비지도학습(차원축소)

홍익 대학교
Hyun-Sun Ryu

차원축소 실습

차원축소 실습

Airline passenger satisfaction data
(항공사 고객 만족도 데이터)

https://www.kaggle.com/datasets/teejmahal20/airline-passenger-satisfaction

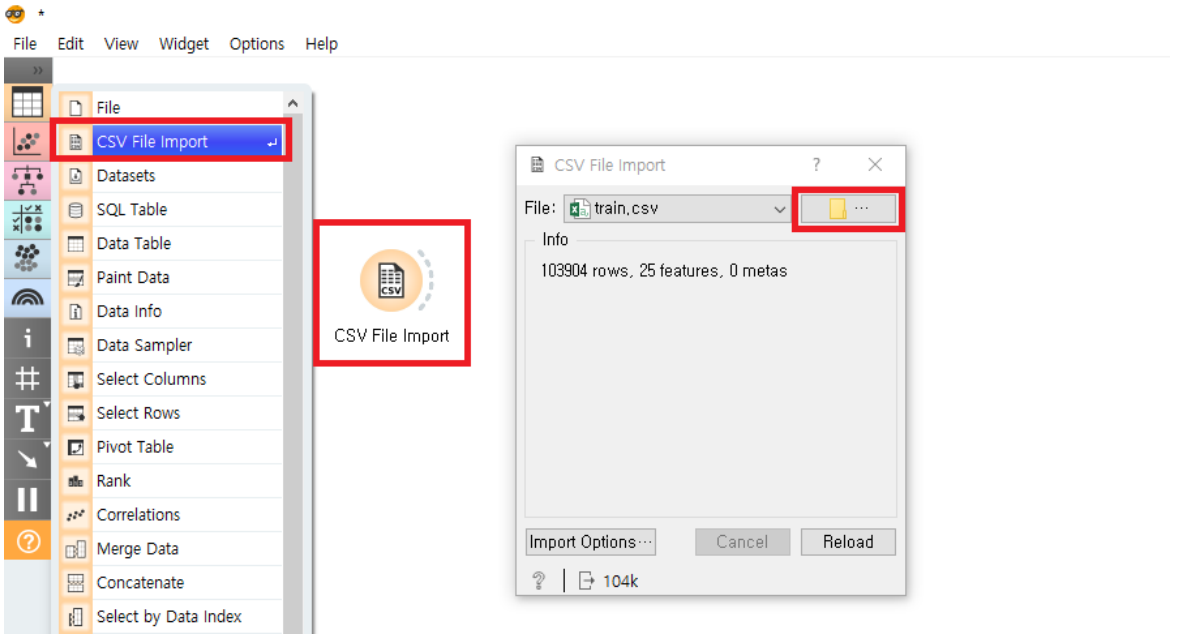
- 캐글에서 제공하는 데이터로 ‘Airline passenger satisfaction’
- 해당 데이터세트는 비행기 승객의 만족도 조사를 나타낸 데이터

Airline passenger satisfaction data

- 각 승객별 ‘만족’ 혹은 ‘불만족’ 결론에 가장 연관되는 요소가 무엇인지 파악하는 것이 이번 데이터 활용의 목표
- 이번 데이터의 경우 우리가 실습한 예제와는 달리 많은 feature값이 있으므로 차원축소를 활용한 데이터 분석이 무엇보다도 필요
- feature값이 많을수록 불필요한 데이터도 많을 뿐더러 다중공선성의 문제도 발생할 수 있어 차원축소-주성분 분석과정을 통해 이를 해결해야 함.
- 기존 데이터의 70%를 훈련데이터로, 30%를 테스트 데이터로 구성하여 무작위로 선정된 30%의 승객의 만족도를 예측
- 예측값에서 각 승객이 만족/불만족 했다면 어떤 부분에서 만족/불만족 했는지 파악

Airline passenger satisfaction data

- 캐글에서 제공하는 데이터를 불러옴
- 해당 데이터는 csv파일이므로 csv data 위젯을 활용



Airline passenger satisfaction data

Data Table

Info
103904 instances
25 features (0,0 % missing data)
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

