WEKA
A Data Mining Tool

By Susan L. Miertschin

Data Mining

Task Types

- Classification
- Clustering
- Discovering Association Rules
- Discovering Sequential Patterns – Sequence Analysis
- Regression
- Detecting Deviations from Normal

Numerous Algorithms

- C4.5 Decision Tree
- K-Means Clustering

http://www.cs.waikato.ac.nz/ml/weka/

WEKA can be freely downloaded by visiting the Web site

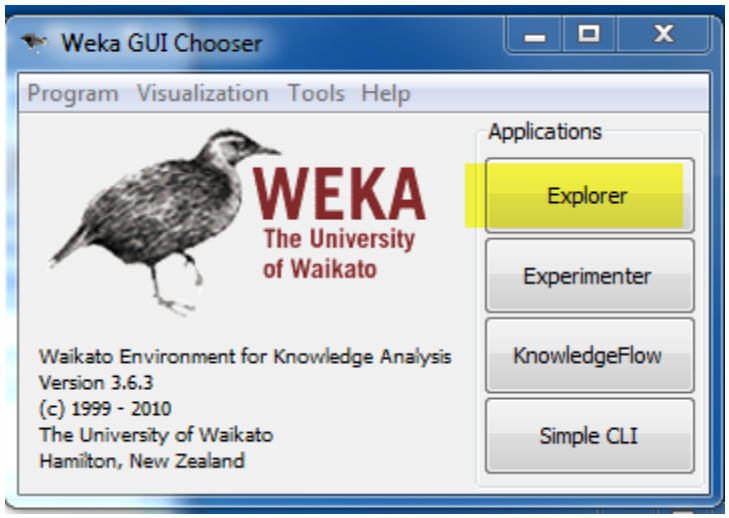
WEKA – Data Mining Software

- Developed by the Machine Learning Group, University of Waikato , New Zealand
- Vision: Build state-of-the-art software for developing machine learning (ML) techniques and apply them to real-world data-mining problems
- Developed in Java

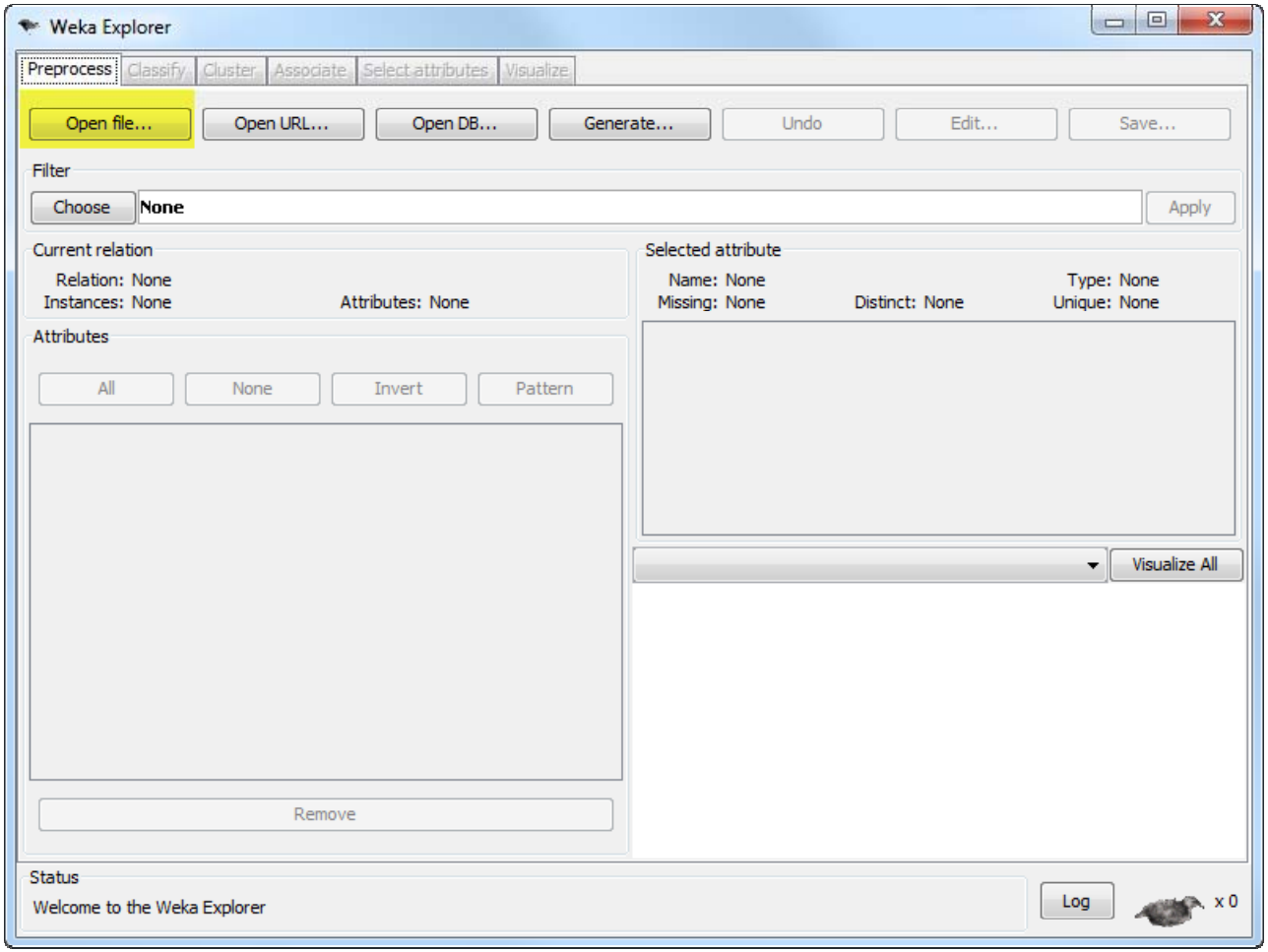
WELA's Collection of Machine Learning Algorithms

- Algorithms for data mining tasks
- WEKA is open source software issued under the GNU General Public License
- Tools for:
  - Data pre-processing
  - Classification
  - Regression
  - Clustering
  - Association rules
  - Visualization

After Installing - Start WEKA



WEKA Main Interface



WEKA Sample Files

- C:\Program Files\weka\data
- WEKA formatted files (.arff)
- Open the contact-lenses file

