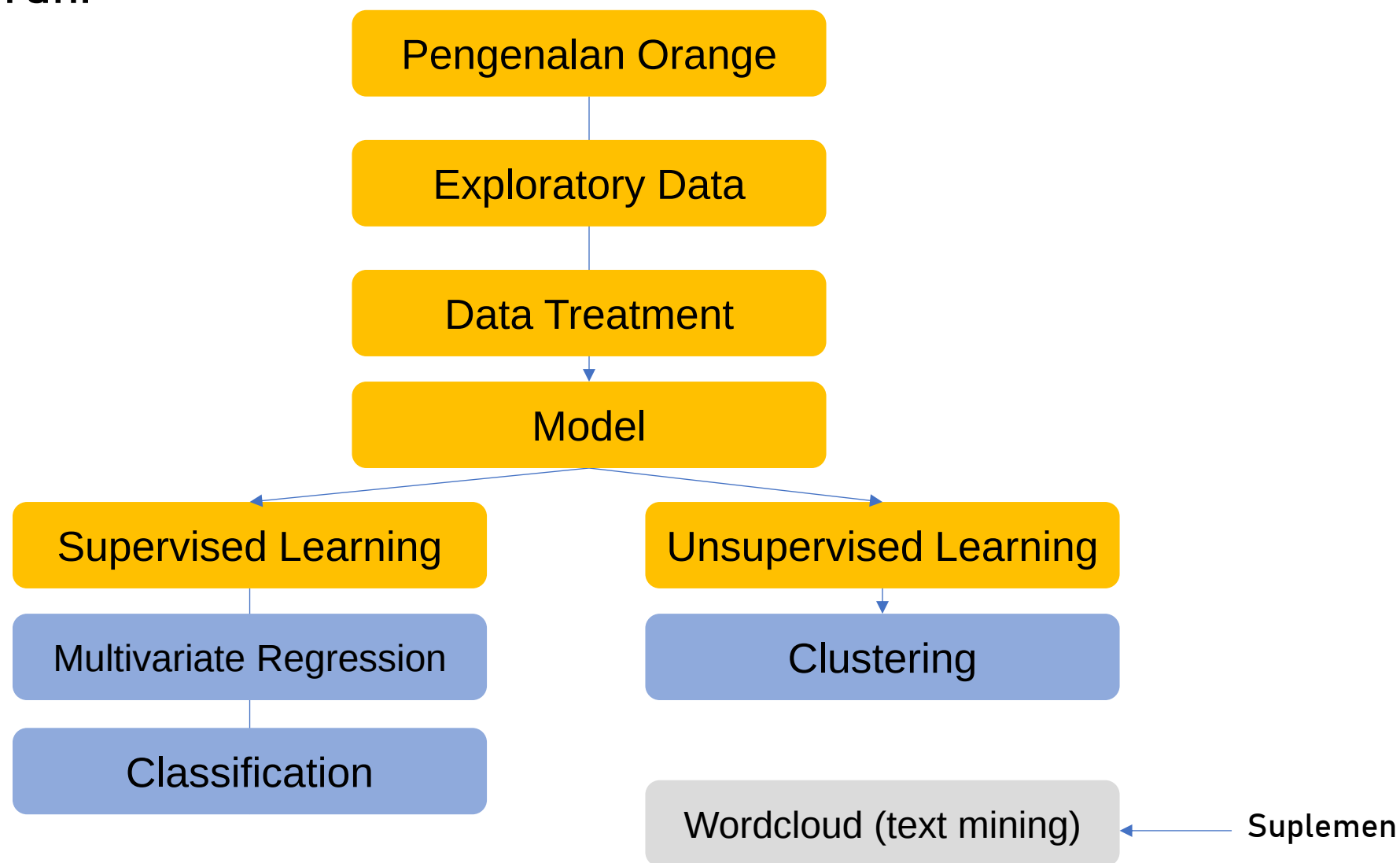

Praktik range Data Mining

Materi Pembelajaran:



A background image showing a business meeting. A person in a suit is holding a coffee cup, and another person's hand is visible on a laptop. The scene is overlaid with a large green semi-circle on the left and a blue semi-circle on the right.

Orange



Screenshots

Workflows

Download

Blog

Docs

Workshops



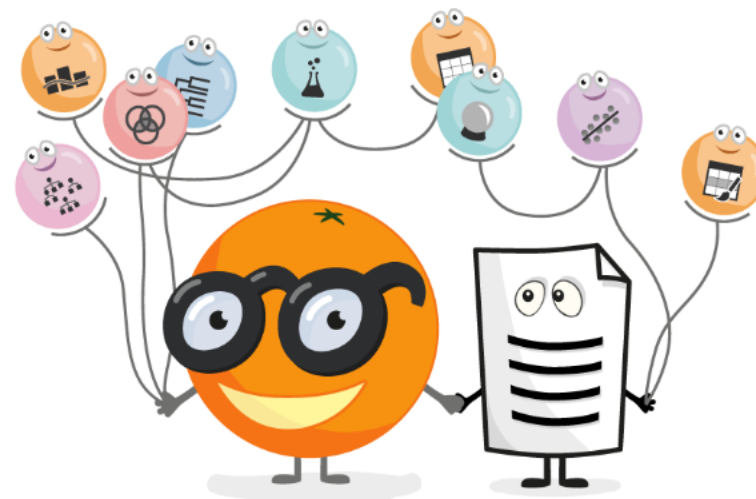
Donate

Data Mining Fruitful and Fun

Open source machine learning and data visualization.

Build data analysis workflows visually, with a large, diverse toolbox.

Download Orange



May 25, 2022

An introduction to the Kaplan-Meier Estimator

Understanding the Kaplan-Meier widget in the Survival Analysis add-on. [Read more >](#)



May 20, 2022

Confusion matrix for regression?

How to display regression error in Orange?

[Read more >](#)



May 13, 2022

Predictive Modelling with Attribute Interactions

A quick introduction to using the new Interactions widget in Orange 3 [Read more >](#)

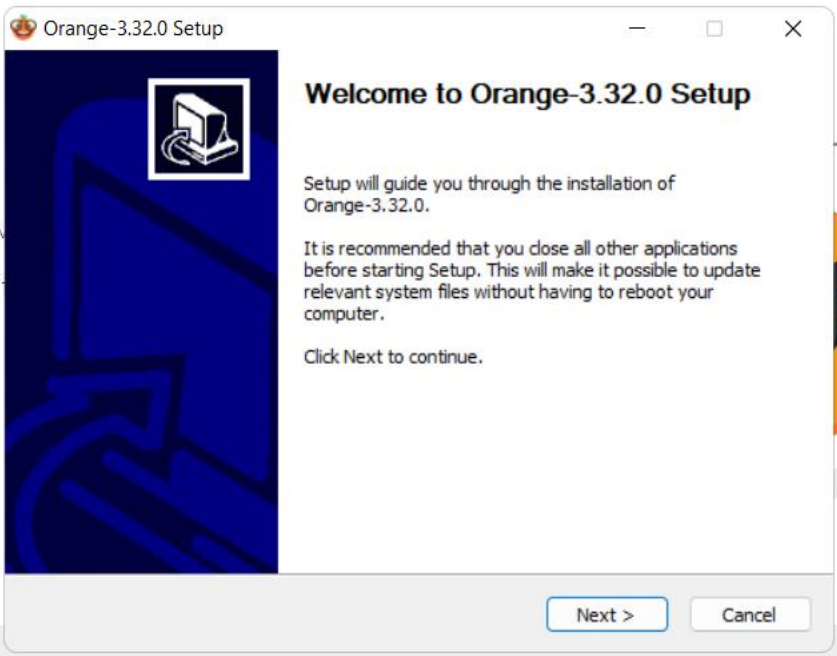




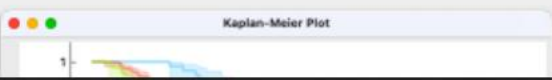
Data Mining Fruitful and Fun

Open source machine learning and data v
Build data analysis workflows visually, with

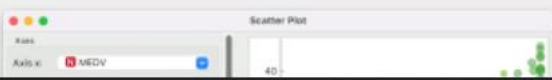
[Download Orange](#)



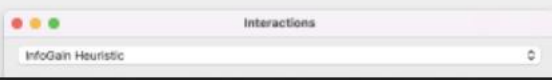
May 25, 2022
An introduction to the Kaplan-Meier Estimator
Understanding the Kaplan-Meier widget in the Survival Analysis add-on. [Read more >](#)



May 20, 2022
Confusion matrix for regression?
How to display regression error in Orange?
[Read more >](#)

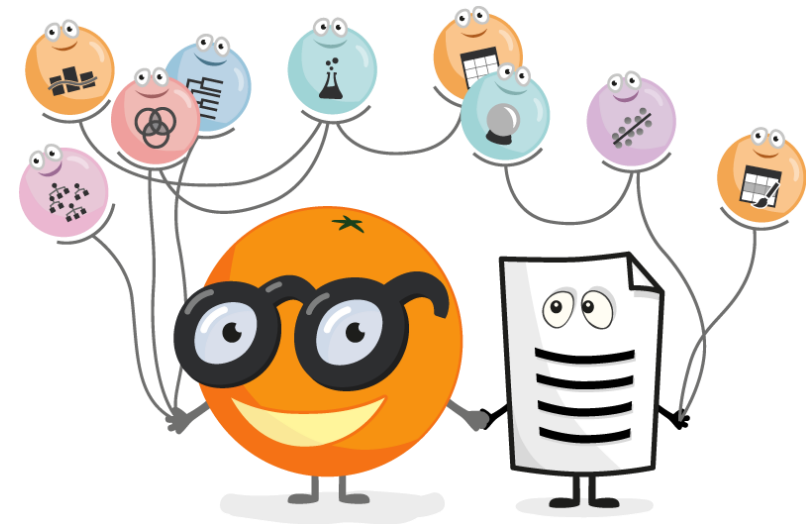


May 13, 2022
Predictive Modelling with Attribute Interactions
A quick introduction to using the new Interactions widget in Orange 3 [Read more >](#)



Kenapa Orange ??

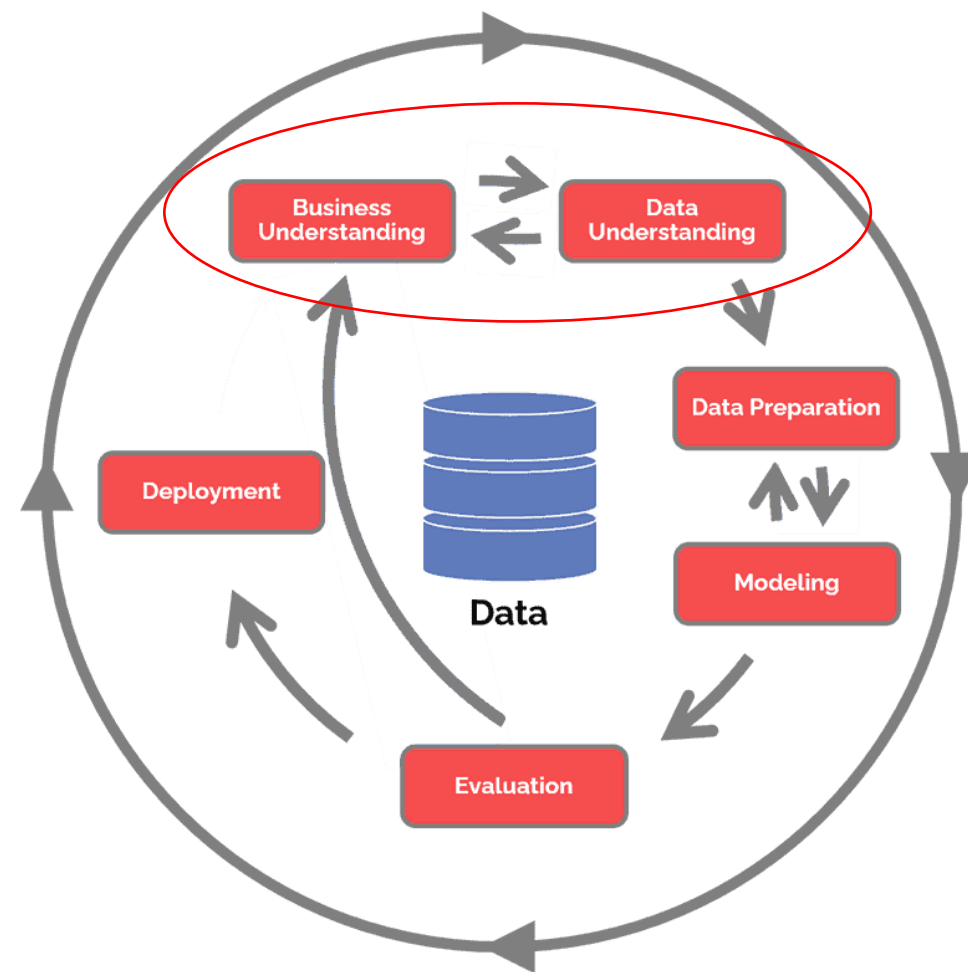
1. Algoritmanya cukup lengkap;
2. Tampilannya menarik;
3. Gratis;
4. Bisa mengolah data tanpa Batasan row;
5. Mudah digunakan (no coding coding club).



Data Mining Lifecycle

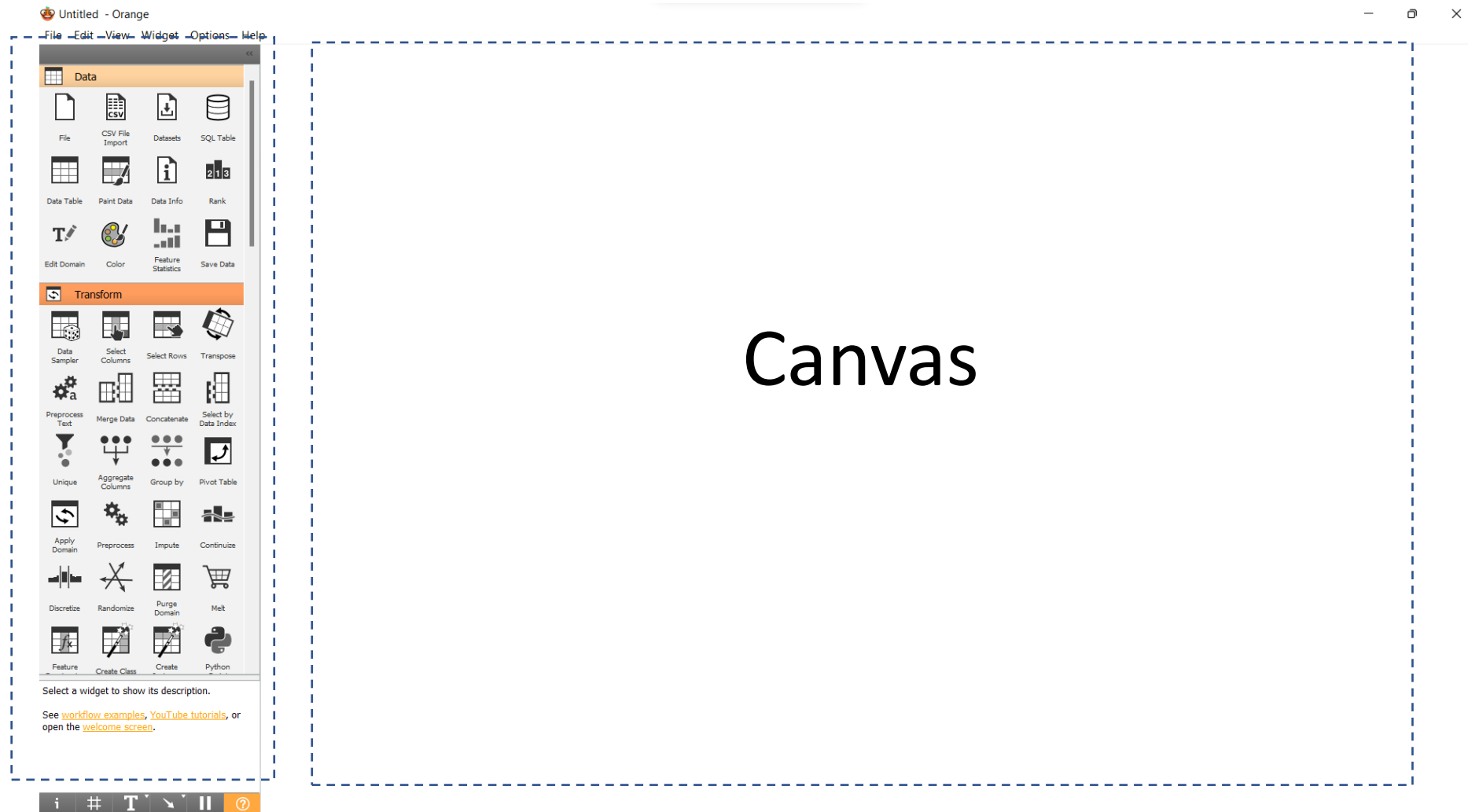
- Business Understanding
- Data Understanding
- Data Preparation
- Modeling
- Evaluation
- Deployment

Notice the iteration!

