



orange 활용 데이터 분석 및 머신 러닝



15차시

데이터분석 실습

탐색적 데이터분석 (EDA- Exploratory Data Analysis)

- 데이터를 분석하고 결과를 내는 과정에 있어서 지속적으로 해당 데이터에 대한 '탐색과 이해'를 기본으로 가져야 한다는 것을 의미
 1. raw data 의 description, dictionary 를 통해 데이터의 각 column들과 row의 의미를 이해
 2. 결측치 처리 및 데이터 필터링 (이상치 발견 및 데이터 정제)
 3. 시각화를 통한 데이터에 대한 다양한 이해와 분석을 함으로서 현상과 수치에 관한 통찰력 획득 및 분석 방향 수립

세계 사회경제 지표를 활용한 기대수명예측

Life Expectancy (WHO) 자료 준비하기

<https://www.kaggle.com/datasets/kumarajarshi/life-expectancy-who>

File - Orange

Source

☒ File: Life Expectancy Data.csv

☐ URL:


Columns (Double click to edit)

	Name	Type	Role	Values
1	Year	N numeric	feature	
2	Status	C categorical	feature	Developed, Developing
3	Life expectancy	N numeric	target	
4	Adult Mortality	N numeric	feature	
5	infant deaths	N numeric	feature	
18	thinness 1-19 ...	N numeric	feature	
19	thinness 5-9 ...	N numeric	feature	
20	Income ...	N numeric	feature	
21	Schooling	N numeric	feature	
22	Country	C categorical	meta	

Life Expectancy (WHO) | Kaggle x +

com/datasets/kumarajarshi/life-expectancy-who

Search

 KUMARRAJARSHI · UPDATED 5 YEARS AGO

Life Expectancy (WHO)

Statistical Analysis on factors influencing Life Expectancy

Feature 설명

- Country: 국가명
- Year: 2000년부터 2015년까지의 연도
- Status: Developed(선진국) or Developing(개발도상국) status
- Life expectancy: 기대수명(나이)
- Adult Mortality: 15세~60세사이의 성인 1000명당 사망자수
- infant deaths: 유아 1000명당 사망자수
- Alcohol: 1인당 알콜 소비량
- percentage expenditure: GDP 대비 보건 예산 지출비율(%)
- Hepatitis B: 1세 아동의 B형 간염 예방 접종률(%)
- Measles: 인구 1000명당 홍역 예방 접종률(%)
- BMI: 전인구 평균 체질량 지수
- Under-five deaths: 5세이하 아동 1000명당 사망자수
- Polio: 1세 아동의 소아마비 면역률(%)
- Total expenditure: 정부 총예산 대비 보건 분야 예산(%)
- Diphtheria: 1세 아동의 디프테리아 예방 접종률(%) HIV/AIDS: HIV/AIDS 감염상태로 태어남 0-4세 인구 1000명당 사망자수
- GDP: 1인당 GDP
- Population: 국가 총인구
- thinness 1-19 years: 1-19 세 청소년 중 저체중 비율
- thinness 5-9 years: 5-9세 사이의 아동의 저체중 비율
- Income composition of resources: 소득 구성에 따른 인간개발지수
- Schooling: 학교 재학 연수

데이터 추출

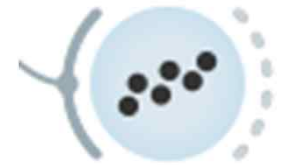
	A	B	C	D	
1	Country	Year	Status	Life expect Ac	
2	Afghanistan	2015	Developing	65	
3	Afghanistan	2014	Developing	59.9	
4	Afghanistan	2013	Developing	59.9	
5	Afghanistan	2012	Developing	59.5	
6	Afghanistan	2011	Developing	59.2	
7	Afghanistan	2010	Developing	58.8	
8	Afghanistan	2009	Developing	58.6	
9	Afghanistan	2008	Developing	58.1	
10	Afghanistan	2007	Developing	57.5	
11	Afghanistan	2006	Developing	57.3	
12	Afghanistan	2005	Developing	57.3	
13	Afghanistan	2004	Developing	57	
14	Afghanistan	2003	Developing	56.7	
15	Afghanistan	2002	Developing	56.2	
16	Afghanistan	2001	Developing	55.3	
17	Afghanistan	2000	Developing	54.8	
18	Albania	2015	Developing	77.8	
19	Albania	2014	Developing	77.5	
20	Albania	2013	Developing	77.2	

