CHAPTER 2

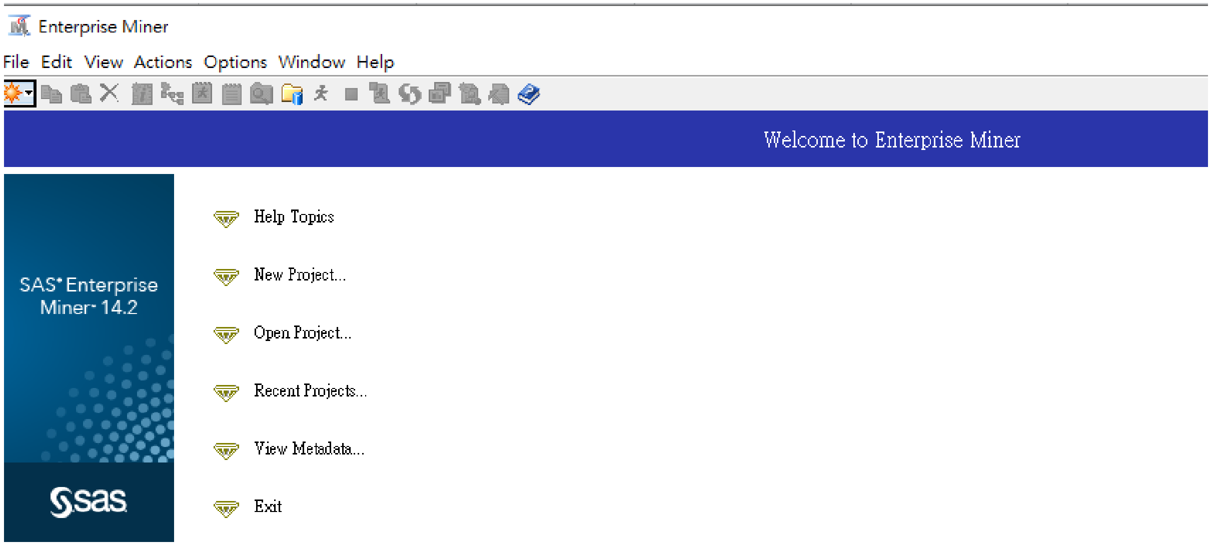
Unsupervised Learning with SAS EM

Content

- Introduction to Unsupervised Learning
- K-means clustering
- Probabilistic clustering via EM algorithm
- Hierarchical clustering
- Unsupervised Learning by Python
- **Unsupervised Learning with SAS EM**
- Number of clusters by Python
- Density-based Spatial Clustering of Applications with Noise (DBSCAN)

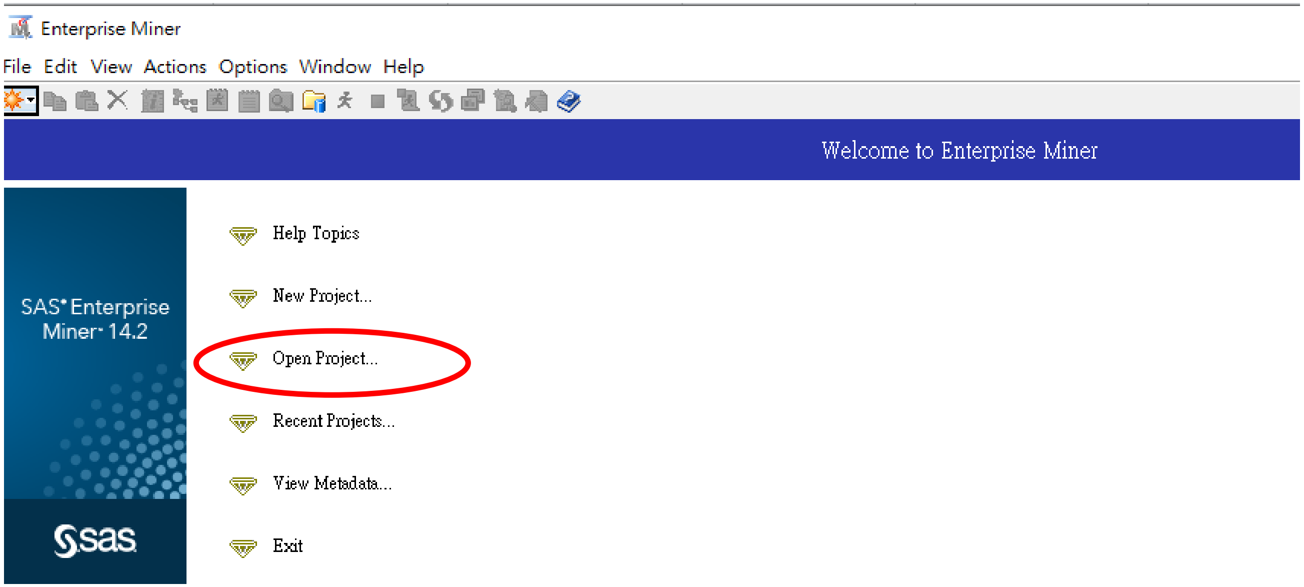
SAS Enterprise Miner (SAS EM)

