
NUDRAT ABBAS · UPDATED 9 DAYS AGO

28
Code
Download

Hospital Records for Data Cleaning (Medium)

clinical terminology with missing values, date anomalies, and inconsistent label

Data Card
Code (3)
Discussion (0)
Suggestions (0)

About Dataset

This dataset is designed for intermediate learners who want hands-on experience cleaning healthcare-style data.


Patient records are synthetic and anonymized, while diagnoses are based on real clinical terminology inspired by public biomedical vocabularies (e.g., MeSH).

The dataset intentionally contains logical inconsistencies and missing fields commonly found in healthcare data systems.

Usability ⓘ
10.00

License
CC0: Public Domain

Expected update frequency
Annually



We will use Azure ML Designer to develop a model that will help us with this dataset for hospitals. In this case we will develop a model to predict the cases in a month given the diagnosis, day and month.

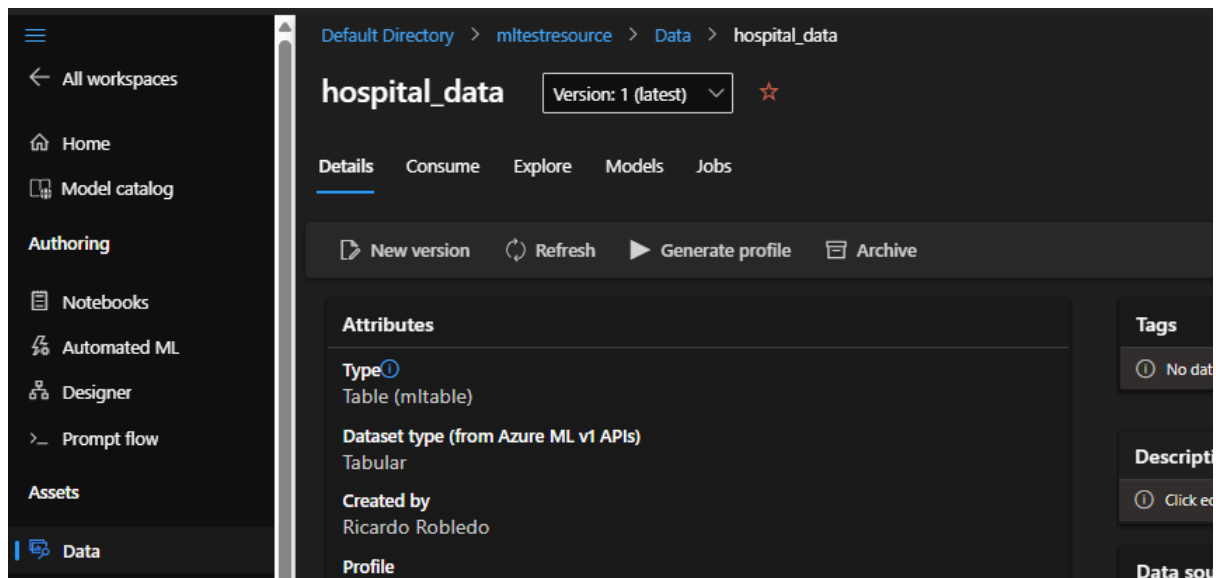
We need to start or create a Compute instance. You need to go to Azure ML Workspace and the select Compute option, then turn on or create your instance.

Choose from a selection of CPU or GPU instances preconfigured with popular tools such as VS Code, JupyterLab, Jupyter, and RStudio, ML packages, deep le
about compute instances