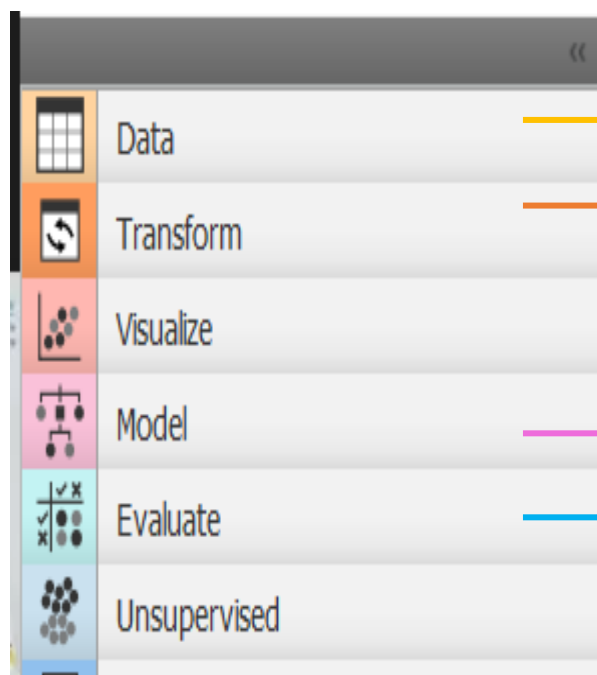


THE FAMOUS CRISP-DM

Widget



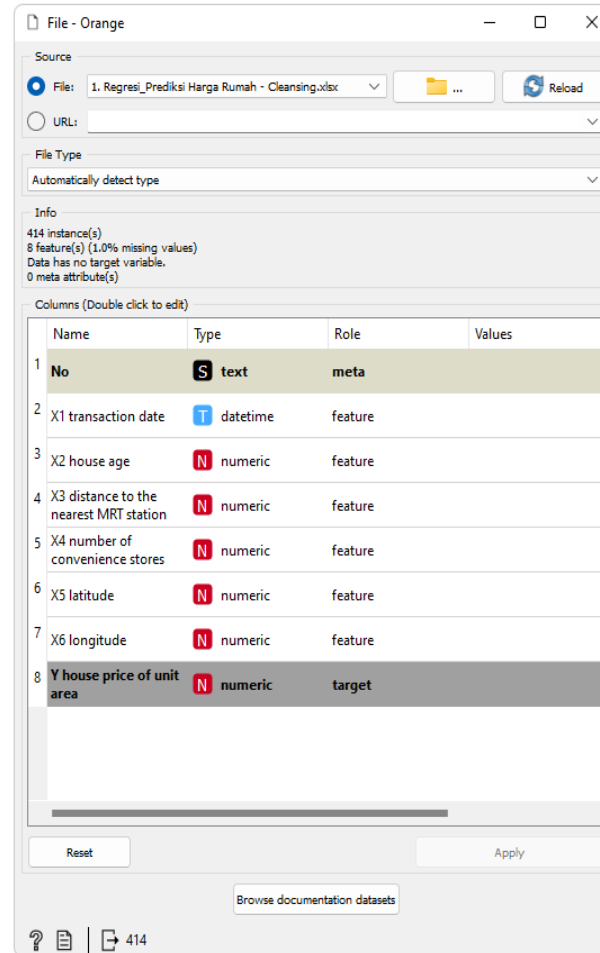
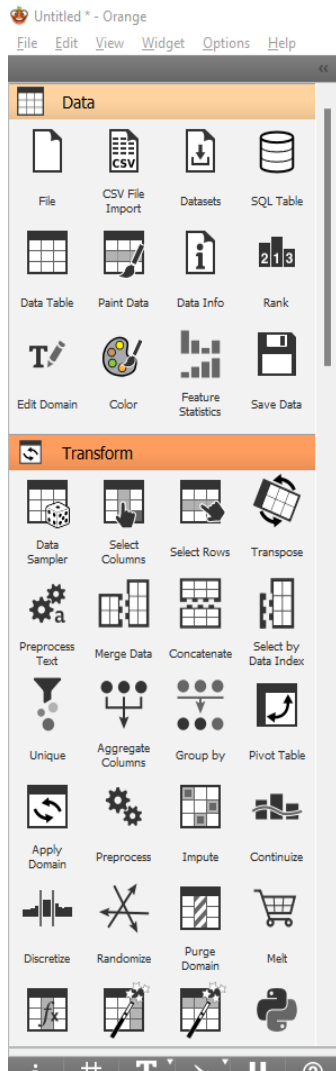
Canvas



Load — Preprocess — Model — Evaluate

Visualize here,
understanding or
explore the data

... also here



Data Type di Orange

1. Categorical

Data kategorik adalah data yang menjelaskan karakteristik dari data tersebut. Misalnya jenis kelamin, Bahasa, Kewarganegaraan, dan lain sebagainya.

2. Numeric

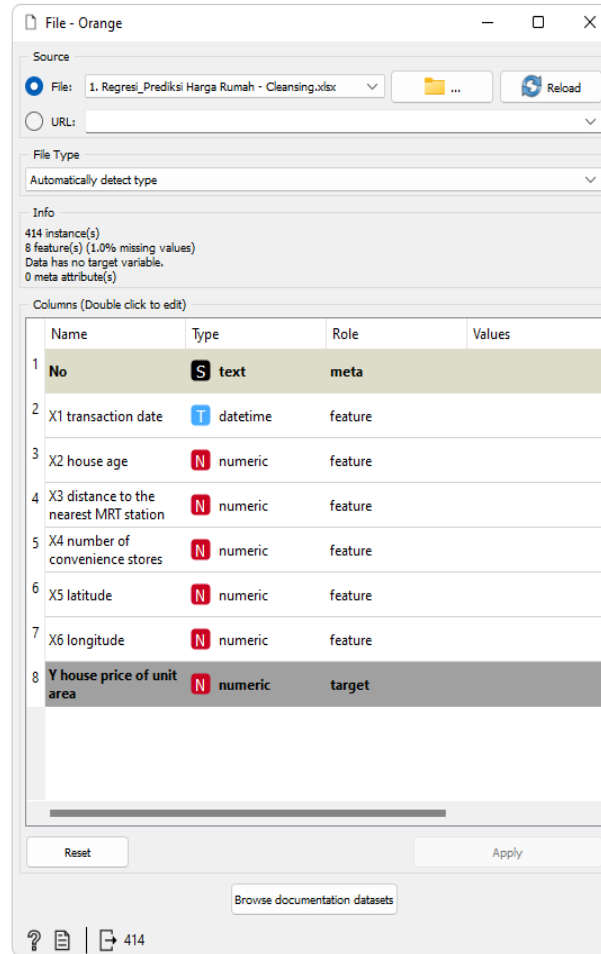
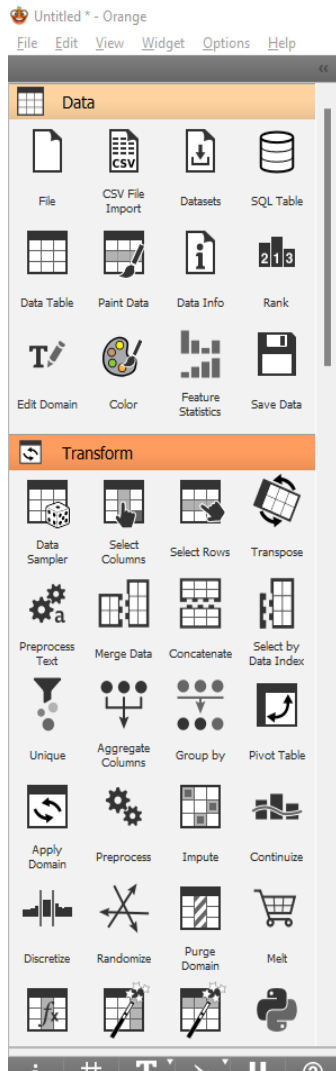
Data numerik adalah data yang menggambarkan kuantitas. Data ini dapat diukur secara kuantitatif sehingga dapat menerima operasi matematik.

3. Datetime

Tipe data tanggal yang berkaitan dengan operasi tanggal dan waktu.

4. Text

Tipe Data Teks (String) merupakan **tipe data** yang bisa kita gunakan untuk menampung banyak karakter.



Type Role di Orange

1. Feature

Variable predictor (independent) / variable X

2. Target

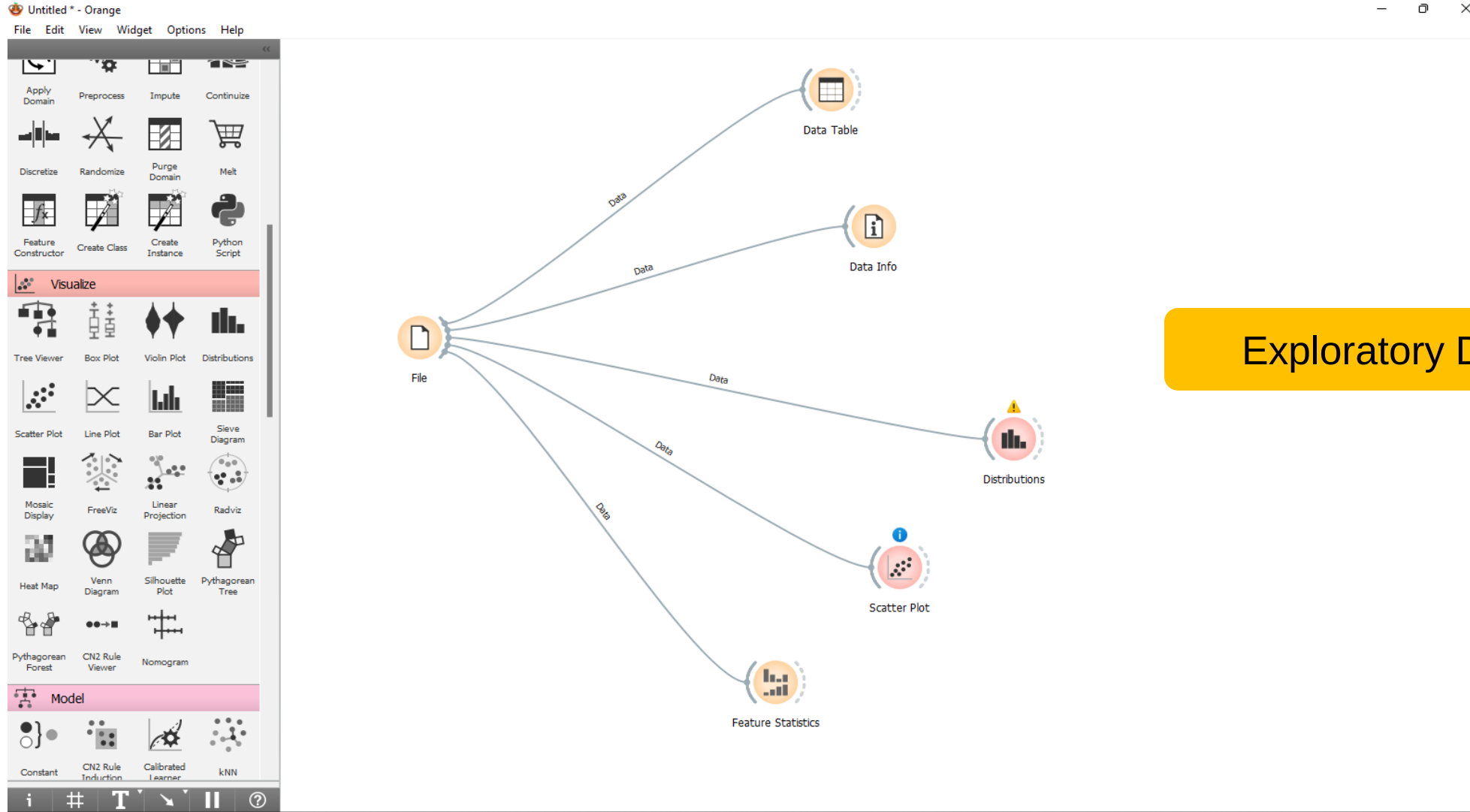
Variable dependen (terikat) / Variable Y

3. Meta

Descriptive field dalam data yang tidak akan digunakan dalam perhitungan apa pun, tetapi akan berfungsi untuk identifikasi atau informasi tambahan.

4. Skip

Field yang tidak akan digunakan dalam perhitungan dan tidak pula berfungsi sebagai informasi tambahan.





Data Preparation

Benar atau Salah

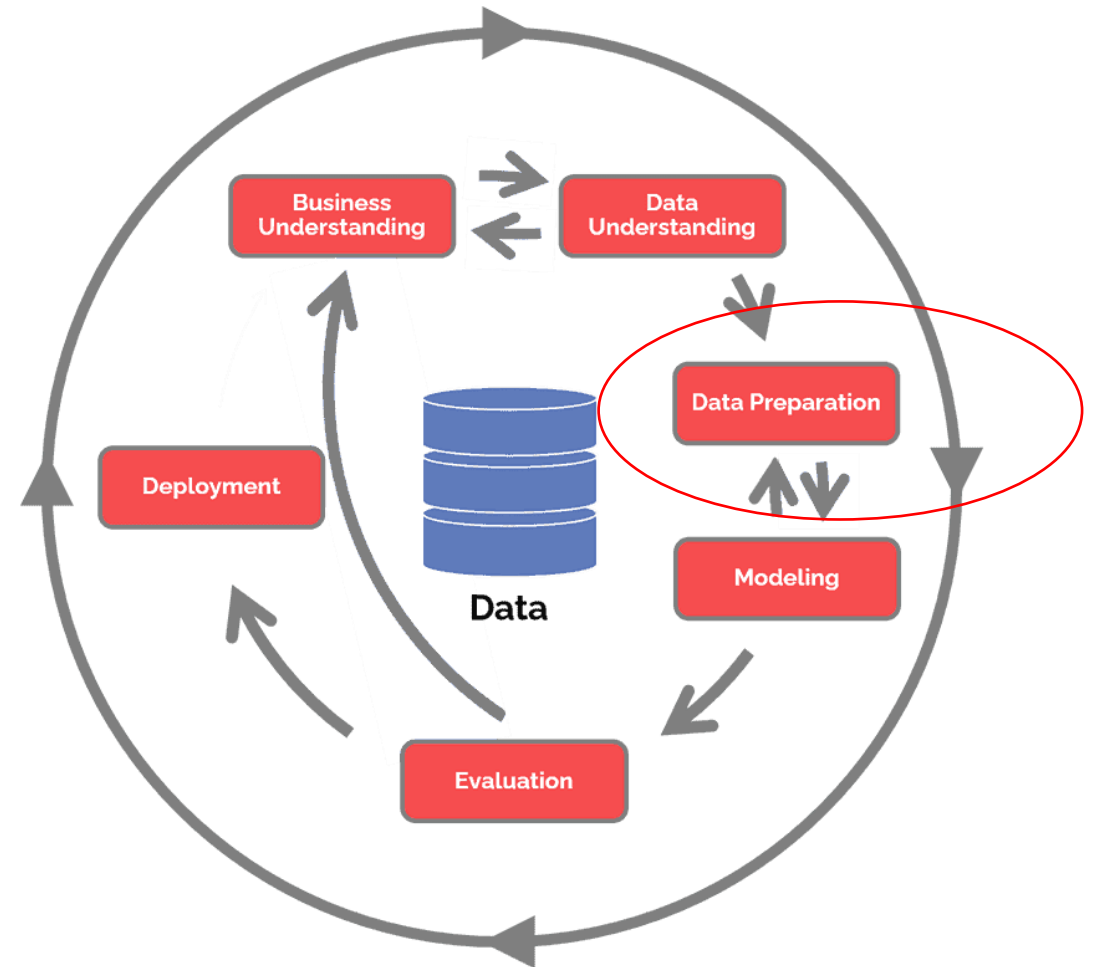
Mengisi data yang kosong (Fill in missing value)
adalah tahap pada data understanding



Data Mining Lifecycle

- Business Understanding
- Data Understanding
- Data Preparation
- Modeling
- Evaluation
- Deployment

Notice the iteration!

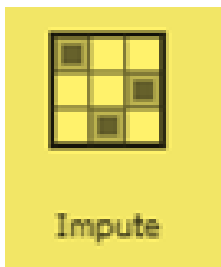


THE FAMOUS CRISP-DM

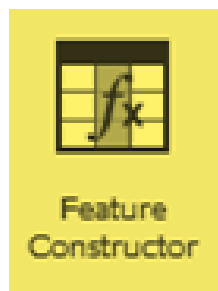


Beberapa treatment Data Preparation di Orange

Fill in Missing value



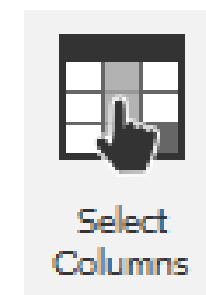
Membuat feature baru



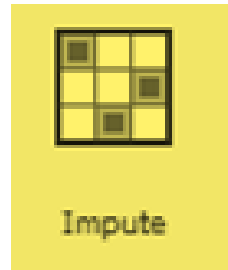
Menghilangkan outliers



Memilih Kolom Tertentu



Fill in Missing value



Data Table - Orange

Info
414 instances
6 features (1.3 % missing data)
Numeric outcome
1 meta attribute

Variables
☒ Show variable labels (if present)
☐ Visualize numeric values
☒ Color by instance classes

Selection
☒ Select full rows

	house price of unit area	No	X1 transaction date	X2 house age	ce to the nearest MRT station	number of convenience stores	X5 latitude	X6 longitude
1	379000000	1	2012-11-01 00:...	32.0	84.87882	10	24.98298	121.54024
2	422000000	2	2012-11-01 00:...	19.5	306.59470	9	?	121.53951
3	473000000	3	2013-06-01 00:...	13.3	561.98450	5	24.98746	121.54391
4	548000000	4	2013-06-01 00:...	13.3	561.98450	5	24.98746	121.54391
5	431000000	5	2012-09-01 00:...	5.0	390.56840	5	24.97937	121.54245
6	321000000	6	2012-08-01 00:...	7.1	2175.03000	3	24.96305	121.52353
7	403000000	7	2012-08-01 00:...	34.5	?	?	24.97933	121.53642
8	467000000	8	2013-05-01 00:...	?	287.60250	6	24.98042	121.54228
9	188000000	9	2013-06-01 00:...	31.7	5512.03800	1	24.95095	121.48458
10	221000000	10	2013-05-01 00:...	17.9	1783.18000	3	24.96731	121.51486
11	414000000	11	2012-12-01 00:...	34.8	405.21340	1	24.97349	121.53372
12	581000000	12	2013-03-01 00:...	6.3	90.45606	9	?	121.54310
13	393000000	13	2012-11-01 00:...	13.0	492.23130	5	24.96515	121.53737
14	238000000	14	2012-08-01 00:...	20.4	2469.64500	4	24.96108	?
15	343000000	15	2013-06-01 00:...	13.2	1164.83800	4	24.99156	121.53406
16	505000000	16	2013-06-01 00:...	?	?	2	24.98240	121.54619
17	701000000	17	2013-03-01 00:...	0.0	292.99780	6	24.97744	121.54458
18	374000000	18	2012-09-01 00:...	17.7	350.85150	1	24.97544	121.53119
19	423000000	19	2013-05-01 00:...	16.9	368.13630	8	24.96750	121.54451
20	477000000	20	2012-08-01 00:...	1.5	23.38284	7	24.96772	121.54102
21	293000000	21	2013-05-01 00:...	4.5	2275.87700	3	24.96314	121.51151
22	516000000	22	2013-05-01 00:...	10.5	279.17260	7	24.97528	121.54541
23	246000000	23	2012-11-01 00:...	?	1360.13900	1	24.95204	?
24	479000000	24	2012-12-01 00:...	10.1	279.17260	7	24.97528	121.54541
25	388000000	25	2012-12-01 00:...	39.6	480.69770	4	?	121.53885
26	270000000	26	2012-12-01 00:...	29.3	1487.86800	2	24.97542	121.51726
27	562000000	27	2012-08-01 00:...	3.1	?	5	24.98085	121.54391
28	336000000	28	2013-03-01 00:...	10.4	276.44900	5	24.95593	121.53913
29	470000000	29	2013-06-01 00:...	19.2	557.47800	4	24.97419	121.53797
30	571000000	30	2012-12-01 00:...	7.1	451.24380	5	24.97563	121.54694
31	221000000	31	2013-06-01 00:...	25.9	5143.6149008	0	24.94826	121.49587
32	250000000	32	2012-09-01 00:...	29.6	769.40340	7	24.98281	121.53408
33	342000000	33	2012-09-01 00:...	37.9	488.57270	1	24.97349	121.53451
34	493000000	34	2013-03-01 00:...	16.5	323.65500	6	24.97841	121.5407821
35	551000000	35	2012-09-01 00:...	15.4	205.36700	7	24.98419	121.54243
36	273000000	36	2013-06-01 00:...	13.9	4079.41800	0	25.01459	121.51816
37	229000000	37	2012-11-01 00:...	14.7	1935.00900	2	24.96386	121.51458
38	253000000	38	2013-02-01 00:...	12.0	1360.13900	1	24.95204	121.54842
39	477000000	39	2012-08-01 00:...	3.1	577.96150	6	24.97201	121.54722
40	462000000	40	2013-02-01 00:...	16.2	289.32480	5	24.98203	121.54348
41	159000000	41	2012-12-01 00:...	13.6	4082.01500	0	24.94155	121.50381
42	182000000	42	2013-06-01 00:...	16.8	4066.58700	0	24.94297	121.5299601
43	347000000	43	2013-05-01 00:...	36.1	519.46170	5	24.96305	121.53758
44	341000000	44	2012-09-01 00:...	34.4	512.78710	6	24.98748	121.54301

