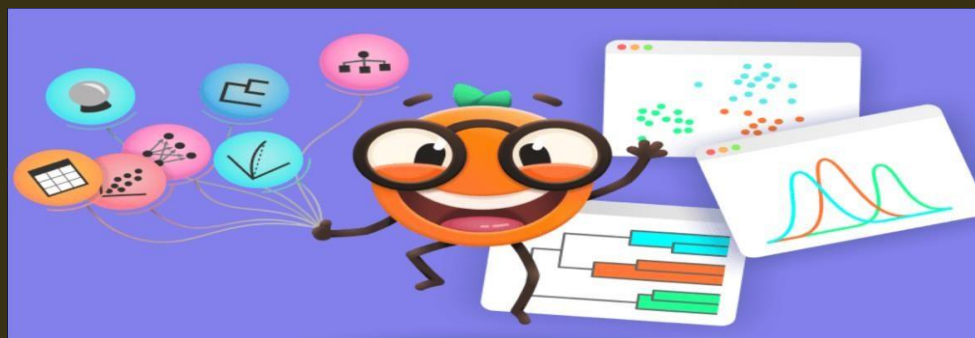


상품 평점 분류하기

Logistic Regression



소프트웨어융합대학원
진혜진

■ 해결해야 할 문제

- 세부 평점이 좋음에도 불구하고 터무니없는 이유로 총 평점을 나쁘게 주는 별점 테러를 방지하는 인공지능 모델을 만들어보자.

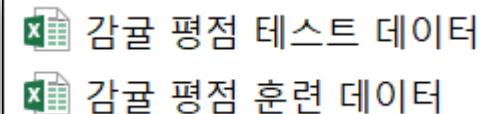
■ 데이터 준비하기

■ 데이터

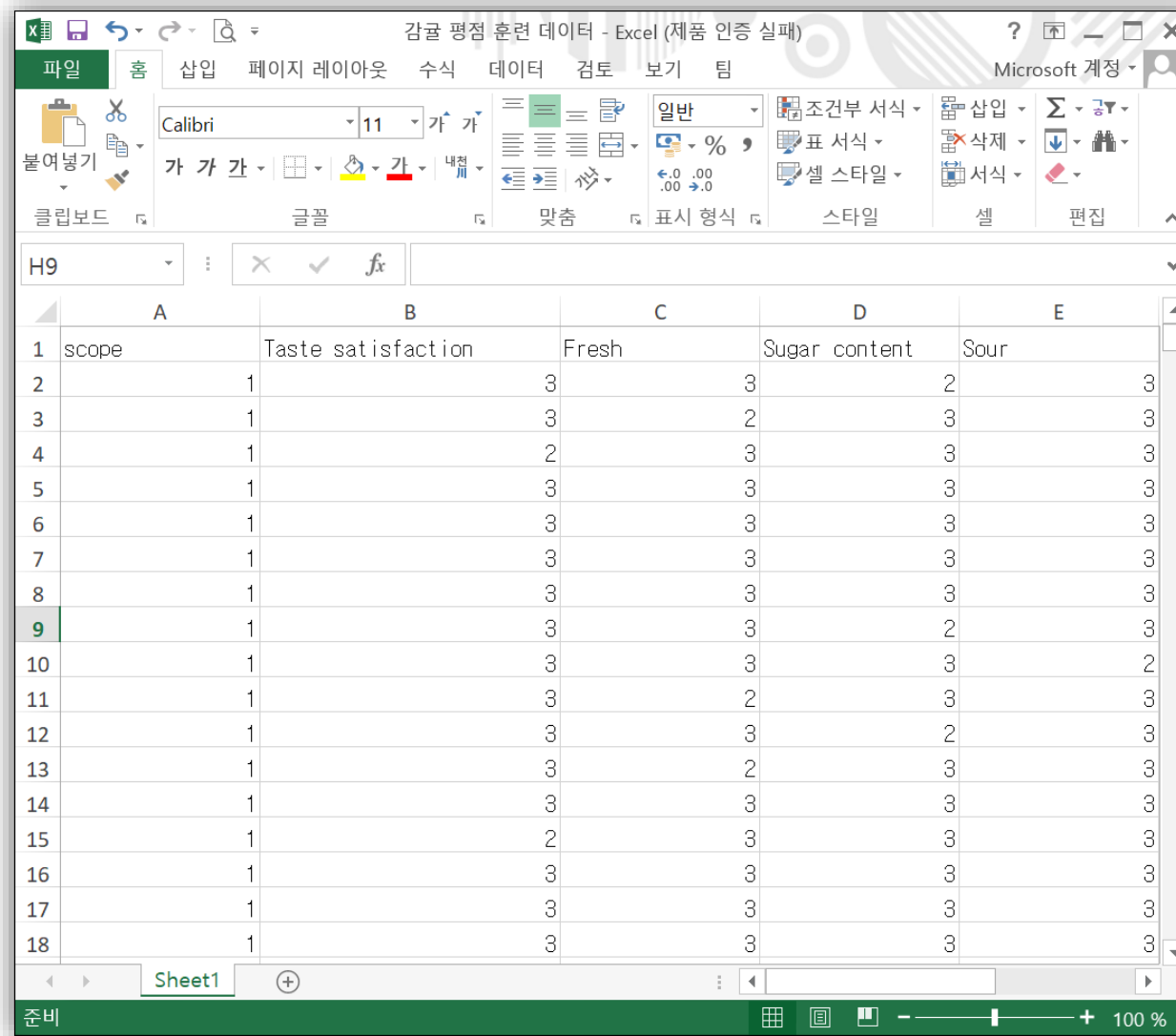
- 감귤 평점 훈련 데이터 : 실제 C사 감귤 판매 데이터 310개 수집한 것
- 감귤 평점 테스트 데이터 : 모델을 만든 후 테스트 하기 위한 데이터 15개를 별도로 수집한 것

■ 외부 데이터 다운로드

- 훈련 데이터와 테스트 데이터를 각각 다운로드 한다.



