Preprocess

Preprocesses data with selected methods.

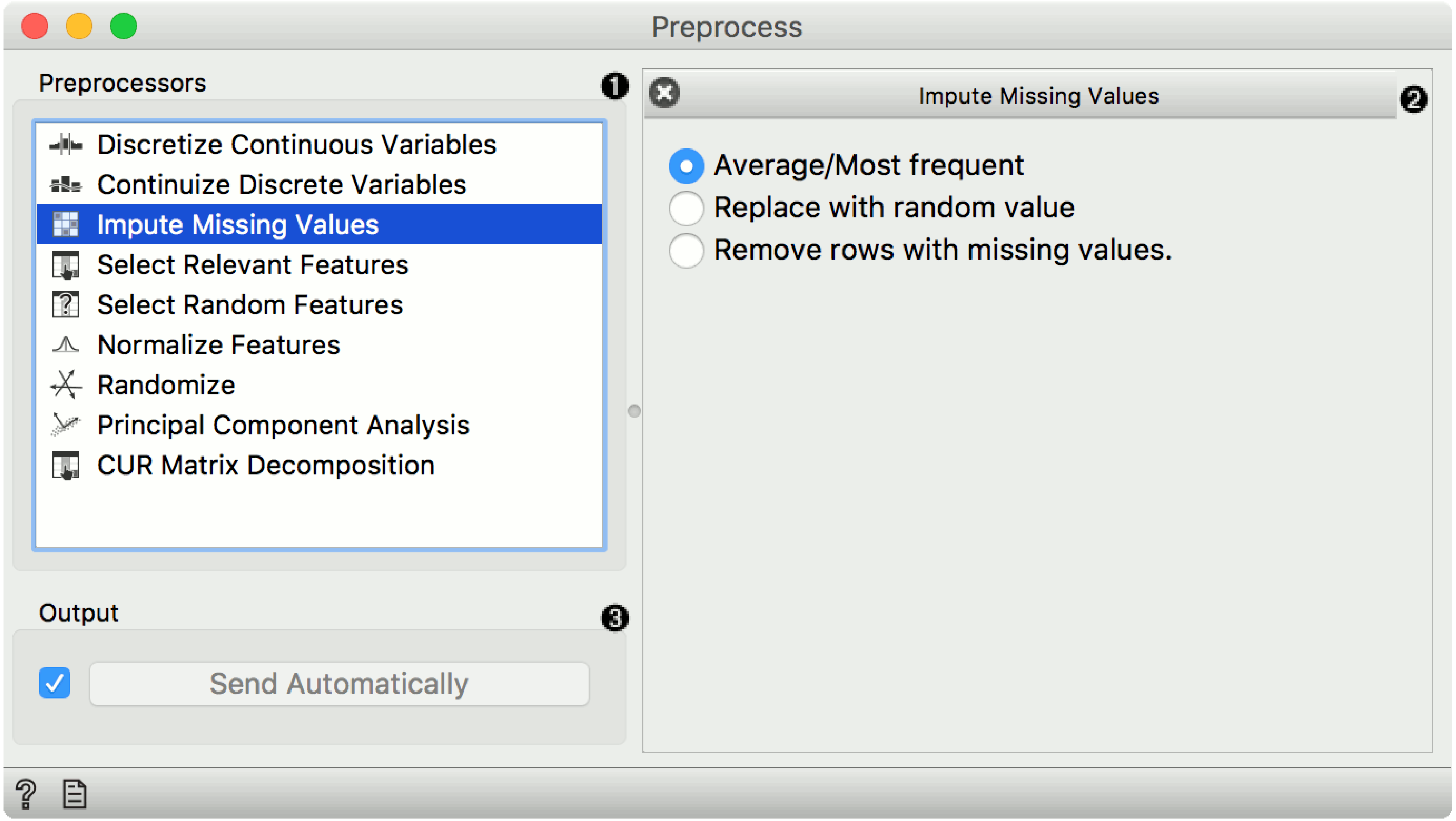
Inputs

- Data: input dataset

Outputs

- Preprocessor: preprocessing method
- Preprocessed Data: data preprocessed with selected methods

Preprocessing is crucial for achieving better-quality analysis results. The **Preprocess** widget offers several preprocessing methods that can be combined in a single preprocessing pipeline. Some methods are available as separate widgets, which offer advanced techniques and greater parameter tuning.