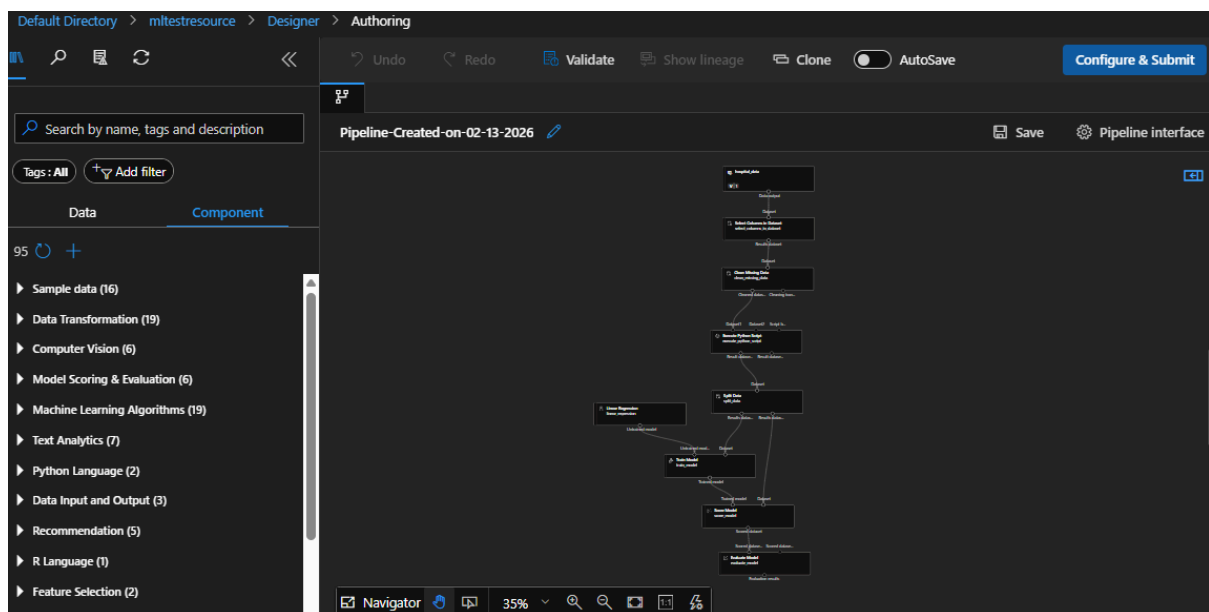
+ New
Refresh
Start
Stop
Restart
Schedule and idle shutdown
Delete
Reset view

<input type="radio"/>	Name ↑	☆ State	Idle shutdown ↓ ⓘ	Applications ↓ ⓘ
<input checked="" type="radio"/>	samplecomputeinstance	Stopped ⓘ	1 hour	JupyterLab Jupyter VS Code (Web) ...
<input type="radio"/>	simplecomputeinstance2	Stopped ⓘ	1 hour	JupyterLab Jupyter VS Code (Web) ...

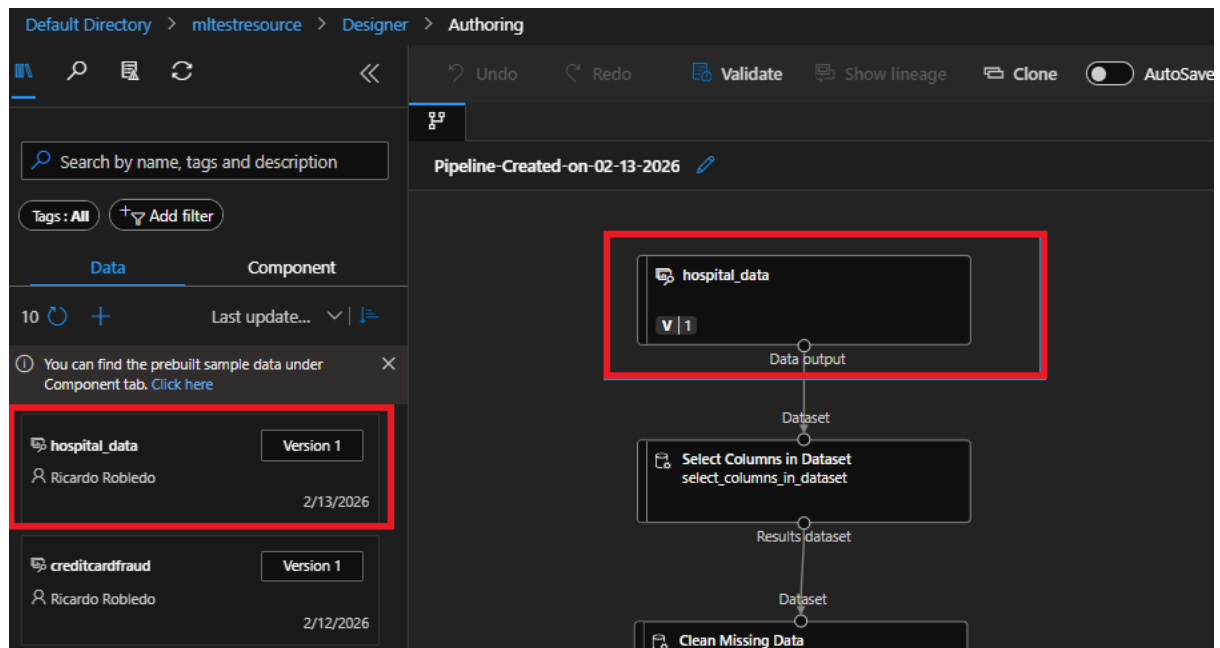
Then you need to have a dataset, for that you need to have a storage account and upload the dataset in a container instance, after that create a data asset with that file.