Data Table (1) - Orange

Info
414 instances (no missing data)
6 features
Numeric outcome
1 meta attribute

Variables
☒ Show variable labels (if present)
☐ Visualize numeric values
☒ Color by instance classes

Selection
☒ Select full rows

	house price of unit area	No	1 transaction date	X2 house age	ce to the nearest MRT station	number of convenience stores	X5 latitude	X6 longitude
1	379000000	1	2012-11-01 00:...	32.0	84.87882	10	24.98298	121.54024
2	422000000	2	2012-11-01 00:...	19.5	306.59470	9	24.9769993	121.53951
3	473000000	3	2013-06-01 00:...	13.3	561.98450	5	24.98746	121.54391
4	548000000	4	2013-06-01 00:...	13.3	561.98450	5	24.98746	121.54391
5	431000000	5	2012-09-01 00:...	5.0	390.56840	5	24.97937	121.54245
6	321000000	6	2012-08-01 00:...	7.1	2175.03000	3	24.96305	121.5123535
7	403000000	7	2012-08-01 00:...	34.5	622.6123803	6.77	24.97933	121.53642
8	467000000	8	2013-05-01 00:...	29.750	287.60250	6	24.98042	121.54228
9	188000000	9	2013-06-01 00:...	31.7	5512.03800	1	24.95095	121.48458
10	221000000	10	2013-05-01 00:...	17.9	1783.18000	3	24.96731	121.51486
11	414000000	11	2012-12-01 00:...	34.8	405.21340	1	24.97349	121.53372
12	581000000	12	2013-03-01 00:...	6.3	90.45606	9	24.9754245	121.54310
13	393000000	13	2012-11-01 00:...	13.0	492.23130	5	24.96515	121.53737
14	238000000	14	2012-08-01 00:...	20.4	2469.64500	4	24.96108	121.5123535
15	343000000	15	2013-06-01 00:...	13.2	1164.83800	4	24.99156	121.53406
16	505000000	16	2013-06-01 00:...	16.583	860.6829398	2	24.98240	121.54619
17	701000000	17	2013-03-01 00:...	0.0	292.99780	6	24.97744	121.54458
18	374000000	18	2012-09-01 00:...	17.7	350.85150	1	24.97544	121.53119
19	423000000	19	2013-05-01 00:...	16.9	368.13630	8	24.96750	121.54451
20	477000000	20	2012-08-01 00:...	1.5	23.38284	7	24.96772	121.54102
21	293000000	21	2013-05-01 00:...	4.5	2275.87700	3	24.96314	121.51151
22	516000000	22	2013-05-01 00:...	10.5	279.17260	7	24.97528	121.54541
23	246000000	23	2012-11-01 00:...	13.175	1360.13900	1	24.95204	121.5397757
24	479000000	24	2012-12-01 00:...	10.1	279.17260	7	24.97528	121.54541
25	388000000	25	2012-12-01 00:...	39.6	480.69770	4	24.9723464	121.53885
26	270000000	26	2012-12-01 00:...	29.3	1487.86800	2	24.97542	121.51726
27	562000000	27	2012-08-01 00:...	3.1	474.5776377	5	24.98085	121.54391
28	336000000	28	2013-03-01 00:...	10.4	276.44900	5	24.95593	121.53913
29	470000000	29	2013-06-01 00:...	19.2	557.47800	4	24.97419	121.53797
30	571000000	30	2012-12-01 00:...	7.1	451.24380	5	24.97563	121.54694
31	221000000	31	2013-06-01 00:...	25.9	5143.6149008	0	24.94826	121.49587
32	250000000	32	2012-09-01 00:...	29.6	769.40340	7	24.98281	121.53408
33	342000000	33	2012-09-01 00:...	37.9	488.57270	1	24.97349	121.53451
34	493000000	34	2013-03-01 00:...	16.5	323.65500	6	24.97841	121.5407821
35	551000000	35	2012-09-01 00:...	15.4	205.36700	7	24.98419	121.54243
36	273000000	36	2013-06-01 00:...	13.9	4079.41800	0	25.01459	121.51816
37	229000000	37	2012-11-01 00:...	14.7	1935.00900	2	24.96386	121.51458
38	253000000	38	2013-02-01 00:...	12.0	1360.13900	1	24.95204	121.54842
39	477000000	39	2012-08-01 00:...	3.1	577.96150	6	24.97201	121.54722
40	462000000	40	2013-02-01 00:...	16.2	289.32480	5	24.98203	121.54348
41	159000000	41	2012-12-01 00:...	13.6	4082.01500	0	24.94155	121.50381
42	182000000	42	2013-06-01 00:...	16.8	4066.58700	0	24.94297	121.5299601
43	347000000	43	2013-05-01 00:...	36.1	519.46170	5	24.96305	121.53758
44	341000000	44	2012-09-01 00:...	34.4	512.78710	6	24.98748	121.54301

Impute - Orange

Default Method
☐ Don't impute
☐ Average/Most frequent
☐ As a distinct value
☐ Fixed values; numeric variables: 0, time: 1970-01-01 07:00:00

Model-based imputer (simple tree)
☐ Random values
☐ Remove instances with unknown values

Individual Attribute Settings
 Filter...
☒ X1 transaction date
☒ X2 house age
☒ X3 distance to the nearest MRT station
☒ X4 number of convenience stores
☒ X5 latitude
☒ X6 longitude
☒ Y house price of unit area

☐ Default (above)
☐ Don't impute
☐ Average/Most frequent
☐ As a distinct value
☐ Model-based imputer (simple tree)
☐ Random values
☐ Remove instances with unknown values
☐ Fixed value

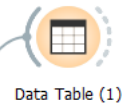
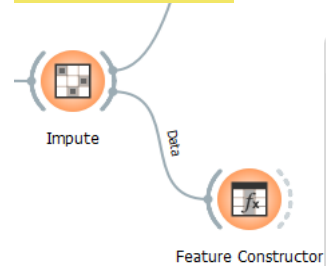
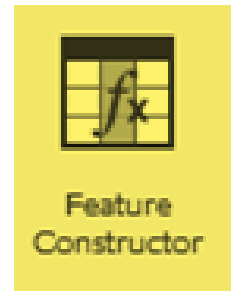
☒ Apply Automatically

414

Restore Original Order

Apply Automatically

Membuat feature baru



Data Table (1)

Feature Constructor - Orange

Variable Definitions

New

Remove

N X3_sqrt := sqrt(X3_distance_to_the_nearest_MRT_station)

N X3_log := log(X3_distance_to_the_nearest_MRT_station)

Send

414 414

Data Table (2) - Orange

Info

g data)

(if present)

lues

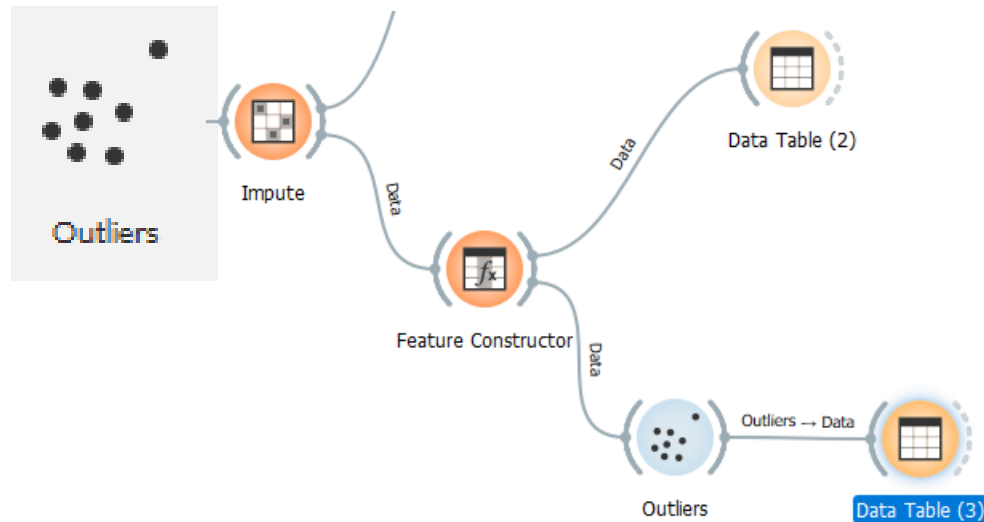
sses

	house price of unit ar	No	X1 transaction date	X2 house age	ce to the nearest MF	nber of convenience	X5 latitude	X6 longitude	X3_sqrt	X3_log
1	379000000	1	2012-11-01 00:...	32.0	84.87882	10	24.98298	121.54024	9.21297	4.44122
2	422000000	2	2012-11-01 00:...	19.5	306.59470	9	24.9769993	121.53951	17.5098	5.72553
3	473000000	3	2013-06-01 00:...	13.3	561.98450	5	24.98746	121.54391	23.7062	6.33147
4	548000000	4	2013-06-01 00:...	13.3	561.98450	5	24.98746	121.54391	23.7062	6.33147
5	431000000	5	2012-09-01 00:...	5.0	390.56840	5	24.97937	121.54245	19.7628	5.9676
6	321000000	6	2012-08-01 00:...	7.1	2175.03000	3	24.96305	121.5123535	46.6372	7.6848
7	403000000	7	2012-08-01 00:...	34.5	622.6123803	6.77	24.97933	121.53642	24.9522	6.43392
8	467000000	8	2013-05-01 00:...	29.750	287.60250	6	24.98042	121.54228	16.9588	5.66158
9	188000000	9	2013-06-01 00:...	31.7	5512.03800	1	24.95095	121.48458	74.2431	8.61469
10	221000000	10	2013-05-01 00:...	17.9	1783.18000	3	24.96731	121.51486	42.2277	7.48615
11	414000000	11	2012-12-01 00:...	34.8	405.21340	1	24.97349	121.53372	20.1299	6.00441
12	581000000	12	2013-03-01 00:...	6.3	90.45606	9	24.9754245	121.54310	9.51084	4.50486
13	393000000	13	2012-11-01 00:...	13.0	492.23130	5	24.96515	121.53737	22.1863	6.19895
14	238000000	14	2012-08-01 00:...	20.4	2469.64500	4	24.96108	121.5123535	49.6955	7.81183
15	343000000	15	2013-06-01 00:...	13.2	1164.83800	4	24.99156	121.53406	34.1297	7.06034
16	505000000	16	2013-06-01 00:...	16.583	860.6829398	2	24.98240	121.54619	29.3374	6.75773
17	701000000	17	2013-03-01 00:...	0.0	292.99780	6	24.97744	121.54458	17.1172	5.68017
18	374000000	18	2012-09-01 00:...	17.7	350.85150	1	24.97544	121.53119	18.731	5.86036
19	423000000	19	2013-05-01 00:...	16.9	368.13630	8	24.96750	121.54451	19.1869	5.90845
20	477000000	20	2012-08-01 00:...	1.5	23.38284	7	24.96772	121.54102	4.83558	3.152
21	293000000	21	2013-05-01 00:...	4.5	2275.87700	3	24.96314	121.51151	47.7062	7.73012
22	516000000	22	2013-05-01 00:...	10.5	279.17260	7	24.97528	121.54541	16.7085	5.63183
23	246000000	23	2012-11-01 00:...	13.175	1360.13900	1	24.95204	121.5397757	36.8801	7.21534
24	479000000	24	2012-12-01 00:...	10.1	279.17260	7	24.97528	121.54541	16.7085	5.63183
25	388000000	25	2012-12-01 00:...	39.6	480.69770	4	24.9723464	121.53885	21.9248	6.17524
26	270000000	26	2012-12-01 00:...	29.3	1487.86800	2	24.97542	121.51726	38.5729	7.3051
27	562000000	27	2012-08-01 00:...	3.1	474.5776377	5	24.98085	121.54391	21.7848	6.16243
28	336000000	28	2013-03-01 00:...	10.4	276.44900	5	24.95593	121.53913	16.6268	5.62203
29	470000000	29	2013-06-01 00:...	19.2	557.47800	4	24.97419	121.53797	23.611	6.32342
30	571000000	30	2012-12-01 00:...	7.1	451.24380	5	24.97563	121.54694	21.2425	6.11201
31	221000000	31	2013-06-01 00:...	25.9	5143.6149008	0	24.94826	121.49587	71.719	8.54551
32	250000000	32	2012-09-01 00:...	29.6	769.40340	7	24.98281	121.53408	27.7381	6.64562
33	342000000	33	2012-09-01 00:...	37.9	488.57270	1	24.97349	121.53451	22.1037	6.19149
34	493000000	34	2013-03-01 00:...	16.5	323.65500	6	24.97841	121.5407821	17.9904	5.77968
35	551000000	35	2012-09-01 00:...	15.4	205.36700	7	24.98419	121.54243	14.3306	5.3248
36	273000000	36	2013-06-01 00:...	13.9	4079.41800	0	25.01459	121.51816	63.8703	8.31371
37	229000000	37	2012-11-01 00:...	14.7	1935.00900	2	24.96386	121.51458	43.9887	7.56787
38	253000000	38	2013-02-01 00:...	12.0	1360.13900	1	24.95204	121.54842	36.8801	7.21534
39	477000000	39	2012-08-01 00:...	3.1	577.96150	6	24.97201	121.54722	24.0408	6.35951
40	462000000	40	2013-02-01 00:...	16.2	289.32480	5	24.98203	121.54348	17.0096	5.66755
41	159000000	41	2012-12-01 00:...	13.6	4082.01500	0	24.94155	121.50381	63.8906	8.31435
42	182000000	42	2013-06-01 00:...	16.8	4066.58700	0	24.94297	121.5299601	63.7698	8.31056
43	347000000	43	2013-05-01 00:...	36.1	519.46170	5	24.96305	121.53758	22.7917	6.25279
44	341000000	44	2012-09-01 00:...	34.4	512.78710	6	24.98748	121.54301	22.6448	6.23986
45	530000000	45	2013-06-01 00:...	2.7	622.47320	4	24.97445	121.54355	22.6274	6.33744

Restore Original Order

☒ Send Automatically

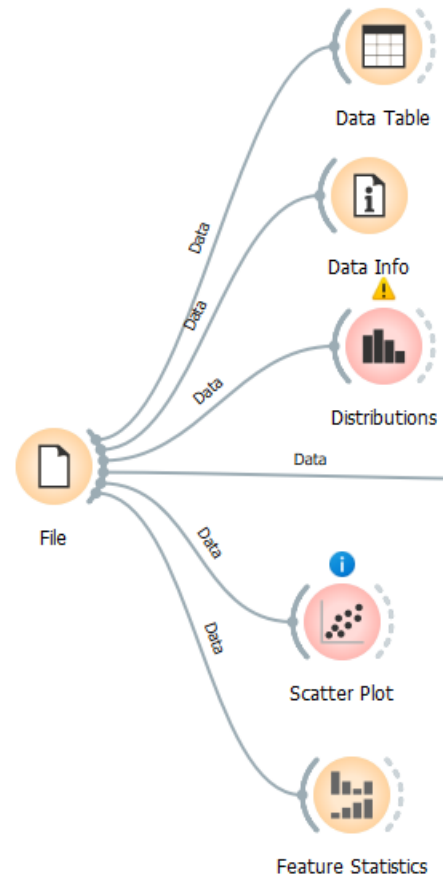
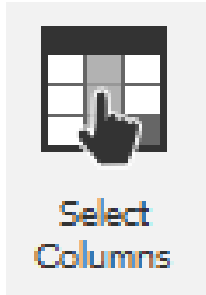
Menghilangkan outliers



Data Table (3) - Orange

	use price of unit	No	1 transaction dat	X2 house age	to the nearest	er of convenienc	X5 latitude	X6 longitude	X3_sqrt	X3_log
1	321000000	6	2012-08-01 00:...	7.1	2175.03000	3	24.96305	121.5123535	46.6372	7.6848
2	238000000	14	2012-08-01 00:...	20.4	2469.64500	4	24.96108	121.5123535	49.6955	7.81183
3	132000000	50	2012-08-01 00:...	29.4	4510.35900	1	24.94925	121.49542	67.1592	8.41413
4	122000000	117	2012-12-01 00:...	30.9	6396.28300	1	24.94375	121.47883	79.9768	8.76347
5	356000000	158	2013-03-01 00:...	16.1	815.93140	4	24.97886	121.53464	28.5645	6.70433
6	144000000	171	2013-03-01 00:...	24.0	4527.68700	0	24.94741	121.49628	67.2881	8.41797
7	155000000	181	2012-08-01 00:...	26.9	4449.27000	0	24.94898	121.49621	66.7028	8.4005
8	205000000	190	2012-11-01 00:...	16.3	4066.58700	0	24.94297	121.50342	63.7698	8.31056
9	174000000	233	2012-11-01 00:...	27.1	4412.76500	1	24.95032	121.49587	66.4286	8.39226
10	150000000	250	2012-09-01 00:...	18.0	6306.15300	1	24.95743	121.47516	79.4113	8.74928
11	174000000	256	2013-05-01 00:...	31.5	5512.03800	1	24.95095	121.48458	74.2431	8.61469
12	167000000	299	2013-03-01 00:...	16.7	4082.01500	0	24.94155	121.50381	63.8906	8.31435
13	156000000	332	2013-03-01 00:...	25.6	4519.69000	0	24.94826	121.49587	67.2286	8.4162
14	228000000	370	2012-08-01 00:...	20.2	2185.12800	3	24.96322	121.51237	46.7454	7.68943

