AI on Cloud

depa Microsoft AIT Asian Institute of Technology SCB ACADEMY

AI on Cloud

Chapter 2

การเรียนรู้ของเครื่องจักร Machine Learning

depa Microsoft AIT Asian Institute of Technology SCB ACADEMY

Lab

4

สร้าง Clustering Model ด้วย Azure Machine Learning Designer

Clustering

- *Clustering* is a form of machine learning that is used to group similar items into clusters based on their features.
- Clustering is an example of *unsupervised* machine learning. There is no previously known cluster value (or *label*) from which to train the model.

https://docs.microsoft.com/en-us/learn/modules/create-clustering-model-azure-machine-learning-designer/1-introduction

Lab4: สร้าง Clustering Model ด้วย Azure Machine Learning Designer

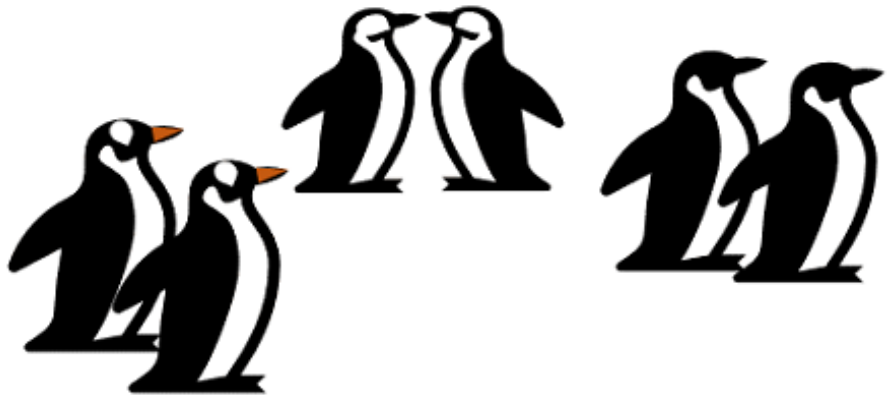
เป้าหมาย:

ใช้ Azure ML Designer สร้าง Clustering Models

- สร้างและ Run Training Pipeline
- สร้าง Inference Pipeline
- Deploy Predictive Model

- Tool: Azure ML Designer
- Model: K-Means Clustering
- Metric:
  - Average Distance to Other Center,
  - Average Distance to Cluster Center,
  - Number of Points,
  - Maximal Distance to Cluster Center

Case Study และ ข้อมูลที่ใช้



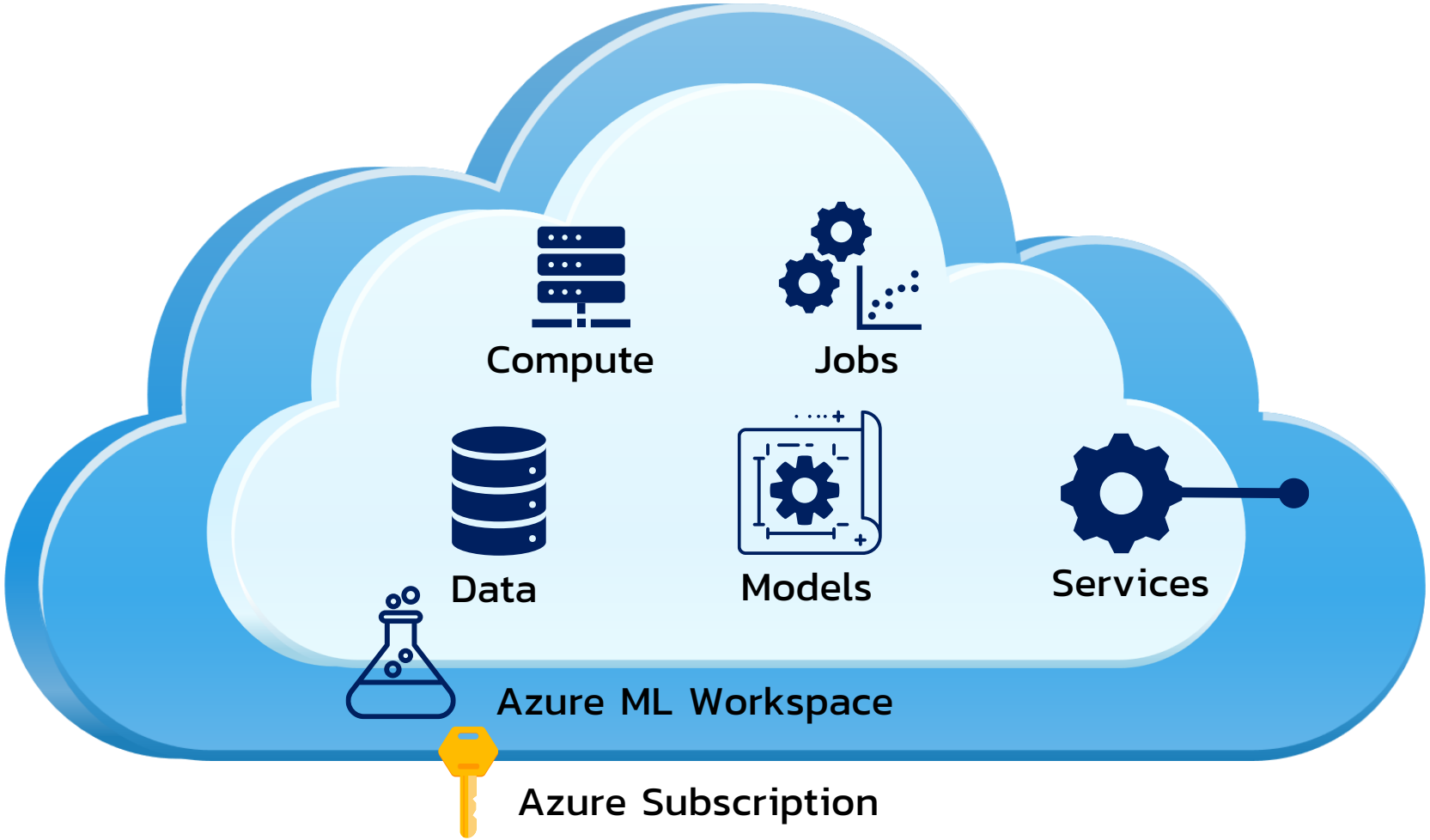
Dataset: penguin-data
(https://aka.ms/penguin-data)