- SAS EM adopts a user-friendly interface and allow user to adopt the drag-and-drop approach to perform unsupervised d learning.
- To start SAS EM: Windows -> SAS -> SAS Enterprise Miner 14.2
- The files are saved in “Unsupervised Learning with SAS EM”



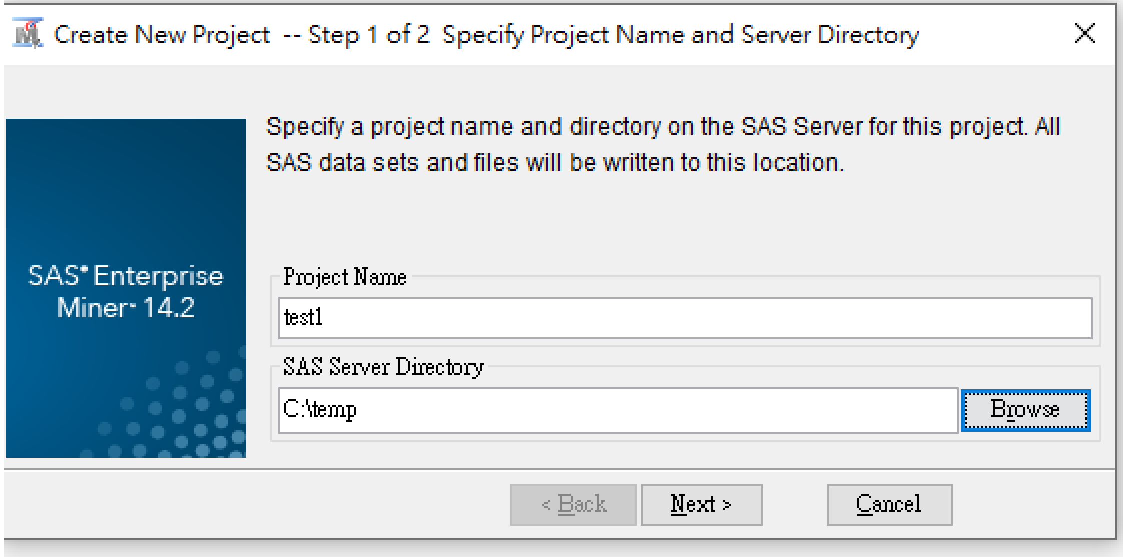
Steps to Clustering

- Step 1: Create a new project



Steps to Clustering

- Step 2: Setup project name and directory
  - ○ Project name: test1
  - ○ Directory: C:\temp
  - ○ Then, click ‘Next’ -> ‘Finish’



Create New Project -- Step 1 of 2 Specify Project Name and Server Directory

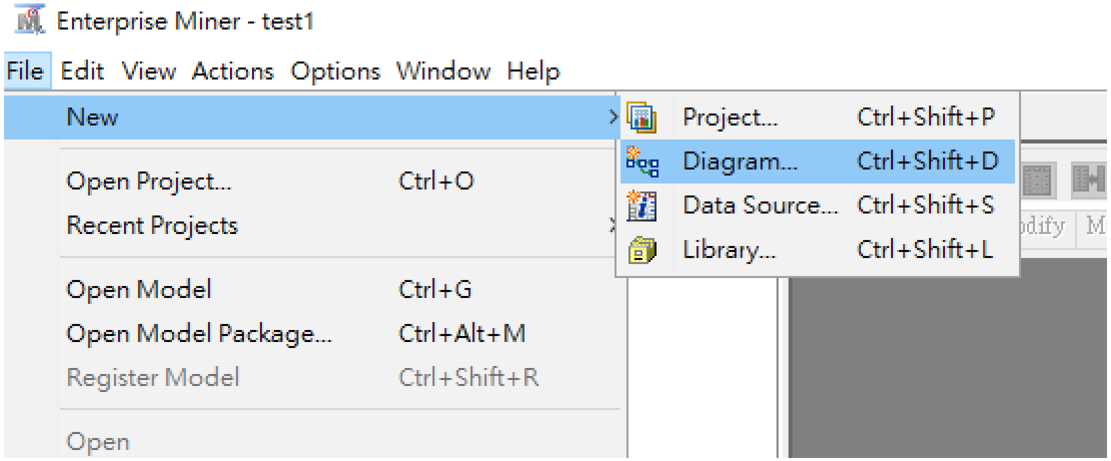
Specify a project name and directory on the SAS Server for this project. All SAS data sets and files will be written to this location.

