ML\_Dataiku\_Pipeline\_project

short line



|  |  |  |
| --- | --- | --- |
| **Version** | **Author** | **Date** |
| 1.0 | abhishek.barandooru | 2023-08-29 14:32:21 |

Table of contents

[I. Project general description 4](#_Toc137204120)

[II. Graphical Representation of the Flow 5](#_Toc137204121)

[III. Datasets 6](#_Toc137204122)

[1. {{$dataset.name}} 6](#_Toc137204123)

[a) Status of the dataset {{$dataset.name}} 6](#_Toc137204124)

[b) Related Recipes 7](#_Toc137204125)

[c) Schema 7](#_Toc137204126)

[IV. Managed folders 9](#_Toc137204127)

[1. {{$folder.name}} 9](#_Toc137204128)

[a) Status 9](#_Toc137204129)

[b) Derived datasets 10](#_Toc137204130)

[c) Related Recipes 10](#_Toc137204131)

[V. Recipes 11](#_Toc137204132)

[1. {{ $recipe.name }} 11](#_Toc137204133)

[a) Status 11](#_Toc137204134)

[b) Inputs / Outputs 11](#_Toc137204135)

[c) Prepare Steps 12](#_Toc137204136)

[d) Join 12](#_Toc137204137)

[e) Data Relationships 13](#_Toc137204138)

[f) Feature Transformations 14](#_Toc137204139)

[g) Code 14](#_Toc137204140)

[h) Windows 15](#_Toc137204141)

[i) Sorted Columns 15](#_Toc137204142)

[j) Top rows and sorting direction 15](#_Toc137204143)

[k) Parameters 16](#_Toc137204144)

[VI. Saved Models 17](#_Toc137204145)

[1. {{ $model.name }} 17](#_Toc137204146)

[a) Status 17](#_Toc137204147)

[b) Related Recipes 17](#_Toc137204148)

[VII. Model Evaluation Stores 19](#_Toc137204149)

[1. {{ $mes.name }} 19](#_Toc137204150)

[a) Related Recipes 19](#_Toc137204151)

[VIII. Labeling Tasks 21](#_Toc137204152)

[1. {{ $labeling\_task.name }} 21](#_Toc137204153)

[a) Status 21](#_Toc137204154)

[b) Inputs / Outputs 21](#_Toc137204155)

[IX. Deployment and Monitoring 23](#_Toc137204156)

[A. Implementation Details 23](#_Toc137204157)

[B. Version Control 23](#_Toc137204158)

