ANOMALY DETECTION

THIS PROJECT

RMUTT
www.rmutt.ac.th ราชมงคลธัญบุรี

Machine Learning (Code : 09152204)
INSTRUCTOR : Assistant Professor Dr.Pichate Kunakornvong

Artificial Intelligence (Code : 09152205)
INSTRUCTOR : Assistant ProfessorPongpon Nilaphruek

FACULTY : Science and Technology
PROGRAM : Big Data Management and Analytics

DATA SCIENCE PROCESS

Obtain Data

การเก็บข้อมูล รับข้อมูล และการดึงข้อมูลจากฐานข้อมูลที่มีอยู่

Scrub Data

การทำความสะอาดข้อมูล เป็นกระบวนการเปลี่ยนข้อมูลดิบให้อยู่ในรูปแบบที่ง่ายต่อการใช้งาน

Explore Data

การสำรวจและค้นหาข้อมูลเพื่อหาความสัมพันธ์ แนวโน้ม หรือสิ่งที่น่าสนใจที่เกี่ยวข้องกัน

Model Data

การจำลองข้อมูลโดยการสร้างโมเดลด้วยการใช้ซอฟต์แวร์และอัลกอริทึม Machine learning

Interpret Results

การแปลงข้อมูลแล้วนำออกมาเล่าเป็นเรื่องราวเพื่อนำเสนอ

TECH STACK

Orange data mining

3.36.2 for Windows
or higher is required.

Anaconda

2023.09
or higher is required.

Python

3.11
or higher is required.

Orange Data Mining

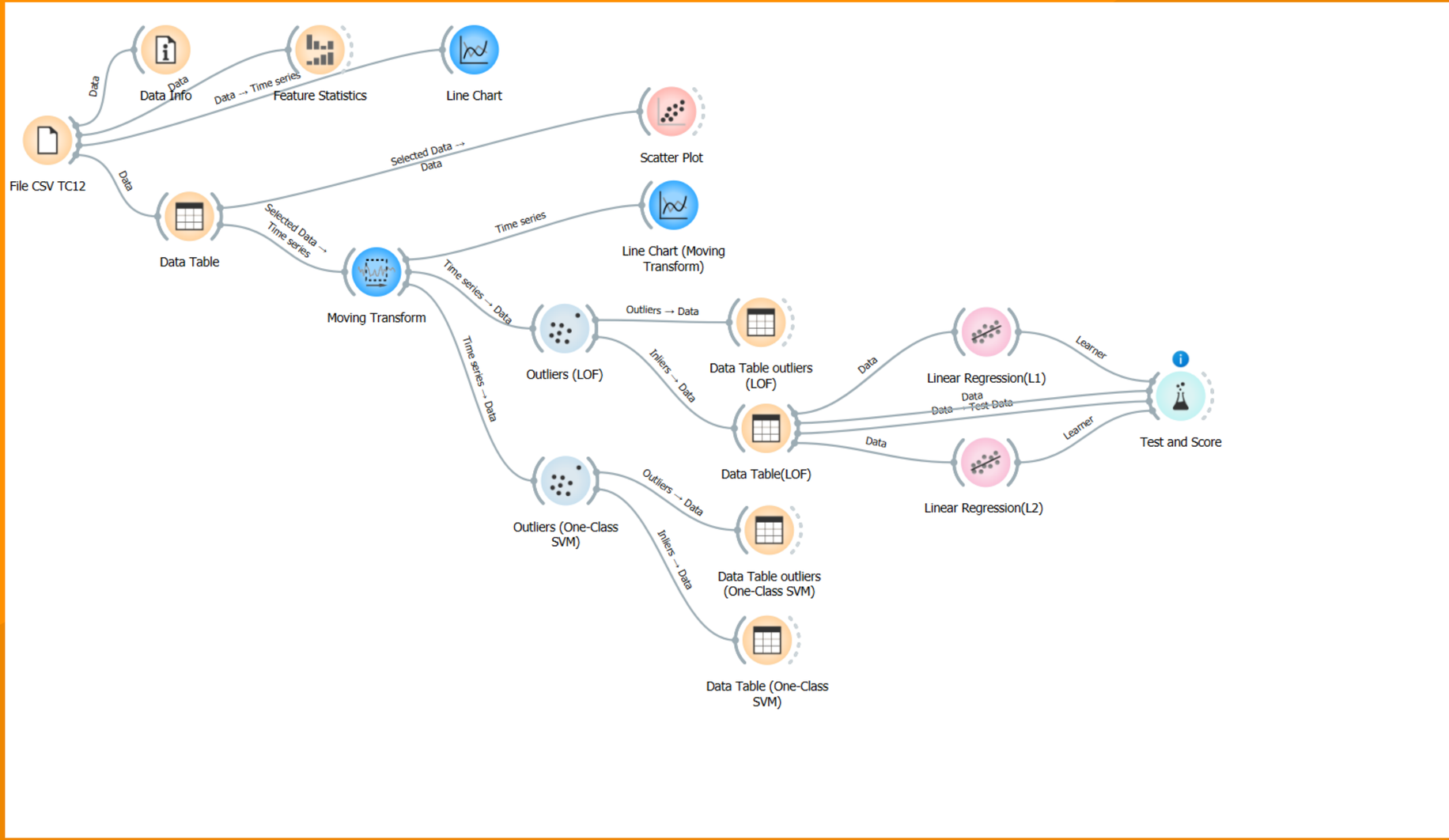
Data Mining Fruitful and Fun

Open source machine learning and data visualization.

RMUTT
www.rmutt.ac.th ราชมงคลธัญบุรี

OVERVIEW

Overview



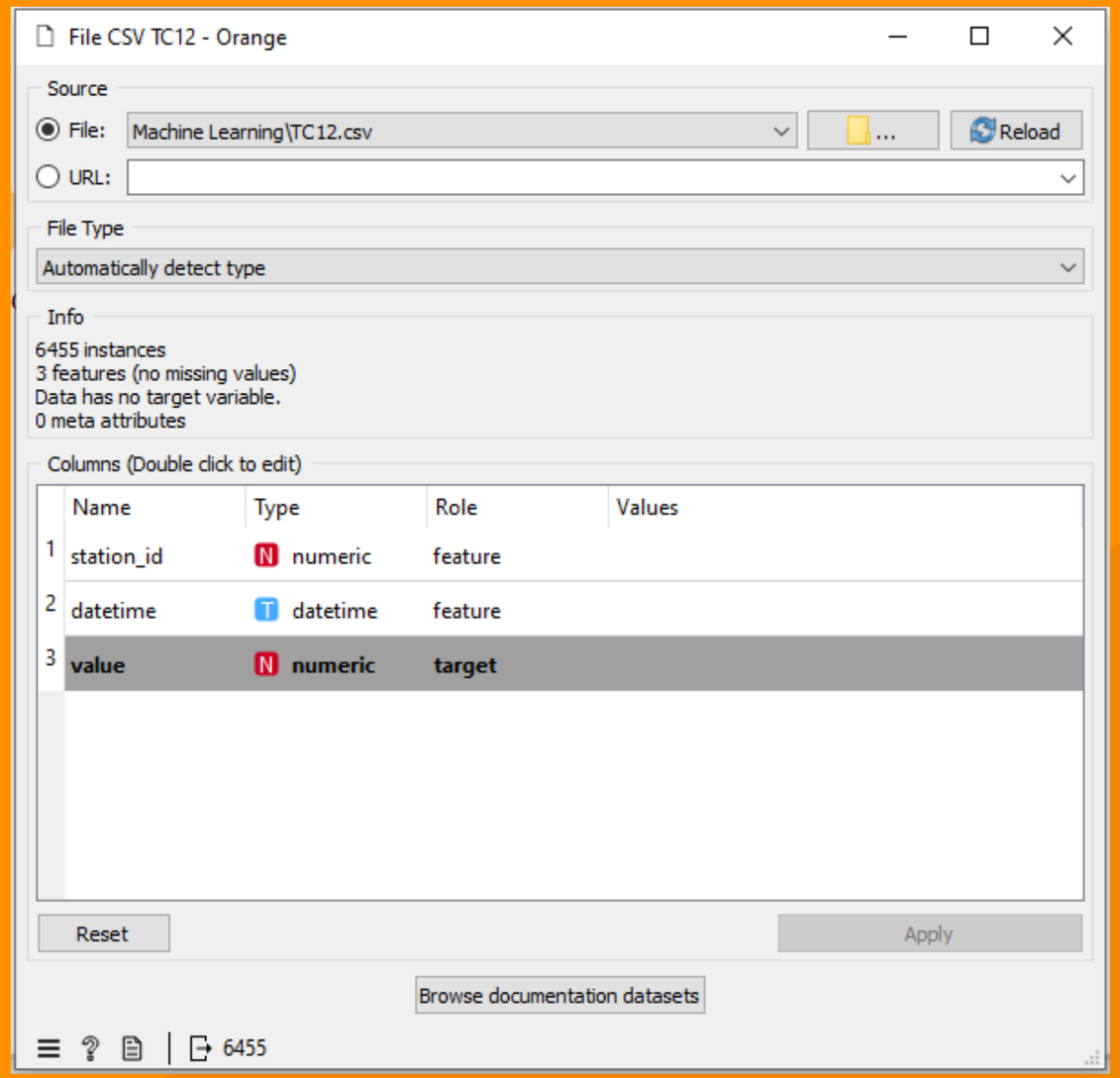
RMUTT
www.rmutt.ac.th ราชมงคลธัญบุรี

Feature Statistics

DEMO

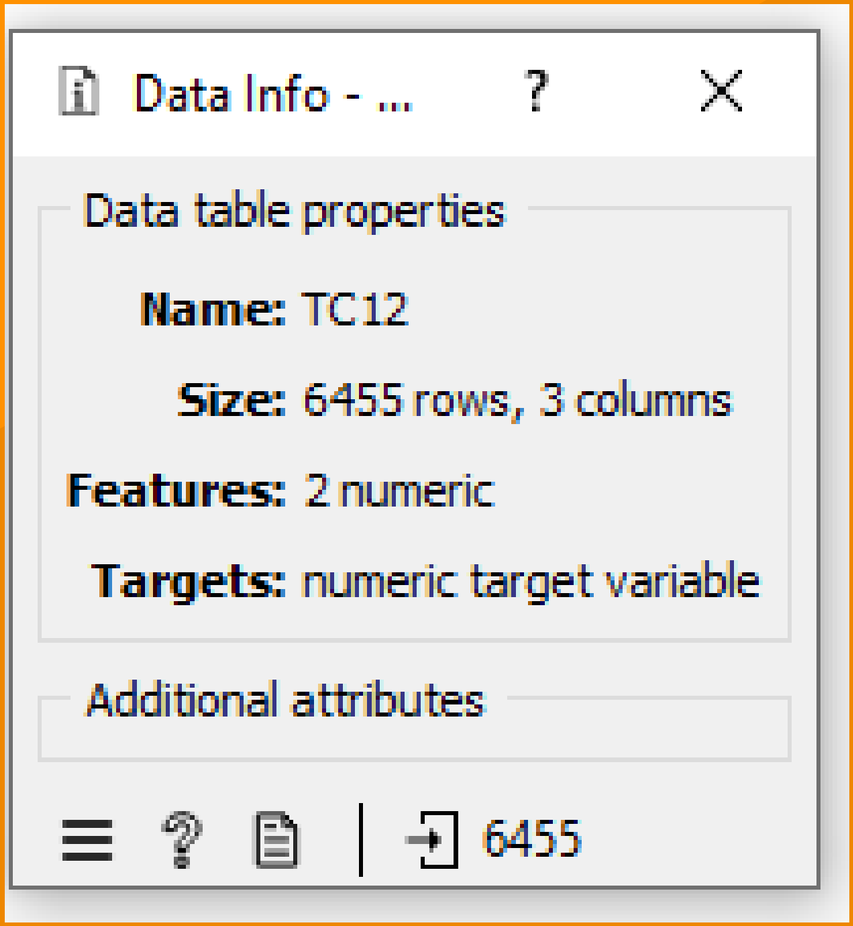
Data Info

RMUTT
www.rmutt.ac.th ราชมงคลธัญบุรี



Import TC12.csv file. and set Role

Information of the data set



Feature Statistics - Orange

	Name	istributio	Mean	Mode	Median	Dispersion	Min.	Max.	Missing
N	station_id		22	22	22	0	22	22	0 (0 %)
N	value		0.2405	-0.70	0.00	4.1127	-0.73	12.31	0 (0 %)
T	datetime		2023-09-16 22:03:52.977334	2023-08-15 22:30:10.008000	2023-09-23 00:45:10.129000	~2 months	2023-08-15 16:45:10.836000	2023-10-15 23:45:10.784000	0 (0 %)