Memilih Kolom Tertentu



Select Columns - Orange

Ignored

Filter

- X3 distance to the nearest MRT st...
- X3_sqrt

Features

Filter

- X4 number of convenience stores
- X5 latitude
- X6 longitude
- X2 house age
- X1 transaction date
- X3_log

Target

Y house price of unit area

Metas

No

Reset ☐ Ignore new variables by default ☒ Send Automatically

? | 400 | - | 400 | 6



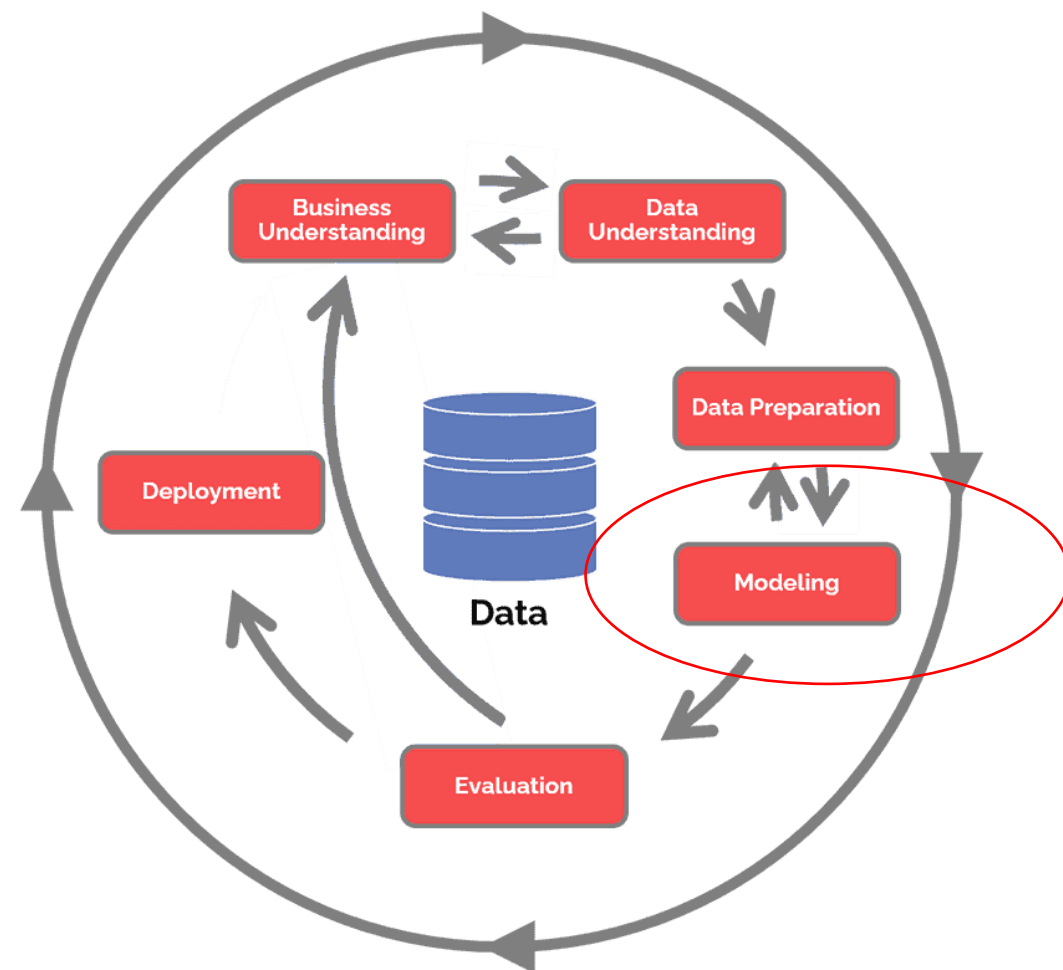
A background image showing a business meeting. On the left, a person's hand is visible, typing on a laptop. On the right, another person's hand is holding a white coffee cup with a black lid. They are sitting at a table with various documents, including one with a bar chart and another with a pie chart. A black pen lies on the table. A large yellow circle is overlaid on the left side of the image, and a blue circle is overlaid on the right side, creating a split-color effect.

Model

Data Mining Lifecycle

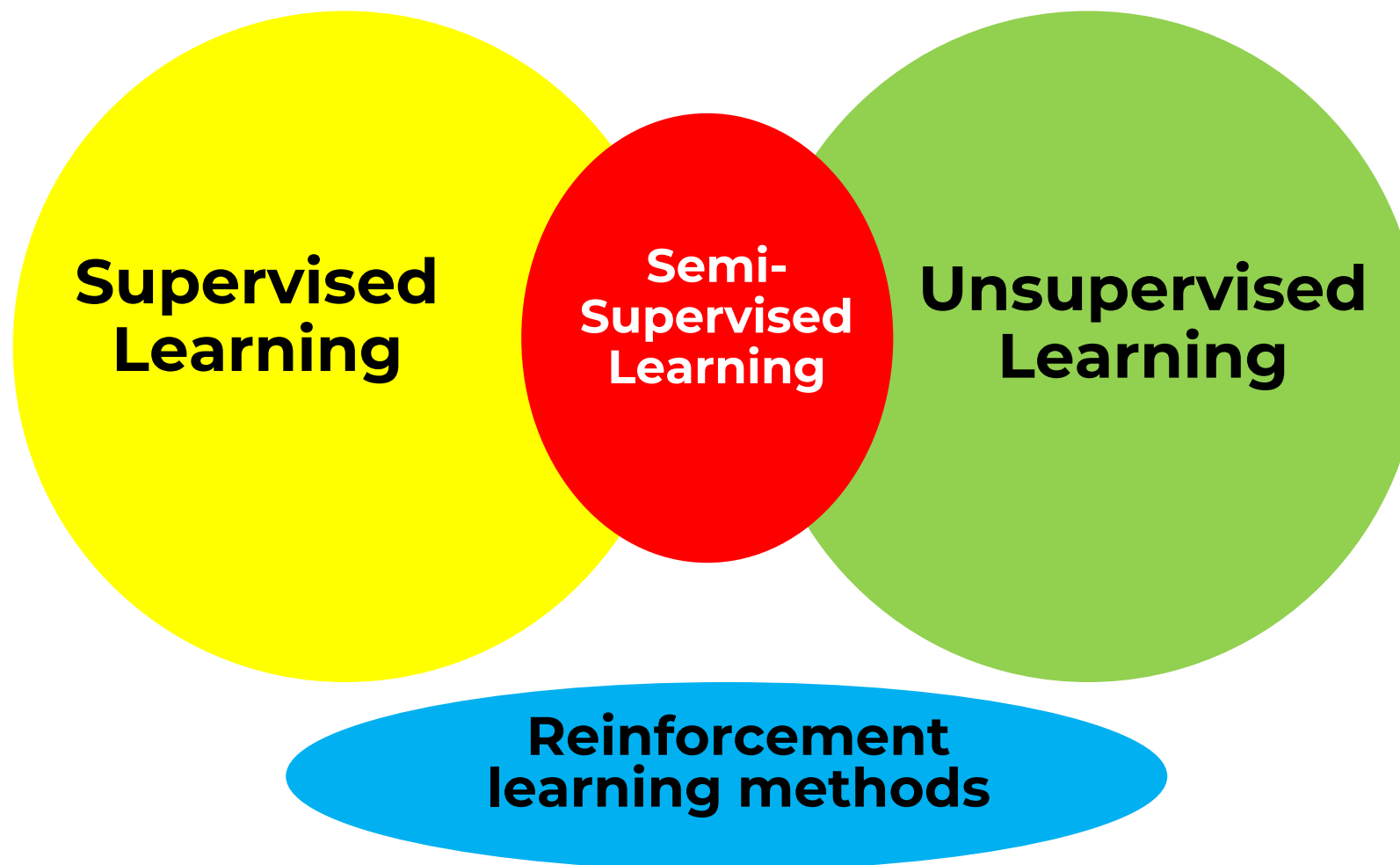
- Business Understanding
- Data Understanding
- Data Preparation
- **Modeling**
- Evaluation
- Deployment

Notice the iteration!



THE FAMOUS CRISP-DM

Kategorisasi Algoritma Data Mining (Learning Methods)

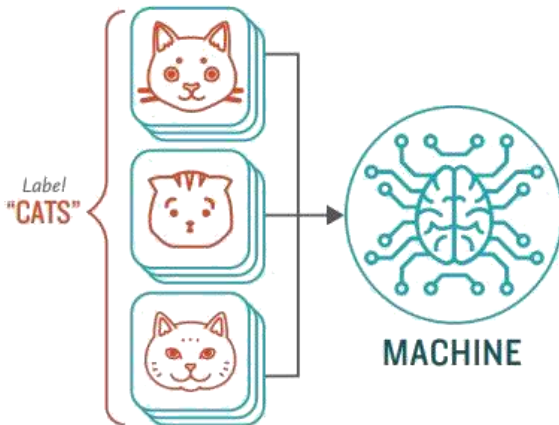


Modeling – Supervised Learning

How Supervised Machine Learning Works

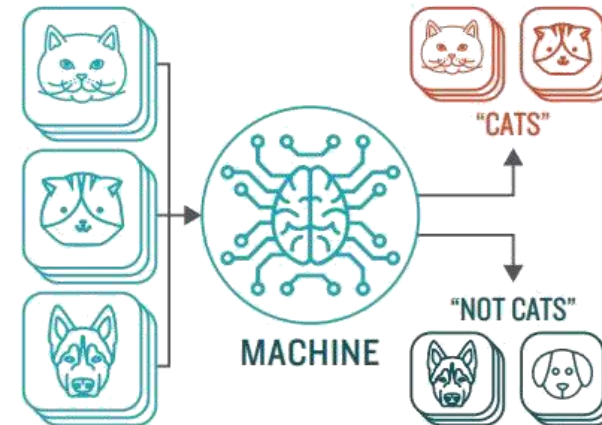
STEP 1

Provide the machine learning algorithm categorized or "labeled" input and output data from to learn

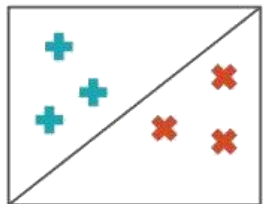


STEP 2

Feed the machine new, unlabeled information to see if it tags new data appropriately. If not, continue refining the algorithm

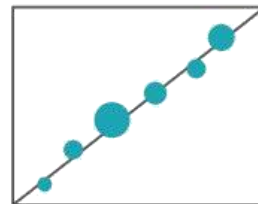


TYPES OF PROBLEMS TO WHICH IT'S SUITED



CLASSIFICATION

Sorting items into categories



REGRESSION

Identifying real values (dollars, weight, etc.)

- Pembelajaran dengan **guru**, data set **memiliki target/label/class**
- Algoritma melakukan proses **belajar berdasarkan nilai dari variabel target**
- Contoh: Algoritma estimation, prediction/forecasting, classification

Modeling – Unsupervised Learning

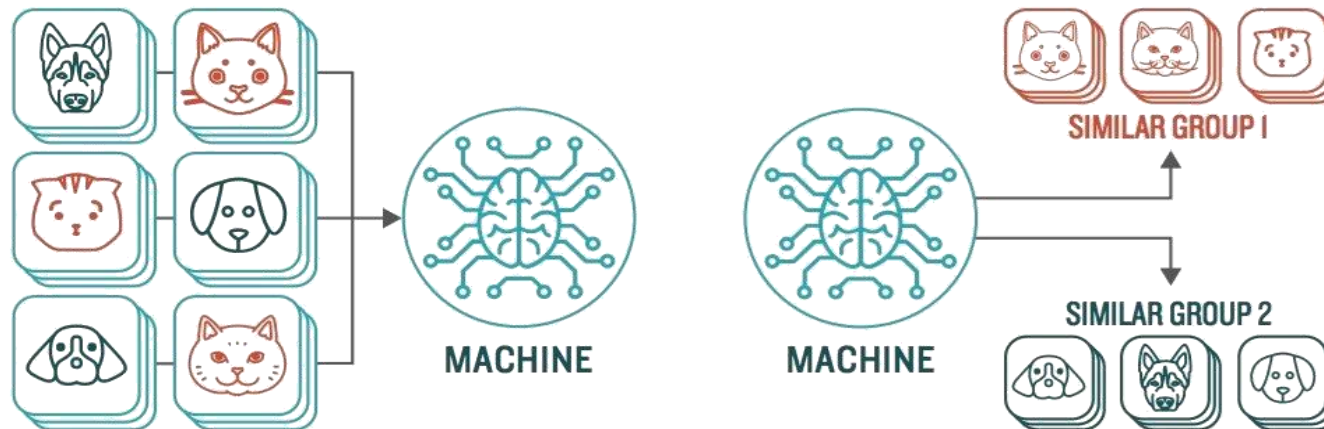
How **Unsupervised** Machine Learning Works

STEP 1

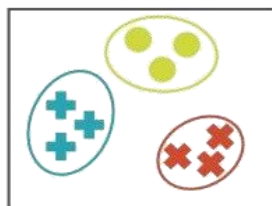
Provide the machine learning algorithm uncategorized, unlabeled input data to see what patterns it finds

STEP 2

Observe and learn from the patterns the machine identifies



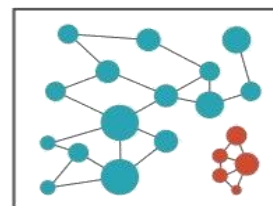
TYPES OF PROBLEMS TO WHICH IT'S SUITED



CLUSTERING

Identifying similarities in groups

For Example: Are there patterns in the data to indicate certain patients will respond better to this treatment than others?



ANOMALY DETECTION

Identifying abnormalities in data

For Example: Is a hacker intruding in our network?

- Variable (atribut) yang menjadi **target/label/class tidak ditentukan (tidak ada)**
- Algoritma data mining **mencari pola** dari semua variable (atribut)
- Contoh: Algoritma **clustering dan asosiasi**

Modeling (CRISP-DM)

1. Estimation

Linear Regression (LR), Neural Network (NN), Deep Learning (DL), Support Vector Machine (SVM), Generalized Linear Model (GLM), etc

2. Forecasting

Linear Regression (LR), Neural Network (NN), Deep Learning (DL), Support Vector Machine (SVM), Generalized Linear Model (GLM), etc

3. Classification

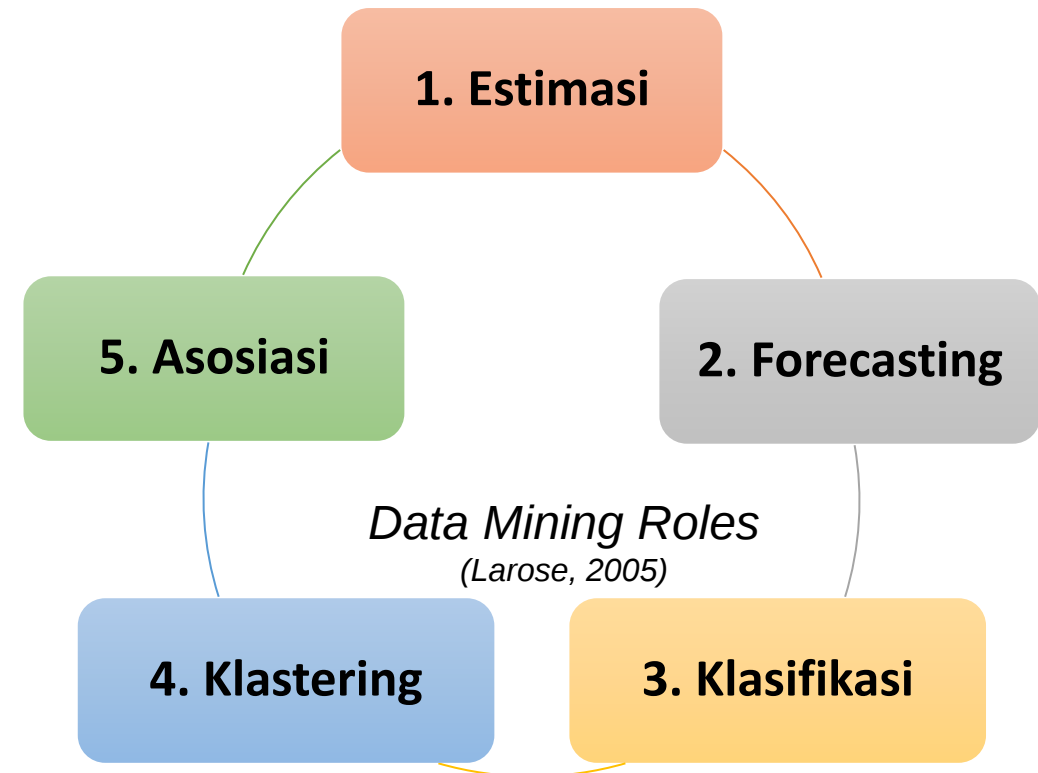
Decision Tree (CART, ID3, C4.5, Credal DT, Credal C4.5, Adaptative Credal C4.5), Naive Bayes (NB), K-Nearest Neighbor (kNN), Linear Discriminant Analysis (LDA), Logistic Regression (LogR), etc

4. Clustering

K-Means, K-Medoids, Self-Organizing Map (SOM), Fuzzy C-Means (FCM), etc

5. Association

FP-Growth, A Priori, Coefficient of Correlation, Chi Square, etc



Supervised Learning

A person in a business suit is holding a coffee cup. The background shows a desk with documents, a pen, and a laptop. The image is overlaid with a large yellow and blue circular graphic.

