

창의자율프로젝트



REFERENCE SITE

https://orangedatamining.com/

https://orange3.readthedocs.io/projects/orange-visual-programming/en/latest/widgets/model/neuralnetwork.html





REFERENCE SITE

Orange

Developer(s) University of Ljubljana

Initial release 10 October 1996; 26 years

ago^[1]

Stable release 3.34.0^[2] / 5 December

2022; 3 months ago

Repository Orange Repository ☑

Written in Python, Cython, C++, C

Operating system Cross-platform

Type Machine learning, Data

mining, Data visualization,

Data analysis

License GPLv3 or later^{[3][4]}

Website orangedatamining.com ☑

P

Visual Programming Front-End 기능 제공

- Machine Learning
- Data Mining
- Data Visualization
- Data Analysis

CONTENTS

1. Abalone flesh weight prediction using linear regression

2. Star terror prevention using logistic regression

3. Courier delivery location clustering using K-Means

Orange Visual Programming



1. Abalone flesh weight prediction using Linear Regression





Problem

 전복의 나이테, 성별, 길이, 직경, 두께, 전체 무게, 내장 무게, 껍질 무게에 해당하는 총 8가지 데이터를 입력하면 AI가 전복의 순살(flesh) 무게를 예측할 수 있을까?





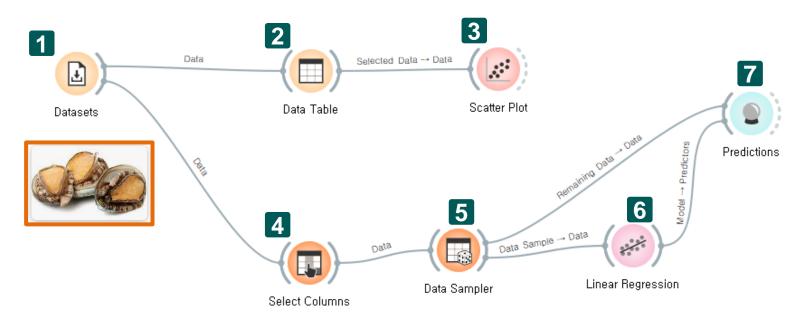




male

female

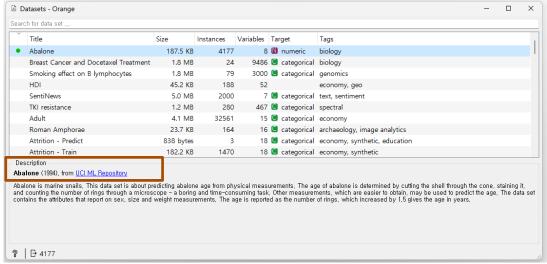
Data Type Al Model Structured data Linear Regression



1. Abalone flesh weight prediction using Linear Regression

1 Datasets





2 Data Table



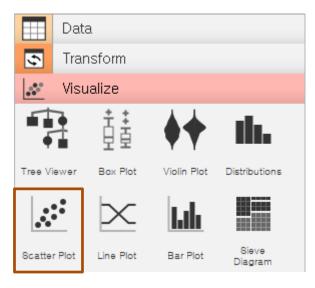
	Rings	Sex	Length	Diameter	Height	Whole weight	Shucked weight	Viscera weight	Shell weight
1	15	M	0.455	0.365	0.095	0.5140	0.2245	0.1010	0.1500
2	7	M	0.350	0.265	0.090	0.2255	0.0995	0.0485	0.0700
3	9	F	0.530	0.420	0.135	0.6770	0.2565	0.1415	0.2100
4	10	M	0.440	0.365	0.125	0.5160	0.2155	0.1140	0.1550
5	7	1	0.330	0.255	0.080	0.2050	0.0895	0.0395	0.0550
6	8	1	0.425	0.300	0.095	0.3515	0.1410	0.0775	0.1200
7	20	F	0.530	0.415	0.150	0.7775	0.2370	0.1415	0.3300
8	16	F	0.545	0.425	0.125	0.7680	0.2940	0.1495	0.2600
9	9	M	0.475	0.370	0.125	0.5095	0.2165	0.1125	0.1650
10	19	F	0.550	0.440	0.150	0.8945	0.3145	0.1510	0.3200
11	14	F	0.525	0.380	0.140	0.6065	0.1940	0.1475	0.2100
12	10	M	0.430	A350	~A110	0.4060	A1675	0.0810	01350
4168	9	M	0.500	0.380	0.125	0.5770	0.2690	0.1265	0.1535
4169	8	F	0.515	0.400	0.125	0.6150	0.2865	0.1230	0.1765
4170	10	M	0.520	0.385	0.165	0.7910	0.3750	0.1800	0.1815
4171	10	M	0.550	0.430	0.130	0.8395	0.3155	0.1955	0.2405
4172	8	M	0.560	0.430	0.155	0.8675	0.4000	0.1720	0.2290
4173	11	F	0.565	0.450	0.165	0.8870	0.3700	0.2390	0.2490
4174	10	M	0.590	0.440	0.135	0.9660	0.4390	0.2145	0.2605
4175	9	M	0.600	0.475	0.205	1.1760	0.5255	0.2875	0.3080
4176	10	F	0.625	0.485	0.150	1.0945	0.5310	0.2610	0.2960
4177	12	M	0.710	0.555	0.195	1.9485	0.9455	0.3765	0.4950

