S3. Data Mining

Caio Fernandes Moreno

POLITÉCNICA

"Ingeniamos el futuro"

Problem Description

- A complex modern **semi-conductor manufacturing process** is normally under consistent surveillance via the monitoring of signals/variables collected from sensors and or process measurement points.

- However, not all of these signals are equally valuable in a specific monitoring system. The measured signals contain **a combination of useful information, irrelevant information as well as noise**.

Problem Description

- Engineers typically have a much larger number of signals than are actually required. **If we consider each type of signal as a feature**, then feature selection may be applied to identify the **most relevant signals**.

- The Process Engineers may then use these signals to determine key factors contributing to yield excursions downstream in the process. This will enable an increase in process throughput, decreased time to learning and reduce the per unit production costs.

SECOM Dataset

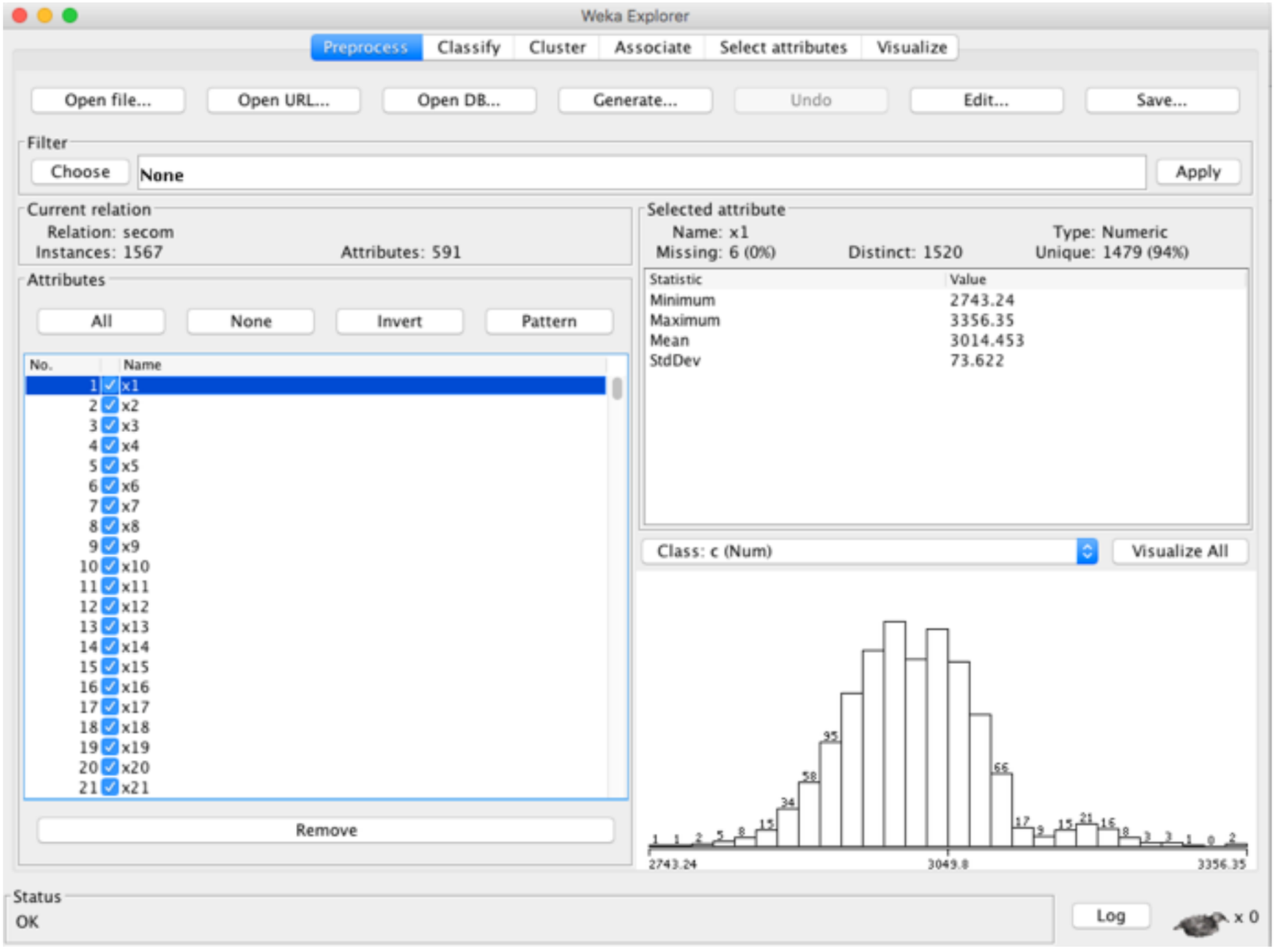
- SECOM Dataset: 1567 examples 591 features, 104 fails