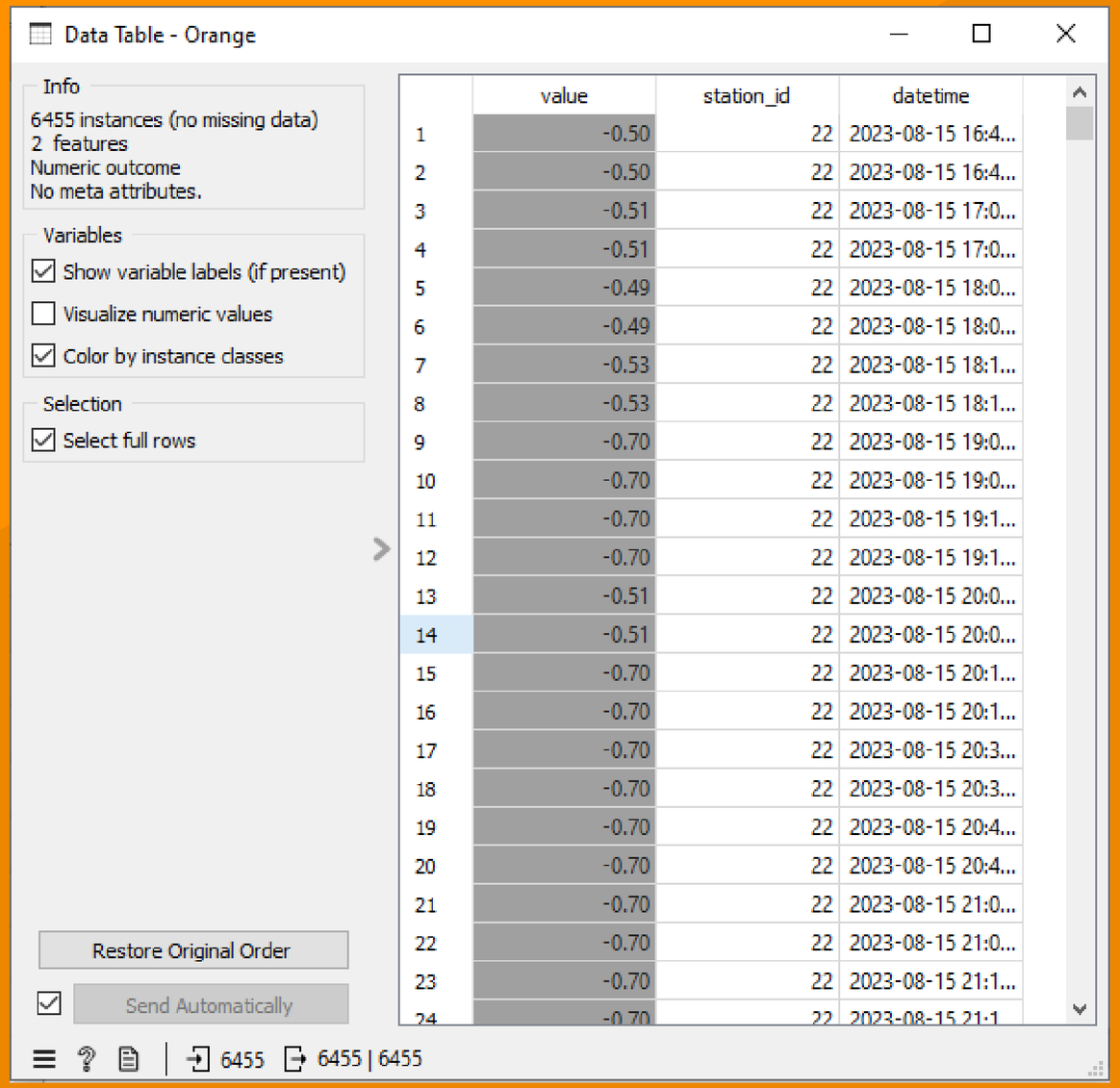
Color: N value

Send Automatically

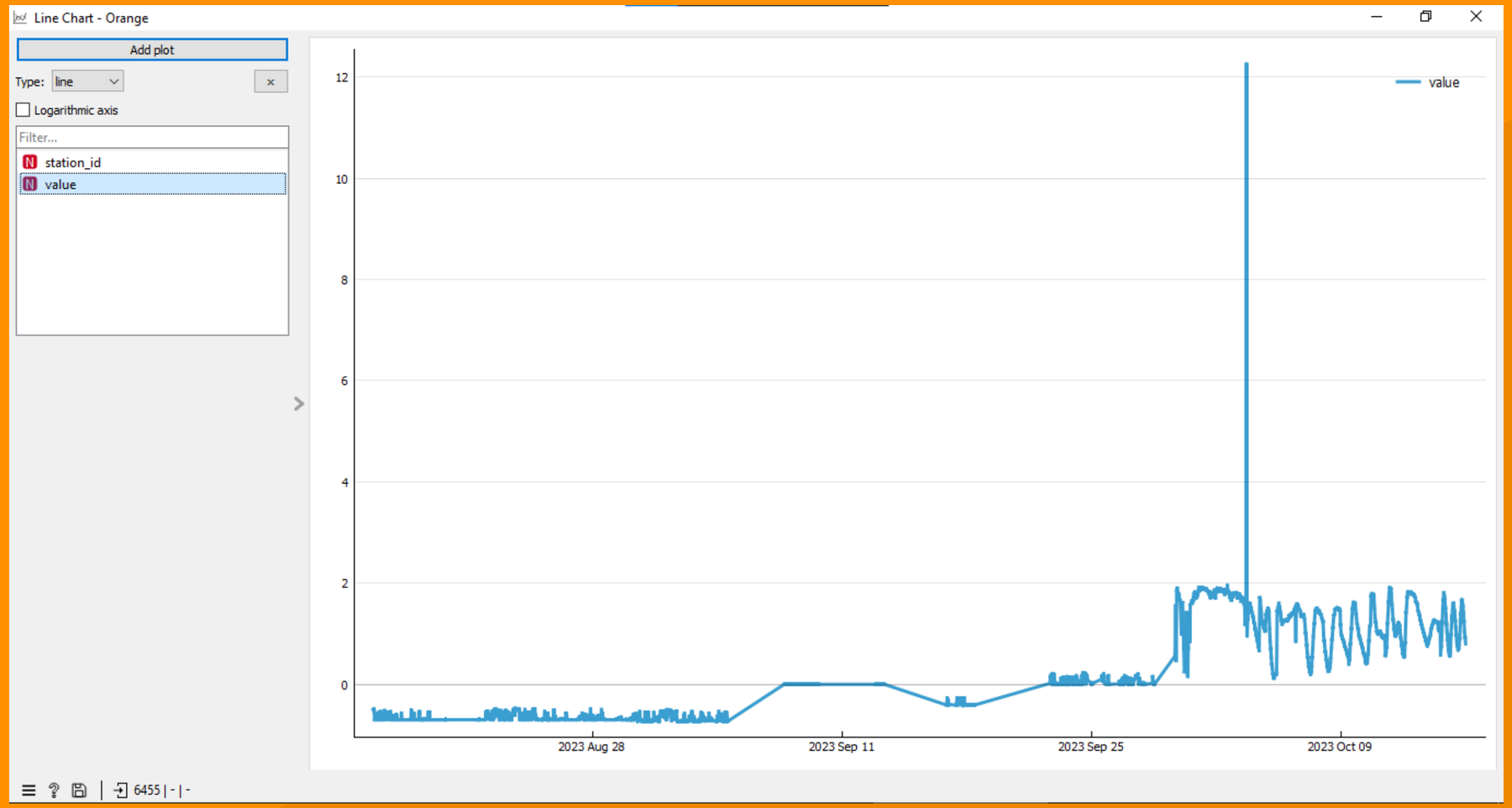
6455 - | 3

Check statistics

View data table



	value	station_id	datetime
1	-0.50	22	2023-08-15 16:4...
2	-0.50	22	2023-08-15 16:4...
3	-0.51	22	2023-08-15 17:0...
4	-0.51	22	2023-08-15 17:0...
5	-0.49	22	2023-08-15 18:0...
6	-0.49	22	2023-08-15 18:0...
7	-0.53	22	2023-08-15 18:1...
8	-0.53	22	2023-08-15 18:1...
9	-0.70	22	2023-08-15 19:0...
10	-0.70	22	2023-08-15 19:0...
11	-0.70	22	2023-08-15 19:1...
12	-0.70	22	2023-08-15 19:1...
13	-0.51	22	2023-08-15 20:0...
14	-0.51	22	2023-08-15 20:0...
15	-0.70	22	2023-08-15 20:1...
16	-0.70	22	2023-08-15 20:1...
17	-0.70	22	2023-08-15 20:3...
18	-0.70	22	2023-08-15 20:3...
19	-0.70	22	2023-08-15 20:4...
20	-0.70	22	2023-08-15 20:4...
21	-0.70	22	2023-08-15 21:0...
22	-0.70	22	2023-08-15 21:0...
23	-0.70	22	2023-08-15 21:1...
24	-0.70	22	2023-08-15 21:1


Line Chart - Orange
Add plot
Type: line
Logarithmic axis
Filter...
station_id
value
value
12
10
8
6
4
2
0
2023 Aug 28
2023 Sep 11
2023 Sep 25
2023 Oct 09
6455 | - | -

Graph plot

ANOMALY DETECTION ALGORITHMS

Local outlier factor

The Local Outlier Factor (LOF) algorithm is an unsupervised anomaly detection method which computes the local density deviation of a given data point with respect to its neighbors.

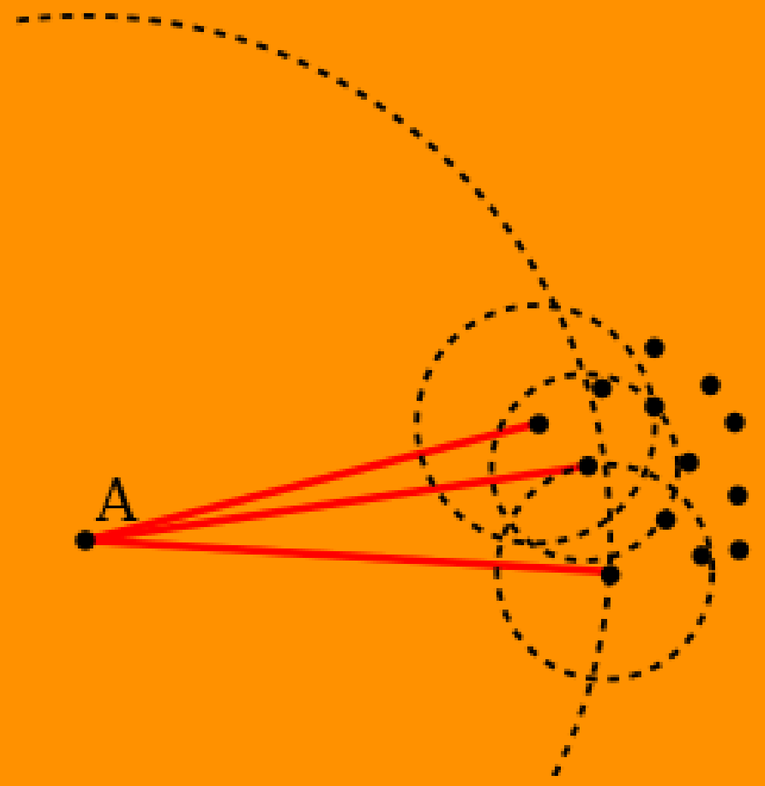
One-Class support vector machine (SVM)

A regular support vector machine algorithm tries to find a hyperplane that best separates the two classes of data points.

RMUTT
www.rmutt.ac.th ราชมงคลธัญบุรี

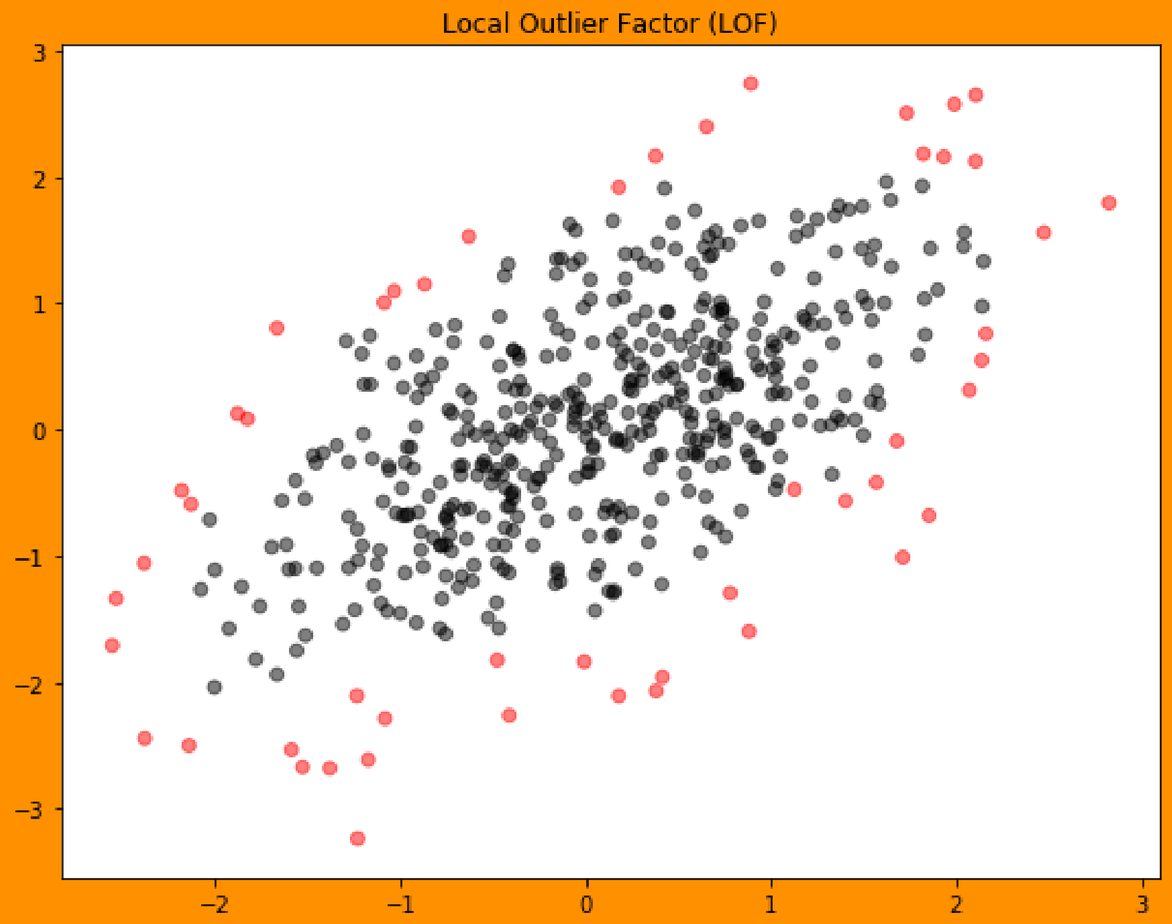
LOCAL OUTLIER FACTOR (LOF)