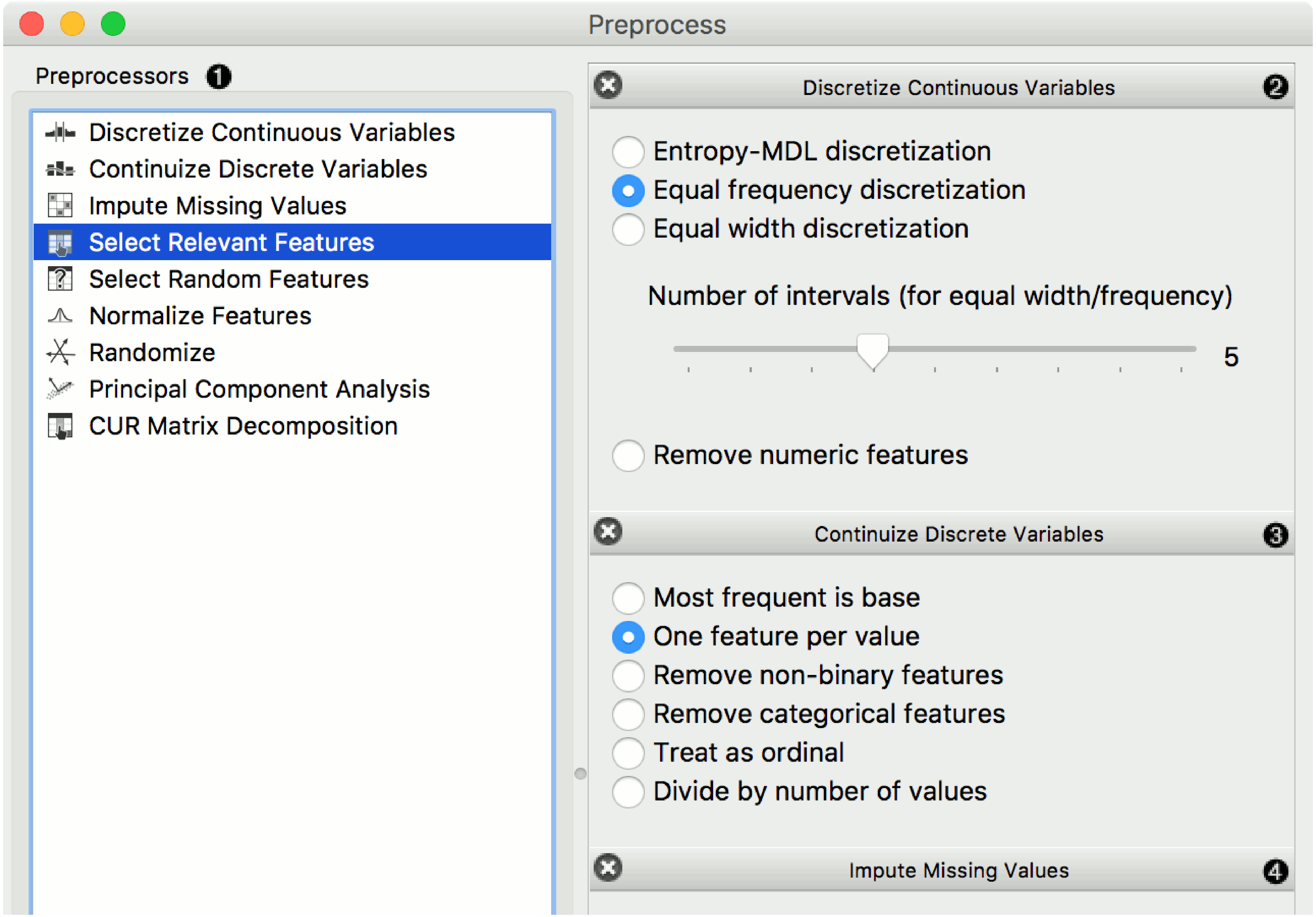
Impute Missing Values

[✓ Average/Most frequent]
Replace with random value
Remove rows with missing values.

[✓] Send Automatically

1. List of preprocessors. Double click the preprocessors you wish to use and shuffle their order by dragging them up or down. You can also add preprocessors by dragging them from the left menu to the right.
2. Preprocessing pipeline.
3. When the box is ticked (*Send Automatically*), the widget will communicate changes automatically. Alternatively, click *Send*.

Preprocessors



Preprocess

Preprocessors ❶

- Discretize Continuous Variables
- Continuize Discrete Variables
- Impute Missing Values
- Select Relevant Features
- Select Random Features
- Normalize Features
- Randomize
- Principal Component Analysis
- CUR Matrix Decomposition

Discretize Continuous Variables ❷

- Entropy-MDL discretization
- [✓ Equal frequency discretization]
- Equal width discretization