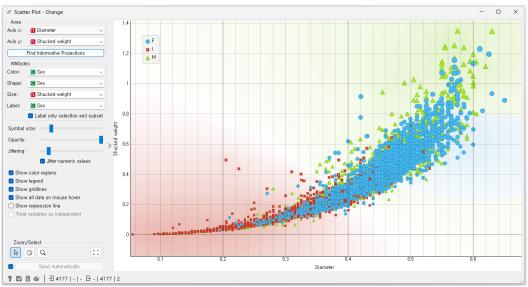
2 Data Table



Attribute name	Attribute		
Rings	나이테: 연도를 나타냄.		
Sex	성별: M(수), F(암), I(유아)		
Length	길이: 최장 껍질 측정(mm)		
Diameter	직경: 길이에 수직(mm)		
Height	두께: 껍질과 살 포함(mm)		
Whole weight	전체 무게: 그램 단위(g)		
Shucked weight	순살 무게: 그램 단위(g)		
Viscera weight	내장 무게: 피를 뺀 후 장 무게(g)		
Shell weight	껍질 무게: 건조 후(g)		

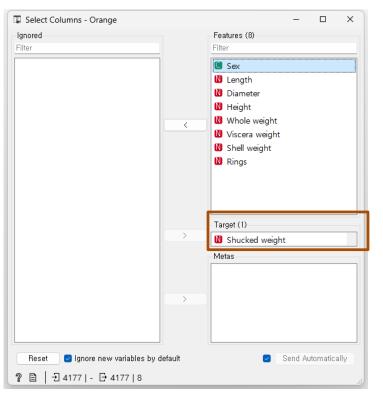
3 Scatter Plot



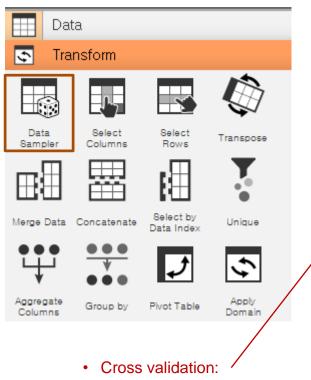


4 Select Columns

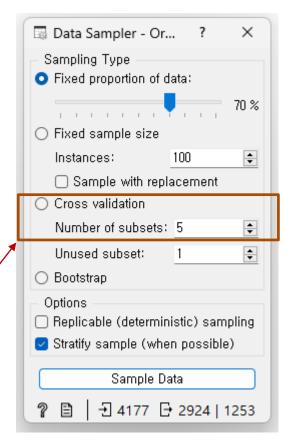




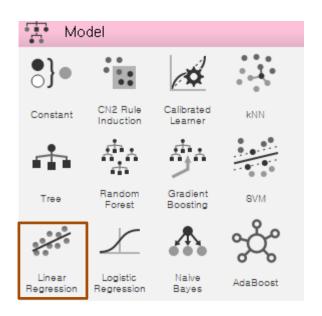
5 Data Sampler



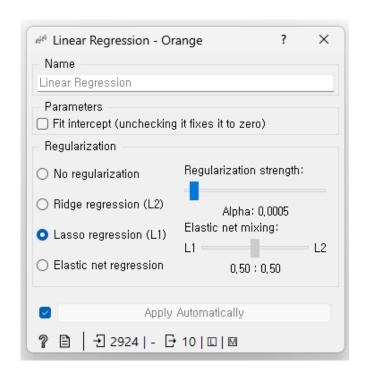
Stratified K-fold Cross Validation



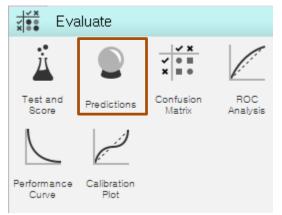
6 Linear Regression

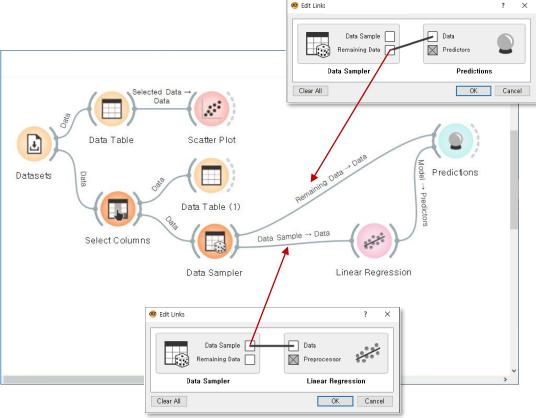


- Regularization
 - Ridge regression(L2): 각 계수의 제곱을 더하는 방식
 - Lasso regression(L1): 각 계수의 절댓값을 더하는 방식
 - Elastic net regression: L2와 L1 방식을 절충한 것



Predictions





1. Abalone flesh weight prediction using Linear Regression

7 Predictions

NIO MILLE	egression error: Diff	erence	~								Restore O	riginal O
	Linear Regression	error	Shucked weight	Sex	Length	Diameter	Height	Whole weight	Viscera weight	Shell weight	Rings	
513	0.7625	-0.0005	0.7630	M	0.640	0.525	0.185	1.7070	0.4205	0.4435	11	
614 。	0.0250	-0.0005	0.0255		0.255	0.195	0.070	0.0735	0.0200	0.0250	6	
315	0.2546	-0.0004	0.2550	M	0.500	0.420	0.125	0.6200	0.1500	0.2050	11	
16	0.1181	-0.0004	0.1185		0.360	0.300	0.085	0.2700	0.0640	0.0745	7	
17 _	0.3122	-0.0003	0.3125	F	0.535	0.420	0.130	0.6990	0.1565	0.2035	8]
