Sampling

The **Sampling** component allows you to select records from a source dataset using multiple methods. Use it to reduce data volume for analysis, avoiding redundant calculations. This is especially important when the complexity of data mining algorithms grows exponentially with larger datasets.

When generating a sample, you can choose from a variety of selection methods: *random*, *random uniform*, *stratified*, *sequencial*, or *biased* sampling.

Ports

Input

- **Input data source** (data table).

Output

- **Output dataset** (data table).

Configuration wizard

At the top of the first wizard page (**Data sampling**), the system shows two read-only fields:

- **Input status**: This field displays the current data use status (`Not activated` upon first access). In some cases, you may need to use the input data, such as for *biased sampling*. Click **Activate** on the right.
- **Total records**: This field displays the number of records in the **Input data source** table.

Output table row count

In the pane below, set the size of the output table. Use the toggle to specify the count of rows as a number (switch to **Absolute (Rows)**) or as a percentage of the source table size (switch to **Relative (%)**).

Sampling method

Select one of the available sampling tecniques and customize it in the **Method parameters** area, if applicable:

- **Random**: The component randomly selects records from the source dataset and places them into the resulting sample. No custom settings apply.
- **Random uniform**: The component divides all records into groups, randomly selects records from each group, and places them into the resulting sample. Set the group size as the method parameter.
- **Stratified**: The component divides all records into homogeneous groups (strata), randomly selects records from each group, and places them into the resulting sample.
  - Use checkboxes to define strata fields.
  - Enable **Complete list of unique values** if you need to include records from all groups in the resulting sample. Note that if this operation fails the node returns an error.
- **Sequence**: Specify whether the component takes the sample from the beginning or the end of the source table. The component maintains the same row order in the sample as in the source table.
- **Biased sampling**:
  - Select a field to use and click **Get values** on the right pane's toolbar.
  - Then, select a unique value from this field and click the **Edit factor** toolbar button.
  - In the window that opens, enter a *positive* factor value. A factor of *N* increases the number of rows with the selected unique value by *N* times in the output dataset.

> **Note:** The number of unique values in a field must not exceed `10 000`.

Random seed

Use a *positive integer* to initialize the pseudorandom number generator. This number fully defines the generator sequence. If you re-initialize the generator with the same number, it produces the same sequence.

This parameter affects the **Random** and **Random uniform** sampling methods. You can reproduce sampling results if you provide the same data and specify the same random seed.

Additionally, the wizard provides the following commands for this parameter:

- **Always randomize**: Enable this so that the node automatically generates a random number each time.
- **Generate**: Generate a new initial number.
- **Copy**: Copy the specified value to the clipboard.

Read on: Data Mining Components