Goal:

- เพื่อจัดกลุ่มเพนกวินออกเป็นกลุ่มๆ โดยอาศัยข้อมูลเกี่ยวกับ ขนาดและสัดส่วนของเพนกวิน
- ไม่มี Label หรือข้อมูลเกี่ยวกับการจัดกลุ่มมาให้

Review: Azure Machine Learning

A cloud-based platform for machine learning



Machine learning in Microsoft Azure

Azure Machine Learning provides the following features and capabilities:

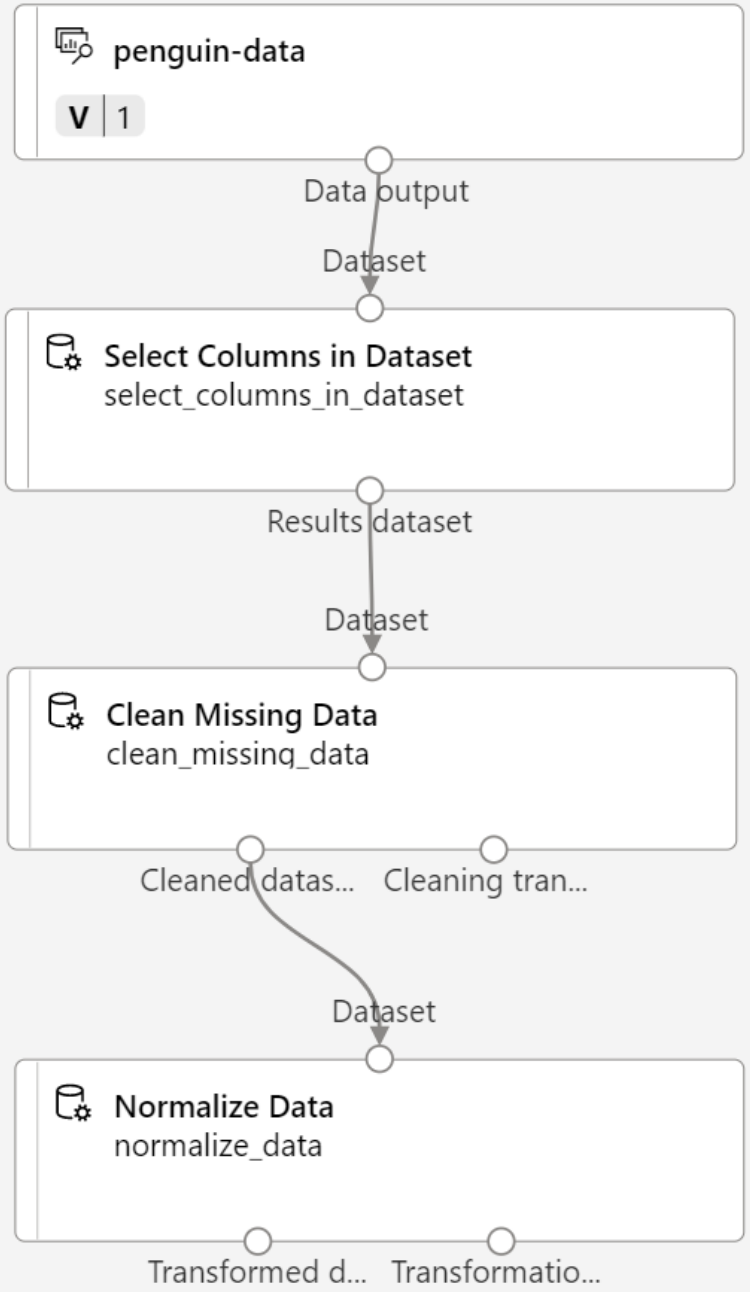
Feature	Capability
Automated machine learning	This feature enables non-experts to quickly create an effective machine learning model from data.
Azure Machine Learning designer	A graphical interface enabling no-code development of machine learning solutions.
Data and compute management	Cloud-based data storage and compute resources that professional data scientists can use to run data experiment code at scale.
Pipelines	Data scientists, software engineers, and IT operations professionals can define pipelines to orchestrate model training, deployment, and management tasks.

https://docs.microsoft.com/en-us/learn/modules/get-started-ai-fundamentals/2-understand-machine-learn

Azure Machine Learning Designer

Visual tool for creating a machine learning *pipeline*

- Part 1: Use a *training pipeline* to train and evaluate a model
- Part 2: Create an *inference pipeline* to predict labels from new data
- Part 3: Deploy the inference pipeline as a *service* for apps to use