18 💂	0.0427	-0.0003	0.0430	M	0.290	0.230	0.075	0.1165	0.0255	0.0400	7	
19 .	0.0293	-0.0002	0.0295	[0.255	0.190	0.050	0.0830	0.0215	0.0270	6	
20 _	0.1749	-0.0001	0.1750	i i	0.415	0.315	0.090	0.3625	0.0835	0.0930	6	
21 _	0.4159	-0.0001	0.4160	M	0.555	0.440	0.150	1.0920	0.2120	0.4405	15	
522	0.1004	-0.0001	0.1005		0.380	0.285	0.090	0.2305	0.0390	0.0775	7	
23	0.3905	-0.0000	0.3905	F	0.570	0.435	0.140	0.8585	0.1960	0.2295	8	
24 _	0.0765	0.0000	0.0765	1	0.330	0.260	0.080	0.1900	0.0385	0.0650	7	
325	0.3120	0.0000	0.3120	l I	0.550	0.440	0.165	0.8605	0.1690	0.3000	17	
526 .	0.0260	0.0000	0.0260	i i	0.230	0.175	0.065	0.0645	0.0105	0.0200	5	
527	0.2061	0.0001	0.2060	F	0.460	0.365	0.125	0.4785	0.1045	0.1410	8	
528 .	0.0252	0.0002	0.0250		0.225	0.160	0.045	0.0465	0.0150	0.0150	4	
529 _	0.2027	0.0002	0.2025	F	0.445	0.335	0.110	0.4355	0.1095	0.1195	6]
630 _	0.2552	0.0002	0.2550	i i	0.465	0.355	0.120	0.5805	0.0915	0.1840	8	
531 _	0.1863	0.0003	0.1860	M	0.460	0.375	0.135	0.4935	0.0845	0.1700	12	
532	0.1864	0.0004	0.1860	F	0.475	0.400	0.115	0.5410	0.1025	0.2100	13	
533	0.1294	0.0004	0.1290		0.410	0.300	0.090	0.3040	0.0710	0.0955	8	
534 _	0.2949	0.0004	0.2945		0.530	0.405	0.130	0.6615	0.1395	0.1900	9	
535	0.2484	0.0004	0.2480		0.505	0.395	0.105	0.5510	0.1030	0.1710	8	
636	0.7179	0.0004	0.7175	F	0.685	0.535	0.175	1.5845	0.3775	0.4215	9	
537 。	0.0125	0.0005	0.0120	l l	0.185	0.130	0.045	0.0290	0.0075	0.0095	4	
638	0.2305	0.0005	0.2300	F	0.465	0.350	0.125	0.4820	0.1060	0.1095	6	
639	0.1610	0.0005	0.1605	M	0.450	0.340	0.130	0.3715	0.0795	0.1050	9	