LOCAL OUTLIER FACTOR(LOF)



Unsupervised learning algorithm สำหรับการตรวจจับความผิดปกติ เปรียบเทียบความหนาแน่นของข้อมูลจุดต่างๆ แล้วแยกจุดที่มีความหนาแน่นน้อยออกเป็น Anomaly โดยความหนาแน่นจะคำนวนจาก K-Nearest neighbors ซึ่งก็คือระยะห่างระหว่างจุดที่เราสนใจ กับจุด "เพื่อนบ้าน" ที่ใกล้ที่สุดจำนวน K จุดตามที่เรากำหนด

LOCAL OUTLIER FACTOR(LOF)



เราสามารถเรียกใช้ Local Outlier Factor จาก Class LocalOutlierFactor ในโมดูล sklearn.neighbors

LOCAL OUTLIER FACTOR(LOF)

Local Outlier Factor, $LOF(x_i)$

Average Local Reachability-Density of datapoints in the neighborhood of x_i

$$LOF(x_i) = \frac{\sum_{x_j \in N(x_i)} lrd(x_j)}{|N(x_i)|} \times \frac{1}{lrd(x_i)}$$

Number of elements in the neighbourhood of x_i

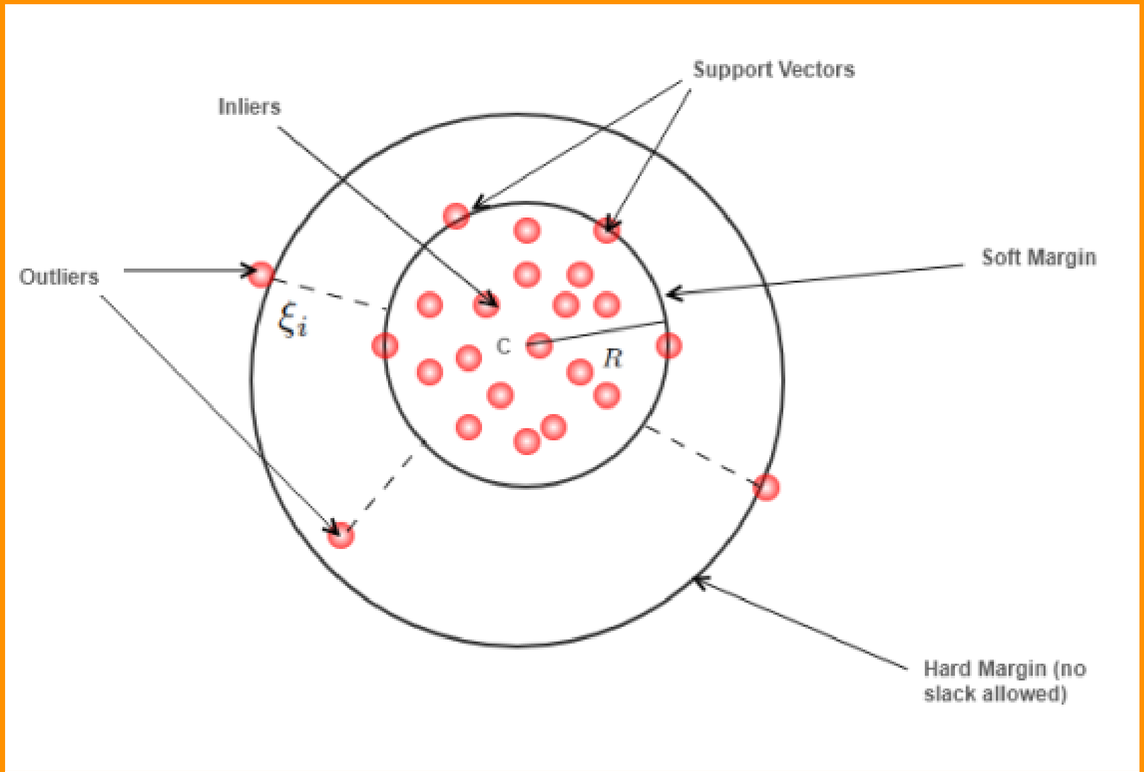
Local Reachability-Density of x_i

www.mlpoint.com

ONE-CLASS SUPPORT VECTOR MACHINE (SVM)

RMUTT
www.rmutt.ac.th ราชมงคลธัญบุรี

ONE-CLASS SUPPORT VECTOR MACHINE (SVM)



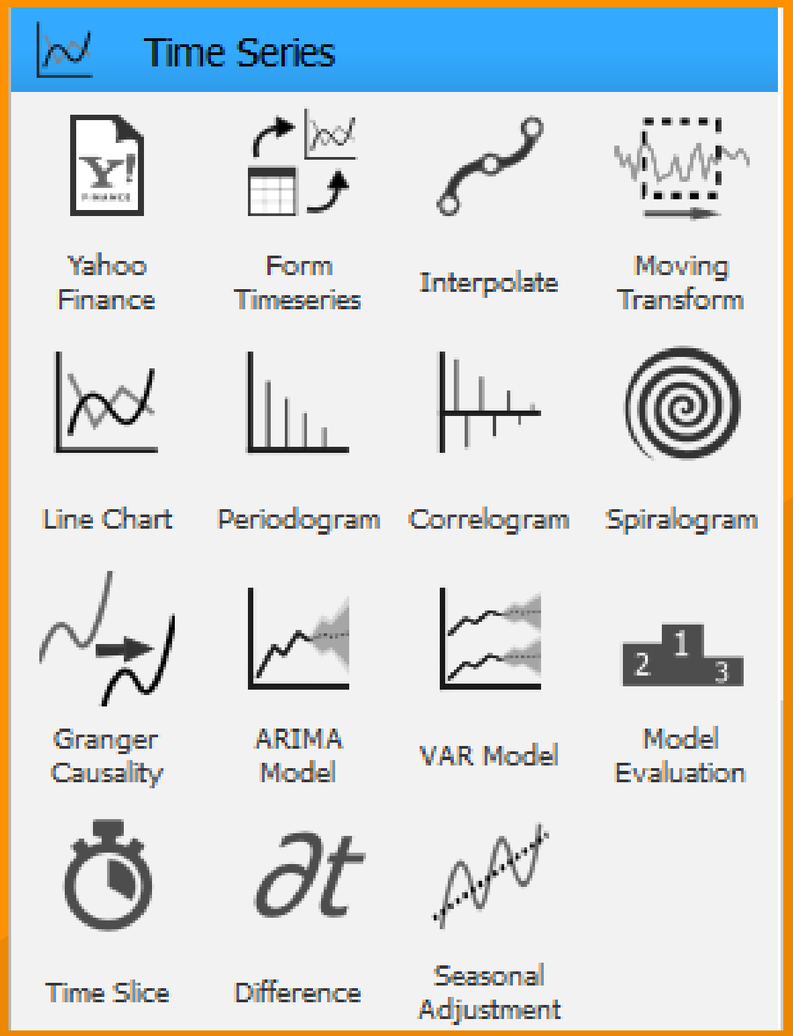
One Class SVM ทำงานโดยการสร้าง Support Vector ด้วย RBF Kernel เพื่อเป็นเส้นขอบเขตการตัดสินใจ และเป็น Supervised Learning ที่มีเป้าหมาย คือ การหา Outlier

