COMP47460 Tutorial

Dimension Reduction

Aonghus Lawlor
Derek Greene

School of Computer Science
Autumn 2018

Reminder - Selection v Transformation

- Two general strategies for dimension reduction:

Feature Selection

- Tries to find a minimum subset of the original features that optimises one or more criteria, rather than producing an entirely new set of dimensions for the data.

  e.g. Filters, Wrappers

Feature Transformation (Feature Extraction)

- Transforms the original features of a dataset to a completely new, smaller, more compact feature set, while retaining as much information as possible.
  e.g. Principal Components Analysis (PCA), Linear Discriminant Analysis (LDA)

Applying Filters and Wrappers

- **General filter approach:**
  1. Choose the feature selection criterion (e.g. Information Gain)
  2. Apply the filter to rank the features based on the criterion.
  3. Select top-*k* ranked features using an appropriate strategy.
  4. Remove the other features from your dataset.
  5. Apply the classifier on this new version of the dataset.

- **General wrapper approach:**
  1. Choose the classifier, search strategy, and evaluation strategy.
  2. Apply the wrapper to select a subset of features.
  3. Remove the other features from your dataset.
  4. Apply the same classifier on this new version of the dataset.

Tutorial Q1

In Weka, apply *filter-based feature selection* with Information Gain to identify the 3 most discriminating and 3 least discriminating features in the *Wine* dataset in the ARFF file provided.

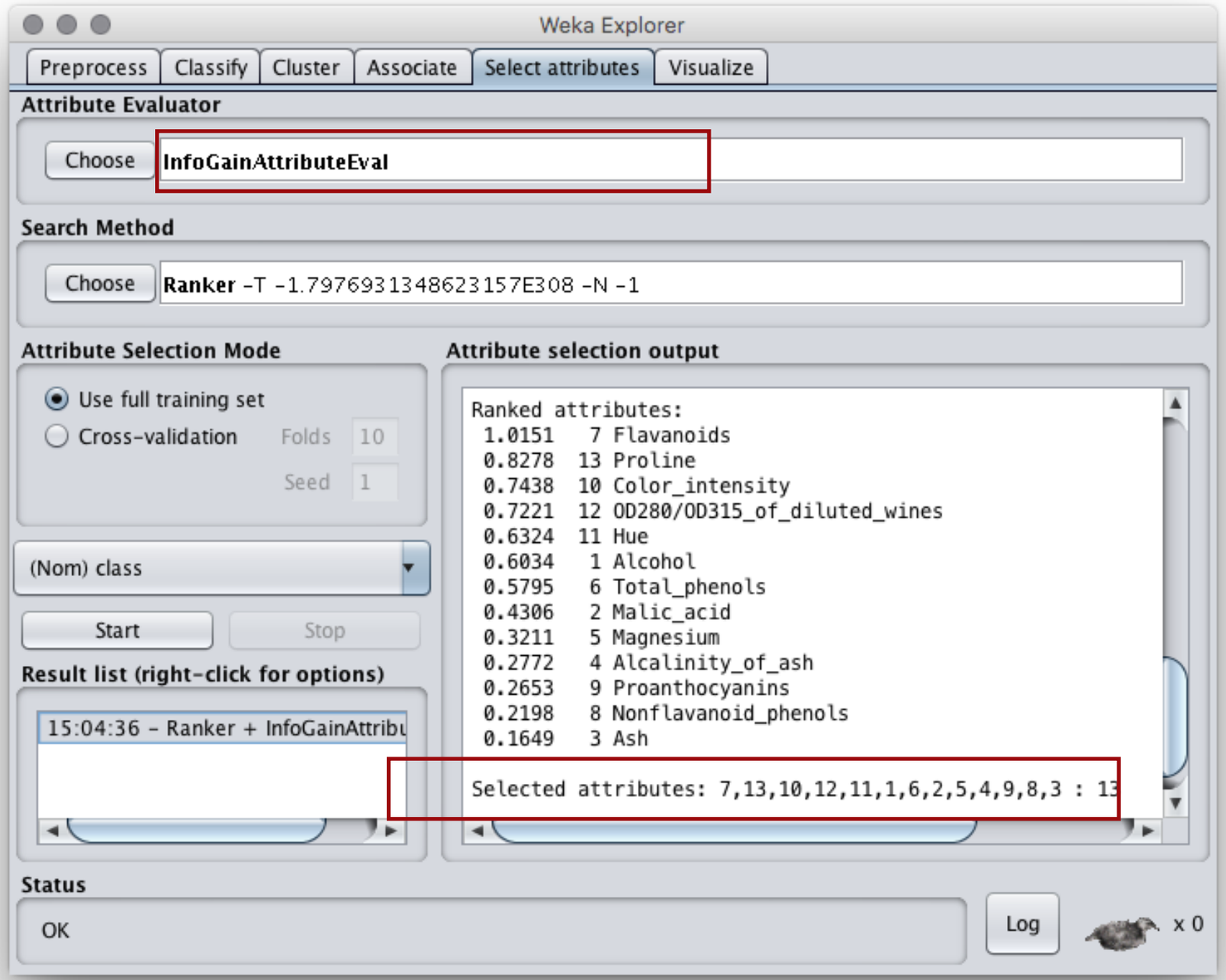
Based on these results, assess the 10-fold cross-validation classification accuracy of a 1-Nearest Neighbour classifier with:

(i) only the 3 most discriminating features included

(ii) only the 3 least discriminating features included

Tutorial Q1

In Weka *Select attributes* tab, choose *InfoGainAttributeEval* as the evaluator, *Ranker* as the method.



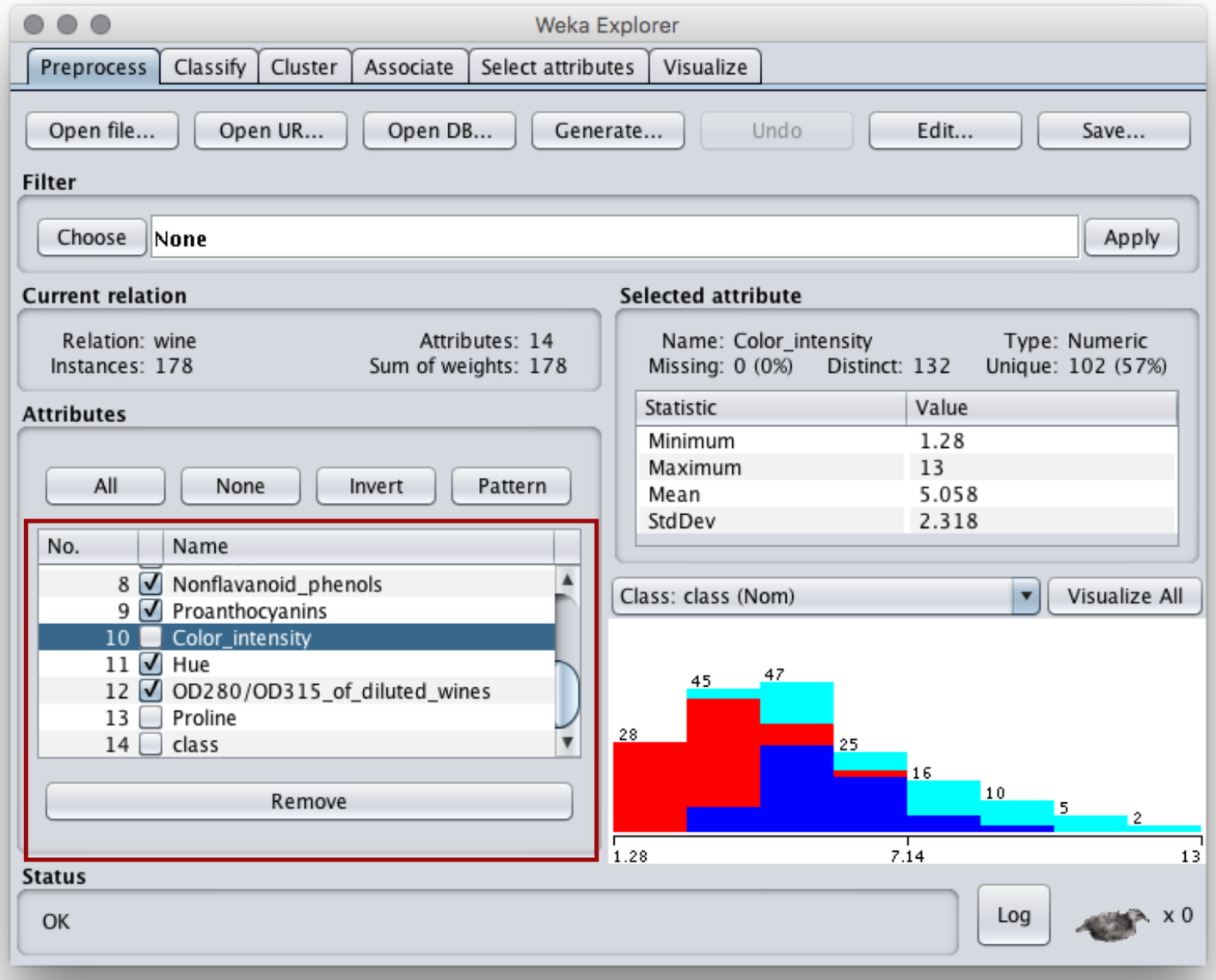
Tutorial Q1(i)

Assess the accuracy of a 1-nearest neighbour classifier with only the 3 most discriminating features included.