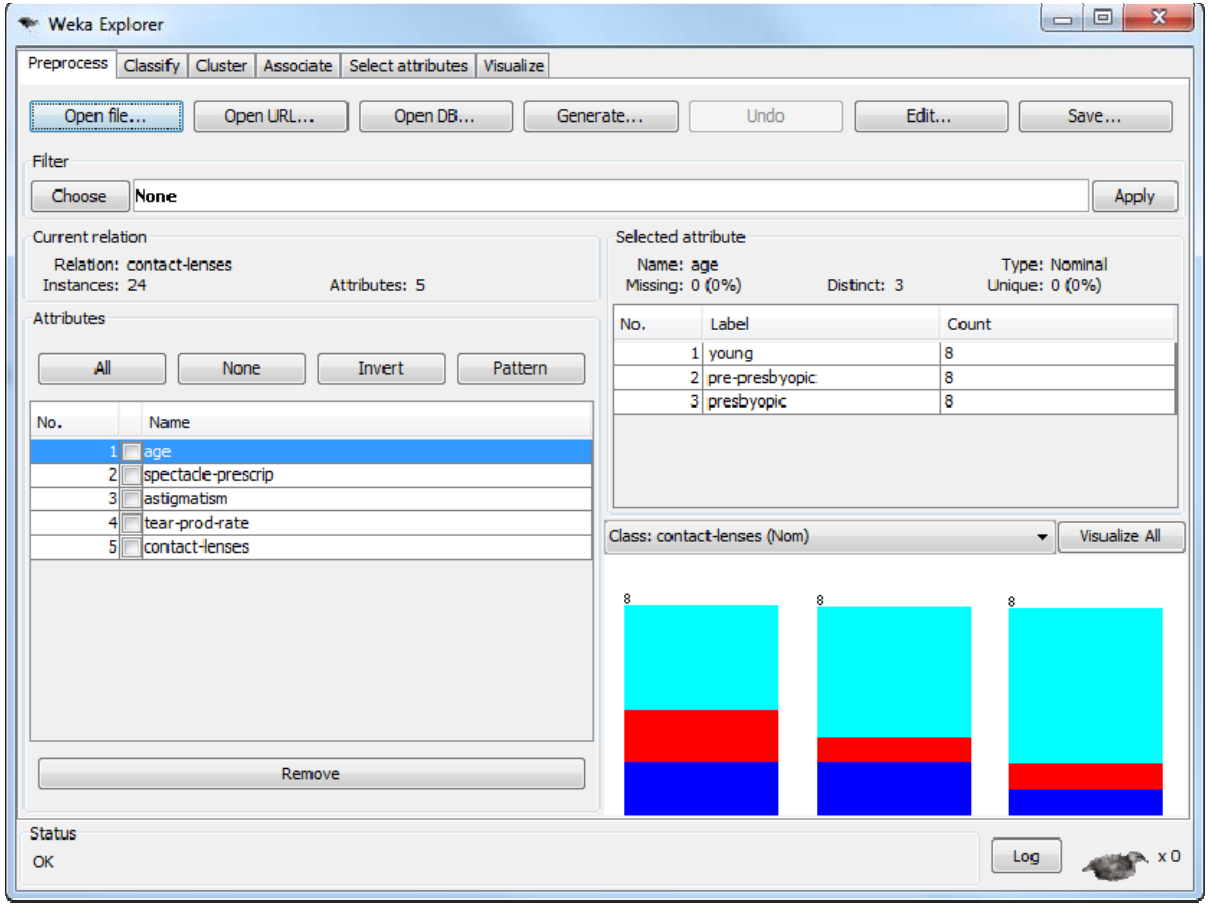
Example – Contact Lens Data

How many data instances are in the file?

How many attributes?

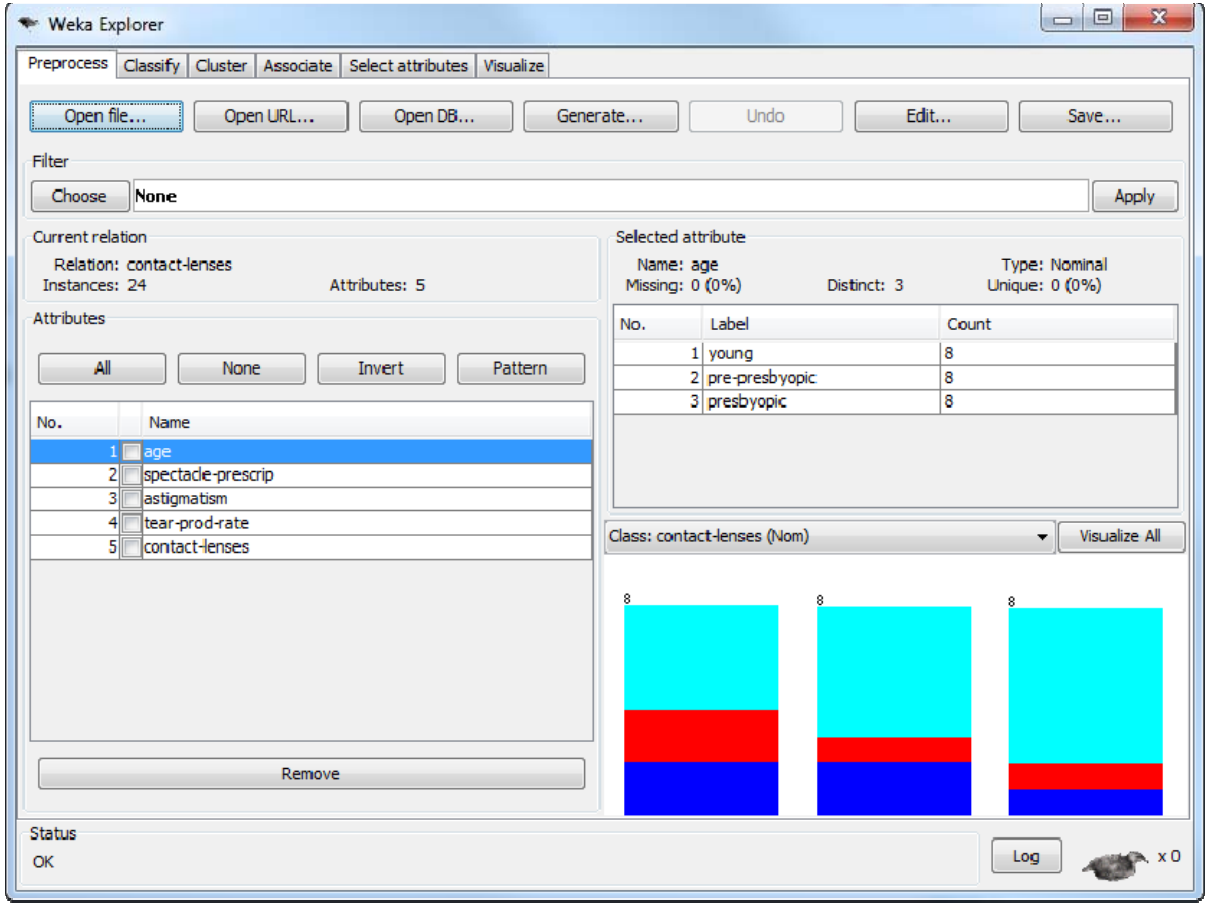
Numerical attributes?

Categorical attributes?



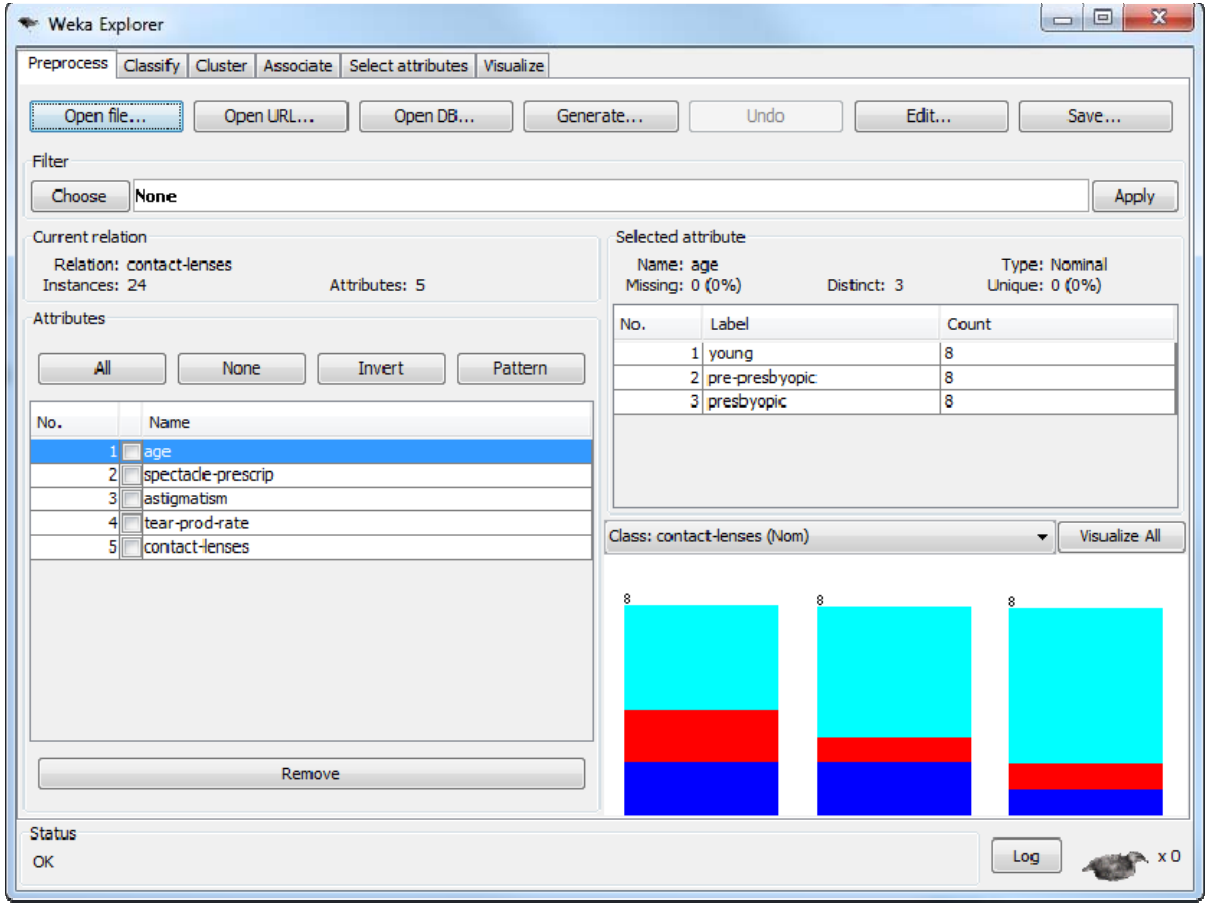
Example – Contact Lens Data

Can you think of problems that might be solved with this data?