Moving
Transform

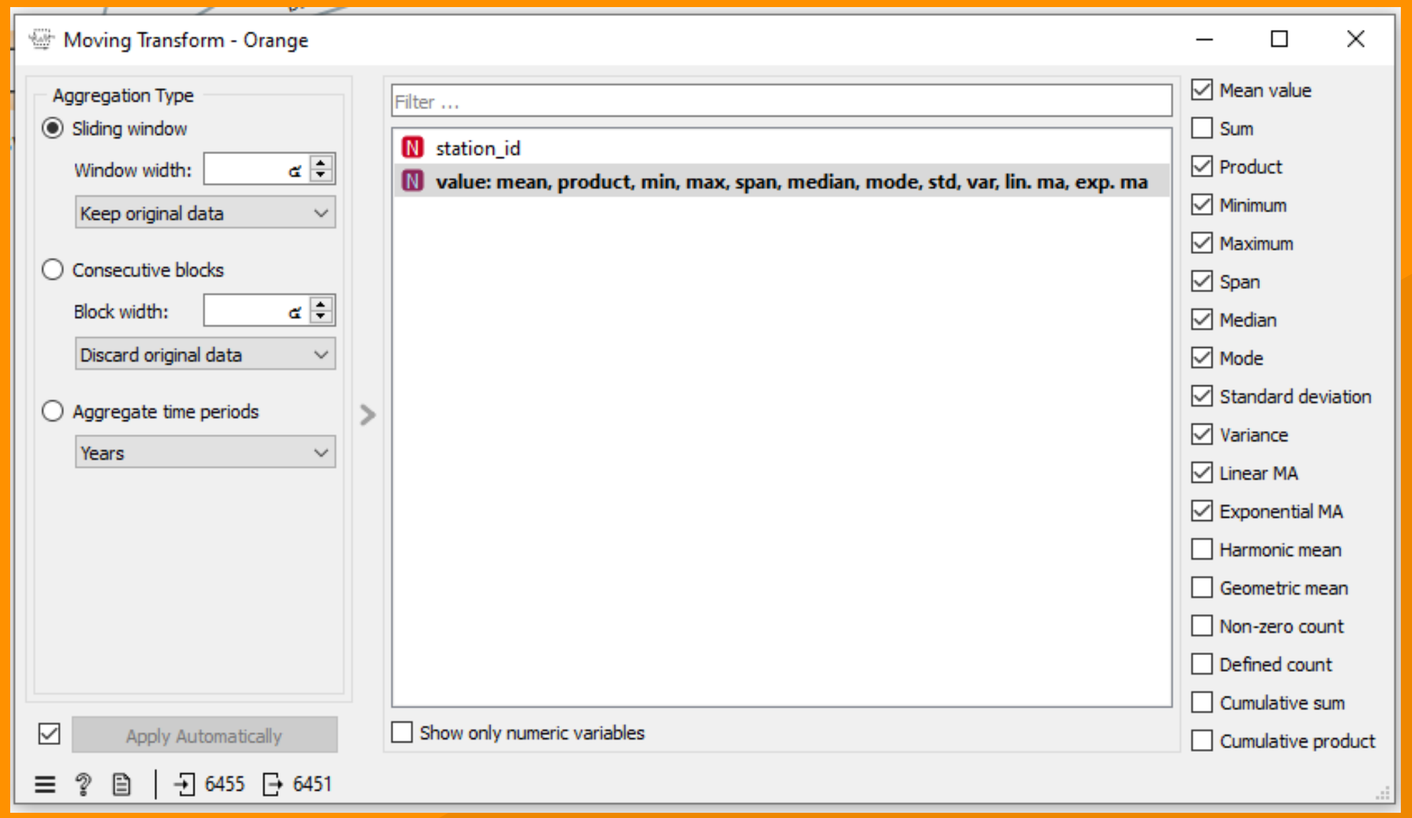
TIME SERIES

Add-on

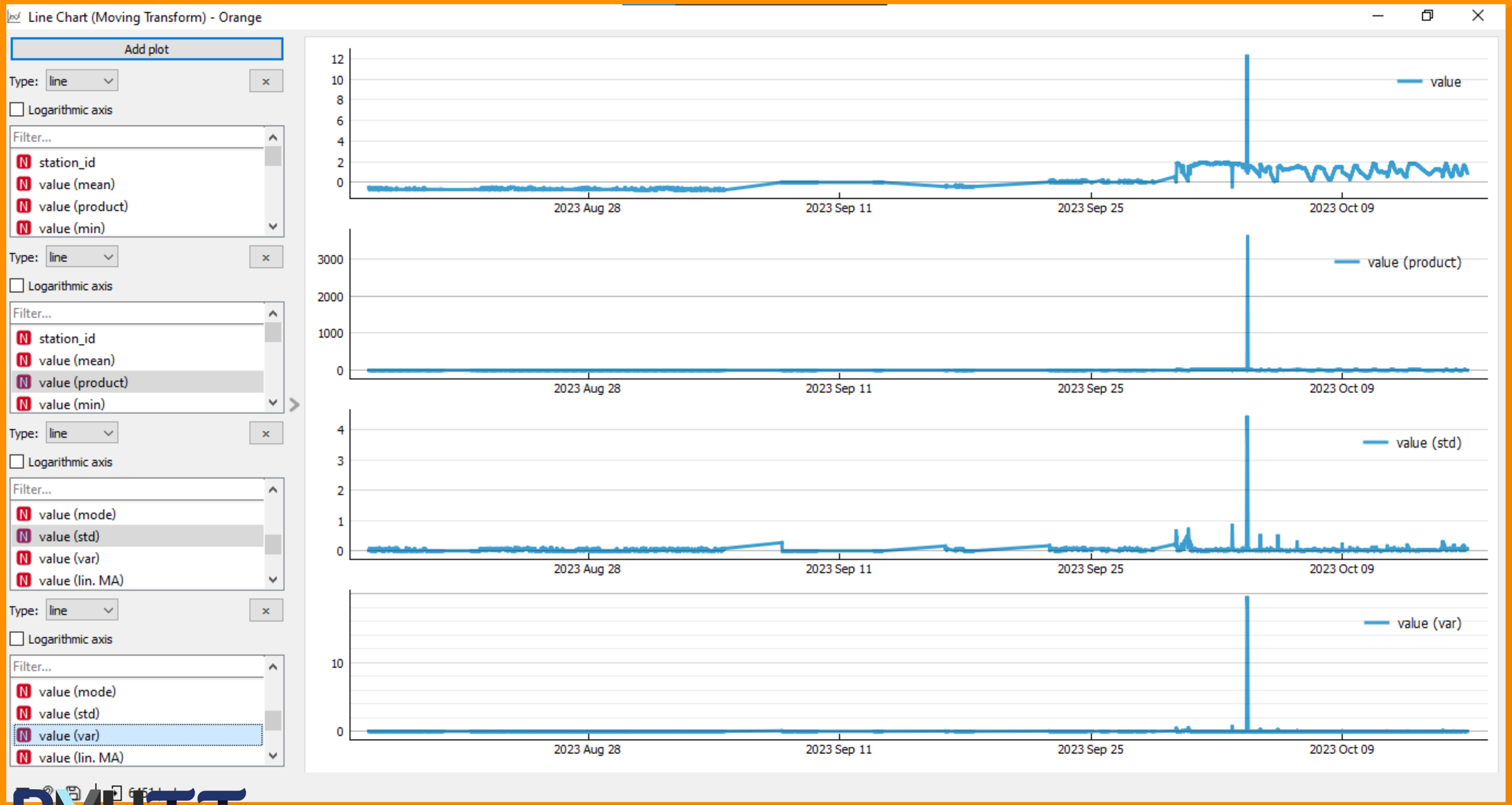
RMUTT
www.rmutt.ac.th ราชมงคลธัญบุรี



Timeseries add-on



Moving Transform



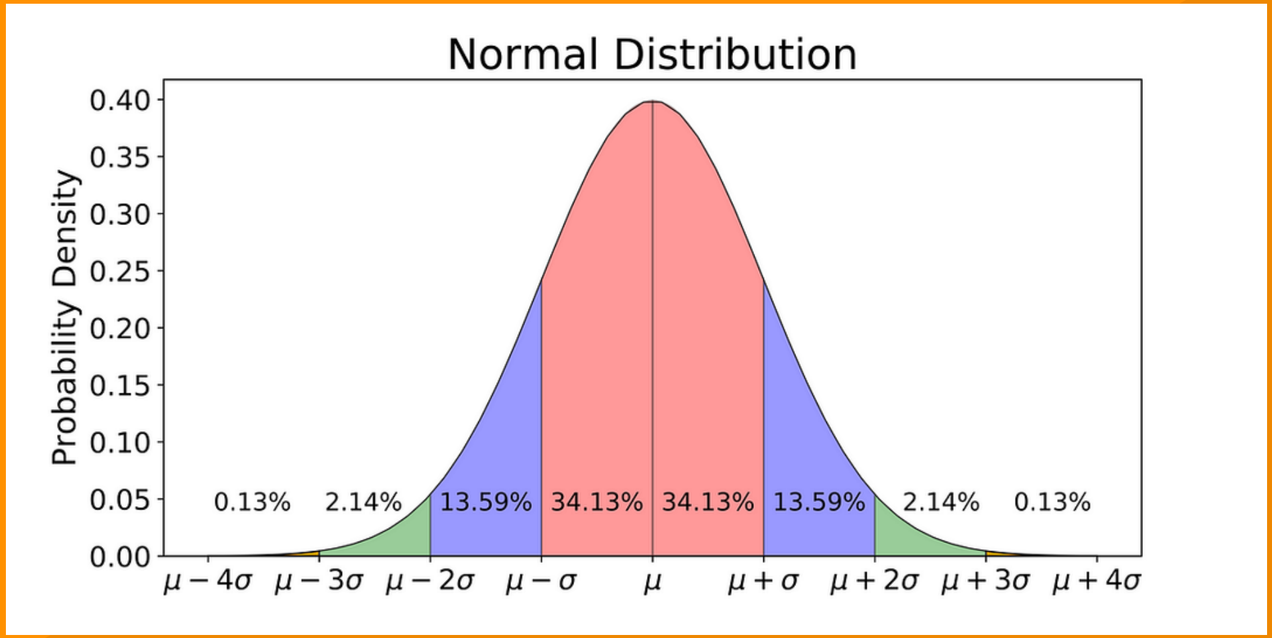
Line Chart (Moving Transform)

Gaussian Distribution (Product)

Normal Distribution

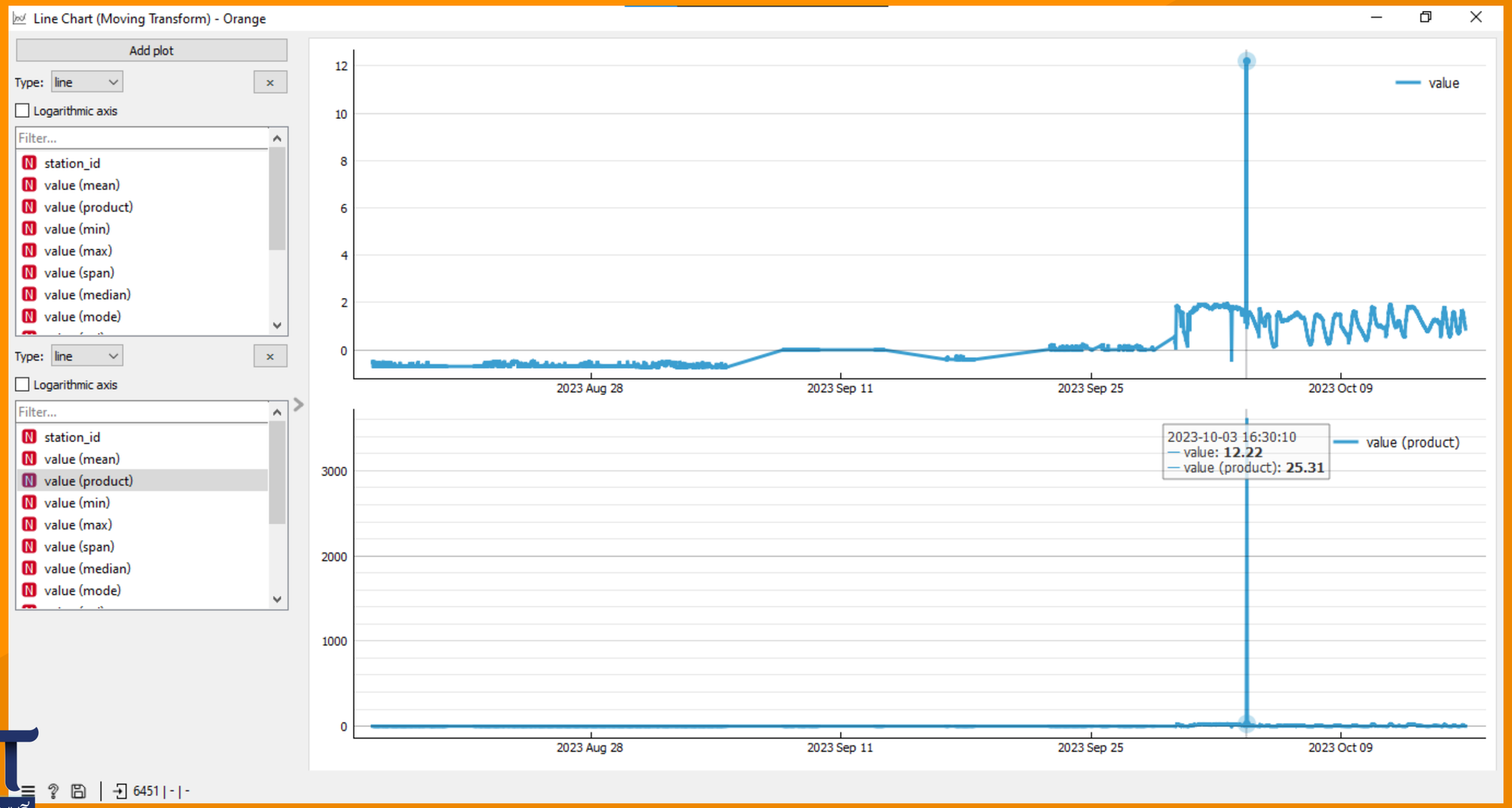
ชื่อไทยคือ การแจกแจงแบบปกติ

มีอีกชื่อหนึ่งคือ Gaussian Distribution



www.rmutt.ac.th ราชมงคลธัญบุรี

Gaussian Distribution (Product)



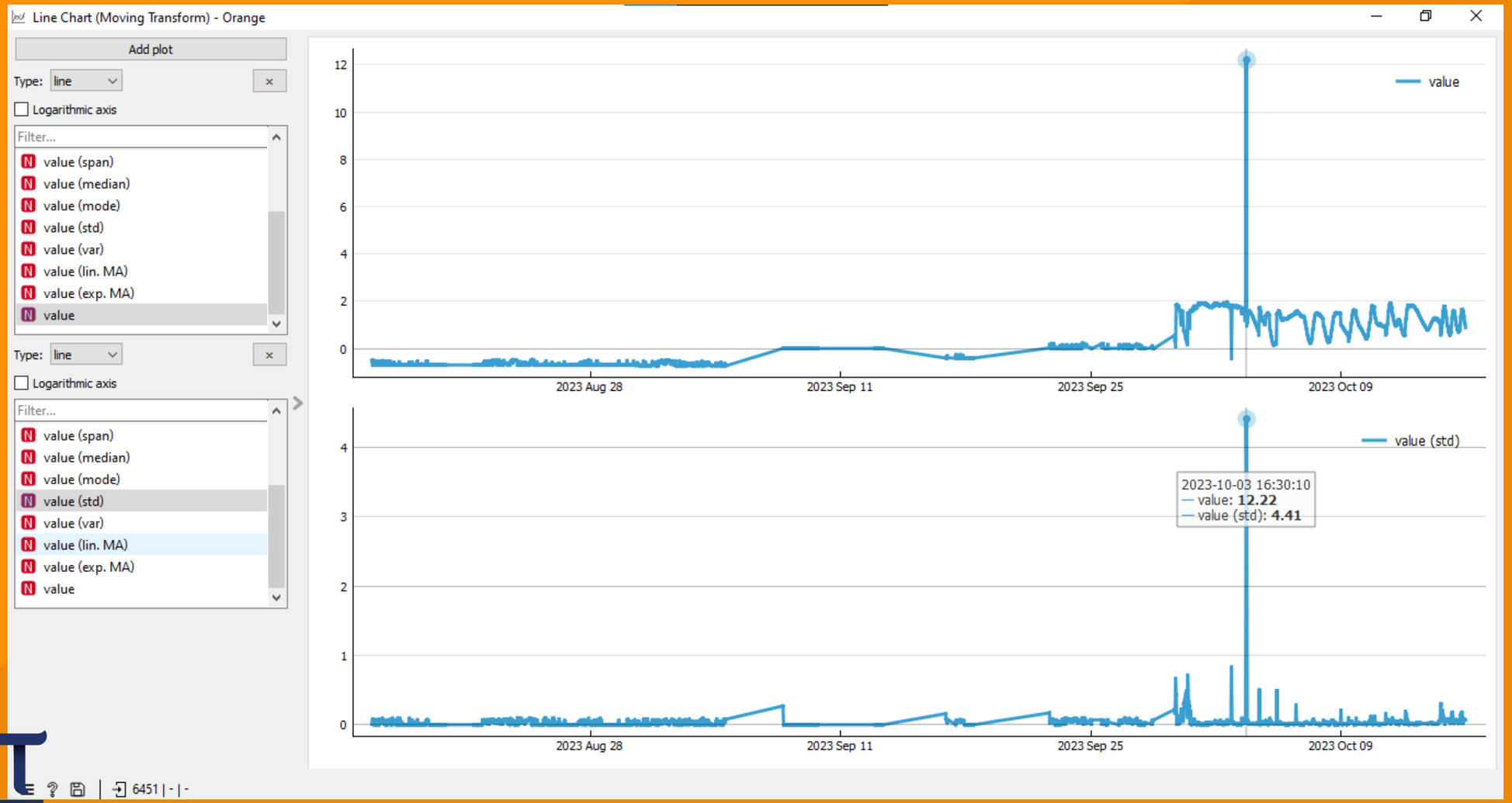
Standard Deviation

ส่วนเบี่ยงเบนมาตรฐาน คือตัววัดระดับความเอนเอียงของข้อมูลจากข้อมูล เนื่องจากคำถามที่ว่า <u>ข้อมูลนี่มันเหวี่ยงแค่ไหนจากข้อมูล</u>