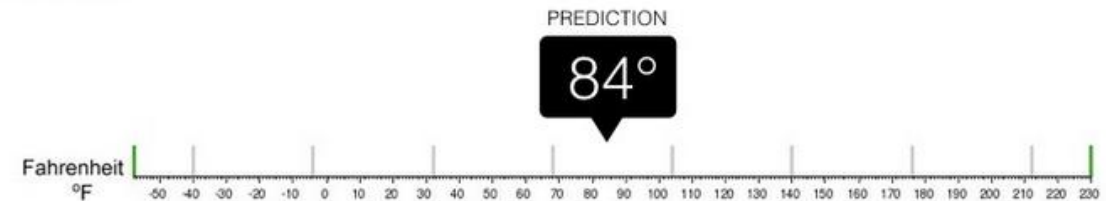
Regression vs Classification

Regression	Classification
Digunakan untuk memprediksi data kontinu (<i>continuous quantity</i>)	Digunakan untuk memprediksi label diskret pada suatu kelas (<i>discrete class label</i>)
Regresi dengan multiple input biasa disebut multivariate regression	Klasifikasi dengan 2 label kelas disebut binary dan lebih dari 2 kelas disebut dengan multi-class
Scoring yang umum digunakan : RMSE, R^2 , MAE, MAPE	Scoring yang umum digunakan : Accuracy, F1-score, ROC-AUC
Contoh : prediksi harga rumah, prediksi GDP, prediksi pertumbuhan penduduk.	Contoh : <i>fraud-detection</i> , <i>email spam filter</i> , <i>image classification</i> .



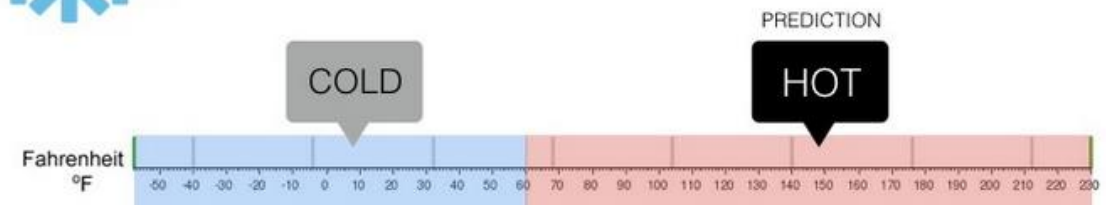
Regression

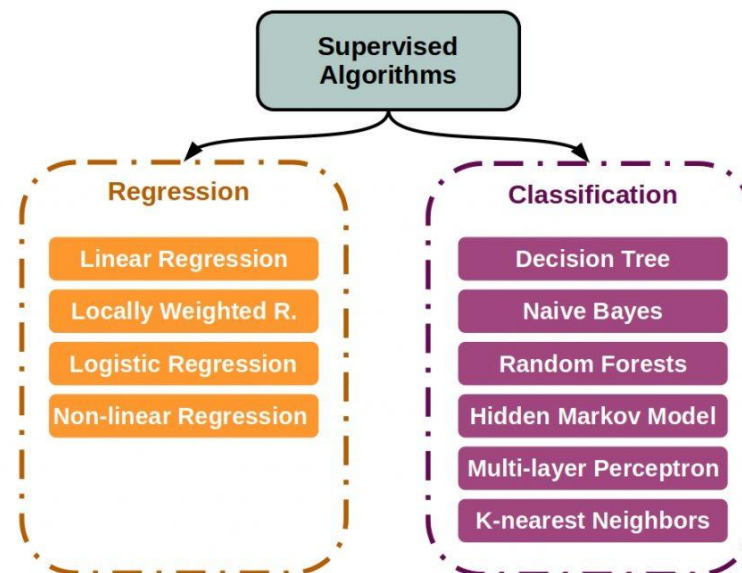
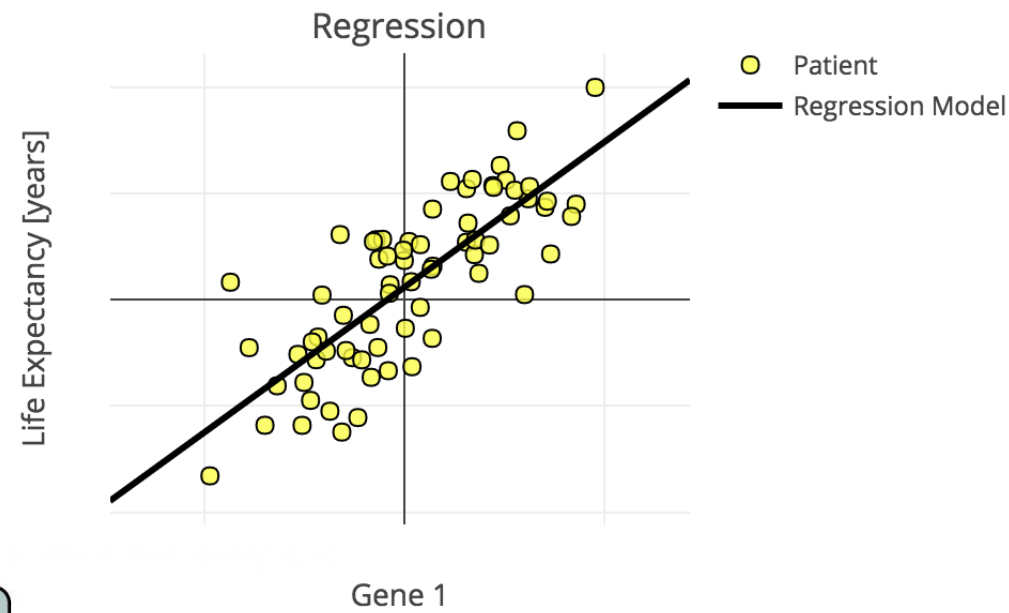
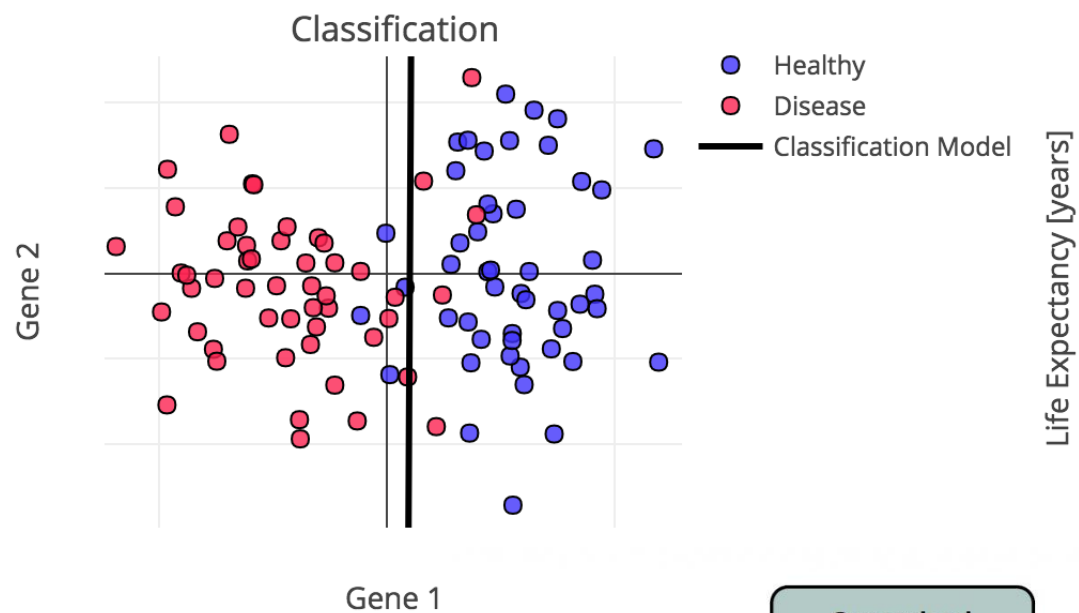
What is the temperature going to be tomorrow?



Classification

Will it be Cold or Hot tomorrow?





Dataset dengan Label

Attribute/Feature/Dimension

Class/Label/Target

	Sepal Length (cm)	Sepal Width (cm)	Petal Length (cm)	Petal Width (cm)	Type
1	5.1	3.5	1.4	0.2	<i>Iris setosa</i>
2	4.9	3.0	1.4	0.2	<i>Iris setosa</i>
3	4.7	3.2	1.3	0.2	<i>Iris setosa</i>
4	4.6	3.1	1.5	0.2	<i>Iris setosa</i>
5	5.0	3.6	1.4	0.2	<i>Iris setosa</i>
...					
51	7.0	3.2	4.7	1.4	<i>Iris versicolor</i>
52	6.4	3.2	4.5	1.5	<i>Iris versicolor</i>
53	6.9	3.1	4.9	1.5	<i>Iris versicolor</i>
54	5.5	2.3	4.0	1.3	<i>Iris versicolor</i>
55	6.5	2.8	4.6	1.5	<i>Iris versicolor</i>
...					
101	6.3	3.3	6.0	2.5	<i>Iris virginica</i>
102	5.8	2.7	5.1	1.9	<i>Iris virginica</i>
103	7.1	3.0	5.9	2.1	<i>Iris virginica</i>

Nominal

Numerik

Regression: Multivariate Regression



Analisis Regresi adalah Teknik statistic untuk pemodelan dan investigasi hubungan dua atau lebih variable.

Santoso Budi dan Ardian Umam (2018), Data Mining dan Big Data Analytics

“Dalam analisis regresi, ada satu atau lebih variable independent (biasanya menggunakan notasi x), dan satu variable response yang diwakili notasi y .”

Evaluasi Model Regresi

- **Do you have outliers in the data?**
 - Use MAE
- **Are you sure they are outliers?**
 - Use MAE
- **Or they are just unexpected values we should still care about?**
 - Use MSE

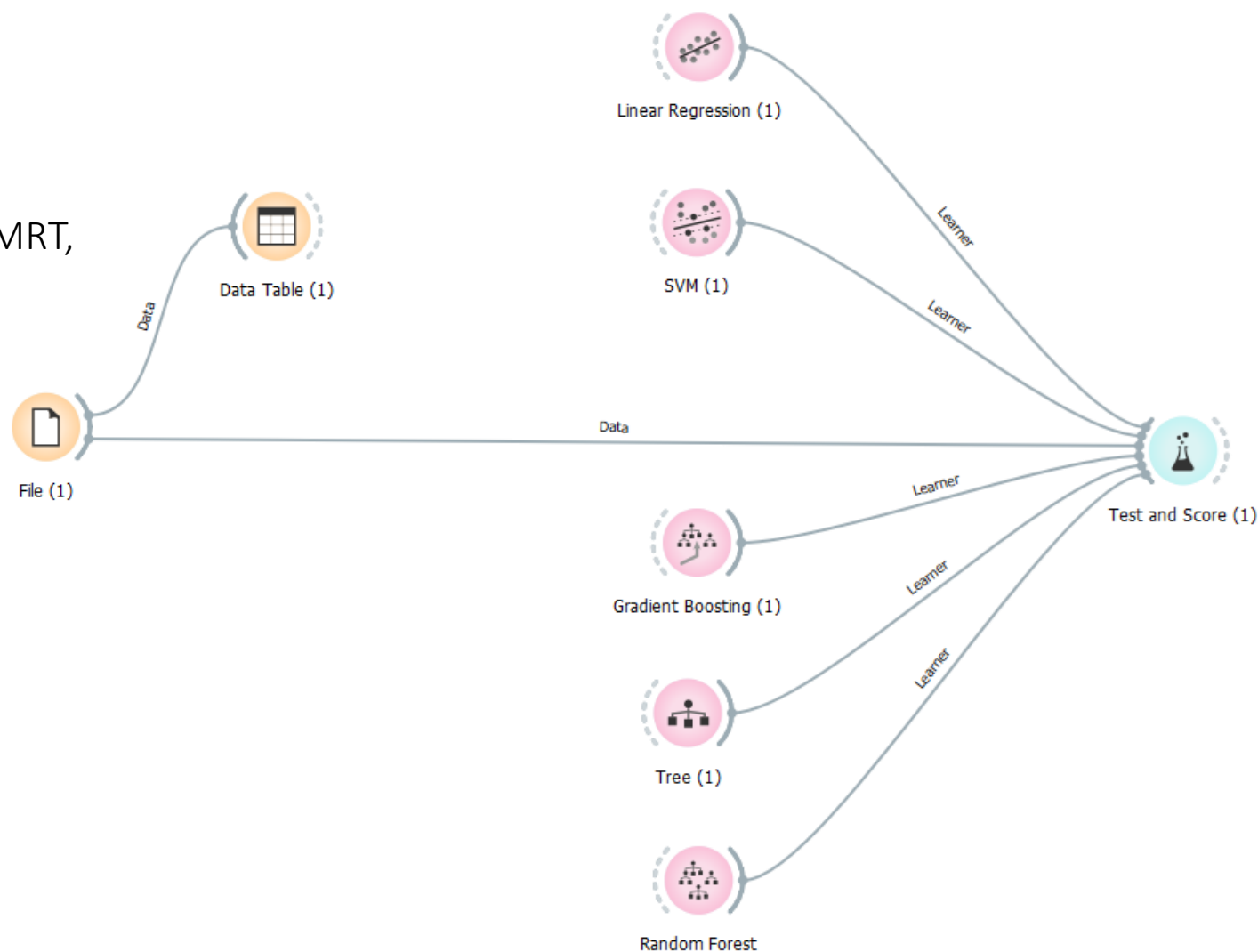
Mean Squared Error (MSE)	Semakin tinggi nilai ini, semakin buruk modelnya.
Root Mean Squared Error (RMSE)	Semakin tinggi nilai ini, semakin buruk modelnya.
Mean Absolute Error (MAE)	Semakin tinggi nilai ini, semakin buruk modelnya.
R Squared (R^2)	Nilai maksimum R^2 adalah 1 tetapi minimum bisa minus tak terhingga.

Use Case: Prediksi Harga Rumah

Dependent Variable: Harga Rumah

Independent Variable:

Transaction date, house age, distance to MRT,
Numbers of Convenience store, Latitude,
Longitude



1 : Load the data

Untitled * - Orange

File Edit View Widget Options Help

Data

File CSV File Import Datasets SQL Table

Data Table Paint Data Data Info Aggregate Columns

Data Sampler Select Columns Select Rows Pivot Table

Rank Correlations Merge Data Concatenate

Select by Data Index Transpose Randomize Preprocess

Apply Domain Impute Outliers Edit Domain

Python Script Create Instance Color Continuize

File

Read data from an input file or network and send a data table to the output.

[more...](#)

i # T ↘ || ?



File - Orange

Source

File: 2. Regresi_Prediksi Harga Rumah.xlsx ... Reload

URL: ...

Info

414 instance(s)
8 feature(s) (no missing values)
Data has no target variable.
0 meta attribute(s)

Columns (Double click to edit)

	Name	Type	Role	Values
1	No	S text	meta	
2	X1 transaction date	T datetime	feature	
3	X2 house age	N numeric	feature	
4	X3 distance to the nearest MRT station	N numeric	feature	
5	X4 number of convenience stores	N numeric	feature	
6	X5 latitude	N numeric	feature	
7	X6 longitude	N numeric	feature	
8	Y house price of unit area	N numeric	target	

Reset Apply

Browse documentation datasets

? | 414

2 : View the data

File Edit View Widget Options Help

Data

File CSV File Import Datasets SQL Table

Data Table Paint Data Data Info Aggregate Columns

Data Sampler Select Columns Select Rows Pivot Table

Rank Correlations Merge Data Concatenate

Select by Data Index Transpose Randomize Preprocess

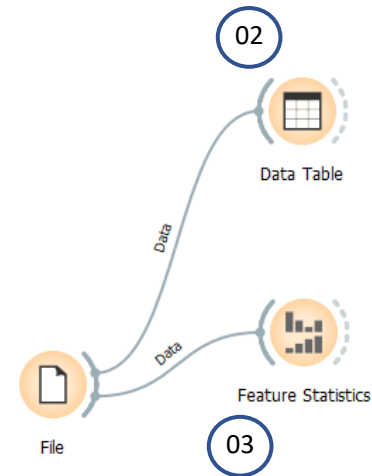
Apply Domain Impute Outliers Edit Domain

Python Script Create Instance Color Continuize

Feature Statistics

Show basic statistics for data features.