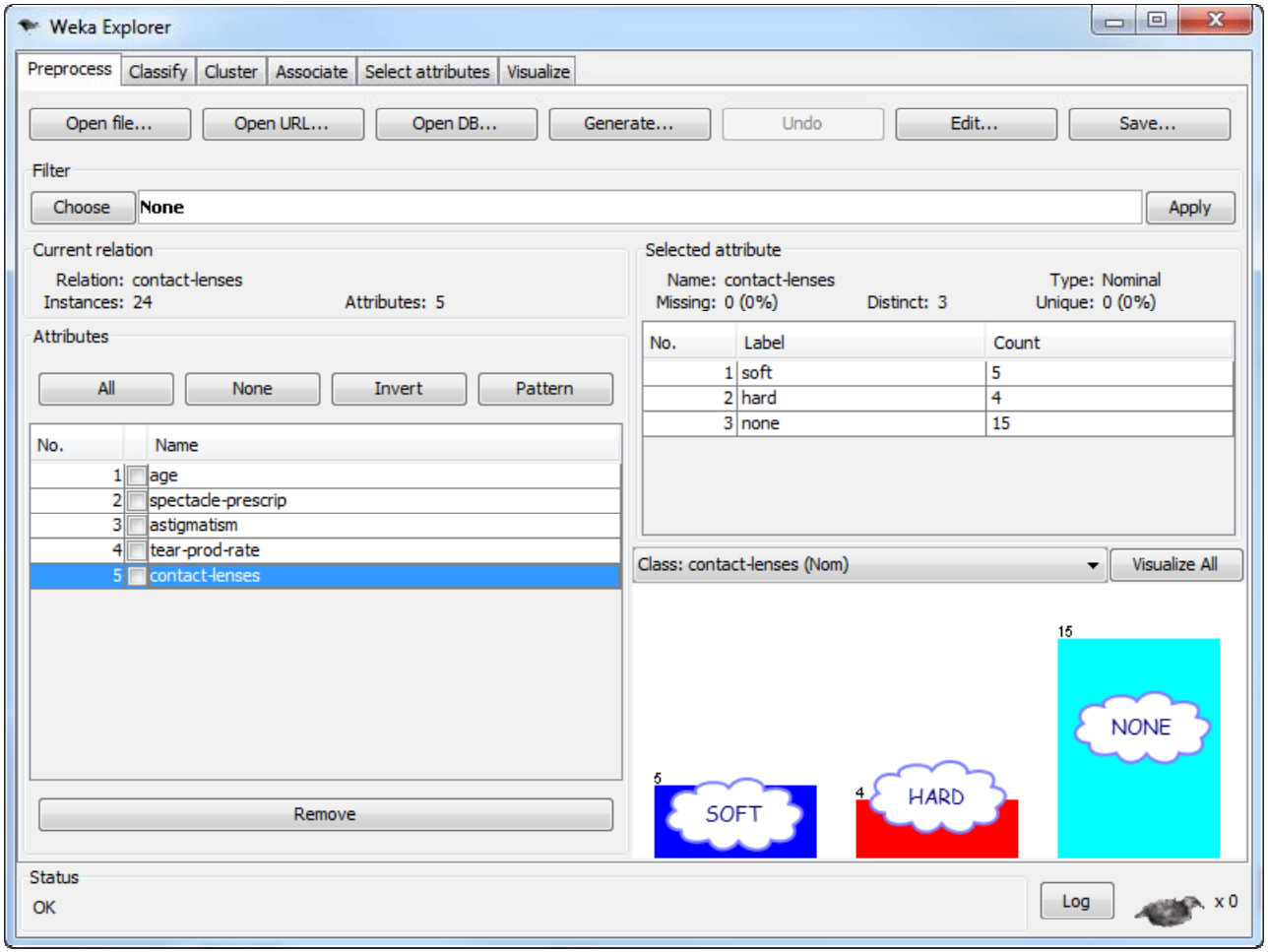


Example – Contact Lens Data

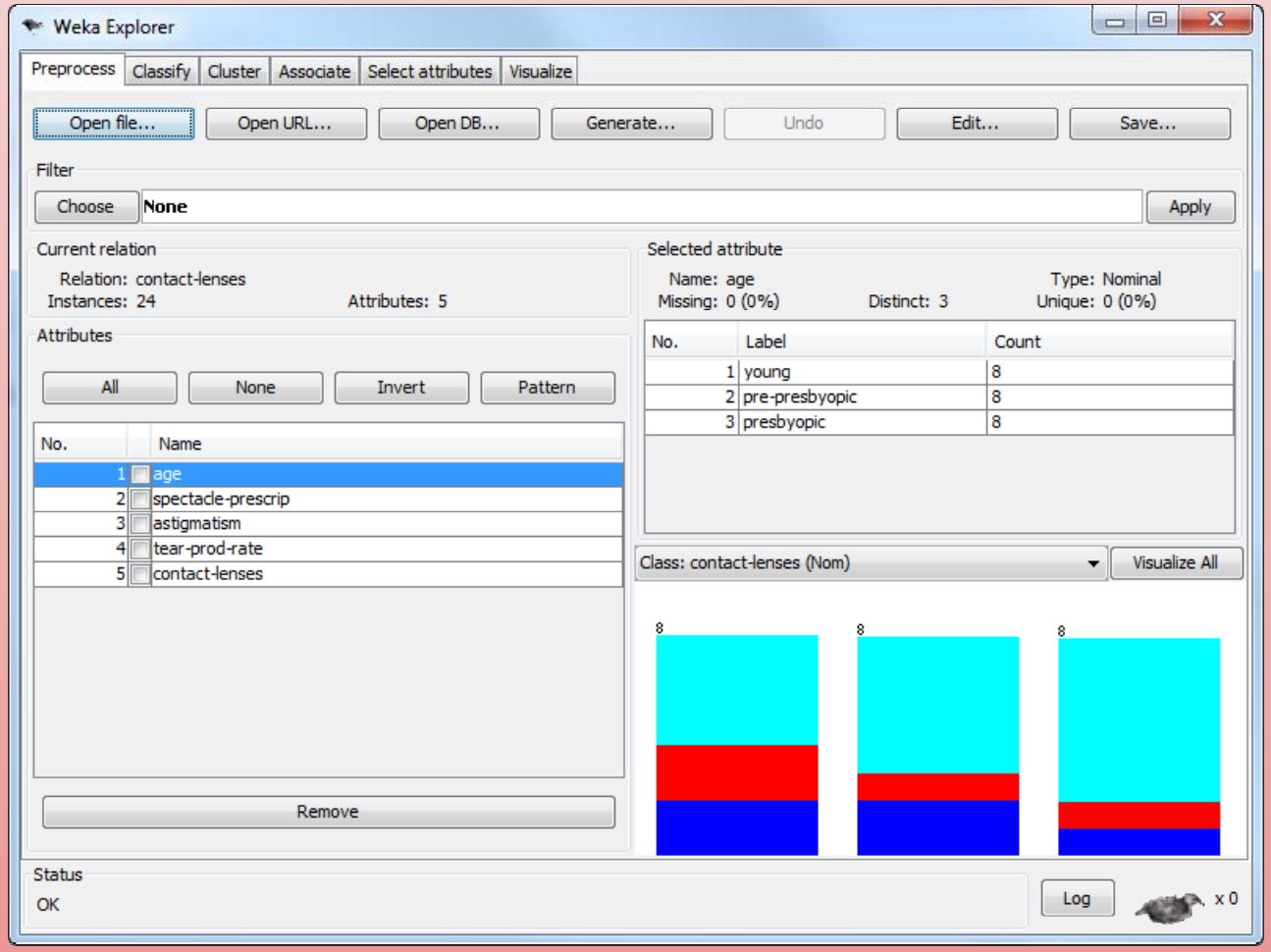
If supervised learning were to be done, which would be the output attribute, do you think?