# Project general description

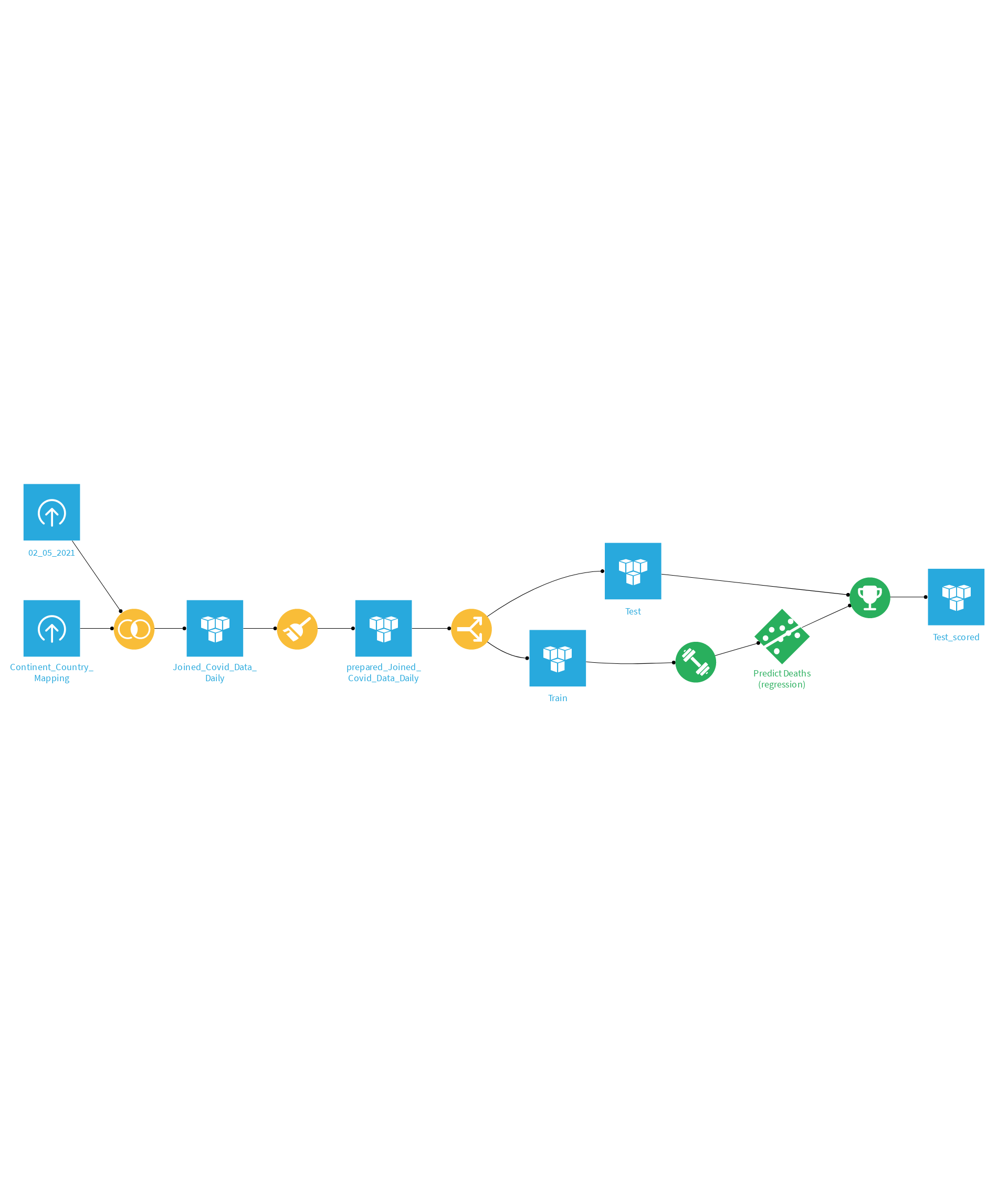
Project references for the created flow:

* Project name: **ML\_Dataiku\_Pipeline\_project**
* Project key: **ML\_DATAIKU\_PIPELINE\_PROJECT**
* Project short description:

The project \*ML\_Dataiku\_Pipeline\_project\* was created by abhishek.barandooru on Aug 29th 2023

* Status: **Sandbox**
* Current git branch: **master**
* Created by abhishek.barandooru on the 2023-08-29 13:09:48
* Last Modification by api:14VF4LADSLYJP81X (deleted) on the 2023-08-29 13:26:34

# Graphical Representation of the Flow



# Datasets

List all the datasets in the flow.

### 02\_05\_2021

Summarized information about your dataset 02\_05\_2021:

#### Status of the dataset 02\_05\_2021

Characteristics of the dataset 02\_05\_2021:

* Type: **Uploaded Files**
* Connection: **dataiku-managed-storage**
* Total Size: **Not computed**
* Record Count: **3,984**
* Created by abhishek.barandooru on the 2023-08-29 13:19:22
* Last Modification by abhishek.barandooru on the 2023-08-29 13:19:22

#### Related Recipes

The dataset 02\_05\_2021 relates to the following recipes.

