ML_Dataiku_Pipeline_project

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

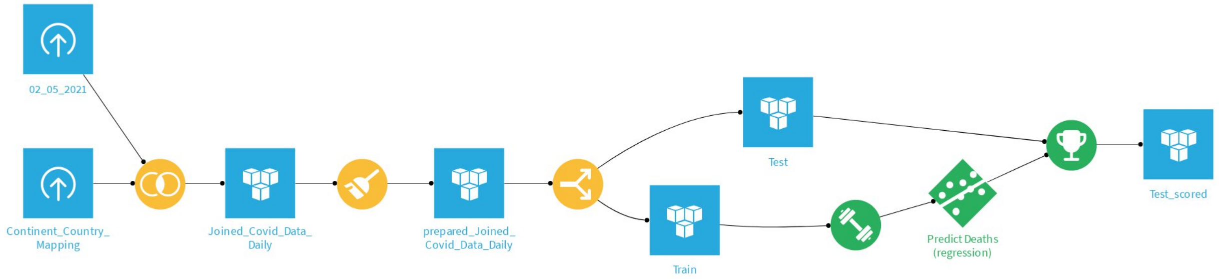
Table of contents

I. 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

II. Graphical Representation of the Flow


02_05_2021
Continent_Country_Mapping
Joined_Covid_Data_Daily
prepared_Joined_Covid_Data_Daily
Test
Train
Predict Deaths (regression)
Test_scored

III. Datasets

List all the datasets in the flow.

1. 02_05_2021

Summarized information about your dataset 02_05_2021:

a) 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

b) 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)*

c) 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	

2. Continent_Country_Mapping

Summarized information about your dataset Continent_Country_Mapping:

a) 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

b) 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)*

c) Schema

Schema details of the dataset.

Column Name	Column Type	Description
Country_Name	string	
Continent_Name	string	
Continent_Code	string	

3. Joined_Covid_Data_Daily

Summarized information about your dataset Joined_Covid_Data_Daily:

a) 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

b) 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)*

c) 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	

4. prepared_Joined_Covid_Data_Daily

Summarized information about your dataset prepared_Joined_Covid_Data_Daily:

a) 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

b) 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)*

c) 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	

5. Train

Summarized information about your dataset Train:

a) 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

b) 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)*

c) 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	

6. Test

Summarized information about your dataset Test:

a) 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

b) 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)*

c) 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	

7. Test_scored

Summarized information about your dataset Test_scored:

a) 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

b) 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)*

c) 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	

IV. Recipes

List all the recipes in the flow.

1. compute_Joined_Covid_Data_Daily

Summarized information about the recipe compute_Joined_Covid_Data_Daily.

a) 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

b) 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)*

c) Join

Left 02_05_2021 - Continent_Country_Mapping And

Left input	Condition	Right input
Country_Region	=	Country_Name

2. compute_prepared_Joined_Covid_Data_Daily

Summarized information about the recipe compute_prepared_Joined_Covid_Data_Daily.

a) 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

b) 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)*

c) Prepare Steps

Type	**Comment**
FilterOnValue	

3. split_prepared_Joined_Covid_Data_Daily

Summarized information about the recipe split_prepared_Joined_Covid_Data_Daily.

a) 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

b) 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

4. train_Predict_Deaths__regression_

Summarized information about the recipe train_Predict_Deaths__regression_.

a) 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

b) 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)*

5. score_Test

Summarized information about the recipe score_Test.

a) 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

b) 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)*

V. Saved Models

List all the saved models in the flow.

1. Predict Deaths (regression)

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

a) Status

Model id: **JW3hAI5S**
Active version: **Gradient Boosted Trees (Session-1) - v1**

b) 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)*

VI. Deployment and Monitoring

A. 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

B. Version Control

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