원본 데이터의 포맷을 확인하여 분석방향 결정

- 1) 2015년 데이터를 활용해 다양한 사회경제지표를 중심 군집화 해보자. (전처리는 어떤 것이 필요할까?)
- 2) 2015년의 데이터를 통해 특정 국가의 기대수명을 예측해보고 실제 값과 비교해보자.
- 3) 우리나라의 2016~2020년간의 기대수명 변화를 예측해보자.

데이터 탐색

- 질문 1. 기대수명과 가장 상관관계가 높은 속성은 무엇일까?



Correlations

- BMI 수치와 기대 수명은 어떤 상관관계가 있을까? (선진국과 개발 도상국을 비교해보기)



Scatter Plot

- 2015년 전 세계의 기대수명 평균은 몇살일까?



Feature Statistics

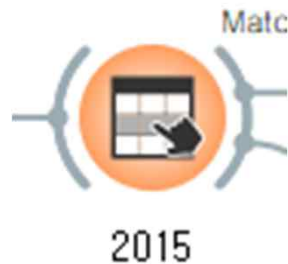
2015년 데이터를 활용하여 군집화 해보자.

1. 결측치 제거는 어떻게 할 것인가? 결측치 처리를 안 했을 때 결과는?

- 필요한 데이터 추출하기

모든 인스턴스의 해당 속성을 모두 사용하지 않고 처리하므로 결측치가 있는 경우 꼭 필요한 속성이라면 반드시 결측치 처리를 해야한다.

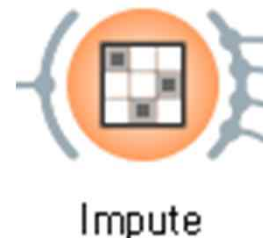
- 속성별 결측치 정보 확인하여 어떤 결측치 처리 방법을 사용할 것인지 결정하자.



Select Row 위젯으로 year=2015인 인스턴스를 추출한다.



결측치 정보를 확인하고 각 속성의 기본적인 특성을 파악한다.

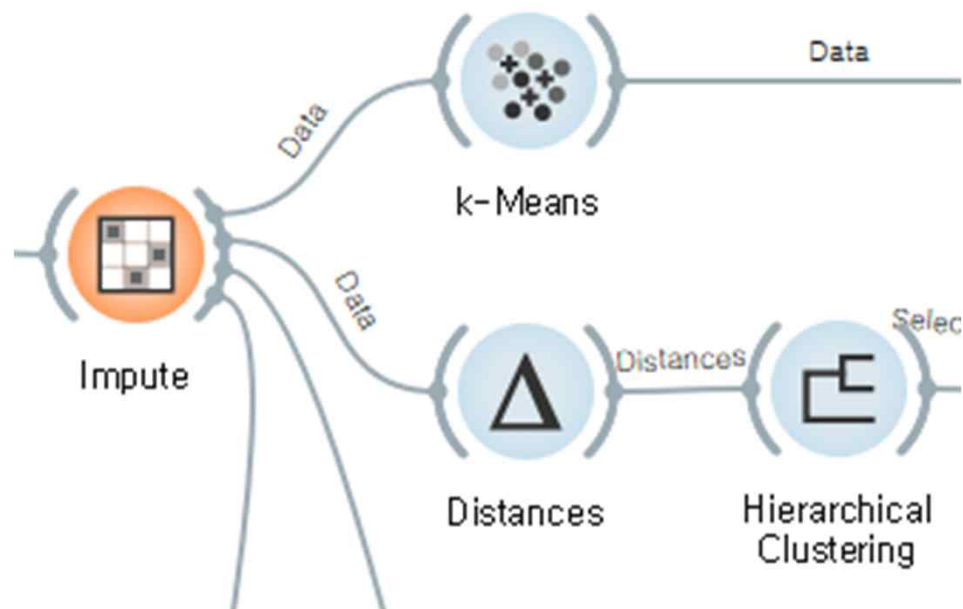


결측치량이 많은 속성은 제거하고 분석에 꼭 필요한 속성은 속성 특성에 따라 처리한다.



