Venn Diagram

Plots a Venn diagram for two or more data subsets.

**Inputs**

- Data: input dataset

Outputs

- Selected Data: instances selected from the plot
- Data: entire data with a column indicating whether an instance was selected or not

The **Venn Diagram** widget displays logical relations between datasets by showing the number of common data instances (rows) or the number of shared features (columns). Selecting a part of the visualization outputs the corresponding instances or features.
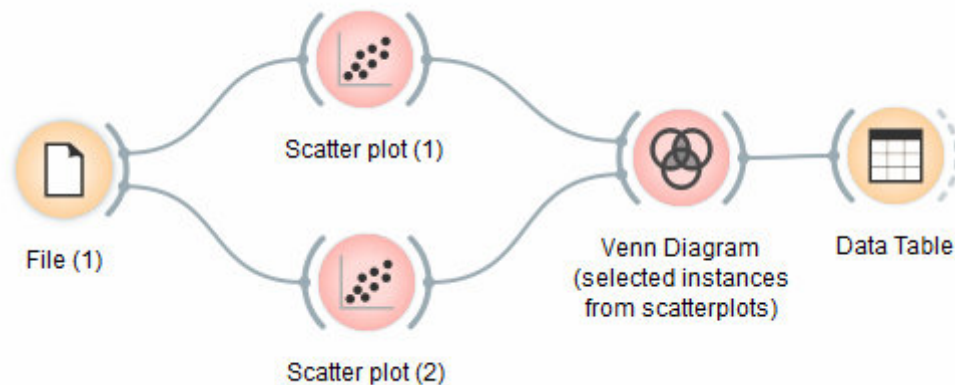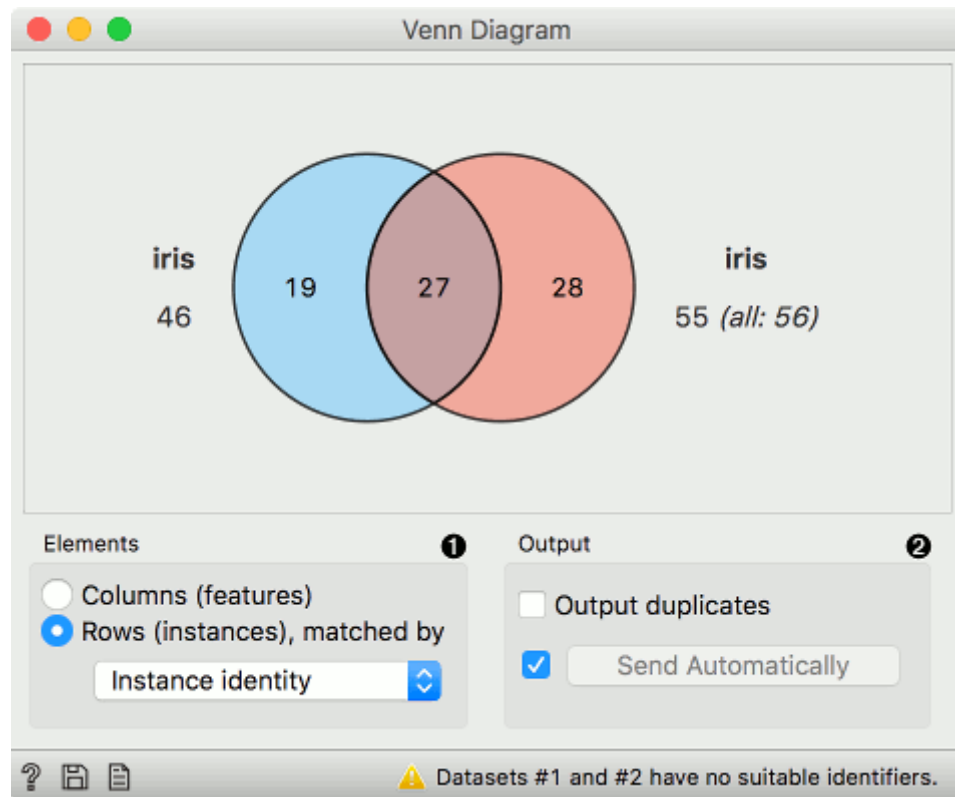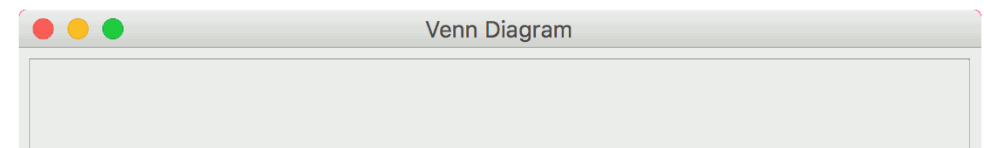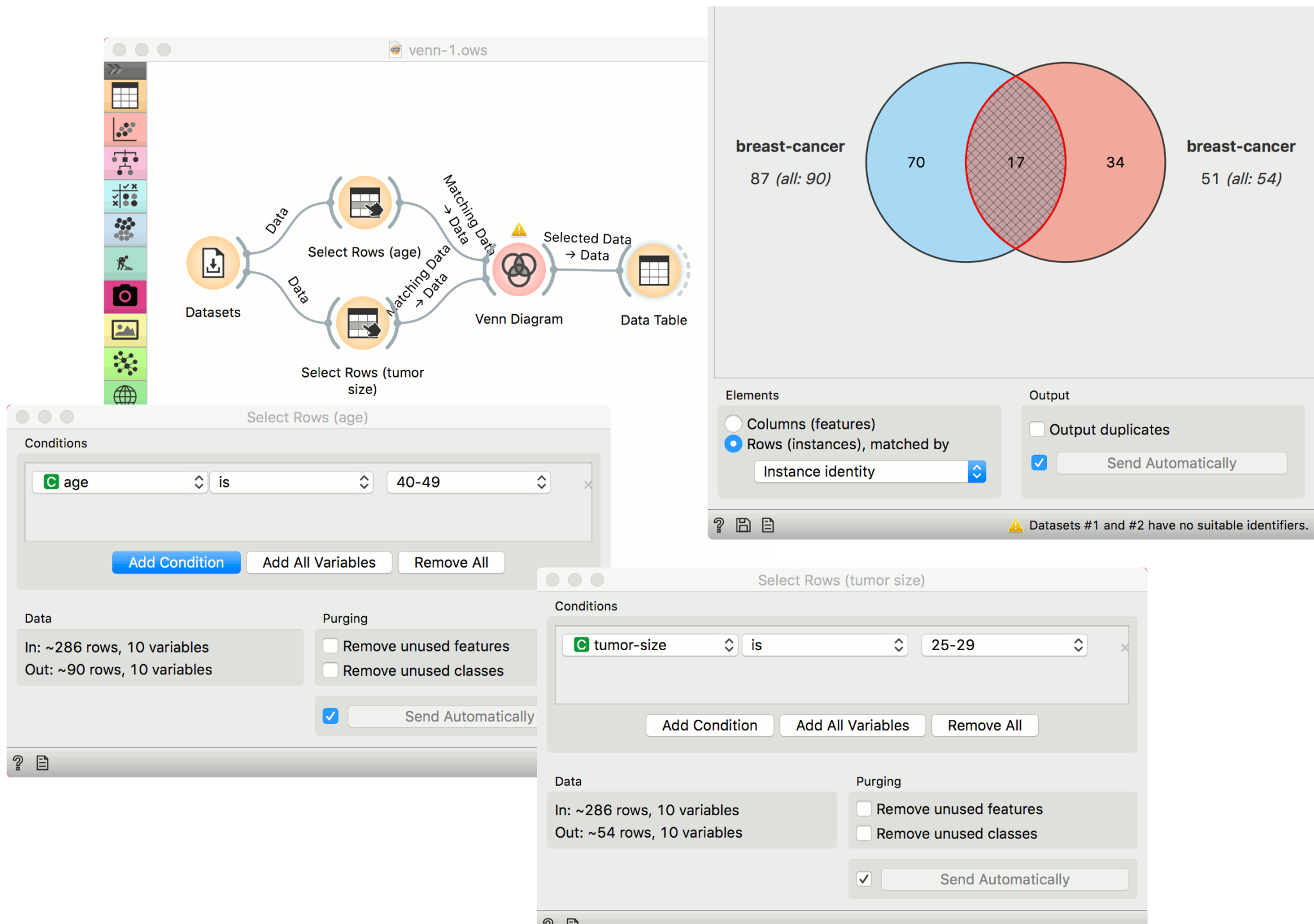
1. Select whether to count common features or instances.
2. Select whether to include duplicates or to output only unique rows (applicable only when matching by instances). If *Auto commit* is on, changes are automatically communicated to other widgets.