[more...](#)



Data Table - Orange

Info
414 instances (no missing data)
6 features
Numeric outcome
1 meta attribute

Variables
☒ Show variable labels (if present)
☐ Visualize numeric values
☒ Color by instance classes

Selection
☒ Select full rows

Restore Original Order

☒ Send Automatically

	Y house price of unit area	No	X1 transaction date	X2 house age	t
1	379000000	1	2012-11-01 00:00:00	32.0	
2	422000000	2	2012-11-01 00:00:00	19.5	
3	473000000	3	2013-06-01 00:00:00	13.3	
4	548000000	4	2013-06-01 00:00:00	13.3	
5	431000000	5	2012-09-01 00:00:00	5.0	
6	321000000	6	2012-08-01 00:00:00	7.1	
7	403000000	7	2012-08-01 00:00:00	34.5	
8	467000000	8	2013-05-01 00:00:00	20.3	
9	188000000	9	2013-06-01 00:00:00	31.7	
10	221000000	10	2013-05-01 00:00:00	17.9	
11	414000000	11	2012-12-01 00:00:00	34.8	
12	581000000	12	2013-03-01 00:00:00	6.3	
13	393000000	13	2012-11-01 00:00:00	13.0	
14	238000000	14	2012-08-01 00:00:00	20.4	
15	343000000	15	2013-06-01 00:00:00	13.2	
16	505000000	16	2013-06-01 00:00:00	35.7	
17	701000000	17	2013-03-01 00:00:00	0.0	
18	374000000	18	2012-09-01 00:00:00	17.7	
19	423000000	19	2013-05-01 00:00:00	16.9	

414 414 | 414

Feature Statistics - Orange

	Name	tributi	Mean	Median	Dispersion	Min.	Max.	Missing
N	X2 house age		17.713	16.1	0.642	0.0	43.8	0 (
N	X3 distance to the nearest MRT station		1083.8856889	492.23130	1.1630234	23.38284	6488.02100	0 (
N	X4 number of convenience stores		4.09	4	0.72	0	10	0 (

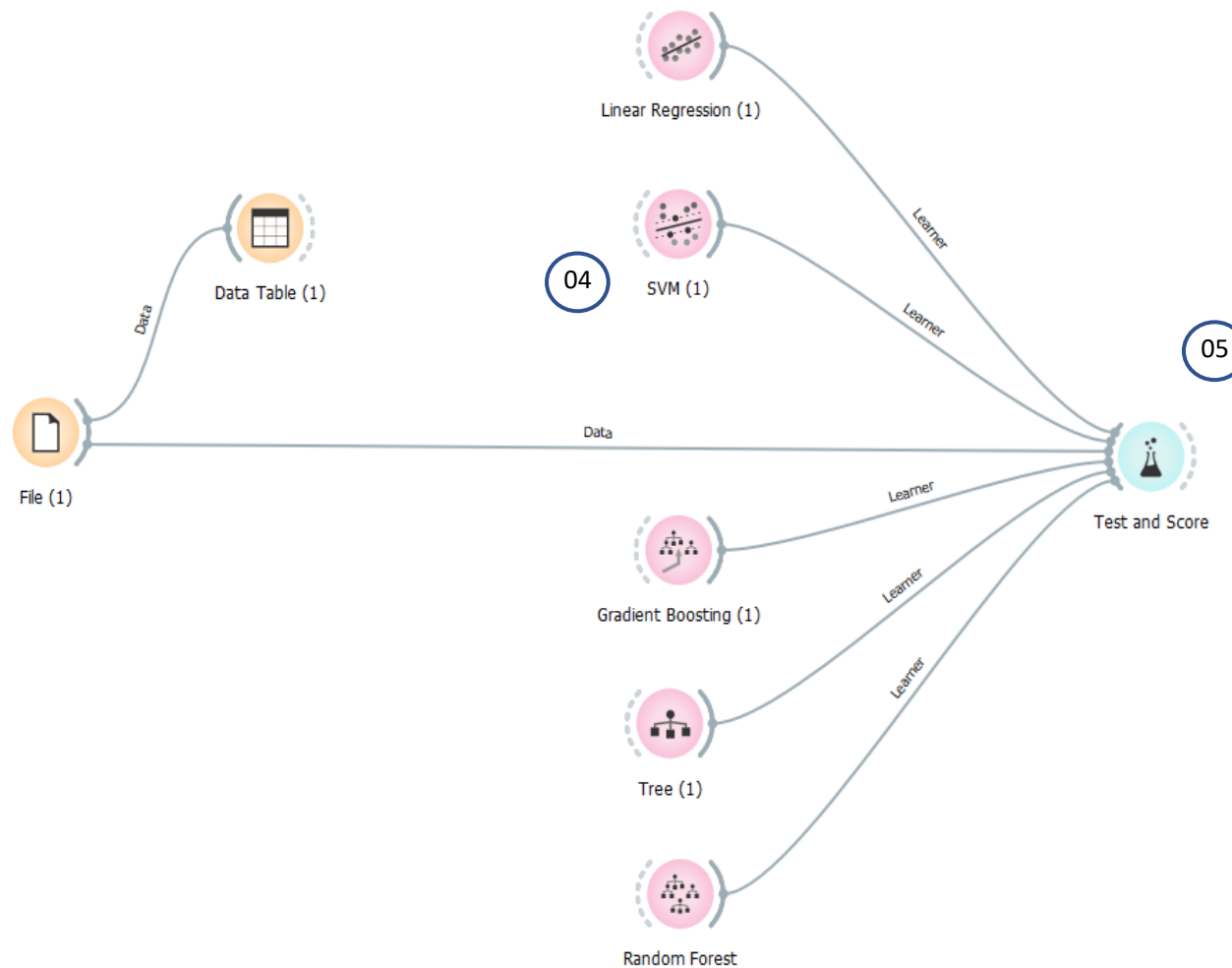
Color: N Y house price of u

☒ Send Automatically

414 -

3 : Melihat statistic dari data yang telah di load

4 : Modelling



5 : Scoring

Test and Score - Orange

Sampling

☒ Cross validation

Number of folds: 10

☐ Stratified

☐ Cross validation by feature

☐ Random sampling

Repeat train/test: 10

Training set size: 60 %

☐ Stratified

☐ Leave one out

☐ Test on train data

☐ Test on test data

Model Comparison

Mean square error

☐ Negligible difference: 0.1

Evaluation Results

Model	MSE	RMSE	MAE	R2
Tree (1)	6789946449127269.000	82401131.358	54997989.746	0.632
SVM (1)	19690001502012976.000	140321065.781	111961349.347	-0.066
Random Forest	5835425427428346.000	76389956.326	49843437.074	0.684
Linear Regression (1)	7972226283110618.000	89287324.314	62161477.931	0.568
Gradient Boosting (1)	6173183329935541.000	78569608.182	49943766.424	0.666

Model Comparison by MSE

	Tree (1)	SVM (1)	Random Forest	Linear Regression (1)	Gradient Boosting (1)
Tree (1)		0.000	0.789	0.128	0.751
SVM (1)	1.000		1.000	1.000	1.000
Random Forest	0.211	0.000		0.025	0.225
Linear Regression (1)	0.872	0.000	0.975		0.963
Gradient Boosting (1)	0.249	0.000	0.775	0.037	

Table shows probabilities that the score for the model in the row is higher than that of the model in the column. Small numbers show the probability that the difference is negligible.

414 | 414 | 5x414

Classification

A person is working at a desk with multiple laptops. They are holding a white coffee cup in their right hand and a blue folder in their left hand. The word "Classification" is overlaid on a large blue circular graphic on the left side of the image.

Classification Model

**K-Nearest
Neighbours**

**Naïve
Bayes**

**Gradient
Boosting**

**Decision
Tree**

SVM

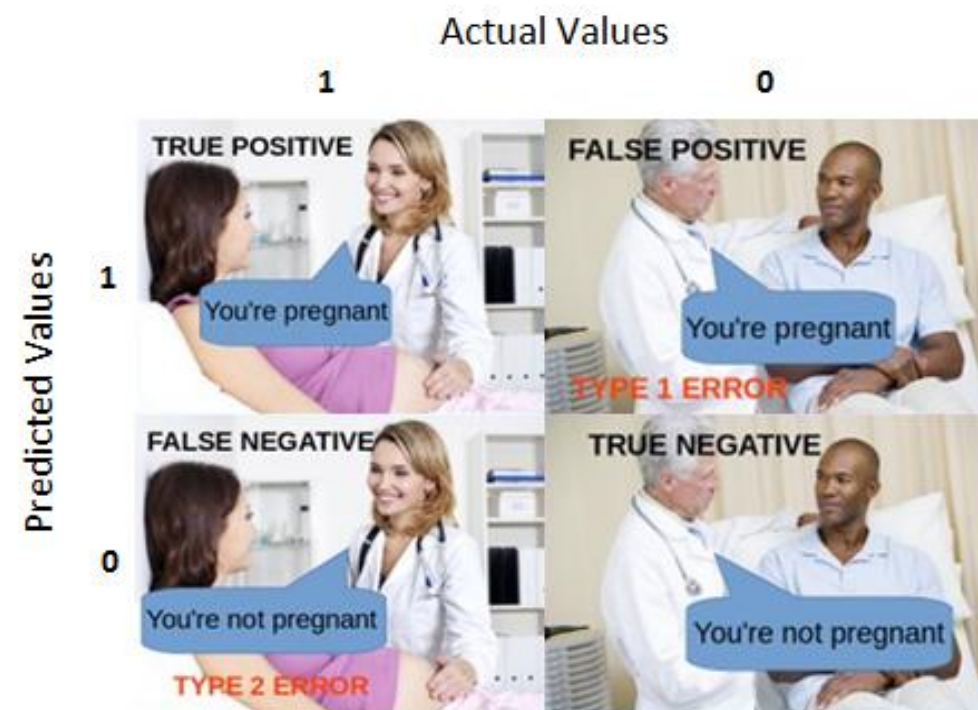
**Random
Forest**

**Logistic
Regression**

Mengukur Performa Model

		Actual Values	
		1 (Positive)	0 (Negative)
Predicted Values	1 (Positive)	TP (True Positive)	FP (False Positive) Type I Error
	0 (Negative)	FN (False Negative) Type II Error	TN (True Negative)

Confusion Matrix



Akurasi : $(TN+TP) / \text{Jumlah Data}$

Recall / True Positive Rate : $TP / (FN+TP)$

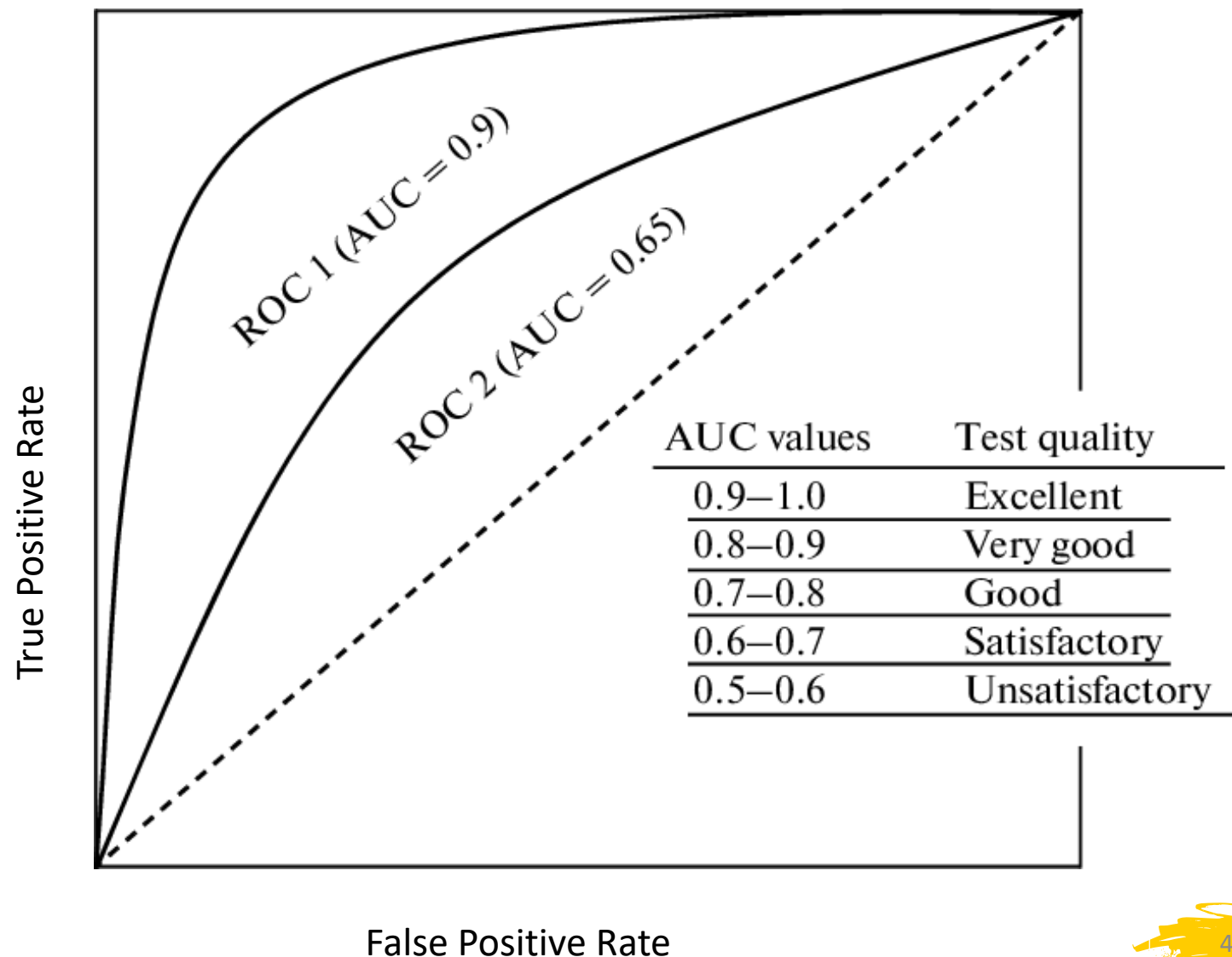
Precision : $TP / (TP+FP)$

Performance Metrics

Classification Cases Area under ROC Curve (AUC)

ROC

Graph that shows performance of a classification model at all classification thresholds

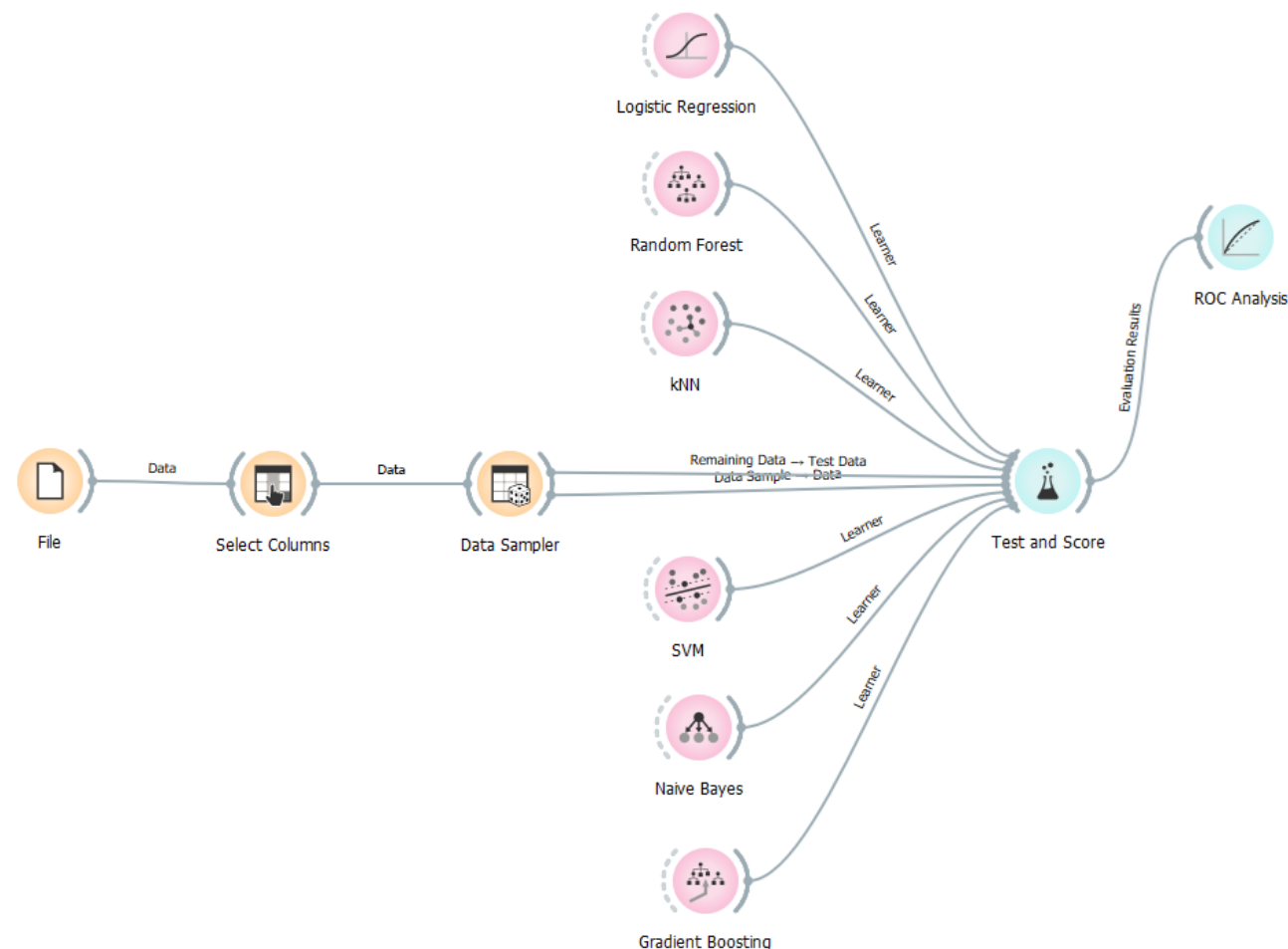


Use Case: Prediksi Penetapan Consignment

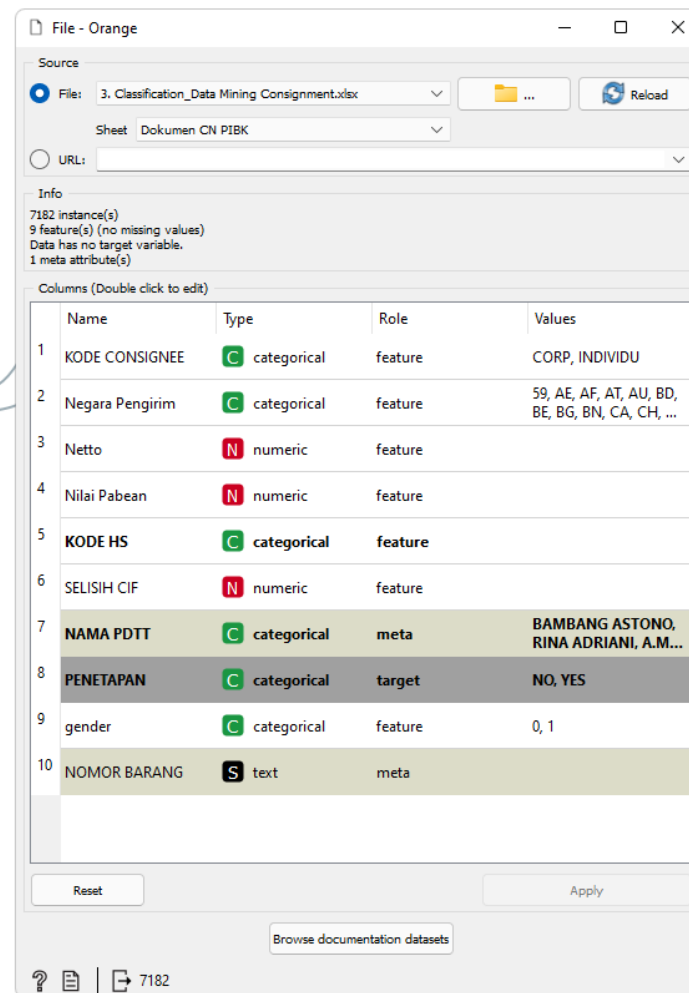
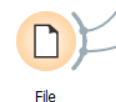
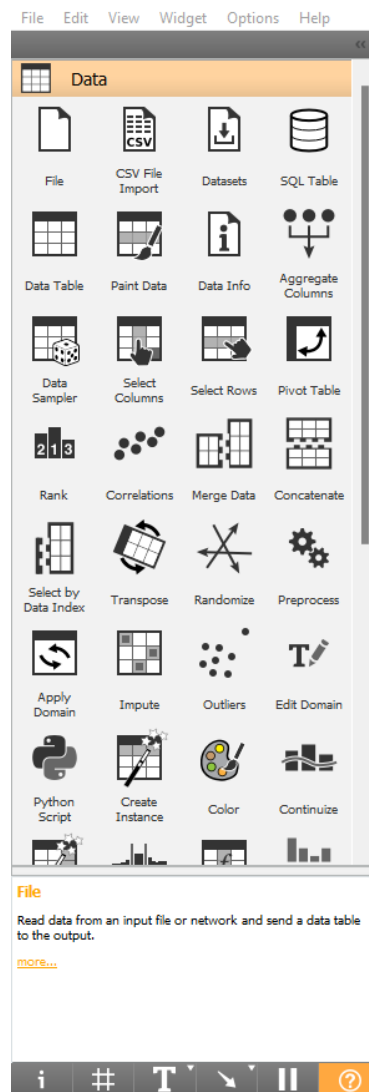
Target: Penetapan (Yes/No)

Feature:

Kode Consignment, Negara Pengirim, Netto, Nilai Pabean, Kode HS, Selisih CIF, Gender Petugas



1 : Load the data



2 : View the data

File Edit View Widget Options Help

Data

File CSV File Import Datasets SQL Table

Data Table Paint Data Data Info Aggregate Columns

Data Sampler Select Columns Select Rows Pivot Table

Rank Correlations Merge Data Concatenate

Select by Data Index Transpose Randomize Preprocess

Apply Domain Impute Outliers Edit Domain

Python Script Create Instance Color Continuize

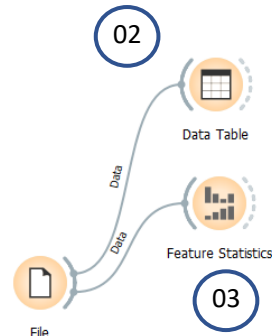
Create Class Discretize Feature Constructor Feature Statistics

Data Table

View the dataset in a spreadsheet.