- Performance Evaluation
 - MSE(Mean Squared Error)
 - RMSE(Root Mean Squared Error)
 - MAE(Mean Absolute Error)
 - R2(R Squared)

MSE, RMSE, MAE는 0에 가까울수록, R2는 1에 가까울수록 정확도가 높음

Conclusion

- 전복의 나이테, 성별, 길이, 직경, 두께, 전체 무게, 내장 무게, 껍질 무게에 해당하는 총 8개의 변수와 전복 순살(flesh)의 상관 관계를 분석함
- Linear regression을 이용하여 8가지 변수 값에 따른 전복의 순살 무게를 예측함

2. Star Terror Prevention Using Logistic Regression





Problem

상품평에 기록된 별(star)을 신뢰할 수 있을까?
 신뢰할 수 없다면, AI가 신뢰할 수 있는 별을 제시할 수 있을까?



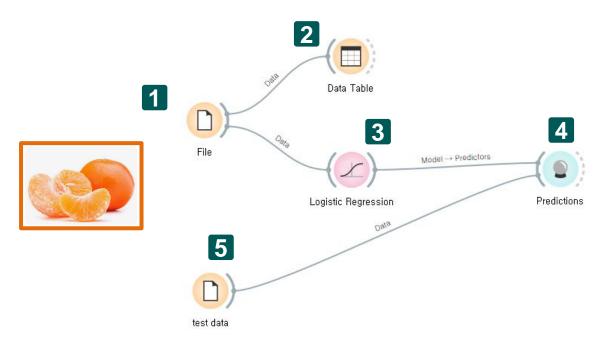


Data Type

Al Model

Structured data

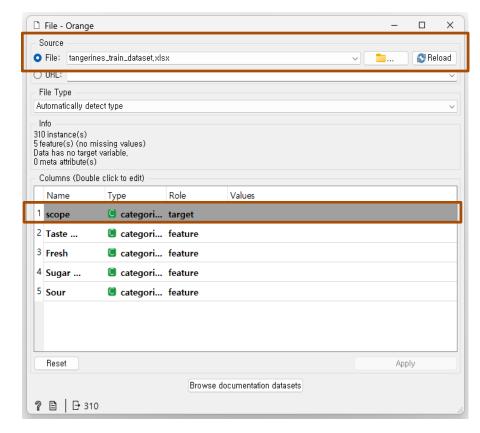
Multinomial Logistic Regression



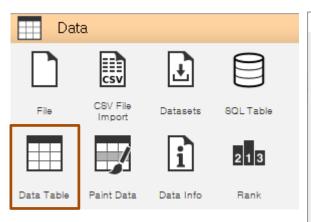
1 Datasets

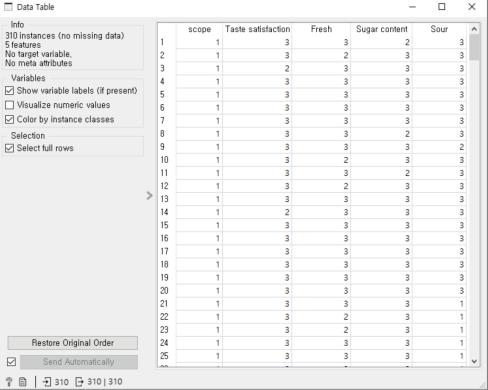


- tangerines_train_dataset.xlsx (310 instances)
- tangerines_test_dataset.xlsx
 (15 instances)