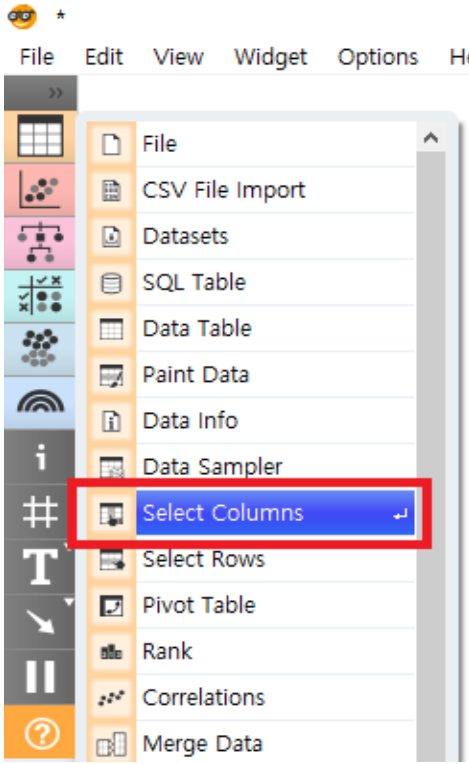
Send Automatically

104k

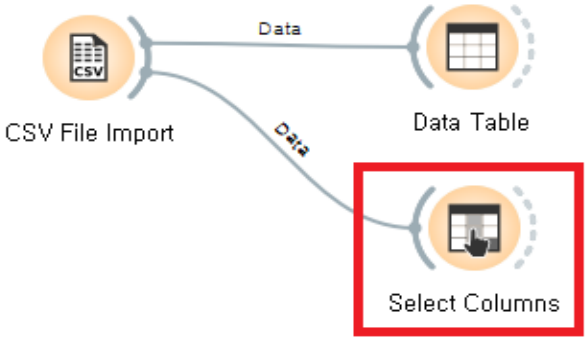
	Unnamed: 0	id	Gender	Customer Type	Age	Type of Travel	Class	Flight
1	0	70172	Male	Loyal Customer	13	Personal Travel	Eco Plus	460
2	1	5047	Male	disloyal Custo...	25	Business travel	Business	235
3	2	110028	Female	Loyal Customer	26	Business travel	Business	1142
4	3	24026	Female	Loyal Customer	25	Business travel	Business	562
5	4	119299	Male	Loyal Customer	61	Business travel	Business	214
6	5	111157	Female	Loyal Customer	26	Personal Travel	Eco	1180
7	6	82113	Male	Loyal Customer	47	Personal Travel	Eco	1276
8	7	96462	Female	Loyal Customer	52	Business travel	Business	2035
9	8	79485	Female	Loyal Customer	41	Business travel	Business	853
10	9	65725	Male	disloyal Custo...	20	Business travel	Eco	1061
11	10	34991	Female	disloyal Custo...	24	Business travel	Eco	1182
12	11	51412	Female	Loyal Customer	12	Personal Travel	Eco Plus	308
13	12	98628	Male	Loyal Customer	53	Business travel	Eco	834
14	13	83502	Male	Loyal Customer	33	Personal Travel	Eco	946
15	14	95789	Female	Loyal Customer	26	Personal Travel	Eco	453
16	15	100580	Male	disloyal Custo...	13	Business travel	Eco	486
17	16	71142	Female	Loyal Customer	26	Business travel	Business	2123

- Data table 위젯을 연결하여 해당 데이터가 어떻게 구성돼 있는지 파악
- 총 103,904명의 승객에 대한 데이터가 있으며 25개의 feature값이 있음
- 25개의 feature중 하나는 satisfaction 즉 만족도이므로 24개의 feature가 만족도에 어떤 영향을 끼치는지 파악

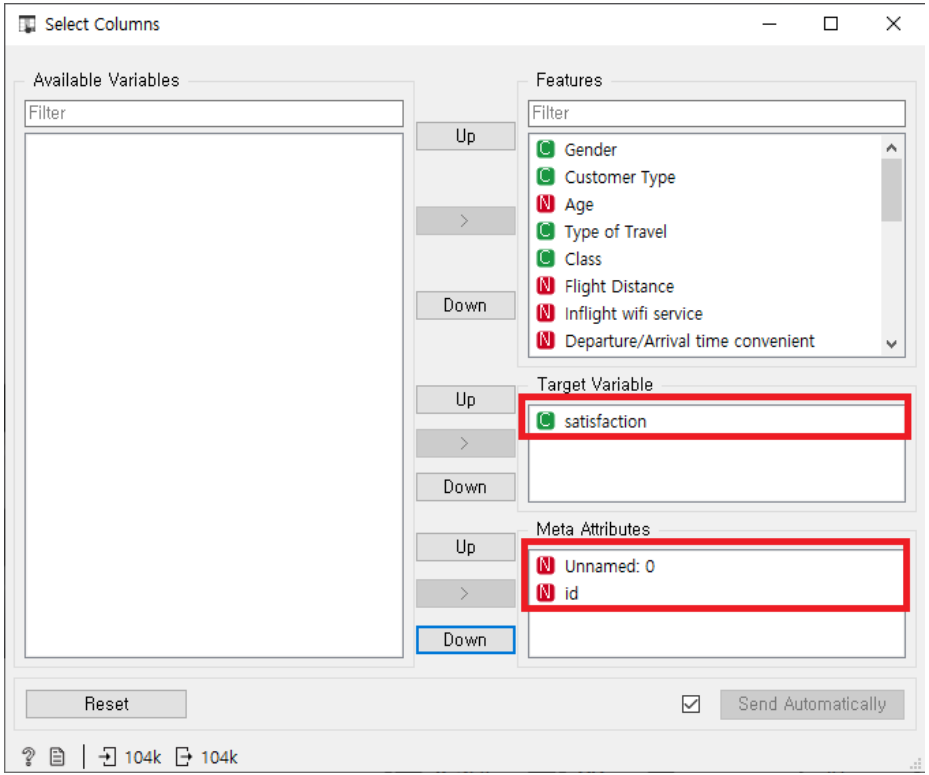
Airline passenger satisfaction data



- 각 feature의 역할(role)을 구분해주기 위해 select columns 위젯을 추가
- Select columns위젯은 데이터의 각 행의 역할을 구분