Most discriminating:
7, 13, 10

In the *Preprocess* tab, remove the unwanted features.

NB: Keep the “class” feature!



Tutorial Q1(i)

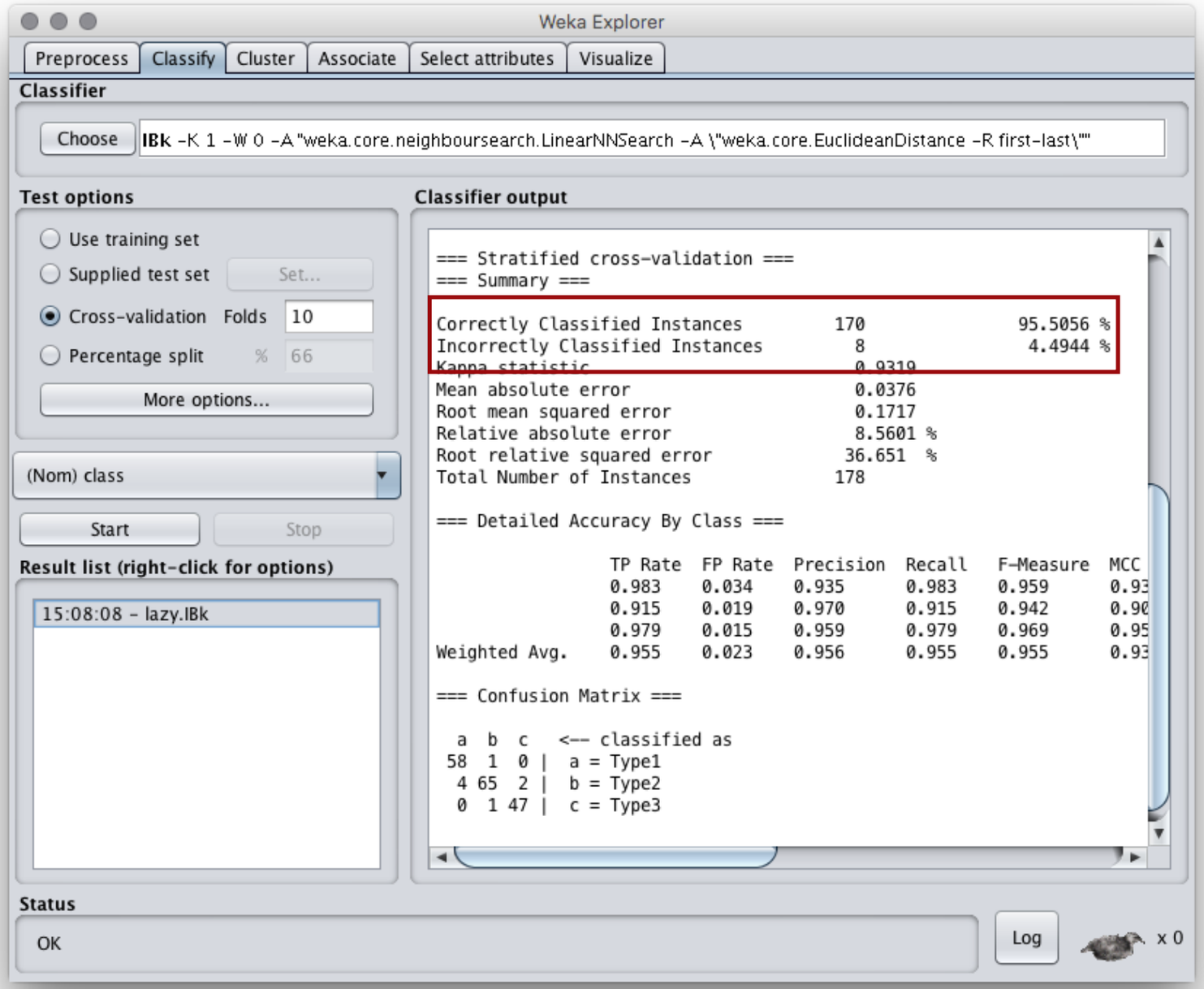
Assess the accuracy of a 1-nearest neighbour classifier with only the 3 most discriminating features included.