- There are missing values;

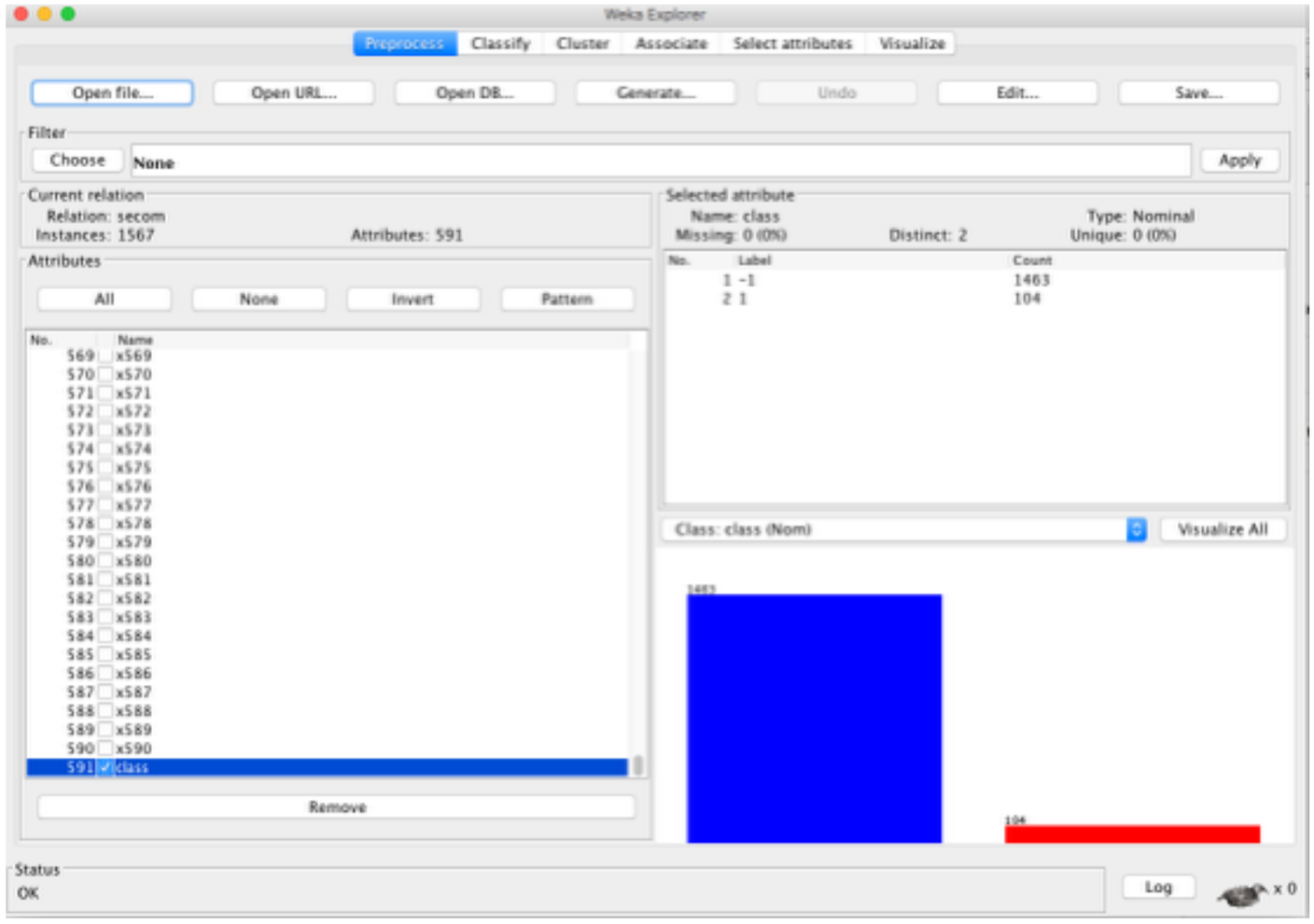
- Where –1 corresponds to a pass and 1 corresponds to a fail and the data time stamp is for that specific test point.