군집화
시작

2. 군집화 알고리즘은 어떤 것으로 선택할까? (두 가지 방법의 군집화 결과를 모두 점검해보자.)



군집화 결과 나타내기

1. Data Table을 Cluster로 정렬하여 확인
2. 지도에 출력하여 나타내기 : 현재 위경도 정보가 없으므로 생성해야 한다.

Geo-code : 영문국가명으로 위도경도 정보 만들기



Geocoding - Orange

☒ Encode region names into geographical coordinates:

Region identifier:

Identifier type:

☐ Decode latitude and longitude:

Latitude:

Longitude:

Administrative level:

☒ Extend coded data with:

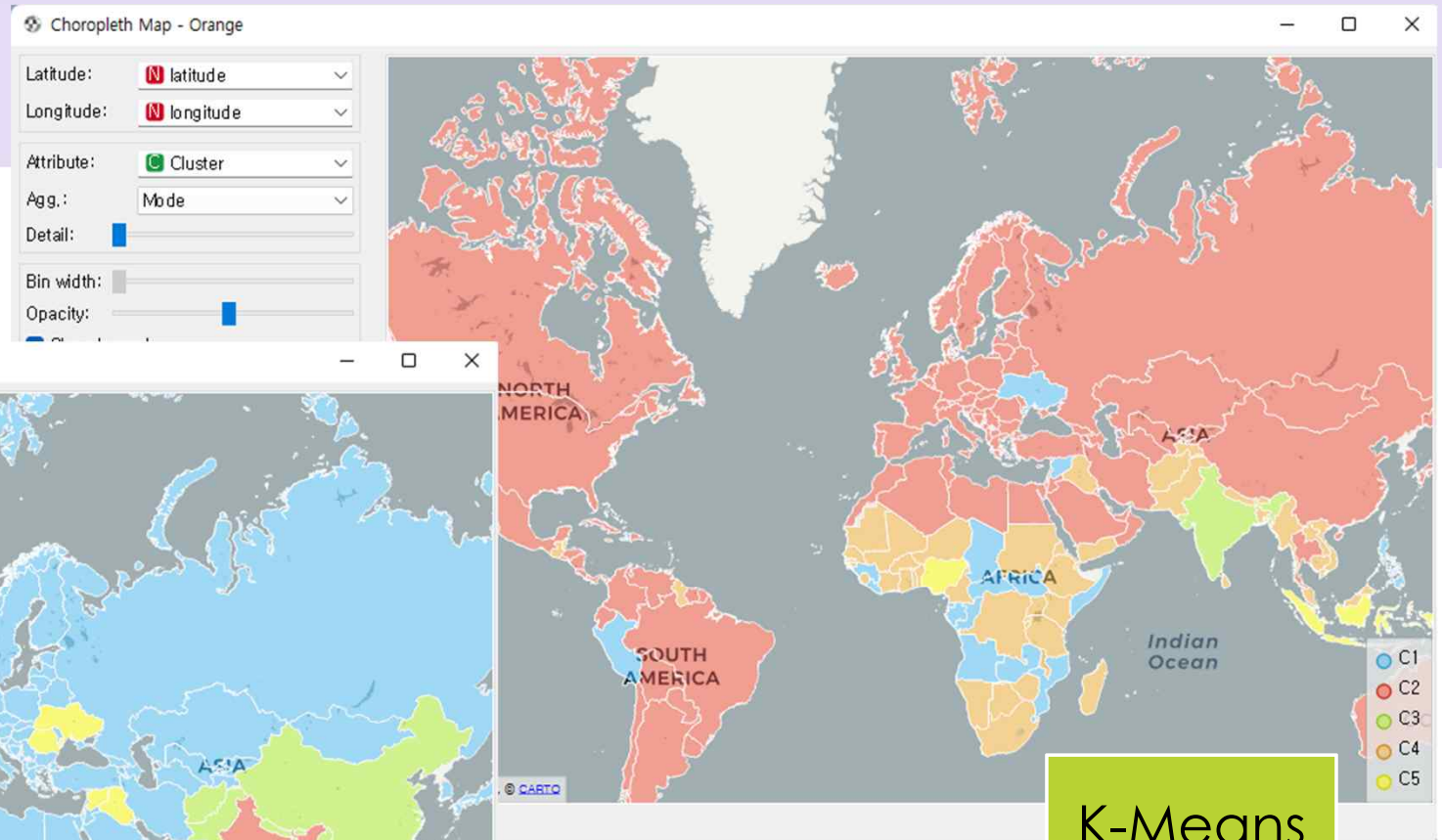
☒ Apply Automatically

Unmatched identifiers: 0 / 183

Unmatched Identifier	Custom Replacement
Antigua and Barbuda	Antigua and Barb.
Bolivia (Plurinational State of)	Bolivia
Côte d'Ivoire	Côte d'Ivoire
Iran (Islamic Republic of)	Iran
Micronesia (Federated States of)	Micronesia
Republic of Moldova	Moldova
Saint Vincent and the Grenadines	St. Vin. and Gren.
Sao Tome and Principe	São Tomé and Príncipe
The former Yugoslav republic of ...	Macedonia
United Republic of Tanzania	Tanzania
Venezuela (Bolivarian Republic of)	Venezuela