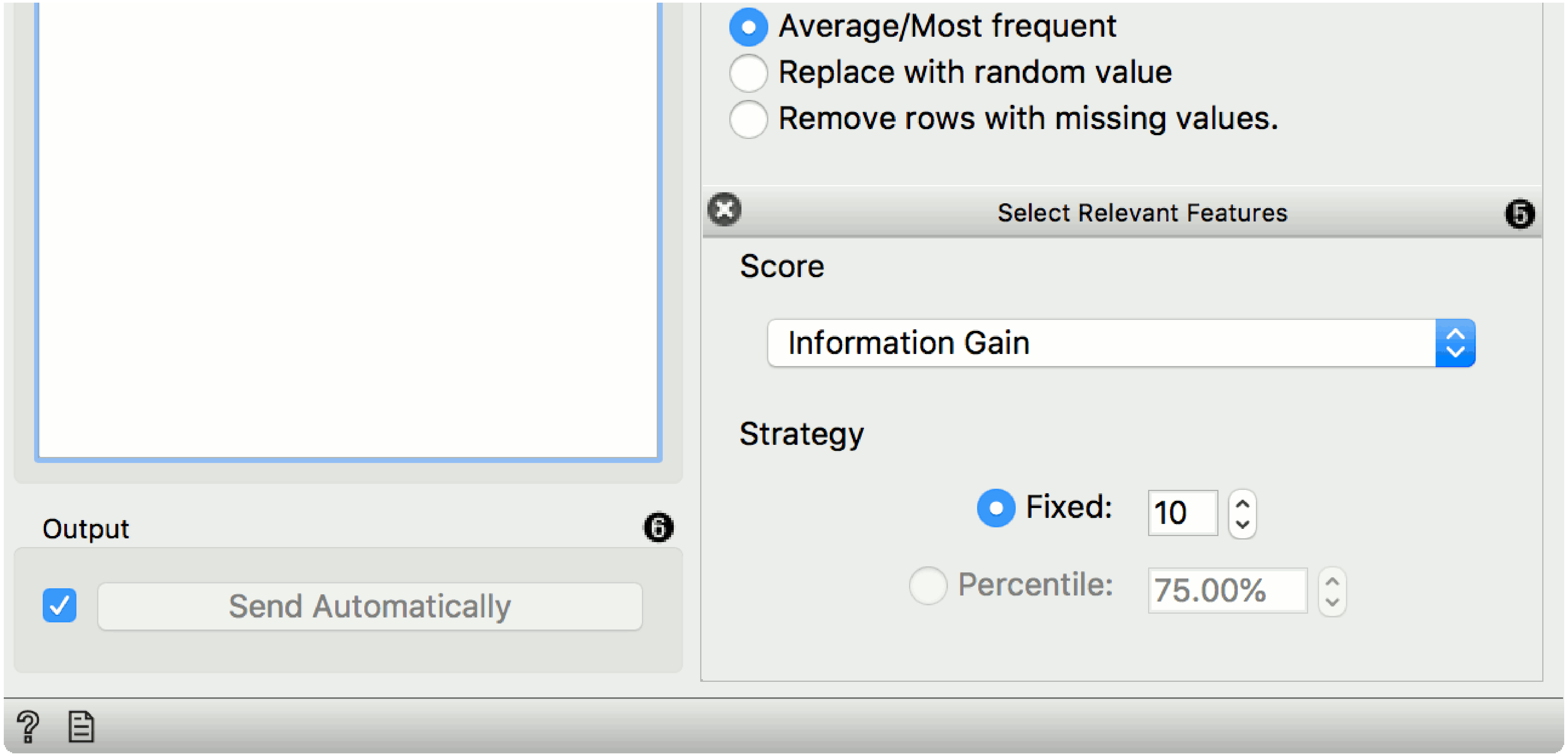
Number of intervals (for equal width/frequency) 5

- Remove numeric features

Continuize Discrete Variables ❸

- Most frequent is base
- [✓ One feature per value]
- Remove non-binary features
- Remove categorical features
- Treat as ordinal
- Divide by number of values

Impute Missing Values ❹



[✓ Average/Most frequent]
Replace with random value
Remove rows with missing values.