- https://archive.ics.uci.edu/ml/datasets/SECOM

Features (variables)



ARFF File



The variable **class** is the binary.

@attribute class {-1,1}

-1 = Pass

1 = Fail

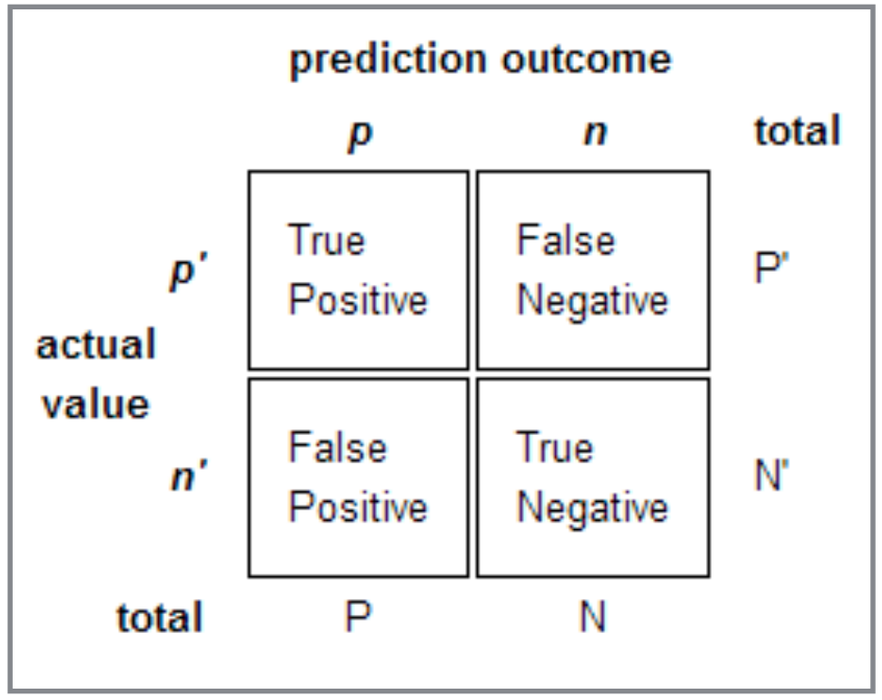
```
@attribute x588 numeric
@attribute x589 numeric
@attribute x590 numeric
@attribute class {-1,1}

@data
3030.93,2564,2187.7333,1411.1265,1.3602,100,97.61
455,202.4396,0,7.9558,414871,10.0433,968,192.3963
```