Part

1

Use a Training Pipeline to Train and Evaluate a Model

Web URL: *https://aka.ms/penguin-data*
Name: *penguin-data*

Step 4: สร้าง Pipeline สำหรับ Manage Dataset

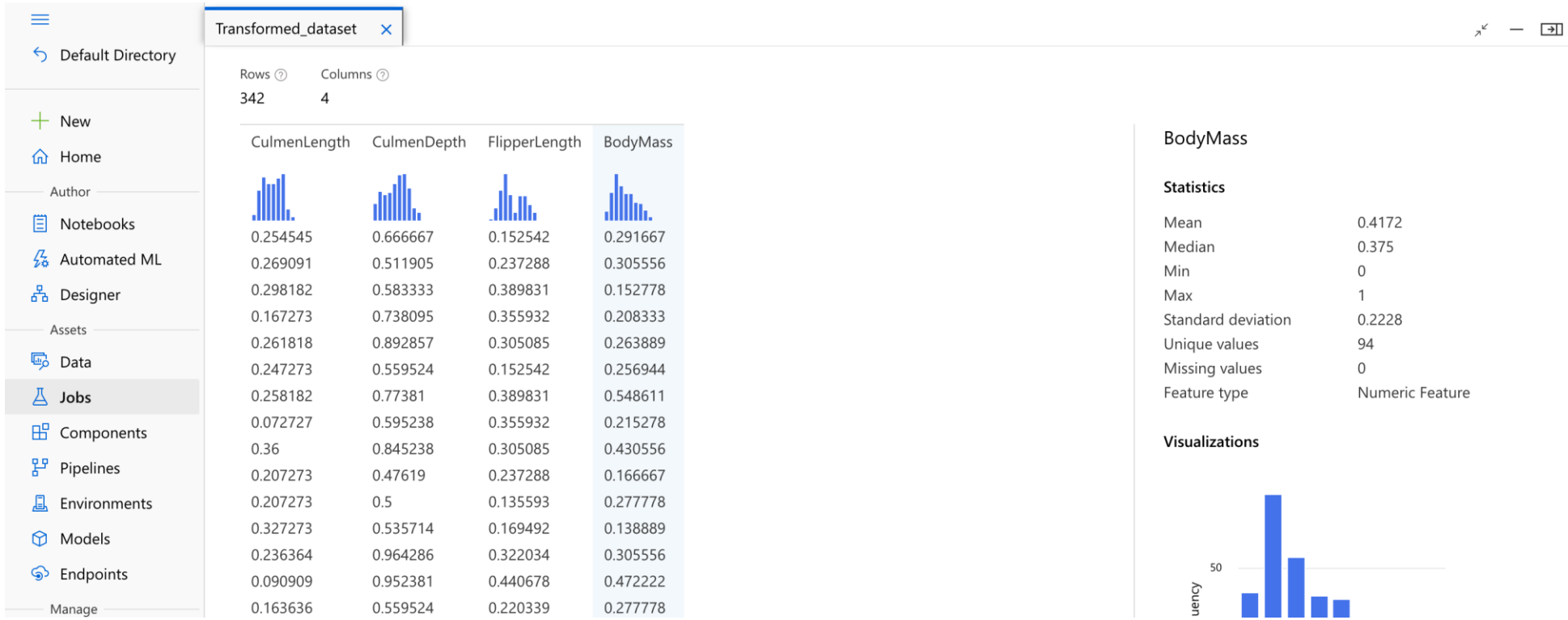
Step 5: ทำ Data Transformation เพื่อจัดการข้อมูลให้พร้อมสำหรับการสร้างโมเดล

- ทำความเข้าใจและสำรวจข้อมูล
- Select Columns:
  - CulmenLength,
  - CulmenDepth,
  - FlipperLength,
  - BodyMass

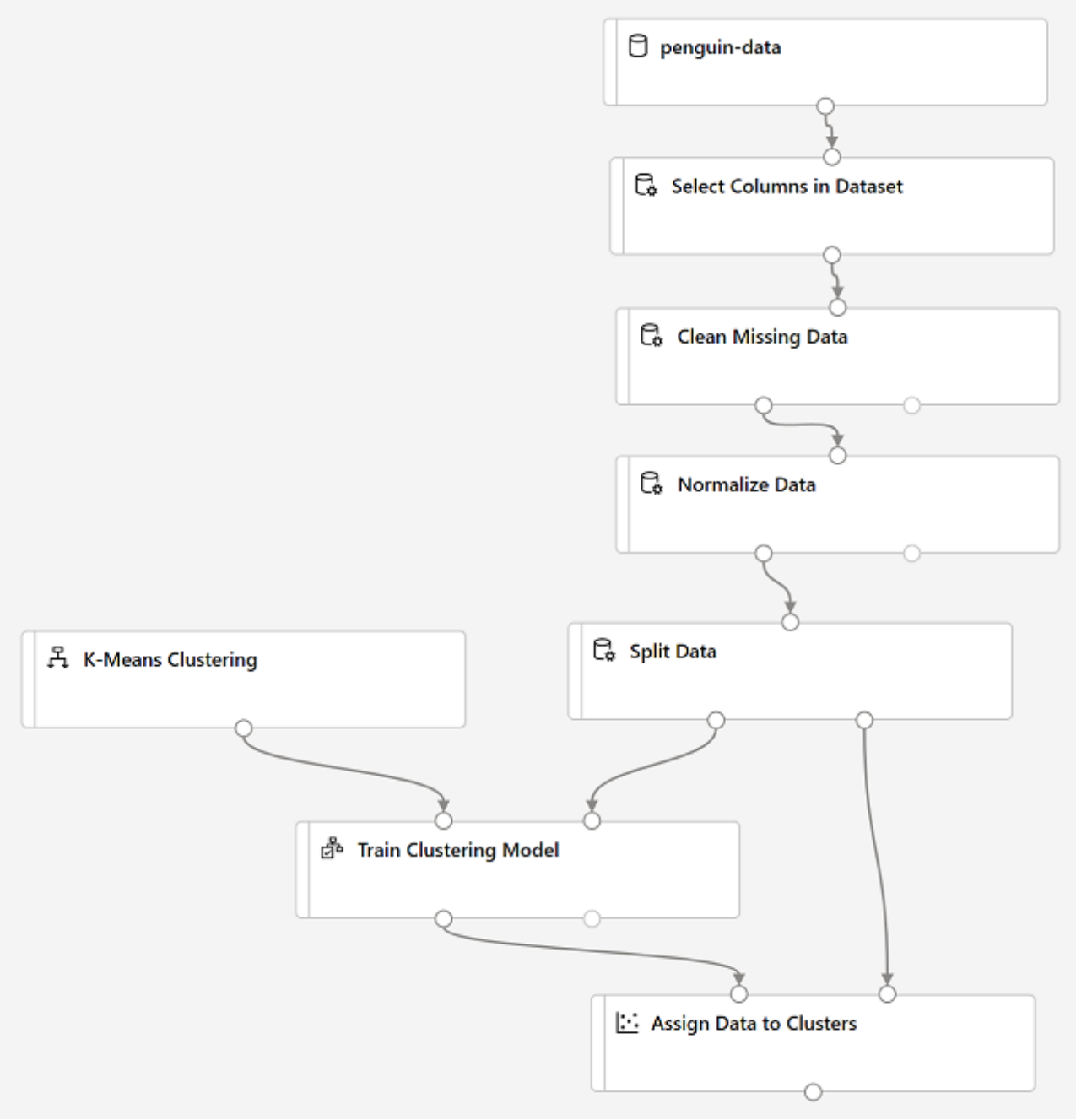
- Clean Missing Values
- Normalize ข้อมูล ด้วยวิธี MinMax สำหรับ Features ต่อไปนี้
  - CulmenLength,
  - CulmenDepth,
  - FlipperLength,
  - BodyMass