2 Data Table



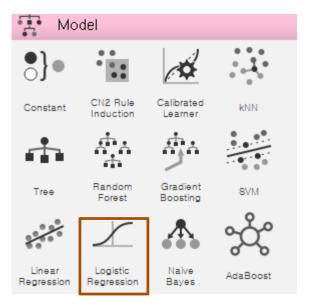


2 Data Table

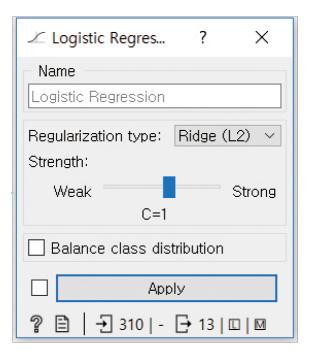


Attribute name	Attribute		
scope	종합 평점 (1~5: 점수가 높을수록 좋은 상품)		
Taste satisfaction	맛 만족도 (1: 예상보다 맛있어요, 2: 괜찮아요, 3: 예상보다 맛이 없어요.)		
Fresh	싱싱함 (1: 예상보다 싱싱해요, 2: 보통이에요, 3: 예상보다 싱싱하지 않아요.)		
Sugar content	당도 (1: 아주 달콤해요, 2: 적당히 달아요, 3: 달지 않아요.)		
Sour	새콤함 (1: 많이 새콤해요, 2: 적당히 새콤해요, 3: 새콤하지 않아요.)		

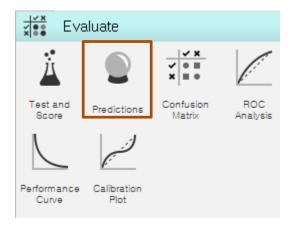
3 Logistic Regression

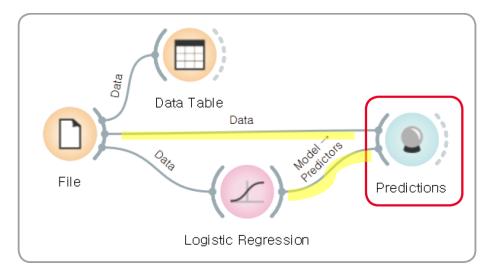


- Regularization
 - Ridge: 분류를 위한 식의 가중치 제곱의 합
 - Lasso: 분류를 위한 식의 가중치 절댓값의 합
 - Week Strong: 데이터를 분류할 때의 강도