Select Relevant Features 5

Score: Information Gain

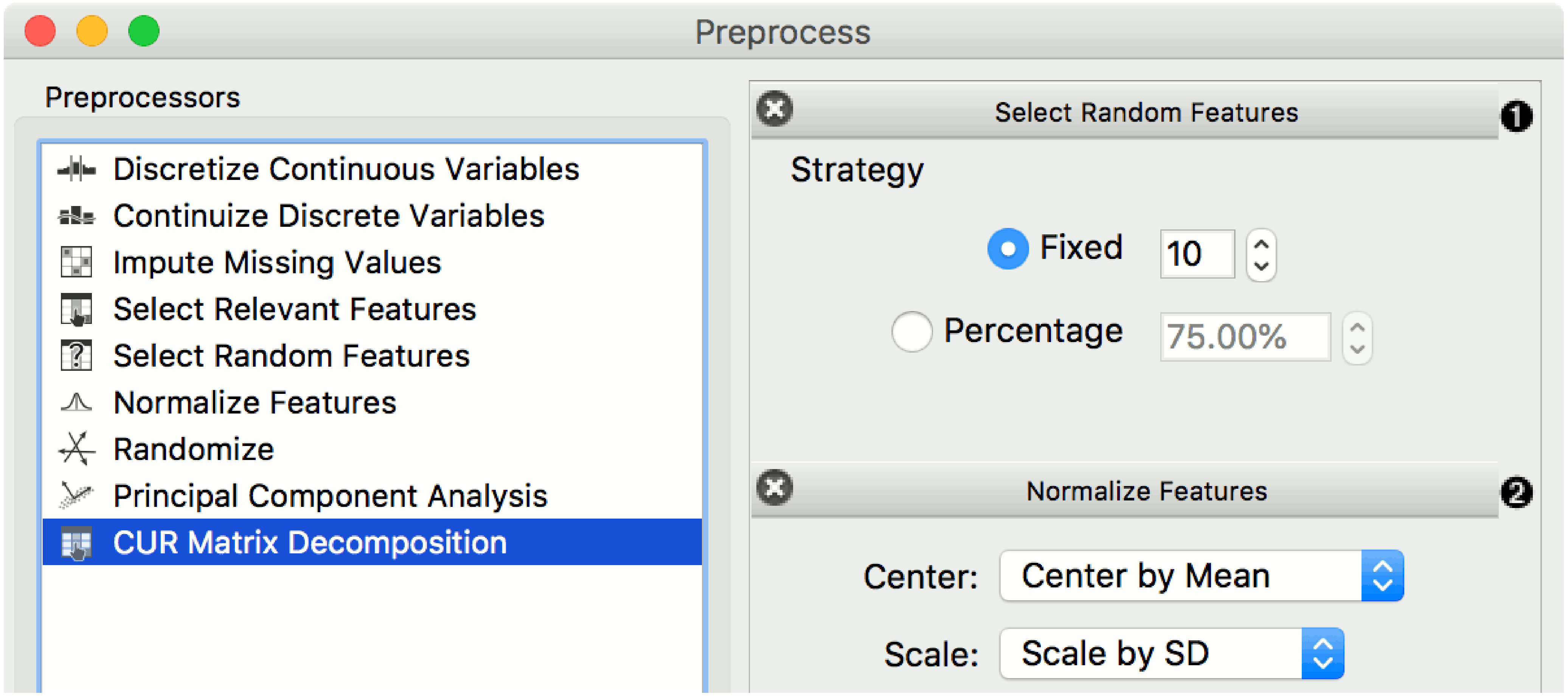
Strategy: [✓ Fixed: 10] Percentile: 75.00%

Output 6

[✓] Send Automatically

1. Discretization of continuous values:
   - Entropy-MDL discretization by Fayyad and Irani that uses expected information to determine bins.
   - *Equal frequency discretization* splits by frequency (same number of instances in each bin.
   - *Equal width discretization* creates bins of equal width (span of each bin is the same).
   - *Remove numeric features* altogether.
2. Continuization of discrete values:
   - *Most frequent as base* treats the most frequent discrete value as 0 and others as 1. The discrete attributes with more than 2 values, the most frequent will be considered as a base and contrasted with remaining values in corresponding columns.
   - *One feature per value* creates columns for each value, place 1 where an instance has that value and 0 where it doesn’t. Essentially One Hot Encoding.
   - *Remove non-binary features* retains only categorical features that have values of either 0 or 1 and transforms them into continuous.

    - *Remove categorical features* removes categorical features altogether.
    - *Treat as ordinal* takes discrete values and treats them as numbers. If discrete values are categories, each category will be assigned a number as they appear in the data.
    - *Divide by number of values* is similar to treat as ordinal, but the final values will be divided by the total number of values and hence the range of the new continuous variable will be [0, 1].
3. Impute missing values:
    - *Average/Most frequent* replaces missing values (NaN) with the average (for continuous) or most frequent (for discrete) value.
    - *Replace with random value* replaces missing values with random ones within the range of each variable.
    - *Remove rows with missing values*.
4. Select relevant features:
    - Similar to Rank, this preprocessor outputs only the most informative features. Score can be determined by information gain, gain ratio, gini index, ReliefF, fast correlation based filter, ANOVA, Chi2, RReliefF, and Univariate Linear Regression.
    - *Strategy* refers to how many variables should be on the output. *Fixed* returns a fixed number of top scored variables, while *Percentile* return the selected top percent of the features.


Preprocess

Preprocessors
- Discretize Continuous Variables
- Continuize Discrete Variables
- Impute Missing Values
- Select Relevant Features
- Select Random Features
- Normalize Features
- Randomize
- Principal Component Analysis
- CUR Matrix Decomposition