$$s = \sqrt{\frac{\sum(x - \bar{x})^2}{n - 1}}$$

RMUTT
www.rmutt.ac.th ราชมงคลธัญบุรี

Standard Deviation



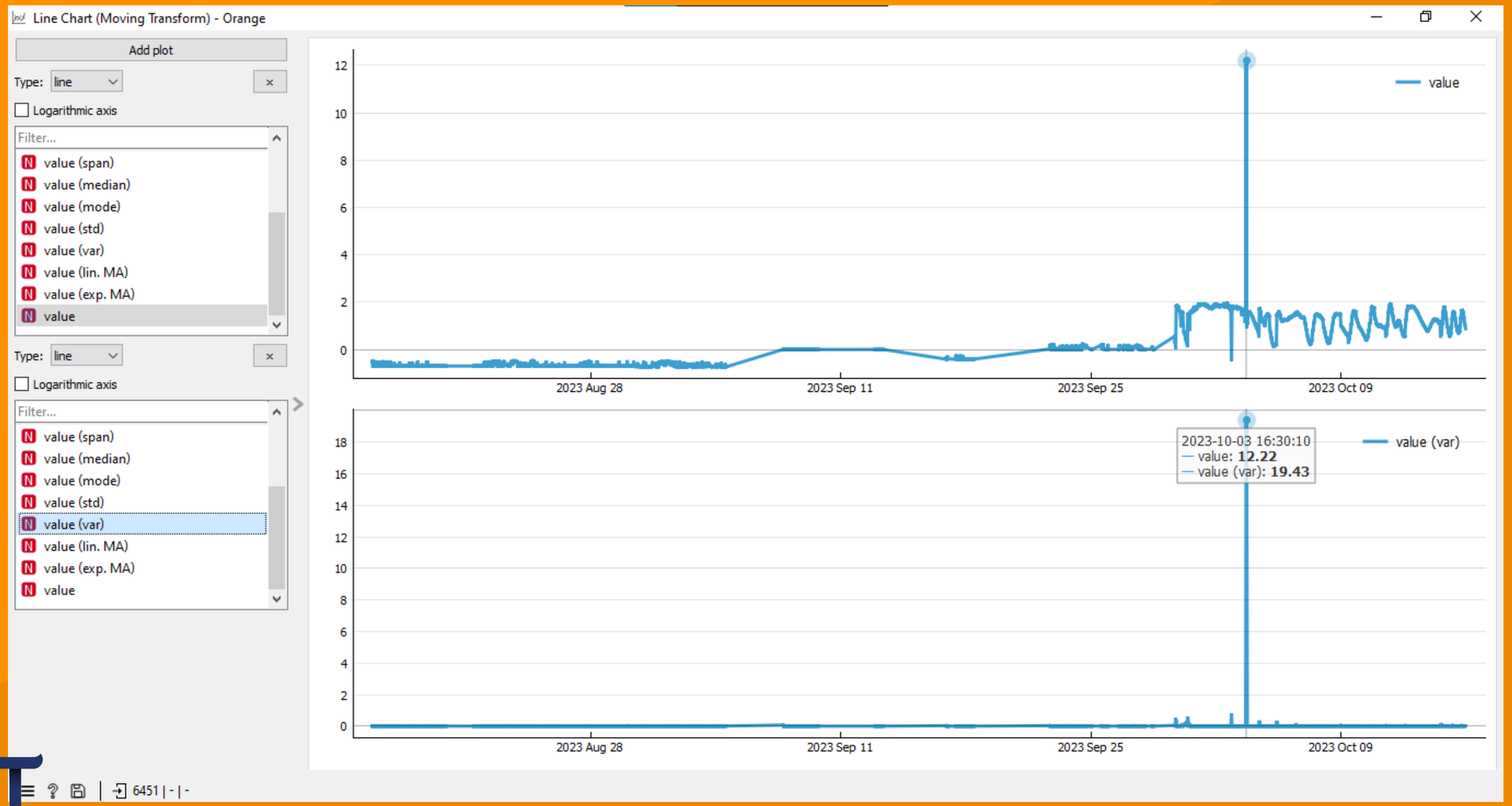
Variance

ความแปรปรวนก็คือการวัดว่าข้อมูลนั้นแจกแจงออกห่างจากค่าเฉลี่ยของข้อมูลมากน้อยแค่ไหนตามสมการนี้

$$S^2 = \frac{\sum (X - \bar{X})^2}{N-1}$$

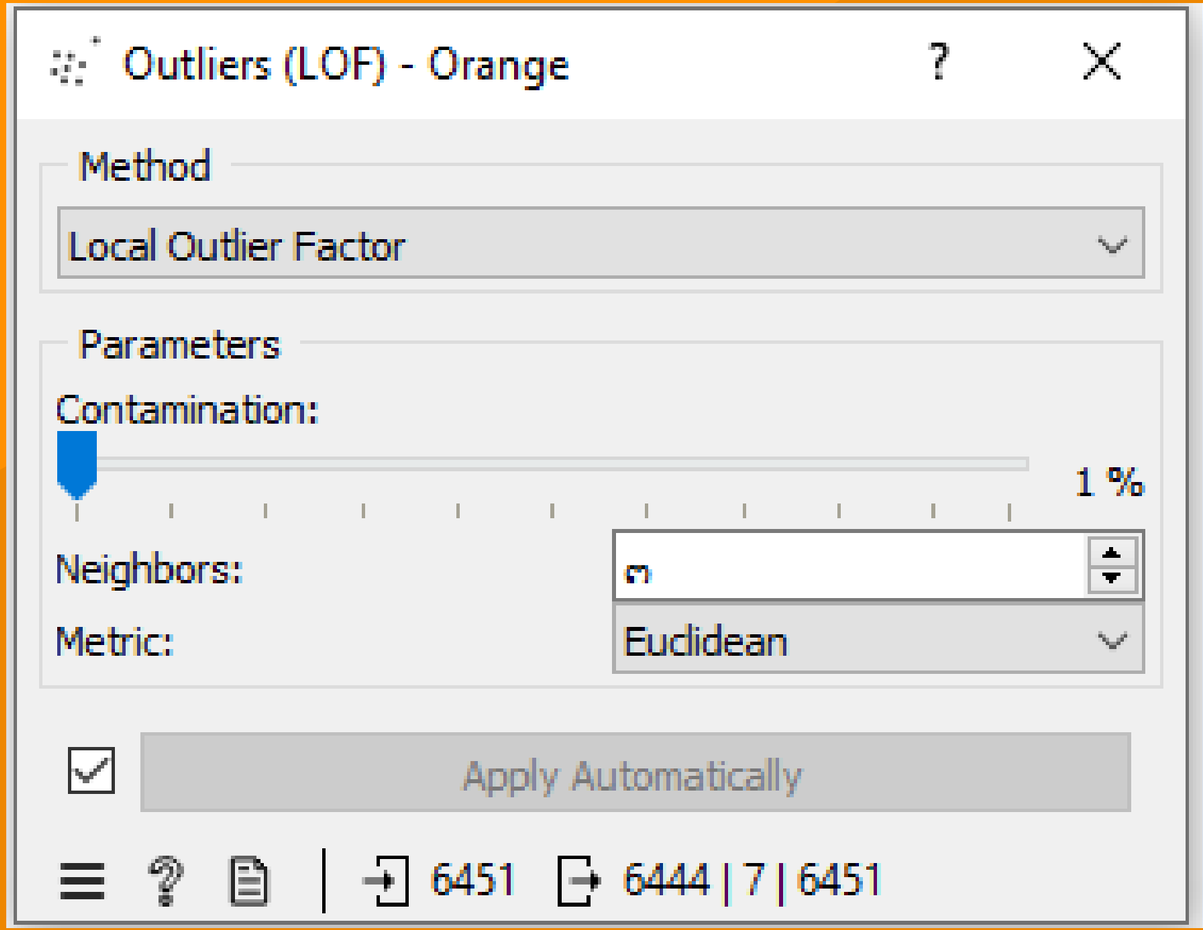
RMUTT
www.rmutt.ac.th ราชมงคลธัญบุรี

Variance


Line Chart (Moving Transform) - Orange
Add plot
Type: line
Logarithmic axis
Filter...
value (span)
value (median)
value (mode)
value (std)
value (var)
value (lin. MA)
value (exp. MA)
value
Type: line
Logarithmic axis
Filter...
value (span)
value (median)
value (mode)
value (std)
value (var)
value (lin. MA)
value (exp. MA)
value
value
2023 Aug 28
2023 Sep 11
2023 Sep 25
2023 Oct 09
2023-10-03 16:30:10
value: 12.22
value (var): 19.43
value (var)
6451 | - | -

OUTLIERS (LOF)

Outliers (LOF)



RMUTT
www.rmutt.ac.th ราชมงคลธัญบุรี

Outliers (LOF)

Inliers: Table with 6444 instances, 14 variables
Features: 13 (12 numeric, 1 time) (no missing values)

	value	station_id	datetime	value (mean)	value (product)	value (min)
1	-0.49	22	2023-08-15	-0.5010	-0.0315	-0.51

Outliers: Table with 7 instances, 14 variables
Features: 13 (12 numeric, 1 time) (no missing values)

	value	station_id	datetime	value (mean)	value (product)	value (min)	v
1	-0.70	22	2023-08-21	-0.70	-0.1681	-0.70	-0.70

Data: Table with 6451 instances, 15 variables
Features: 13 (12 numeric, 1 time) (no missing values)

	value	Outlier	station_id	datetime	value (mean)	value (product)
1	-0.49	No	22	2023-08-15	-0.5010	-0.0315

Data Table outliers (LOF)

Data Table outliers (LOF) - Orange

Info

7 instances (no missing data)
13 features
Numeric outcome
No meta attributes.

Variables

- [x] Show variable labels (if present)
- [ ] Visualize numeric values
- [x] Color by instance classes

Selection

- [x] Select full rows

Restore Original Order

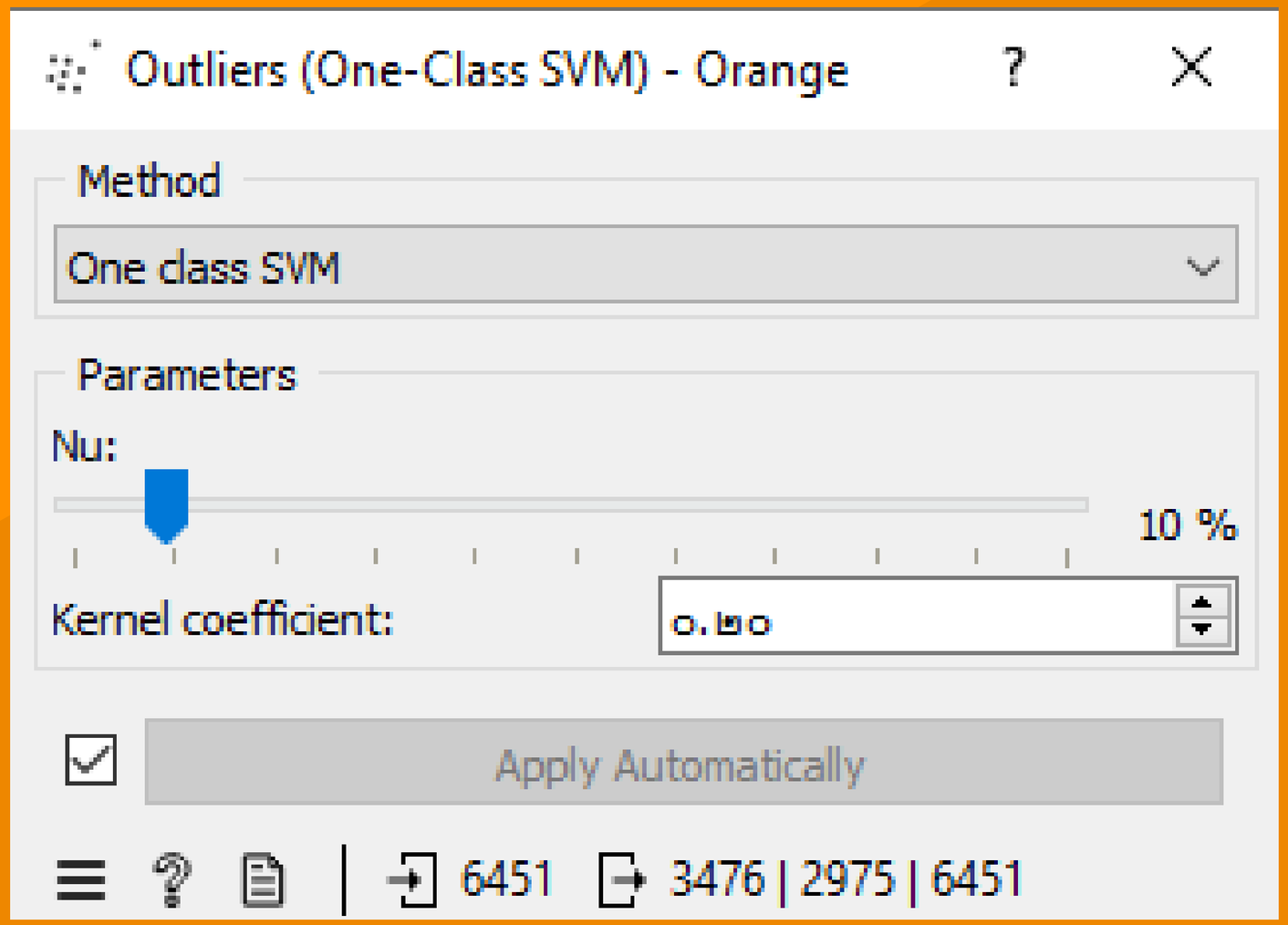
[x] Send Automatically

7 | 7

	value	station_id	datetime	value (mean)	value (product)	v
1	-0.70	22	2023-08-21 19:3...	-0.70	-0.1681	
2	-0.73	22	2023-09-04 14:1...	-0.6580	-0.1187	
3	-0.73	22	2023-09-04 14:1...	-0.66	-0.1203	
4	12.25	22	2023-10-03 16:4...	7.8060	2504.0917	
5	12.25	22	2023-10-03 16:4...	10.0220	26218.0538	
6	12.31	22	2023-10-03 17:0...	12.25	275849.7801	
7	12.31	22	2023-10-03 17:0...	12.2680	277881.4070	

OUTLIERS
(ONE-CLASS SVM)

Outliers (one-class SVM)



RMUTT
www.rmutt.ac.th ราชมงคลธัญบุรี

Outliers (one-class SVM)

Inliers: Table with 3476 instances, 14 variables
Features: 13 (12 numeric, 1 time) (no missing values)

	value	station_id	datetime	value (mean)	value (product)	value (min)
1	-0.49	22	2023-08-15	-0.5010	-0.0315	-0.51

Outliers: Table with 2975 instances, 14 variables
Features: 13 (12 numeric, 1 time) (no missing values)

	value	station_id	datetime	value (mean)	value (product)	value (min)
1	-0.49	22	2023-08-15	-0.5010	-0.0315	-0.51

Data: Table with 6451 instances, 15 variables
Features: 13 (12 numeric, 1 time) (no missing values)

	value	Outlier	station_id	datetime	value (mean)	value (product)
1	-0.49	No	22	2023-08-15	-0.5010	-0.0315

RMUTT
www.rmutt.ac.th ราชมงคลธัญบุรี

Data Table outliers (one-class SVM)

Data Table outliers (One-Class SVM) - Orange

Info
2975 instances (no missing data)
13 features
Numeric outcome
No meta attributes.

