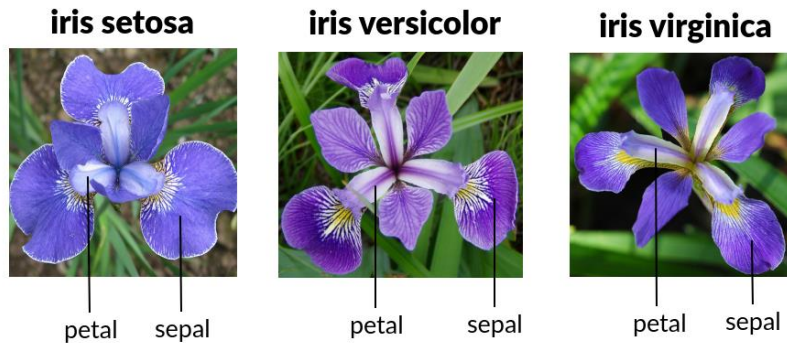


วิเคราะห์ฐานข้อมูล iris ด้วยโปรแกรม Weka

Iris dataset เป็นฐานข้อมูลเกี่ยวกับดอกไม้ iris กว่า 150 ดอก โดยรายละเอียดจะมี ความยาวกลีบเลี้ยง (Sepal length), ความกว้างกลีบเลี้ยง (Sepal width), ความยาวกลีบดอก (petal length) และ ความกว้างกลีบดอก (Petal width) หน่วยวัดเป็นหน่วยเซนติเมตร (cm) และ คลาสสายพันธุ์ (class) ของดอกไม้ iris แบ่งเป็น 3 กลุ่ม คือ Setosa, Versicolor และ Virginica.



