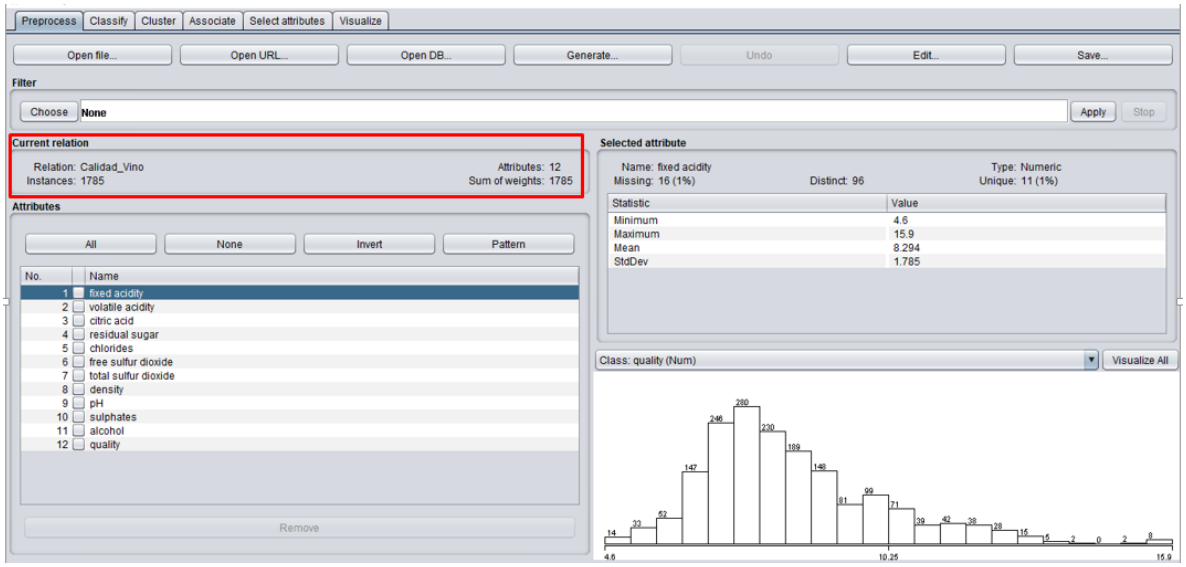
Laura Cristina López Bedoya

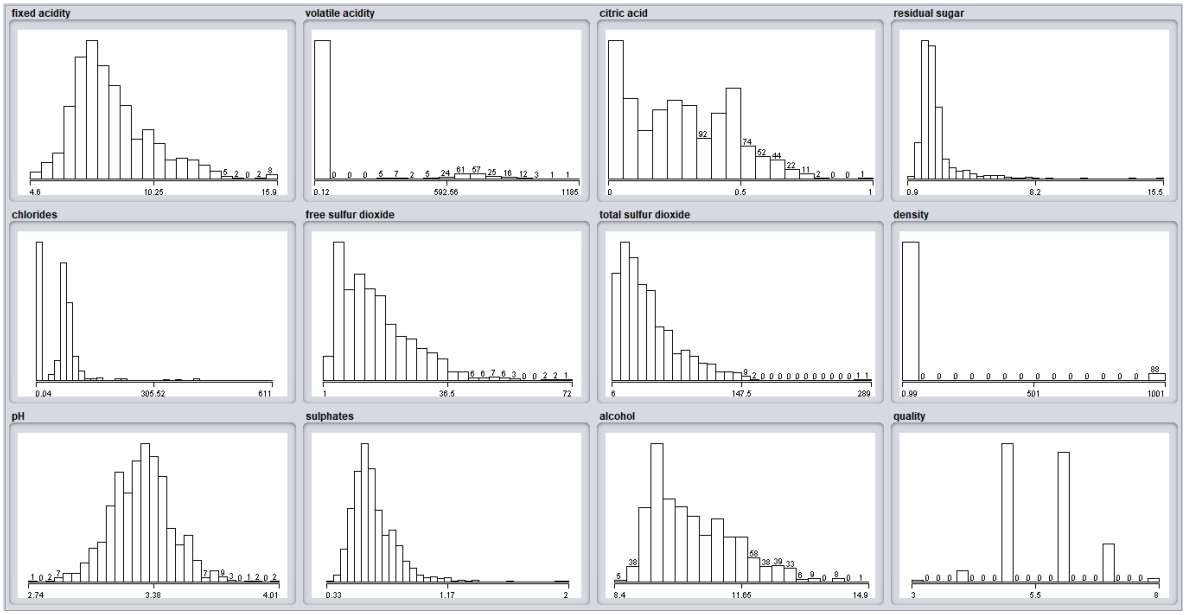
WEKA

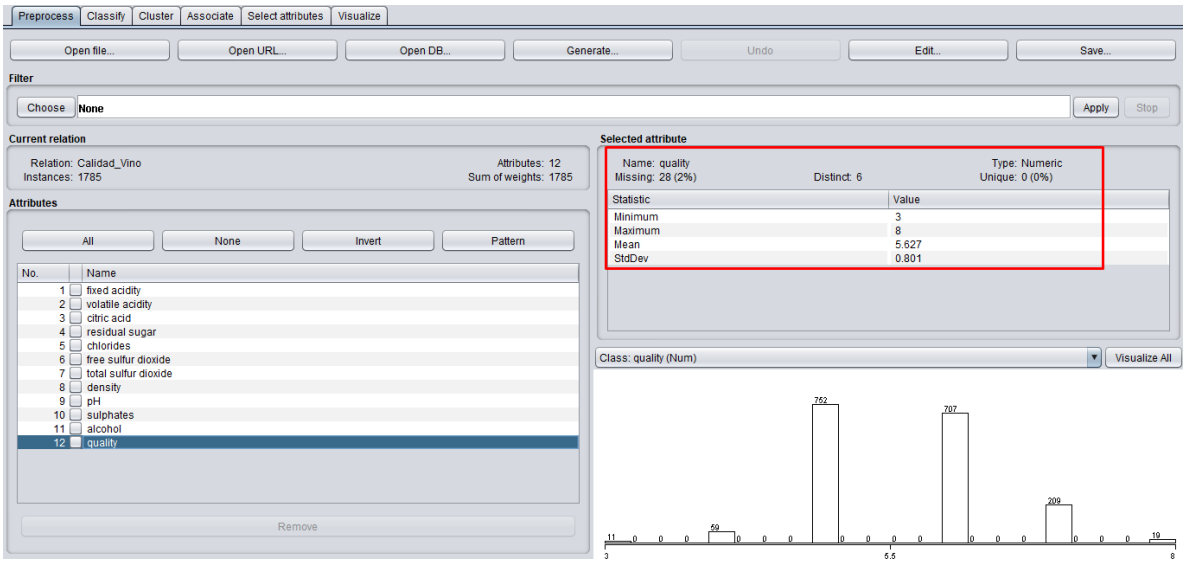
PREPARACIÓN DE DATOS

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

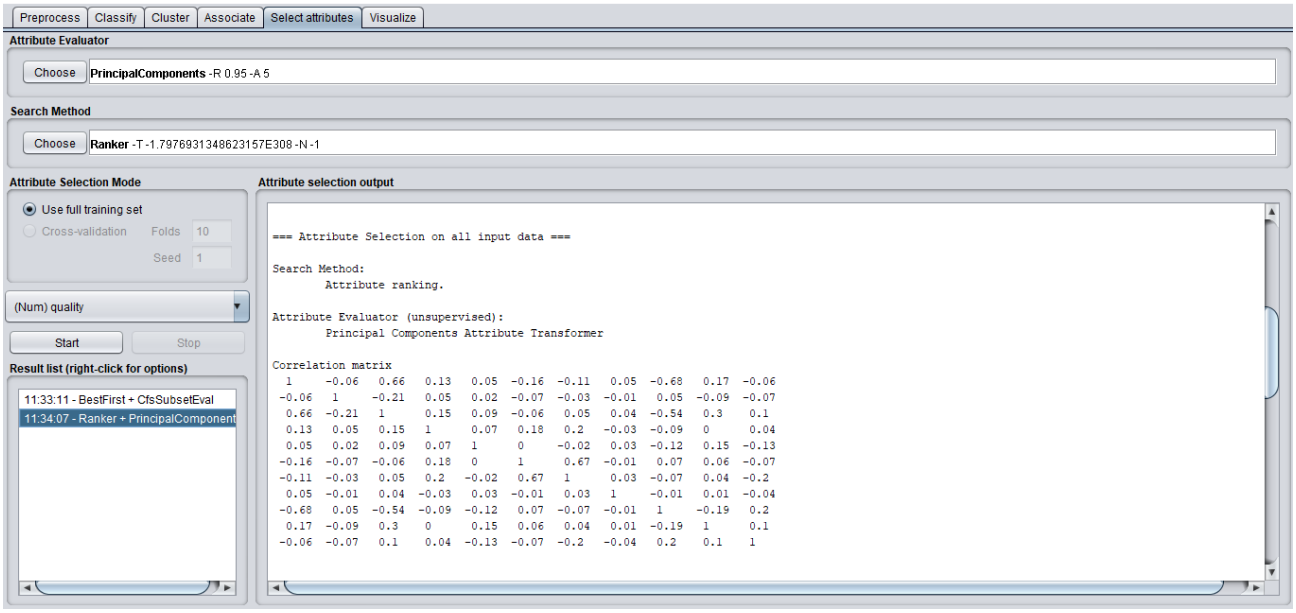
- Visualización de los datos:

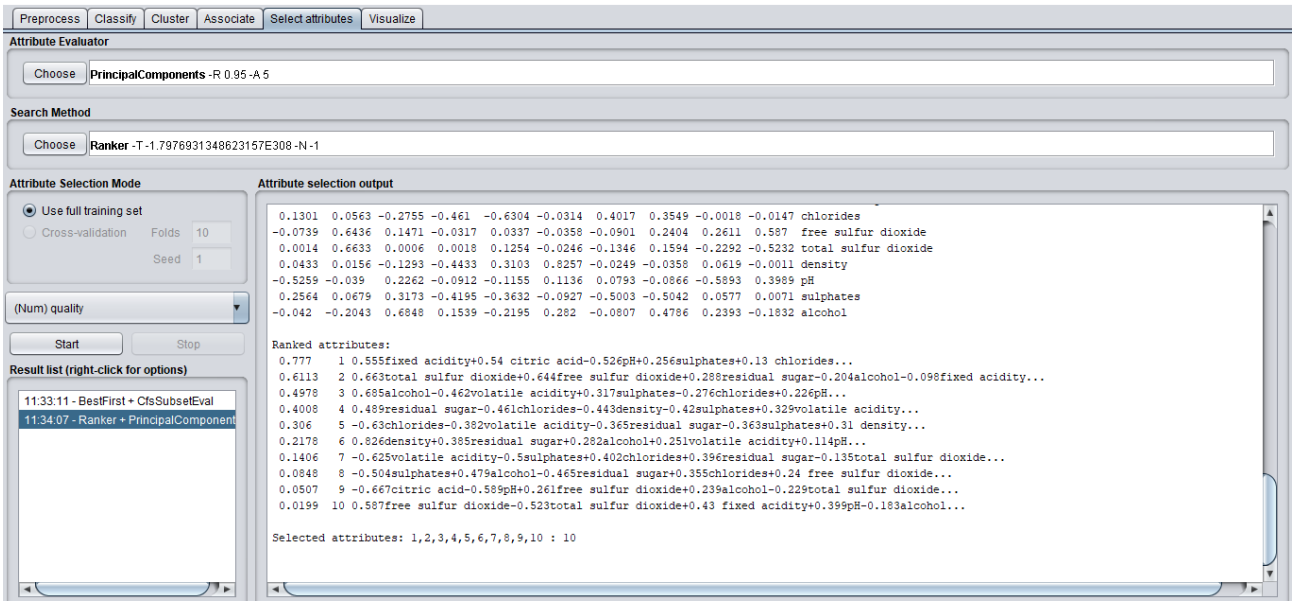




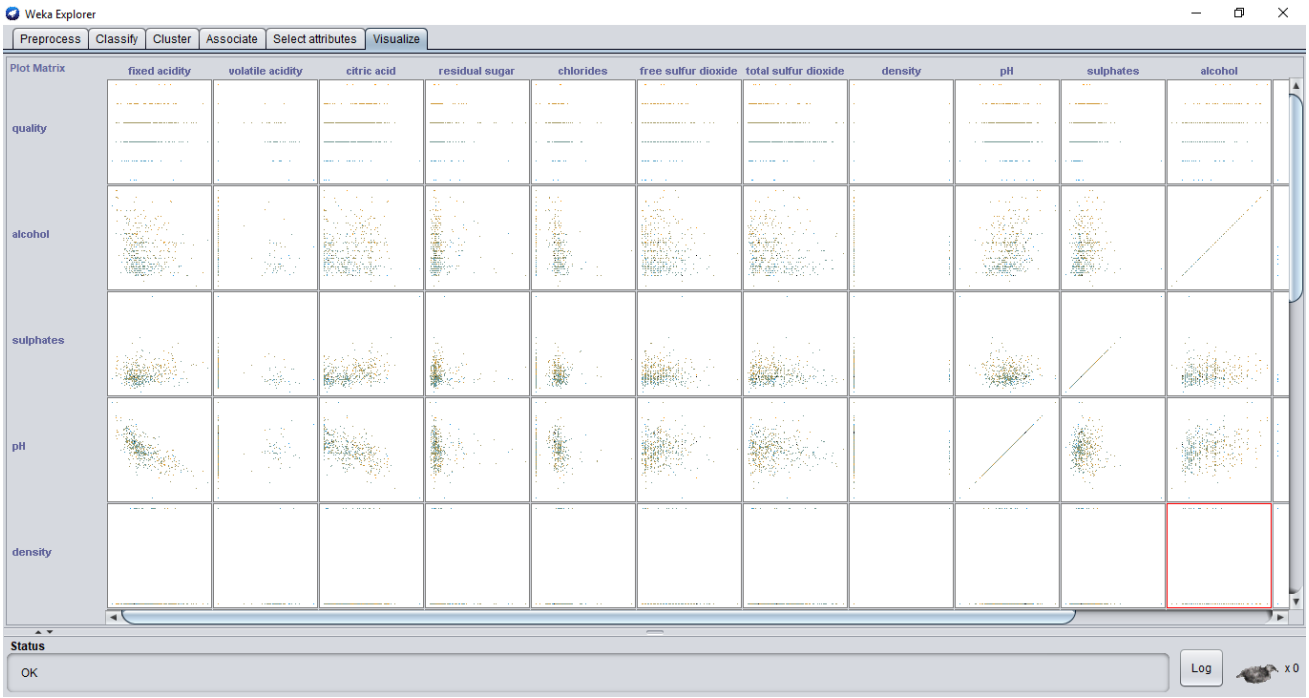


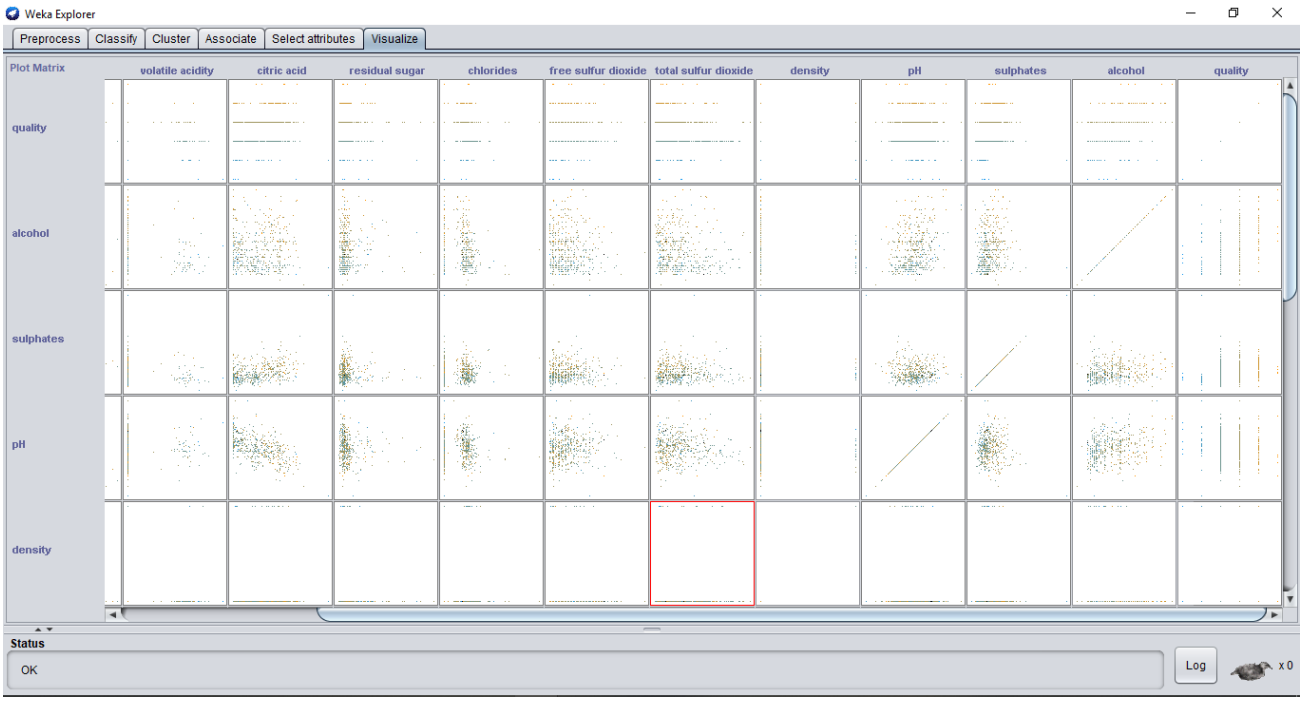
- Matriz de correlación:

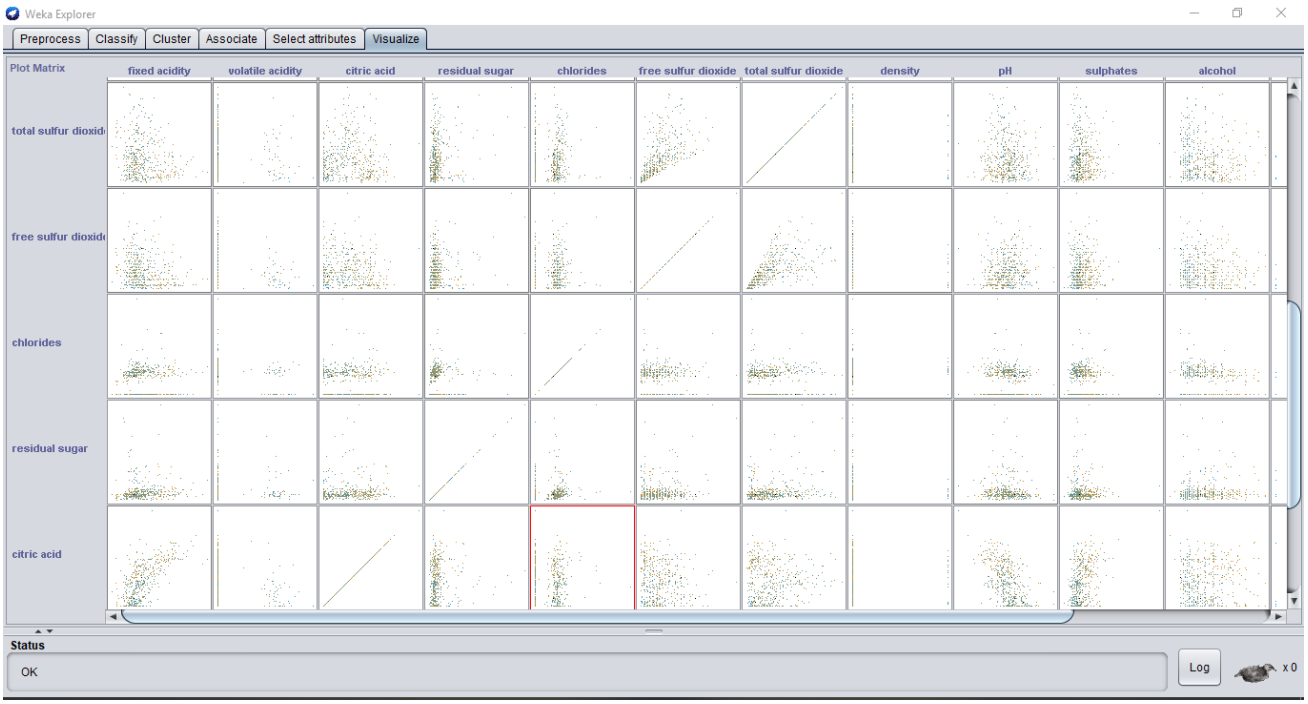


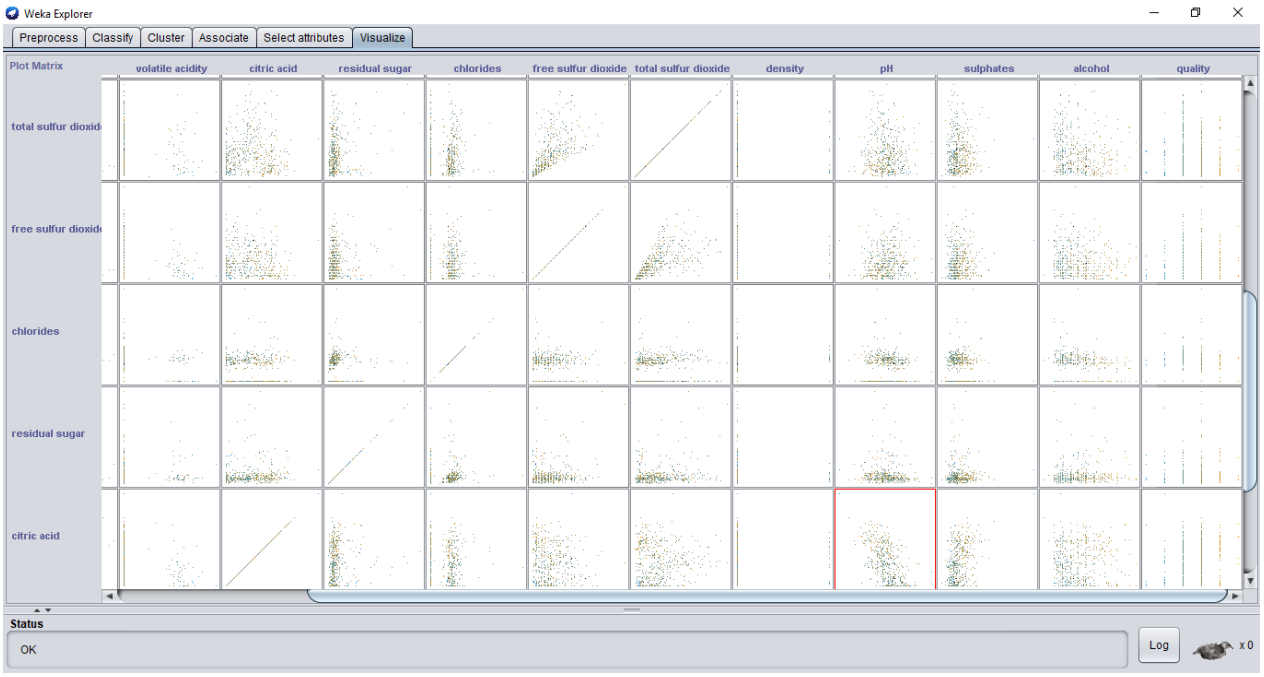


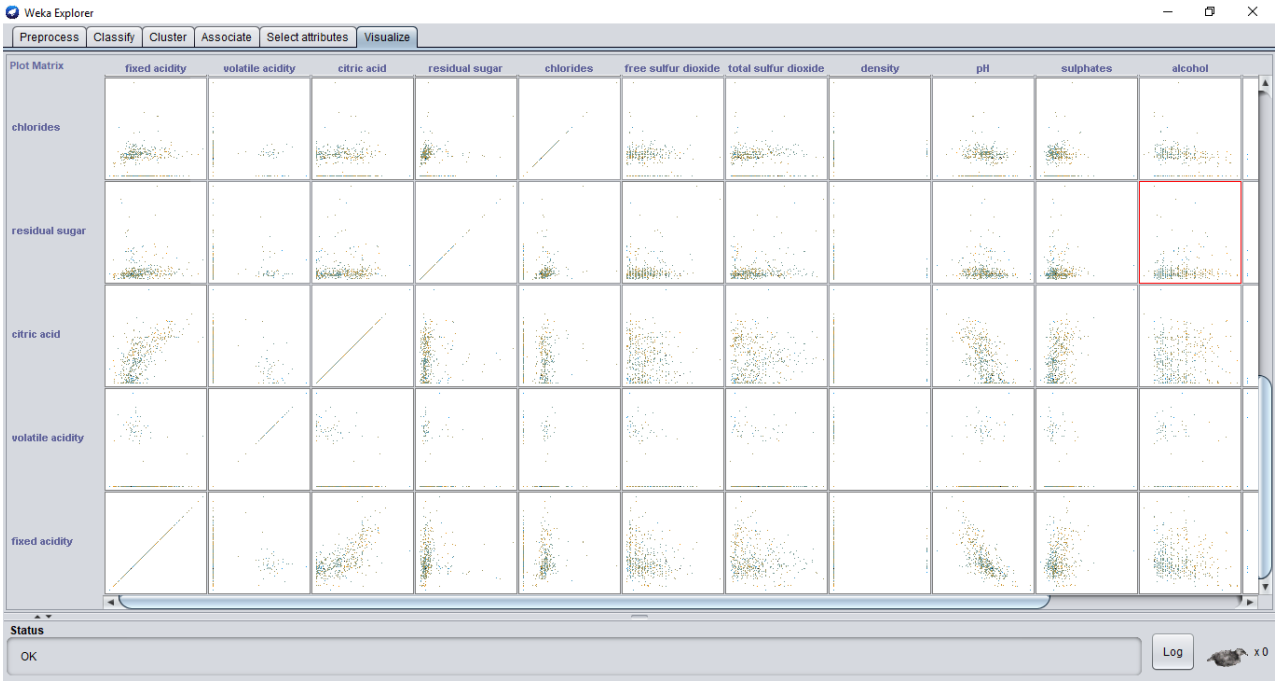
- Visualizar el scatter:

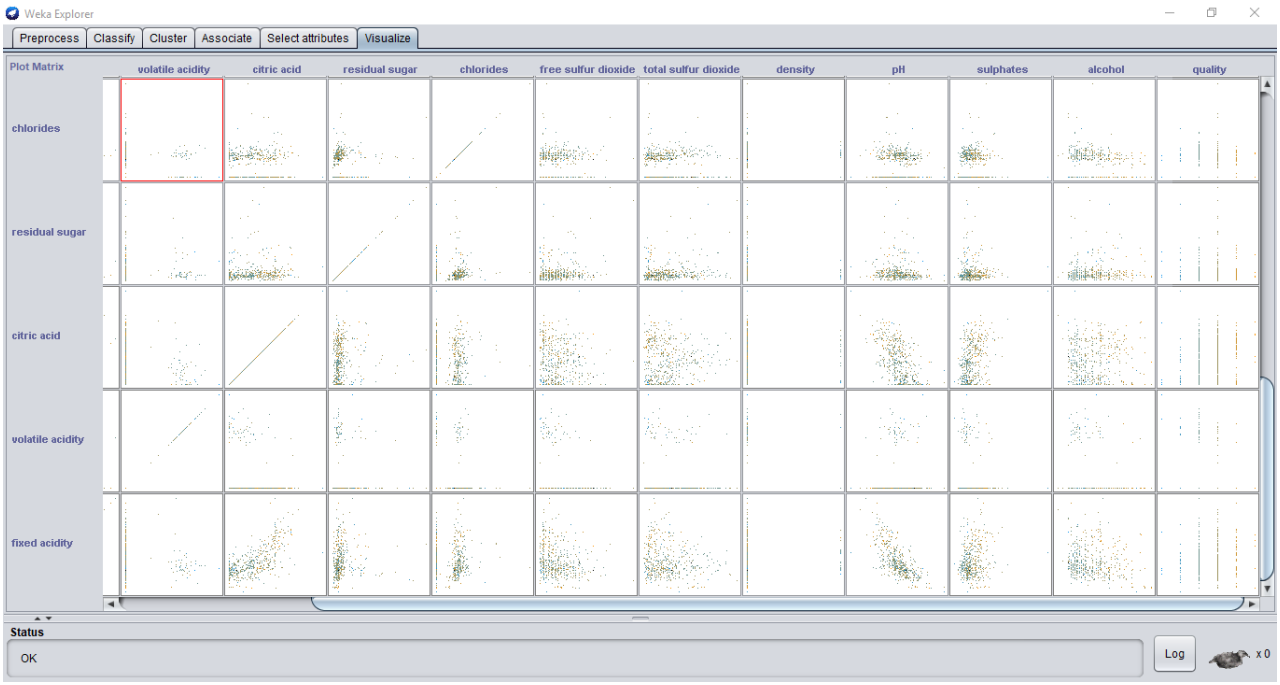



Weka Explorer
Preprocess
Classify
Cluster
Associate
Select attributes
Visualize
Plot Matrix
volatile acidity
citric acid
residual sugar
chlorides
free sulfur dioxide
total sulfur dioxide
density
pH
sulphates
alcohol
quality
Status
OK
Log
x 0

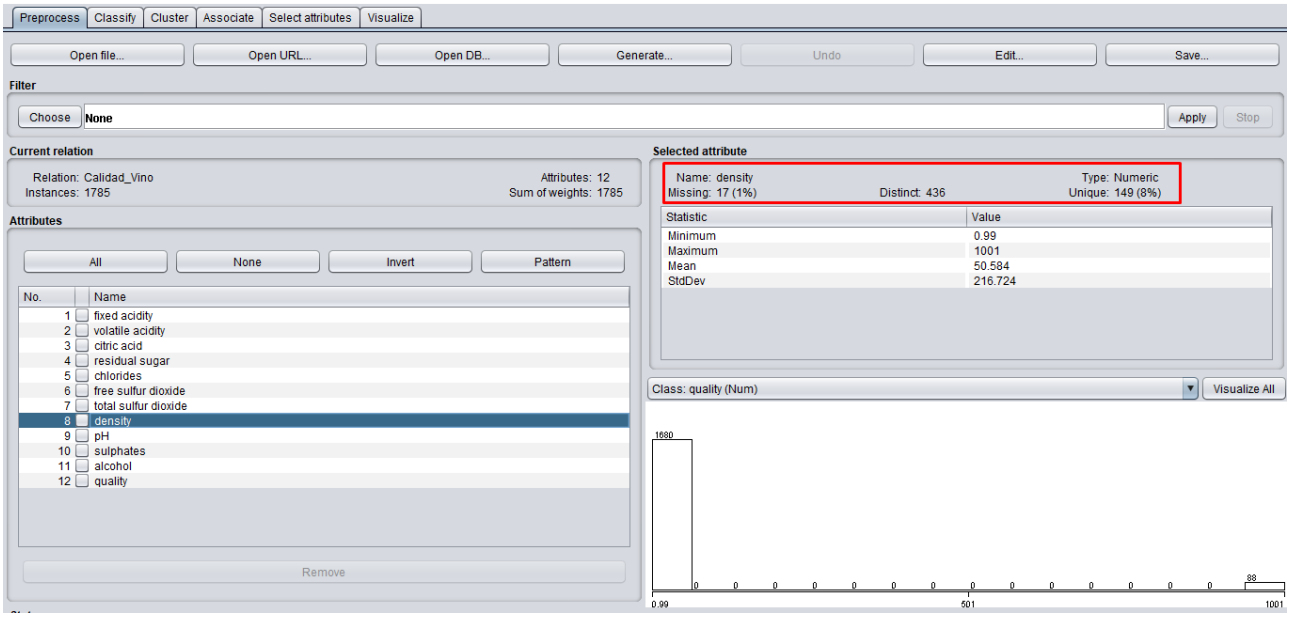

Weka Explorer
Preprocess
Classify
Cluster
Associate
Select attributes
Visualize
Plot Matrix
fixed acidity
volatile acidity
citric acid
residual sugar
chlorides
free sulfur dioxide
total sulfur dioxide
density
pH
sulphates
alcohol
Status
OK
Log
x 0

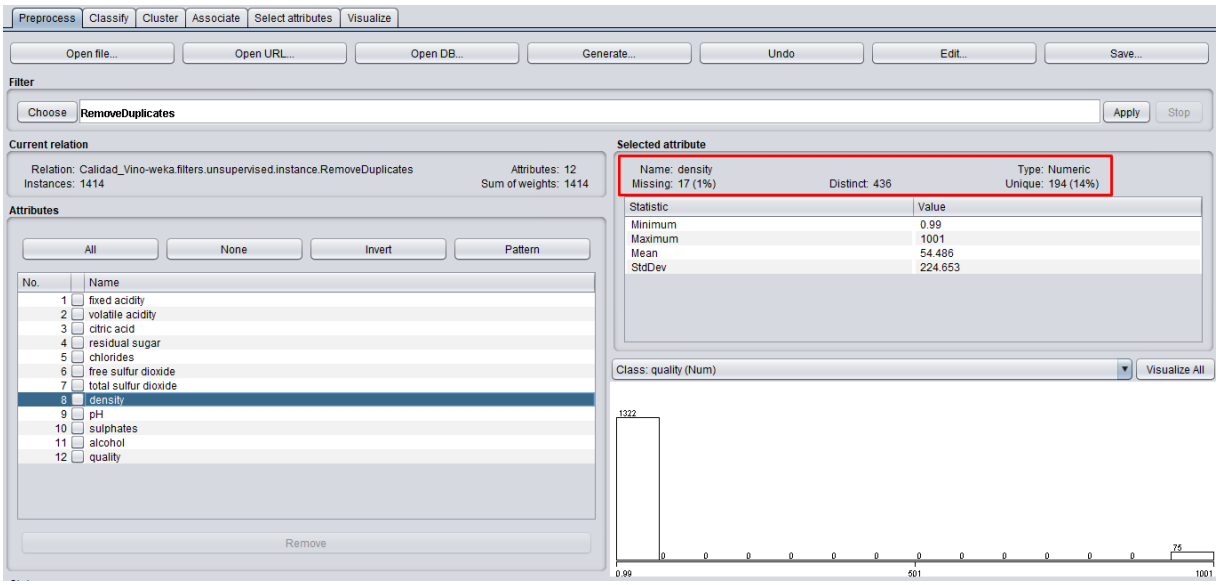

Weka Explorer
Preprocess
Classify
Cluster
Associate
Select attributes
Visualize
Plot Matrix
volatile acidity
citric acid
residual sugar
chlorides
free sulfur dioxide
total sulfur dioxide
density
pH
sulphates
alcohol
quality
total sulfur dioxid
free sulfur dioxid
chlorides
residual sugar
citric acid
Status
OK
Log
x 0


Weka Explorer
Preprocess
Classify
Cluster
Associate
Select attributes
Visualize
Plot Matrix
fixed acidity
volatile acidity
citric acid
residual sugar
chlorides
free sulfur dioxide
total sulfur dioxide
density
pH
sulphates
alcohol
chlorides
residual sugar
citric acid
volatile acidity
fixed acidity
Status
OK
Log
x 0