Variables
- [x] Show variable labels (if present)
- [ ] Visualize numeric values
- [x] Color by instance classes

Selection
- [x] Select full rows

Restore Original Order

Send Automatically

	value	station_id	datetime	value (mean)	value (product)	value (min)	value (max)
2427	1.67	22	2023-10-03 14:1...	1.6060	10.6659	1.56	1.67
2428	1.66	22	2023-10-03 14:3...	1.6480	12.1499	1.62	1.67
2429	1.22	22	2023-10-03 14:4...	1.5760	9.3758	1.22	1.67
2430	1.18	22	2023-10-03 15:0...	1.3880	4.8397	1.18	1.66
2431	0.98	22	2023-10-03 15:3...	1.14	1.9000	0.98	1.18
2432	0.98	22	2023-10-03 15:3...	1.10	1.5780	0.98	1.18
2433	1.67	22	2023-10-03 15:4...	1.1980	2.2332	0.98	1.67
2434	1.67	22	2023-10-03 15:4...	1.2960	3.1606	0.98	1.67
2435	1.23	22	2023-10-03 16:0...	1.3060	3.2945	0.98	1.67
2436	1.23	22	2023-10-03 16:0...	1.3560	4.1349	0.98	1.67
2437	1.17	22	2023-10-03 16:1...	1.3940	4.9366	1.17	1.67
2438	1.17	22	2023-10-03 16:1...	1.2940	3.4586	1.17	1.67
2439	12.25	22	2023-10-03 16:4...	7.8060	2504.0917	1.17	12.25
2440	12.31	22	2023-10-03 17:0...	12.25	275849.7801	12.22	12.31
2441	0.95	22	2023-10-03 17:1...	4.7157	3595.2702	0.88	8.5433
2442	0.95	22	2023-10-03 17:3...	4.7157	3595.2702	0.88	8.5433
2443	0.95	22	2023-10-03 17:4...	4.7157	3595.2702	0.88	8.5433
2444	1.11	22	2023-10-03 18:0...	1.01	1.0435	0.96	1.11
2445	1.29	22	2023-10-03 18:3...	1.1860	2.3271	1.11	1.29
2446	1.36	22	2023-10-03 18:4...	1.2720	3.3135	1.21	1.36
2447	1.46	22	2023-10-03 19:1...	1.40	5.3687	1.36	1.46
2448	1.51	22	2023-10-03 19:3...	1.45	6.3991	1.41	1.51
2449	1.54	22	2023-10-03 19:4...	1.4960	7.4848	1.46	1.54

2975 | 2 | 2975

COMPARE

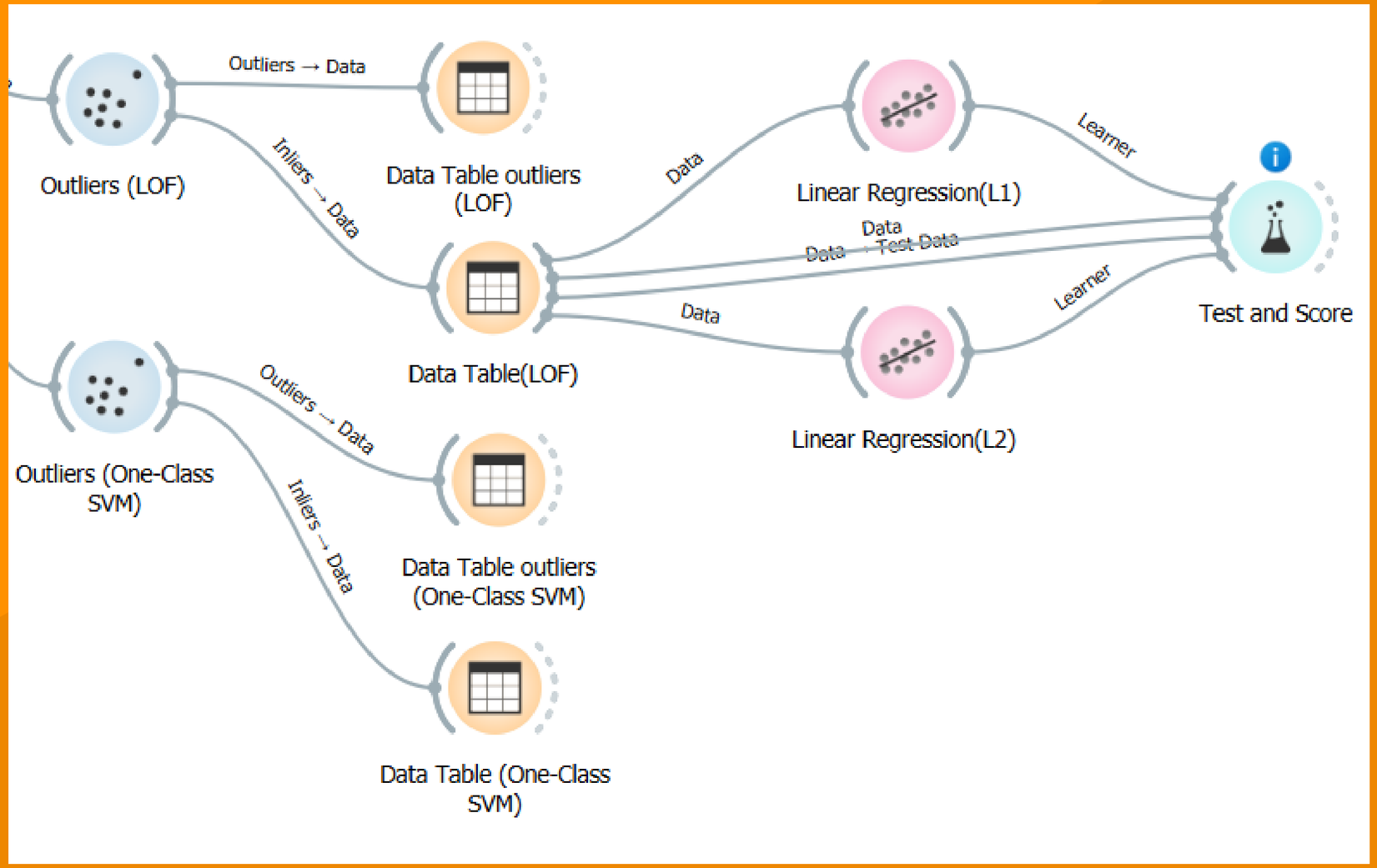
COMPARE MODEL

Model/Algorithm	Prediction	Actual	Score
Product Standard Deviation Product	**6**	**6**	**100**
LOF	4	6	66.66
one-class SVM	2	6	33.33

REGRESSION

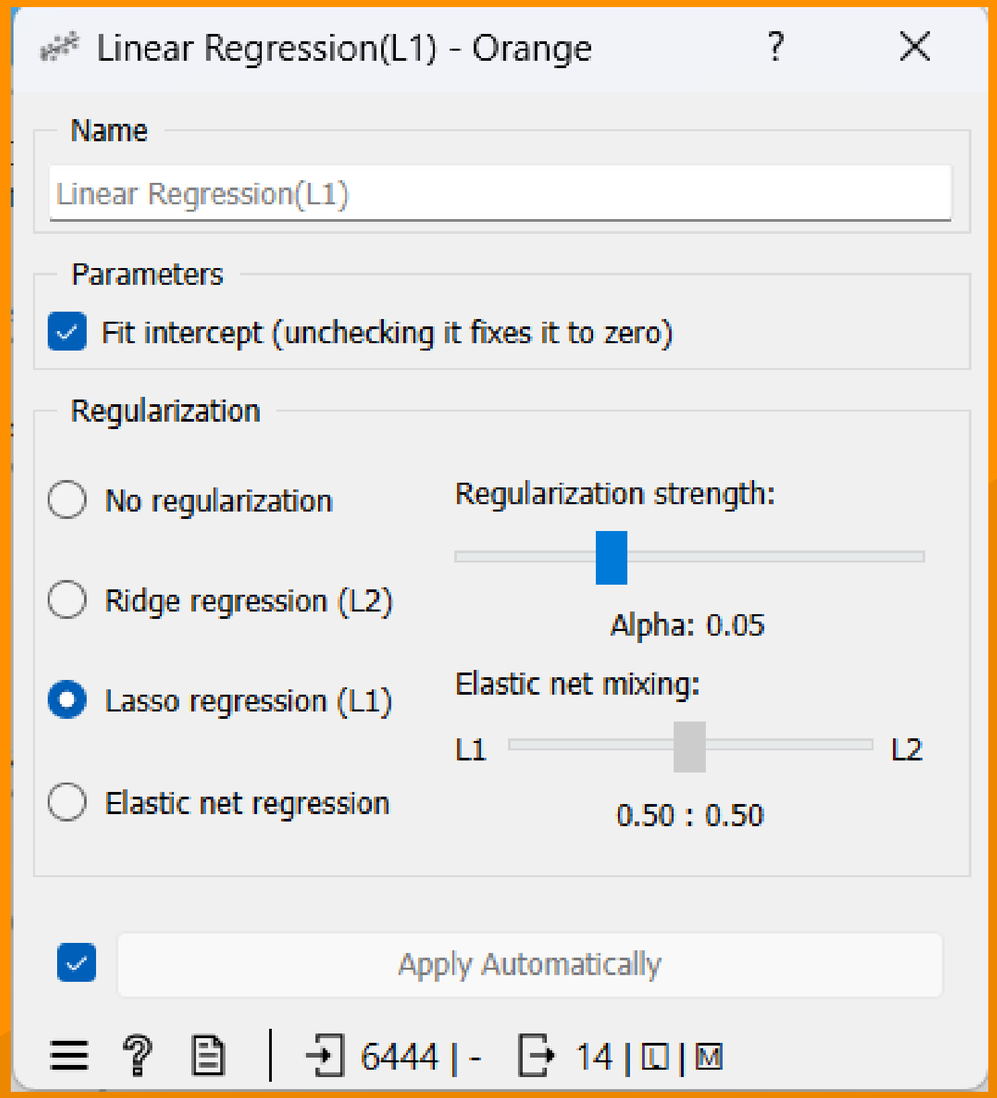
RMUTT
www.rmutt.ac.th ราชมงคลธัญบุรี

Overview



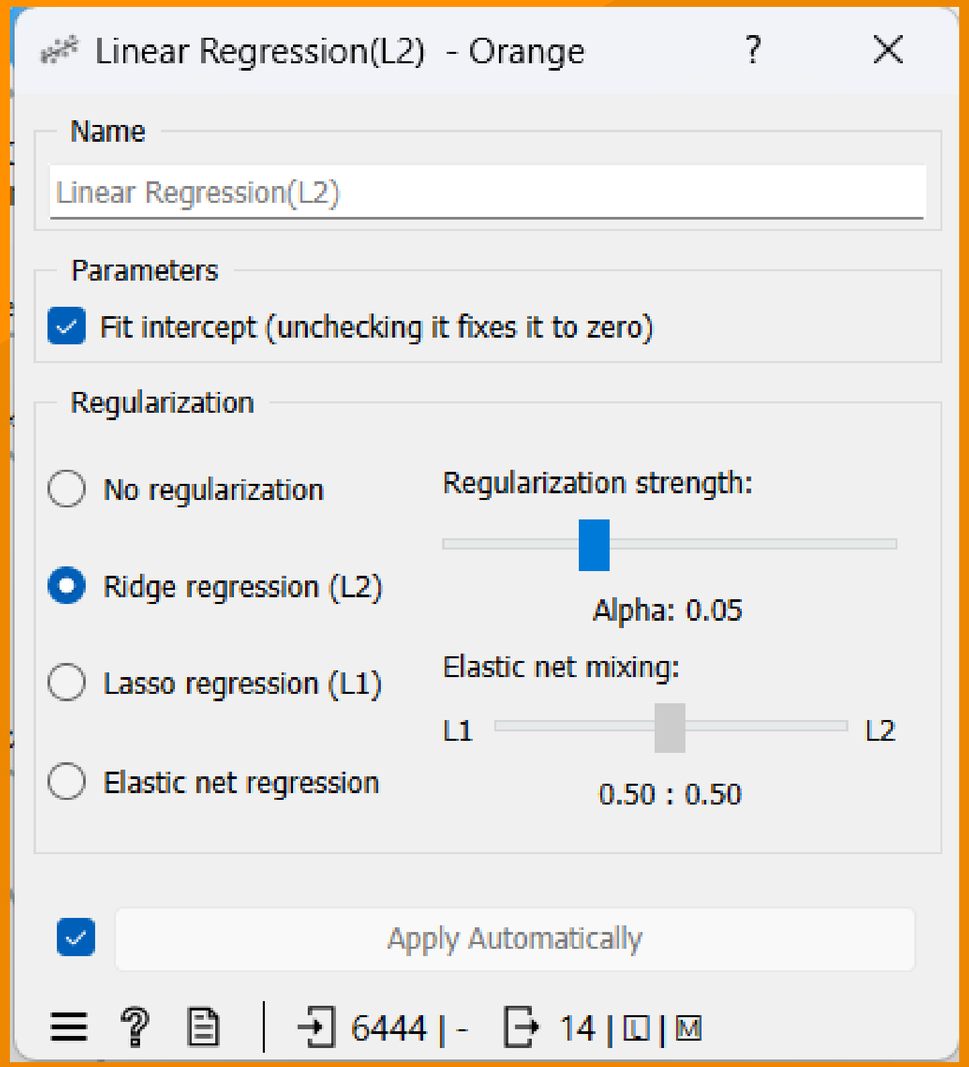
RMUTT
www.rmutt.ac.th ราชมงคลธัญบุรี

Linear Regression(L1)