Step 6: Run the Pipeline และ Explore ผลลัพธ์

Dataset ได้ถูก Prepare และพร้อมสำหรับการสร้าง Model

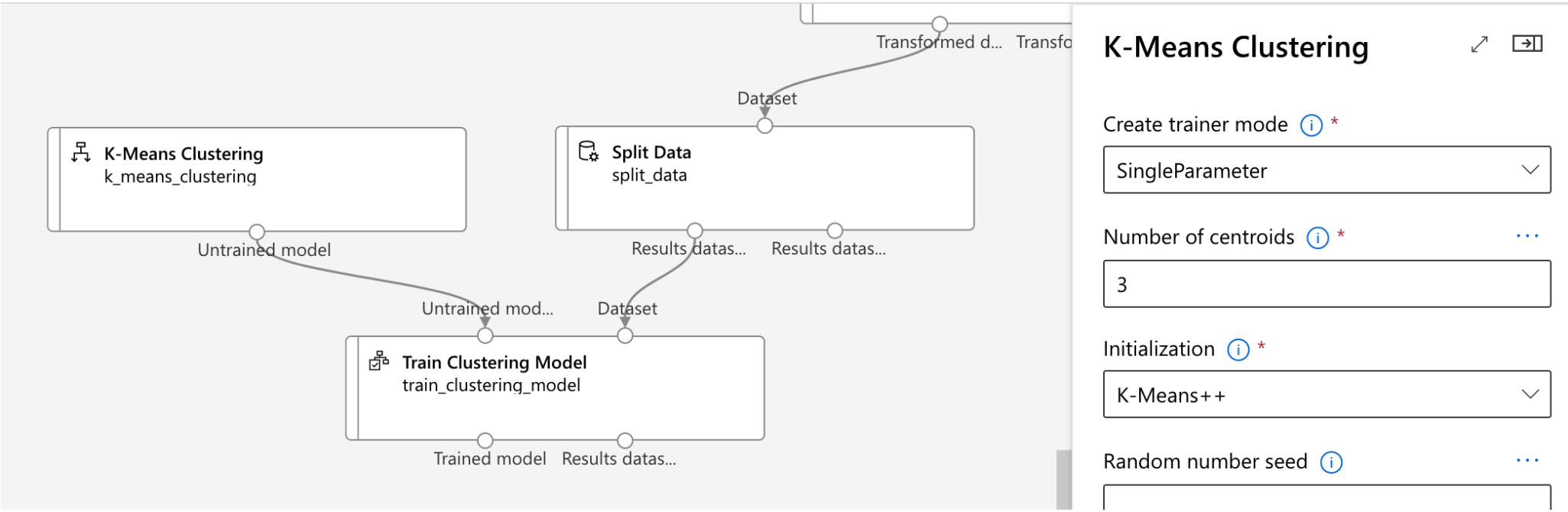


Step 7: Create Training Pipeline for Model Training



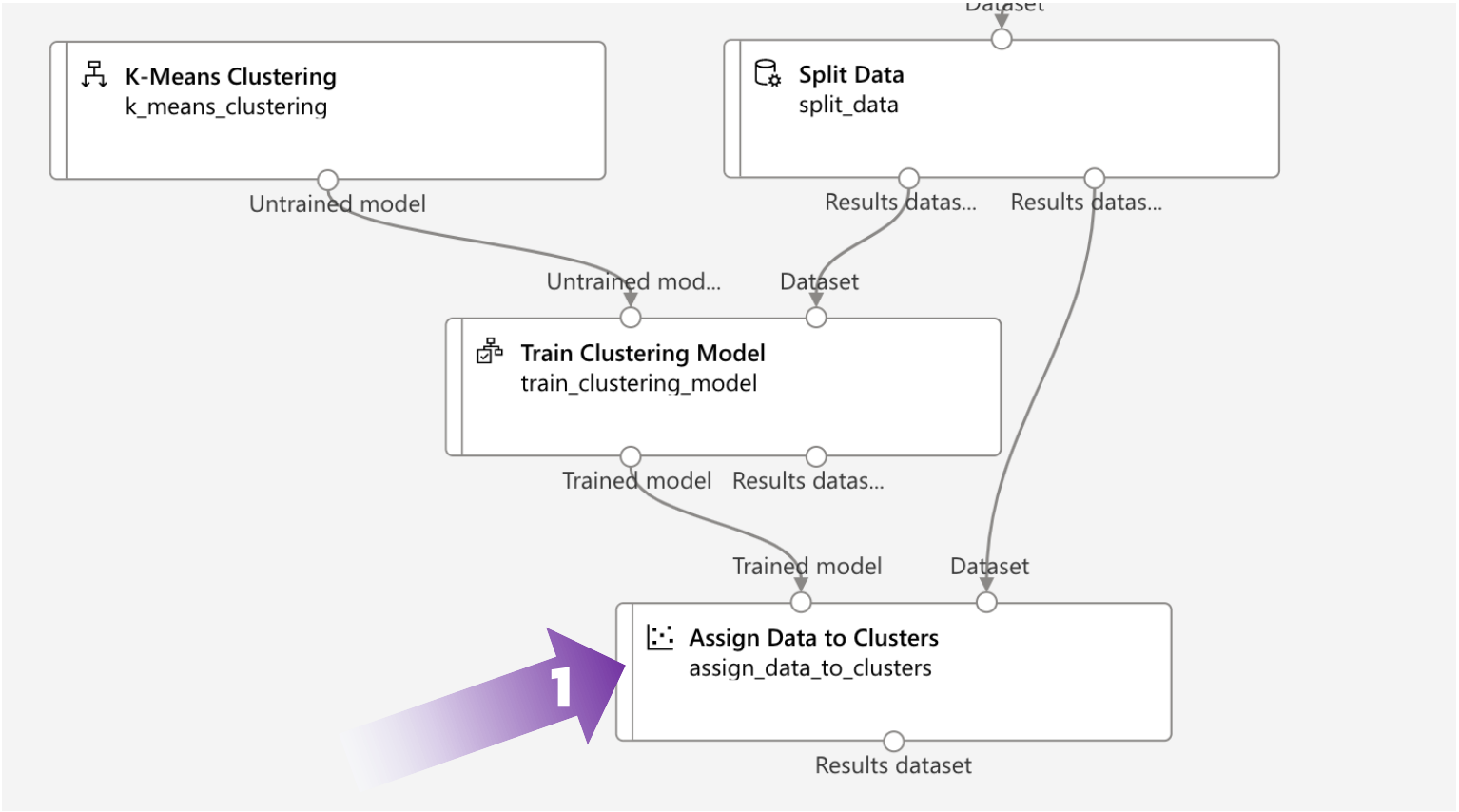
Using a K-Means Clustering Algorithm

- The *K-Means* algorithm groups items into the number of clusters you specify.
- Set the **Number of centroids** parameter to **3** (3 types of penguins).