Project Name: test1

SAS Server Directory: C:\temp

Steps to Clustering

- Step 3: Create a new diagram
- File->New->Diagram



- Step 4: diagram name: clustering

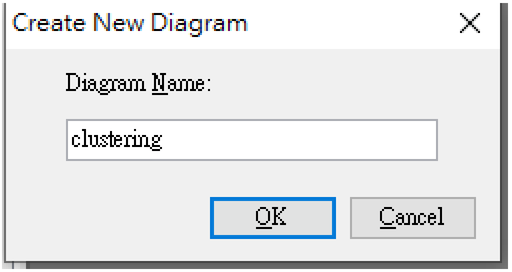
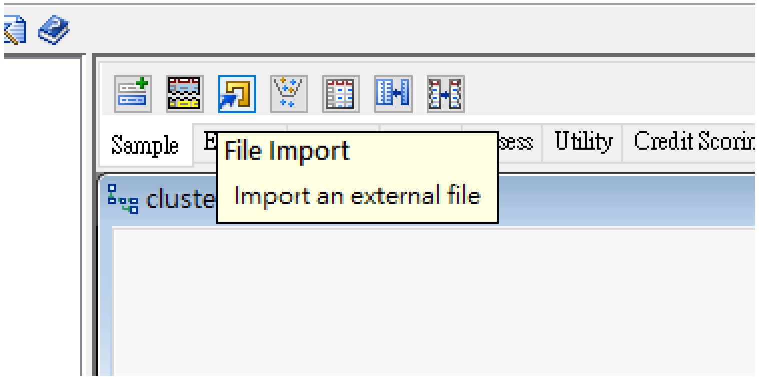
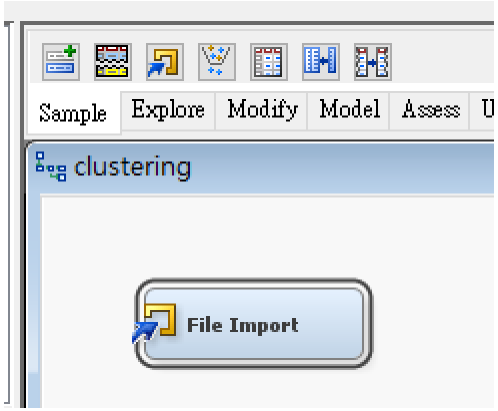


Diagram Name: clustering

Steps to Clustering

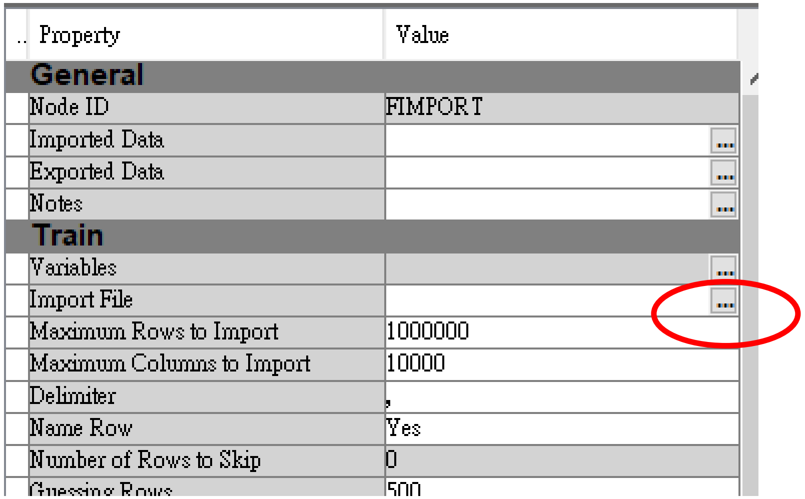
- Step 5:Click onto the icon and drag to the diagram
- Icon: Sample-> File Import





Steps to Clustering

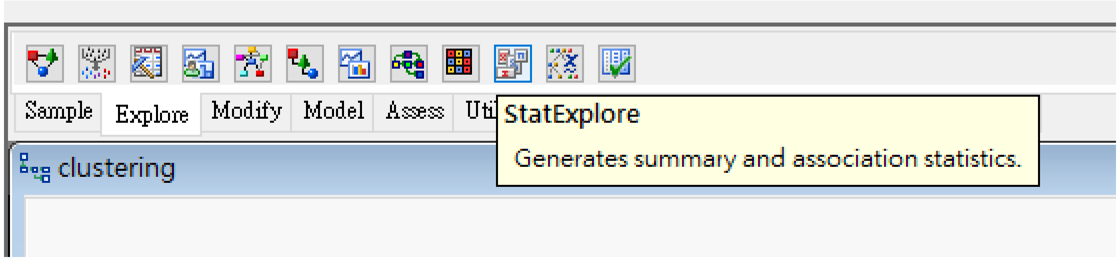
- Step 6: Import file
  - Click “File Import” node
  - On the left panel, select train-> import file
  - Select the iris dataset from your computer