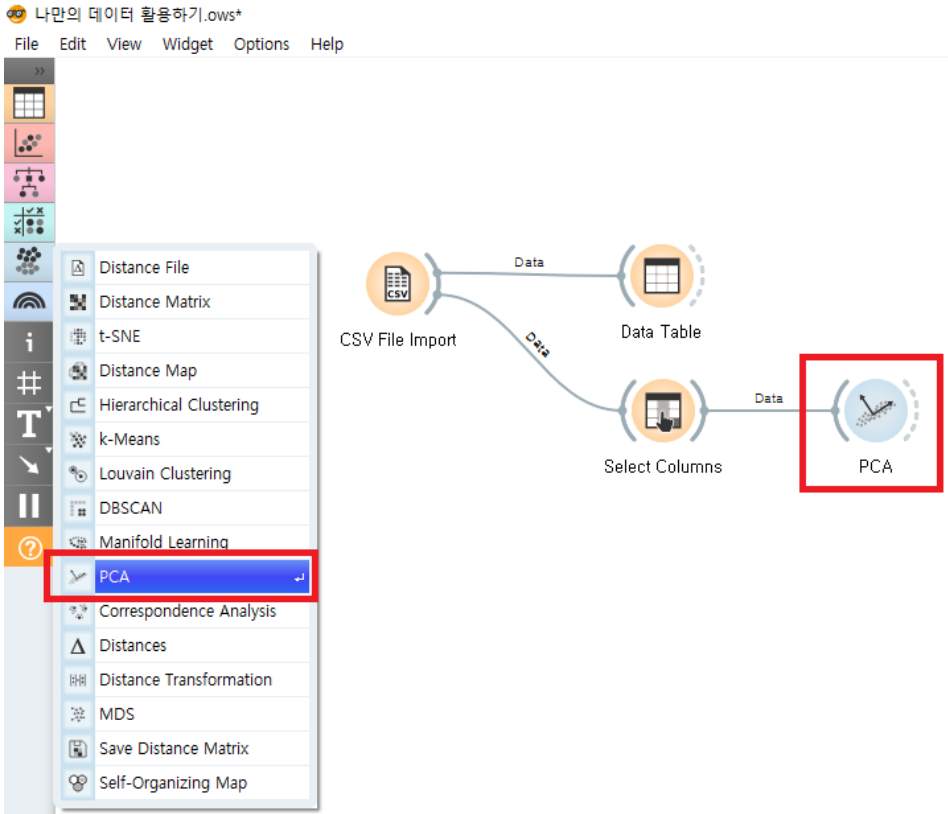


Airline passenger satisfaction data



- Select columns위젯을 실행하여 각 행의 역할을 구분
- ‘Unnamed:0’은 순번을, ‘id’는 승객의 identification을 나타내므로 meta로 분류
- ‘satisfaction’은 승객의 만족도를 나타내므로 target variable로 분류

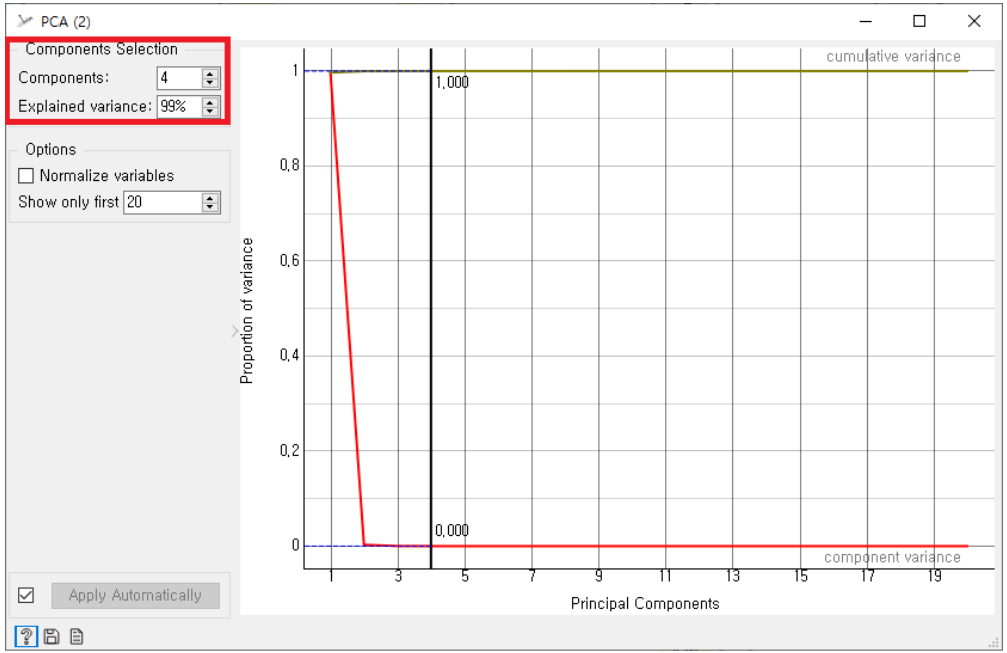
Airline passenger satisfaction data



- 24개의 feature값을 토대로 예측을 한다면 각 feature가 어떻게 작용하는지 파악하기 어려움
- 또한 다중공선성의 문제가 발생할 수 있으므로 주성분 분석을 함
- 왼쪽 unsupervised learning 메뉴에서 PCA를 추가하여 연결