Most discriminating:
7, 13, 10

In the *Preprocess* tab, remove the unwanted features.

NB: Keep the “class” feature!

Run the 1NN classifier with the new feature subset.



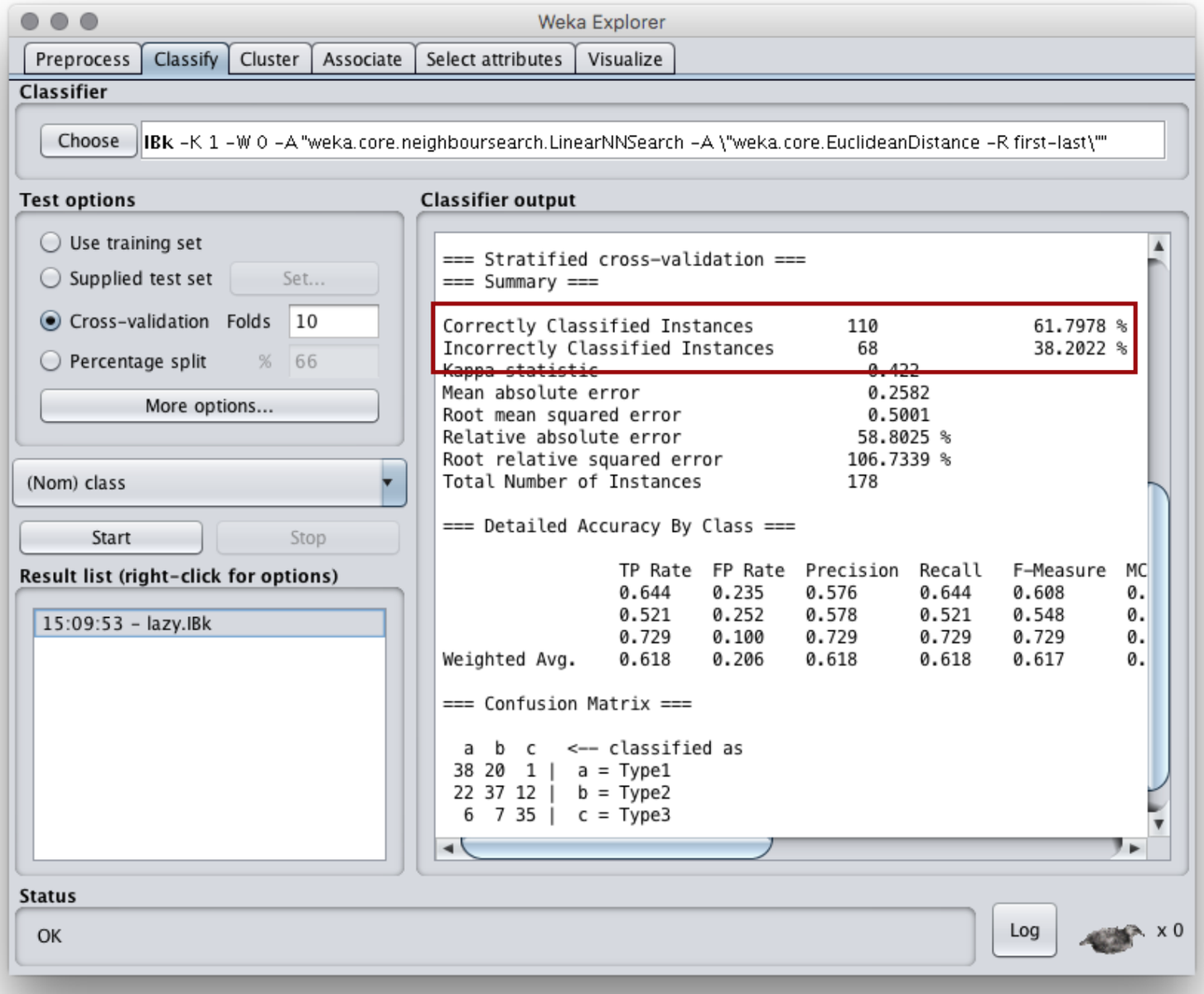
Tutorial Q1(ii)

Assess the accuracy of a 1-nearest neighbour classifier with only the 3 least discriminating features included.

Least discriminating:
3, 8, 9

Reload the ARFF file again, remove the unwanted features.

Run the 1NN classifier with the new feature subset.