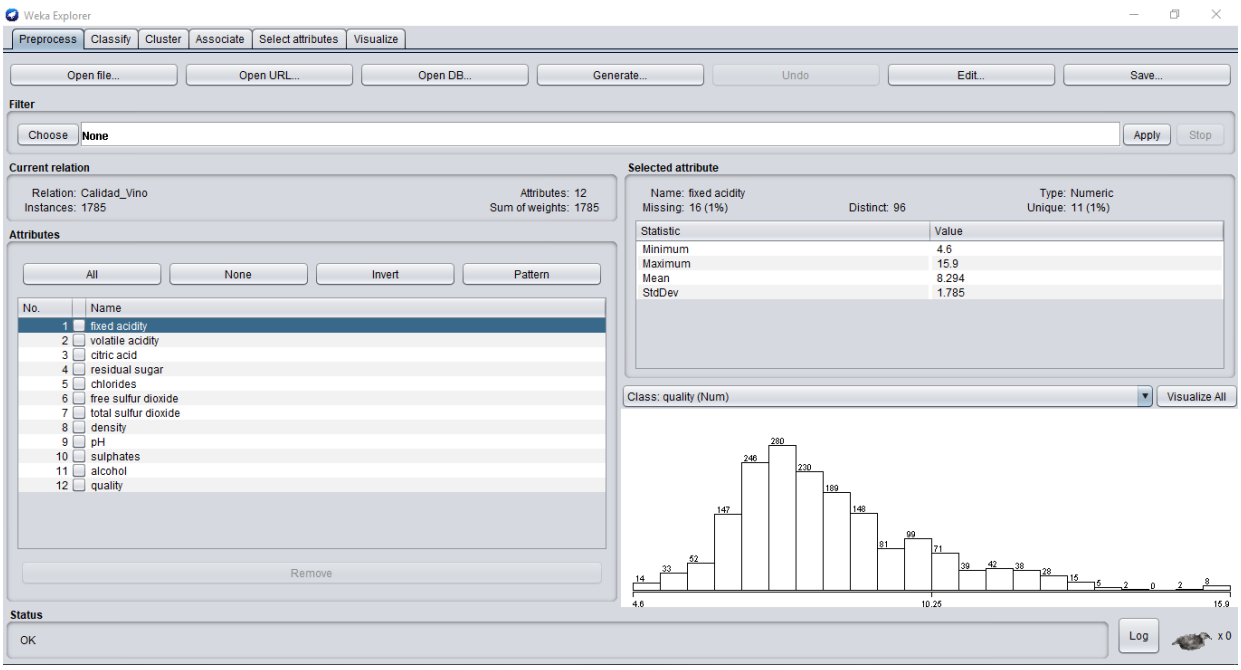


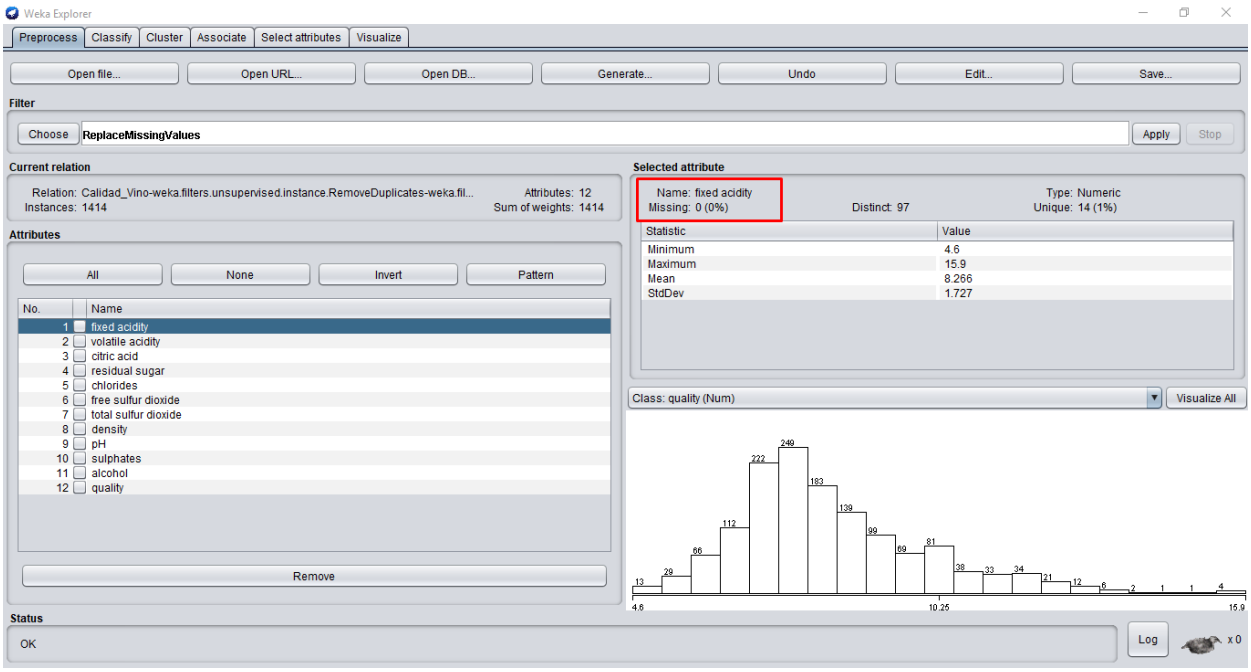
- Remover variables duplicadas:





- Remover valores faltantes:





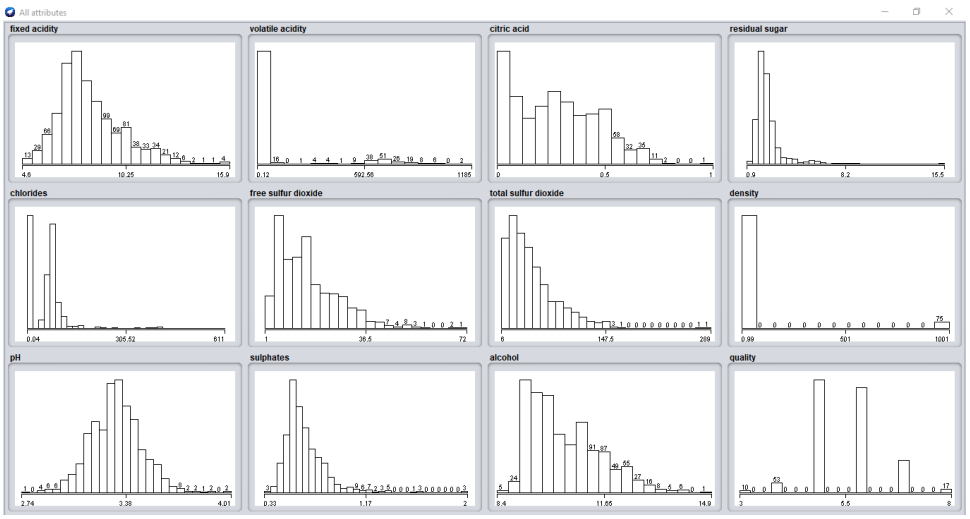
- Visualizar de nuevo la matriz de correlación:

```
=== Attribute Selection on all input data ===

Search Method:
        Attribute ranking.

Attribute Evaluator (unsupervised):
        Principal Components Attribute Transformer

Correlation matrix
  1      -0.08    0.66    0.12    0.05   -0.15   -0.1     0.07   -0.68    0.19   -0.07
 -0.08    1      -0.21    0.04    0.02   -0.07   -0.01    0       0.07   -0.08   -0.08
  0.66   -0.21    1       0.15    0.1    -0.05    0.05    0.06   -0.55    0.32    0.1
  0.12    0.04    0.15    1       0.03    0.15    0.2    -0.02   -0.09   -0.01    0.06
  0.05    0.02    0.1     0.03    1      -0.01   -0.02    0.02   -0.14    0.18   -0.13
 -0.15   -0.07   -0.05    0.15   -0.01    1       0.66   -0.02    0.06    0.05   -0.07
 -0.1    -0.01    0.05    0.2    -0.02    0.66    1       0.03   -0.08    0.03   -0.21
  0.07    0       0.06   -0.02    0.02   -0.02    0.03    1      -0.03    0.01   -0.04
 -0.68    0.07   -0.55   -0.09   -0.14    0.06   -0.08   -0.03    1      -0.21    0.21
  0.19   -0.08    0.32   -0.01    0.18    0.05    0.03    0.01   -0.21    1       0.09
 -0.07   -0.08    0.1     0.06   -0.13   -0.07   -0.21   -0.04    0.21    0.09    1
```

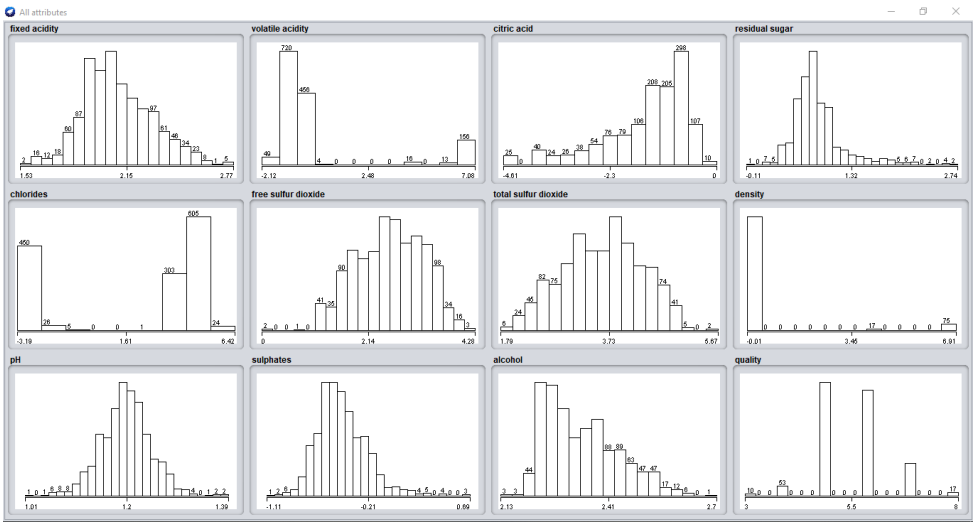
```
Ranked attributes:
 0.7738    1 0.545fixed acidity+0.538citric acid-0.524pH+0.28 sulphates+0.143chlorides...
 0.6101    2 0.673total sulfur dioxide+0.645free sulfur dioxide+0.263residual sugar-0.217alcohol-0.096fixed acidity...
 0.4973    3 -0.712alcohol+0.429volatile acidity+0.291chlorides-0.223sulphates-0.218pH...
 0.3988    4 -0.638chlorides-0.581sulphates+0.371residual sugar+0.243fixed acidity+0.172volatile acidity...
 0.305     5 -0.563residual sugar-0.55volatile acidity+0.415density-0.368chlorides-0.207alcohol...
 0.2156    6 0.886density+0.254residual sugar+0.248volatile acidity+0.241alcohol+0.142pH...
 0.1392    7 0.588volatile acidity-0.468chlorides+0.454sulphates-0.422residual sugar+0.125total sulfur dioxide...
 0.0846    8 -0.545sulphates+0.472alcohol-0.426residual sugar+0.352chlorides+0.241free sulfur dioxide...
 0.0512    9 0.673citric acid+0.549pH-0.304free sulfur dioxide+0.26 total sulfur dioxide-0.232alcohol...
 0.0199   10 0.548free sulfur dioxide-0.507total sulfur dioxide+0.454pH+0.428fixed acidity-0.213alcohol...

Selected attributes: 1,2,3,4,5,6,7,8,9,10 : 10
```

- Normalizar las variables para eliminar datos atípicos:

Antes:



Después:

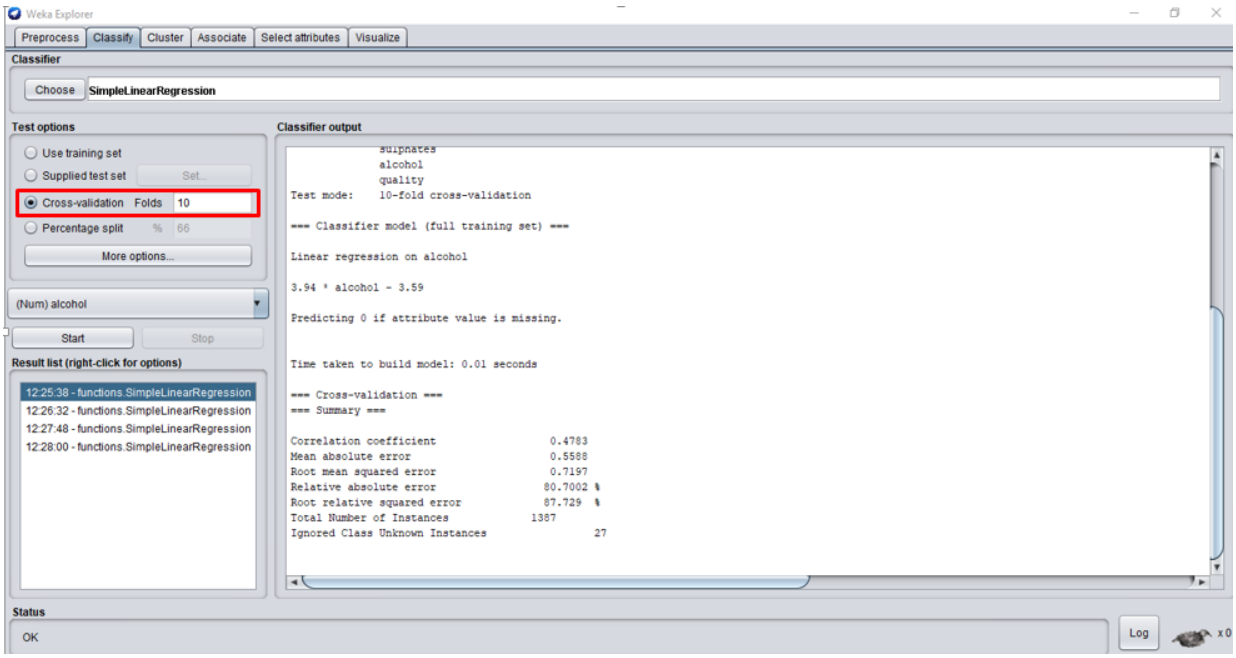


MÉTODOS SUPERVISADOS

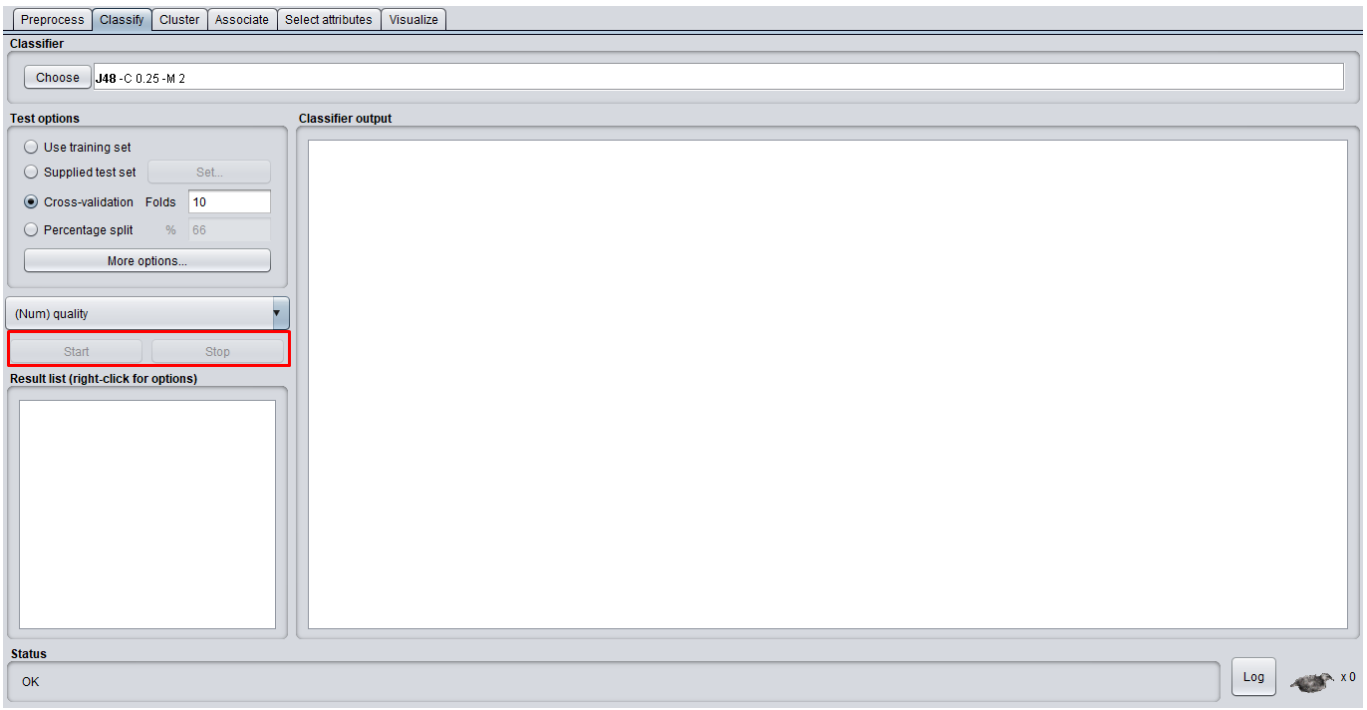
PREDICCIÓN:

A cada método se le aplica Cross-validation con Folds de 10.

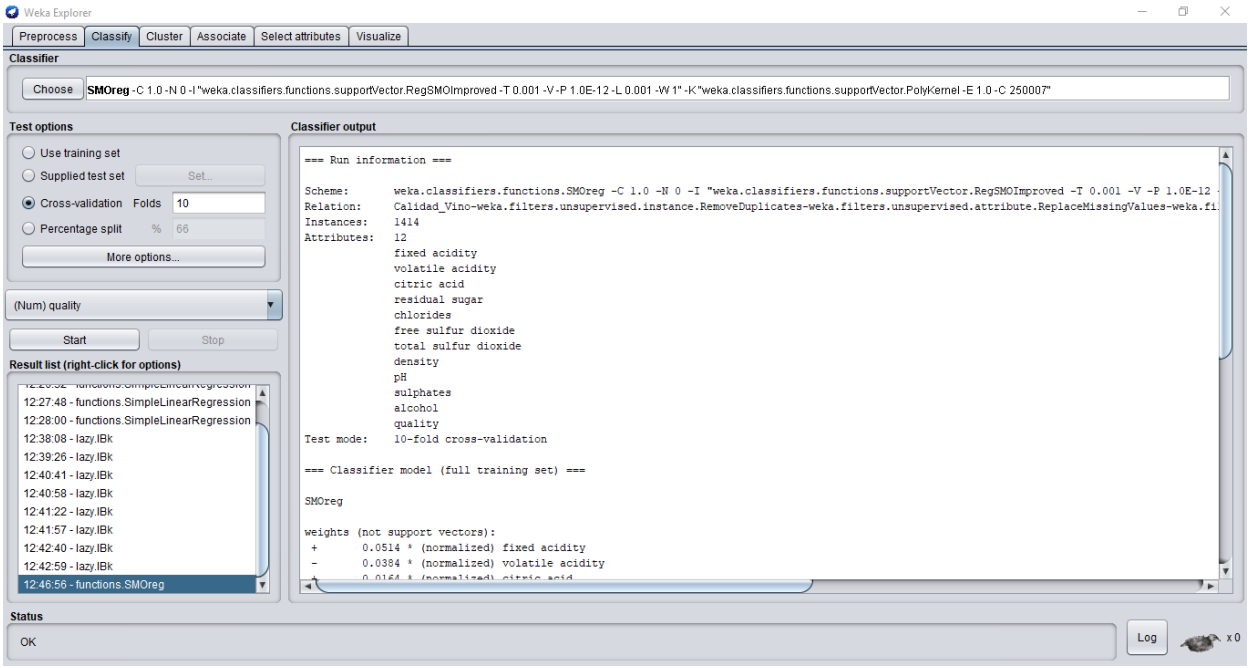
- REGRESIÓN LINEAL:



- DECISION TREE: El método no funciona

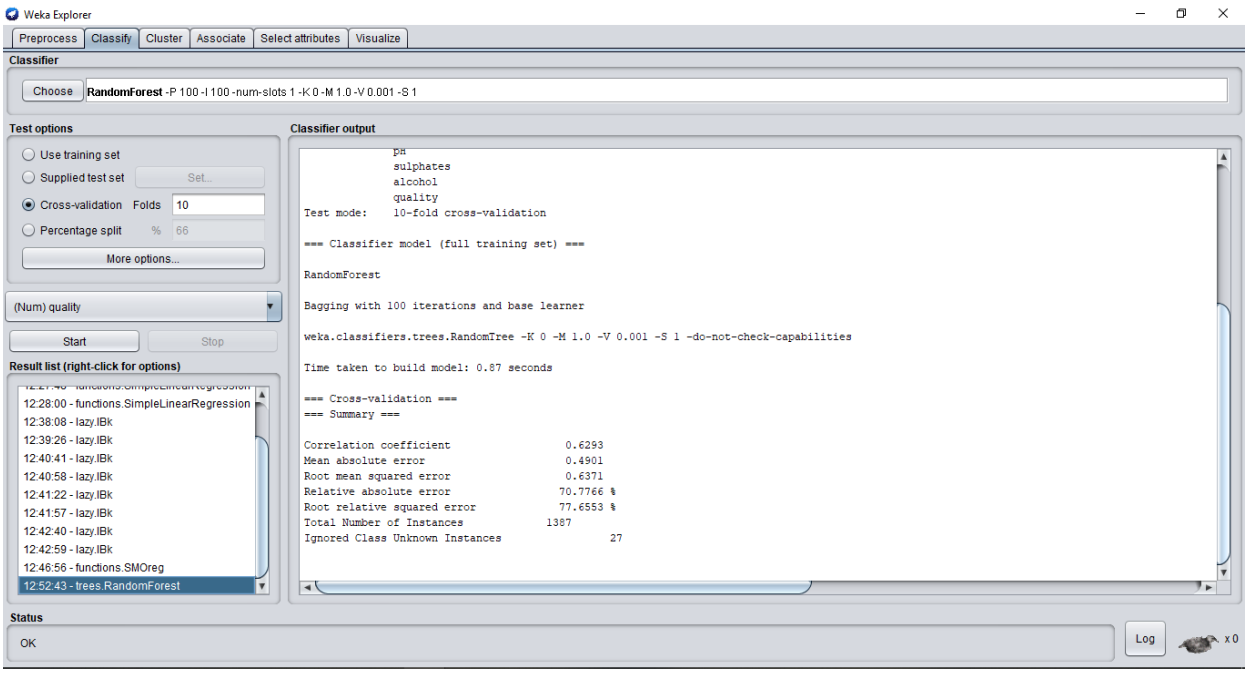


- SVM:

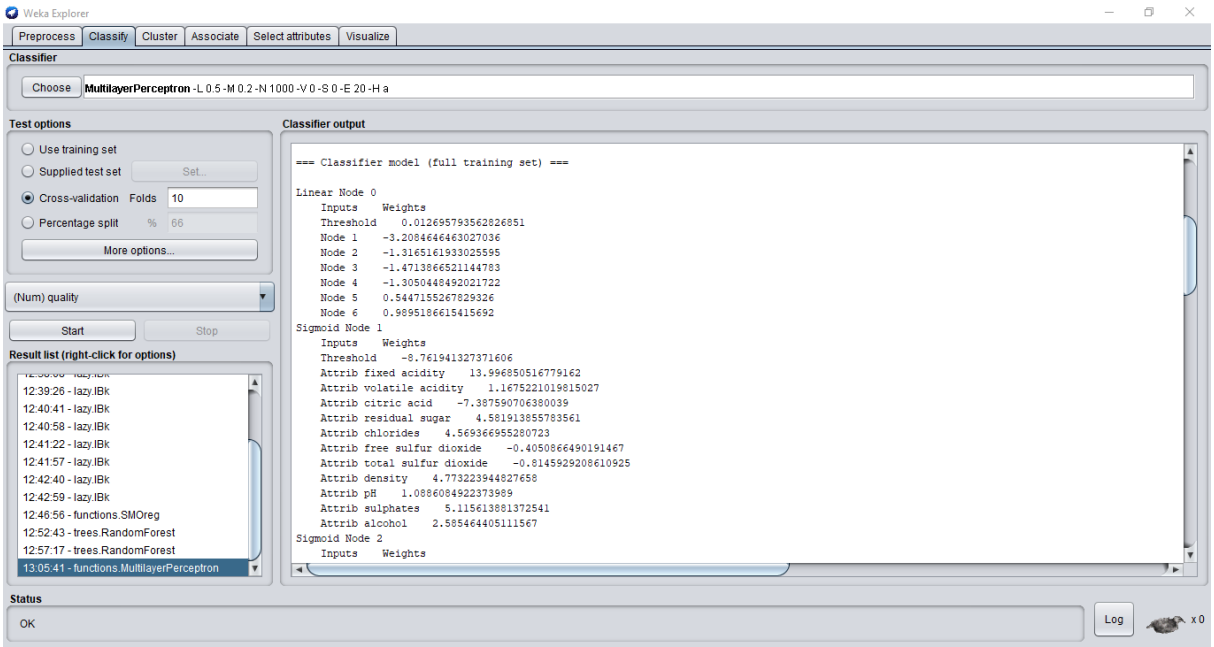



```
=== Classifier model (full training set) ===

SMOreg

weights (not support vectors):
 +       0.0514 * (normalized) fixed acidity
 -       0.0384 * (normalized) volatile acidity
 +       0.0164 * (normalized) citric acid
 +       0.0019 * (normalized) residual sugar
 -       0.0067 * (normalized) chlorides
 +       0.1144 * (normalized) free sulfur dioxide
 -       0.1458 * (normalized) total sulfur dioxide
 -       0.0082 * (normalized) density
 -       0.0582 * (normalized) pH
 +       0.2813 * (normalized) sulphates
 +       0.4321 * (normalized) alcohol
 +       0.2614


Number of kernel evaluations: 1581649 (87.677% cached)

Time taken to build model: 0.89 seconds

=== Cross-validation ===
=== Summary ===

Correlation coefficient                  0.5583
Mean absolute error                      0.5133
Root mean squared error                  0.6811
Relative absolute error                 74.1252 %
Root relative squared error             83.0226 %
Total Number of Instances             1387
Ignored Class Unknown Instances                27
```