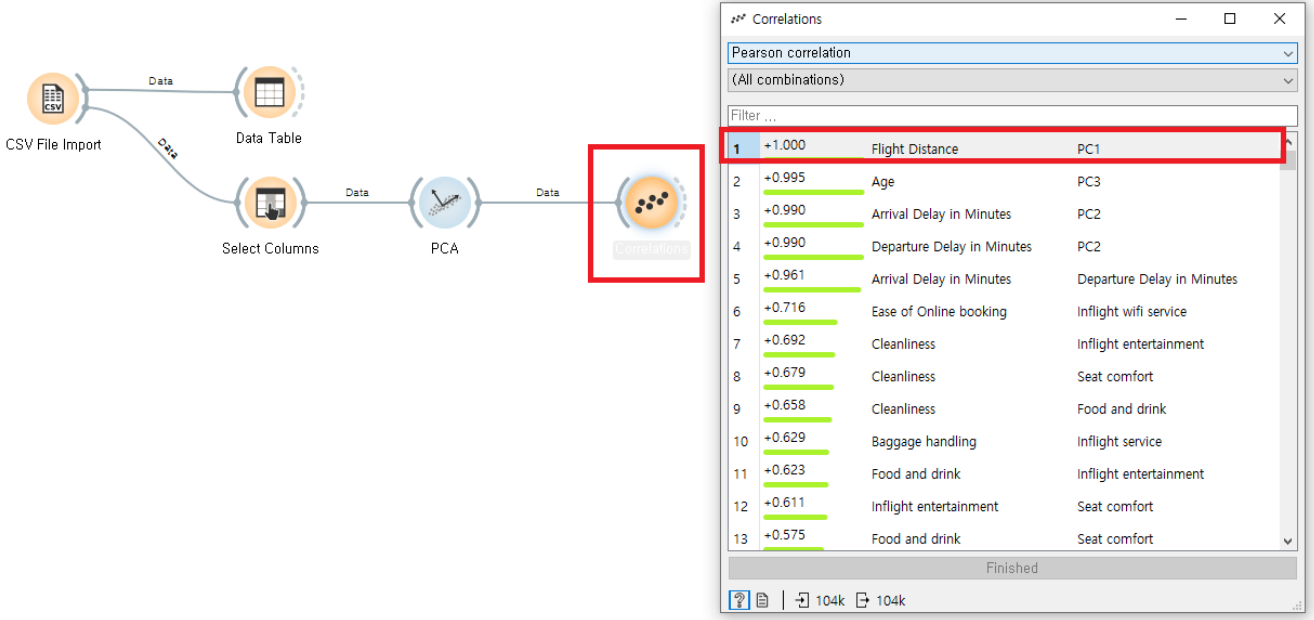
Airline passenger satisfaction data

- PCA를 통해 사용자가 원하는 차원으로 축소
- 4개의 주성분(PC1~PC4) 설정을 하니 비율이 99%로 설정



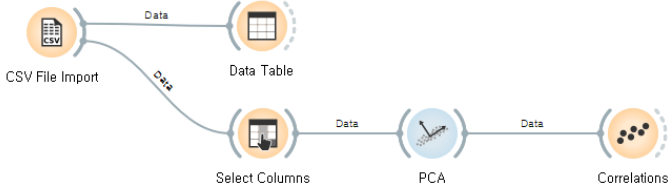
Airline passenger satisfaction data

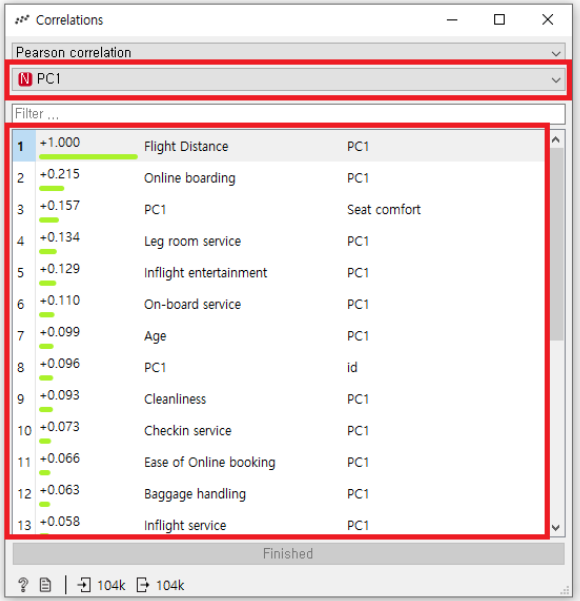
- 각 주성분이 어떤 상관관계를 가지는지 파악하기 위해 Correlations 위젯을 연결
- 대표적으로 PC1의 경우 기존 'Flight Distance' 와 상관관계가 큼