ภาพจาก <https://kongruksiamza.medium.com>

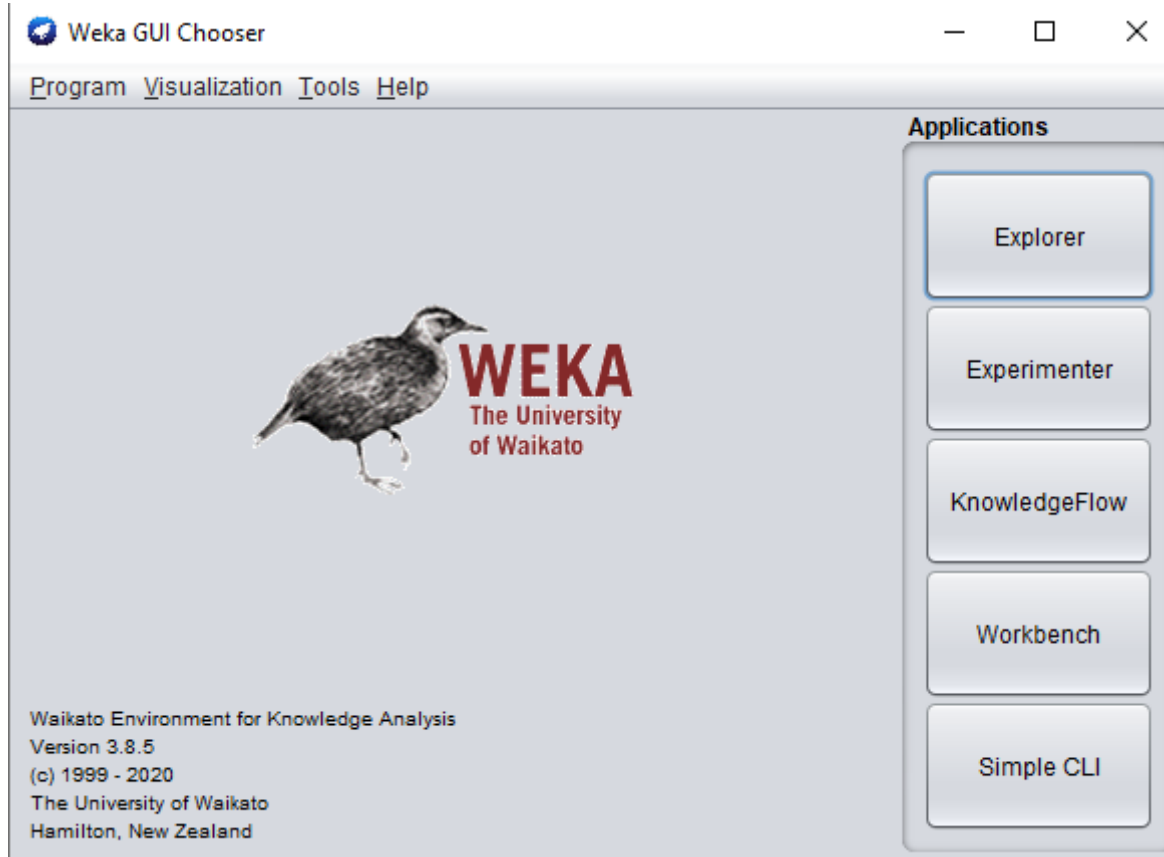
ตัวอย่างชุดข้อมูล

sepal_length	sepal_width	petal_length	petal_width	class
5.1	3.5	1.4	0.2	Iris-setosa
4.9	3	1.4	0.2	Iris-setosa
4.7	3.2	1.3	0.2	Iris-setosa
4.6	3.1	1.5	0.2	Iris-setosa
5	3.6	1.4	0.2	Iris-setosa

Attribute/ features = 5 (sepal length, sepal width, petal length, petal width, class)

Class Label to predict = 3 (Iris Setosa, Iris Versicolor, Iris Virginica)

Instances = 150 records (แต่ละ class จะมี 50 records) และไม่มี missing value

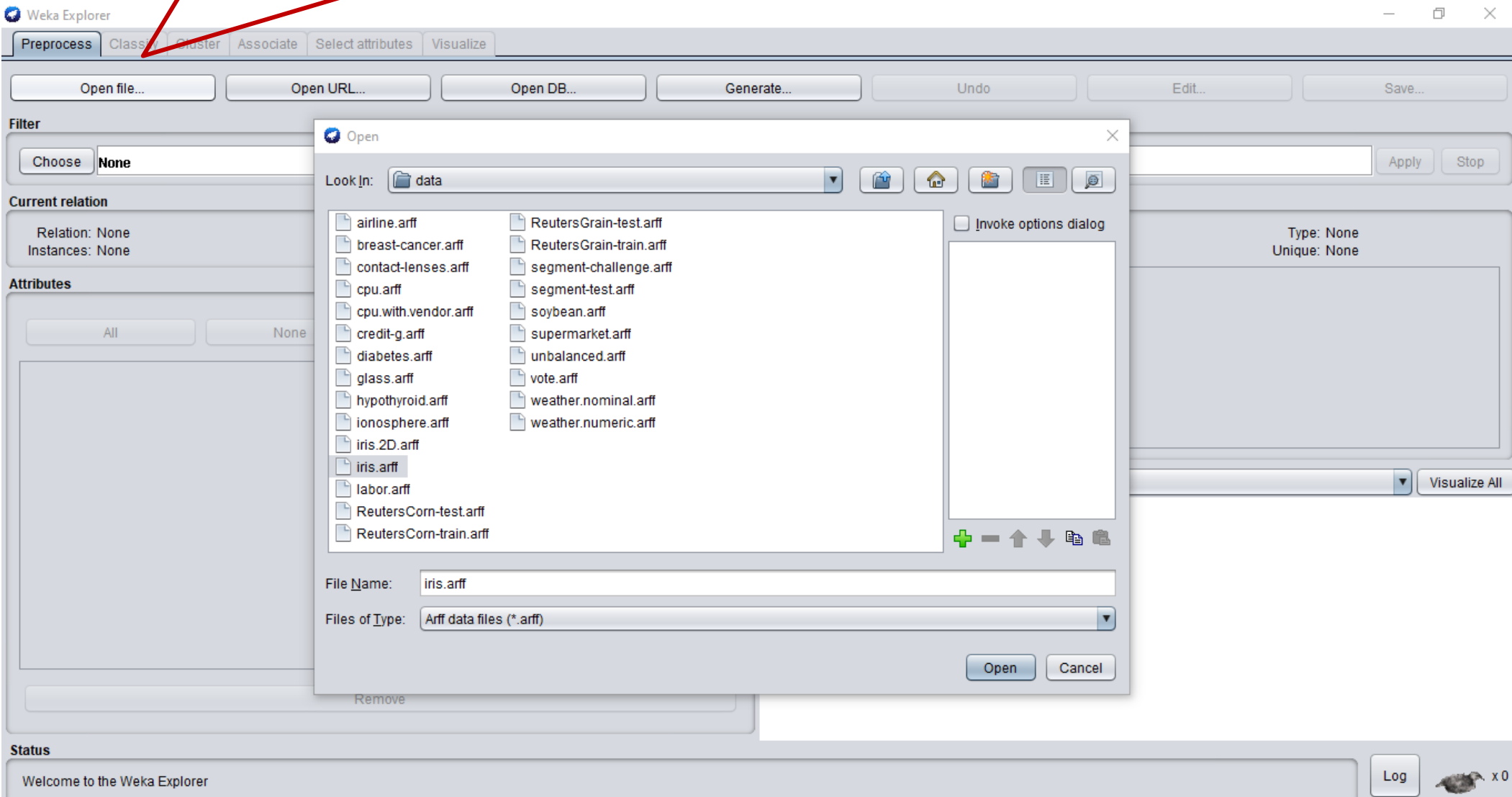


โปรแกรม weka เป็นซอฟต์แวร์ที่ใช้ทำ machine learning, pattern recognition หรือ data mining เป็นต้น เริ่มการทำงาน โดยกดที่ปุ่ม Explorer ในหน้าต่างหลัก