- RANDOM FOREST:



- REDES NEURONALES:



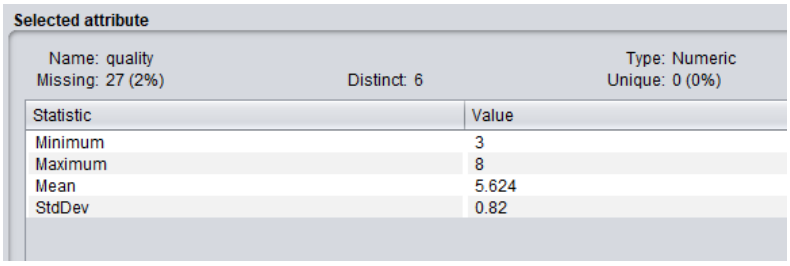
```
Sigmoid Node 2
    Inputs    Weights
    Threshold    -4.4582187647328615
    Attrib fixed acidity    -0.40563349185038833
    Attrib volatile acidity    0.38152171453821054
    Attrib citric acid    -0.14287853987183338
    Attrib residual sugar    0.1203710803056117
    Attrib chlorides    -0.24385525170380576
    Attrib free sulfur dioxide    -0.11809116754141137
    Attrib total sulfur dioxide    0.910065993319902
    Attrib density    -0.07074849595974292
    Attrib pH    -0.8825664082064211
    Attrib sulphates    0.2405677066480379
    Attrib alcohol    -3.917477268740148
Sigmoid Node 3
    Inputs    Weights
    Threshold    -15.622965477467762
    Attrib fixed acidity    -5.152845927826507
    Attrib volatile acidity    0.26351094575866046
    Attrib citric acid    -0.702123649679872
    Attrib residual sugar    5.831301562108013
    Attrib chlorides    -5.863624141343358
    Attrib free sulfur dioxide    -6.945431117627385
    Attrib total sulfur dioxide    -3.86897149987818
    Attrib density    -1.6314231057301567
    Attrib pH    6.932647003888969
    Attrib sulphates    -3.2629426967619573
    Attrib alcohol    0.18653309320209305
Sigmoid Node 4
    Inputs    Weights
    Threshold    -6.019053323698511
    Attrib fixed acidity    0.4139115077122772
    Attrib volatile acidity    -1.148845115282567
    Attrib citric acid    -0.4582379085529743
    Attrib residual sugar    1.3132393643927647
    Attrib chlorides    1.0304434576557744
    Attrib free sulfur dioxide    0.3227842080923362
    Attrib total sulfur dioxide    -1.5888857079315335
    Attrib density    0.4062341582723168
    Attrib pH    2.6126419690110674
    Attrib sulphates    -6.3971314504976435
    Attrib alcohol    0.023178605085421737
Sigmoid Node 5
    Inputs    Weights
    Threshold    -11.336210847742683
    Attrib fixed acidity    0.1916720980259231
    Attrib volatile acidity    -11.262973052452631
    Attrib citric acid    -1.570416974169361
    Attrib residual sugar    1.3917140686252036
    Attrib chlorides    0.08611115191186484
```

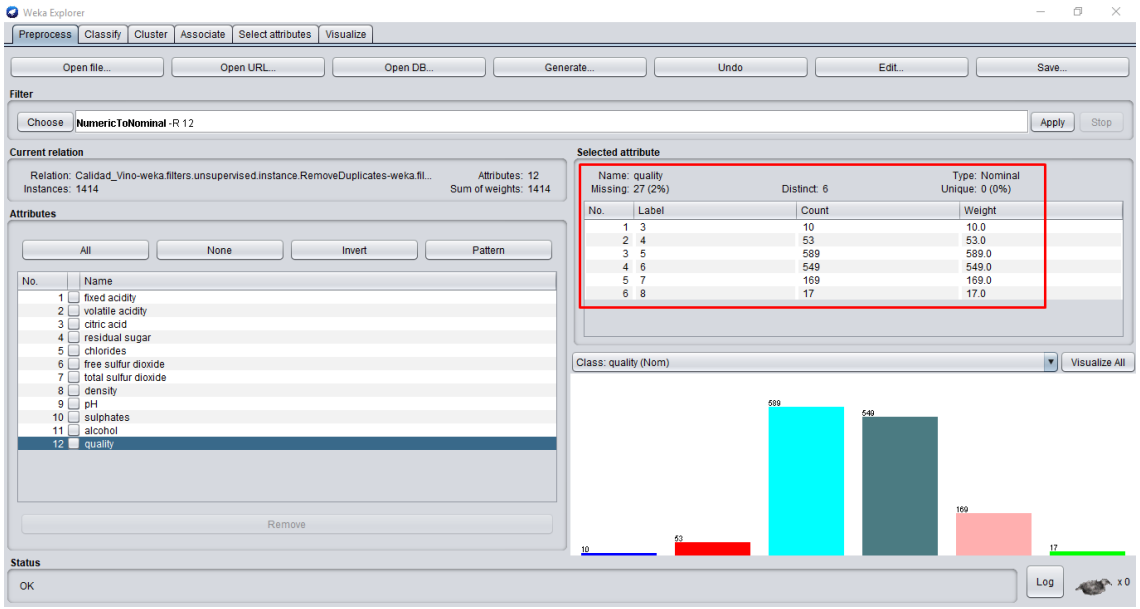
```
    Attrib chlorides    0.08611115191186484
    Attrib free sulfur dioxide    2.400292572993844
    Attrib total sulfur dioxide    -2.6694629112133286
    Attrib density    -2.02727021205916
    Attrib pH    -1.3673652922028878
    Attrib sulphates    0.8984799694911312
    Attrib alcohol    3.0305405140055304
Sigmoid Node 6
    Inputs    Weights
    Threshold    -2.819087750489133
    Attrib fixed acidity    0.8405585530794177
    Attrib volatile acidity    1.8302617803648014
    Attrib citric acid    0.12029925528830675
    Attrib residual sugar    -3.2833847768113804
    Attrib chlorides    -0.6122214332882024
    Attrib free sulfur dioxide    -4.697114610587977
    Attrib total sulfur dioxide    0.4888348412169438
    Attrib density    2.466668422813461
    Attrib pH    5.862370602749059
    Attrib sulphates    8.435860987812159
    Attrib alcohol    5.814858175506903
Class
    Input
    Node 0


Time taken to build model: 2.64 seconds
```

```
Time taken to build model: 2.64 seconds

=== Cross-validation ===
=== Summary ===

Correlation coefficient                  0.3702
Mean absolute error                      0.6503
Root mean squared error                  0.8812
Relative absolute error                 93.916  %
Root relative squared error            107.4137 %
Total Number of Instances             1387
Ignored Class Unknown Instances                27
```

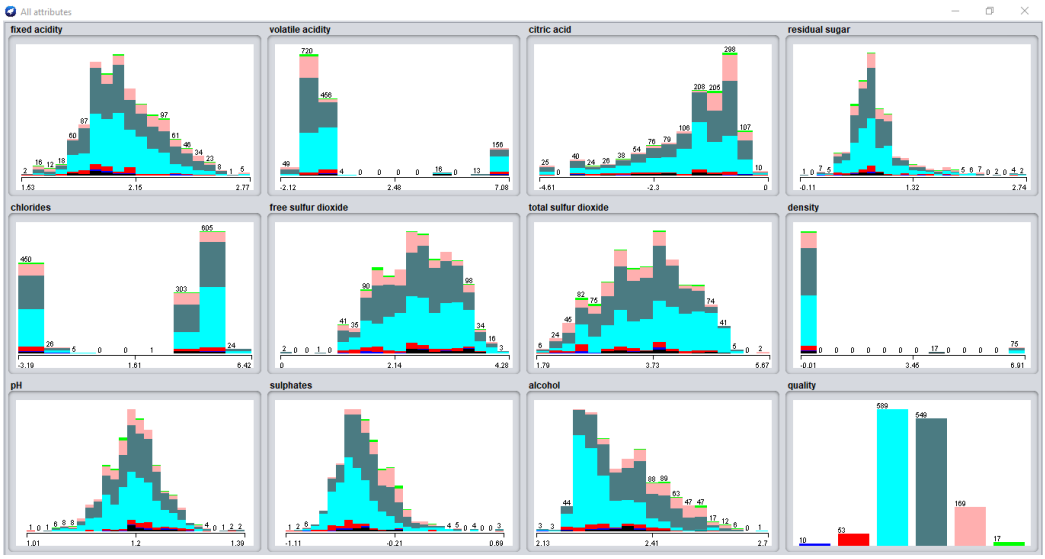
CLASIFICACIÓN:

- Convertir variables numéricas a categóricas:

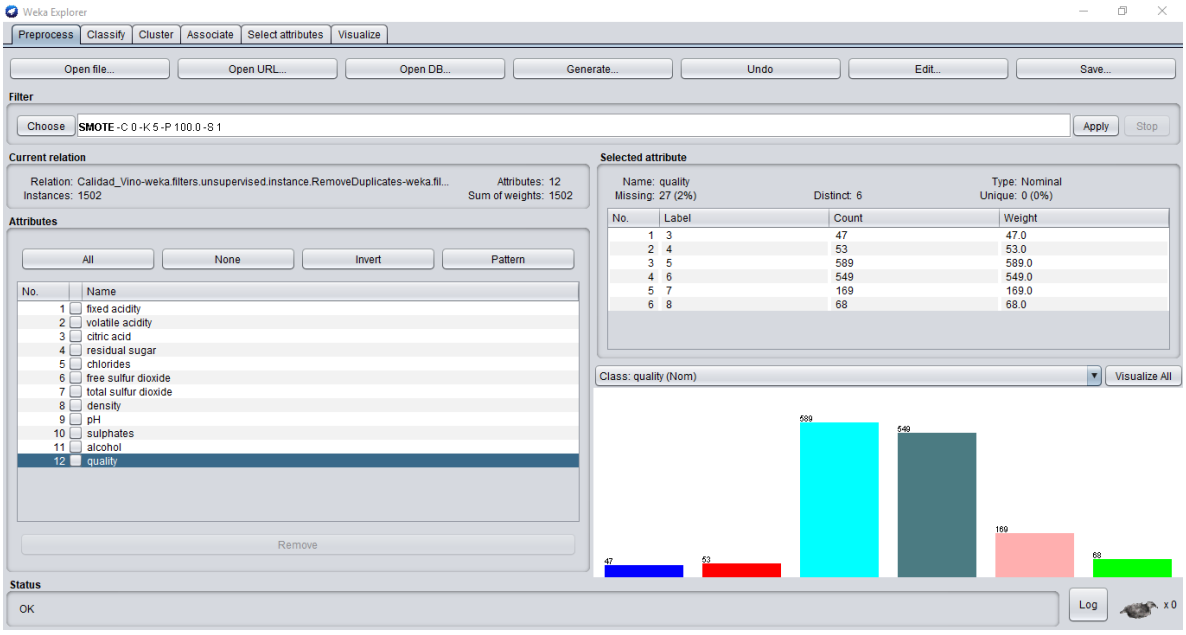




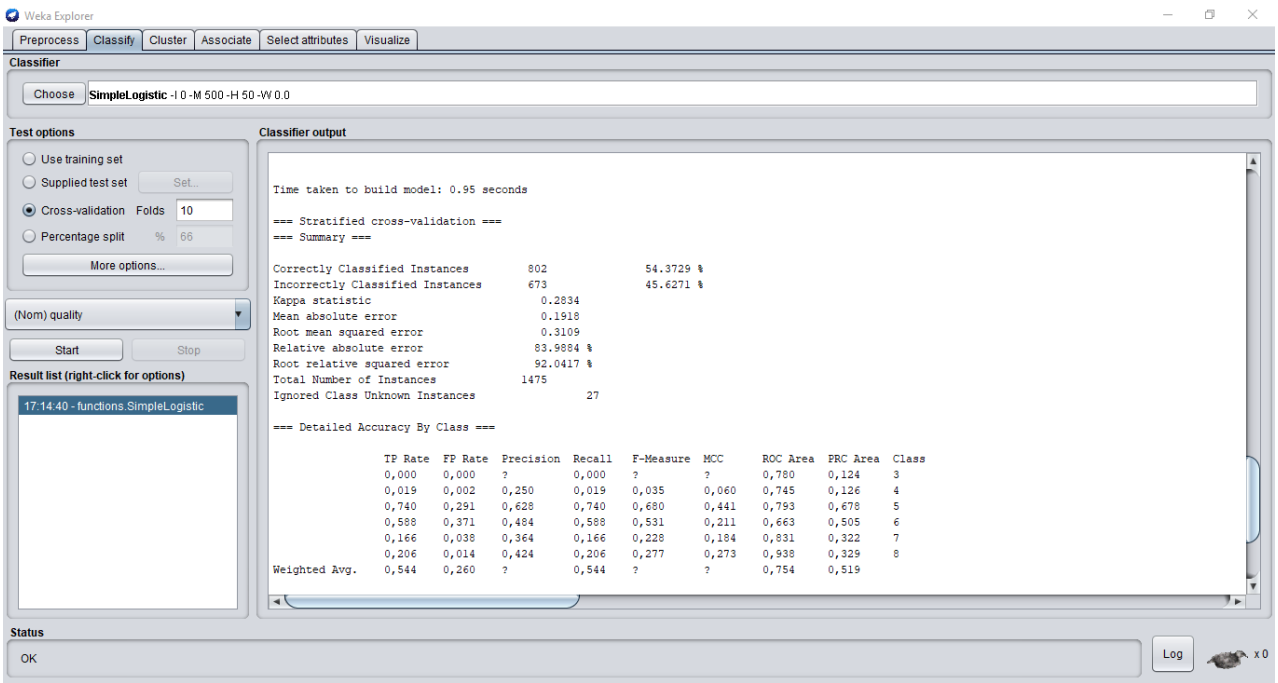
- Visualizar los datos:



- Balancear los datos:

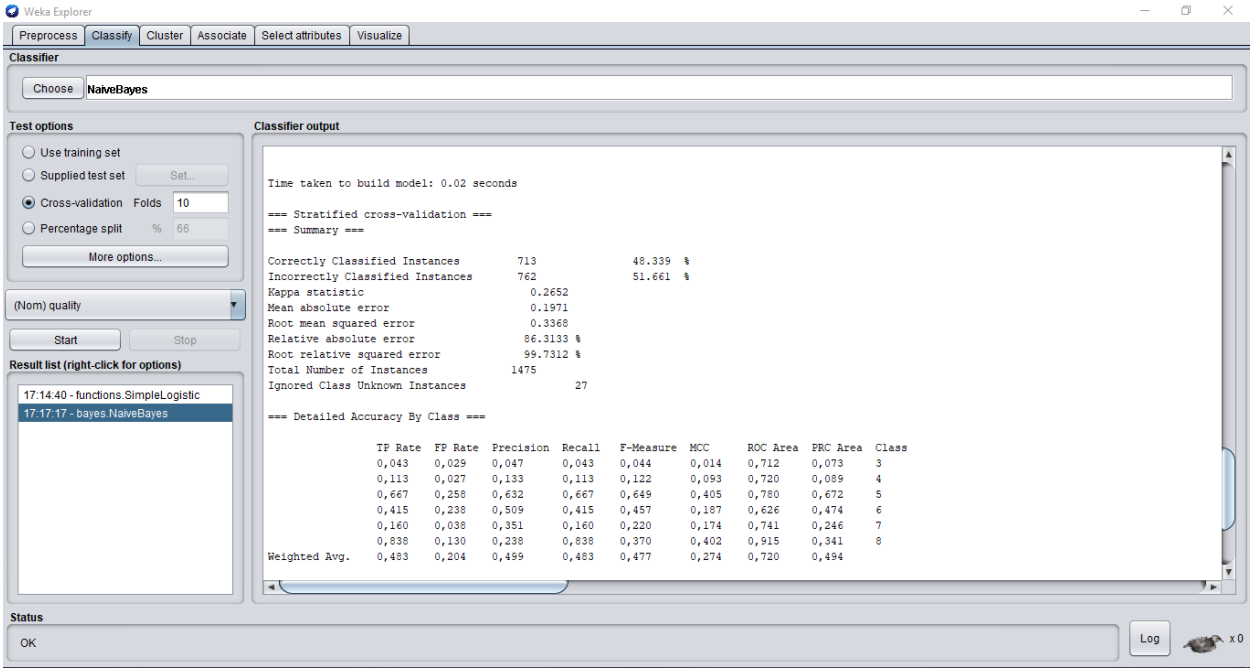


- REGRESIÓN LOGÍSTICA:



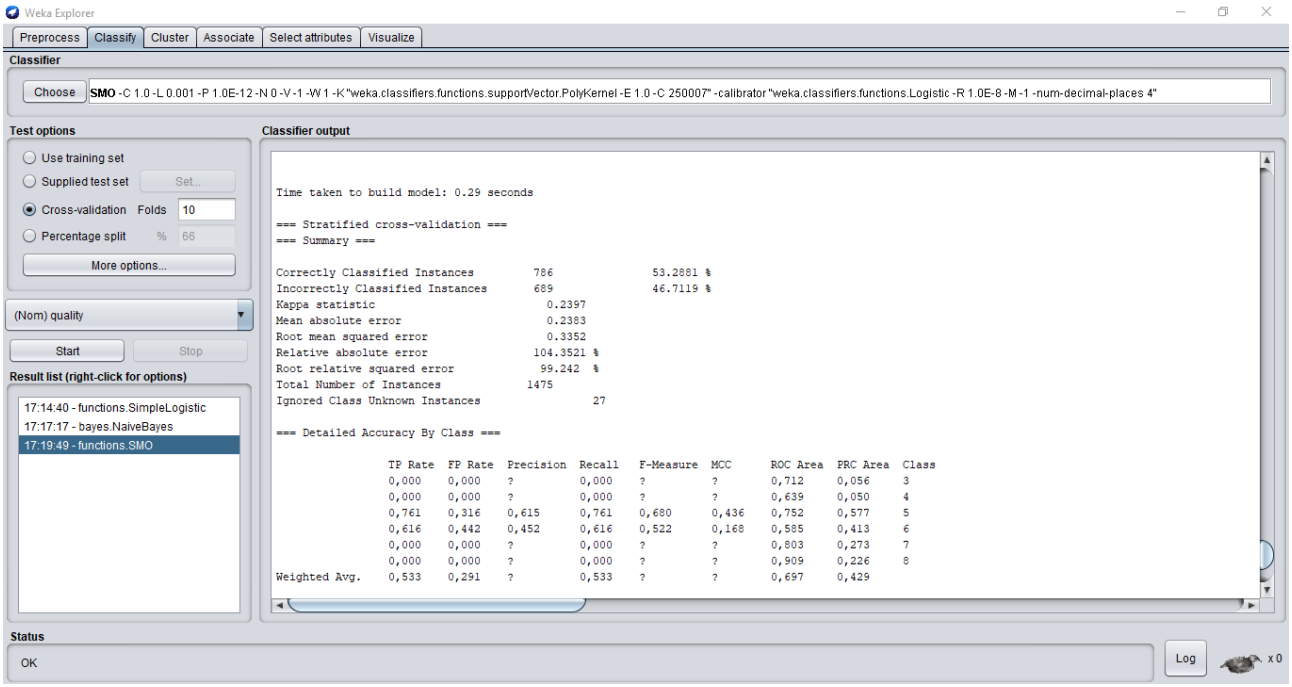
```
=== Confusion Matrix ===

   a   b   c   d   e   f   <-- classified as
   0   2  24  21   0   0 |   a = 3
   0   1  28  23   0   1 |   b = 4
   0   1 436 147   3   2 |   c = 5
   0   0 196 323  23   7 |   d = 6
   0   0  10 122  28   9 |   e = 7
   0   0   0  31  23  14 |   f = 8
```

- NAIVE –BAYES:




```
=== Confusion Matrix ===

   a   b   c   d   e   f   <-- classified as
   2   6  24  14   0   1 |   a = 3
   2   6  27  15   3   0 |   b = 4
  18  18 393 135  11  14 |   c = 5
  19  11 167 228  33  91 |   d = 6
   2   4  10  49  27  77 |   e = 7
   0   0   1   7   3  57 |   f = 8
```

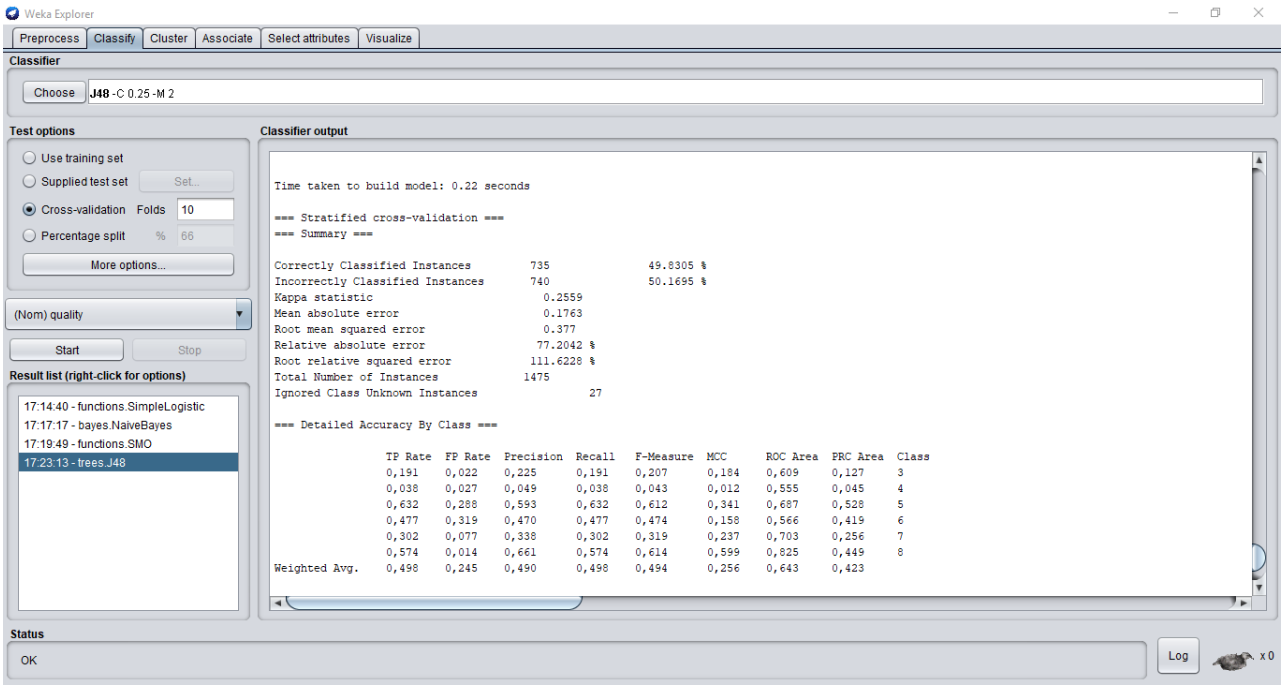
- SVM:



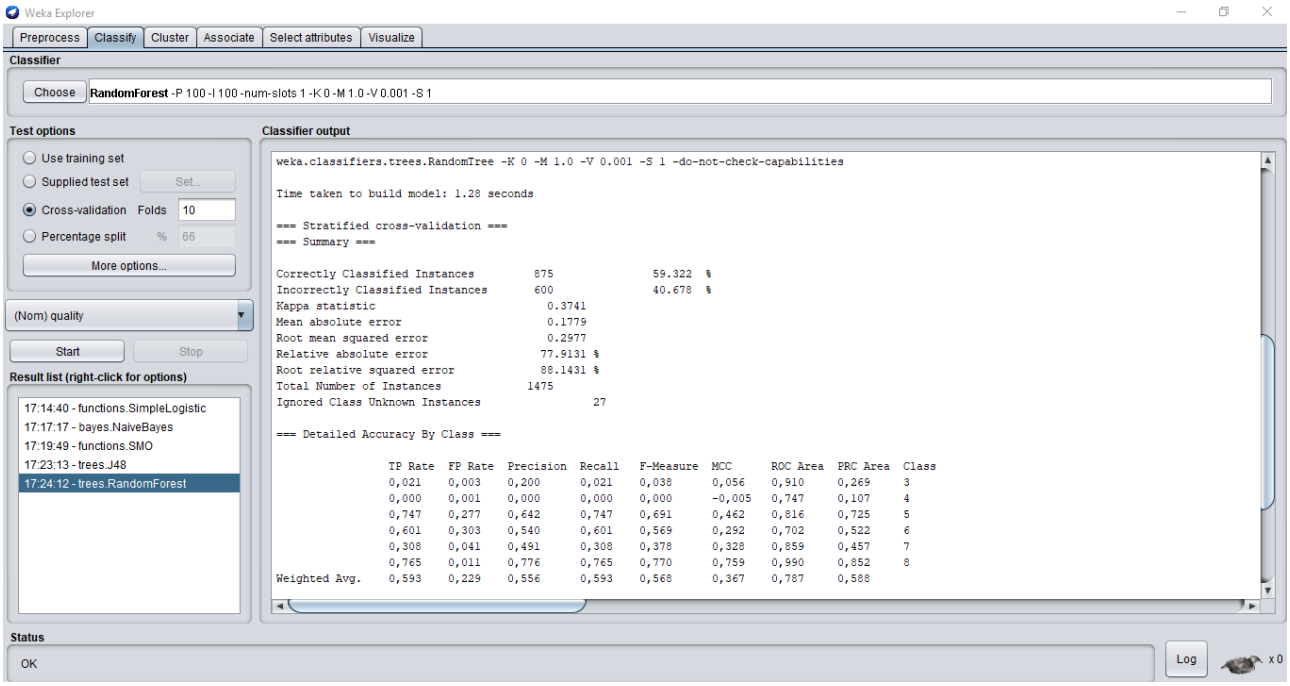

```
=== Confusion Matrix ===

   a   b   c   d   e   f   <-- classified as
   0   0  28  19   0   0 |   a = 3
   0   0  29  24   0   0 |   b = 4
   0   0 448 141   0   0 |   c = 5
   0   0 211 338   0   0 |   d = 6
   0   0  12 157   0   0 |   e = 7
   0   0   0  68   0   0 |   f = 8
```

- DECISION TREE:

```
=== Confusion Matrix ===

  a   b   c   d   e   f   <-- classified as
  9   1  17  19   1   0 |   a = 3
  1   2  30  20   0   0 |   b = 4
 16  16 372 166  18   1 |   c = 5
 13  18 182 262  65   9 |   d = 6
  1   4  25  78  51  10 |   e = 7
  0   0   1  12  16  39 |   f = 8
```



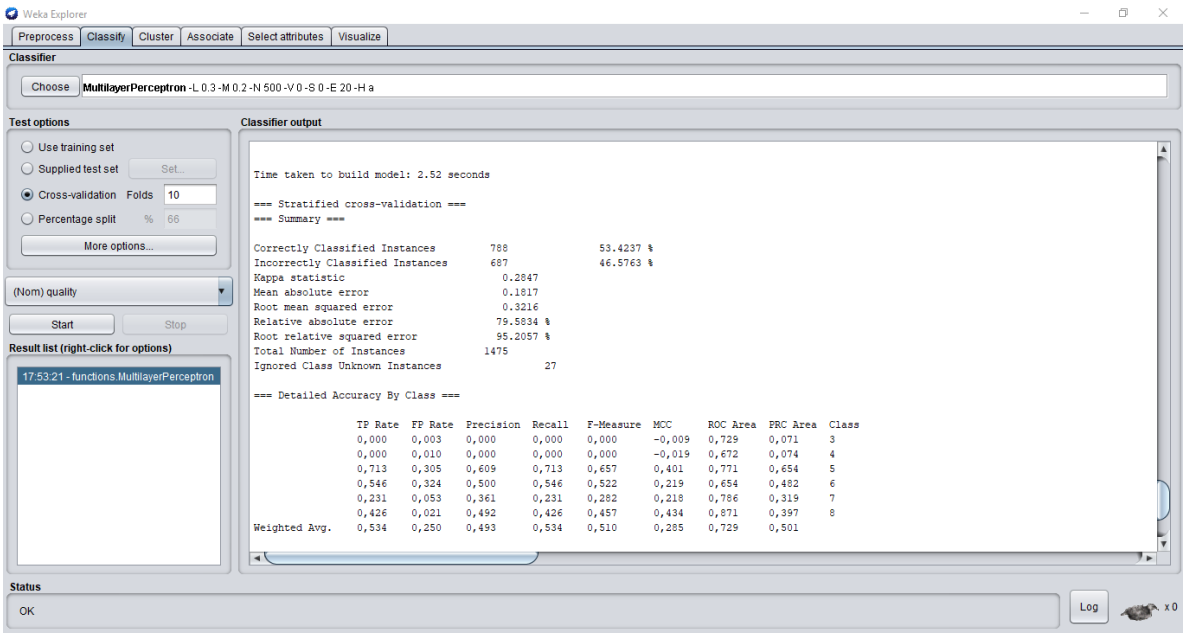
- RANDOM FOREST:



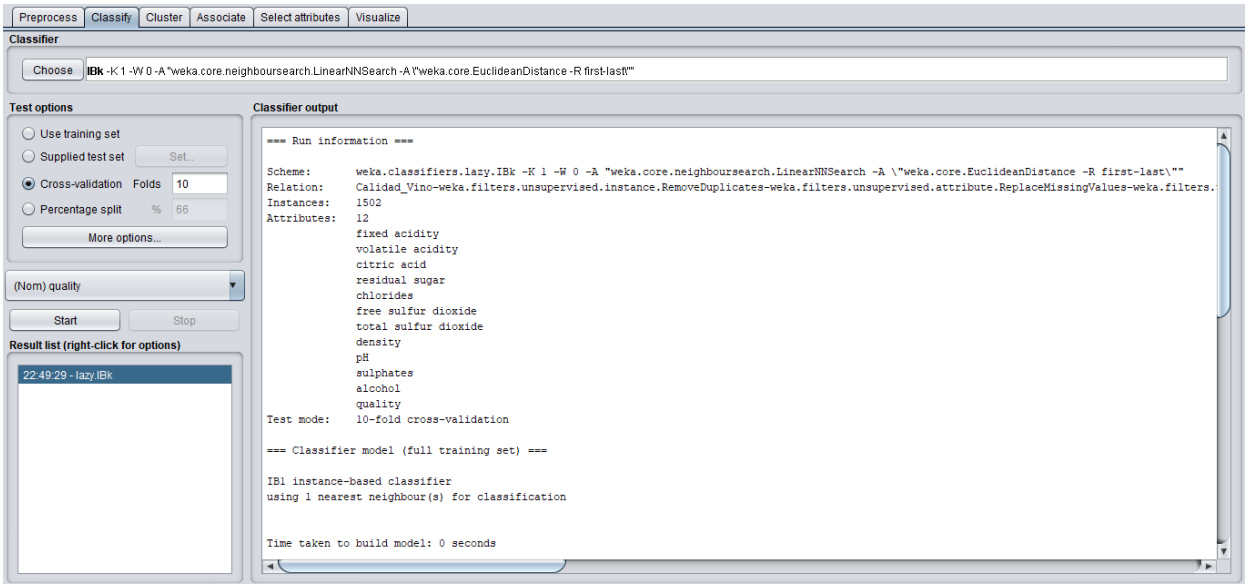
```
=== Confusion Matrix ===

   a   b   c   d   e   f   <-- classified as
   1   0  31  15   0   0 |   a = 3
   0   0  33  19   1   0 |   b = 4
   3   0 440 142   4   0 |   c = 5
   1   1 173 330  39   5 |   d = 6
   0   0   8  99  52  10 |   e = 7
   0   0   0   6  10  52 |   f = 8
```

- REDES NEURONALES:

```
=== Confusion Matrix ===

   a   b   c   d   e   f   <-- classified as
   0   1  32  14   0   0 |   a = 3
   0   0  36  16   1   0 |   b = 4
   4   3 420 156   6   0 |   c = 5
   0   6 188 300  43  12 |   d = 6
   0   4  13  95  39  18 |   e = 7
   0   0   1  19  19  29 |   f = 8
```



- KNN:




```
=== Stratified cross-validation ===
=== Summary ===

Correctly Classified Instances         745               50.5085 %
Incorrectly Classified Instances       730               49.4915 %
Kappa statistic                          0.2816
Mean absolute error                      0.1655
Root mean squared error                  0.4053
Relative absolute error                 72.4599 %
Root relative squared error            119.9862 %
Total Number of Instances             1475
Ignored Class Unknown Instances                27

=== Detailed Accuracy By Class ===

               TP Rate  FP Rate  Precision  Recall   F-Measure  MCC      ROC Area  PRC Area  Class
               0,383    0,022    0,367      0,383    0,375      0,354    0,652     0,139     3
               0,075    0,023    0,111      0,075    0,090      0,064    0,541     0,045     4
               0,560    0,269    0,581      0,560    0,570      0,294    0,647     0,509     5
               0,468    0,313    0,470      0,468    0,469      0,155    0,574     0,414     6
               0,444    0,087    0,397      0,444    0,419      0,340    0,674     0,241     7
               0,897    0,018    0,709      0,897    0,792      0,787    0,942     0,624     8
Weighted Avg.  0,505    0,236    0,501      0,505    0,502      0,264    0,633     0,420
```

```
=== Confusion Matrix ===

   a   b   c   d   e   f   <-- classified as
  18   0  17  12   0   0 |   a = 3
   3   4  25  18   3   0 |   b = 4
  15  20 330 189  34   1 |   c = 5
  13  11 179 257  74  15 |   d = 6
   0   1  16  68  75   9 |   e = 7
   0   0   1   3   3  61 |   f = 8
```

CONCLUSIÓN:

- Predicción:

METODO	MEAN ABSOLUTE ERROR	ROOT MEAN SQUARED ERROR
Regresión Lineal	0.5588	0.7197
SVM	0.5133	0.6811
Random Forest	0.4901	0.6371
Redes Neuronales	0.6503	0.8812

Para los métodos de predicción aplicados el que presenta un mayor Root mean squared error y mean absolute error son las redes neuronales y el menor es Random Forest. Se puede concluir que el mejor método usado para la predicción de la calidad de vino es el Random Forest, ya que presenta valores de error menores a los demás.

- Clasificación:

METODO	MEAN ABSOLUTE ERROR	ROOT MEAN SQUARED ERROR
Regresión Logística	0.1971	0.3368
Naive-Bayes	0.1971	0.3368
SVM	0.2383	0.3352
Decision tree	0.1763	0.377
Random Forest	0.1779	0.2977
Redes Neuronales	0.1817	0.3216
KNN	0.1655	0.4653

Para los métodos de clasificación aplicados el que presenta un mayor mean absolute error es SVM y con el mayor Root mean squared error es KNN; el menor mean absolute error es KNN y el menor Root mean squared error es Random Forest. Se puede concluir que el mejor método usado para la predicción de la calidad de vino es el Random Forest, ya que presenta valores de error menores a los demás.