Then we go to Designer and create a template to create our model. In this case I have a pipeline that I will explain.



Go to data and select the dataset.



Ok, it is important to know which columns we will use, in my case I used Azure Data Explorer and Kusto to explore data, we will make a model that predicts the total count of cases depending on the diagnosis.

Run Recall KQL tools MyFreeCluster/MyDatabase

```

1 hospitaltesttable
2 | extend Year = toint(format_datetime(AdmissionDate, "yyyy")),
3 | extend Month = toint(format_datetime(AdmissionDate, "MM"))
4 | summarize Cases = count()
5 | by Diagnosis, Year, Month

```

Table 1 Add visual Stats Search 2026-02-13 15:38 (UTC)

Diagnosis	Year	Month	Cases
> Myocardial Infarction	2,024	3	11
> Pneumonia	2,024	8	8
> Influenza	2,024	11	12
> Acute Bronchitis	2,025	7	14
> Type 2 Diabetes	2,023	8	9
> Gastroenteritis	2,024	8	10

Now we must go to select components and drag Select Columns in Dataset, after that select the next columns.

select columns

Tags: All Add filter

Data Component

Most relevant

Select Columns Transform
Microsoft
Create a transformation that selects the same subset of columns as in the given dataset. [Learn More]
(https://aka.ms/aml/transformations/selectcolumns)
azureml.Design: true 3/20/2025

Select Columns in Dataset
Microsoft
Select columns to include or exclude from a dataset in an operation. [Learn More](https://aka.ms/aml/selectcolumns)
azureml.Design: true 3/20/2025

2/12/2026

Pipeline-Created-on-02-13-2026

hospital_data

Data output

Dataset

Select Columns in Dataset
select_columns_in_dataset

Results dataset

Dataset

Clean Missing Data
clean_missing_data

Clean Missing Data

Select Columns in Dataset

Select columns

Column names: AdmissionDate,Diagnosis

Edit column

Output settings

Input settings

Run settings

Node information

Component information

Now we need to use the component Clean Missing Data. Select column Diagnosis, replace with mode and not generate another column.

Pipeline-Created-on-02-13-2026

Save Pipeline interface

Data output

Dataset

Select Columns in Dataset
select_columns_in_dataset

Results dataset

Dataset

Clean Missing Data
clean_missing_data

Cleaned dataset... Clearing tran...

Dataset1 Dataset2 Script b...