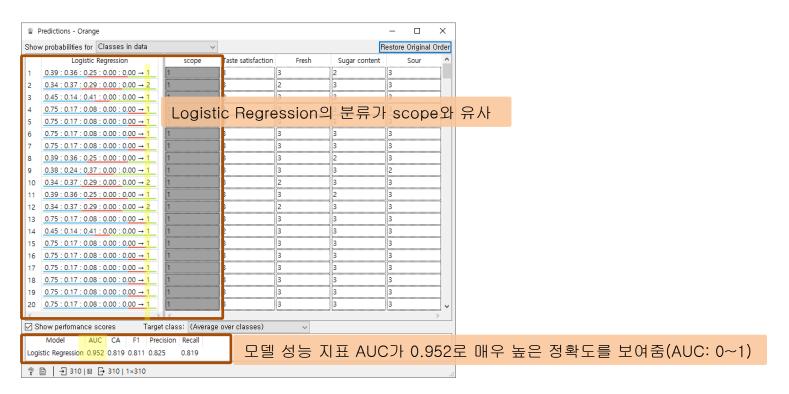


Predictions





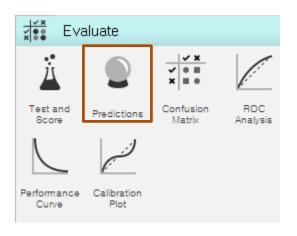
Predictions

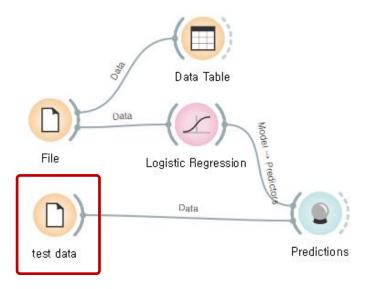


Ex) Star Terror

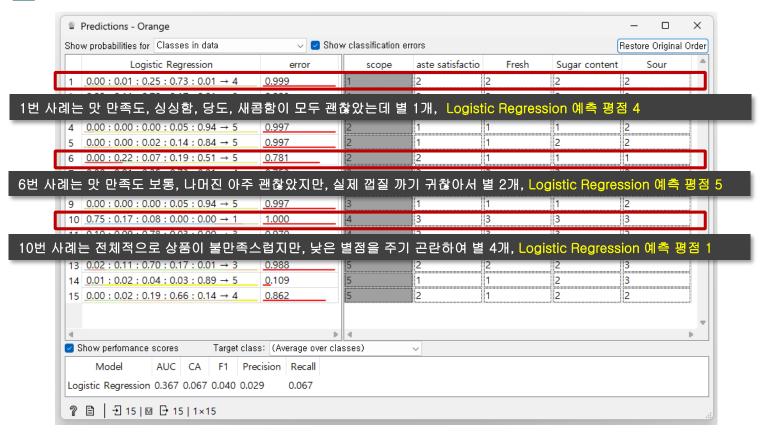


Predictions





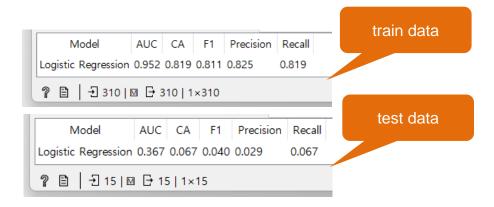
4 Predictions



딥러닝 모델 평가 지표

Confusion Matrix

		예측값			
		Positive	Negative		
실젯값	True	TP	FN		
	False	FP	TN		



Positive: 1로 예측, Negative: 0으로 예측

- TP(True Positive): 실젯값이 True인 것을 Positive라고 예측
- TN(True Negative): 실젯값이 False인 것을 Negative라고 예측
- FP(False Positive): 실젯값이 False인 것을 Positive라고 예측
- FN(False Negative): 실젯값이 True인 것을 Negative라고 예측

딥러닝 모델 평가 지표

지표	의미	그래프또는식
AUC (Area Under the ROC Curve)	재현율(Recall, 실제 True인 것 중에서 모델이 True라고 분류한 것)과 위양성률(Fall-out, 실제 False인 것 중에서 모델이 True라고 분류한 것의 비율 관계를 나타낸 ROC(Receiver Operating Characteristic) 그래프의 아래쪽 면적	재현율 (Recall) (Area Under the Curve) 이 위앙성률 (Fall-out)
분류 정확도 (CA)	모델이 입력된 데이터에 대해 얼마나 정확하게 분류 하는지를 나타내는 값 (1에 가까울수록 정확도가 높음)	$\frac{TP + TN}{TP + FP + TN + FN}$
정밀도 (Precision)	모델이 True라고 분류한 것 중에서 실제 True인 것의 비율	$\frac{TP}{TP + FP}$
재현율 (Recall)	실제 True인 것 중에서 모델이 True라고 분류한 것의 비율	$\frac{TP}{TP + FN}$
F1	정밀도와 재현율, 두 값의 조화 평균으로 하나의 수치로 나타낸 지표	

Conclusion

- 맛 만족도, 싱싱함, 당도, 새콤함 등의 변수가 모두 좋더라도 별 개수가 작게 나올 수 있음이 분석되었음
- 사용자의 별 평가 개수가 전체적인 맛 평가와 유사하다고 할 수 없음
- Logistic Regression은 맛 만족도, 싱싱함, 당도, 새콤함 등의 변수를 기반으로 전체적인 맛 평가 결과를 신뢰할 수 있게 예측함

3. Courier delivery location clustering using k-Means

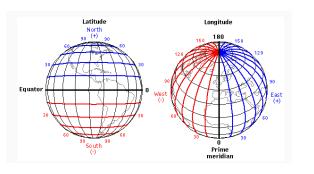




Problem

- 주소지 중심으로 택배를 분류하면, 인근 거리임에도 행정 구역상 주소지가
 다를 경우 택배 배달원의 배달 업무가 비효율적으로 수행될 수 있음
- AI가 택배 배달원의 효율적인 배달 업무를 위해서 주소지 중심으로 택배를 분류하지 않고, 인근 거리 위주로 택배 배달 물품을 분류할 수 있을까?