Airline passenger satisfaction data

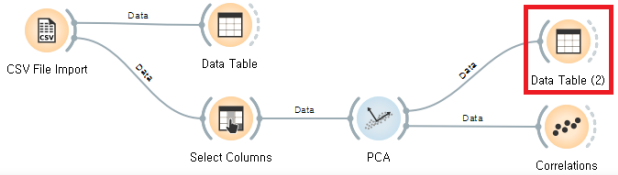
- 각 주성분의 가중치만을 보고싶다면 all combinations 가 아닌 해당 feature를 설정
- PC1만을 설정한 결과 PC1에 영향을 주는 각 feature들의 상관계수가 나타나 있음

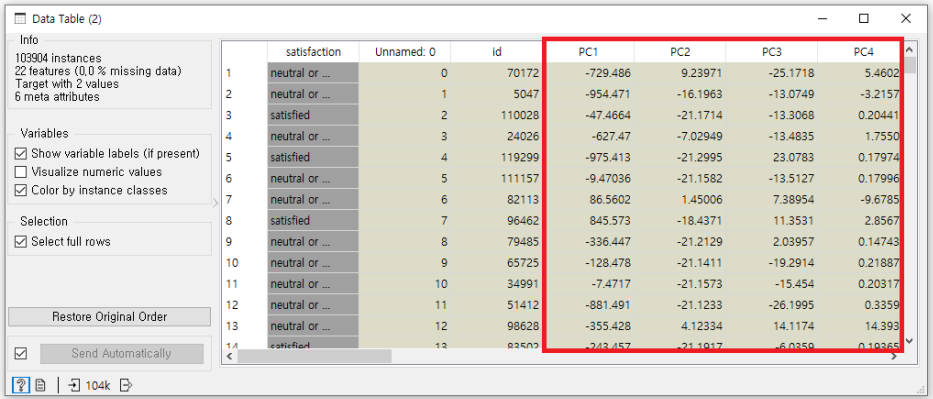




Airline passenger satisfaction data

- 차원축소를 통해 데이터가 어떤 식으로 구성됐는지 data table 위젯을 추가하여 확인
- 각 instance가 어떤 PC와 양/음의 상관관계인지 확인

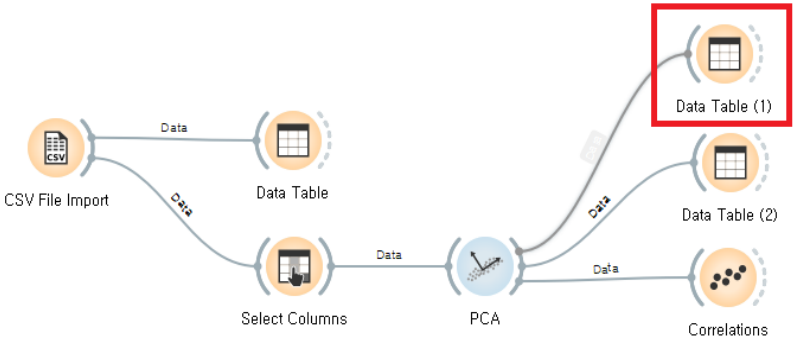


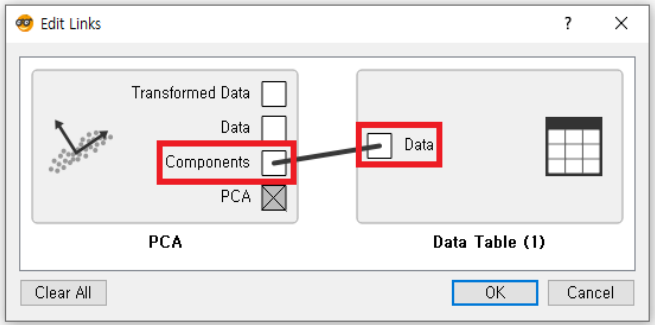


	satisfaction	Unnamed: 0	id	PC1	PC2	PC3	PC4
1	neutral or ...	0	70172	-729.486	9.23971	-25.1718	5.4602
2	neutral or ...	1	5047	-954.471	-16.1963	-13.0749	-3.2157
3	satisfied	2	110028	-47.4664	-21.1714	-13.3068	0.20441
4	neutral or ...	3	24026	-627.47	-7.02949	-13.4835	1.7550
5	satisfied	4	119299	-975.413	-21.2995	23.0783	0.17974
6	neutral or ...	5	111157	-9.47036	-21.1582	-13.5127	0.17996
7	neutral or ...	6	82113	86.5602	1.45006	7.38954	-9.6785
8	satisfied	7	96462	845.573	-18.4371	11.3531	2.8567
9	neutral or ...	8	79485	-336.447	-21.2129	2.03957	0.14743
10	neutral or ...	9	65725	-128.478	-21.1411	-19.2914	0.21887
11	neutral or ...	10	34991	-7.4717	-21.1573	-15.454	0.20317
12	neutral or ...	11	51412	-881.491	-21.1233	-26.1995	0.3359
13	neutral or ...	12	98628	-355.428	4.12334	14.1174	14.393
14	satisfied	13	83502	-243.457	-21.1917	-6.0359	0.19365