Property	Value
General	
Node ID	FIMPORT
Imported Data	...
Exported Data	...
Notes	...
Train	
Variables	...
Import File	...
Maximum Rows to Import	1000000
Maximum Columns to Import	10000
Delimiter	,
Name Row	Yes
Number of Rows to Skip	0
Guessing Rows	500

Steps to Clustering

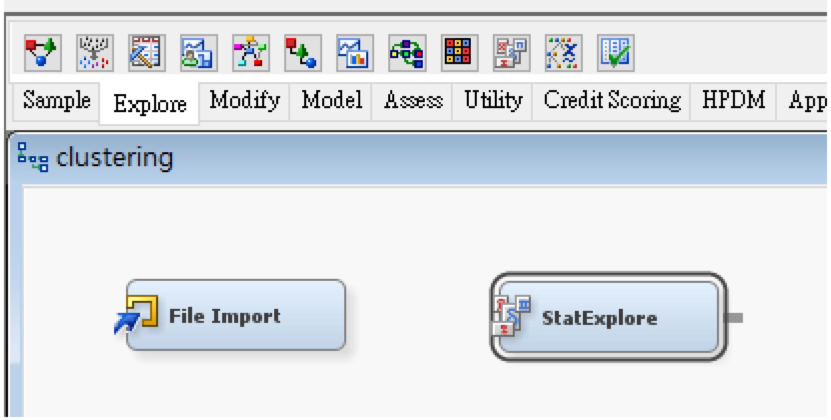
- Step 7: Import descriptive statistics tool
- Icon: Explore -> StatExplore



StatExplore

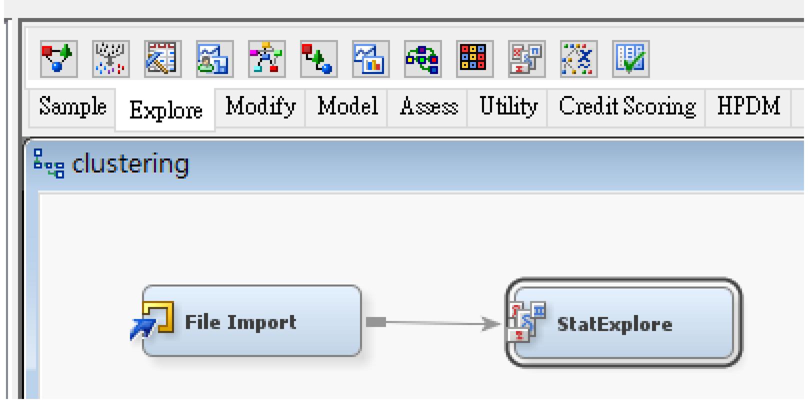
Generates summary and association statistics.

- Drag and drop to the diagram

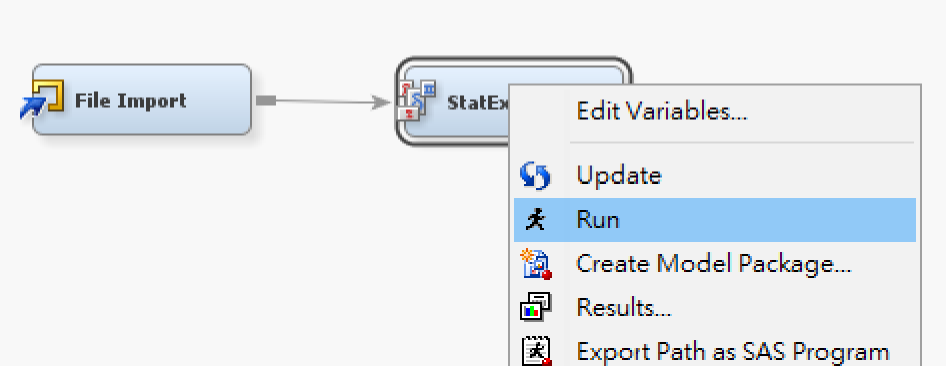


Steps to Clustering

- Step 8: Connect the two nodes

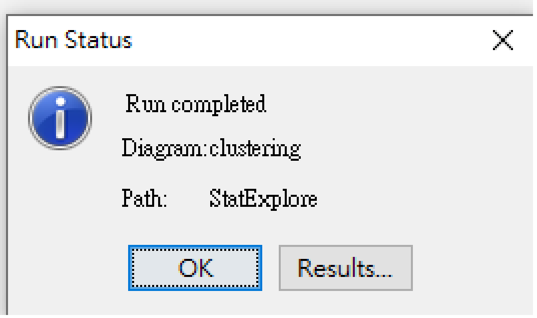


- Step 9: Right-click onto StatExplore and select Run



Steps to Clustering

- Step 10: Click Results (or on the menu bar: Actions->View Results)