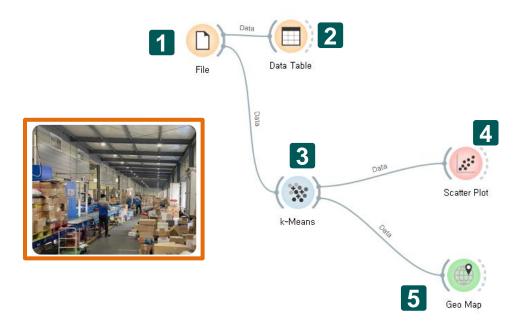




Data Type Structured data

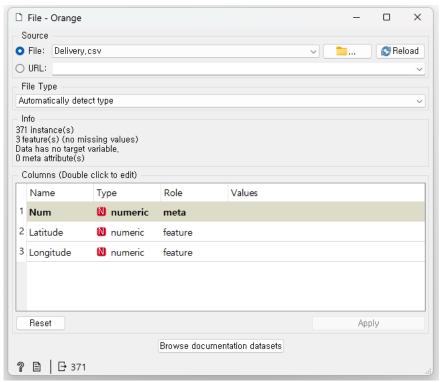
Al Model

K-Means



1 File

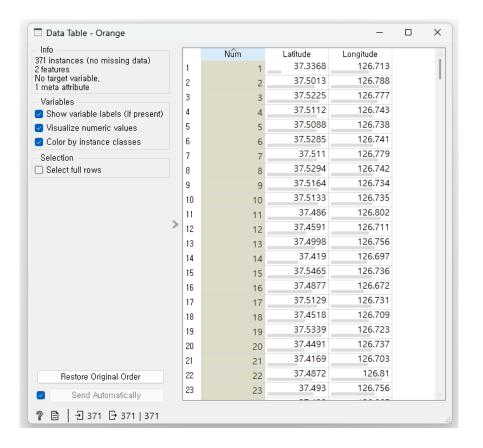




3. Courier delivery location clustering using k-Means

2 Data Table



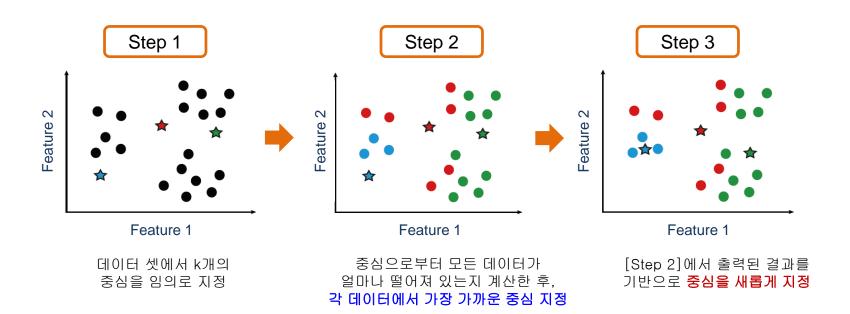


2 Data Table

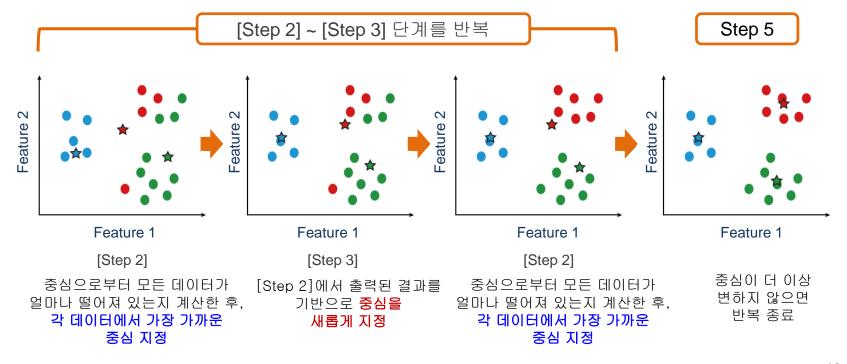


Attribute name	Attribute
Num	일련번호
Latitude	위도
Longitude	경도

Clustering process of k-Means

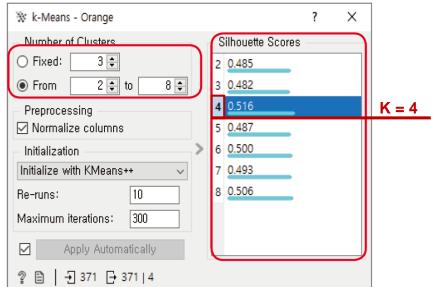


Clustering process of k-Means



3 K-Means

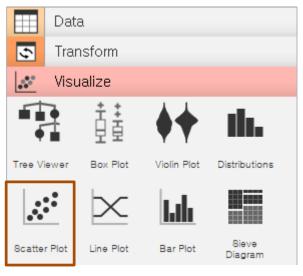