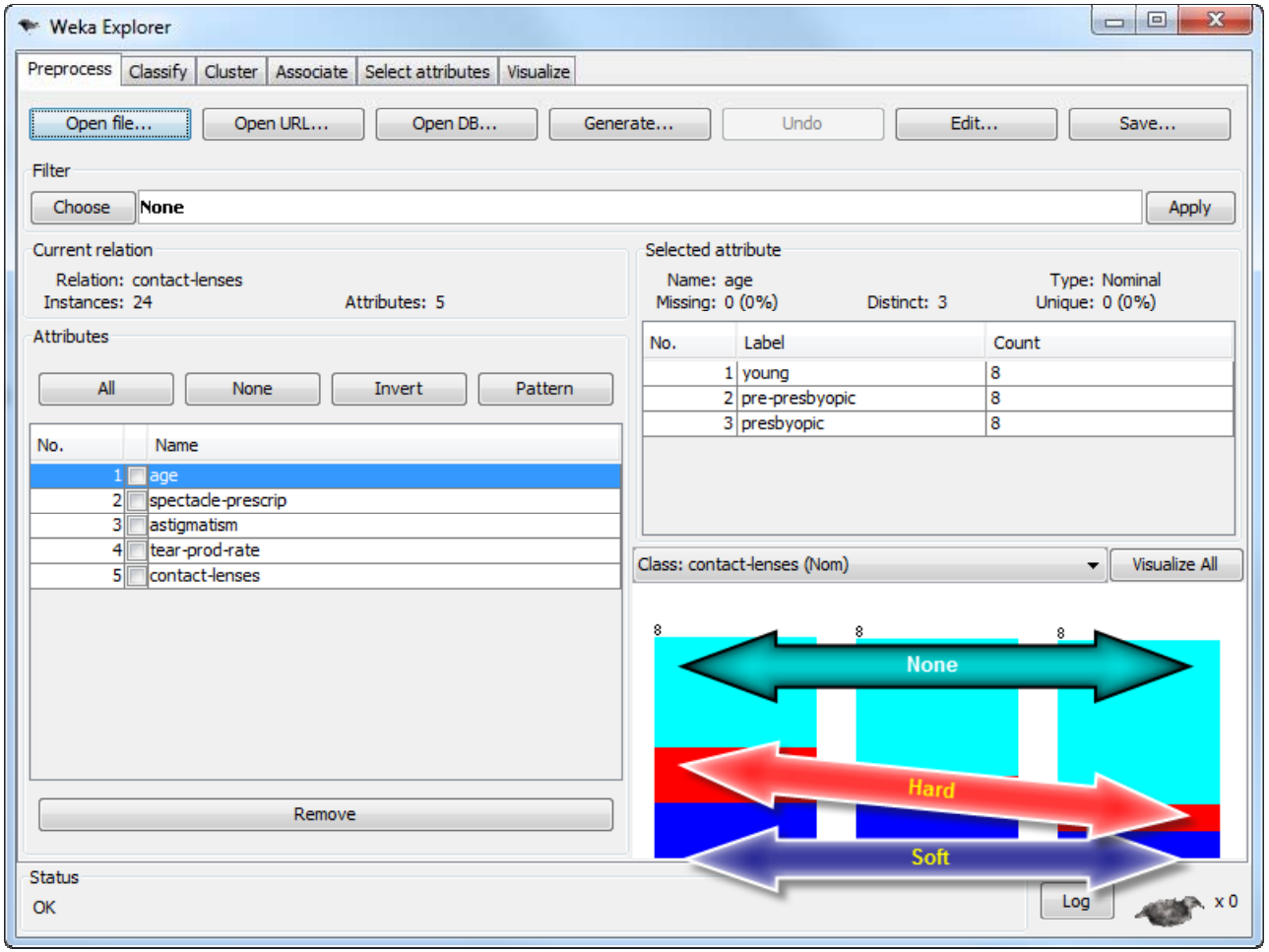
Example – Contact Lens Data



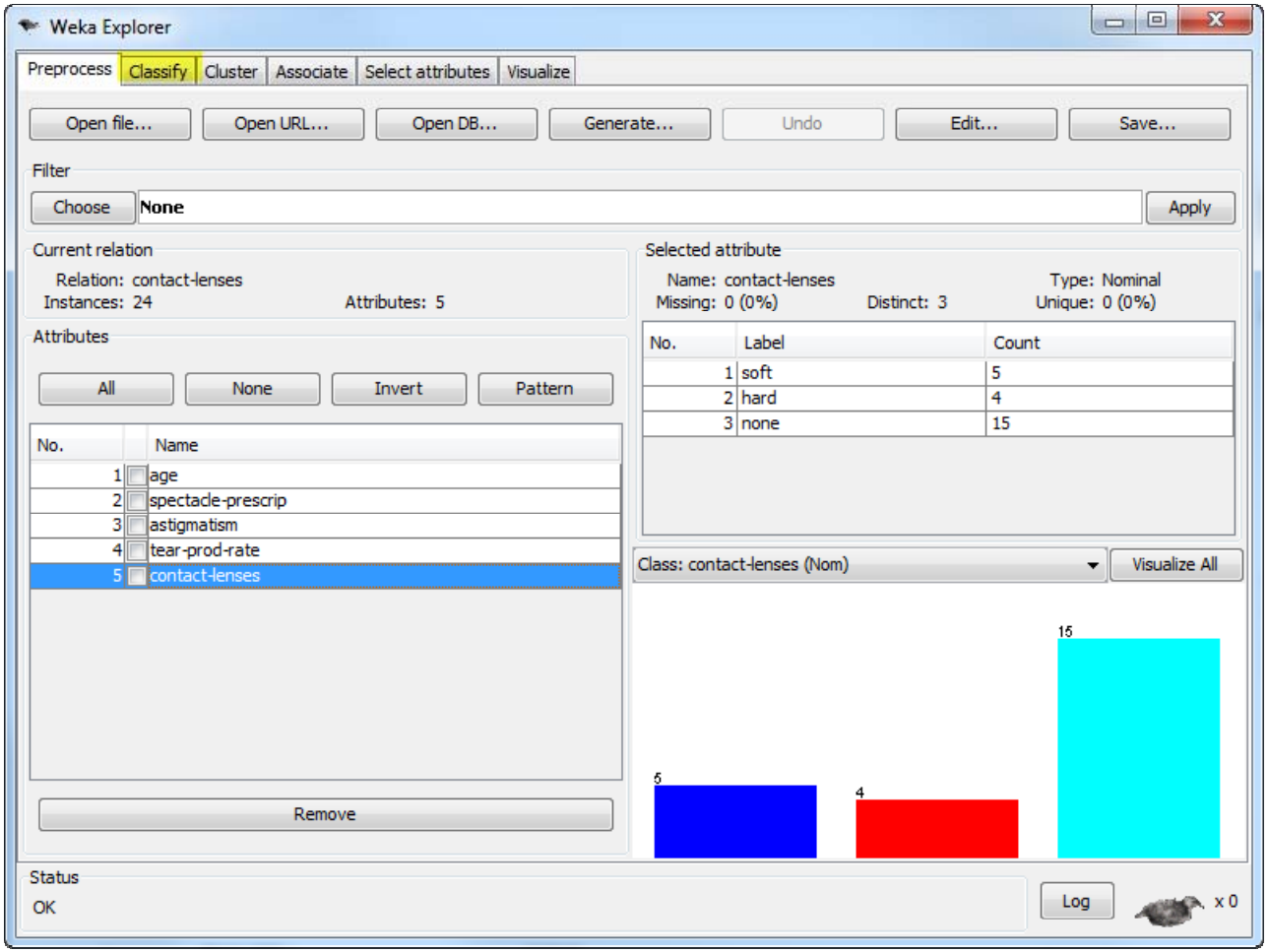
Example – Contact Lens Data



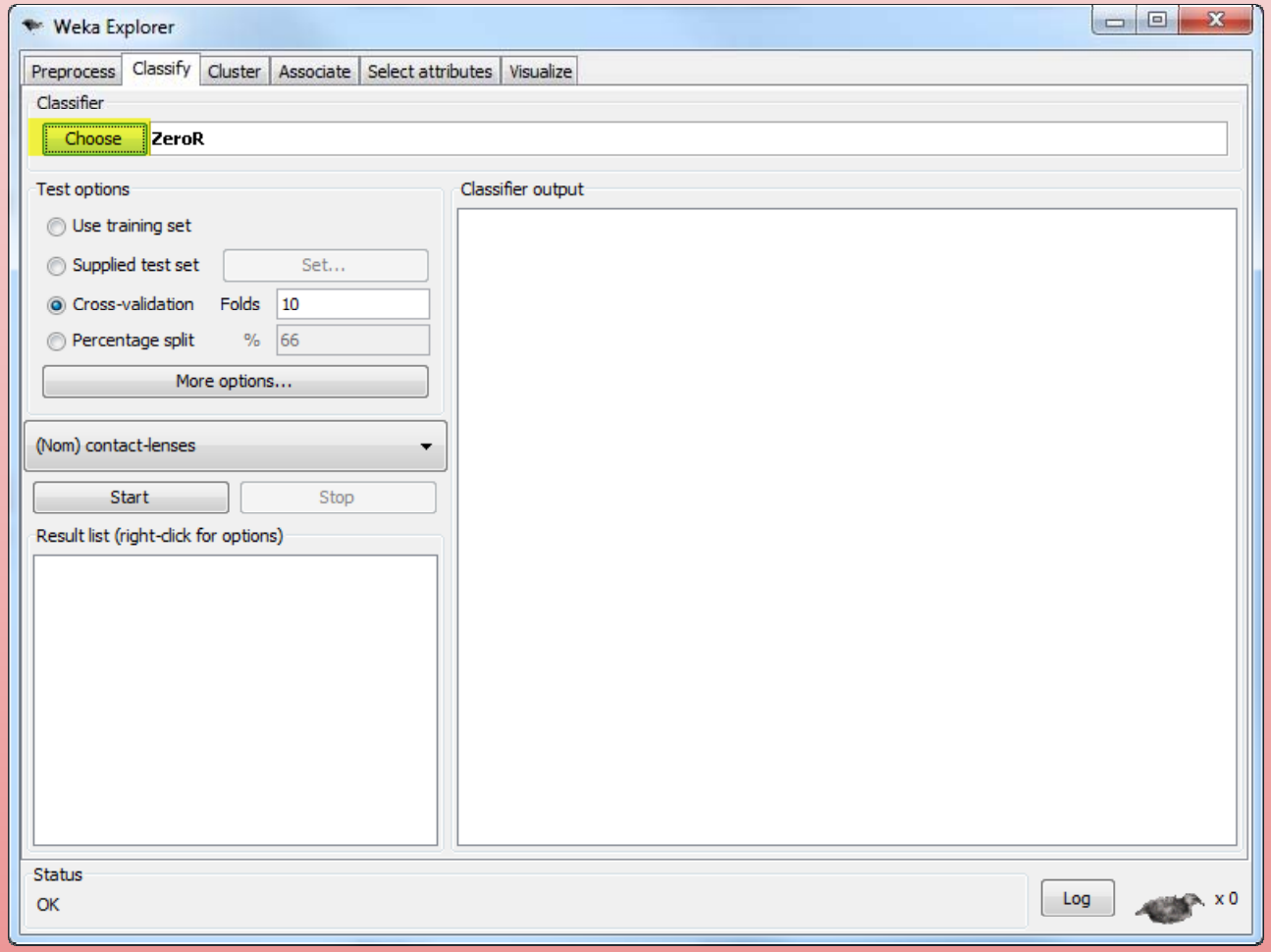
Example – Contact Lens Data



Example – Classify - Contact Lens Data

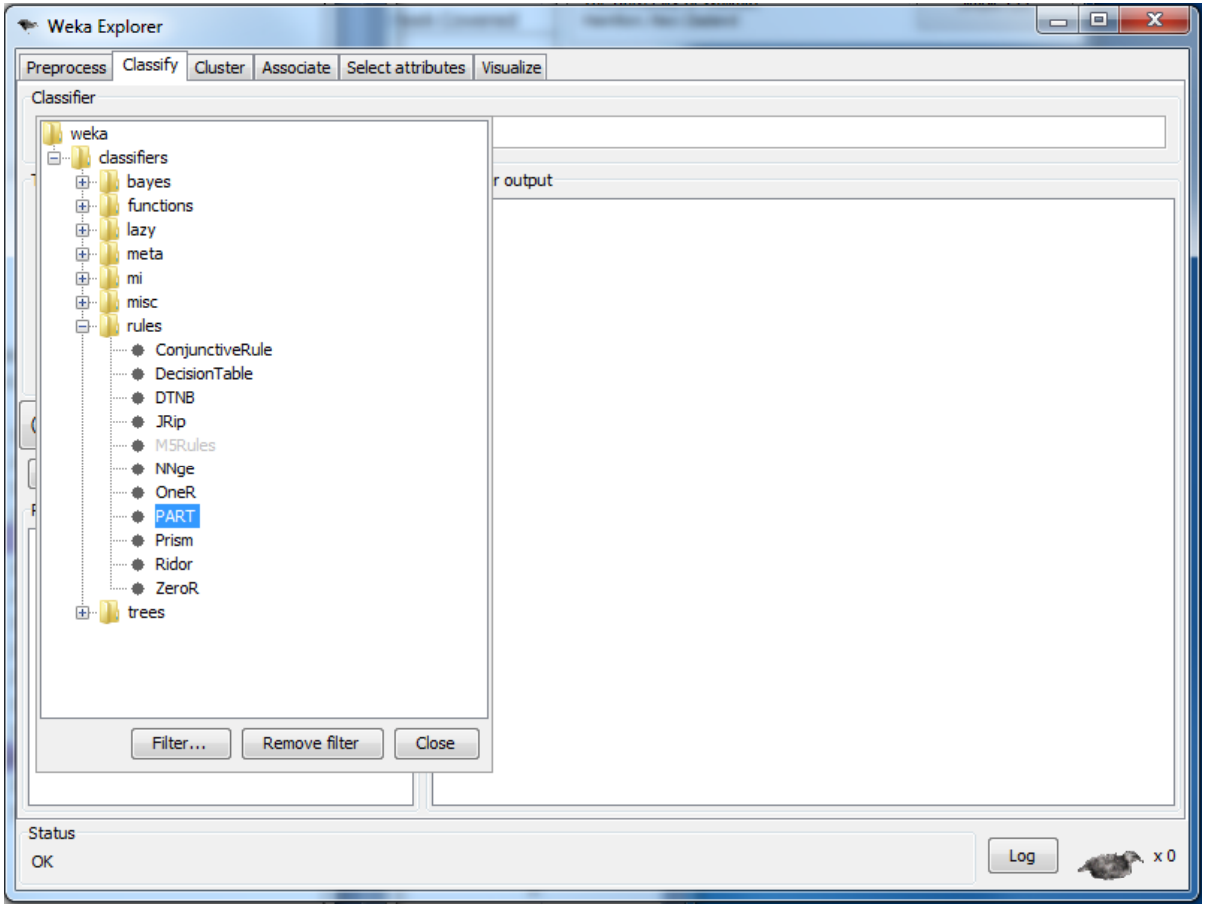


Example – Classify - Contact Lens Data