Airline passenger satisfaction data

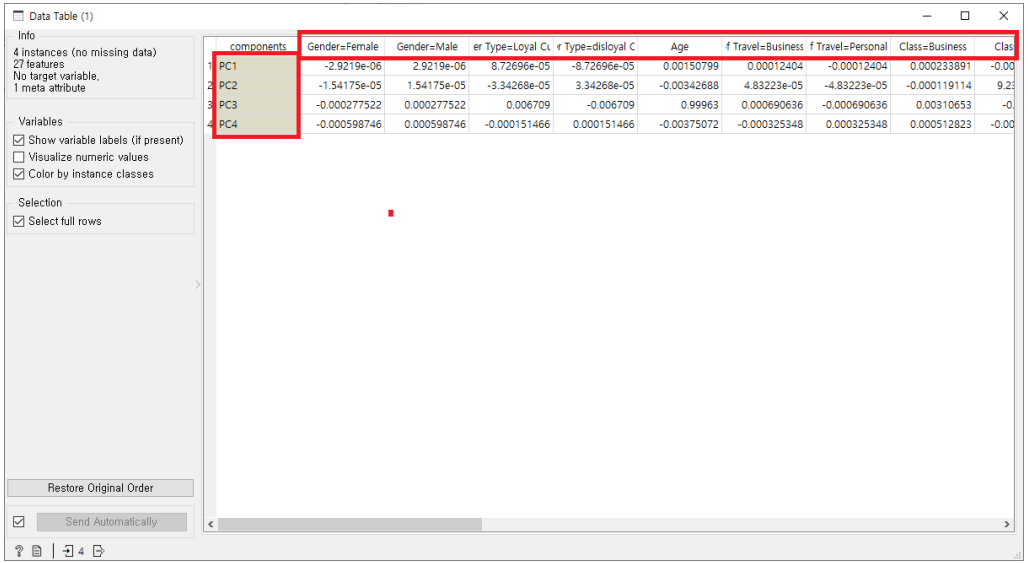
- 각 PC가 feature값과 어떤 상관관계가 파악하기 위해 data table위젯을 추가
- 이때 PCA와 data table사이의 링크는 components to data로 지정





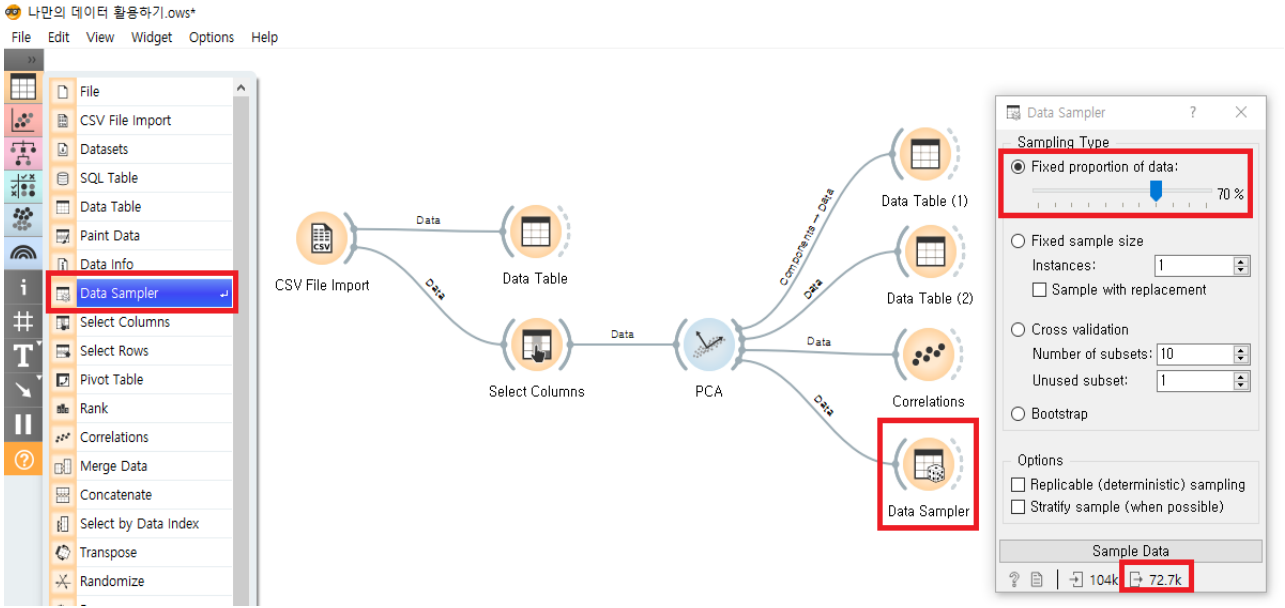
Airline passenger satisfaction data

- Data table 위젯을 활성화하여 PC와 feature간의 상관관계를 파악
- PC별로 상관관계가 큰 feature를 확인
- 그 값이 양(+)의 값이며 클 수록 상관관계가 크며 그 값이 음(-)이며 클 수록 관련이 없이 상관관계가 크다고 할 수 있음