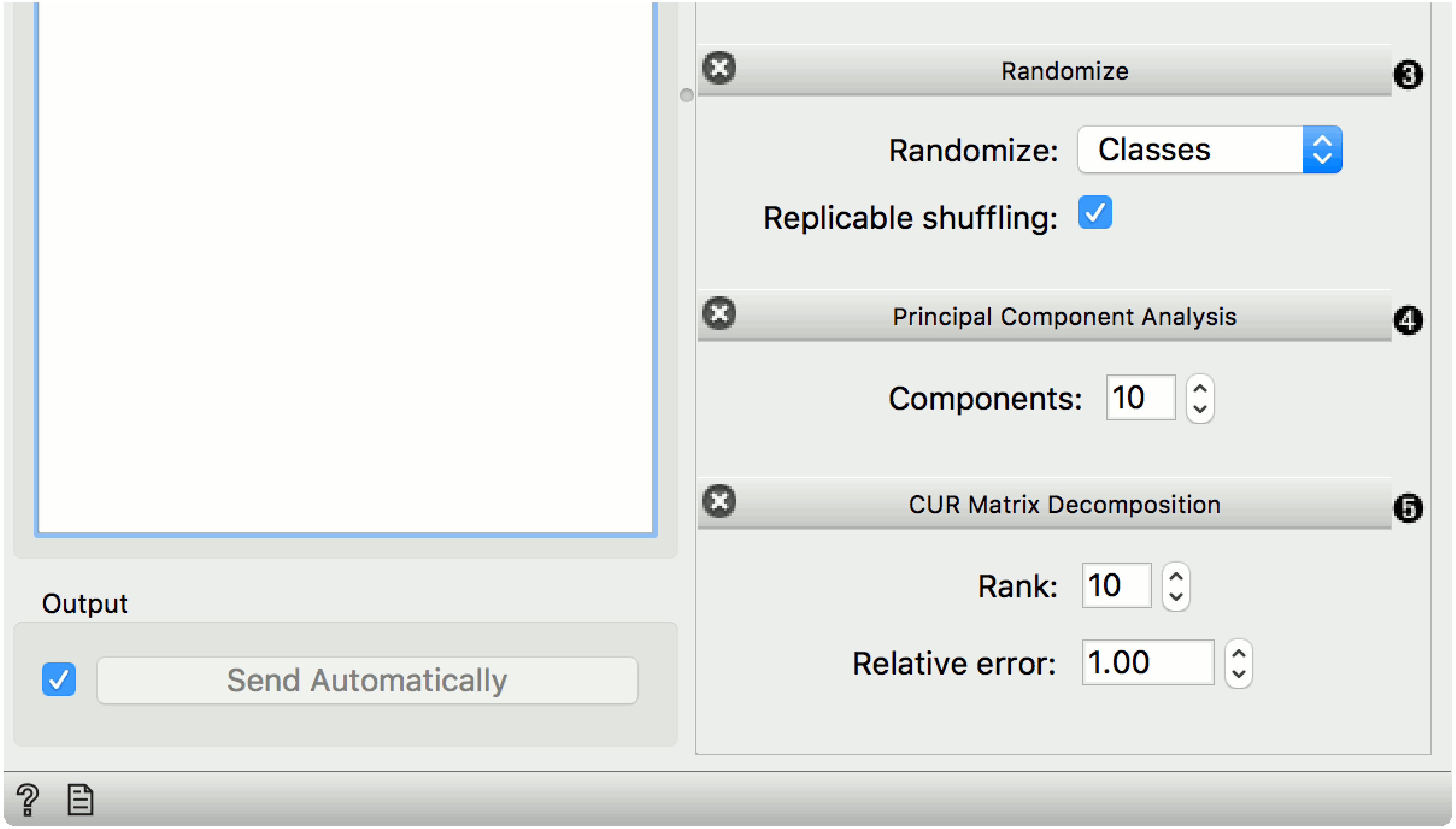
1 Select Random Features

Strategy
- [✓ Fixed] 10
- Percentage 75.00%

2 Normalize Features

Center: Center by Mean

Scale: Scale by SD


Randomize 3

Randomize: Classes

Replicable shuffling: ✓

Principal Component Analysis 4

Components: 10

CUR Matrix Decomposition 5

Rank: 10

Relative error: 1.00

Output

✓ Send Automatically

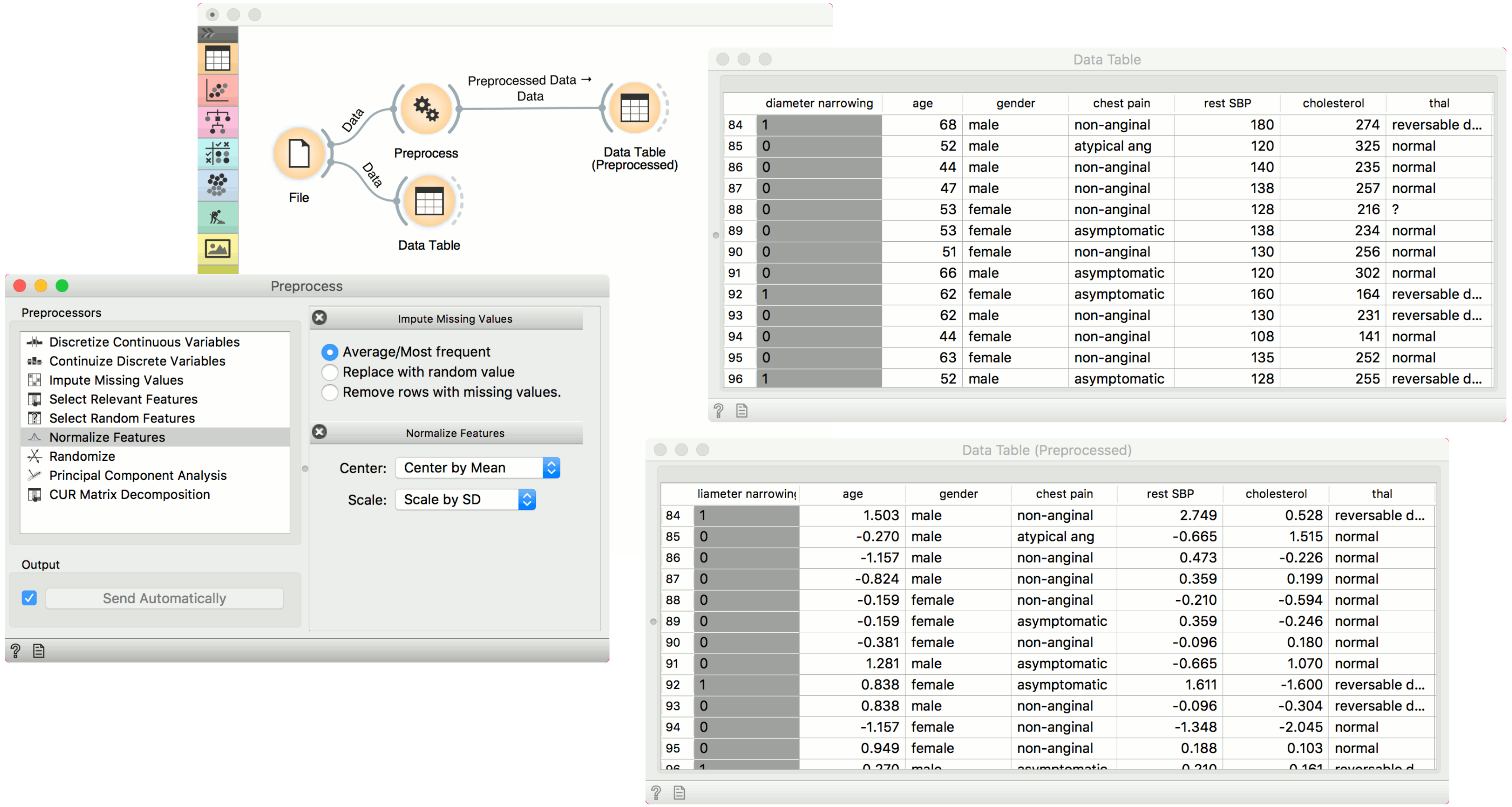
5. *Select random features* outputs either a fixed number of features from the original data or a percentage. This is mainly used for advanced testing and educational purposes.
6. Normalize adjusts values to a common scale. Center values by mean or median or omit centering altogether. Similar for scaling, one can scale by SD (standard deviation), by span or not at all.
7. Randomize instances. Randomize classes shuffles class values and destroys connection between instances and class. Similarly, one can randomize features or meta data. If replicable shuffling is on, randomization results can be shared and repeated with a saved workflow. This

is mainly used for advanced testing and educational purposes.

8. *Remove sparse features* retains features that have more than user-defined threshold percentage of non-zero values. The rest are discarded.
9. Principal component analysis outputs results of a PCA transformation. Similar to the PCA widget.
10. CUR matrix decomposition is a dimensionality reduction method, similar to SVD.