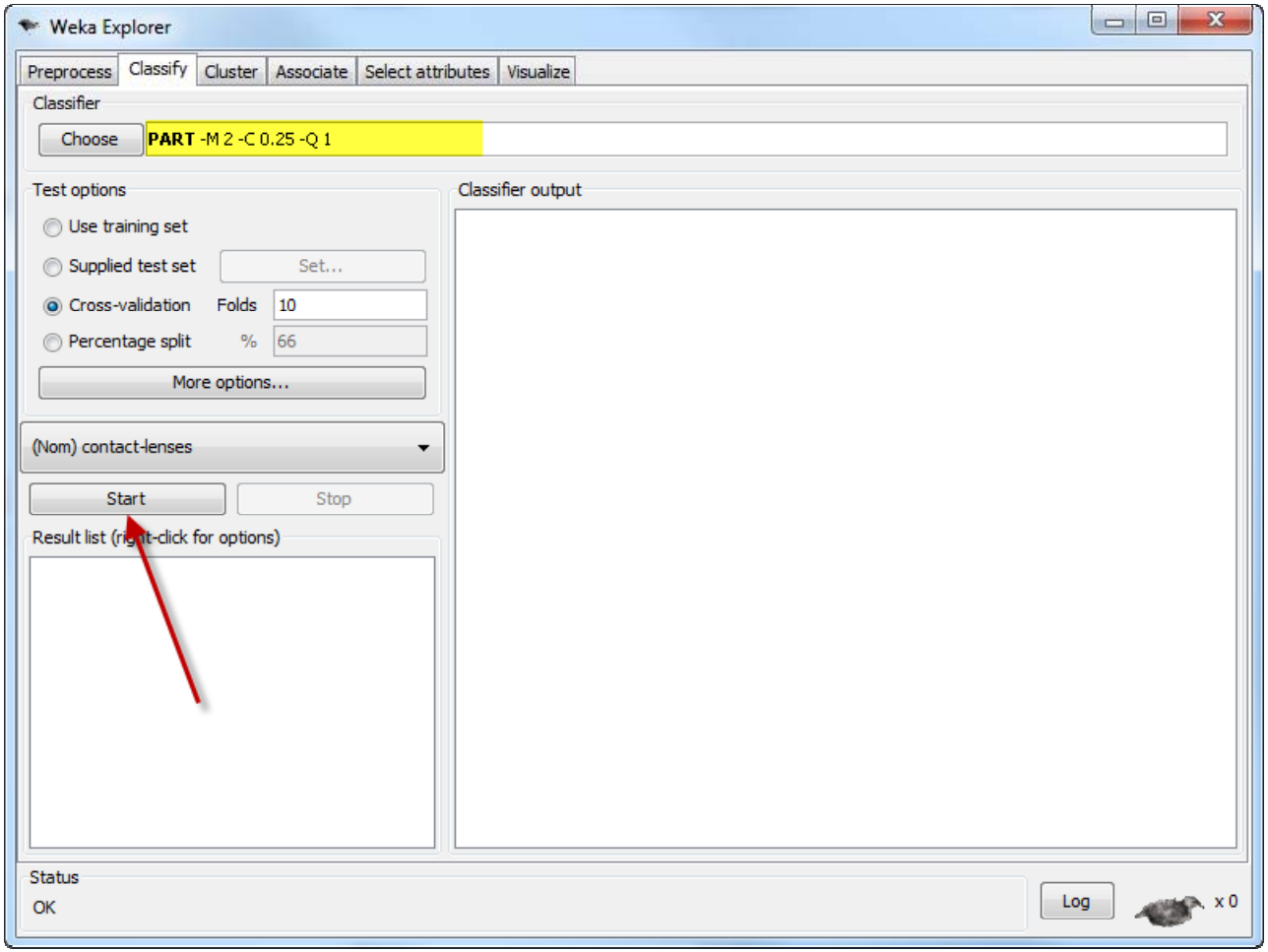


Example – Classify - Contact Lens Data

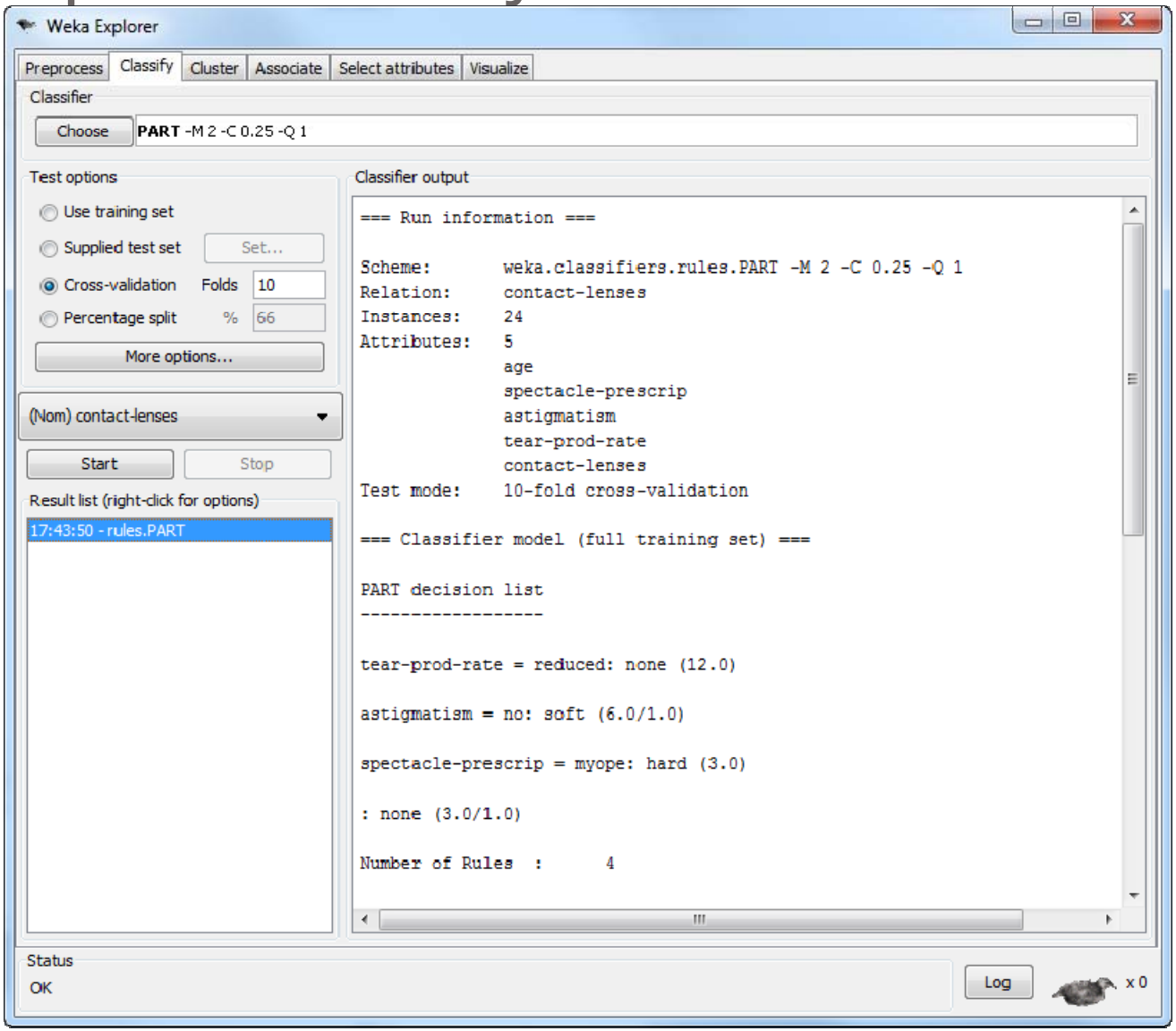
Select the rule generator named PART from the list that shows up after you select Choose



Example – Classify - Contact Lens Data



Example – Classify - Contact Lens Data