Successor Recipes

List of recipes with 02\_05\_2021 as an input:

* compute\_Joined\_Covid\_Data\_Daily (Join)

#### Schema

Schema details of the dataset.

|  |  |  |
| --- | --- | --- |
| ****Column Name**** | Column Type | Description |

|  |  |  |
| --- | --- | --- |
| ****FIPS**** | string |  |

|  |  |  |
| --- | --- | --- |
| ****Admin2**** | string |  |

|  |  |  |
| --- | --- | --- |
| ****Province\_State**** | string |  |

|  |  |  |
| --- | --- | --- |
| ****Country\_Region**** | string |  |

|  |  |  |
| --- | --- | --- |
| ****Last\_Update**** | string |  |

|  |  |  |
| --- | --- | --- |
| ****Lat**** | string |  |

|  |  |  |
| --- | --- | --- |
| ****Long\_**** | string |  |

|  |  |  |
| --- | --- | --- |
| ****Confirmed**** | string |  |

|  |  |  |
| --- | --- | --- |
| ****Deaths**** | string |  |

|  |  |  |
| --- | --- | --- |
| ****Recovered**** | string |  |

|  |  |  |
| --- | --- | --- |
| ****Active**** | string |  |

|  |  |  |
| --- | --- | --- |
| ****Combined\_Key**** | string |  |

|  |  |  |
| --- | --- | --- |
| ****Incident\_Rate**** | string |  |

|  |  |  |
| --- | --- | --- |
| ****Case\_Fatality\_Ratio**** | string |  |

### Continent\_Country\_Mapping

Summarized information about your dataset Continent\_Country\_Mapping:

#### Status of the dataset Continent\_Country\_Mapping

Characteristics of the dataset Continent\_Country\_Mapping:

* Type: **Uploaded Files**
* Connection: **dataiku-managed-storage**
* Total Size: **Not computed**
* Record Count: **182**
* Created by abhishek.barandooru on the 2023-08-29 13:16:05
* Last Modification by abhishek.barandooru on the 2023-08-29 13:16:08

#### Related Recipes

The dataset Continent\_Country\_Mapping relates to the following recipes.

Successor Recipes

List of recipes with Continent\_Country\_Mapping as an input:

* compute\_Joined\_Covid\_Data\_Daily (Join)

#### Schema

Schema details of the dataset.

|  |  |  |
| --- | --- | --- |
| ****Column Name**** | Column Type | Description |

|  |  |  |
| --- | --- | --- |
| ****Country\_Name**** | string |  |

|  |  |  |
| --- | --- | --- |
| ****Continent\_Name**** | string |  |

|  |  |  |
| --- | --- | --- |
| ****Continent\_Code**** | string |  |

### Joined\_Covid\_Data\_Daily

Summarized information about your dataset Joined\_Covid\_Data\_Daily:

#### Status of the dataset Joined\_Covid\_Data\_Daily

Characteristics of the dataset Joined\_Covid\_Data\_Daily:

* Type: **Amazon S3**
* Connection: **dataiku-managed-storage**
* Total Size: **271.2 KB**
* Record Count: **3,984**
* Created by abhishek.barandooru on the 2023-08-29 13:26:29
* Last Modification by abhishek.barandooru on the 2023-08-29 13:26:29

#### Related Recipes

The dataset Joined\_Covid\_Data\_Daily relates to the following recipes.

Parent Recipe

List of recipes with Joined\_Covid\_Data\_Daily as an output:

* compute\_Joined\_Covid\_Data\_Daily (Join)

Successor Recipes

List of recipes with Joined\_Covid\_Data\_Daily as an input:

* compute\_prepared\_Joined\_Covid\_Data\_Daily (Prepare)

#### Schema

Schema details of the dataset.