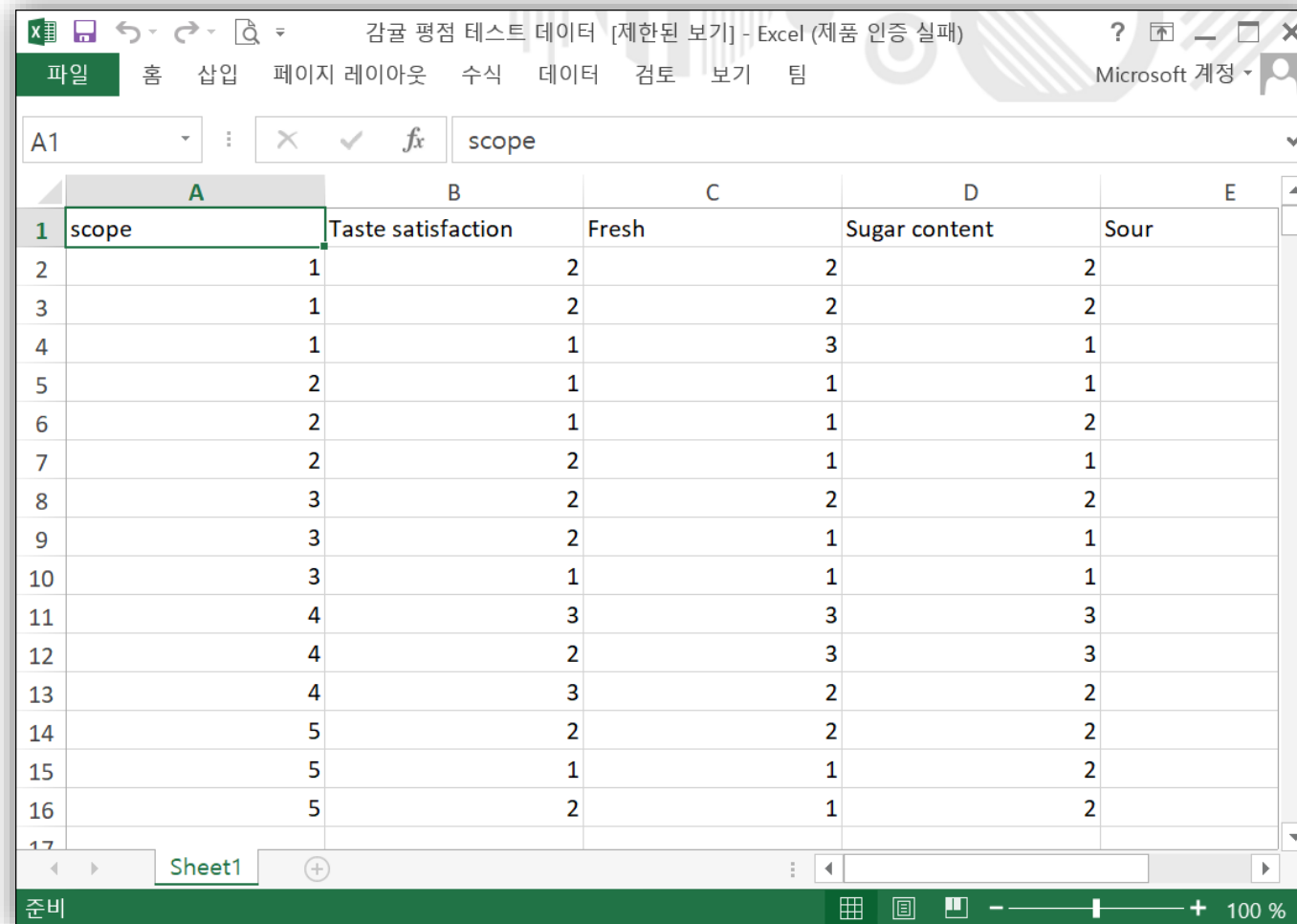
■ 감귤 평점 훈련 데이터



감귤 평점 훈련 데이터 - Excel (제품 인증 실패)

	A	B	C	D	E
1	scope	Taste satisfaction	Fresh	Sugar content	Sour
2	1	3	3	2	3
3	1	3	2	3	3
4	1	2	3	3	3
5	1	3	3	3	3
6	1	3	3	3	3
7	1	3	3	3	3
8	1	3	3	3	3
9	1	3	3	2	3
10	1	3	3	3	2
11	1	3	2	3	3
12	1	3	3	2	3
13	1	3	2	3	3
14	1	3	3	3	3
15	1	2	3	3	3
16	1	3	3	3	3
17	1	3	3	3	3
18	1	3	3	3	3

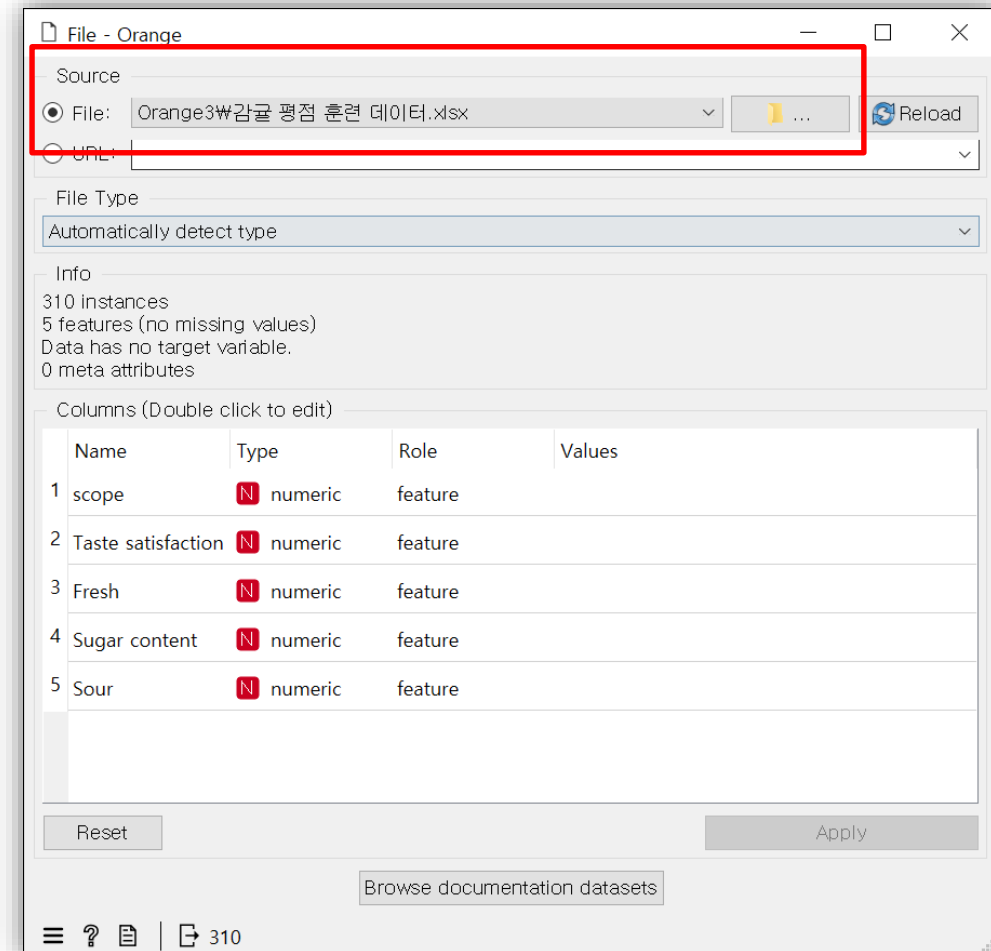
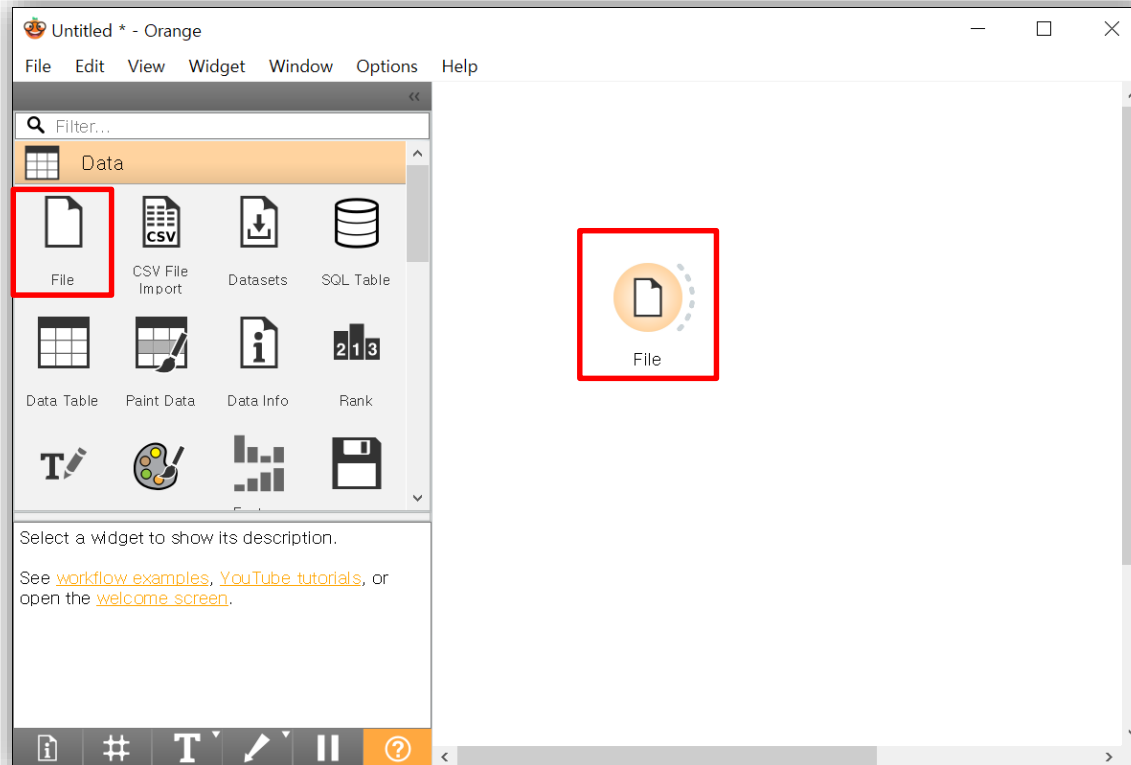
■ 감귤 평점 테스트 데이터