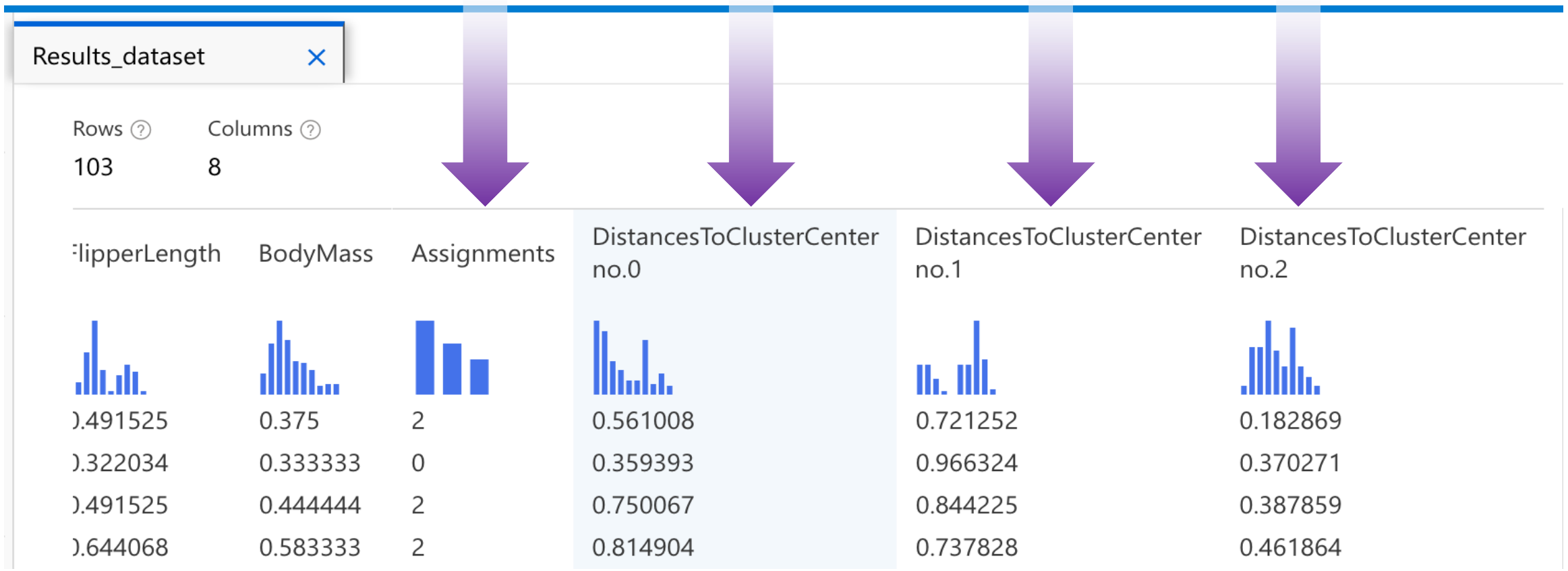
Step 8. Assign Data to Clusters Module

- To test the trained model, use the **Assign Data to Clusters** module
- **Run**
  - using the existing experiment.



