Duplicates and Inconsistencies

Description

Use this component to detect duplicate and inconsistent records in the source dataset.

Duplicates: Records where all input and output fields match exactly. They create redundancy and inflate dataset size without adding value.

Inconsistencies: Records where all input fields match, but at least one output field differs. They distort analysis and reduce model quality by violating expected data patterns.

The algorithm identifies records with identical input fields and checks whether their output fields match (duplicates) or differ (inconsistencies).

▼ Example:

Let's analyze the following dataset for duplicates and inconsistencies. In the input port settings, we set the **Input** *usage type* for **Field 1** and **Field 2**, and **Output** *usage type* for **Field 3** and **Field 4**.

Source table:

Field 1	Field 2	Field 3	Field 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

Output table:

Duplicate	Duplicate group	Inconsistency	Inconsistency group	Field 1	Field 2	Field 3	Field 4
false		false		01.01.2019	2	1000	1500

Duplicate	Duplicate group	Inconsistency	Inconsistency group	Field 1	Field 2	Field 3	Field 4
false		true	1	21.05.2019	3	1 000	1 500
true	1	true	1	21.05.2019	3	700	1 500
true	1	true	1	21.05.2019	3	700	1 500
true	2	false		01.09.2019	4	1 200	1 700
true	2	false		01.09.2019	4	1 200	1 700

The analysis found two duplicate groups and one inconsistency group.

Ports

Input

- **Input data source** (data table): Accepts a data table. In the port settings, select **Input** and **Output** usage type for fields that you want to investigate.

> **Note:** You must define at least one **Input** column.

Output

- **Output dataset**: Returns a table with the following structure:
  - Required fields:
    - **Duplicate**: A `logical` value. Indicates if the row is a duplicate.
    - **Duplicate group**: Groups identical records (where both inputs and outputs match).
    - **Inconsistency**: A `logical` value. Indicates if the row is inconsist with another row (i.e., their inputs match but outputs differ).
    - **Inconsistency group**: Groups records with matching inputs but differing outputs.

> **Note:** Group numbering starts at *1*.

Read on: Correlation Analysis