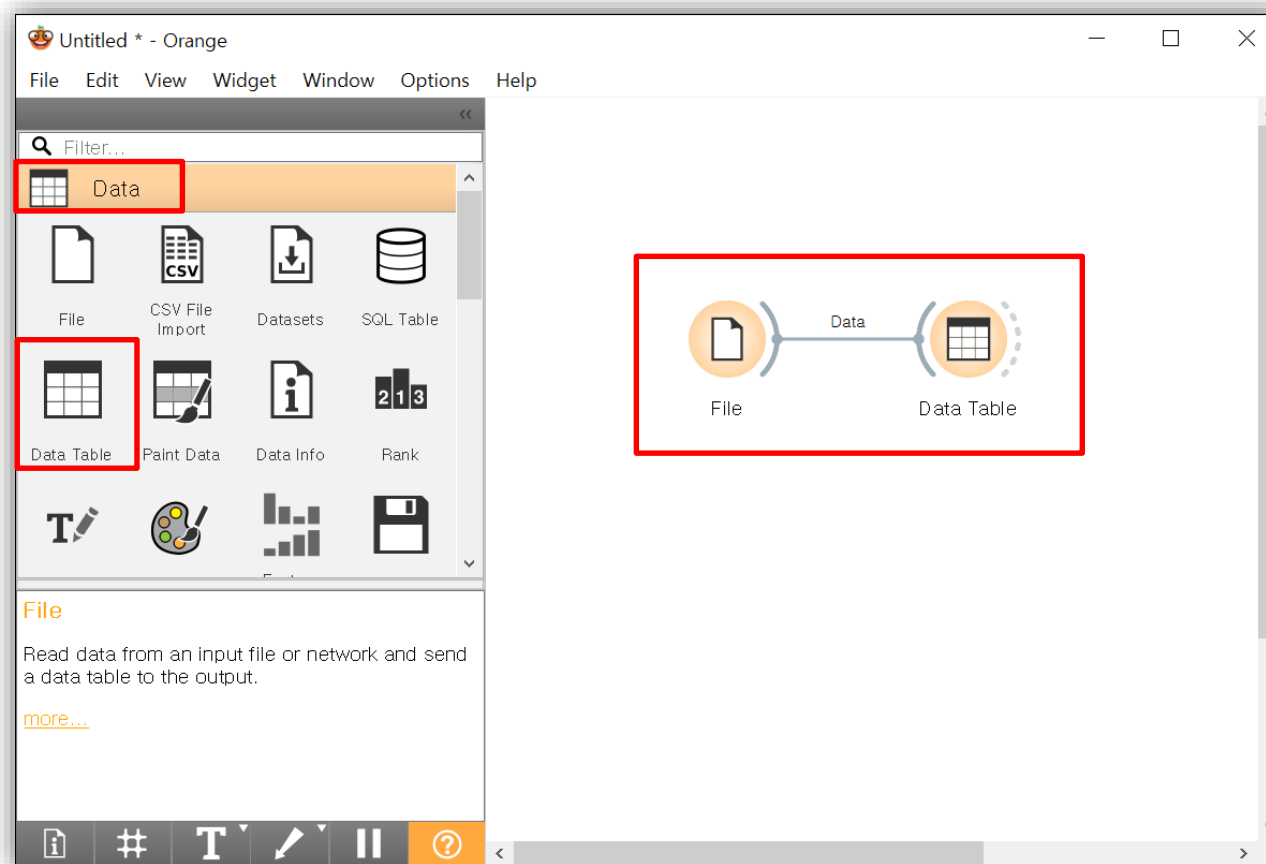
	A	B	C	D	E
1	scope	Taste satisfaction	Fresh	Sugar content	Sour
2	1	2	2	2	2
3	1	2	2	2	2
4	1	1	3	1	1
5	2	1	1	1	1
6	2	1	1	1	2
7	2	2	1	1	1
8	3	2	2	2	2
9	3	2	1	1	1
10	3	1	1	1	1
11	4	3	3	3	3
12	4	2	3	3	3
13	4	3	2	2	2
14	5	2	2	2	2
15	5	1	1	1	2
16	5	2	1	1	2

■ 데이터 불러오기

- Data 카테고리에서 File 위젯을 캔버스로 가져와서 더블 클릭한 후 훈련 데이터 파일을 연다.



- Data 카테고리에서 [Data Table] 위젯을 가져와서 [File] 위젯과 연결한 후 더블 클릭하면, 각각 5개의 속성으로 구성된 310개의 감귤 평점 데이터 정보를 확인할 수 있다.



■ Data Table로 본 감귤 평점 데이터

Data Table - Orange

Info
310 instances (no missing data)
5 features
No target variable.
No meta attributes.

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

Selection
☒ Select full rows

Restore Original Order

☒ Send Automatically

310 | 310

	scope	Taste satisfactor	Fresh	Sugar content	Sour
1	1.0	3.0	3.0	2.0	3.0
2	1.0	3.0	2.0	3.0	3.0
3	1.0	2.0	3.0	3.0	3.0
4	1.0	3.0	3.0	3.0	3.0
5	1.0	3.0	3.0	3.0	3.0
6	1.0	3.0	3.0	3.0	3.0
7	1.0	3.0	3.0	3.0	3.0
8	1.0	3.0	3.0	2.0	3.0
9	1.0	3.0	3.0	3.0	2.0
10	1.0	3.0	2.0	3.0	3.0
11	1.0	3.0	3.0	2.0	3.0
12	1.0	3.0	2.0	3.0	3.0
13	1.0	3.0	3.0	3.0	3.0
14	1.0	2.0	3.0	3.0	3.0
15	1.0	3.0	3.0	3.0	3.0
16	1.0	3.0	3.0	3.0	3.0
17	1.0	3.0	3.0	3.0	3.0
18	1.0	3.0	3.0	3.0	3.0
19	1.0	3.0	3.0	3.0	3.0
20	1.0	3.0	3.0	3.0	3.0
21	1.0	3.0	3.0	3.0	1.0

■ 데이터 속성 정보 확인하기

■ 종합 평점(scope)

- 1~5
- 점수가 높을수록 좋은 상품

■ 맛 만족도(Taste satisfaction)

- 1: 예상보다 맛있어요. 2: 괜찮아요. 3: 예상보다 맛없어요.