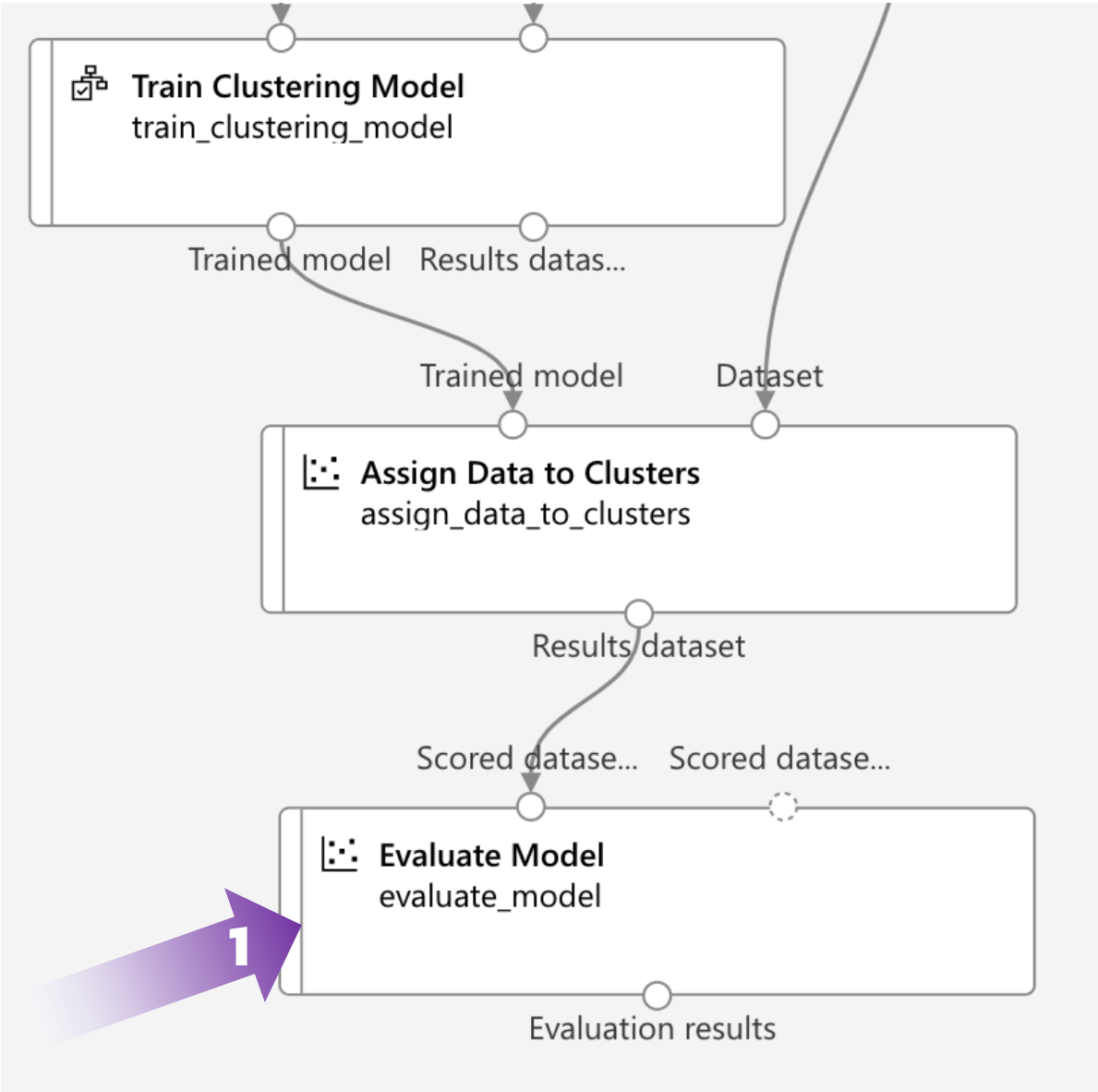
See Result

- At Job detail, select **Assign Data to Clusters** and view the results.
- See the **Assignments** column, which contains the cluster (0, 1, or 2) to which each row is assigned.
- There are also new columns indicating the distance from the point representing this row to the centers of each of the clusters - the cluster to which the point is closest is the one to which it is assigned.



6. Evaluate the model

- Use the **Evaluate Model** module to evaluate the performance of the model.
- Connect the **Evaluate Model** module.
- **Run**



Metrics for Measure the Separation

They include a row of metrics for each cluster, and a summary row for a combined evaluation. The metrics in each row are:

Metric	Description
Average Distance to Other Center	This indicates how close, on average, each point in the cluster is to the centroids of all other clusters.
Average Distance to Cluster Center	This indicates how close, on average, each point in the cluster is to the centroid of the cluster.
Number of Points	The number of points assigned to the cluster.
Maximal Distance to Cluster Center	The maximum of the distances between each point and the centroid of that point's cluster. If this number is high, the cluster may be widely dispersed. This statistic in combination with the **Average Distance to Cluster Center** helps you determine the cluster's spread.

https://docs.microsoft.com/en-us/learn/modules/create-clustering-model-azure-machine-learning-designer/6-evaluate-model

Metrics for Measure the Separation

Result

Evaluation_results

Evaluate Model, Evaluation_results

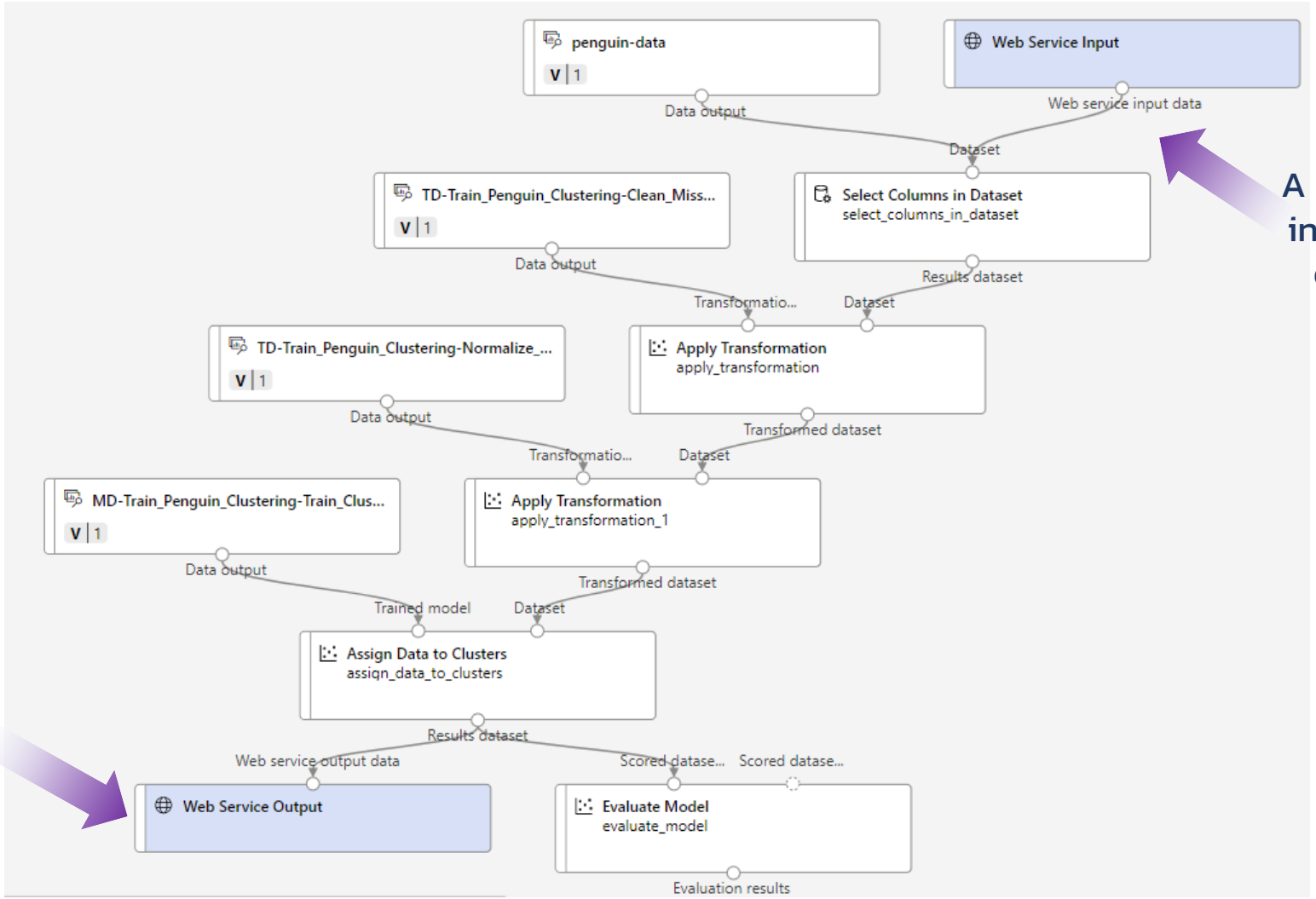
Result Description	Average Distance to Other Center	Average Distance to Cluster Center	Number of Points	Maximal Distance to Cluster Center
Evaluation For Cluster No.0	0.431223	0.193987	47	0.457334
Evaluation For Cluster No.1	0.725174	0.22048	33	0.399237
Evaluation For Cluster No.2	0.469312	0.256118	23	0.461864
Combined Evaluation	0.533907	0.216349	103	0.461864