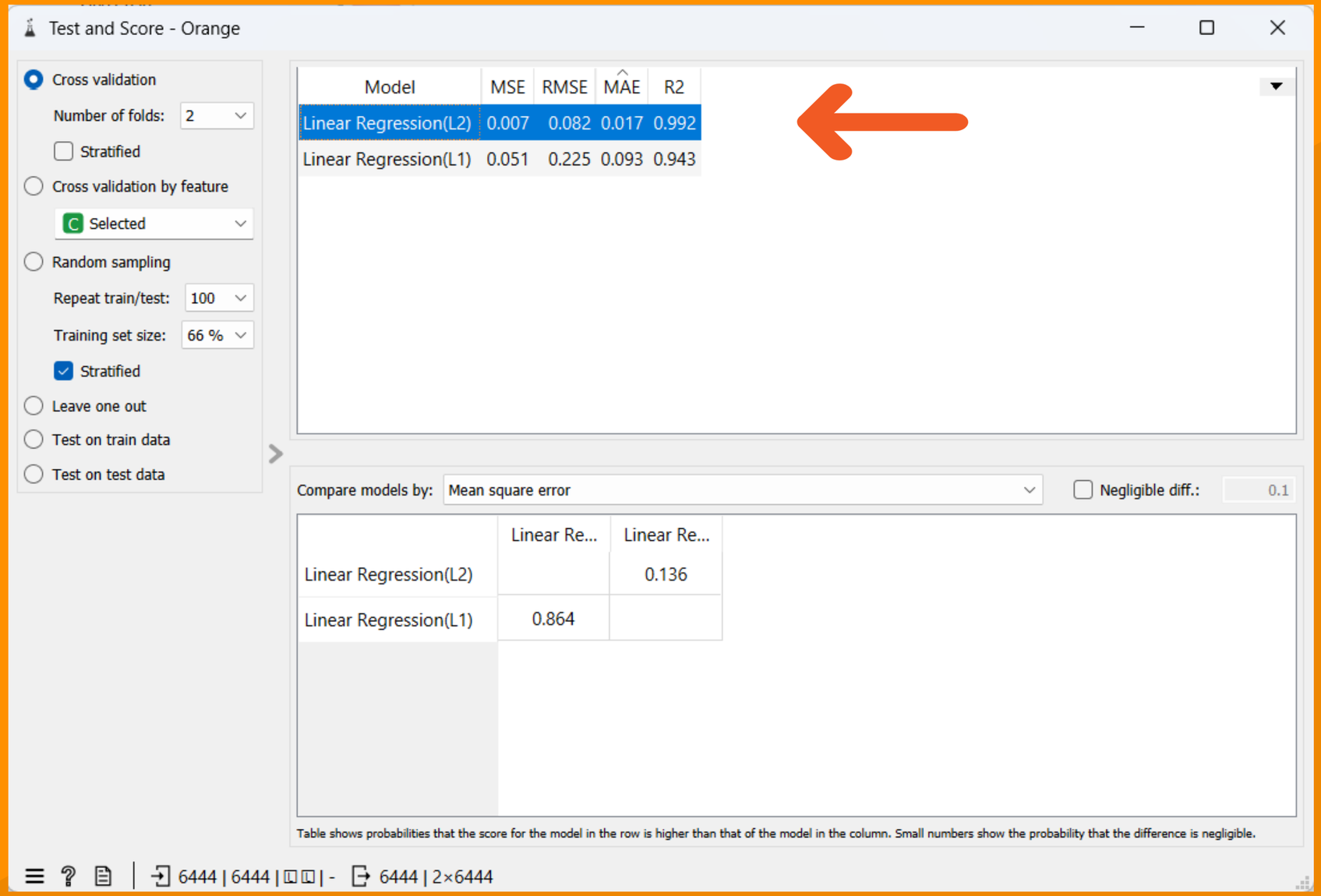
RMUTT
www.rmutt.ac.th ราชมงคลธัญบุรี

Linear Regression(L2)



TEST AND SCORE

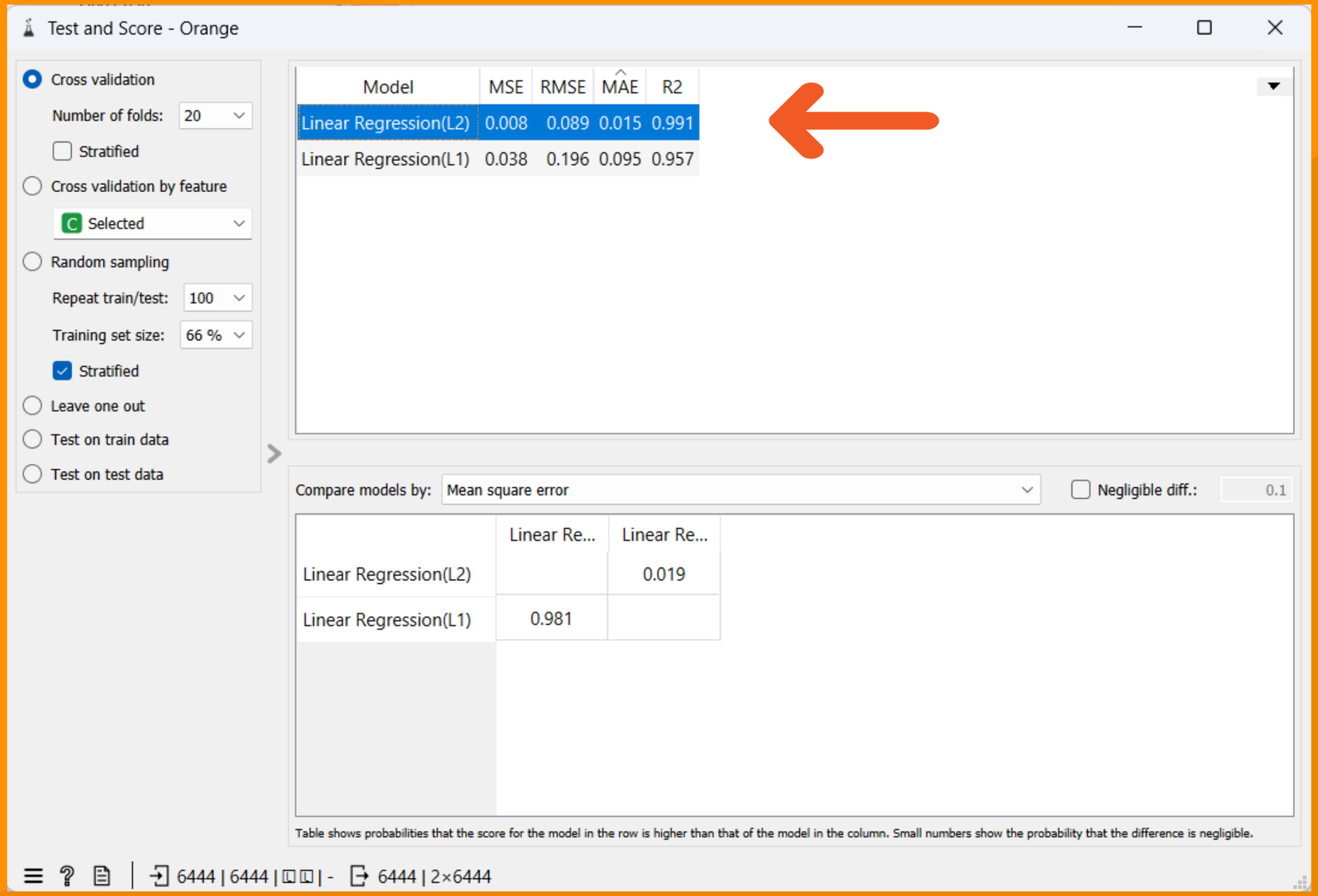
Test and Score : Cross-validation



k-fold = 2

RMUTT
www.rmutt.ac.th ราชมงคลธัญบุรี

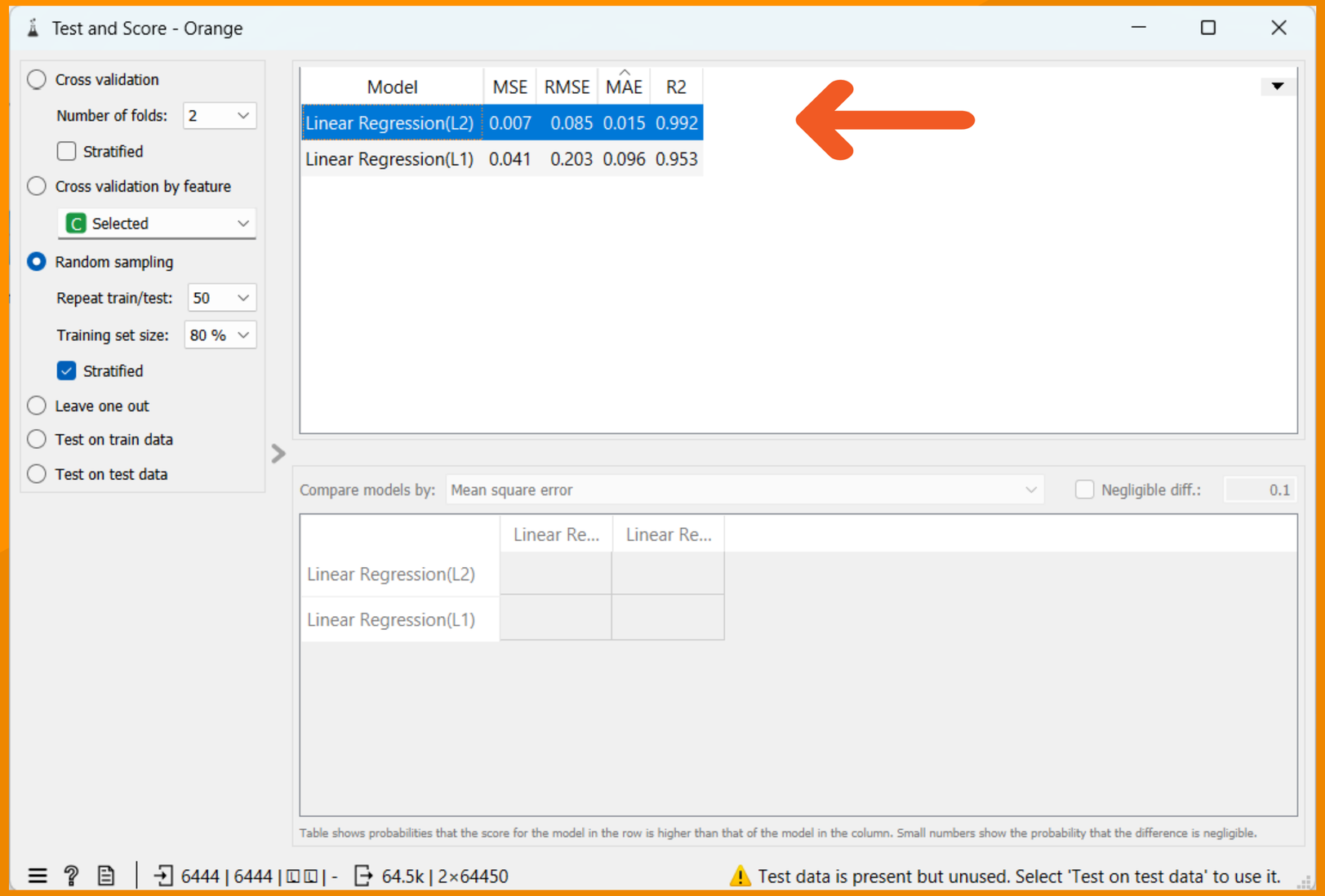
Test and Score : Cross-validation



k-fold = 20

RMUTT
www.rmutt.ac.th ราชมงคลธัญบุรี

Test and Score : Random sampling



RMUTT
www.rmutt.ac.th ราชมงคลธัญบุรี

SUMMARY

Summary

สำหรับข้อมูลระดับน้ำแม่น้ำในระบบชลประทาน
ไฟล์ : TC12.csv

- ควรใช้เพื่อตรวจจับค่าผิดปกติ(Anomaly Detection)
- ควรใช้ อัลกอลิทิม Local outlier factor(LOF).