Algorithms used

- Naive Bayes
- TAN
- IB1
- Idk
- RIPPER
- ID3
- C4.5 (J48)
- Logistic

Measures

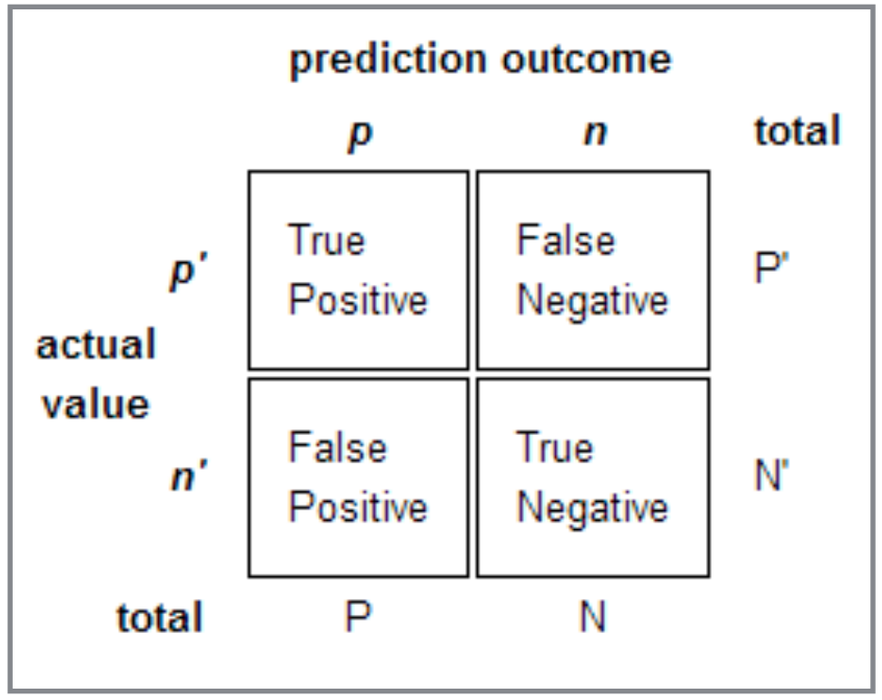
```
=== Stratified cross-validation ===
=== Summary ===

Correctly Classified Instances        1012               64.582  %
Incorrectly Classified Instances       555               35.418  %
Kappa statistic                          0.0077
Mean absolute error                      0.3534
Root mean squared error                  0.59
Relative absolute error                283.9367 %
Root relative squared error            237.0124 %
Total Number of Instances             1567

=== Detailed Accuracy By Class ===

               TP Rate   FP Rate   Precision   Recall  F-Measure   ROC Area  Class
                 0.666     0.644      0.936     0.666     0.778      0.505    -1
                 0.356     0.334      0.07      0.356     0.118      0.497    1
Weighted Avg.    0.646     0.624      0.878     0.646     0.735      0.505

=== Confusion Matrix ===

   a   b   <-- classified as
 975 488 |   a = -1
  67  37 |   b = 1
```



It is possible to see that 975 are True Positive, 488 False Negative, 67 False Positive and 37 True Negative.

Naive Bayes with all variables.

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

Correctly Classified Instances        1012               64.582  %
Incorrectly Classified Instances       555               35.418  %
Kappa statistic                          0.0077
Mean absolute error                      0.3534
Root mean squared error                  0.59
Relative absolute error                283.9367 %
Root relative squared error            237.0124 %
Total Number of Instances             1567

=== Detailed Accuracy By Class ===

               TP Rate   FP Rate   Precision   Recall  F-Measure   ROC Area  Class
                 0.666     0.644      0.936     0.666     0.778      0.505    -1
                 0.356     0.334      0.07      0.356     0.118      0.497    1
Weighted Avg.    0.646     0.624      0.878     0.646     0.735      0.505