Part

2

Create an Inference Pipeline to predict clusters from new data

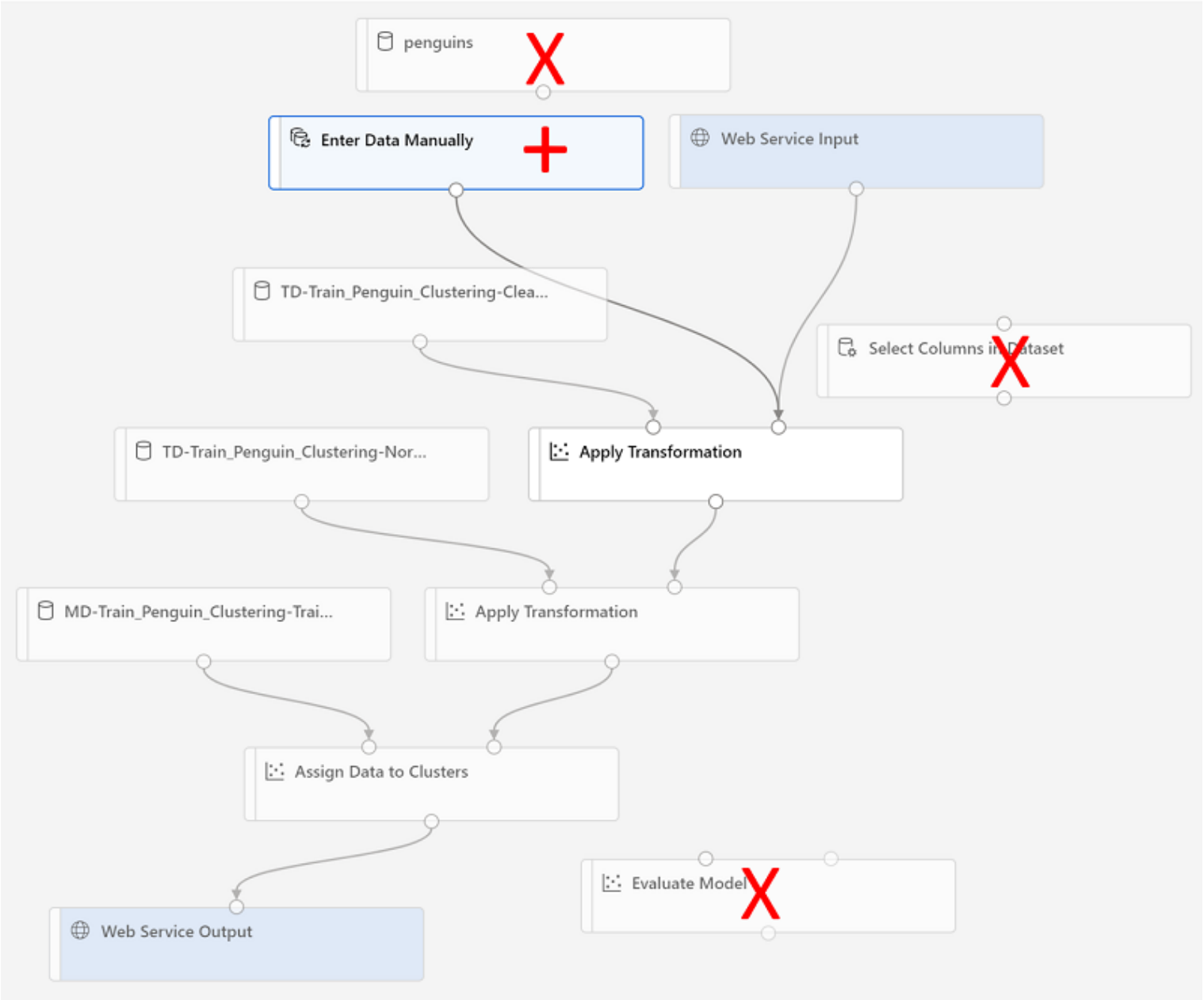
Step 1: สร้าง Inference Pipeline



A web service input for new data to be submitted

A web service output to return results

Step 2: แก้ไข the Inference Pipeline

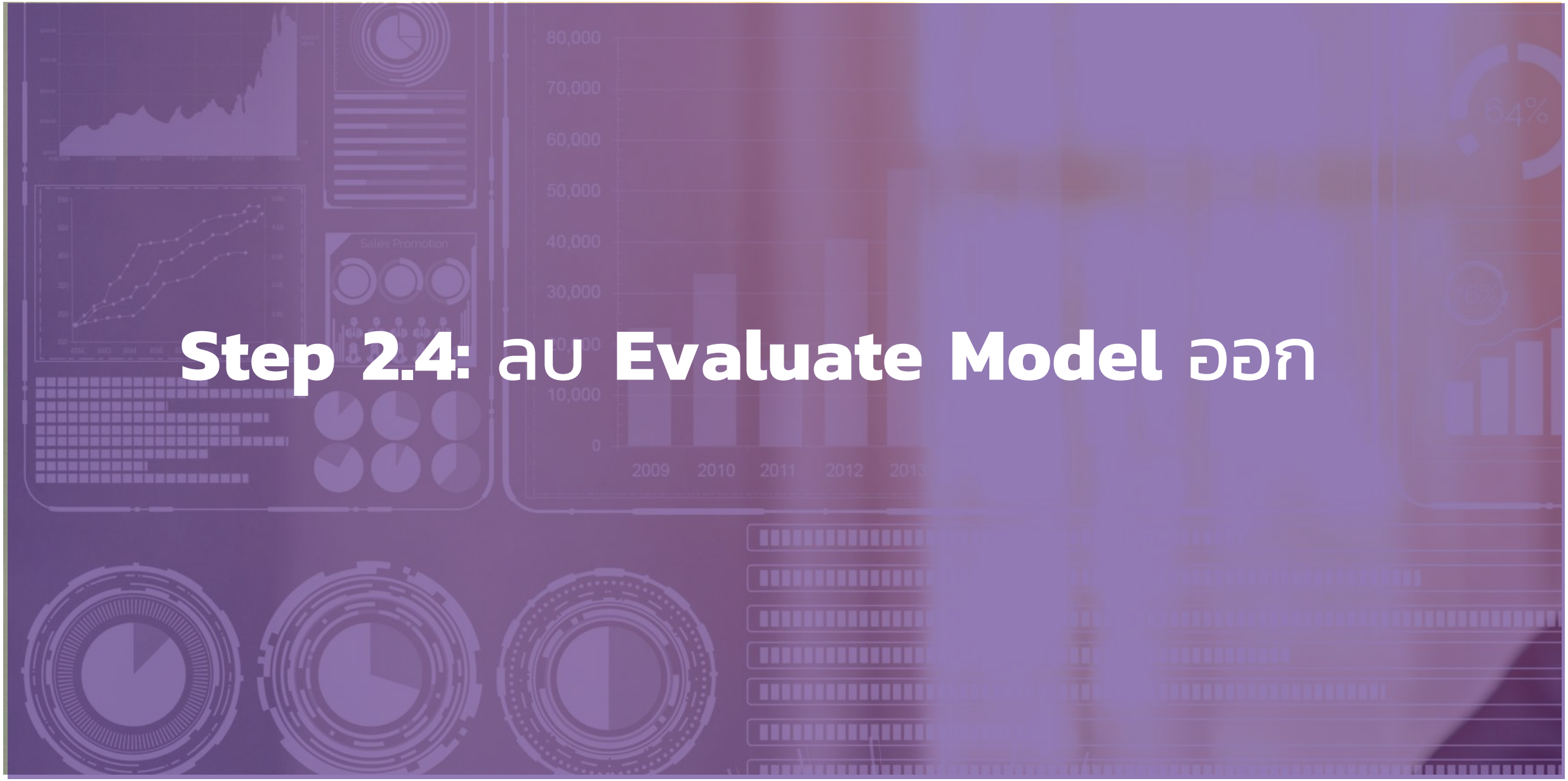


Step 2.1: ลบ Penguin Data

Step 2.2: ลบ Select Columns in Data

Step 2.3: เพิ่ม **Enter Data Manually**
โดย Copy&Paste ข้อมูล CSV Data ของ Penguin 3 ตัว ดังนี้

CulmenLength,CulmenDepth,FlipperLength,BodyMass
39.1,18.7,181,3750 49.1,14.8,220,5150 46.6,17.8,193,3800



Step 2.4: ลบ Evaluate Model ออก

Step 3: Check and Run Your Inference Pipeline

Results_dataset

Rows 3

Columns 8

CulmenLength	CulmenDepth	FlipperLength	BodyMass	Assignments	DistancesToClusterCenter no.0	DistancesToC no.1
0.254545	0.666667	0.152542	0.291667	0	0.161106	0.945574
0.618182	0.202381	0.813559	0.680556	1	0.93688	0.081267
0.527273	0.559524	0.355932	0.305556	2	0.323781	0.683342

Part

3

Deploy the Inference Pipeline as a Service for apps to use

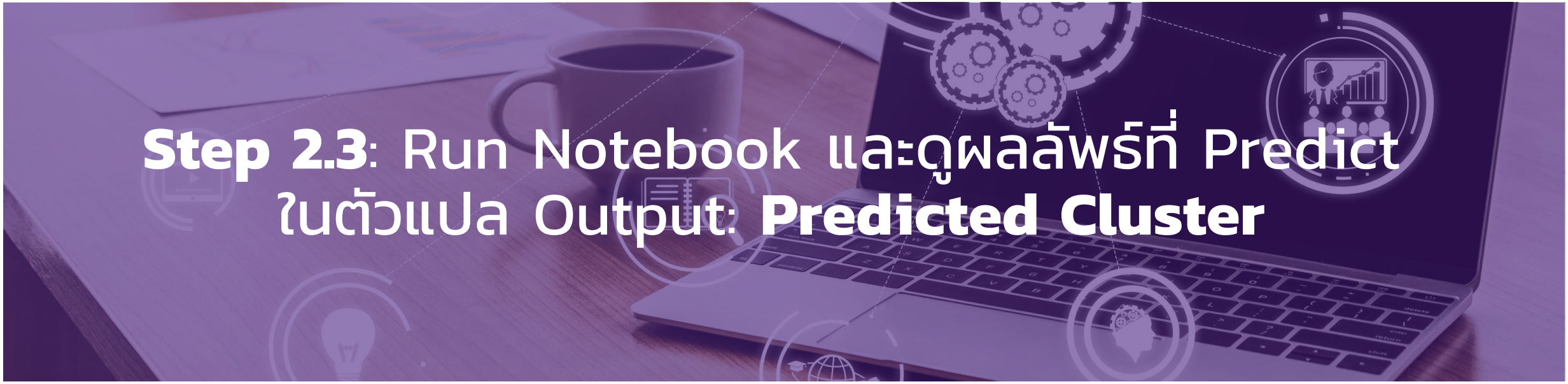
Step 1: Deploy Real-time endpoint สำหรับ Predictive Service (ใช้เวลา)

Step 2: ทดสอบ Service ที่สร้างขึ้น ผ่าน Tab Test และ Notebook

Step 2.1: Code สำหรับทดสอบ Notebook

```
endpoint = 'YOUR_ENDPOINT' #Replace with your endpoint