[more...](#)

i # T ↵ || ?



Data Table - Orange

Info
7182 instances (no missing data)
9 features
No target variable.
1 meta attribute

Variables
☒ Show variable labels (if present)
☐ Visualize numeric values
☒ Color by instance classes

Selection
☒ Select full rows

Restore Original Order

☒ Send Automatically

? | 7182 | 7182 | 7182

	NOMOR BARANG	KODE CONSIGNEE	Negara Pengirim	Netto	Nilai Pabean	KODE HS
1	EP861874667US	INDIVIDU	US	1.00	712650	
2	EM000565696JP	INDIVIDU	JP	4.00	641385	
3	CP315569511MY	INDIVIDU	MY	10.00	425670	
4	CP123869512MY	INDIVIDU	MY	10.00	425670	
5	CJ201882293US	INDIVIDU	US	10.00	425670	
6	CP391946576CN	INDIVIDU	CN	1.00	141890	
7	CP002642206MM	INDIVIDU	MM	10.00	425670	
8	CF337143853US	INDIVIDU	US	10.00	709450	
9	CX126707686FR	INDIVIDU	FR	0.80	28378	
10	CP135430238MY	INDIVIDU	MY	10.00	439859	
11	CP318402507MY	INDIVIDU	MY	10.00	425670	
12	LP138359705SG	INDIVIDU	SG	0.40	695261	
13	LP138024786SG	INDIVIDU	SG	0.30	255402	
14	EN040094730JP	INDIVIDU	JP	1.20	212835	
15	EY104087359AU	INDIVIDU	AU	1.00	425670	
16	RS329542479LV	INDIVIDU	LV	1.00	283780	

3 : Melihat statistic dari data yang telah diload

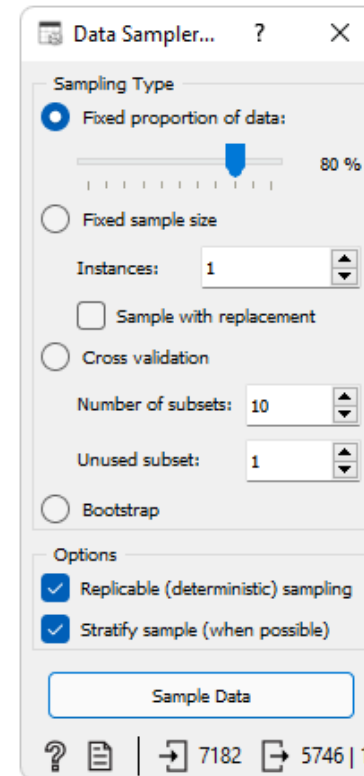
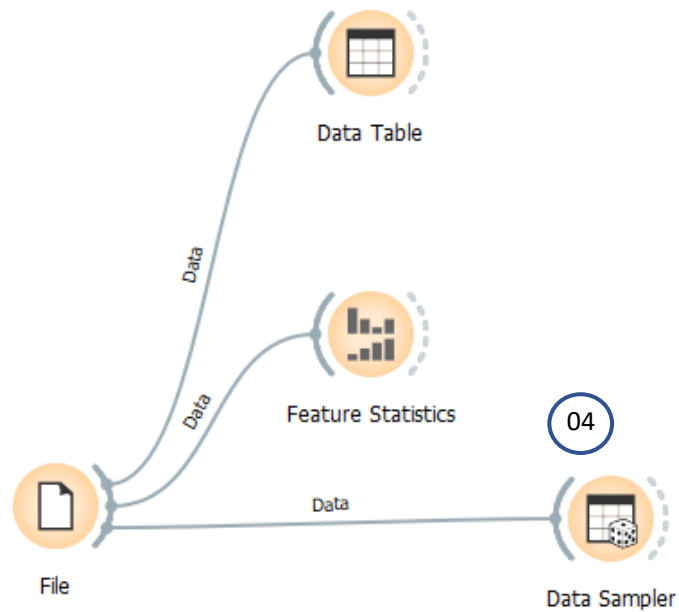
Feature Statistics - Orange

Name	Distribution	Mean	Median	Dispersion	Min.	Max.	Missing
NAMA PDDT			BAMBANG ASTONO	0.981			0 (0%)
PENETAPAN			NO	0.222			0 (0%)
gender			1	0.489			0 (0%)

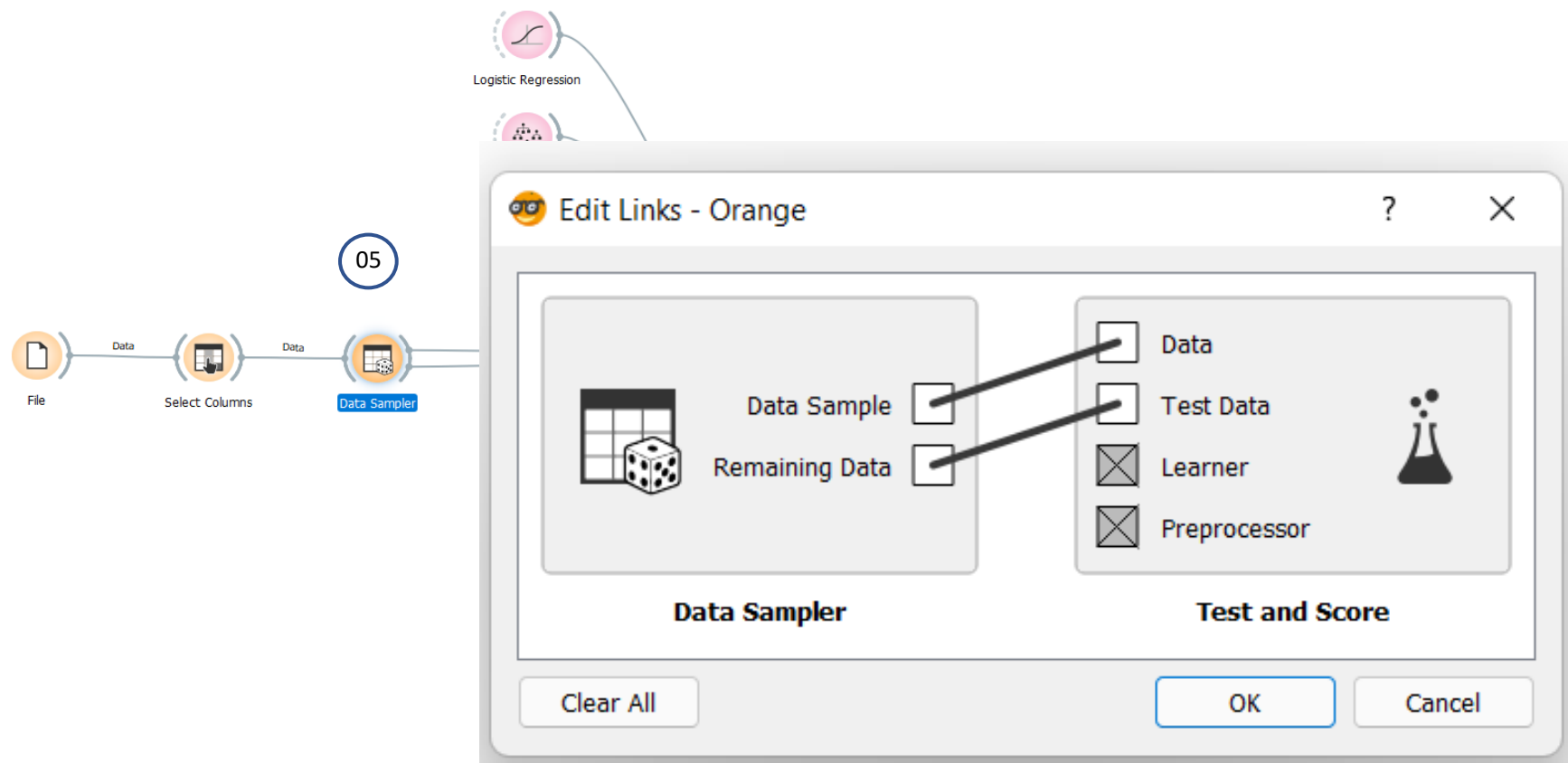
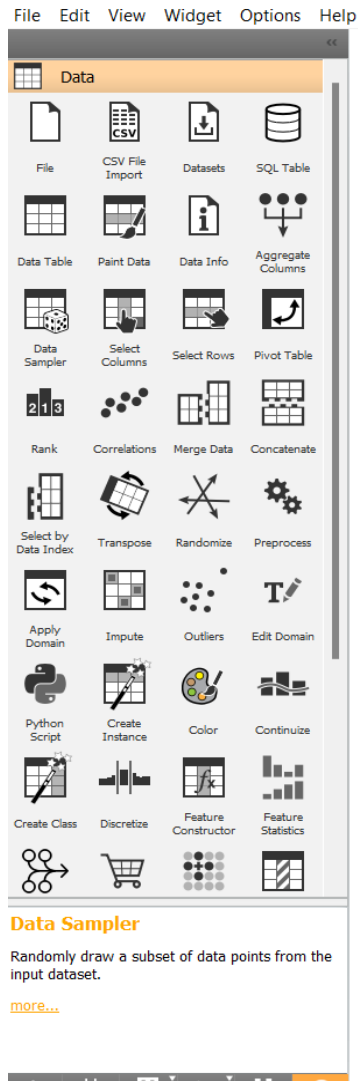
Color: **C** PENETAPAN

☒ Send Automatically

? | 7182 | -

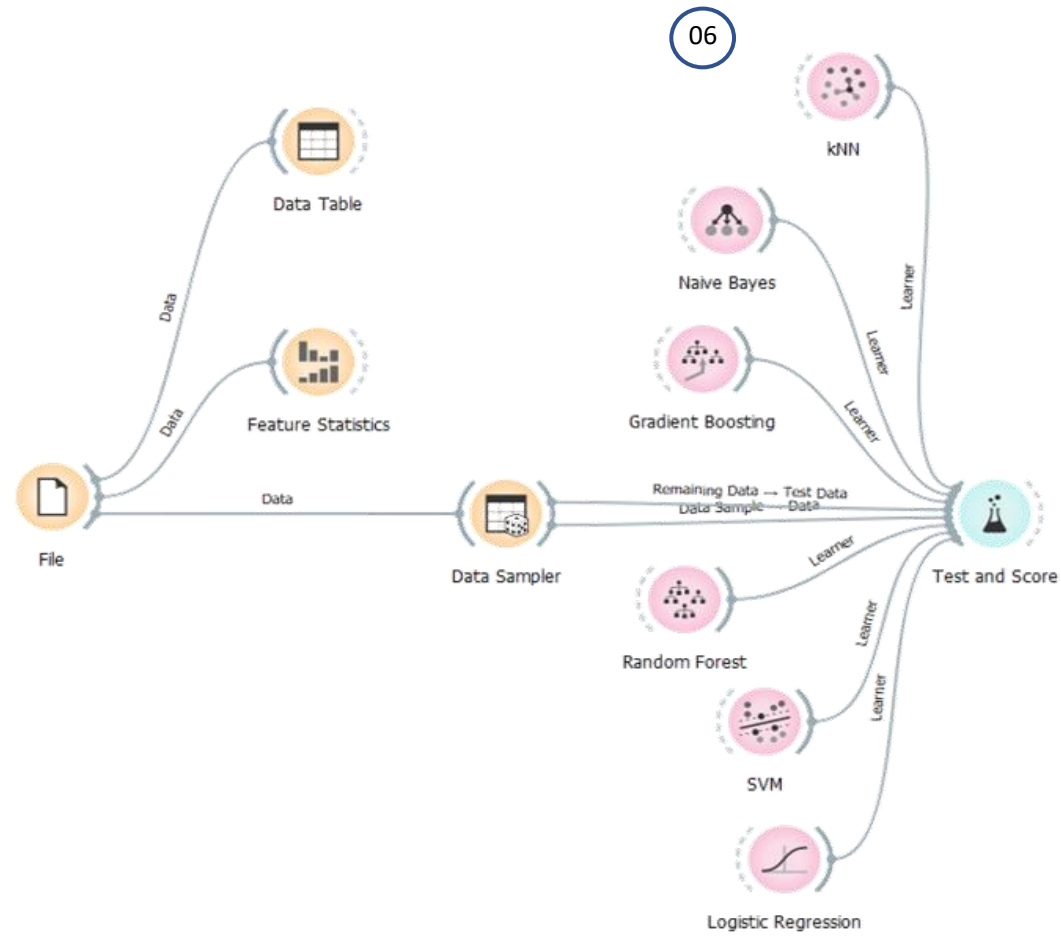


4 : Membagi data sampler

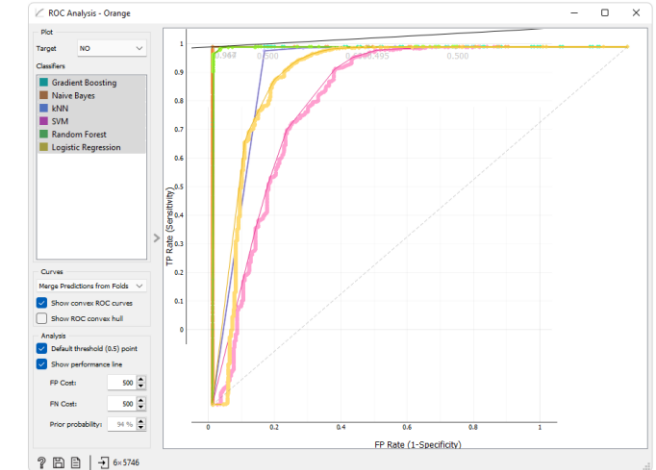


5 : Pembagian data training dan data testing

6 : Modelling



8: ROC Analysis



7: Scoring

Test and Score - Orange window showing evaluation results for various models.

Model	AUC	CA	F1	Precision	Recall
kNN	0.935	0.980	0.978	0.980	0.980
SVM	0.862	0.765	0.826	0.938	0.765
Random Forest	0.998	0.994	0.993	0.994	0.994
Naive Bayes	1.000	1.000	1.000	1.000	1.000
Logistic Regression	0.919	0.977	0.975	0.977	0.977
Gradient Boosting	1.000	1.000	1.000	1.000	1.000

	kNN	SVM	Random Forest	Naive Bayes	Logistic Regres...	Gradient Boost...
kNN		0.979	0.001	0.000	0.739	0.000
SVM	0.021		0.001	0.000	0.048	0.000
Random Forest	0.999	0.999		0.217	0.997	0.217
Naive Bayes	1.000	1.000	0.783		0.998	0.500
Logistic Regression	0.261	0.952	0.003	0.002		0.002
Gradient Boosting	1.000	1.000	0.783	0.500	0.998	

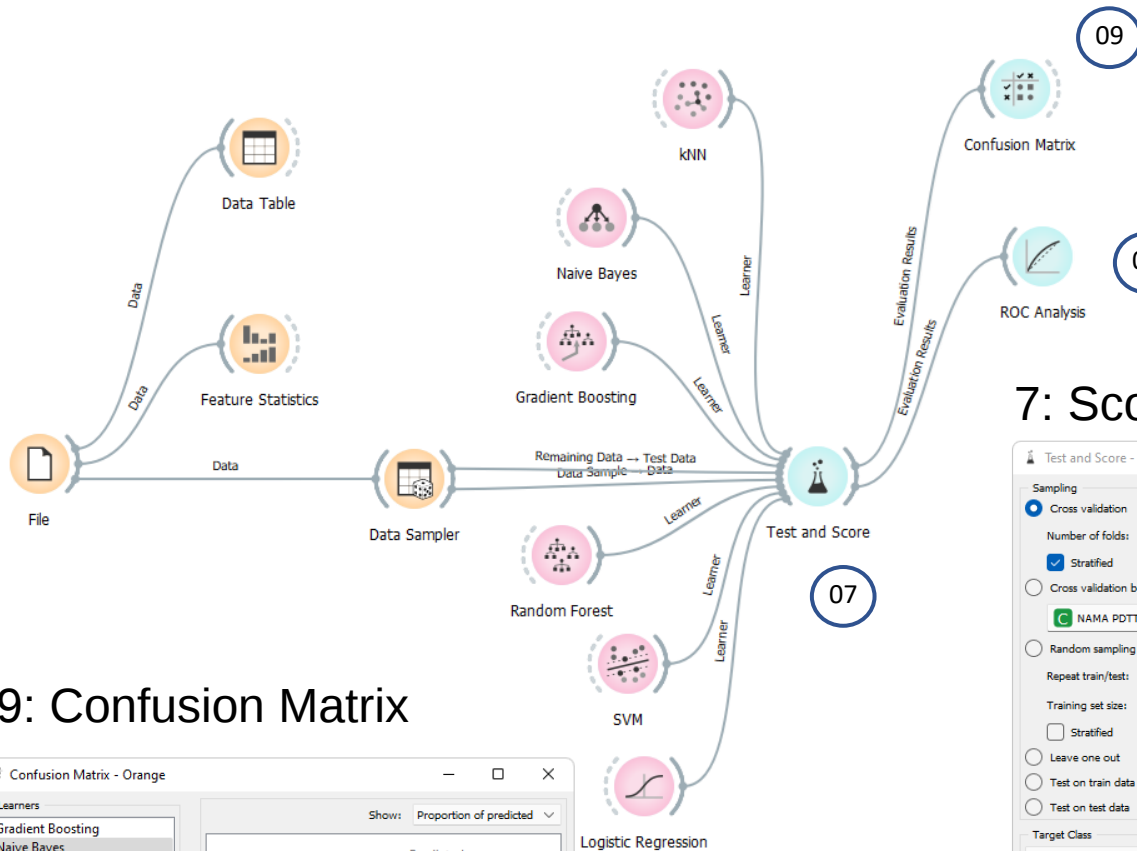
Table shows probabilities that the score for the model in the row is higher than that of the model in the column. Small numbers show the probability that the difference is negligible.