|  |  |  |
| --- | --- | --- |
| ****Column Name**** | Column Type | Description |

|  |  |  |
| --- | --- | --- |
| ****FIPS**** | string |  |

|  |  |  |
| --- | --- | --- |
| ****Admin2**** | string |  |

|  |  |  |
| --- | --- | --- |
| ****Province\_State**** | string |  |

|  |  |  |
| --- | --- | --- |
| ****Country\_Region**** | string |  |

|  |  |  |
| --- | --- | --- |
| ****Last\_Update**** | string |  |

|  |  |  |
| --- | --- | --- |
| ****Lat**** | string |  |

|  |  |  |
| --- | --- | --- |
| ****Long\_**** | string |  |

|  |  |  |
| --- | --- | --- |
| ****Confirmed**** | string |  |

|  |  |  |
| --- | --- | --- |
| ****Deaths**** | string |  |

|  |  |  |
| --- | --- | --- |
| ****Recovered**** | string |  |

|  |  |  |
| --- | --- | --- |
| ****Active**** | string |  |

|  |  |  |
| --- | --- | --- |
| ****Combined\_Key**** | string |  |

|  |  |  |
| --- | --- | --- |
| ****Incident\_Rate**** | string |  |

|  |  |  |
| --- | --- | --- |
| ****Case\_Fatality\_Ratio**** | string |  |

|  |  |  |
| --- | --- | --- |
| ****Country\_Name**** | string |  |

|  |  |  |
| --- | --- | --- |
| ****Continent\_Name**** | string |  |

|  |  |  |
| --- | --- | --- |
| ****Continent\_Code**** | string |  |

### prepared\_Joined\_Covid\_Data\_Daily

Summarized information about your dataset prepared\_Joined\_Covid\_Data\_Daily:

#### Status of the dataset prepared\_Joined\_Covid\_Data\_Daily

Characteristics of the dataset prepared\_Joined\_Covid\_Data\_Daily:

* Type: **Amazon S3**
* Connection: **dataiku-managed-storage**
* Total Size: **214.3 KB**
* Record Count: **Not computed**
* Created by abhishek.barandooru on the 2023-08-29 13:33:23
* Last Modification by abhishek.barandooru on the 2023-08-29 13:33:23

#### Related Recipes

The dataset prepared\_Joined\_Covid\_Data\_Daily relates to the following recipes.

Parent Recipe

List of recipes with prepared\_Joined\_Covid\_Data\_Daily as an output:

* compute\_prepared\_Joined\_Covid\_Data\_Daily (Prepare)

Successor Recipes

List of recipes with prepared\_Joined\_Covid\_Data\_Daily as an input:

* split\_prepared\_Joined\_Covid\_Data\_Daily (Split)

#### Schema

Schema details of the dataset.

|  |  |  |
| --- | --- | --- |
| ****Column Name**** | Column Type | Description |

|  |  |  |
| --- | --- | --- |
| ****FIPS**** | string |  |

|  |  |  |
| --- | --- | --- |
| ****Admin2**** | string |  |

|  |  |  |
| --- | --- | --- |
| ****Province\_State**** | string |  |

|  |  |  |
| --- | --- | --- |
| ****Country\_Region**** | string |  |

|  |  |  |
| --- | --- | --- |
| ****Last\_Update**** | string |  |

|  |  |  |
| --- | --- | --- |
| ****Lat**** | string |  |

|  |  |  |
| --- | --- | --- |
| ****Long\_**** | string |  |

|  |  |  |
| --- | --- | --- |
| ****Confirmed**** | string |  |

|  |  |  |
| --- | --- | --- |
| ****Deaths**** | string |  |

|  |  |  |
| --- | --- | --- |
| ****Recovered**** | string |  |

|  |  |  |
| --- | --- | --- |
| ****Active**** | string |  |

|  |  |  |
| --- | --- | --- |
| ****Combined\_Key**** | string |  |

|  |  |  |
| --- | --- | --- |
| ****Incident\_Rate**** | string |  |

|  |  |  |
| --- | --- | --- |
| ****Case\_Fatality\_Ratio**** | string |  |

|  |  |  |
| --- | --- | --- |
| ****Country\_Name**** | string |  |

|  |  |  |
| --- | --- | --- |
| ****Continent\_Name**** | string |  |

|  |  |  |
| --- | --- | --- |
| ****Continent\_Code**** | string |  |

### Train

Summarized information about your dataset Train:

#### Status of the dataset Train

Characteristics of the dataset Train:

* Type: **Amazon S3**
* Connection: **dataiku-managed-storage**
* Total Size: **156.5 KB**
* Record Count: **2,331**
* Created by abhishek.barandooru on the 2023-08-29 13:38:10
* Last Modification by abhishek.barandooru on the 2023-08-29 13:40:04

#### Related Recipes

The dataset Train relates to the following recipes.