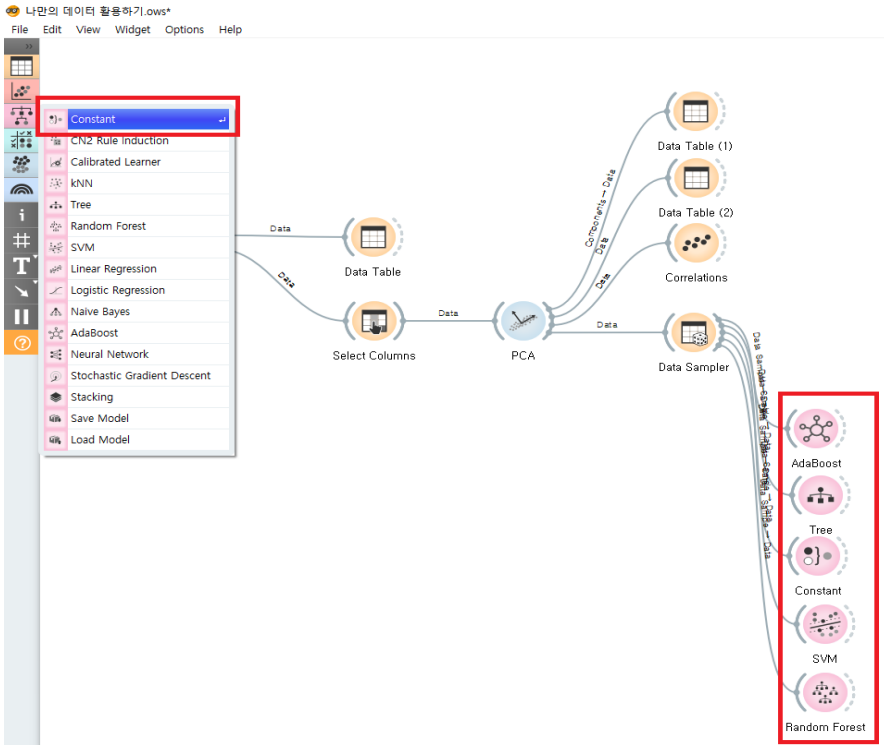


Airline passenger satisfaction data

- 승객들의 선호도에 따른 예측 모델을 만듬
- 예측을 위해서는 훈련 데이터와 예측 데이터가 있어야 함
- 예측 데이터를 분류하기 위해 data sampler 위젯을 활용하며 70%비율을 설정



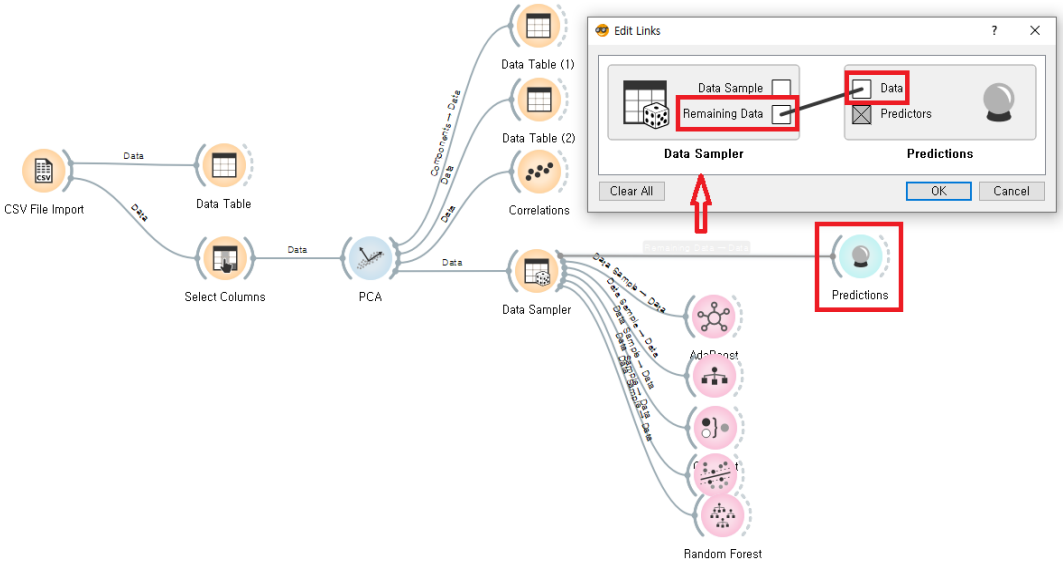
Airline passenger satisfaction data



- 다양한 모델을 활용해 학습 모델을 만들도록 함
- 이번 활동에서 활용할 모델은 AdaBoost, Tree, Constant, SVM, Random Forest임
- 연결을 통해 학습 모델을 생성하고 각 모델은 왼쪽 메뉴 Prediction에서 끌어옴