ข้อควรระวัง ต้องดูค่าสำคัญทางสถิติ เพื่อประกอบการตัดสินใจ

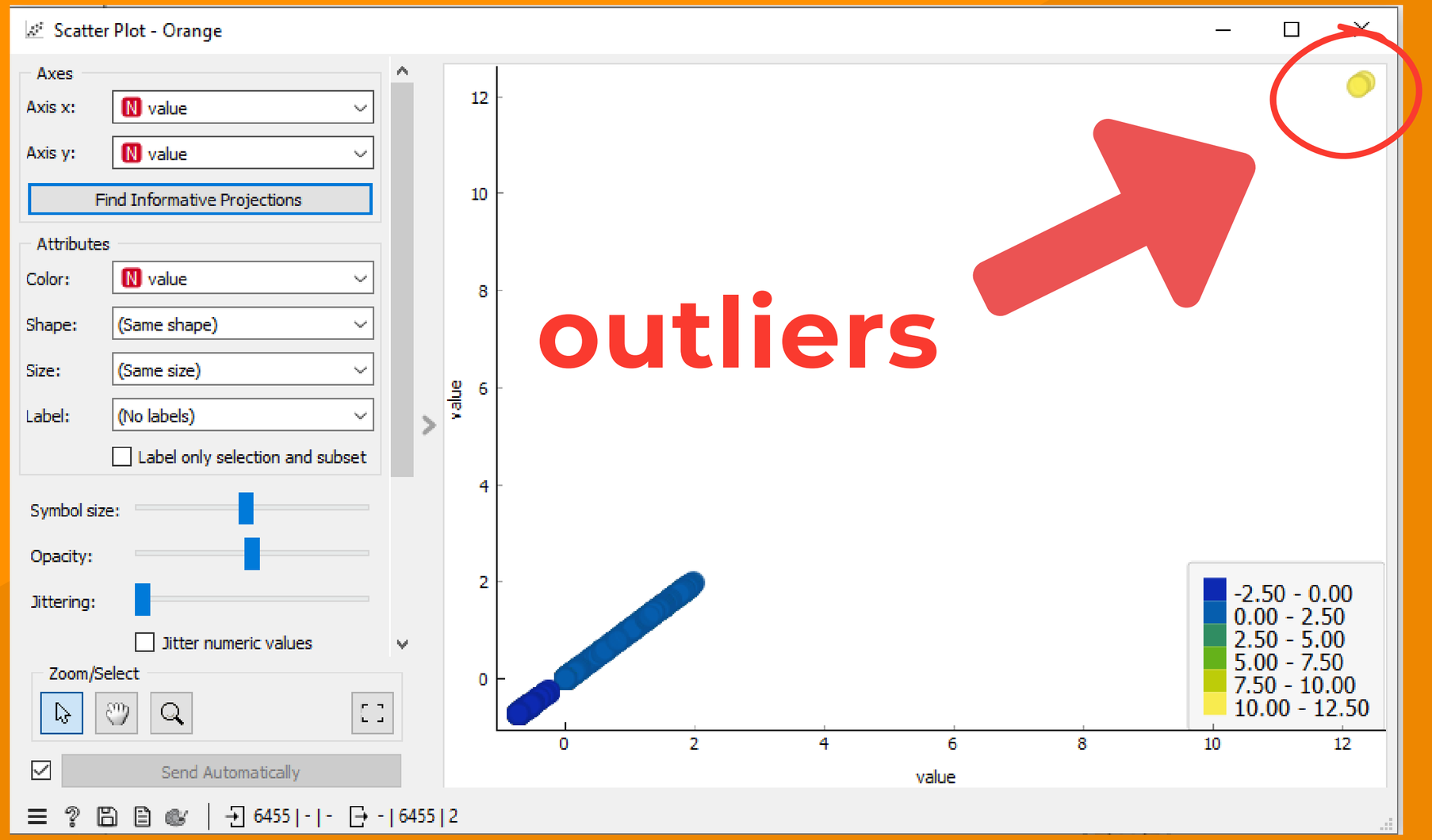
Note

REMARK

Scatter Plot

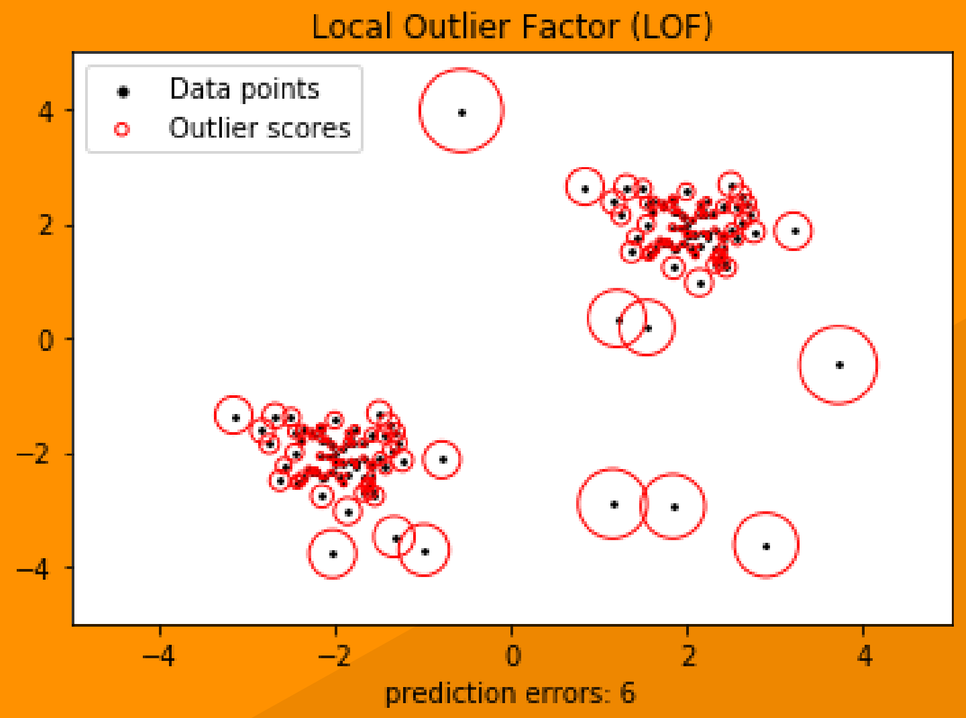
RMUTT
www.rmutt.ac.th ราชมงคลธัญบุรี

Scatter Plot



RMUTT
www.rmutt.ac.th ราชมงคลธัญบุรี

NOTE : Local outlier factor Disadvantages

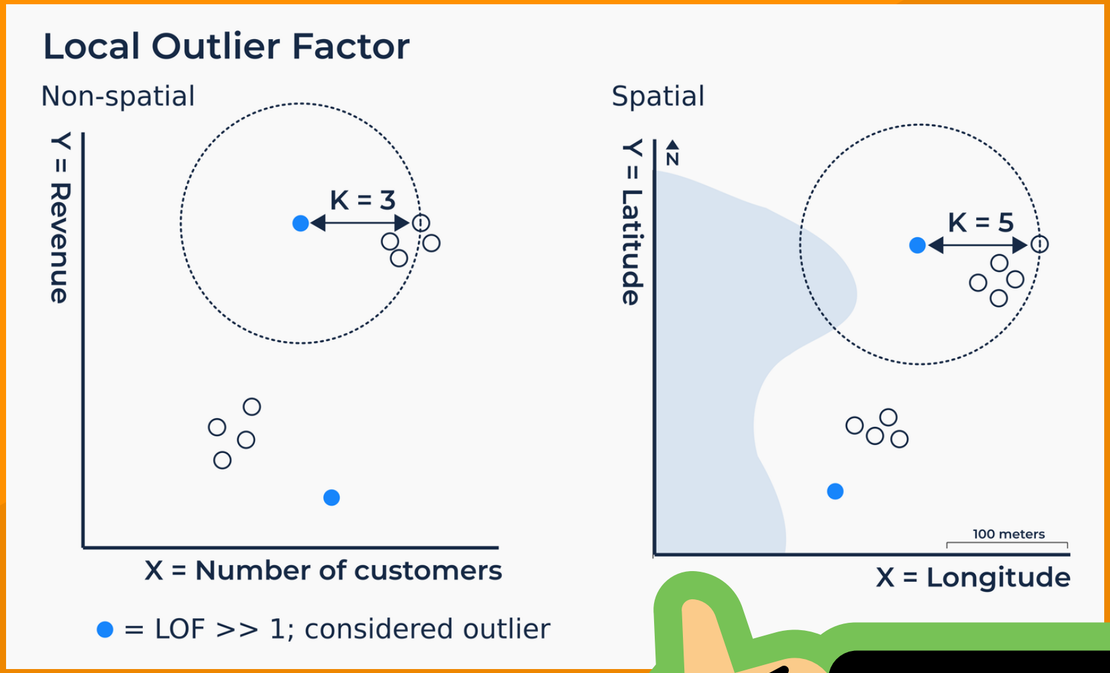


✕ FAILED

In higher dimensions, the LOF algorithm detection accuracy gets effected.

RMUTT
www.rmutt.ac.th ราชมงคลธัญบุรี

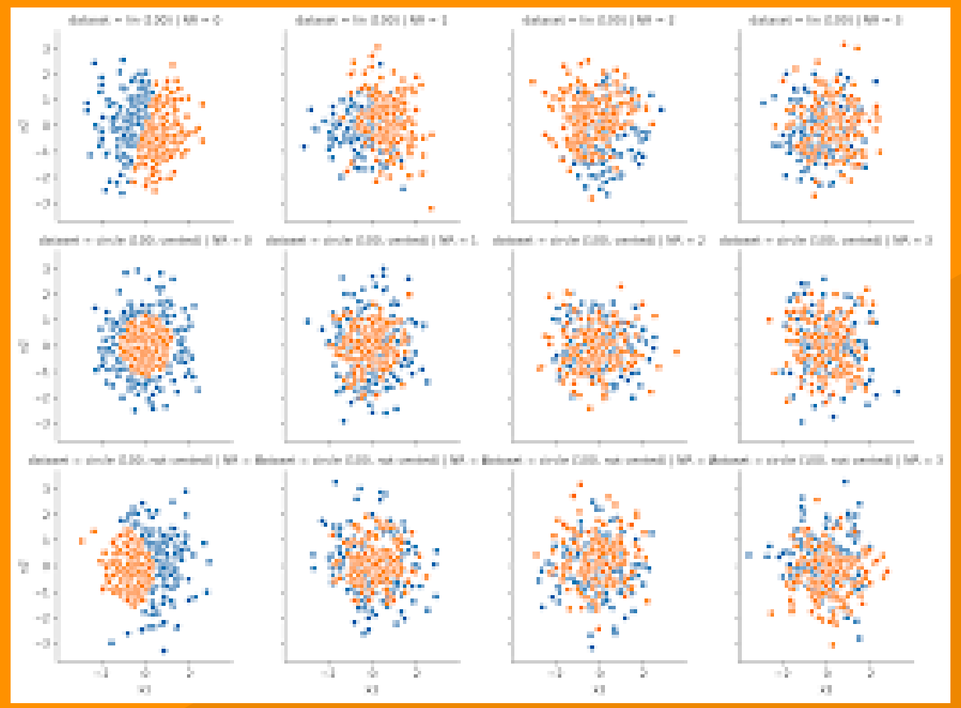
NOTE : Local outlier factor Advantages



The method used in LOF can be applied in many other fields to solve problems of detecting outliers like geographic data, video streams, etc.

RMUTT
www.rmutt.ac.th ราชมงคลธัญบุรี

NOTE : Support Vector Machine Disadvantages

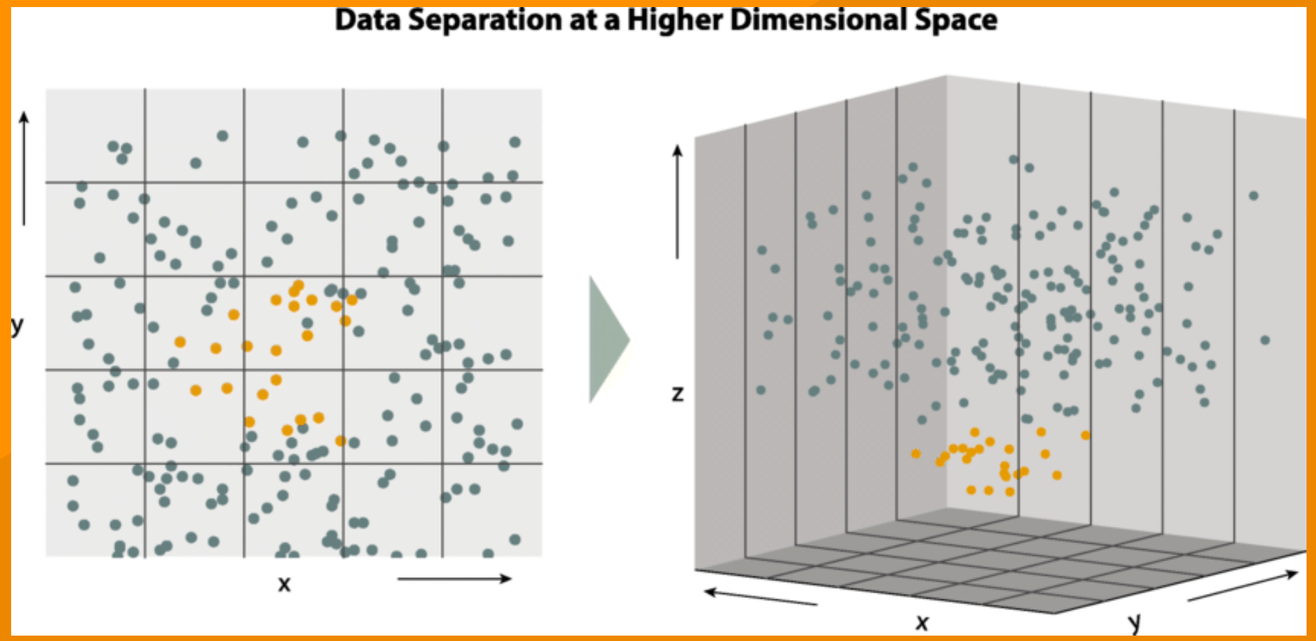


✕ FAILED

SVM algorithm is not suitable for large data sets.

RMUTT
www.rmutt.ac.th ราชมงคลธัญบุรี

NOTE : Support Vector Machine Advantages



SVM is more effective in high dimensional spaces.

RMUTT
www.rmutt.ac.th ราชมงคลธัญบุรี

COMPARE MODEL

Model	high dimensional	large data sets
LOF	Bad	**Good** GOOD JOB!
One-class SVM	**Good** GOOD JOB!	Bad

RMUTT
www.rmutt.ac.th ราชมงคลธัญบุรี

GITHUB LINK

www.rmutt.ac.th ราชมงคลธัญบุรี

THANK YOU

RMUTT
www.rmutt.ac.th ราชมงคลธัญบุรี

นาย เฉลิมวุฒิ สุดประเสริฐ
นาย ธนาพล นันทเกษตร