Rows can be matched by their identity, e.g. rows from different data sets match if they came from the same row in a file. Instead of using identities, we can choose a string variable to match the rows by. A warning is shown if data sets have no common string variable.

Examples

The easiest way to use the **Venn Diagram** is to select data subsets and find matching instances in the visualization. We use the *breast-cancer* dataset to select two subsets with Select Rows widget - the first subset is that of breast cancer patients aged between 40 and 49 and the second is that of patients with a tumor size between 20 and 29. The **Venn Diagram** helps us find instances that correspond to both criteria, which can be found in the intersection of the two circles.

venn-1.ows

Select Rows (age)

Data

Matching Data → Data

Datasets

Data

Matching Data → Data

Select Rows (tumor size)

Selected Data → Data

Venn Diagram

Data Table

breast-cancer
87 *(all: 90)*

70   17   34

breast-cancer
51 *(all: 54)*

Elements

○ Columns (features)
● Rows (instances), matched by
Instance identity

Output

☐ Output duplicates

☑   Send Automatically

⚠ Datasets #1 and #2 have no suitable identifiers.

Select Rows (age)

Conditions

| C age | is | 40-49 | ✕ |

Add Condition   Add All Variables   Remove All

Data
In: ~286 rows, 10 variables
Out: ~90 rows, 10 variables

Purging
☐ Remove unused features
☐ Remove unused classes

☑   Send Automatically

Select Rows (tumor size)

Conditions

| C tumor-size | is | 25-29 | ✕ |

Add Condition   Add All Variables   Remove All

Data
In: ~286 rows, 10 variables
Out: ~54 rows, 10 variables

Purging
☐ Remove unused features
☐ Remove unused classes

☑   Send Automatically

The **Venn Diagram** widget can be also used for exploring different prediction models. In the following example, we analysed 3 prediction methods, namely Naive Bayes, SVM and Random Forest, according to their misclassified instances.

By selecting misclassifications in the three Confusion Matrix widgets and sending them to Venn diagram, we can see all the misclassification instances visualized per method used. Then we open **Venn Diagram** and select, for example, the misclassified instances that were identified by all three methods. This is represented as an intersection of all three circles. Click on the intersection to see this two instances marked in the Scatter Plot widget. Try selecting different diagram sections to see how the scatter plot visualization changes.

		Iris-setosa	Iris-versicolor	Iris-virginica	Σ
Act	Iris-virginica	0	3	47	
	Σ	50	50	50	

	Select Correct	Select Misclassified	Clea

		Iris-setosa	Iris-versicolor	Iris-virginica	Σ
	Iris-setosa	50	0	0	50
Actual	Iris-versicolor	0	41	9	50
	Iris-virginica	0	7	43	50
	Σ	50	48	52	150

Select Correct	Select Misclassified	Clear Selection