Execute Python Script

Clean Missing Data

Columns to be cleaned

Column names: Diagnosis

Edit column

Minimum missing value ratio

0.0

Maximum missing value ratio

1.0

Cleaning mode

Replace with mode

Generate missing value indicator column

False

Cols with all missing values

Remove

The next is to put a Execute Python Script module, the code is the next:

```
import pandas as pd

def azureml_main(dataframe1=None, dataframe2=None):

    # Copiar solo columnas necesarias
    df = dataframe1[["Diagnosis", "AdmissionDate"]].copy()
```

```

# Convertir a datetime
df["AdmissionDate"] = pd.to_datetime(df["AdmissionDate"])

# Crear Year y Month
df["Year"] = df["AdmissionDate"].dt.year
df["Month"] = df["AdmissionDate"].dt.month

# Agrupar y contar
result = (
    df.groupby(["Diagnosis", "Year", "Month"])
      .size()
      .reset_index(name="Cases")
      .sort_values(["Year", "Month", "Cases"],
                  ascending=[True, True, False])
)

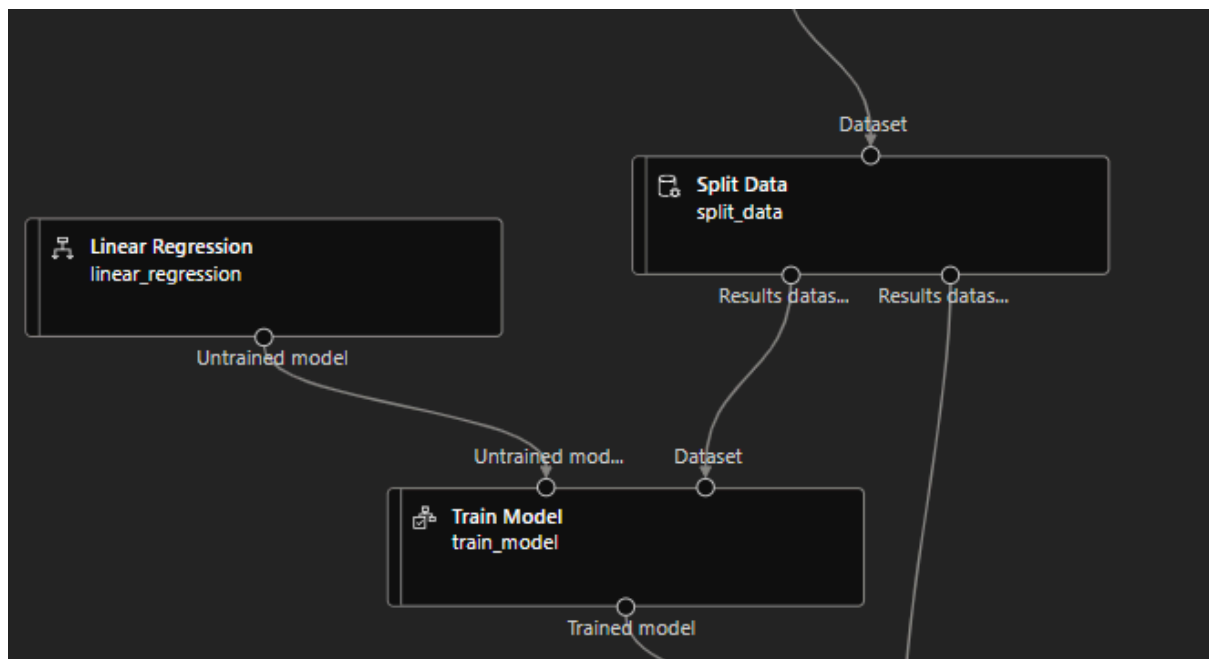
# Devolver SOLO las columnas deseadas explícitamente
result = result[["Diagnosis", "Year", "Month", "Cases"]]

return result,

```

In a brief summary it only group the diagnosis by month and year, then we only return the columns that we will use to predict the Cases.

Ok, we will use 3 modules, they are: Split Data, Liner Regression and Train Model.



These are the settings for Split Data. Like we have temporal data, we don't need randomized split neither stratified split.

The screenshot shows the 'Split Data' module settings. On the left, a workflow diagram shows a 'Dataset' node connected to a 'Split Data' module (labeled 'split_data'). The module has two output ports labeled 'Results datas...'. The right panel shows the following settings:

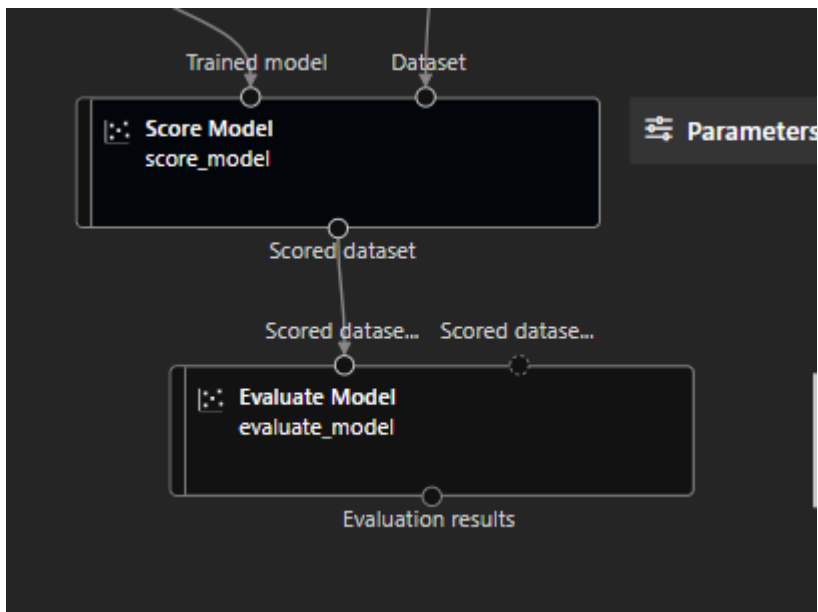
- Splitting mode**: Split Rows
- Fraction of rows in the first output dataset**: 0.7
- Randomized split**: True
- Random seed**: 0
- Stratified split**: False

Train Model is very simple, only select the target that is Cases.

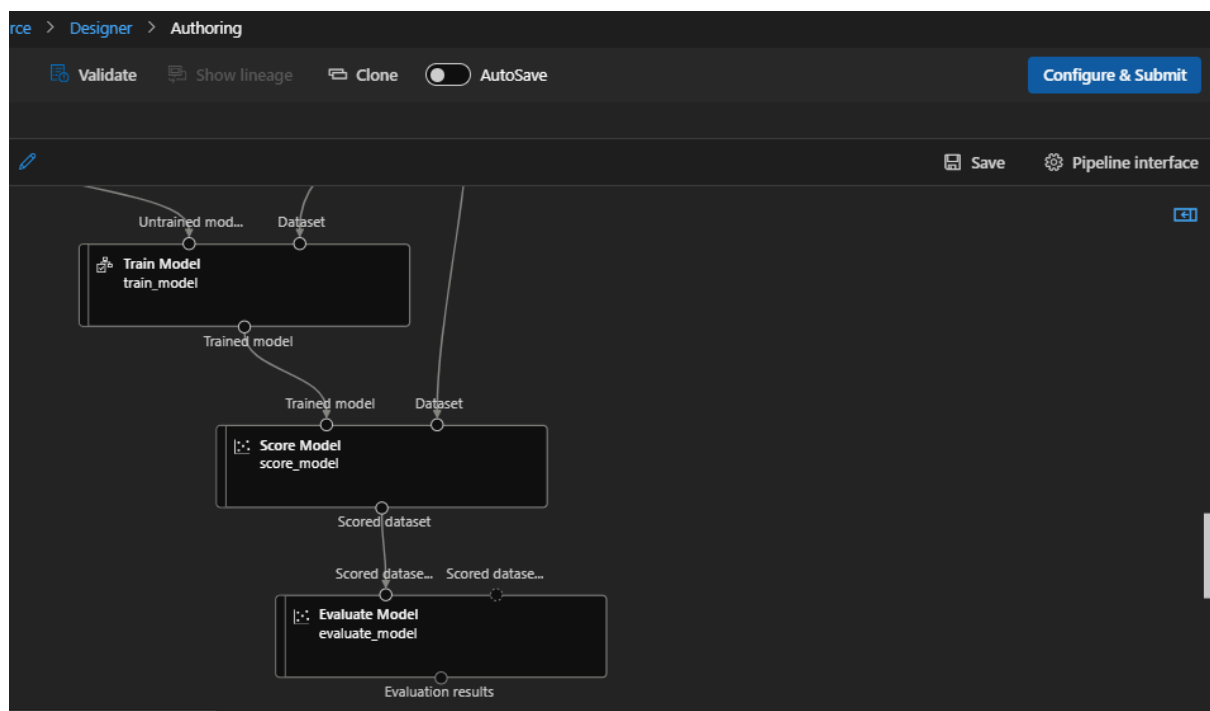
The screenshot shows the 'Train Model' module settings. On the left, a workflow diagram shows an 'Untrained mod...' node connected to a 'Train Model' module (labeled 'train_model'). The module has two output ports labeled 'Results datas...' and 'Trained model'. The right panel shows the following settings:

- Label column**: Column names: Cases
- Model explanations**: False
- Output settings**: >

After that we need only 2 modules that are Score Model and Evaluate Model.