Examples

In the first example, we have used the *heart_disease.tab* dataset available in the dropdown menu of the File widget. then we used **Preprocess** to impute missing values and normalize features. We can observe the changes in the Data Table and compare it to the non-processed data.



Preprocess

Preprocessors
- Discretize Continuous Variables
- Continuize Discrete Variables
- Impute Missing Values
- Select Relevant Features
- Select Random Features
- Normalize Features
- Randomize
- Principal Component Analysis
- CUR Matrix Decomposition

Impute Missing Values

- [✓ Average/Most frequent]
- Replace with random value
- Remove rows with missing values.

Normalize Features

Center: Center by Mean

Scale: Scale by SD

Output

[✓] Send Automatically

Data Table

	diameter narrowing	age	gender	chest pain	rest SBP	cholesterol	thal
84	1	68	male	non-anginal	180	274	reversable d...
85	0	52	male	atypical ang	120	325	normal
86	0	44	male	non-anginal	140	235	normal
87	0	47	male	non-anginal	138	257	normal
88	0	53	female	non-anginal	128	216	?
89	0	53	female	asymptomatic	138	234	normal
90	0	51	female	non-anginal	130	256	normal
91	0	66	male	asymptomatic	120	302	normal
92	1	62	female	asymptomatic	160	164	reversable d...
93	0	62	male	non-anginal	130	231	reversable d...
94	0	44	female	non-anginal	108	141	normal
95	0	63	female	non-anginal	135	252	normal
96	1	52	male	asymptomatic	128	255	reversable d...