Parent Recipe

List of recipes with Train as an output:

* split\_prepared\_Joined\_Covid\_Data\_Daily (Split)

Successor Recipes

List of recipes with Train as an input:

* train\_Predict\_Deaths\_\_regression\_ (Prediction Training)

#### Schema

Schema details of the dataset.

|  |  |  |
| --- | --- | --- |
| ****Column Name**** | Column Type | Description |

|  |  |  |
| --- | --- | --- |
| ****FIPS**** | string |  |

|  |  |  |
| --- | --- | --- |
| ****Admin2**** | string |  |

|  |  |  |
| --- | --- | --- |
| ****Province\_State**** | string |  |

|  |  |  |
| --- | --- | --- |
| ****Country\_Region**** | string |  |

|  |  |  |
| --- | --- | --- |
| ****Last\_Update**** | string |  |

|  |  |  |
| --- | --- | --- |
| ****Lat**** | string |  |

|  |  |  |
| --- | --- | --- |
| ****Long\_**** | string |  |

|  |  |  |
| --- | --- | --- |
| ****Confirmed**** | string |  |

|  |  |  |
| --- | --- | --- |
| ****Deaths**** | string |  |

|  |  |  |
| --- | --- | --- |
| ****Recovered**** | string |  |

|  |  |  |
| --- | --- | --- |
| ****Active**** | string |  |

|  |  |  |
| --- | --- | --- |
| ****Combined\_Key**** | string |  |

|  |  |  |
| --- | --- | --- |
| ****Incident\_Rate**** | string |  |

|  |  |  |
| --- | --- | --- |
| ****Case\_Fatality\_Ratio**** | string |  |

|  |  |  |
| --- | --- | --- |
| ****Country\_Name**** | string |  |

|  |  |  |
| --- | --- | --- |
| ****Continent\_Name**** | string |  |

|  |  |  |
| --- | --- | --- |
| ****Continent\_Code**** | string |  |

### Test

Summarized information about your dataset Test:

#### Status of the dataset Test

Characteristics of the dataset Test:

* Type: **Amazon S3**
* Connection: **dataiku-managed-storage**
* Total Size: **69.6 KB**
* Record Count: **Not computed**
* Created by abhishek.barandooru on the 2023-08-29 13:38:27
* Last Modification by abhishek.barandooru on the 2023-08-29 13:40:04

#### Related Recipes

The dataset Test relates to the following recipes.

Parent Recipe

List of recipes with Test as an output:

* split\_prepared\_Joined\_Covid\_Data\_Daily (Split)

Successor Recipes

List of recipes with Test as an input:

* score\_Test (Prediction Scoring)

#### Schema

Schema details of the dataset.

|  |  |  |
| --- | --- | --- |
| ****Column Name**** | Column Type | Description |