Very well, we only need to execute our pipeline, to that push Configure & Submit.



Give it a name to the experiment.

Set up pipeline job

1 Basics

2 Inputs & outputs

3 Runtime settings

4 Review + Submit

Basics

Experiment name
☐ Select existing ☒ Create new

New experiment name *

Job display name

Job description

Job tags

:

Add

Review + Submit

Back

Next

Close

Select Compute instance, and Review + Submit.

Set up pipeline job

✓ Basics

✓ Inputs & outputs

3 Runtime settings

4 Review + Submit

Runtime settings

Default compute ⓘ

✗ The pipeline compute target samplecomputeinstance is invalid.

Select compute type

Select Azure ML compute instance

[Create Azure ML compute instance](#) [Refresh Compute](#)

> Identity

Default datastore ⓘ

Select datastore *

Advanced settings
☒ Continue on step failure ⓘ

Review + Submit

Back

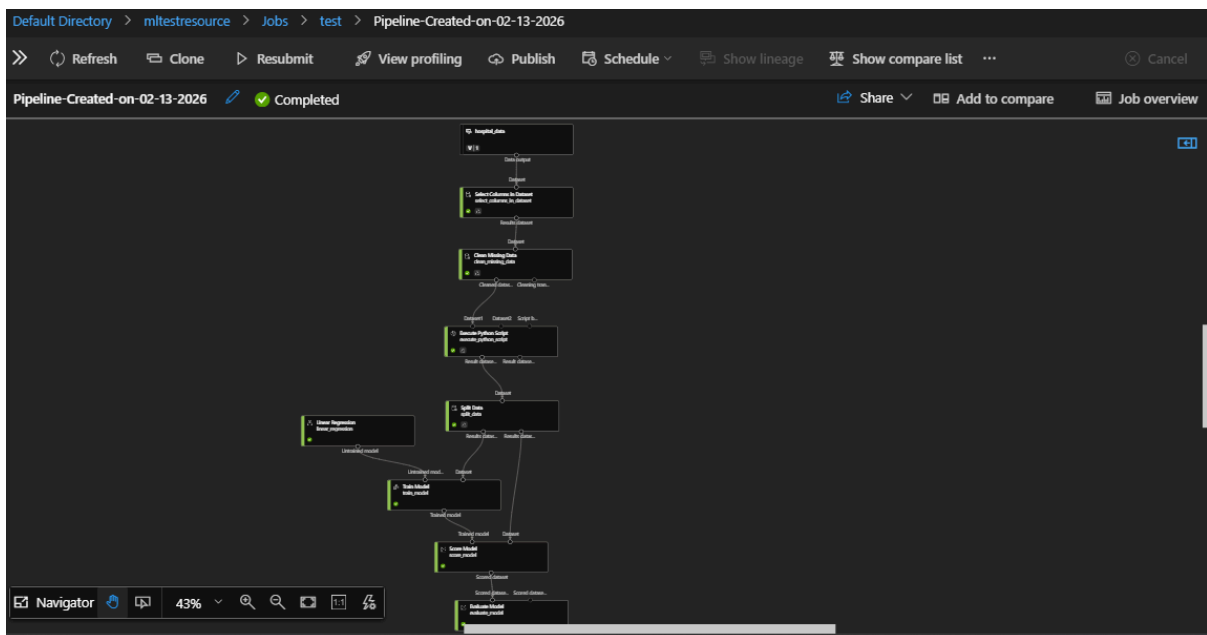
Next

Close

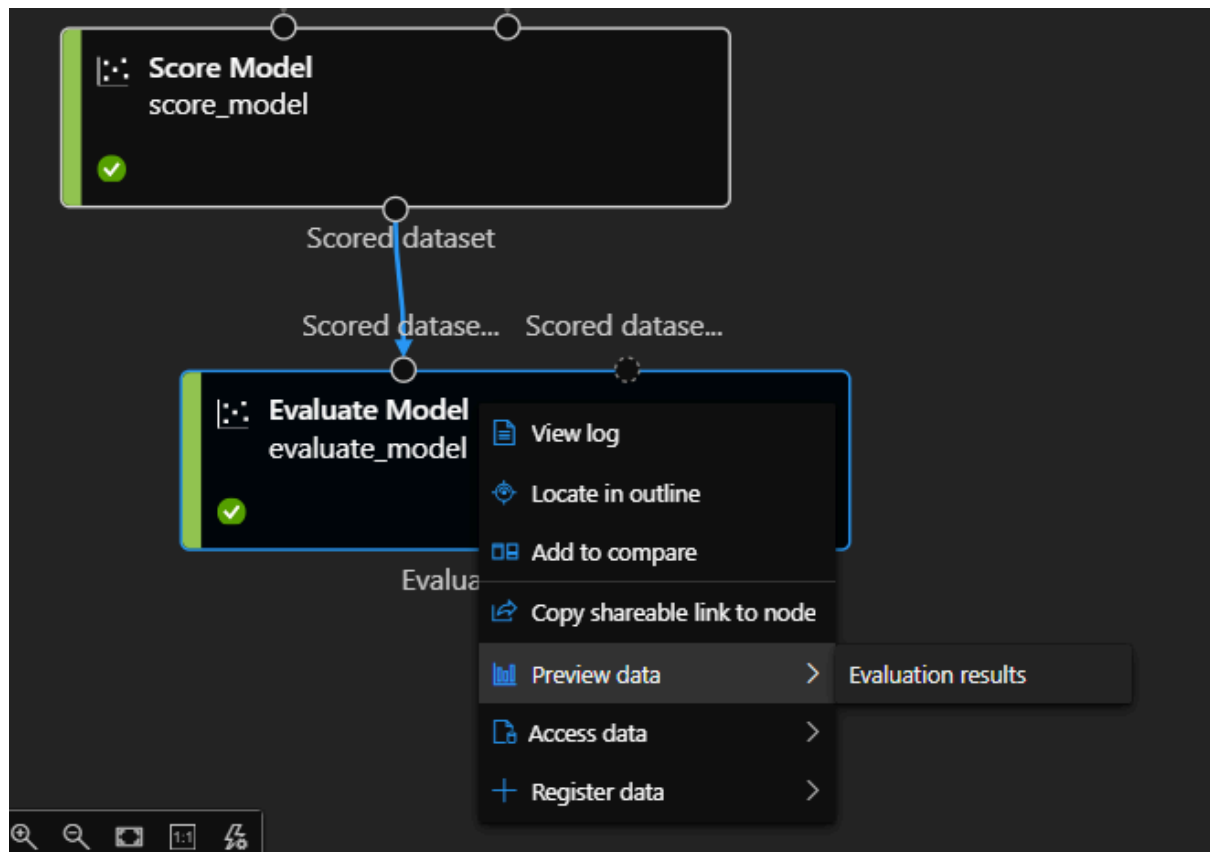
after that the pipeline will be created, so after some minutes we can go to Jobs in the menu and Select pipeline.

