Splitting tables into training data sets and test data sets

Last Updated: 2021-03-01

If you want to create a prediction model, you typically use a table or a view that contains historical data. If you decide to use the Easy Mining procedures for classification or regression, you might want to split this table into the following disjoint data sets:

- One data set to train the prediction model
- One data set to test the prediction model

Stratified samples

Last Updated: 2021-03-01

If you specify a value that is different from `NULL` or an empty string for the parameter *<stratSampleColumn>*, a stratified sample is created for the training data set. This means that, based on the complement of the test data set, a sample is created. In this sample, the values of the `Stratified Sample` column occur with approximately the same frequency.

Syntax

Last Updated: 2021-03-01

```
IDMMX.SplitData(<inputTable>,
                <trainViewName>,
                <testViewName>,
                <testSampleSize>
               [<stratSampleColumn>]
```

Input parameters

Last Updated: 2021-03-01

With the **SplitData** procedure, you must specify the following parameters:

<inputTable>

The name of the input table or the view
This parameter is of type VARCHAR. Its size is 240.

<trainViewName>

The name of the view that contains the training data set.
This view contains the same columns as the input table.

<testViewName>

The name of the view that contains the test data set.

This view contains the same columns as the input table. It contains a random sample of the records in the input table. The sample size is approximately <testSampleSize>%.

This parameter is of type VARCHAR. Its size is 240.

<testSampleSize>

This column contains a percentage that indicates the size of the test data set.

This parameter is of type REAL.

Example

Last Updated: 2021-03-01

You might want to split the table `BANK.BANKCUSTOMERS` into equal portions.

- For the portion that includes the training data set, you want to specify the name `BANK.BANKCUSTOMERS_TRAIN`.
- For the portion that includes the test data set, you want to specify the name `BANK.BANKCUSTOMERS_TEST`.
- Because you want to split the table into equal portions, you specify `50.0` for the size of each portion.

Use the following command to run the **SplitData** procedure:

```
DB2 "call IDMMX.SplitData('BANK.BANK
                          'BANK.BANK
                          'BANK.BANK
                           50.0)"
```