Tutorial Q2

In Weka, apply *wrapper-based feature selection* to the *Wine* dataset using a 3-Nearest Neighbour classifier and the following search strategies:

(i) forward sequential search
(ii) backward elimination

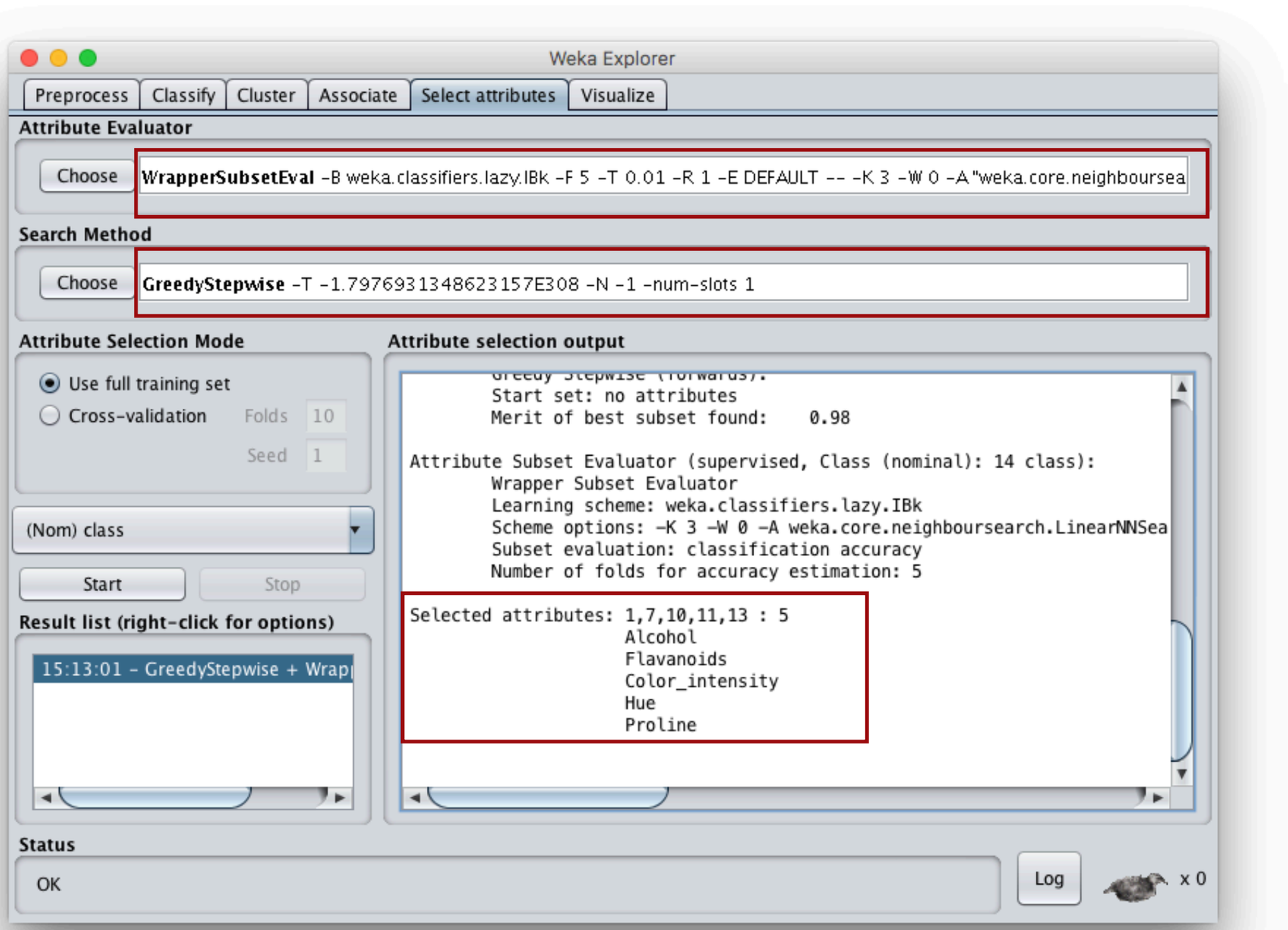
Which common features were selected by both search strategies?

Would it be appropriate to use either of the resulting feature subsets in conjunction with a Decision Tree classifier? Justify your answer.

NB: Reload the Wine dataset with all features.