■ 싱싱함(Fresh)

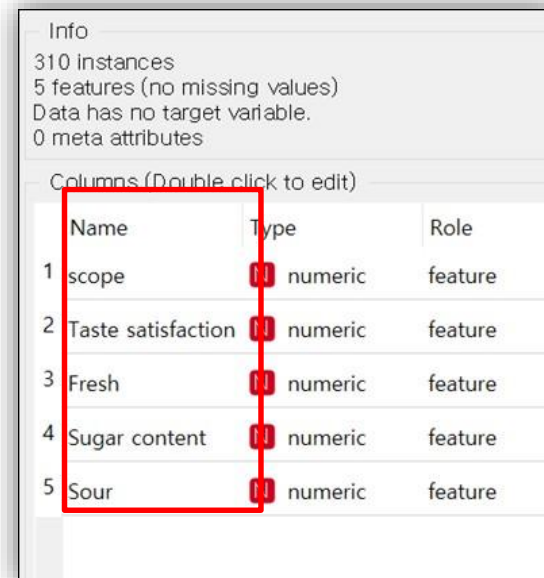
- 1: 예상보다 싱싱해요. 2: 보통이에요. 3: 예상보다 싱싱하지 않아요.

■ 당도(Sugar content)

- 1: 아주 달콤해요. 2: 적당히 달아요. 3: 달지 않아요.

■ 새콤함(Sour)

- 1: 많이 새콤해요. 2: 적당히 새콤해요. 3: 새콤하지 않아요.



Info

310 instances
5 features (no missing values)
Data has no target variable.
0 meta attributes

Columns (Double click to edit)

	Name	Type	Role
1	scope	numeric	feature
2	Taste satisfaction	numeric	feature
3	Fresh	numeric	feature
4	Sugar content	numeric	feature
5	Sour	numeric	feature

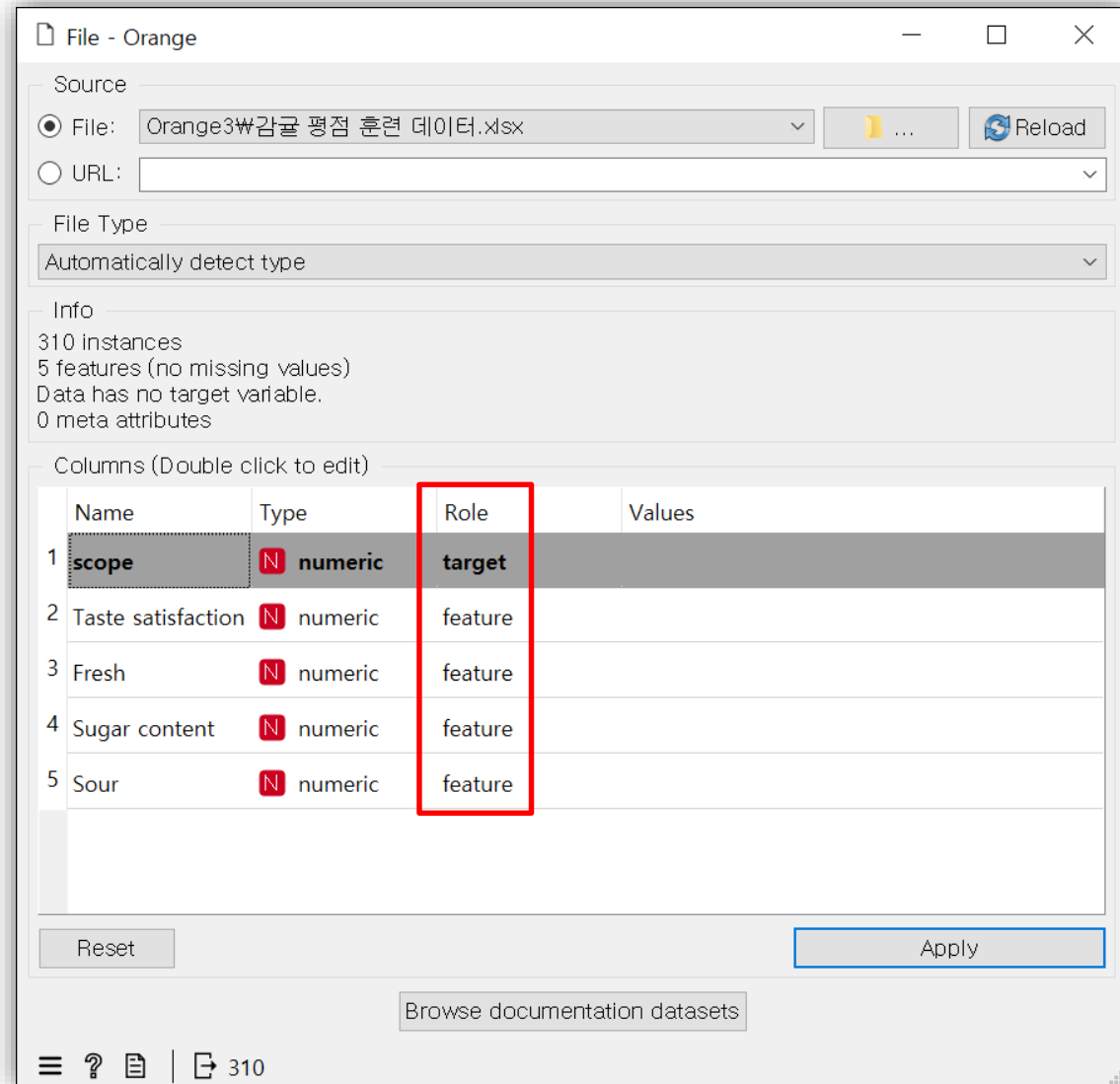
■ 데이터 전처리하기

- 데이터를 분석 및 처리에 적합한 형태로 만드는 과정

- [File] 위젯에서 가져온 데이터를 모델 학습에 사용할 때는 사용할 데이터의 역할(Role)과 형식(Type)을 변경하는 과정이 필요하다.