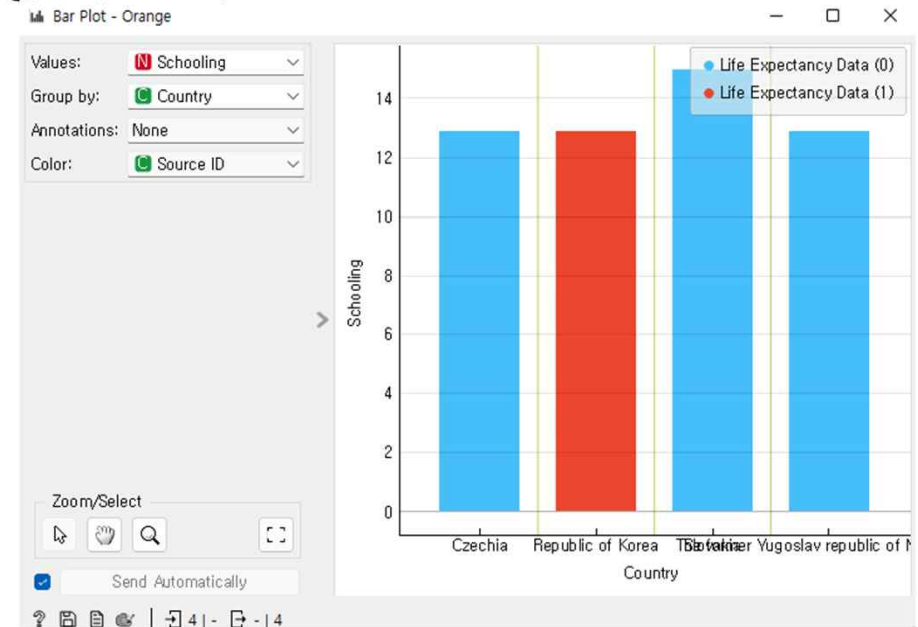
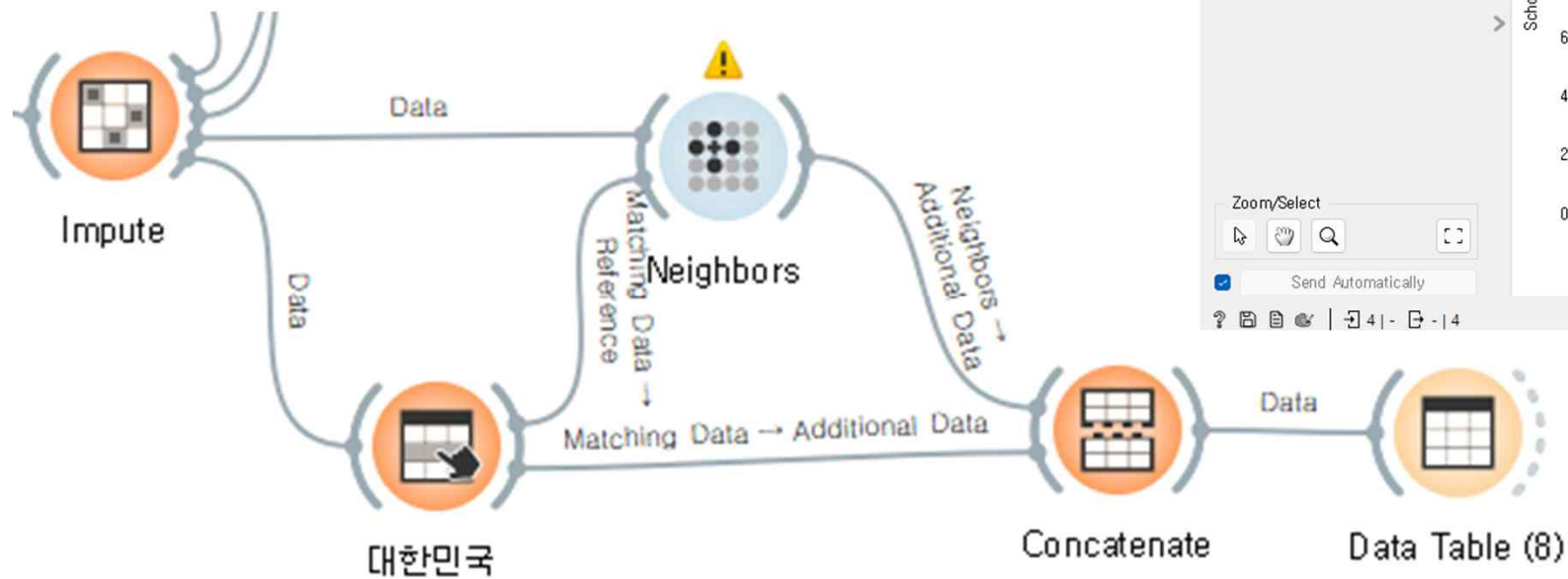
? | 183 | 183