ใช้เปิดไฟล์ที่โปรแกรมรองรับ เช่น *.csv , *.arff เป็นต้น สามารถลองเปิดไฟล์ที่ตัวโปรแกรมแถมมาได้ที่

C:\Program Files\Weka-3-8-5\data

@Github/Zuwannn



หน้าต่างแบบ Preprocess

Weka Explorer

Preprocess Classify Cluster Associate Select attributes Visualize

Open file... Open URL... Open DB... Generate... Undo Edit... Save...

Filter

Choose None Apply Stop

Current relation

Relation: iris Instances: 150 Attributes: 5 Sum of weights: 150

Attributes

All None Invert Pattern

No.	Name
1	<input checked="" type="checkbox"/> sepalength
2	<input type="checkbox"/> sepalwidth
3	<input type="checkbox"/> petallength
4	<input type="checkbox"/> petalwidth
5	<input type="checkbox"/> class

Remove

Status

OK Log

Selected attribute

Name: sepalength Missing: 0 (0%) Distinct: 35 Type: Numeric Unique: 9 (6%)

Statistic	Value
Minimum	4.3
Maximum	7.9
Mean	5.843
StdDev	0.828

Class: class (Nom)

Visualize All

แบบ Selected attribute ใช้แสดงใช้แสดงค่าสถิติของพีเจอร์
ที่เลือกจาก แบบ Attribute

แบบ Attribute ใช้แสดงพีเจอร์ของไฟล์ที่เปิด

แบบ Class ใช้แสดงการกระจายตัวของพีเจอร์ที่เลือกจากแบบ
Attribute โดยแกนนอนแทนค่าของพีเจอร์จากน้อยไปมาก แกนตั้ง
แทนค่าแทนจำนวน สีแทนชั้นข้อมูล (class)

หน้าต่างแถบ Classify โดย Classify เป็นส่วนที่ใช้ในการวิเคราะห์ข้อมูล
ด้วยวิธีการจำแนกข้อมูล (classification) หรือทำนายข้อมูล (prediction)
ซึ่งมีวิธีการต่างๆ ให้เลือกมากมาย

Weka Explorer

Preprocess **Classify** Cluster Associate Select attributes Visualize

Classifier

Choose **IBk** -K 1 -W 0 -A "weka.core.neighboursearch.LinearNNSearch -A "weka.core.EuclideanDistance -R first-last"" กด Choose เพื่อเลือกตัวจำแนกที่ต้องการใช้

Test options

☐ Use training set
☐ Supplied test set Set...
☒ Cross-validation Folds **10**
☐ Percentage split % 66
More options...

(Nom) class

Start Stop

Result list (right-click for options)

00:25:49 - lazy.IBk

แสดงผลการประเมิน
เมื่อกดปุ่ม Start ในแต่
ละครั้ง

Classifier output

=== Stratified cross-validation ===
=== Summary ===

Correctly Classified Instances	143	95.3333 %
Incorrectly Classified Instances	7	4.6667 %
Kappa statistic	0.93	
Mean absolute error	0.0399	
Root mean squared error	0.1747	
Relative absolute error	8.9763 %	
Root relative squared error	37.0695 %	
Total Number of Instances	150	

=== Detailed Accuracy By Class ===

	TP Rate	FP Rate	Precision	Recall	F-Measure	MCC	ROC Area	PRC Area	Class
	1.000	0.000	1.000	1.000	1.000	1.000	1.000	1.000	Iris-setosa
	0.940	0.040	0.922	0.940	0.931	0.896	0.952	0.887	Iris-versicolor
	0.920	0.030	0.939	0.920	0.929	0.895	0.947	0.894	Iris-virginica
Weighted Avg.	0.953	0.023	0.953	0.953	0.953	0.930	0.966	0.927	

=== Confusion Matrix ===

a	b	c	<-- classified as
50	0	0	a = Iris-setosa
0	47	3	b = Iris-versicolor
0	4	46	c = Iris-virginica