Display name (7 visualized)	Parent job name	Status	Created on ↓	Duration
Pipeline-Created-on-02-13-202 (8)		Completed	Feb 13, 2026 9:33 AM	2m 7s
Pipeline-Created-on-02-13-202 (8)		Completed	Feb 13, 2026 9:28 AM	2m 56s
Pipeline-Created-on-02-13-202 (8)		Completed	Feb 13, 2026 9:15 AM	4m 0s
Pipeline-Created-on-02-11-202 (8)		Completed	Feb 13, 2026 9:13 AM	3m 5s
epic_jicama_c2ldq2oxmm (7)		Completed	Feb 12, 2026 8:29 AM	17m 28s
Pipeline-Created-on-02-11-202 (6)		Failed	Feb 11, 2026 8:20 AM	3m 34s
Pipeline-Created-on-01-29-202 (1)		Completed	Jan 31, 2026 10:42 AM	55s

OK!, this is good, all the modules have executed successfully.



Find the Evaluate Model module and do right click and select the next option..



These are our results.

Evaluation_results					
Rows		Columns			
1		5			
Mean_Absolute_Error		Root_Mean_Squared_Error		Relative_Squared_Error	
2.0259		2.806342		0.299339	
Relative_Absolute_Error		Coefficient_of_Det			
0.471088		0.700661			