계층적 군집화
결과

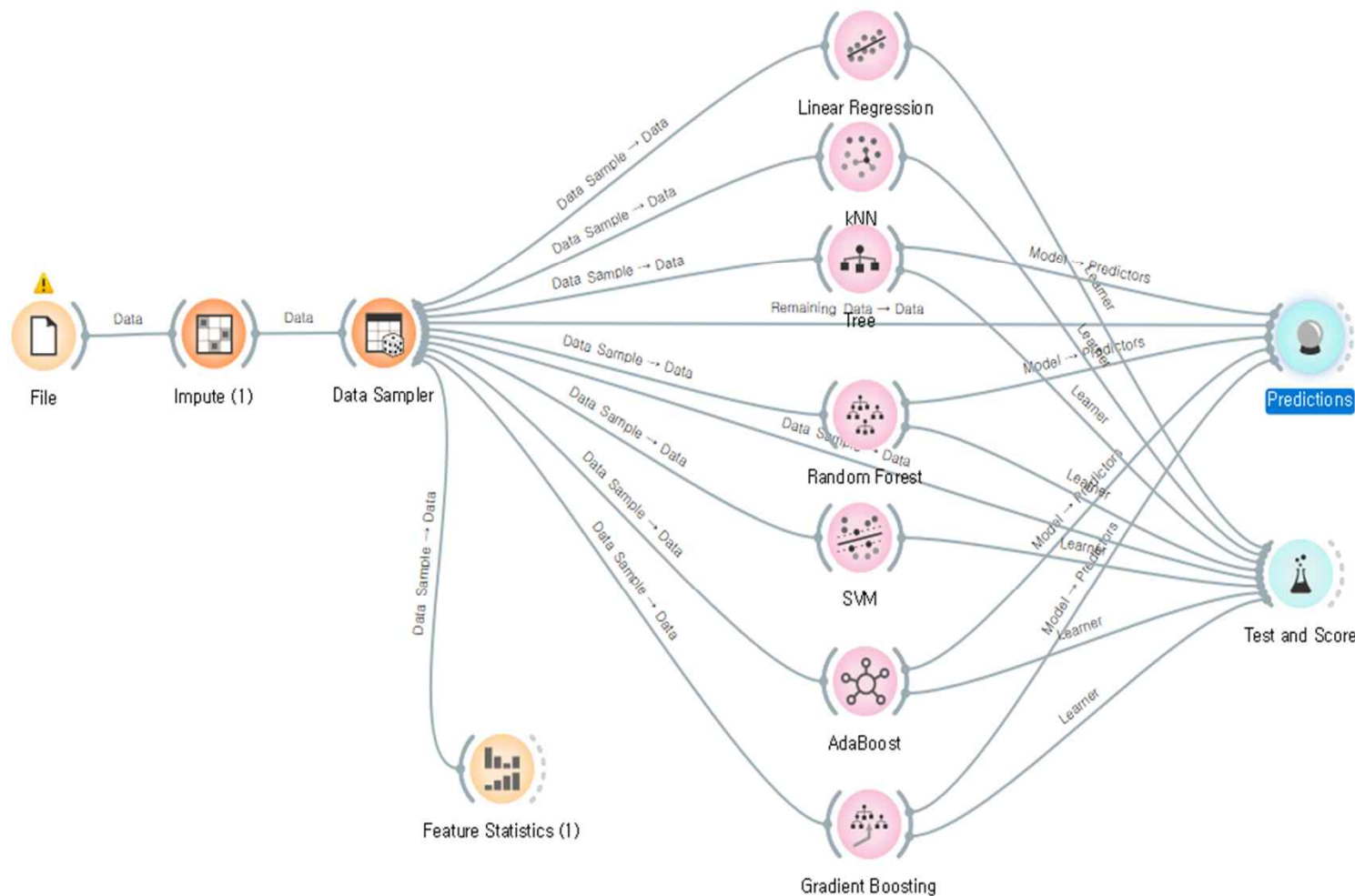


K-Means
결과

3. 우리나라와 가장 유사한 국가 3곳을 선정해보자.



국가의 기대수명을 예측해보고 실제 값과 비교해보자.



