

Дубликаты и противоречия

Описание

Компонент позволяет выявить в исходной выборке данных дублирующие и противоречивые записи.

Дубликаты — записи в таблице, все входные и выходные поля которых одинаковые. Дубликаты приводят к избыточности, увеличивают объем выборки, при этом не повышая информативность данных.

Противоречия — записи в таблице с одинаковыми входными полями, но отличающиеся хотя бы по одному выходному полю. Противоречия приводят к искажению результата анализа и снижают качество моделей, поскольку нарушают общие закономерности в данных, обнаружение которых и является целью исследования.

Алгоритм ищет в наборе данных записи, для которых одинаковым входным полям соответствуют одинаковые (дубликаты) или разные (противоречия) выходные поля.

▼ Пример:

Исследуем следующий набор данных на дубликаты и противоречия, для этого зададим назначение  «Входное» для полей «Поле 1» и «Поле 2», назначение  «Выходное» для полей «Поле 3» и «Поле 4» в настройках входного порта.

Исходная таблица:

Поле 1	Поле 2	Поле 3	Поле 4
01.01.2019	2	1000	1500
21.05.2019	3	1000	1500
21.05.2019	3	700	1500
21.05.2019	3	700	1500
01.09.2019	4	1200	1700
01.09.2019	4	1200	1700

Выходная таблица:

Дубликат	Группа дубликата	Противоречие	Группа противоречия	Поле 1	Поле 2	Поле 3	Поле 4
false		false		01.01.2019	2	1000	1500
false		true	1	21.05.2019	3	1000	1500
true	1	true	1	21.05.2019	3	700	1500
true	1	true	1	21.05.2019	3	700	1500
true	2	false		01.09.2019	4	1200	1700
true	2	false		01.09.2019	4	1200	1700

В результате найдены две группы дубликатов и одна группа противоречий.

Порты

Вход

-  Входной источник данных (таблица данных). В настройках этого порта следует выставить назначение  «Входное» и  «Выходное» полям, которые будут использоваться для исследования на дубликаты и противоречия.

Примечание: обязательно должна быть задана хотя бы одна входная колонка.

Выход

-  Выходной набор данных. Таблица имеет следующую структуру:
 - Обязательные поля:
 - Дубликат** — значение логического типа, определяющее, является исходная строка дубликатом или нет.
 - Группа дубликата** — в одну группу дубликатов объединяются дубликаты, у которых входные и выходные колонки совпадают.

- **Противоречие** — значение логического типа, определяющее, является исходная строка противоречием или нет.
- **Группа противоречия** — в одну группу противоречий объединяются все противоречия, у которых входные колонки совпадают.

Примечание: нумерация групп начинается с 1.