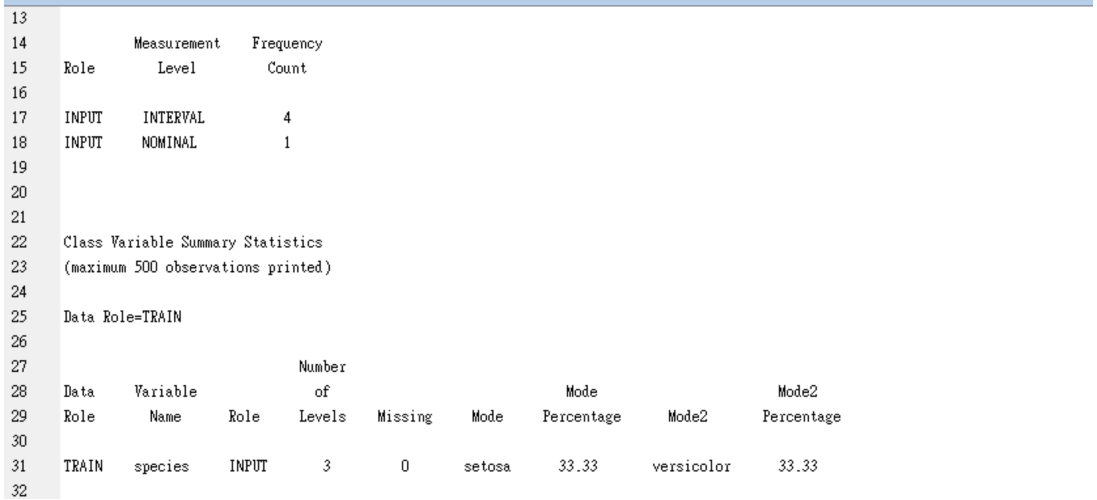
Run Status

Run completed

Diagram:clustering

Path: StatExplore

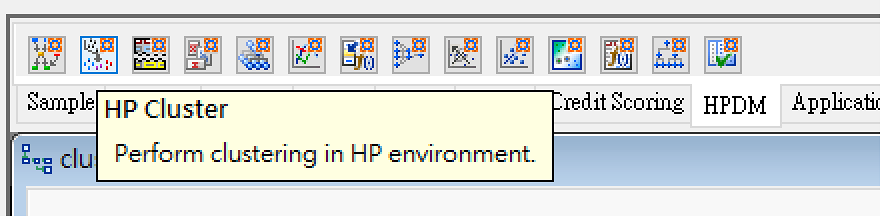
- Basics statistics of the data are shown.



```

         Measurement    Frequency
Role        Level         Count

INPUT     INTERVAL          4
INPUT     NOMINAL           1



Class Variable Summary Statistics
(maximum 500 observations printed)

Data Role=TRAIN

                                Number
Data     Variable                 of                             Mode                    Mode2
Role       Name        Role     Levels    Missing    Mode     Percentage     Mode2      Percentage

TRAIN    species       INPUT       3         0      setosa      33.33      versicolor     33.33

```

Steps to Clustering

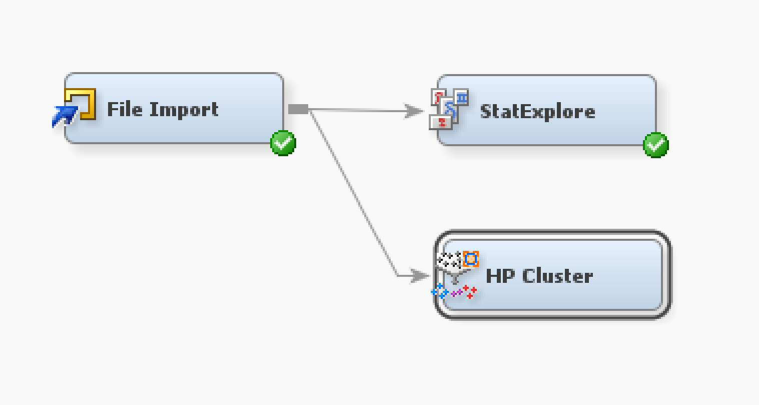
- Step 11: Add Clustering node
  - Icon: HPDM->HP Cluster



HP Cluster

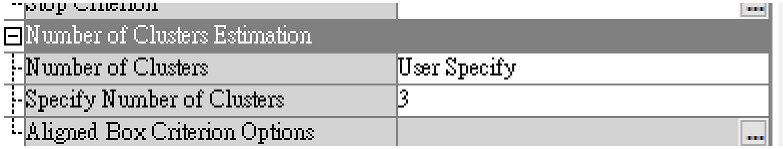
Perform clustering in HP environment.