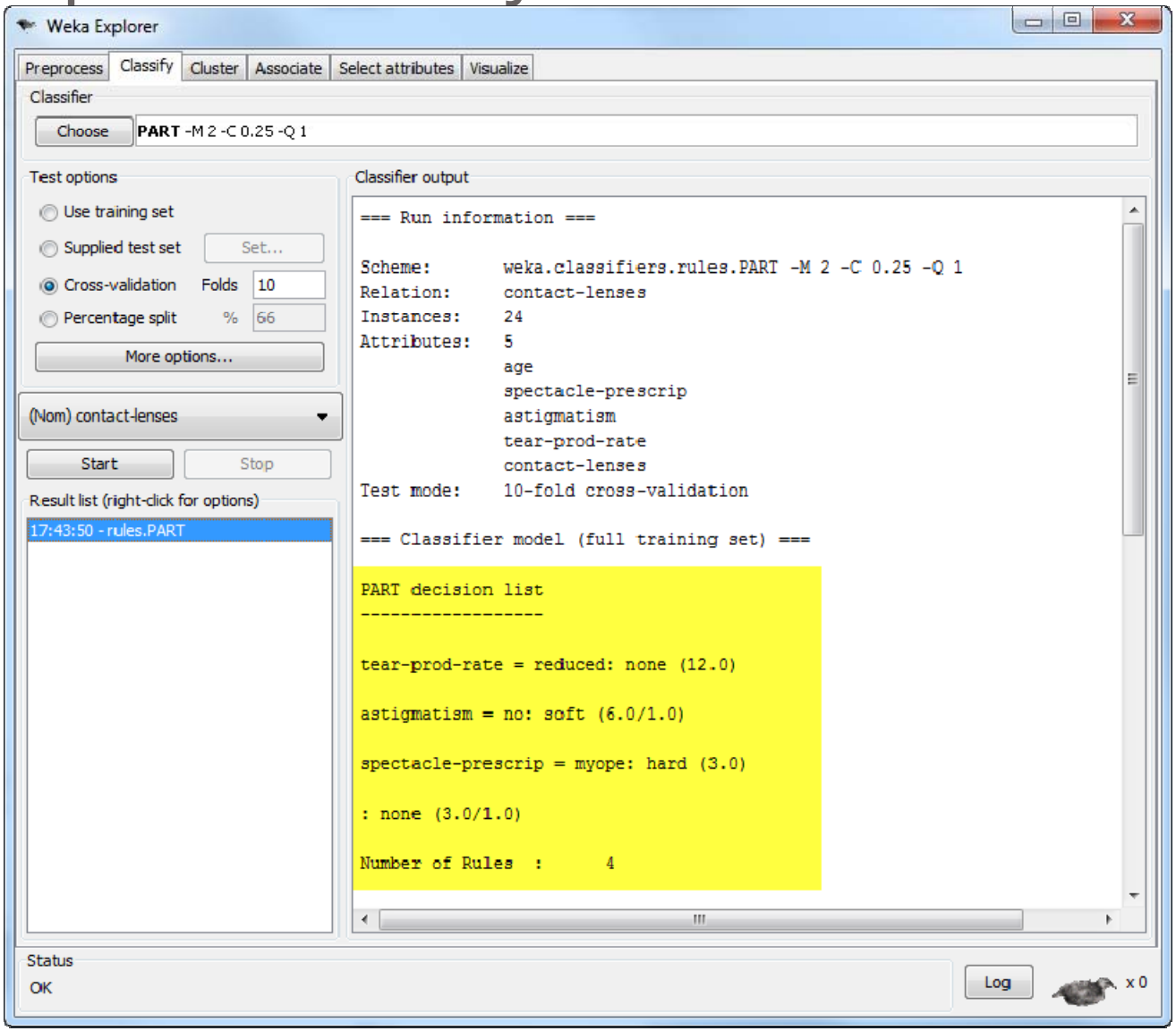


10-Fold Cross-Validation

- Data is partitioned into 10 equally (or nearly equally) sized segments or folds
- 10 iterations of training and validation are completed
- In each iteration a different fold of the data is held out for validation, with the remaining 9 folds used for learning