■ 데이터의 역할(Role) 변경하기

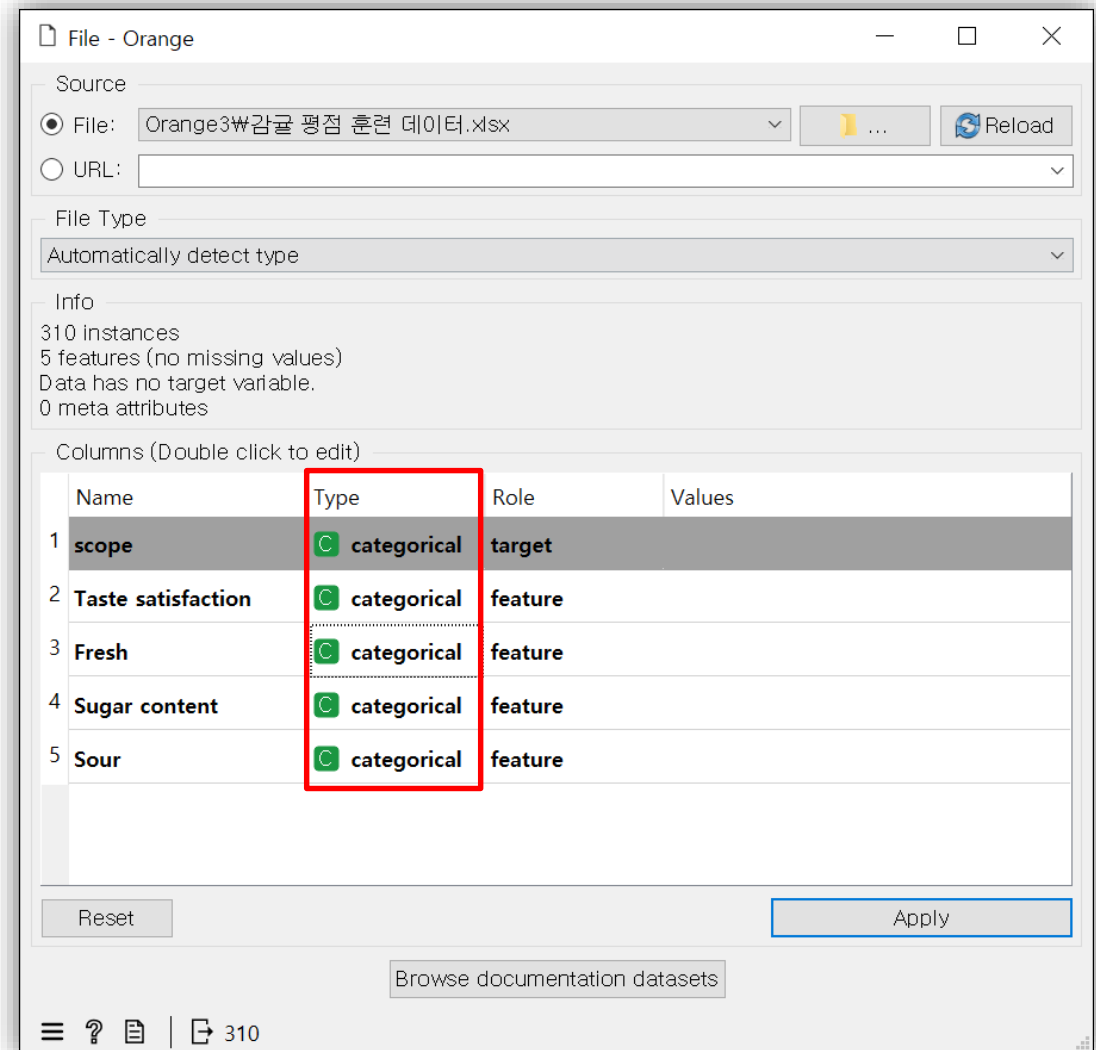
- 네 가지 속성(맛 만족도, 싱싱함, 당도, 새콤함)을 이용하여 종합 평점을 분류하는 모델을 만들어야 하므로
- [File] 위젯을 더블 클릭한 후 scope 속성만 target으로 설정하고, 나머지 속성은 feature로 설정한다.



■ 데이터 형식(Type) 변경하기

- 각 데이터 형식은 수치가 아니라 별점의 개수이므로 모두 categorical로 설정한다.

- numeric으로 설정하면 소수점까지 포함하여 1.4, 2.89 등으로 예측되기 때문
- categorical로 설정하면 1, 2, 3, 4, 5로 예측할 수 있다.

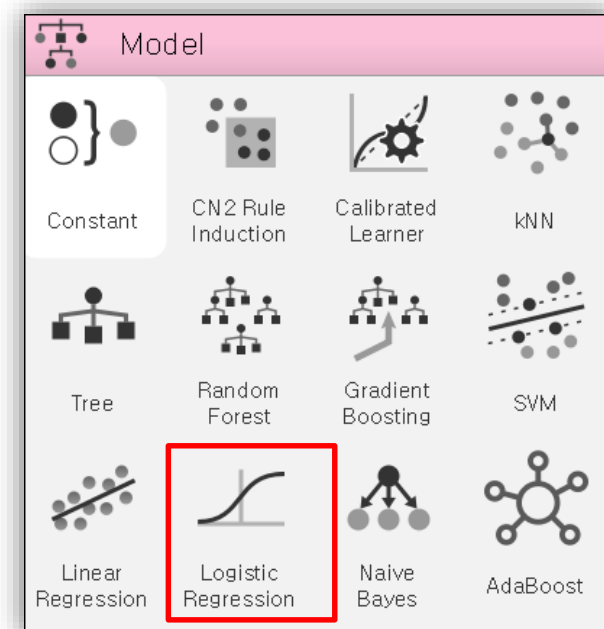


■ 어떤 모델을 선택하고 학습시킬까?

- 분류 모델 중 Logistic Regression(로지스틱 회귀) 모델을 이용

■ Logistic Regression

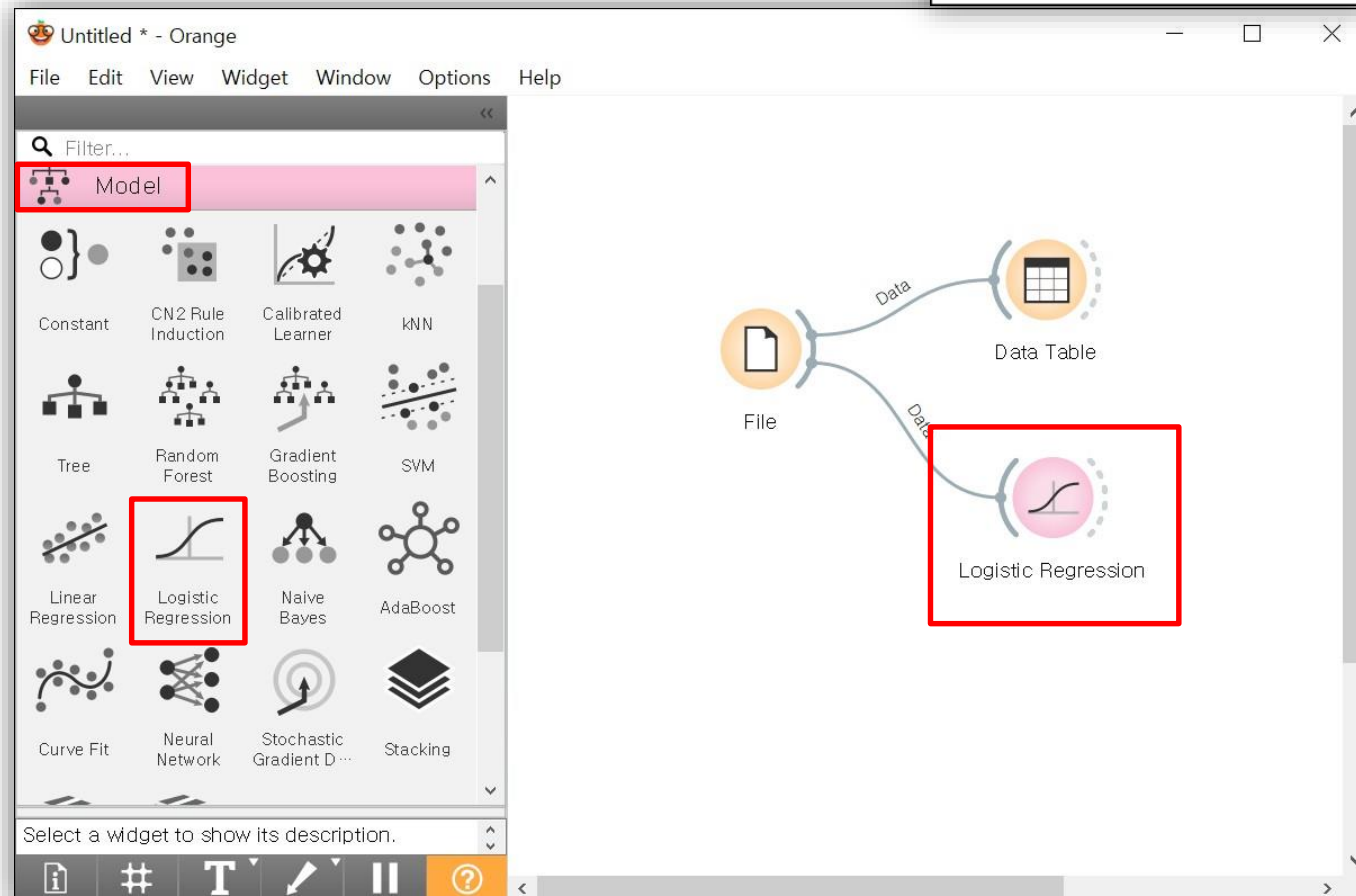
- 입력 값을 결합하여 어떤 사건의 발생 가능성을 확률적으로 예측하여 분류하는 모델
- 독립 변수를 입력 받아 종속 변수의 레이블이 두 범주 중 어디에 해당하는지 분류한다.
- 스팸 메일 필터, 텍스트 분류, 감정 분석, 추천 시스템 등에 광범위하게 활용된다.