key = 'YOUR_KEY' #Replace with your key

import urllib.request
import json
import os

data = {
   "Inputs": {
      "WebServiceInput0":
      [
         {
               'CulmenLength': 49.1,
               'CulmenDepth': 4.8,
               'FlipperLength': 1220,
               'BodyMass': 5150,
         },
      ],
   },
   "GlobalParameters":  {
   }
}

body = str.encode(json.dumps(data))


headers = {'Content-Type':'application/json', 'Authorization':('Bearer '+ key)}

req = urllib.request.Request(endpoint, body, headers)

try:
   response = urllib.request.urlopen(req)
   result = response.read()
   json_result = json.loads(result)
   output = json_result["Results"]["WebServiceOutput0"][0]
   print('Cluster: {}'.format(output["Assignments"]))

except urllib.error.HTTPError as error:
   print("The request failed with status code: " + str(error.code))

   # Print the headers to help debug
   print(error.info())
   print(json.loads(error.read().decode("utf8", 'ignore')))
```

Step 2.2: Copy REST Endpoint และ Primary Key ของ Service ที่ Deploy สำเร็จ ใส่ใน Notebook



Step 2.3: Run Notebook และดูผลลัพธ์ที่ Predict ในตัวแปล Output: **Predicted Cluster**


Clean up หลังจากทดลองเสร็จ โดยการ
Delete Endpoint ที่สร้างขึ้น และ Stop
Compute Instance


End of Lab4

Thank you