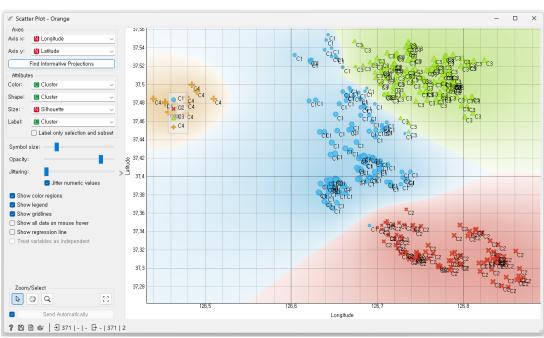


Number of Clusters

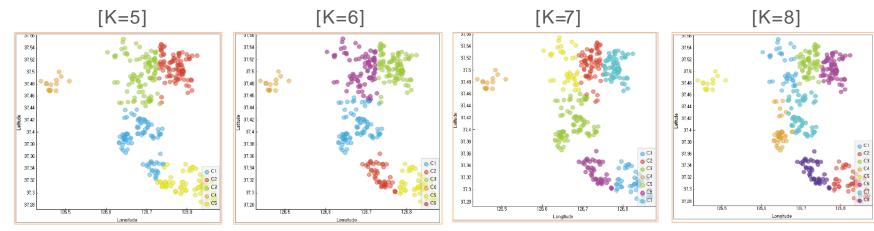
- Fixed: 원하는 cluster의 개수를 설정
- From: Silhouette Scores를 보여 주는 범위를 설정
 [Silhouette Scores] 해당 범위 내에서 가장 높은 점수의 cluster 개수를 추천

4 Scatter Plot

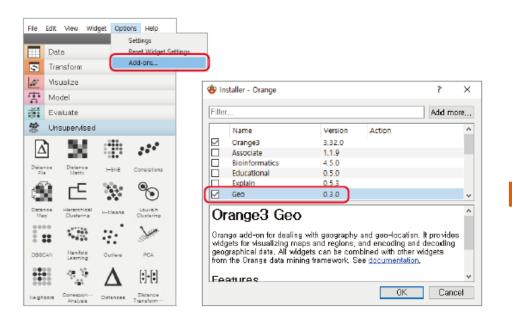




4 Scatter Plot

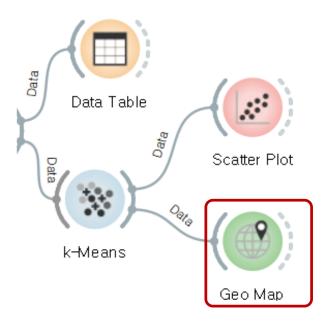


Geo Map



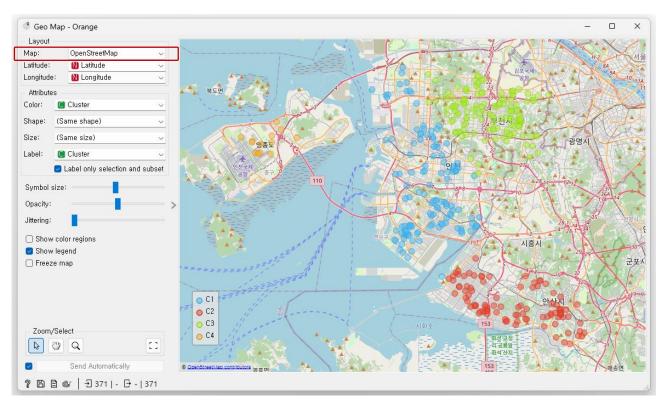


Geo Map



5 Geo Map

- Map
- OpenStreetMap
- Black and white
- Topographic
- Satellite
- Print
- Dark



Conclusion

 인천광역시, 안산시 등의 행정 구역 기반의 택배 거점을 클러스터 기반으로 변경한다면, 비용을 절약할 뿐만 아니라 택배 집하장 선정 등을 최적화할 수 있을 것으로 기대함