  - Drag the icon and drop to the diagram
  - Connect HP Cluster with File Import



Steps to Clustering

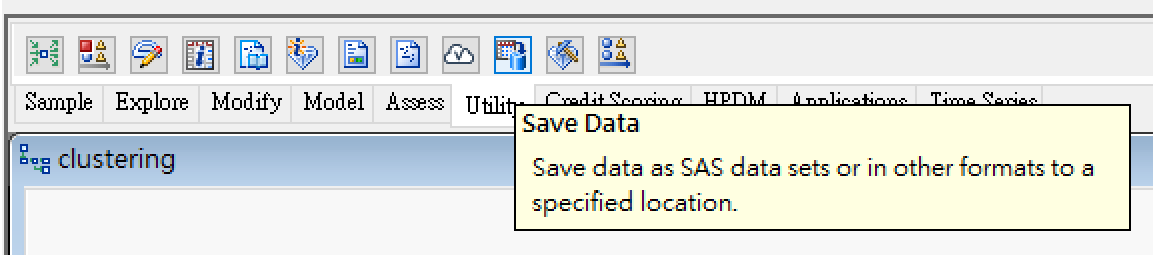
- Step 12: Specify the number of clusters.
  - Click “HP Cluster” node.
  - On the left panel, “Number of Clusters Estimation”>”Number of Clusters”>”User Specify”
  - Specify Number of Clusters: 3



Stop Criterion	...
⊟ Number of Clusters Estimation	
Number of Clusters	User Specify
Specify Number of Clusters	3
Aligned Box Criterion Options	...

Steps to Clustering

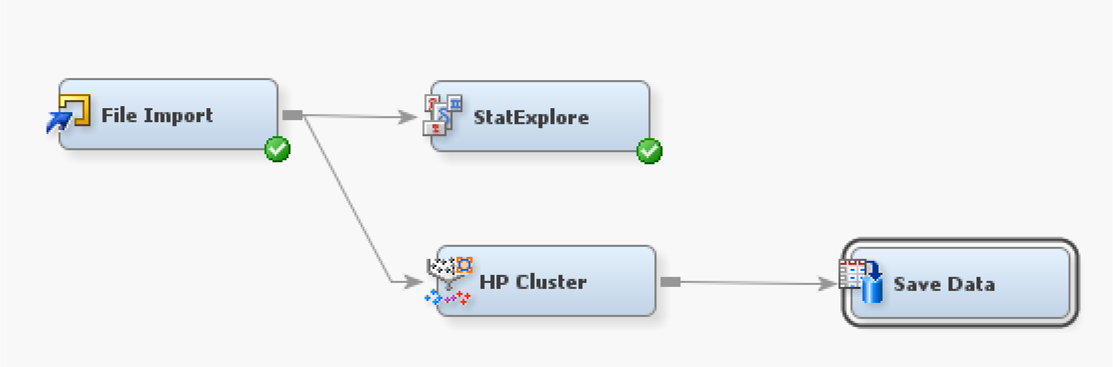
- Step 13: Add saved data node
- Icon: Utility -> Save Data



Save Data

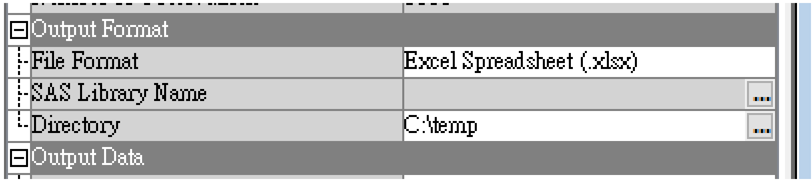
Save data as SAS data sets or in other formats to a specified location.

- Drag and drop to the diagram
- Connect it to HP Cluster



Steps to Clustering

- Step 14: Setup the output format



⊟ Output Format	
File Format	Excel Spreadsheet (.xlsx)
SAS Library Name	...
Directory	C:\temp ...
⊟ Output Data	