|  |  |  |
| --- | --- | --- |
| ****FIPS**** | string |  |

|  |  |  |
| --- | --- | --- |
| ****Admin2**** | string |  |

|  |  |  |
| --- | --- | --- |
| ****Province\_State**** | string |  |

|  |  |  |
| --- | --- | --- |
| ****Country\_Region**** | string |  |

|  |  |  |
| --- | --- | --- |
| ****Last\_Update**** | string |  |

|  |  |  |
| --- | --- | --- |
| ****Lat**** | string |  |

|  |  |  |
| --- | --- | --- |
| ****Long\_**** | string |  |

|  |  |  |
| --- | --- | --- |
| ****Confirmed**** | string |  |

|  |  |  |
| --- | --- | --- |
| ****Deaths**** | string |  |

|  |  |  |
| --- | --- | --- |
| ****Recovered**** | string |  |

|  |  |  |
| --- | --- | --- |
| ****Active**** | string |  |

|  |  |  |
| --- | --- | --- |
| ****Combined\_Key**** | string |  |

|  |  |  |
| --- | --- | --- |
| ****Incident\_Rate**** | string |  |

|  |  |  |
| --- | --- | --- |
| ****Case\_Fatality\_Ratio**** | string |  |

|  |  |  |
| --- | --- | --- |
| ****Country\_Name**** | string |  |

|  |  |  |
| --- | --- | --- |
| ****Continent\_Name**** | string |  |

|  |  |  |
| --- | --- | --- |
| ****Continent\_Code**** | string |  |

### Test\_scored

Summarized information about your dataset Test\_scored:

#### Status of the dataset Test\_scored

Characteristics of the dataset Test\_scored:

* Type: **Amazon S3**
* Connection: **dataiku-managed-storage**
* Total Size: **59.6 KB**
* Record Count: **946**
* Created by abhishek.barandooru on the 2023-08-29 14:02:00
* Last Modification by abhishek.barandooru on the 2023-08-29 14:02:00

#### Related Recipes

The dataset Test\_scored relates to the following recipes.

Parent Recipe

List of recipes with Test\_scored as an output:

* score\_Test (Prediction Scoring)

#### Schema

Schema details of the dataset.

|  |  |  |
| --- | --- | --- |
| ****Column Name**** | Column Type | Description |

|  |  |  |
| --- | --- | --- |
| ****FIPS**** | bigint |  |

|  |  |  |
| --- | --- | --- |
| ****Admin2**** | string |  |

|  |  |  |
| --- | --- | --- |
| ****Province\_State**** | string |  |

|  |  |  |
| --- | --- | --- |
| ****Country\_Region**** | string |  |

|  |  |  |
| --- | --- | --- |
| ****Last\_Update**** | string |  |

|  |  |  |
| --- | --- | --- |
| ****Lat**** | double |  |

|  |  |  |
| --- | --- | --- |
| ****Long\_**** | double |  |

|  |  |  |
| --- | --- | --- |
| ****Confirmed**** | bigint |  |

|  |  |  |
| --- | --- | --- |
| ****Deaths**** | bigint |  |

|  |  |  |
| --- | --- | --- |
| ****Recovered**** | bigint |  |

|  |  |  |
| --- | --- | --- |
| ****Active**** | bigint |  |

|  |  |  |
| --- | --- | --- |
| ****Combined\_Key**** | string |  |

|  |  |  |
| --- | --- | --- |
| ****Incident\_Rate**** | double |  |

|  |  |  |
| --- | --- | --- |
| ****Case\_Fatality\_Ratio**** | double |  |

|  |  |  |
| --- | --- | --- |
| ****Country\_Name**** | string |  |

|  |  |  |
| --- | --- | --- |
| ****Continent\_Name**** | string |  |

|  |  |  |
| --- | --- | --- |
| ****Continent\_Code**** | string |  |

|  |  |  |
| --- | --- | --- |
| ****prediction**** | float |  |

# Recipes

List all the recipes in the flow.

### compute\_Joined\_Covid\_Data\_Daily

Summarized information about the recipe compute\_Joined\_Covid\_Data\_Daily.

#### Status

Characteristics of the recipe compute\_Joined\_Covid\_Data\_Daily:

* Type: **Join**
* Created by abhishek.barandooru on the 2023-08-29 13:26:29
* Last Modification by abhishek.barandooru on the 2023-08-29 13:30:03

#### Inputs / Outputs

List of the inputs of the recipe:

* Dataset 02\_05\_2021 (Uploaded Files)
* Dataset Continent\_Country\_Mapping (Uploaded Files)

List of the outputs of the recipe:

* Dataset Joined\_Covid\_Data\_Daily (Amazon S3)

#### Join

Left 02\_05\_2021 - Continent\_Country\_Mapping And

|  |  |  |
| --- | --- | --- |
| Left input | Condition | Right input |

|  |  |  |
| --- | --- | --- |
| Country\_Region | = | Country\_Name |

### compute\_prepared\_Joined\_Covid\_Data\_Daily

Summarized information about the recipe compute\_prepared\_Joined\_Covid\_Data\_Daily.

