

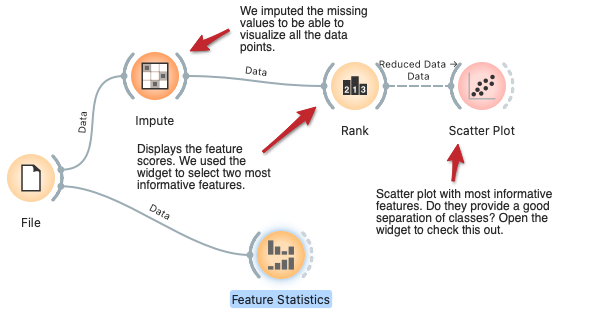
Lab 3: Feature selection with Orange

Feature selection is one of the main steps in data mining. It allows eliminating unnecessary and redundant features so we can focus on the important ones. It also helps to create models quicker and speed up the response time.

**Exercise I: Ranking the best features**

Find the most important features in orange for the following datasets:

Brown-selected, heard\_desease, housing, iris, titanic, zoo.



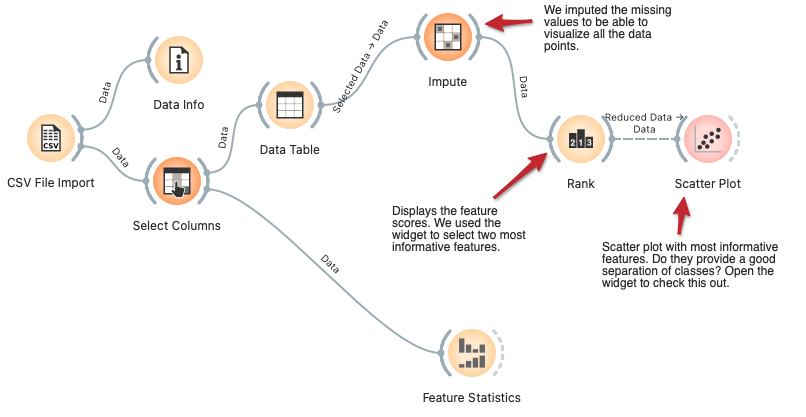
Write down the two most important features for each dataset, the accuracy using all features, the accuracy using only two features, and the accuracy using the main two components of PCA.

Use the widget Feature statistics to see if the best features have some kind of pattern when they are visualized. \*Hint: See if the best features have a class that predominates over the other for an attribute with a good number of instances.

Build models using logistic regression, random forest, and decision tree and see the performance difference when the best two features are selected and when all the features are selected. You need to use Cross-Validation to calculate the results in a more reliable way.

| Num | Dataset name | First most important features | Second most important features | Accuracy with all features | Accuracy with two most important features | Accuracy with PCA one and two |
| --- | --- | --- | --- | --- | --- | --- |
| 1 | Brown-selected |  |  | Example:  LG: 86%, RF: 87%, DT:93% |  |  |
| 2 | Heard\_desease |  |  |  |  |  |
| 3 | Housing |  |  |  |  |  |
| 4 | Iris |  |  |  |  |  |
| 5 | Titanic |  |  |  |  |  |
| 6 | Zoo |  |  |  |  |  |

**Exercise II - Using select columns and CSV File Import widgets:**

Now, try to use the widgets “CSV File Import” and “Select Columns”. You need to use the widget “Select Columns” to create the class by labelling it as the target. The file can be downloaded from Brightspace where it says Labs.

Fill the following table with the following results:

| Num | Dataset name | First most important features | Second most important features | Accuracy with all features | Accuracy with two most important features | Accuracy with PCA one and two |
| --- | --- | --- | --- | --- | --- | --- |
| 1 | Contact lenses |  |  |  |  |  |

**Exercise III - Extracting insights from the data:**

Get three important insights and support them with some evidence of the following dataset. You will need to use the widget called “Datasets”.



Bank Marketing (2014), from [UCI ML Repository](http://archive.ics.uci.edu/ml/datasets/Bank+Marketing)

*“Data from direct marketing campaigns (phone calls) of a Portuguese bank. The classification goal is to predict if the client will subscribe to a term deposit given the profile of a client that contains attributes such as age, job type, martial status, education, information on previous loans, and other.”*

References

Moro S, Cortez P, and Rita P (2014) A Data-Driven Approach to Predict the Success of Bank Telemarketing. Decision Support Systems 62:22-31.