: 0.7 (700 instances)
Minimum metric <confidence>: 0.9
Number of cycles performed: 6

Generated sets of large itemsets:

Size of set of large itemsets L(1): 6

Size of set of large itemsets L(2): 5

Size of set of large itemsets L(3): 2

Best rules found:

 1. other_parties=none num_dependents=1 767 ==> foreign_worker=yes 749    <conf:(0.98)> lift:(1.01) lev:(0.01) [10] conv:(1.49)
 2. other_parties=none 907 ==> foreign_worker=yes 880    <conf:(0.97)> lift:(1.01) lev:(0.01) [6] conv:(1.2)
 3. num_dependents=1 845 ==> foreign_worker=yes 819    <conf:(0.97)> lift:(1.01) lev:(0.01) [5] conv:(1.16)
 4. other_parties=none other_payment_plans=none 742 ==> foreign_worker=yes 718    <conf:(0.97)> lift:(1) lev:(0) [3] conv:(1.1)
 5. other_payment_plans=none 814 ==> foreign_worker=yes 782    <conf:(0.96)> lift:(1) lev:(-0) [-1] conv:(0.91)
 6. other_payment_plans=none foreign_worker=yes 782 ==> other_parties=none 718    <conf:(0.92)> lift:(1.01) lev:(0.01) [8] conv:(1.12)
 7. num_dependents=1 foreign_worker=yes 819 ==> other_parties=none 749    <conf:(0.91)> lift:(1.01) lev:(0.01) [6] conv:(1.07)
 8. foreign_worker=yes 963 ==> other_parties=none 880    <conf:(0.91)> lift:(1.01) lev:(0.01) [6] conv:(1.07)
 9. other_payment_plans=none 814 ==> other_parties=none 742    <conf:(0.91)> lift:(1.01) lev:(0) [3] conv:(1.04)
10. num_dependents=1 845 ==> other_parties=none 767    <conf:(0.91)> lift:(1) lev:(0) [0] conv:(0.99)
```