Tutorial Q2

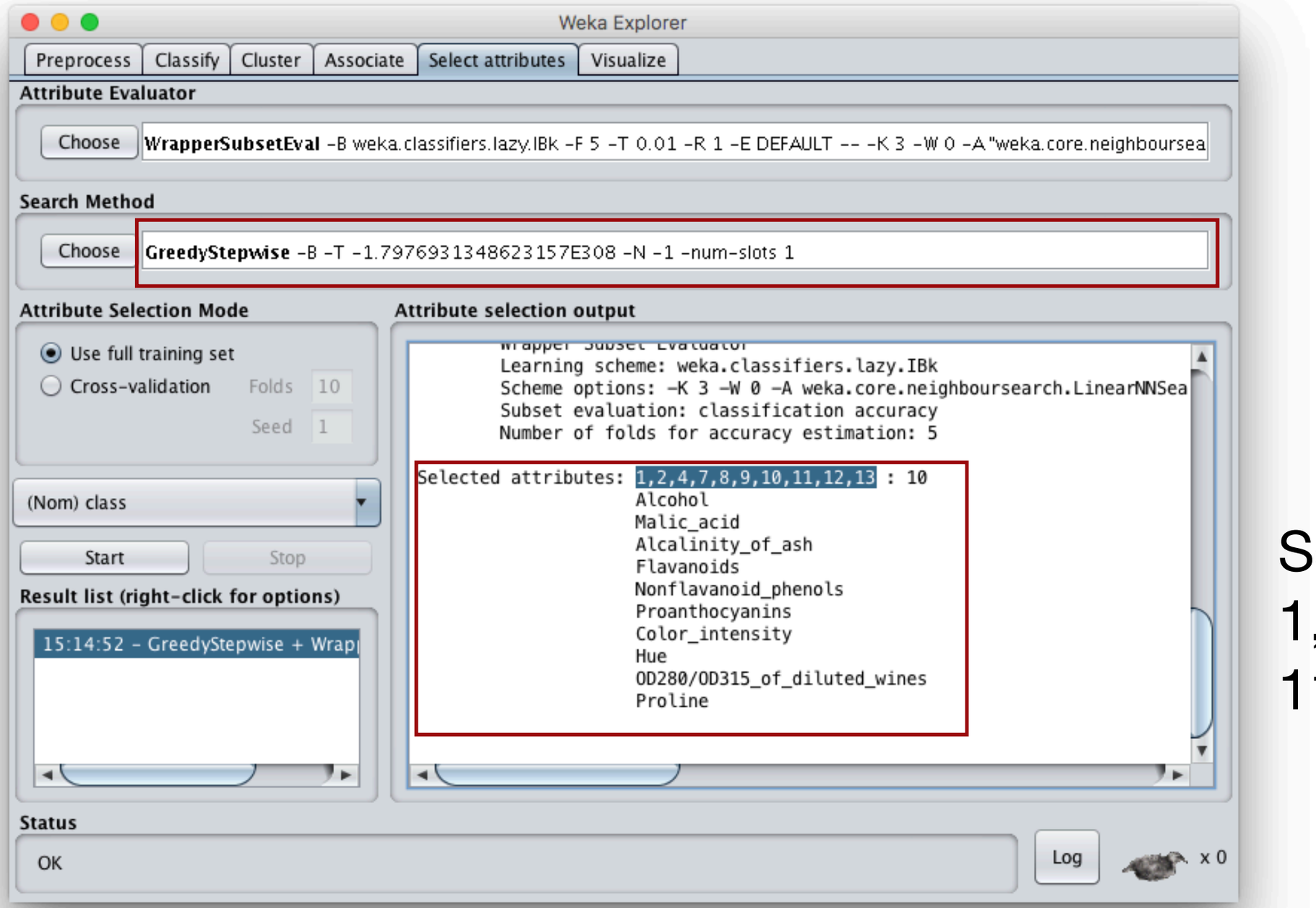
For **Forward Sequential** search: In *Select attributes* tab, choose *WrapperSubsetEval* as evaluator. Change options for wrapper to *IBk* (k=3). Choose *GreedyStepwise* as the search method, with option *SearchBackwards = False*.



Selected:
1,7,10,11,13

Tutorial Q2

For **Backward Elimination**: In *Select attributes* tab, choose *WrapperSubsetEval* as evaluator. Change options for wrapper to *IBk* (k=3). Choose *GreedyStepwise* as the search method, with option *SearchBackwards* = *True*.