Data Table (Preprocessed)

	liameter narrowin	age	gender	chest pain	rest SBP	cholesterol	thal
84	1	1.503	male	non-anginal	2.749	0.528	reversable d...
85	0	-0.270	male	atypical ang	-0.665	1.515	normal
86	0	-1.157	male	non-anginal	0.473	-0.226	normal
87	0	-0.824	male	non-anginal	0.359	0.199	normal
88	0	-0.159	female	non-anginal	-0.210	-0.594	normal
89	0	-0.159	female	asymptomatic	0.359	-0.246	normal
90	0	-0.381	female	non-anginal	-0.096	0.180	normal
91	0	1.281	male	asymptomatic	-0.665	1.070	normal
92	1	0.838	female	asymptomatic	1.611	-1.600	reversable d...
93	0	0.838	male	non-anginal	-0.096	-0.304	reversable d...
94	0	-1.157	female	non-anginal	-1.348	-2.045	normal
95	0	0.949	female	non-anginal	0.188	0.103	normal
96	1	0.270	male	asymptomatic	0.210	0.161	reversable d

In the second example, we show how to use **Preprocess** for predictive modeling.

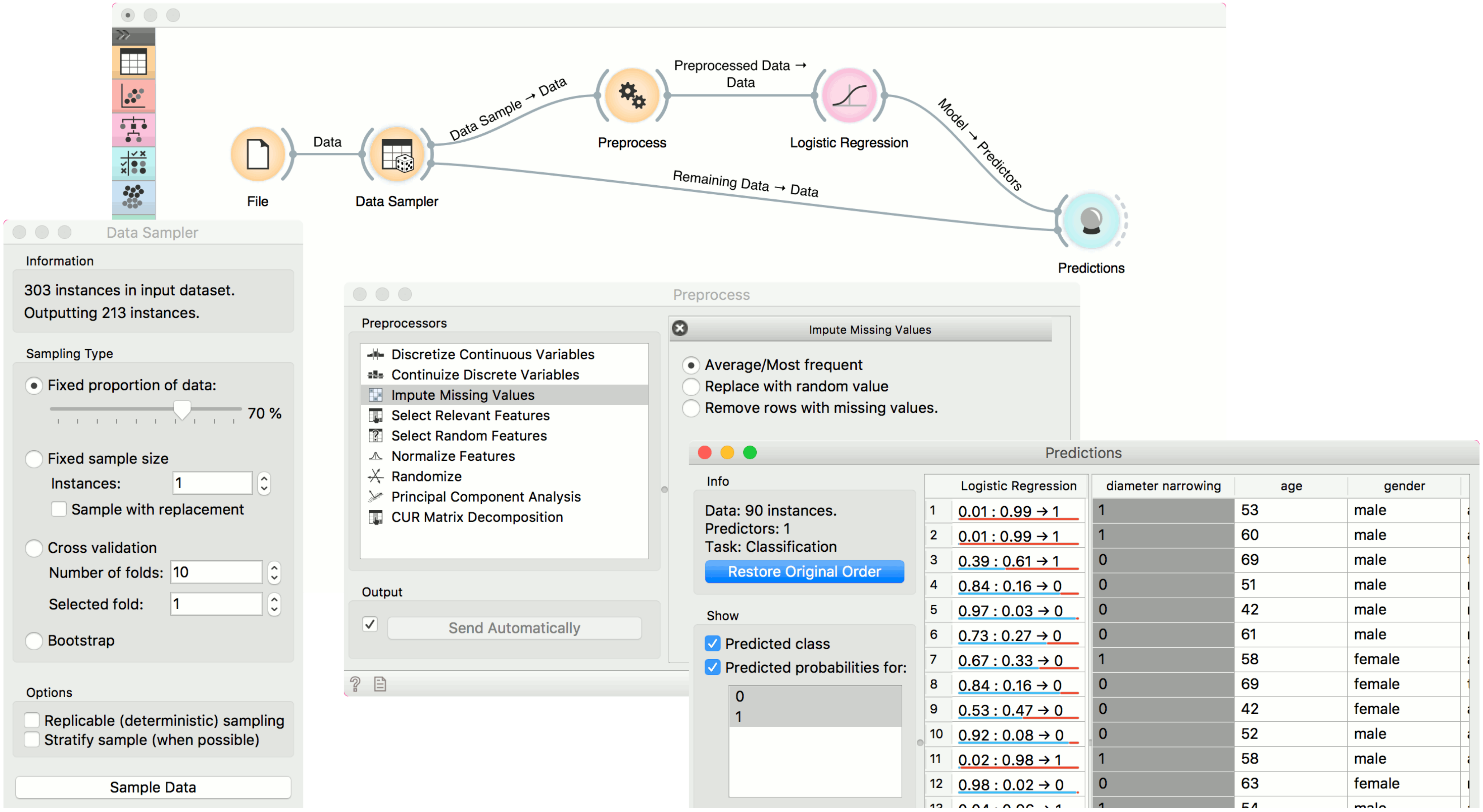
This time we are using the *heart_disease.tab* data from the File widget. You can access the data through the dropdown menu. This is a dataset with 303 patients that came to the doctor suffering from a chest pain. After the tests were done, some patients were found to have diameter narrowing and others did not (this is our class variable).