=== Confusion Matrix ===

   a   b   <-- classified as
 975 488 |   a = -1
  67  37 |   b = 1
```



It is possible to see that 975 are True Positive, 488 False Negative, 67 False Positive and 37 True Negative.
64.58% Correctly Classified Instances.

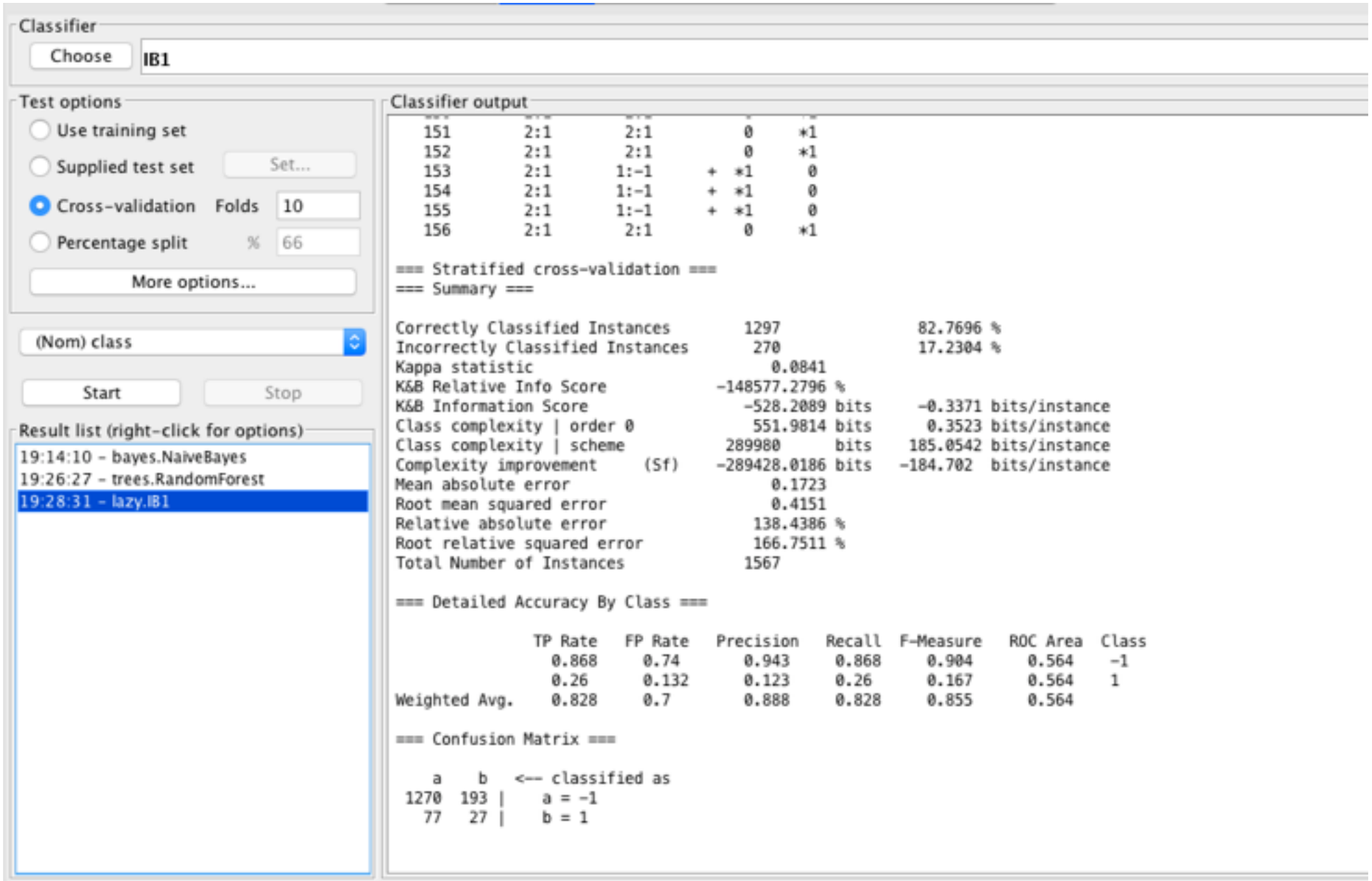
TAN with all variables.

Where is it?

% Correctly Classified Instances.

Caio Fernandes Moreno

IB1 with all variables.