Selected:
1,2,4,7,8,9,10,
11,12,13

Reminder - PCA

- Principal Components (PCs): New features constructed as linear combinations of the original features, uncorrelated with one another.
- **PCA Process:**

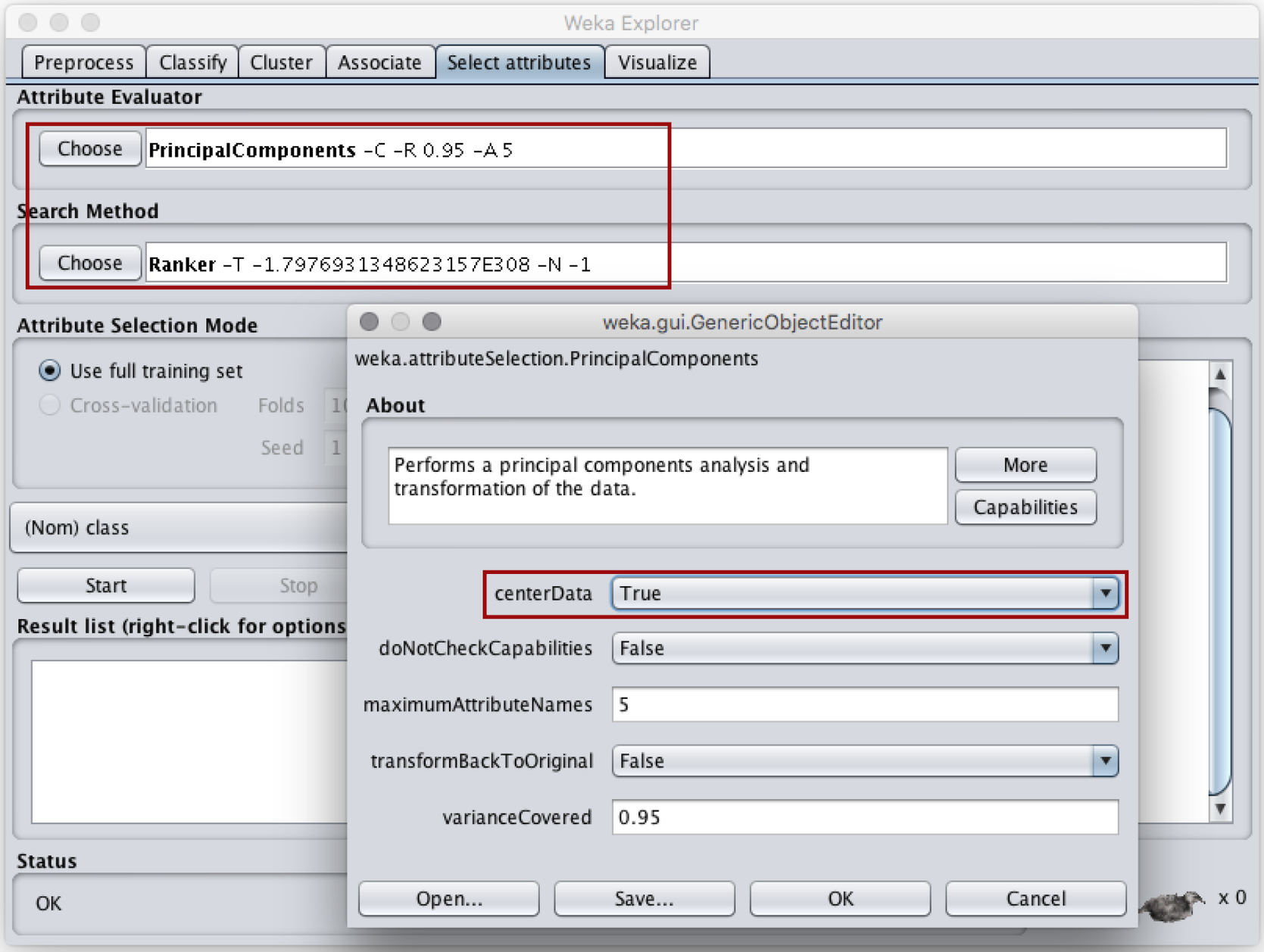
1. Calculate the mean of the columns of $\mathbf{X}$.
2. Subtract the column means from each row of $\mathbf{X}$, to create the centred matrix $\mathbf{Y}$.
3. Calculate the covariance matrix $\mathbf{C} = \mathbf{Y}^\mathsf{T}\mathbf{Y}/(n-1)$
4. Calculate the eigenvectors of the covariance matrix $\mathbf{C}$.
5. The Principal Components (PCs) are given by the eigenvectors of $\mathbf{C}$. The *i*-th PC is given by the eigenvector corresponding to the *i*-th largest eigenvalue of $\mathbf{C}$.
6. Select an appropriate number of PCs k and use them as a new reduced n x k representation of the dataset.

Tutorial Q3

- Use Weka to apply PCA *feature transformation* to the Diabetes dataset in the ARFF file provided.
  (Use the Ranker search method in the ‘Select Attributes’ tab).
- How many Principal Components are selected by Weka?
- Produce a visualisation of the 2 leading Principal Components.