The heart disease data have some missing values and we wish to account for that. First, we will split the data set into train and test data with Data Sampler.

Then we will send the *Data Sample* into Preprocess. We will use *Impute Missing Values*, but you can try any combination of preprocessors on your data. We will send preprocessed data to Logistic Regressionand the constructed model to **Predictions**.

Finally, **Predictions** also needs the data to predict on. We will use the output of Data Sampler for prediction, but this time not the *Data Sample*, but the *Remaining Data*, this is the data that wasn't used for training the model.

Notice how we send the remaining data directly to **Predictions** without applying any preprocessing. This is because Orange handles preprocessing on new data internally to prevent any errors in the model construction. The exact same preprocessor that was used on the training data will be used for predictions. The same process applies to Test & Score.



Data Sampler

Information

303 instances in input dataset.
Outputting 213 instances.

Sampling Type

[✓ Fixed proportion of data:] 70 %

Fixed sample size
Instances: 1
Sample with replacement

Cross validation
Number of folds: 10
Selected fold: 1

Bootstrap

Options

Replicable (deterministic) sampling
Stratify sample (when possible)

Sample Data

Preprocess

Preprocessors

- Discretize Continuous Variables
- Continuize Discrete Variables
- Impute Missing Values
- Select Relevant Features
- Select Random Features
- Normalize Features
- Randomize
- Principal Component Analysis
- CUR Matrix Decomposition

Impute Missing Values

[✓ Average/Most frequent]
Replace with random value
Remove rows with missing values.

Output

✓ Send Automatically

Predictions

Info

Data: 90 instances.
Predictors: 1
Task: Classification

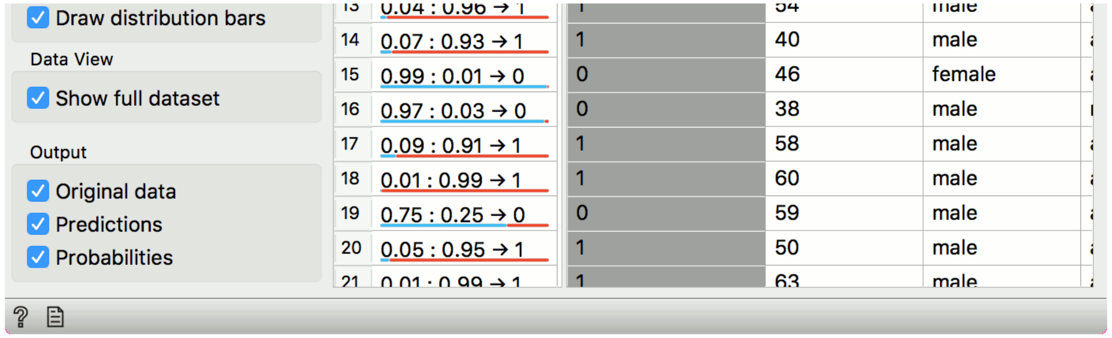
Restore Original Order

Show

✓ Predicted class
✓ Predicted probabilities for:

0
1

	Logistic Regression	diameter narrowing	age	gender
1	0.01 : 0.99 → 1	1	53	male
2	0.01 : 0.99 → 1	1	60	male
3	0.39 : 0.61 → 1	0	69	male
4	0.84 : 0.16 → 0	0	51	male
5	0.97 : 0.03 → 0	0	42	male
6	0.73 : 0.27 → 0	0	61	male
7	0.67 : 0.33 → 0	1	58	female
8	0.84 : 0.16 → 0	0	69	female
9	0.53 : 0.47 → 0	0	42	female
10	0.92 : 0.08 → 0	0	52	male
11	0.02 : 0.98 → 1	1	58	male
12	0.98 : 0.02 → 0	0	63	female



☑ Draw distribution bars

Data View

☑ Show full dataset

Output

☑ Original data
☑ Predictions
☑ Probabilities

13	0.04 : 0.96 → 1	1	54	male
14	0.07 : 0.93 → 1	1	40	male
15	0.99 : 0.01 → 0	0	46	female
16	0.97 : 0.03 → 0	0	38	male
17	0.09 : 0.91 → 1	1	58	male
18	0.01 : 0.99 → 1	1	60	male
19	0.75 : 0.25 → 0	0	59	male
20	0.05 : 0.95 → 1	1	50	male
21	0.01 : 0.99 → 1	1	63	male