http://www.public.asu.edu/~ltang9/papers/ency-cross-validation.pdf

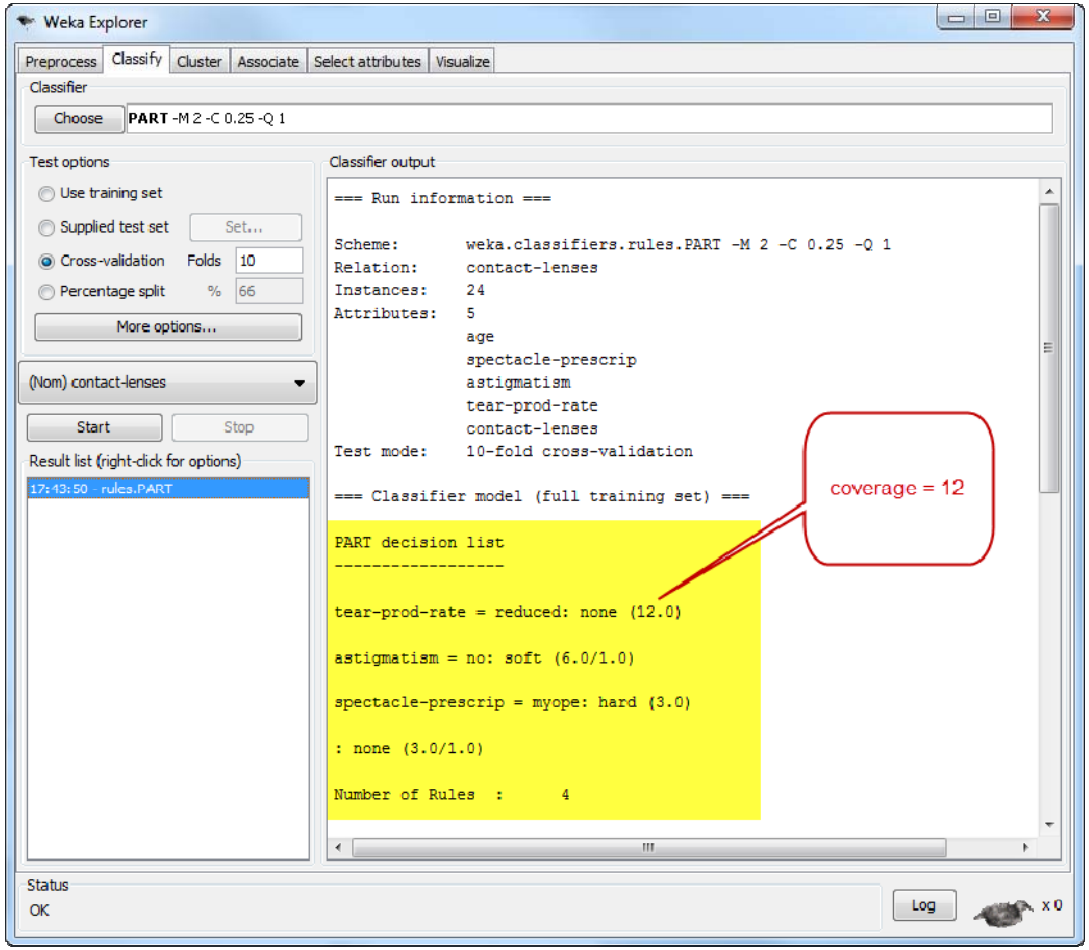
Example – Classify - Contact Lens Data



Example – Classify - Contact Lens Data

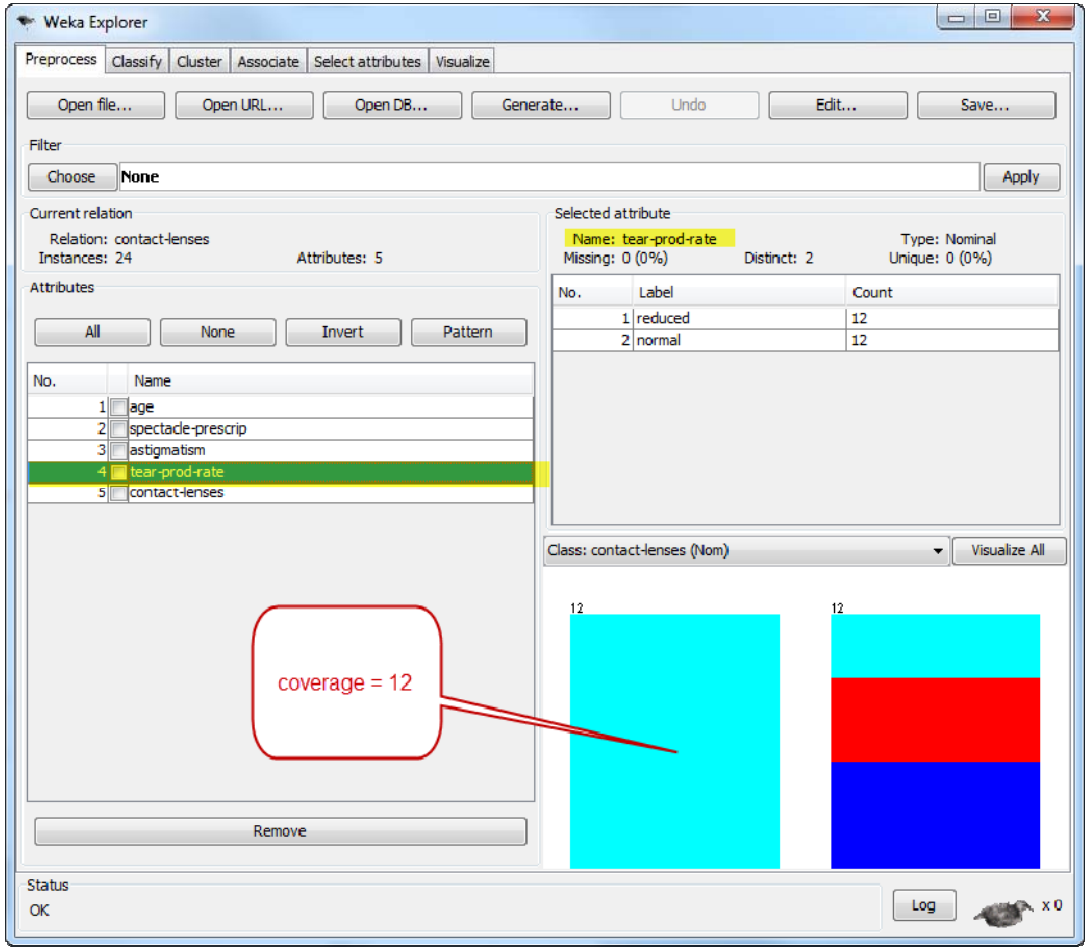
IF tear-prod-rate = reduced THEN contact-lenses = none