Test and Score - Orange

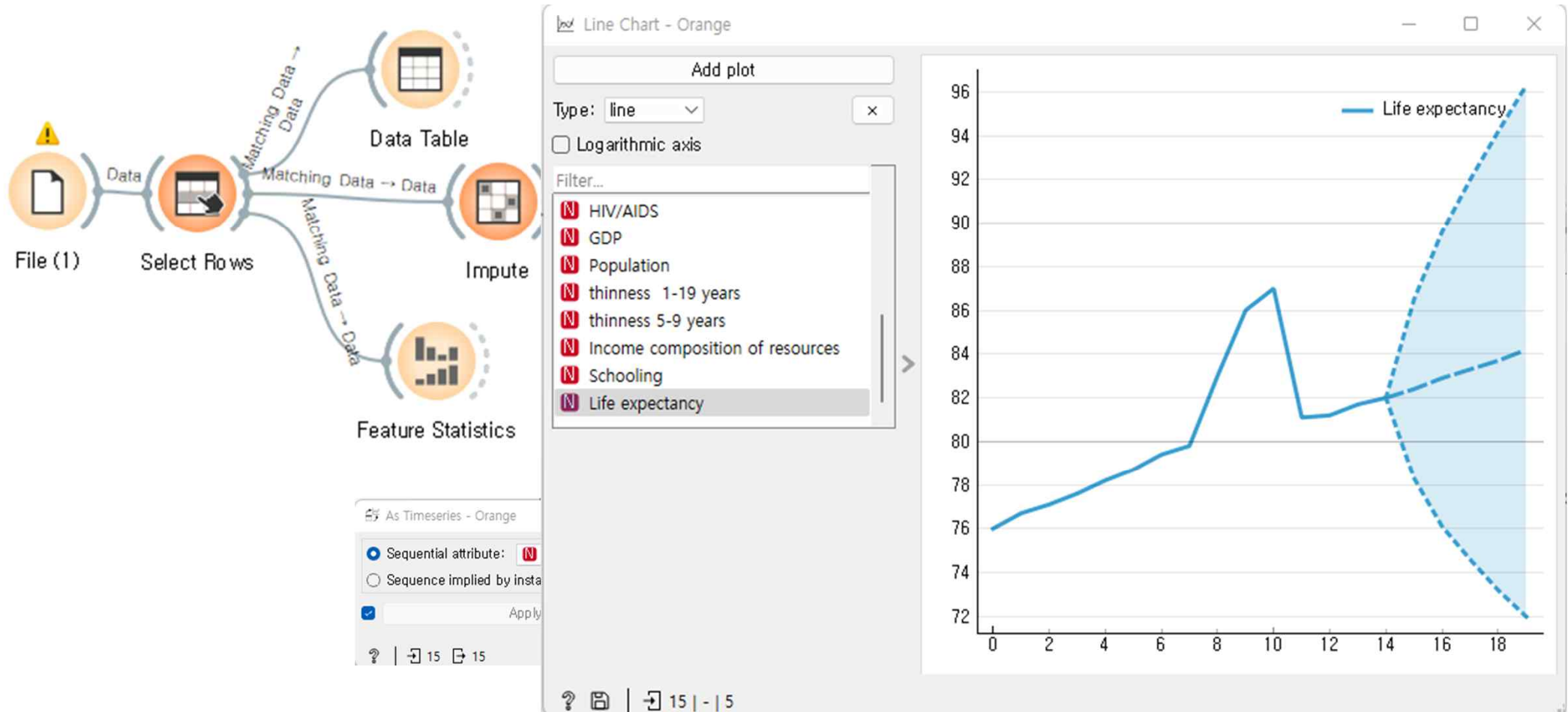
Model	MSE	RMSE	MAE	R2
AdaBoost	3.289	1.813	1.049	0.964
Random Forest	4.223	2.055	1.326	0.953
Gradient Boosting	5.075	2.253	1.616	0.944
Tree	7.322	2.706	1.691	0.919
Linear Regression	16.770	4.095	3.039	0.815
SVM	39.339	6.272	5.210	0.565
kNN	75.247	8.674	6.436	0.168

Predictions - Orange

Shown regression error: (None)

	Random Forest	AdaBoost	Gradient Boosting	Tree	Life expectancy
1	73.9	72.8	72.7	73.0	72.4
2	73.5	73.8	74.2	76.5	73.9
3	75.7	75.0	74.7	75.3	75.6
4	77.4	77.5	78.1	77.7	77.5
5	76.2	77.9	75.5	77.7	78.2
6	74.7	74.5	74.6	73.9	76.1
7	68.9	68.9	69.9	68.7	68.7
8	58.6	59.7	59.4	60.0	65.0
9	47.2	45.3	46.1	45.5	44.6

우리나라의 2016~2020년간의 기대수명 변화를 예측해보자.





긴 과정을 함께 해 주셔서 감사합니다.

오렌지 데이터분석도구를 통해 여러분의 데이터분석 능력과 업무를 이해하는 시야가 더욱더 넓어지는 도움이 되길 바랍니다. 감사합니다.

오렌지라는 새로운 도구가 많이 낫설고 데이터분석과 머신러닝의 개념도 생소하셨을 텐데 긴 교육 열심히 참여해주셔서 감사합니다.

baejteacher@gmail.com