1. 학습 모델 선택하기

- Model 카테고리에서 [Logistic Regression] 위젯을 캔버스로 가져와서 [file] 위젯과 연결한다.
- 이때 사용하는 데이터는 훈련 데이터이다.

엑셀 아이콘 감귤 평점 훈련 데이터



2. 학습시키기

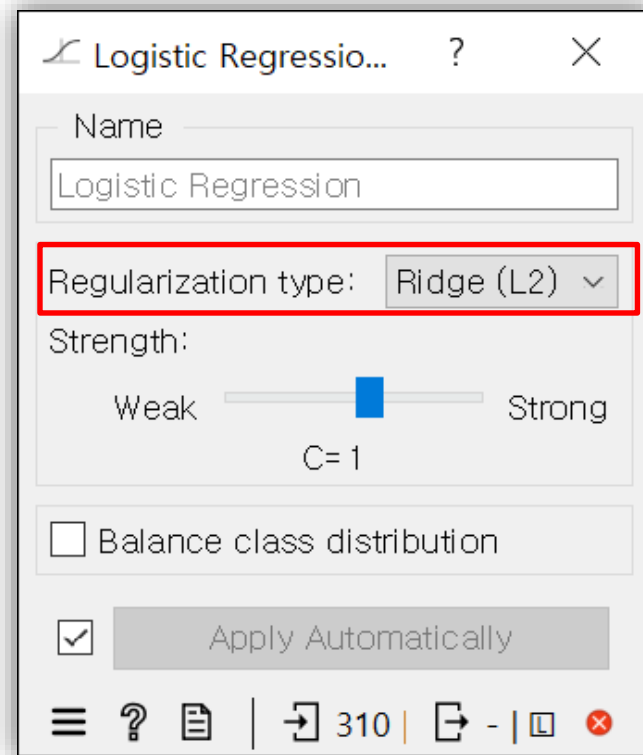
- [Logistic Regression] 위젯을 더블 클릭 한다.
- 설정 변경에 따라 인공지능 모델 성능은 달라질 수 있다.
- 데이터 특성에 맞게 설정하도록 한다.

• 과적합 방지를 위한 정규화

- Ridge : 분류를 위한 식의 가중치 제곱의 합
- Lasso : 분류를 위한 식의 가중치 절대값의 합
- Weak와 Strong : 데이터를 분류할 때의 강도

• 학습 데이터에 모델이 과적합되는 현상

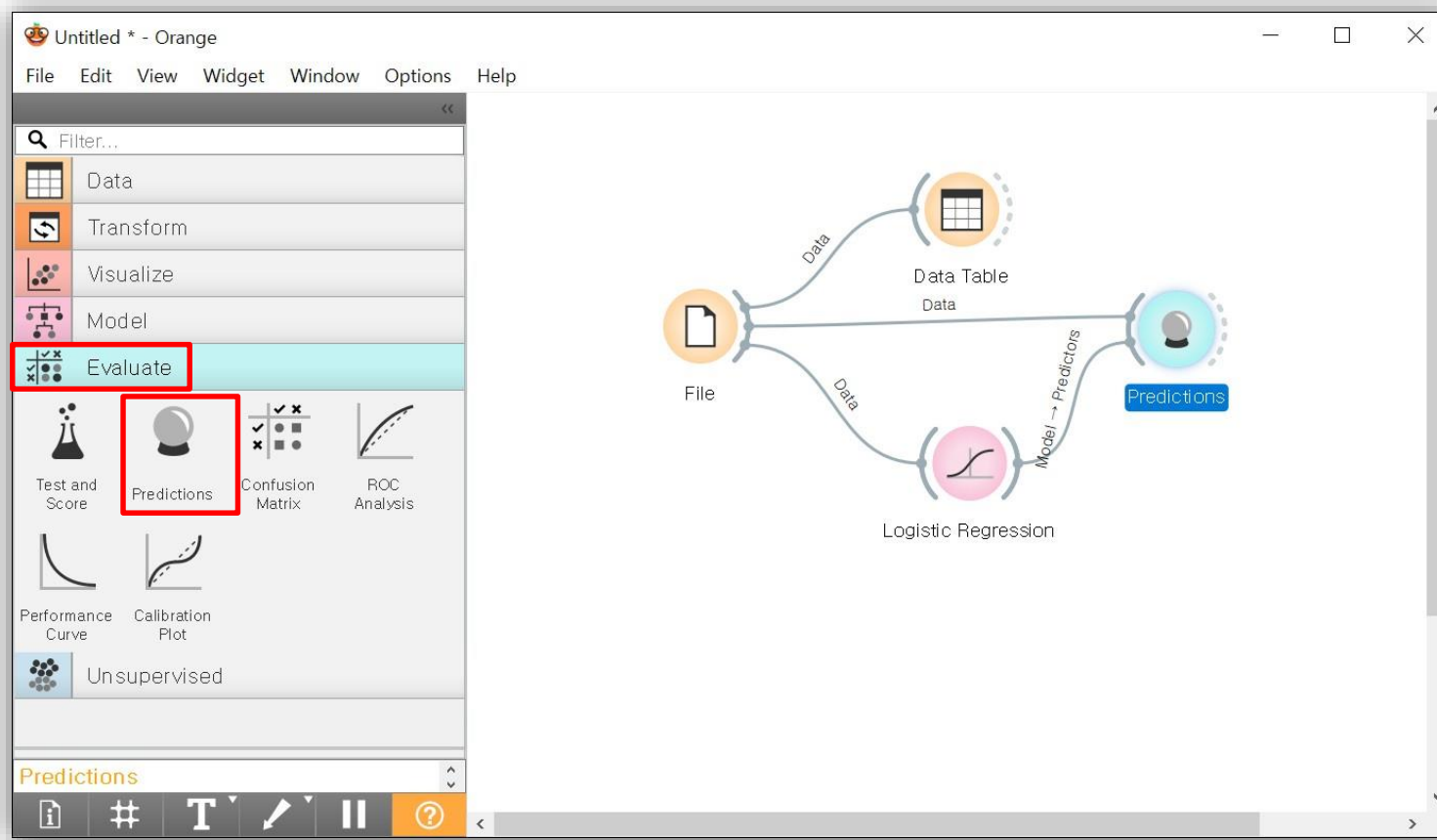
- 모델의 성능을 떨어트리는 주요 이슈
- 모델이 과적합되면 훈련 데이터에 대한 정확도는 높을지라도
- 새로운 데이터(검증 데이터나 테스트 데이터)에 대해서는 제대로 동작하지 않는다.
- 이는 모델이 학습 데이터를 불필요할 정도로 과하게 암기하여 훈련 데이터에 포함된 노이즈까지 학습한 상태이다.



■ 모델의 성능을 확인해보자

■ [Predictions] 위젯 연결하기

- Evaluate 카테고리에서 [Predictions] 위젯을 가져와서 [File] 위젯과 [Logistic Regression] 위젯에 각각 연결한다.



로지스틱 회귀 모델 평가하기

- [Predictions] 위젯을 더블 클릭하면 결과를 살펴 볼 수 있다.
- Logistic Regression의 분류가 scope와 유사하게 나온 것을 확인할 수 있다.
- 평가지표 중 0~1 사이의 값을 갖는 모델의 성능 지표인 AUC가 0.952로 매우 높은 정확도를 보여주고 있다.