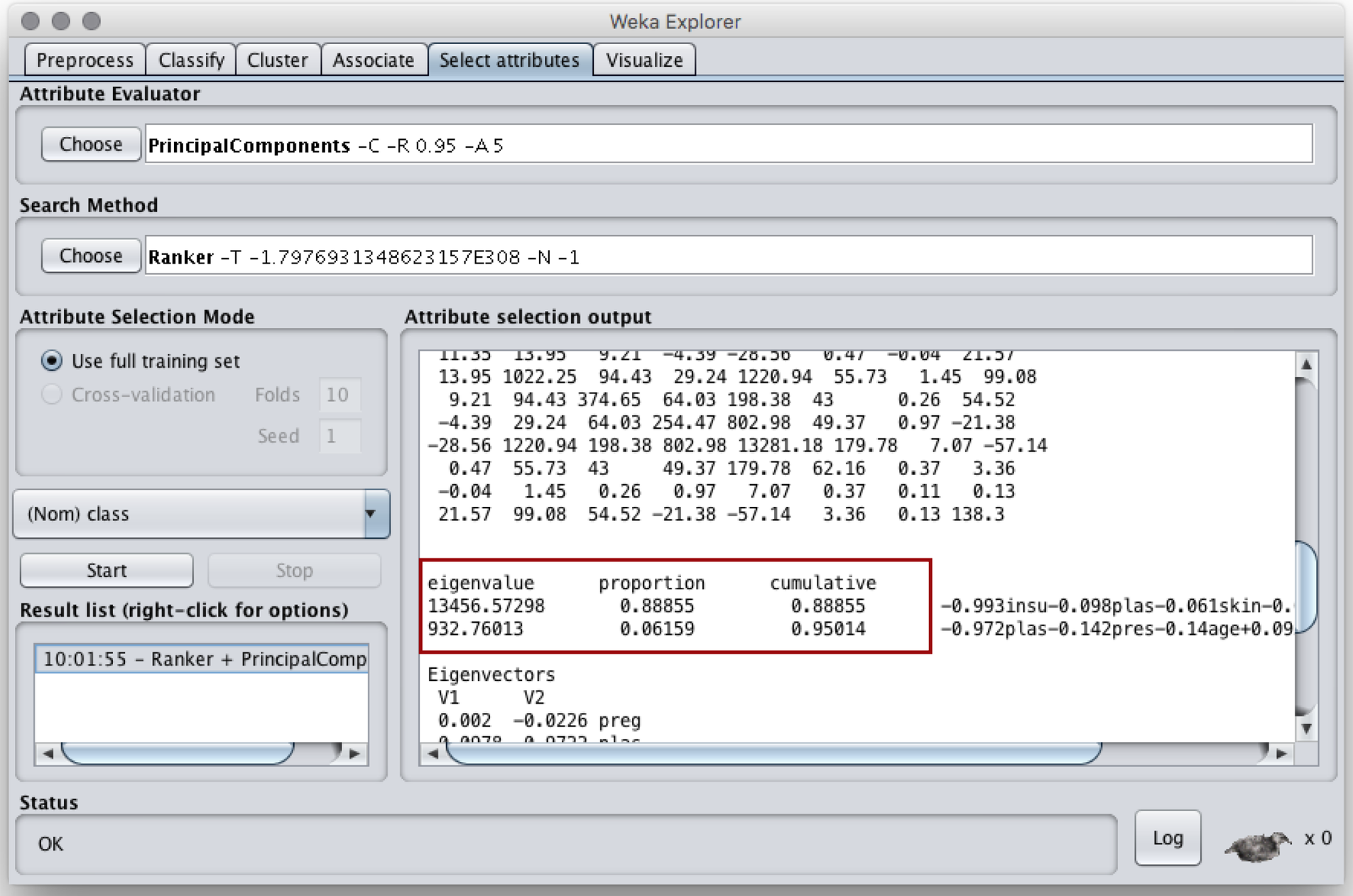
Tutorial Q3

In *Select attributes* tab, choose *PrincipalComponents* as the evaluator, and *Ranker* as the search method. Change options for PCA, set *centerData* to True to use the standard covariance matrix approach.



Tutorial Q3

PCA in Weka selects 2 PCs, which account for > 95% of the variance of the data (proportion is 0.88855 + 0.06159 = 0.95014).



Tutorial Q3

To visualise the PCs, right click the result on the *Result List* and choose *Visualize transformed data*. In the new window, click one of the small pairwise plots to zoom in.