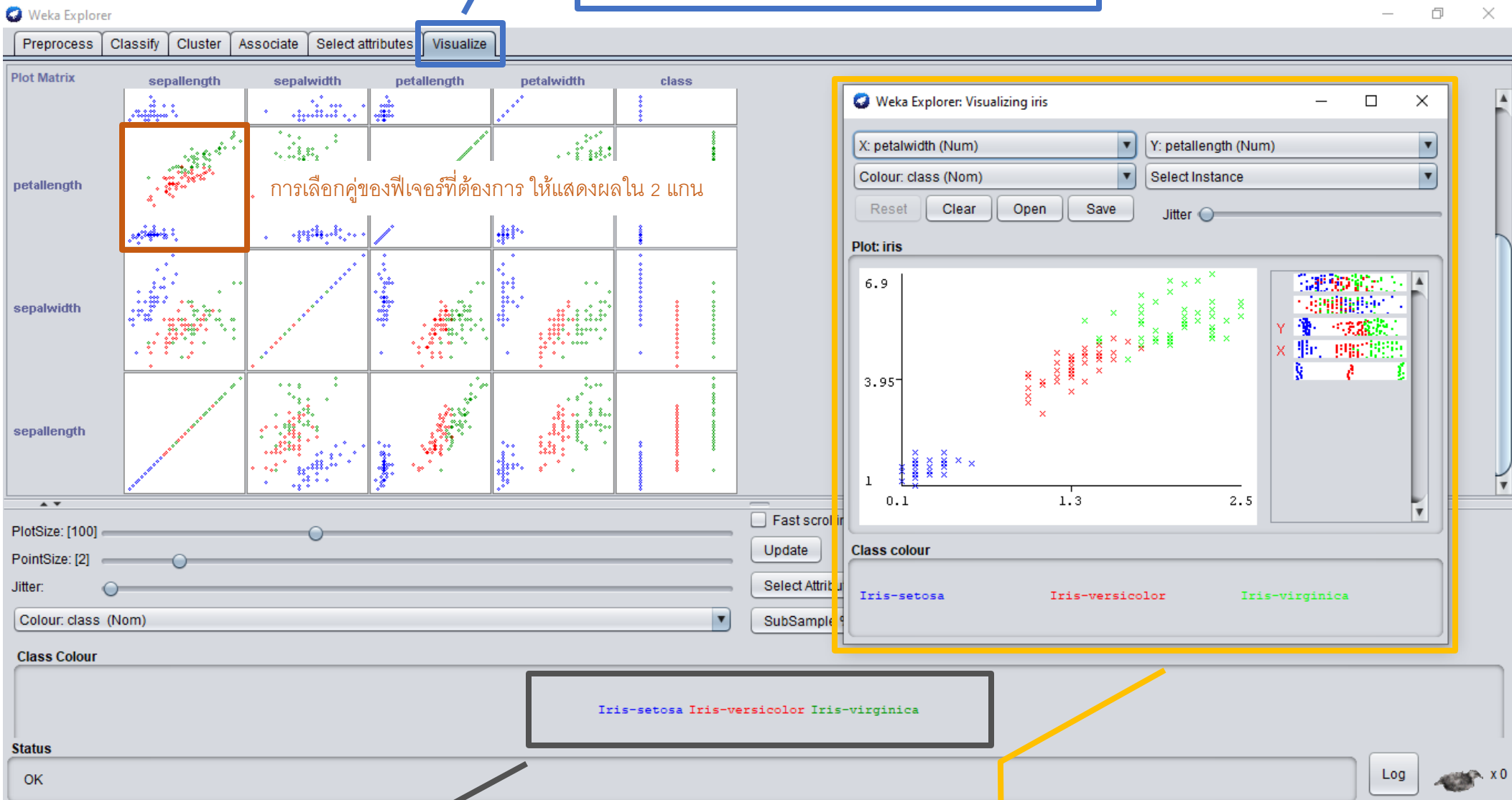
Status

OK Log x 0

แถบ Test options มีไว้สำหรับเลือกวิธีประเมินตัวจำแนก โดย Use Training set คือ ใช้เซตข้อมูลทั้งหมดเป็นทั้ง train set และ test set , Supplied test set คือเลือก test set ต่างจาก train set , Cross-validation คือ , Percentage Split คือแบ่ง train/test set ตามเปอร์เซ็นต์ จากนั้นกดปุ่ม Start เพื่อเริ่มการประเมิน

หน้าต่างแบบ Visualize สำหรับการแสดงผล การกระจาย
ของแบบ จากแต่ละชั้นข้อมูล

@Github/Zuwannn



สีสัญลักษณ์ที่ระบุว่า แบบโนสีใดอยู่ในชั้นข้อมูลไหน

หน้าต่าง Visualizing สำหรับการแสดงผล การกระจายของ
แบบ จากแต่ละชั้นข้อมูล