82.76% Correctly Classified Instances.

Caio Fernandes Moreno

Idk with all variables.

Where is it?

% Correctly Classified Instances.

Caio Fernandes Moreno

RIPPER with all variables.

Where is it?

% Correctly Classified Instances.

Caio Fernandes Moreno

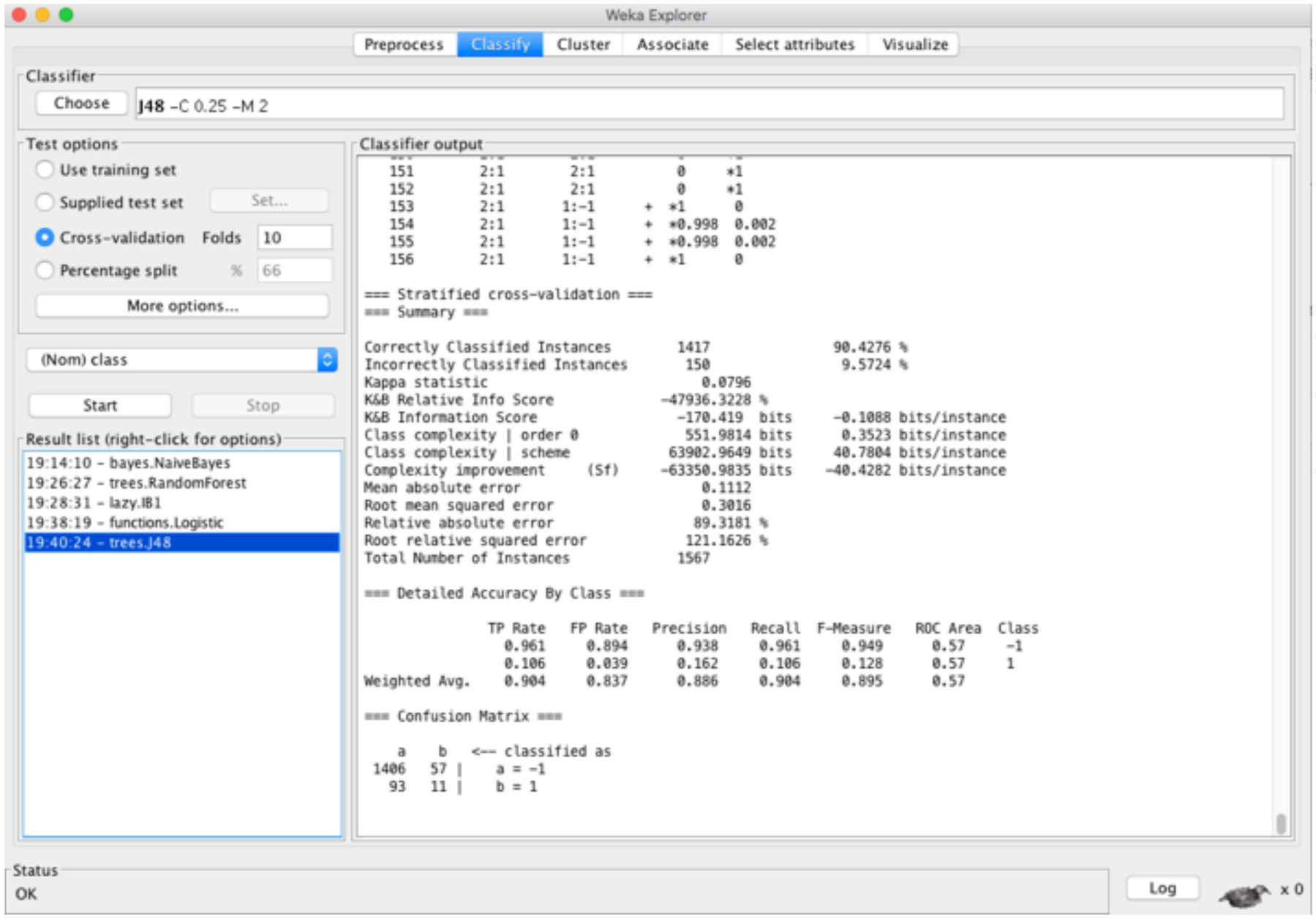
ID3 with all variables.

Where is it?

% Correctly Classified Instances.

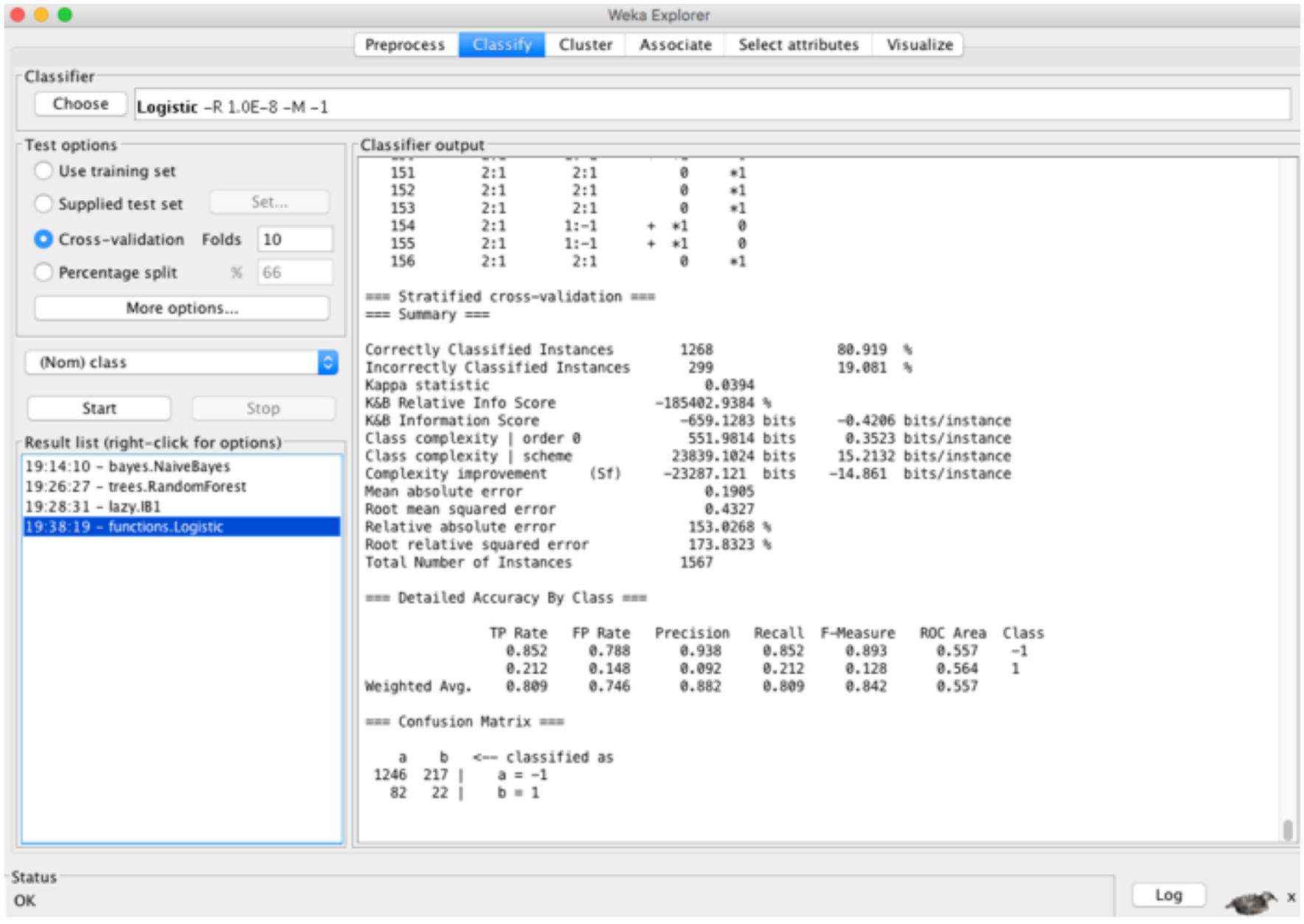
Caio Fernandes Moreno

C4.5 (J48) with all variables.



90.42% Correctly Classified Instances.

Logistic with all variables.



80.91% Correctly Classified Instances.

Performance

	All variables	FSS1	FSS2	Wrapper
Naive Bayes	64.58%			
TAN				
IB1	82.76%			
IBK				
RIPPER				
ID3				
C4.5 (J48)	90.42%			
Logistic	80.91%			

Conclusions

- We can see…

Thanks

- Caio Fernandes Moreno
  caio.fmoreno@alumnos.upm.es