Predictions - Orange

Show probabilities for: Classes in data ☒ Show classification errors Restore Original Order

	Logistic Regression	error	scope	Taste satisfaction	Fresh	Sugar content	Sour
1	0.39 : 0.36 : 0.25 : 0.00 : 0.00 → 1	0.614	1	3	3	2	3
2	0.34 : 0.37 : 0.29 : 0.00 : 0.00 → 2	0.663	1	3	2	3	3
3	0.45 : 0.14 : 0.41 : 0.00 : 0.00 → 1	0.551	1	2	3	3	3
4	0.75 : 0.17 : 0.08 : 0.00 : 0.00 → 1	0.253	1	3	3	3	3
5	0.75 : 0.17 : 0.08 : 0.00 : 0.00 → 1	0.253	1	3	3	3	3
6	0.75 : 0.17 : 0.08 : 0.00 : 0.00 → 1	0.253	1	3	3	3	3
7	0.75 : 0.17 : 0.08 : 0.00 : 0.00 → 1	0.253	1	3	3	3	3
8	0.39 : 0.36 : 0.25 : 0.00 : 0.00 → 1	0.614	1	3	3	2	3
9	0.38 : 0.24 : 0.37 : 0.00 : 0.00 → 1	0.617	1	3	3	3	2
10	0.34 : 0.37 : 0.29 : 0.00 : 0.00 → 2	0.663	1	3	2	3	3
11	0.39 : 0.36 : 0.25 : 0.00 : 0.00 → 1	0.614	1	3	3	2	3
12	0.34 : 0.37 : 0.29 : 0.00 : 0.00 → 2	0.663	1	3	2	3	3
13	0.75 : 0.17 : 0.08 : 0.00 : 0.00 → 1	0.253	1	3	3	3	3
14	0.45 : 0.14 : 0.41 : 0.00 : 0.00 → 1	0.551	1	2	3	3	3
15	0.75 : 0.17 : 0.08 : 0.00 : 0.00 → 1	0.253	1	3	3	3	3
16	0.75 : 0.17 : 0.08 : 0.00 : 0.00 → 1	0.253	1	3	3	3	3
17	0.75 : 0.17 : 0.08 : 0.00 : 0.00 → 1	0.253	1	3	3	3	3
18	0.75 : 0.17 : 0.08 : 0.00 : 0.00 → 1	0.253	1	3	3	3	3

☒ Show performance scores Target class: (Average over classes)

Model	AUC	CA	F1	Prec	Recall	MCC
Logistic Regression	0.952	0.819	0.811	0.825	0.819	0.735

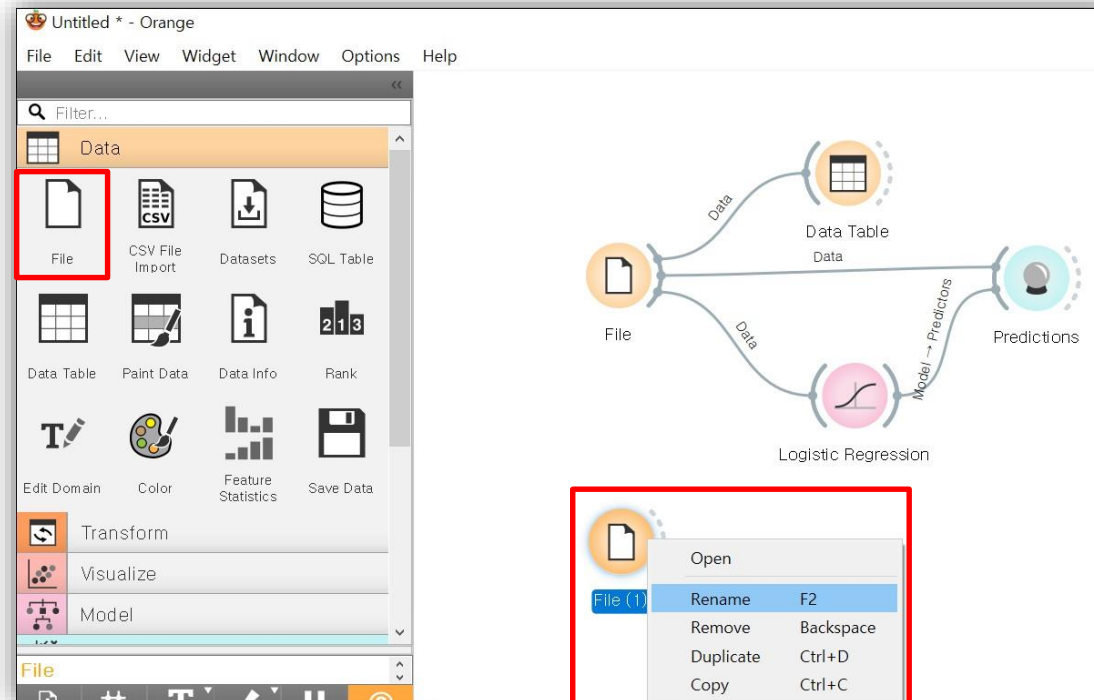
310 | 310 | 1x310

■ 성능 결과 확인하기

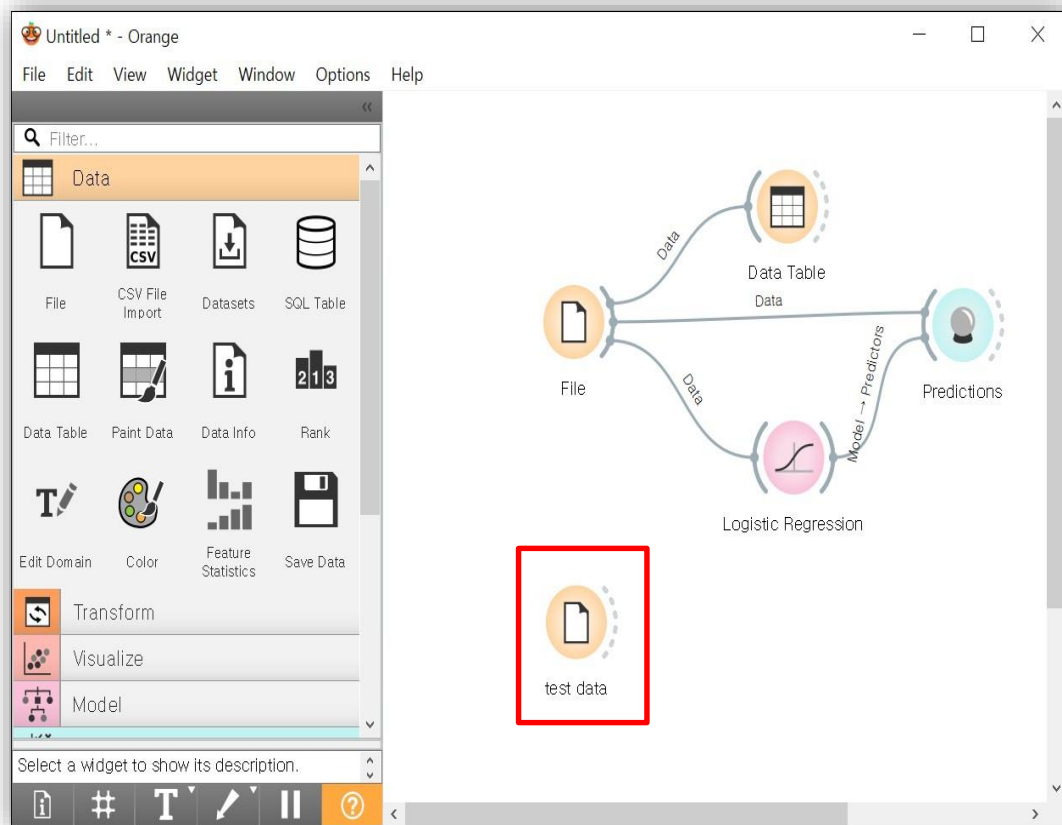
- 실제 별점 테러를 당한 몇 개의 평점들을 조사한 테스트 데이터인 감귤 평점 테스트 데이터.xlsx로 별점 테러를 방지해보자.