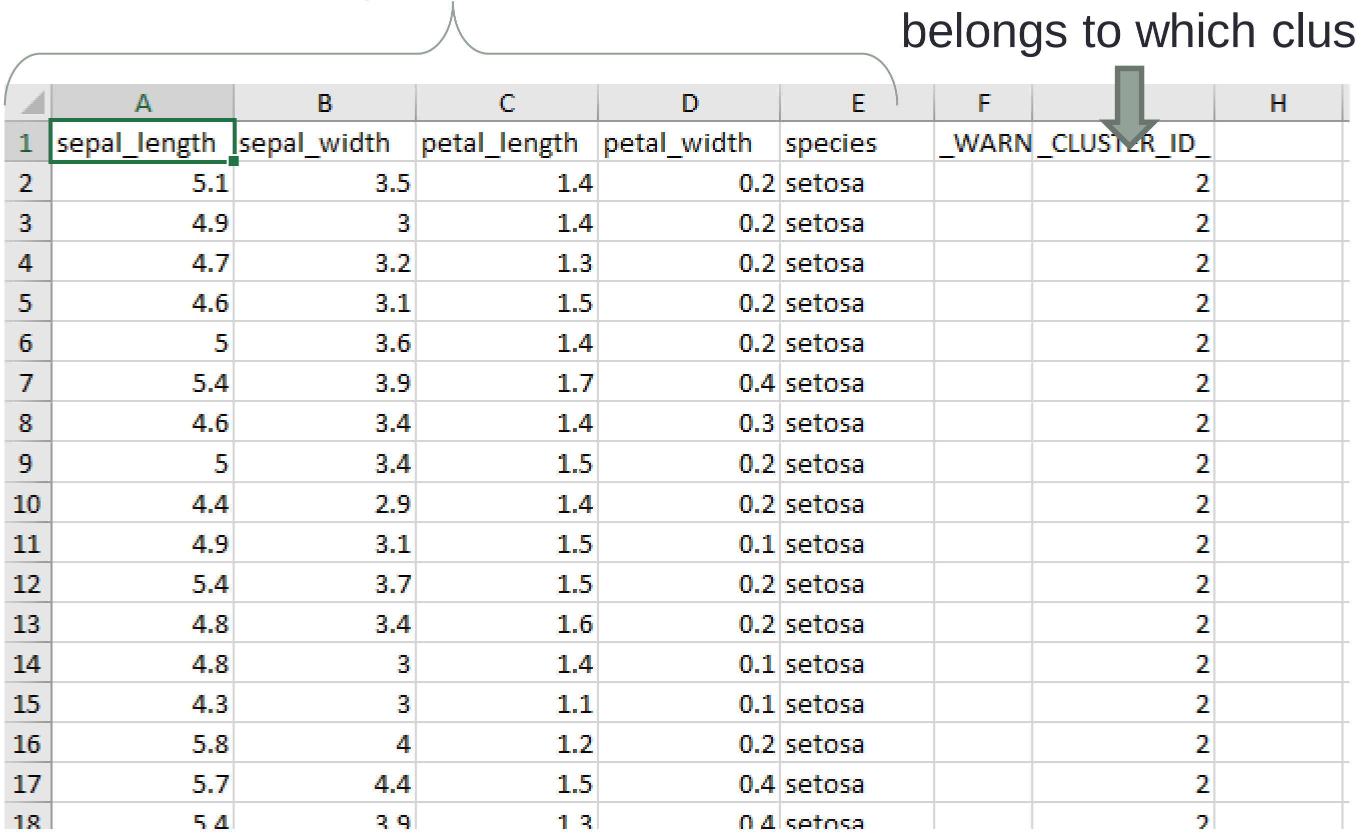
  - Click “Save Data” node
  - On the left panel, select “output format”
  - Set File Format: Excel
  - Directory: your path (the result will be stored in this directory)

Steps to Clustering

- Step 15: The output file is em_save_TRAIN.xlsx

Input data

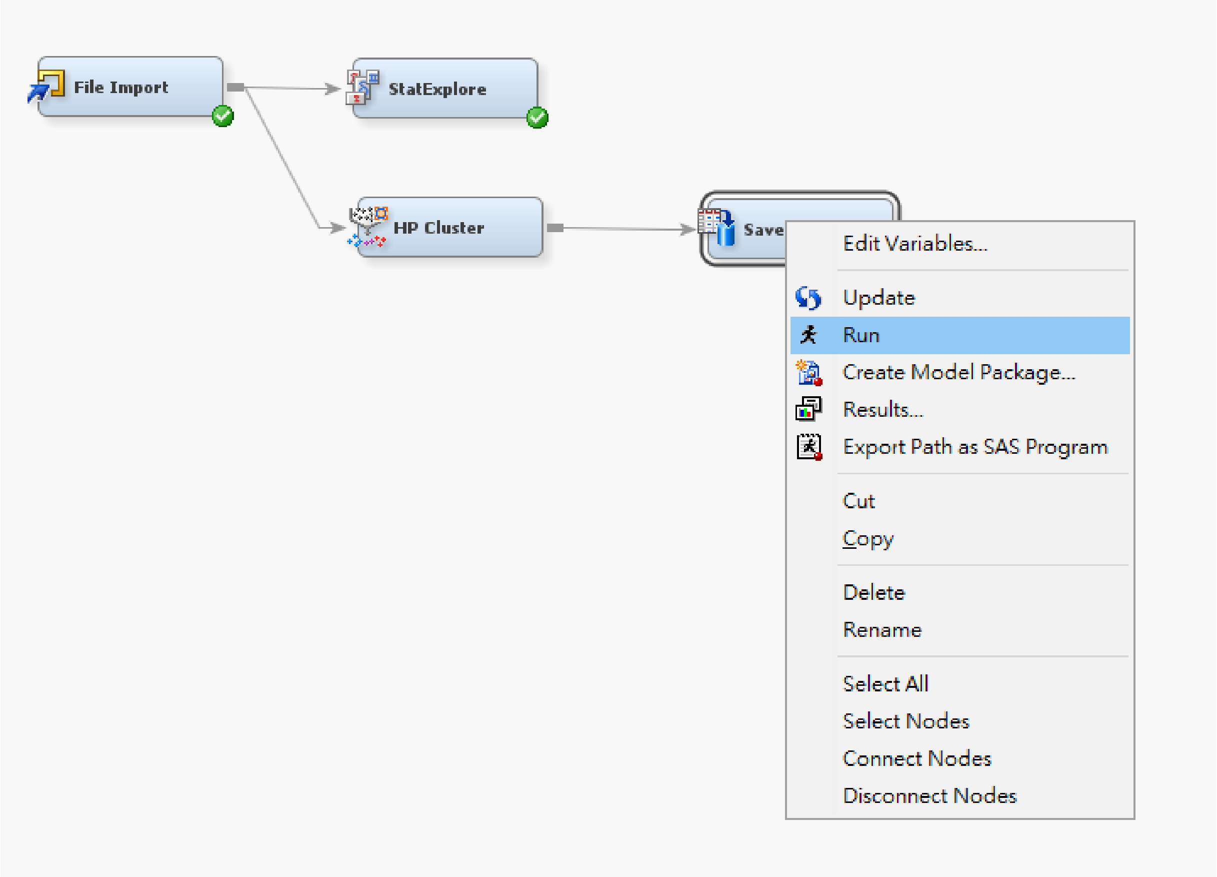
The observation belongs to which cluster