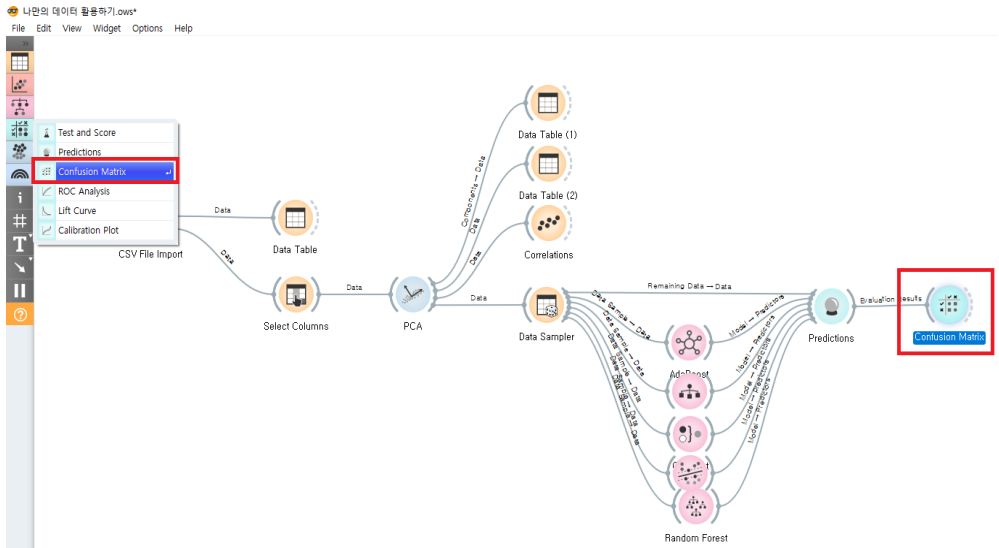
Airline passenger satisfaction data

- 학습한 모델을 토대로 예측을 하도록 함
- 70%의 샘플데이터로 훈련을 시켜 학습을 하고 30%의 데이터로 예측
- Predictions 위젯을 추가하여 연결하고 위젯사이의 연결을 remaining data to data로 변경



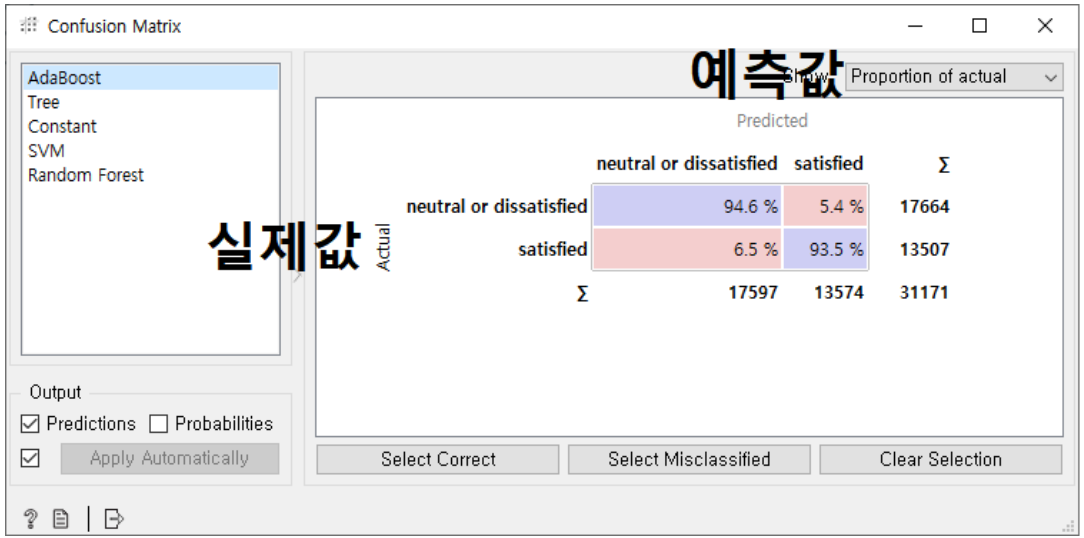
Airline passenger satisfaction data

- 각 모델별로 Predictions 위젯에 연결하여 예측을 실행하고 이에 대한 결과를 확인하기 위해 Confusion Matrix 위젯을 추가
- Confusion Matrix는 평가 결과에서 생성된 혼동 행렬을 표시하는 위젯
- 실제값은 어떠했고 예측값은 어떠한지 비교할 수 있음



Airline passenger satisfaction data

- Adaboost를 예로 들면 실제로 neutral or dissatisfied였는데 test data로 예측한 결과 neutral or dissatisfied라고 예측한 비율은 94.6%. 반면 satisfied라고 예측한 비율은 5.4%.
- 실제 조사 결과 Satisfied였는데 test data로 예측한 결과 neutral or dissatisfied 라고 예측한 비율은 6.5%. 반면 satisfied 라고 예측한 비율은 93.5%.



Airline passenger satisfaction data

- 차원 축소 과정을 통해 뽑아낸 주성분을 가지고 해당 승객들이 어떤 특성을 갖는지 해석할 수 있음
- 다만 주성분의 의미는 연구자의 자의적 해석에 따라 달라질 수 있으며, 주성분 분석을 통해 각각의 개인에 대한 특성을 쉽게 확인할 수 있는 장점이 있음
- 데이터 샘플을 구성하지 않고 raw data를 활용하여 해당 승객들이 어떤 만족도를 느끼는지 확인
- 특히 만족도가 낮게 나온 승객 instance를 조사하여 –값이 큰 PC를 알아내고 이를 위해 어떤 feature를 보완 해야할지 생각해 볼 수 있음

질문 있나요?

hsryu13@hongik.ac.kr