• 테스트 데이터 불러오기 감귤 평점 테스트 데이터

- 실제 맛 만족도, 싱싱함, 당도, 새콤함 정도는 우수하지만, 껍질을 까지 귀찮다는 이유만으로 종합 평점이 낮은 사례가 있다.
- 이러한 사례를 앞서 만든 분류 모델에 적용하기 위해 테스트 데이터인 감귤 평점 테스트 데이터.xlsx를 불러오자.



- [File] 위젯으로 테스트 데이터를 불러와 데이터의 형식과 역할을 변경한다.



test data - Orange

Source

☒ File: Orange3₩감굴 평점 테스트 데이터.xlsx

☐ URL:

File Type

Automatically detect type

Info

15 instances
5 features (no missing values)
Data has no target variable.
0 meta attributes

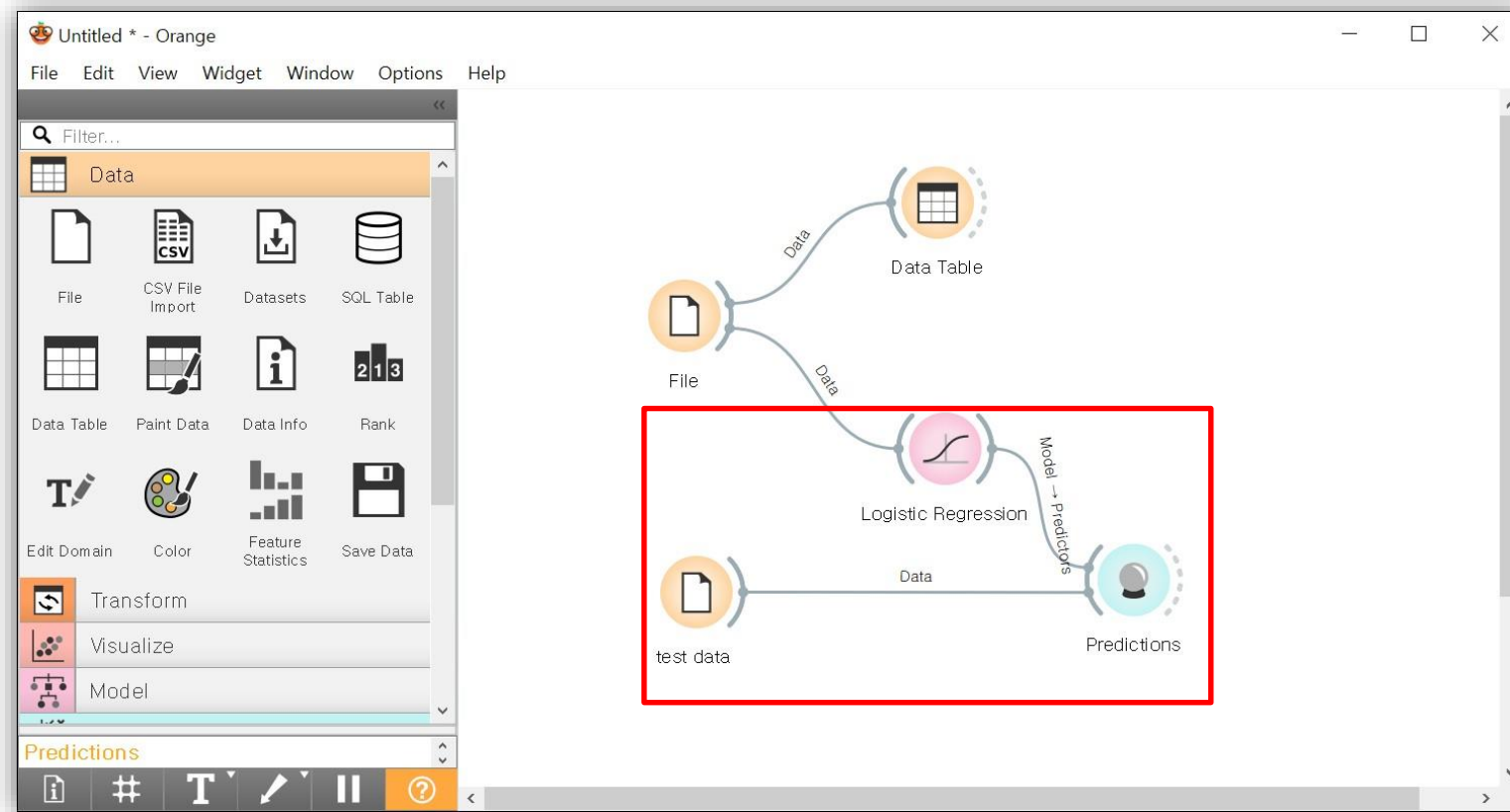
Columns (Double click to edit)

	Name	Type	Role	Values
1	scope	Ⓢ categorical	target	
2	Taste ...	Ⓢ categorical	feature	
3	Fresh	Ⓢ categorical	feature	
4	Sugar content	Ⓢ categorical	feature	
5	Sour	Ⓢ categorical	feature	

≡ ? 📄 | 📄 15

■ 결과 확인하기

- [Predictions] 위젯을 더블 클릭하면 예측 값을 확인할 수 있다.



- 맛 만족도, 싱싱함, 당도, 새콤함이 모두 괜찮았는데 분류 모델로 적용해 보니 평점 4가 나왔다.
- 맛 만족도만 보통이었고 나머지는 아주 괜찮았지만, 실제 껍질을 까지 귀찮아서 별점 2를 주어 분류 모델에서 평점 5가 나왔다.
- 이렇게 우리는 실생활에서 데이터를 직접 수집하여 인공지능 분류 모델을 만든 다음, 이를 통해 주어진 문제를 해결할 수 있음을 알 수 있다.

Predictions - Orange

Show probabilities for (None) ☒ Show classification errors [Restore Original Order](#)

	Logistic Regression	error	scope	Taste satisfaction	Fresh	Sugar content	Sour
1	4	<u>0.999</u>	1	2	2	2	2
2	3	<u>0.983</u>	1	2	2	2	3
3	5	<u>0.904</u>	1	1	3	1	1
4	5	<u>0.997</u>	2	1	1	1	2
5	5	<u>0.997</u>	2	1	1	2	2
6	5	<u>0.781</u>	2	2	1	1	1
7	4	<u>0.753</u>	3	2	2	2	2
8	5	<u>0.930</u>	3	2	1	1	1
9	5	<u>0.997</u>	3	1	1	1	2
10	1	<u>1.000</u>	4	3	3	3	3
11	3	<u>0.970</u>	4	2	3	3	2
12	3	<u>0.969</u>	4	3	2	2	3
13	3	<u>0.988</u>	5	2	2	2	3
14	5	<u>0.109</u>	5	1	1	2	3
15	4	<u>0.862</u>	5	2	1	2	2

☒ Show performance scores Target class: (Average over classes)

Model	AUC	CA	F1	Prec	Recall	MCC
Logistic Regression	0.367	0.067	0.040	0.029	0.067	-0.183

≡ ? | 15 | ↩ -