	A	B	C	D	E	F		H
1	sepal_length	sepal_width	petal_length	petal_width	species	_WARN_	_CLUSTER_ID_	
2	5.1	3.5	1.4	0.2	setosa		2	
3	4.9	3	1.4	0.2	setosa		2	
4	4.7	3.2	1.3	0.2	setosa		2	
5	4.6	3.1	1.5	0.2	setosa		2	
6	5	3.6	1.4	0.2	setosa		2	
7	5.4	3.9	1.7	0.4	setosa		2	
8	4.6	3.4	1.4	0.3	setosa		2	
9	5	3.4	1.5	0.2	setosa		2	
10	4.4	2.9	1.4	0.2	setosa		2	
11	4.9	3.1	1.5	0.1	setosa		2	
12	5.4	3.7	1.5	0.2	setosa		2	
13	4.8	3.4	1.6	0.2	setosa		2	
14	4.8	3	1.4	0.1	setosa		2	
15	4.3	3	1.1	0.1	setosa		2	
16	5.8	4	1.2	0.2	setosa		2	
17	5.7	4.4	1.5	0.4	setosa		2	
18	5.4	3.9	1.3	0.4	setosa		2	

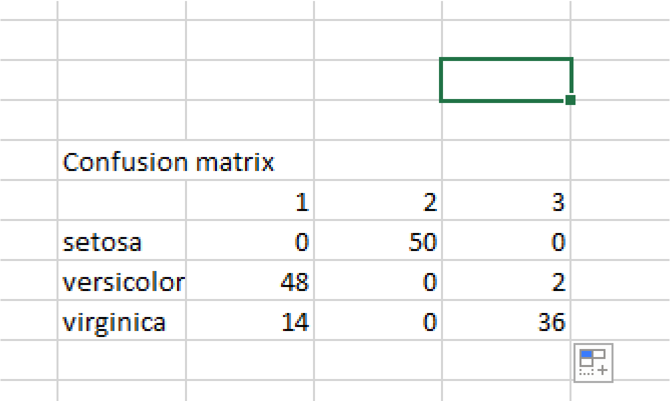
Steps to Clustering

- Right-click onto the Save Data node and click Run



Steps to Clustering

- Step 16: Step Construction of confusion matrix
- You may use the excel function (countifs) to construct the confusion matrix.

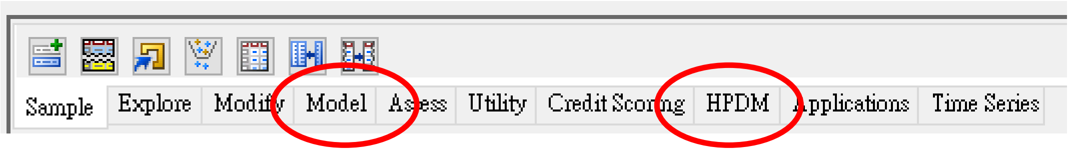


Confusion matrix

	1	2	3
setosa	0	50	0
versicolor	48	0	2
virginica	14	0	36

Remarks

- Other clustering algorithms can be found in either “Model” or “HDPM’.


Sample
Explore
Modify
Model
Assess
Utility
Credit Scoring
HPDM
Applications
Time Series