9: Confusion Matrix

Confusion Matrix - Orange window showing the confusion matrix for the Naive Bayes classifier.

		Predicted		Σ
		NO	YES	
Actual	NO	100.0 %	0.0 %	5412
	YES	0.0 %	100.0 %	334
Σ		5412	334	5746

Buttons: Select Correct, Select Misclassified, Clear Selection



A background image showing a business meeting. A person in a suit and tie is holding a coffee cup. Another person's hand is visible on a laptop. There are papers with charts and graphs on the table. A large blue circle is overlaid on the left side of the image.

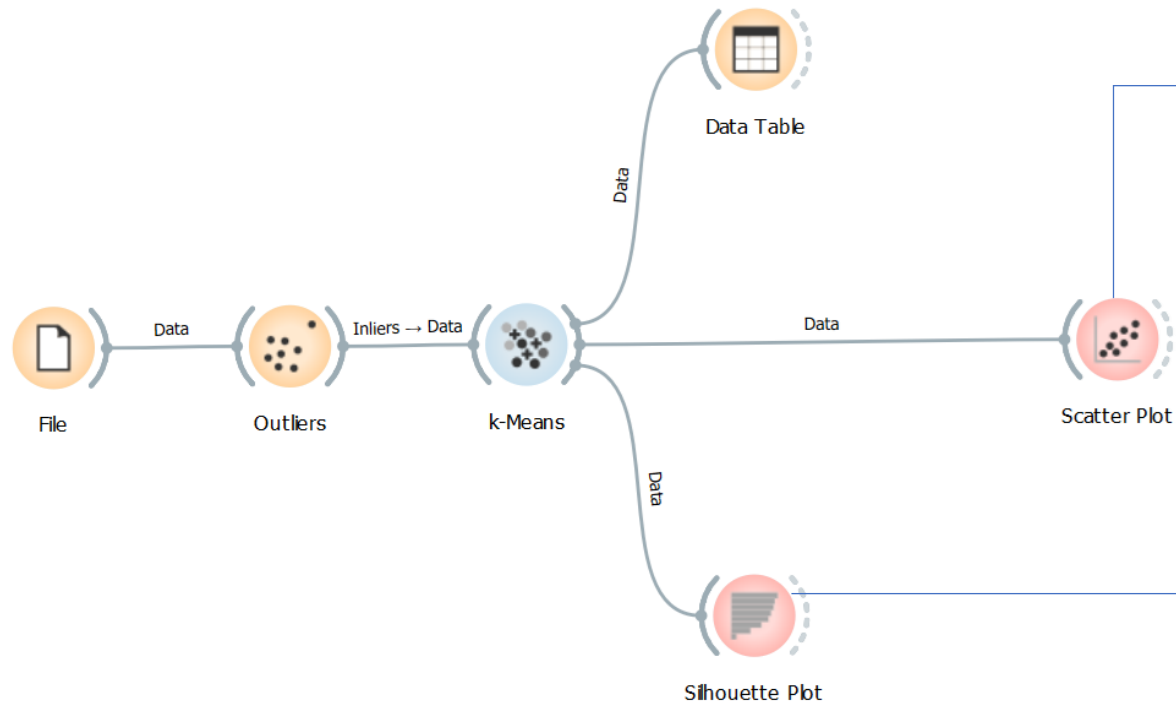
Unsupervised Learning

Unsupervised Learning

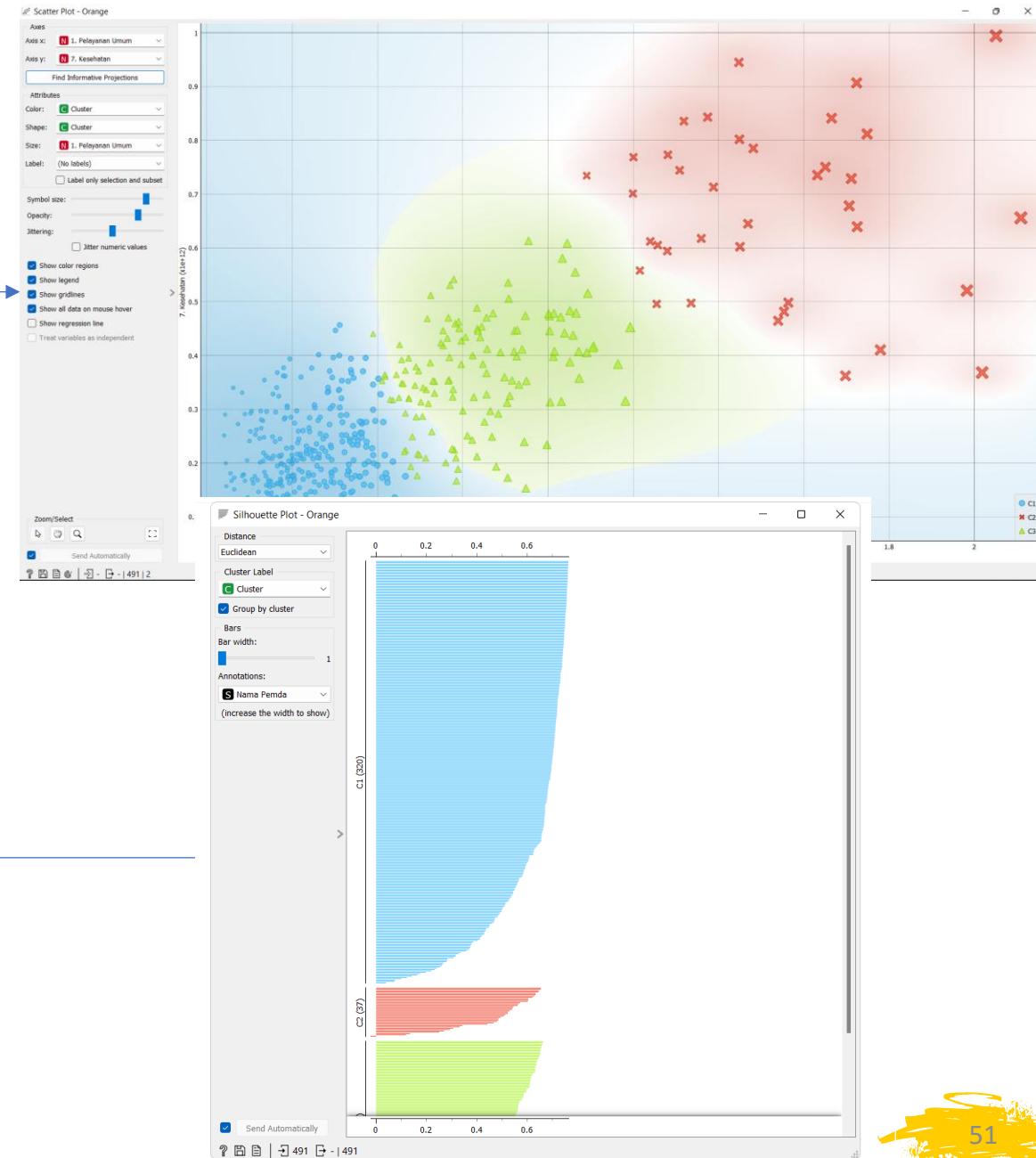
- Algoritma data mining mencari pola dari **semua variable** (atribut)
- Variable (atribut) yang menjadi **target/label/class** tidak ditentukan (tidak ada)
- Algoritma **clustering** adalah algoritma unsupervised learning



Use Case: APBD (Pelayanan Umum dan Kesehatan)



Silhouette score adalah ukuran seberapa mirip suatu objek dengan clusternya sendiri dibandingkan dengan cluster lain dan sangat penting dalam pembuatan Silhouette plot. Silhouette score mendekati 1 menunjukkan bahwa instance data dekat dengan pusat cluster dan instance yang memiliki Silhouette score mendekati 0 berada di perbatasan antara dua cluster.



Suplemen

Text Mining: Word Cloud





Tujuan *text mining*

untuk menggali informasi yang dapat berguna dari beberapa Dokumen, Selain itu *text mining* dapat mendukung proses *knowledge discovery* pada beberapa Dokumen yang besar.

Beberapa area penerapan *text mining*

1. Ekstraksi Informasi (*Information Extraction*)
2. Pelacakan Topik (*Topic Tracking*)
3. Perangkuman (*Summarization*)
4. Kategorisasi (*Categorization*)
5. Penggugusan (*Clustering*)
6. Penautan Konsep (*Concept Linking*)
7. Penjawaban Pertanyaan (*Question Answering*)



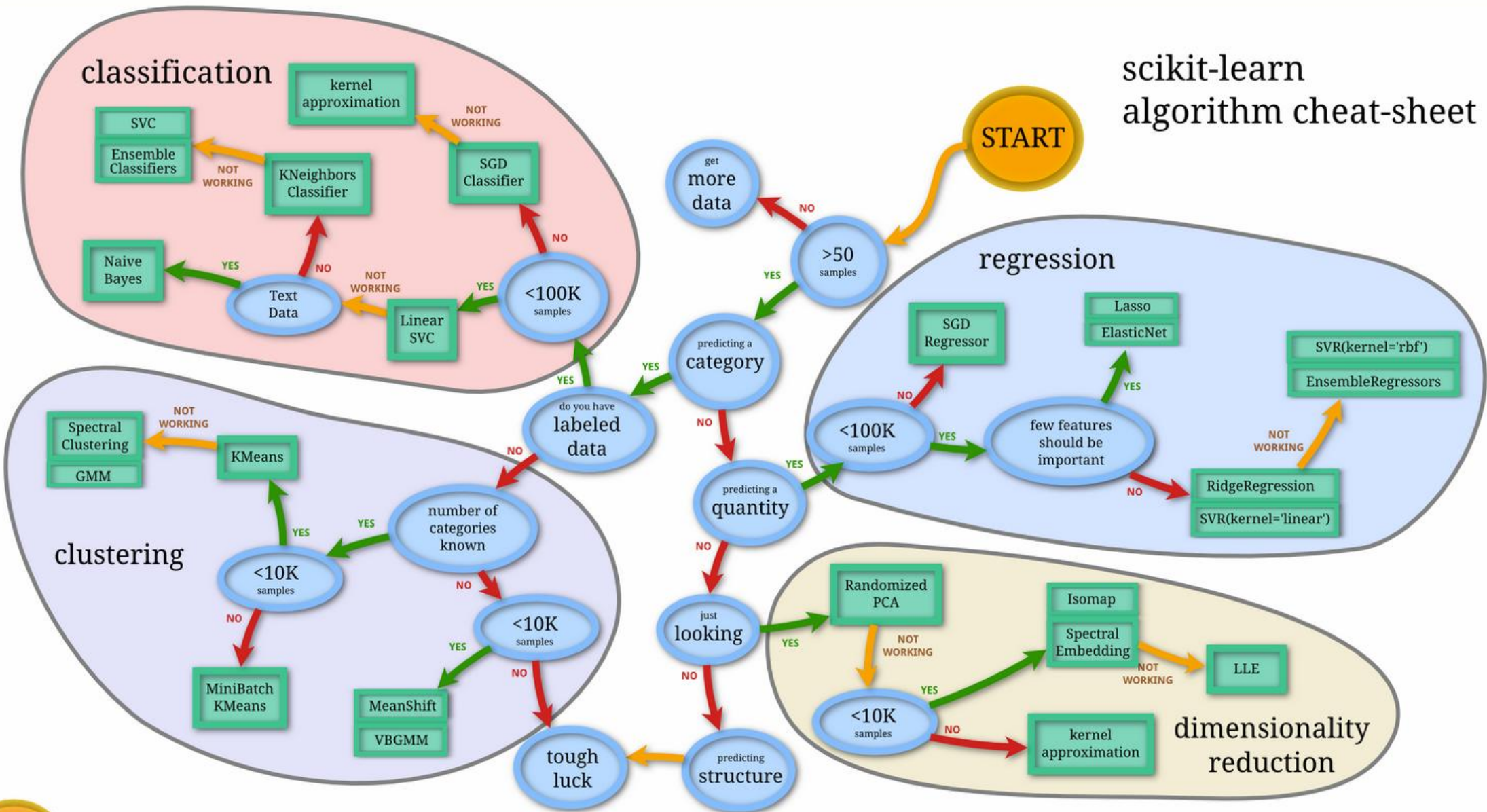
Word cloud

metode untuk menampilkan data teks secara visual. Gambaran frekuensi kata-kata dapat ditampilkan dalam bentuk yang menarik namun tetap informatif. Semakin sering satu kata digunakan, maka semakin besar pula ukuran kata tersebut ditampilkan dalam *word cloud*.

A person wearing a red and white checkered shirt is sitting at a desk, writing in a notebook with a blue pen. The desk is cluttered with various items, including a stack of books, a small globe, and several colorful sticky notes. The background is a blurred view of a city through a window. The text "Terima Kasih." is overlaid on the right side of the image in a large, white, outlined font.

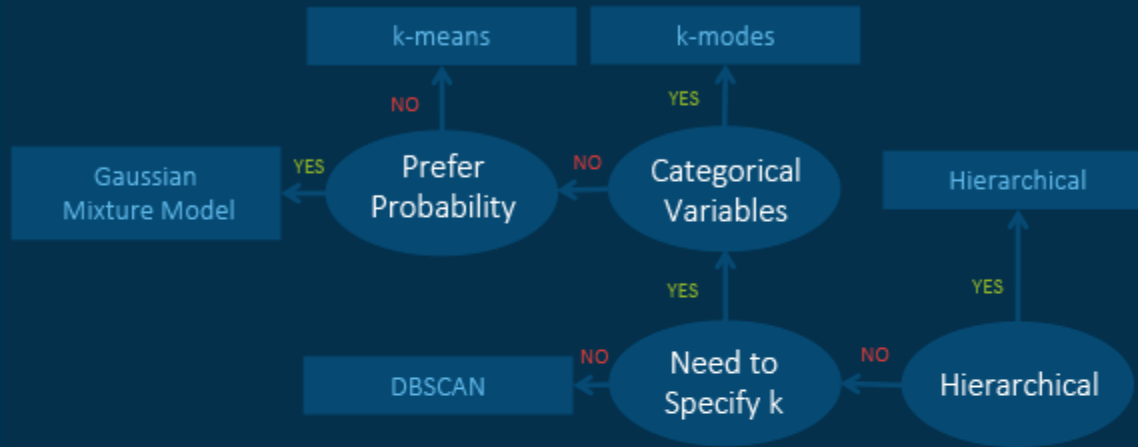
Terima Kasih.

scikit-learn
algorithm cheat-sheet

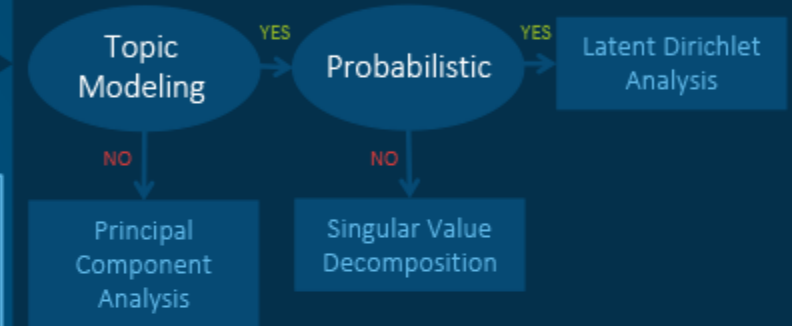


Machine Learning Algorithms Cheat Sheet

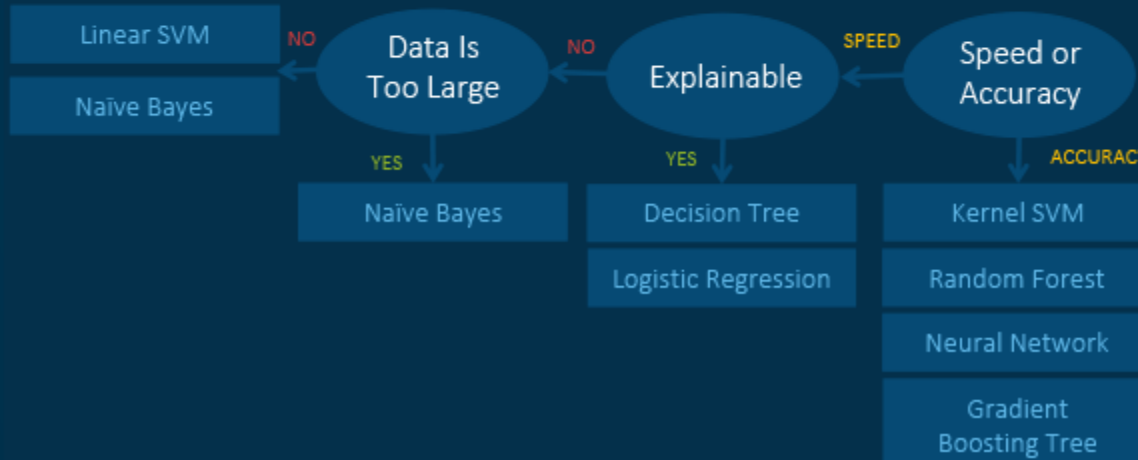
Unsupervised Learning: Clustering



Unsupervised Learning: Dimension Reduction



Supervised Learning: Classification



Supervised Learning: Regression