#### Status

Characteristics of the recipe compute\_prepared\_Joined\_Covid\_Data\_Daily:

* Type: **Prepare**
* Created by abhishek.barandooru on the 2023-08-29 13:33:23
* Last Modification by abhishek.barandooru on the 2023-08-29 13:35:37

#### Inputs / Outputs

List of the inputs of the recipe:

* Dataset Joined\_Covid\_Data\_Daily (Amazon S3)

List of the outputs of the recipe:

* Dataset prepared\_Joined\_Covid\_Data\_Daily (Amazon S3)

#### Prepare Steps

|  |  |
| --- | --- |
| Type | Comment |

|  |  |
| --- | --- |
| FilterOnValue |  |

### split\_prepared\_Joined\_Covid\_Data\_Daily

Summarized information about the recipe split\_prepared\_Joined\_Covid\_Data\_Daily.

#### Status

Characteristics of the recipe split\_prepared\_Joined\_Covid\_Data\_Daily:

* Type: **Split**
* Created by abhishek.barandooru on the 2023-08-29 13:38:46
* Last Modification by abhishek.barandooru on the 2023-08-29 13:40:04

#### Inputs / Outputs

List of the inputs of the recipe:

* Dataset prepared\_Joined\_Covid\_Data\_Daily (Amazon S3)

List of the outputs of the recipe:

* Dataset Train (Amazon S3)
* Dataset Test (Amazon S3)

Split mode: Randomly split to the output datasets with exact ratio

### train\_Predict\_Deaths\_\_regression\_

Summarized information about the recipe train\_Predict\_Deaths\_\_regression\_.

#### Status

Characteristics of the recipe train\_Predict\_Deaths\_\_regression\_:

* Type: **Prediction Training**
* Created by abhishek.barandooru on the 2023-08-29 13:53:16
* Last Modification by abhishek.barandooru on the 2023-08-29 13:53:16

#### Inputs / Outputs

List of the inputs of the recipe:

* Dataset Train (Amazon S3)

List of the outputs of the recipe:

* Saved Model Predict Deaths (regression) (Gradient Boosted Trees (Session-1) - v1)

### score\_Test

Summarized information about the recipe score\_Test.

#### Status

Characteristics of the recipe score\_Test:

* Type: **Prediction Scoring**
* Created by abhishek.barandooru on the 2023-08-29 14:02:00
* Last Modification by abhishek.barandooru on the 2023-08-29 14:02:24

#### Inputs / Outputs

List of the inputs of the recipe:

* Dataset Test (Amazon S3)
* Saved Model Predict Deaths (regression) (Gradient Boosted Trees (Session-1) - v1)

List of the outputs of the recipe:

* Dataset Test\_scored (Amazon S3)

# Saved Models

List all the saved models in the flow.

### Predict Deaths (regression)

Summarized information about the saved model Predict Deaths (regression):.

#### Status

Model id: JW3hAI5S

Active version: Gradient Boosted Trees (Session-1) - v1

#### Related Recipes

List of the recipes related to the saved model.

Parent Recipe

Recipe with Predict Deaths (regression) as an output:

* train\_Predict\_Deaths\_\_regression\_ (Prediction Training)

Successor Recipes

List of recipes with Predict Deaths (regression) as an input:

* score\_Test (Prediction Scoring)

# Deployment and Monitoring

## Implementation Details

* The flow can be found here: https://dss-81f24b5d-126eabfb-dku.eu-west-3.app.dataiku.io/projects/ML\_DATAIKU\_PIPELINE\_PROJECT/
* The name of the generated file is: Dataiku Flow Documentation - ML\_DATAIKU\_PIPELINE\_PROJECT.docx

## Version Control

* The flow was executed with the following version of DSS: 12.1.3