IF astigmatism = no THEN contact-lenses = soft



Example – Classify - Contact Lens Data

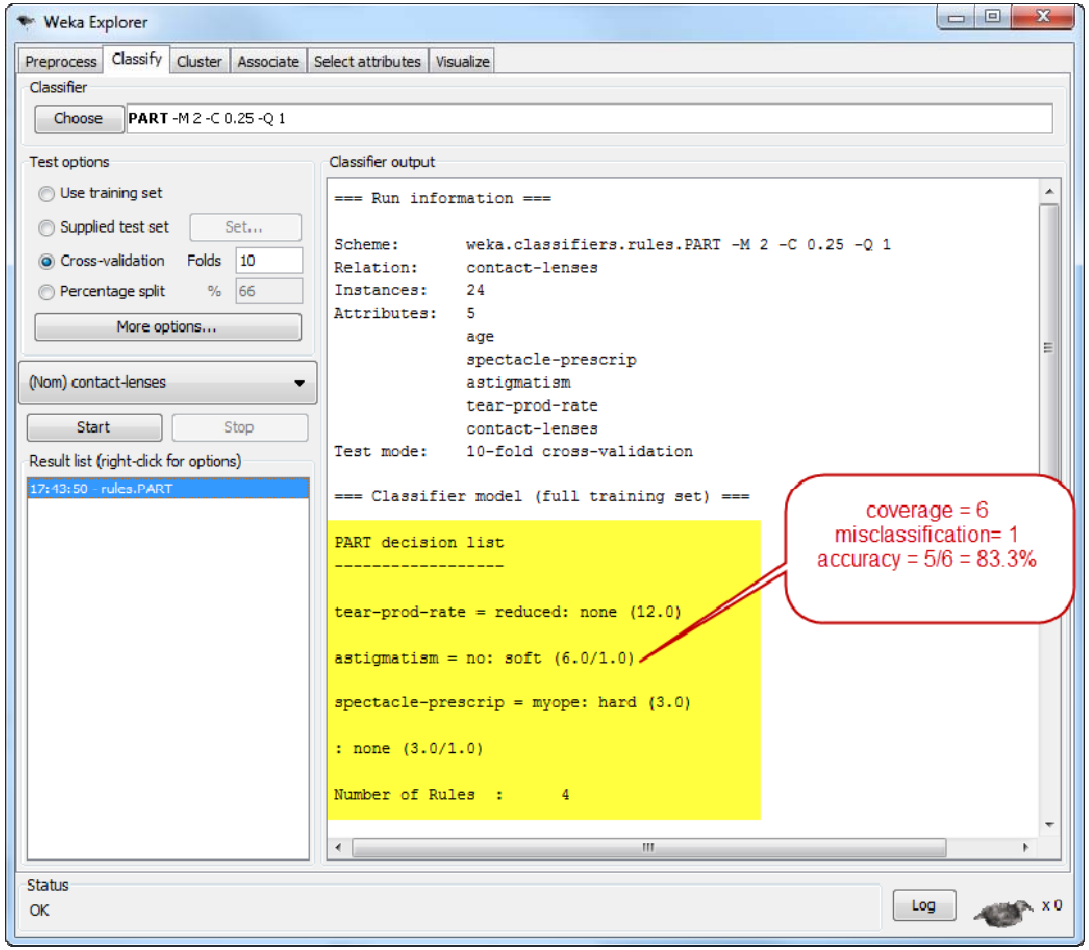
Coverage = 12



Example – Classify - Contact Lens Data

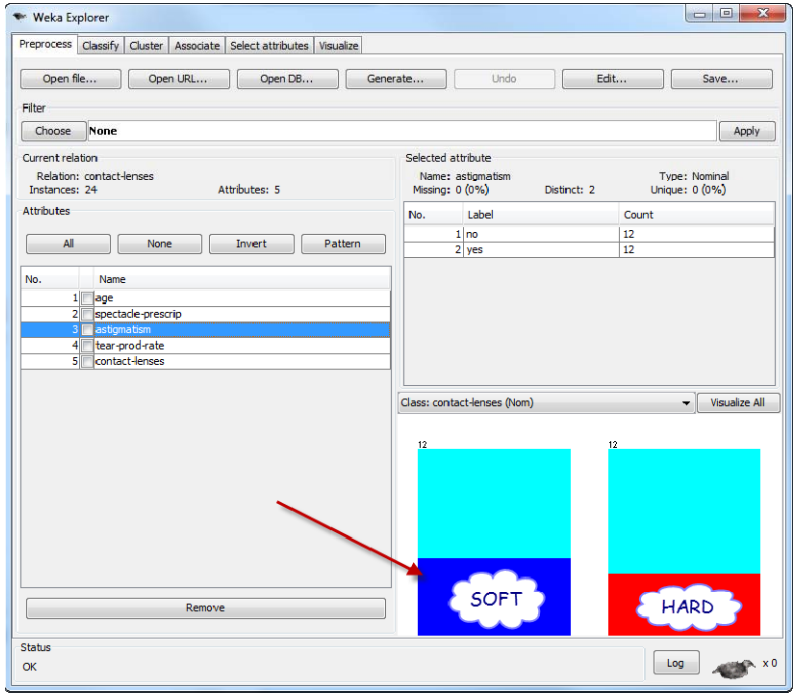
IF tear-prod-rate = reduced THEN contact-lenses = none

IF astigmatism = no THEN contact-lenses = soft

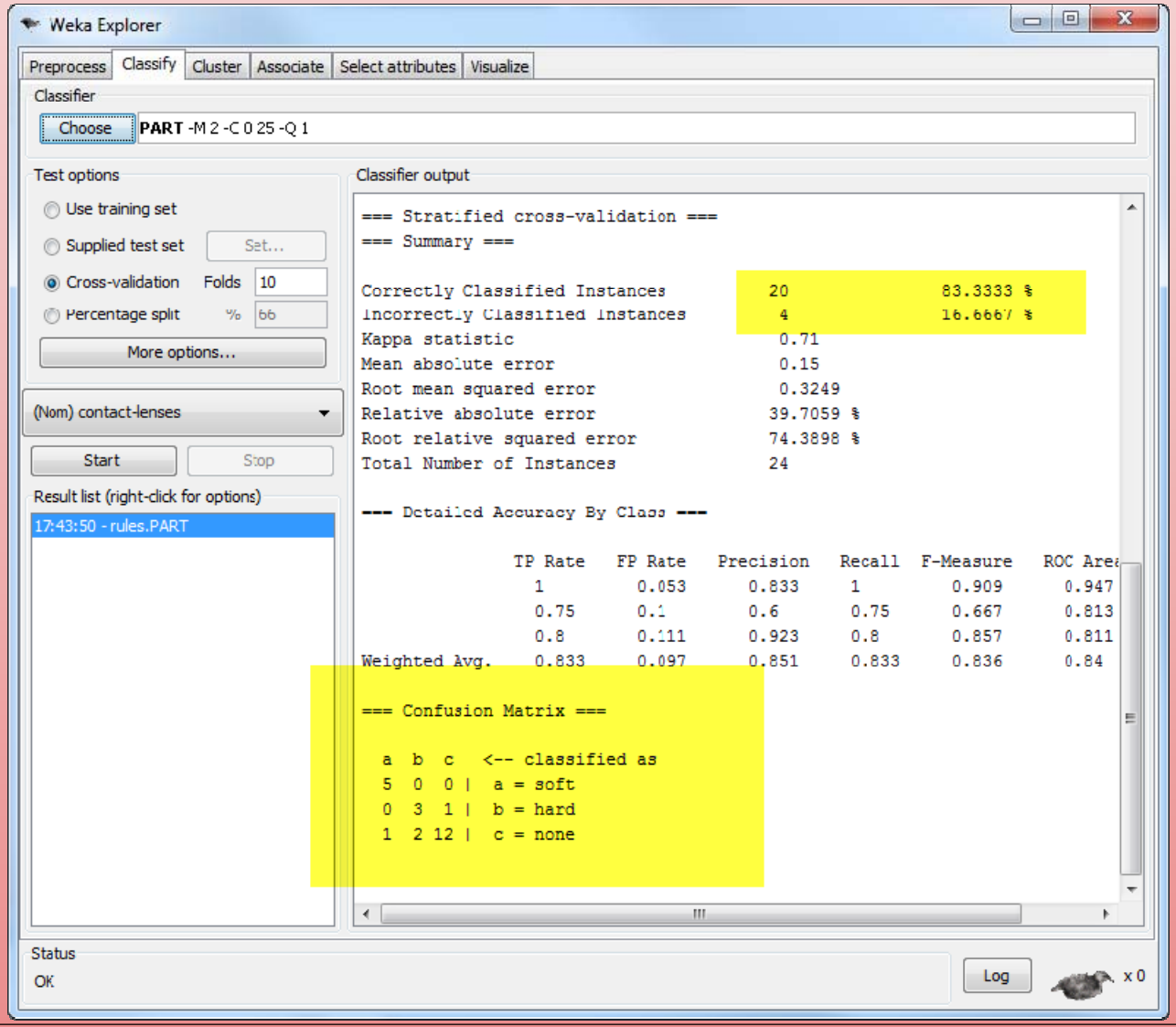


Example – Classify - Contact Lens Data

- Coverage = 6
- Misclassification = 1
- Accuracy = 5/6 = 83.3%



Example – Classify - Contact Lens Data



WEKA
A Data Mining Tool

By Susan L. Miertschin