실습 3

Naive Bayes를 사용한
심장 마비 발생 예측

[Lecture] Dr. HeeSuk Kim

3

내 심장은 건강할까?

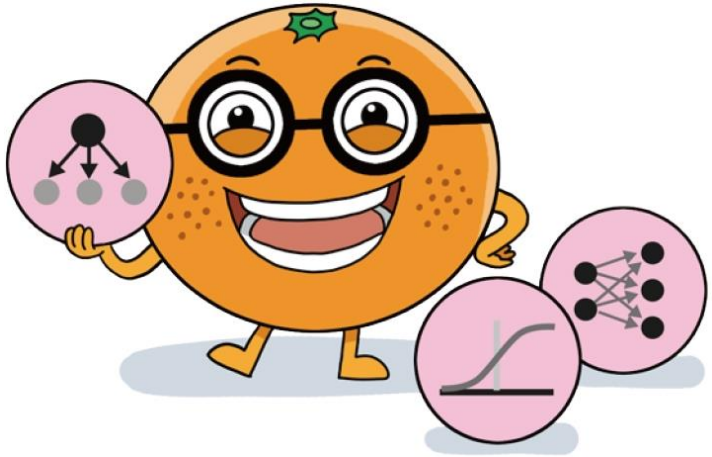
Naive Bayes를 사용하여 심장 마비가 발생할 사람을 예측해 보자.

데이터 종류:

정형 데이터

사용하는 모델:

Naive Bayes



1 해결해야 할 문제는 무엇일까?

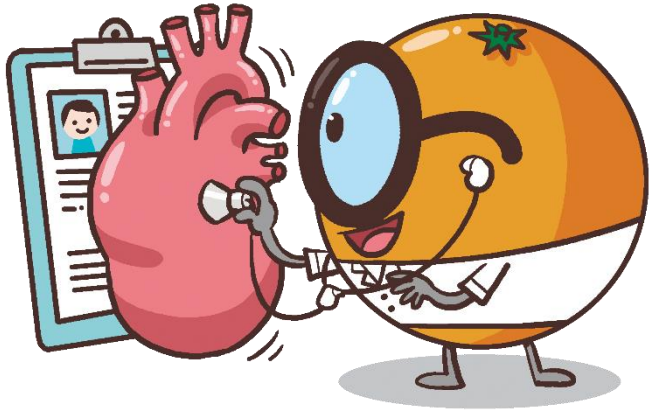
문제 상황

미국에서는 매년 90만 명 이상의 사람이 심장 질환으로 사망한다.
미국보다 그 수는 많지 않지만 우리나라도 많은 이들이 심장마비나
급성 심근 경색 등의 심장 질환으로 사망하고 있다. 이러한 상황에서
심장 질환으로 인한 사망자 수를 줄이기 위해 환자의 건강 정보를
활용하여 환자에게 심장 마비가 발생할 확률을 예측해 보는 것은 어떨까?

환자의 건강 정보 데이터를 분석하고

심장 마비가 발병할지를 예측할 수 있는

인공지능 모델을 만들어 보자.



2 데이터를 준비하자!

데이터 다운로드 링크

https://bit.ly/3ch5DWc

1 외부 데이터 다운로드

① 캐글 데이터 다운로드하기

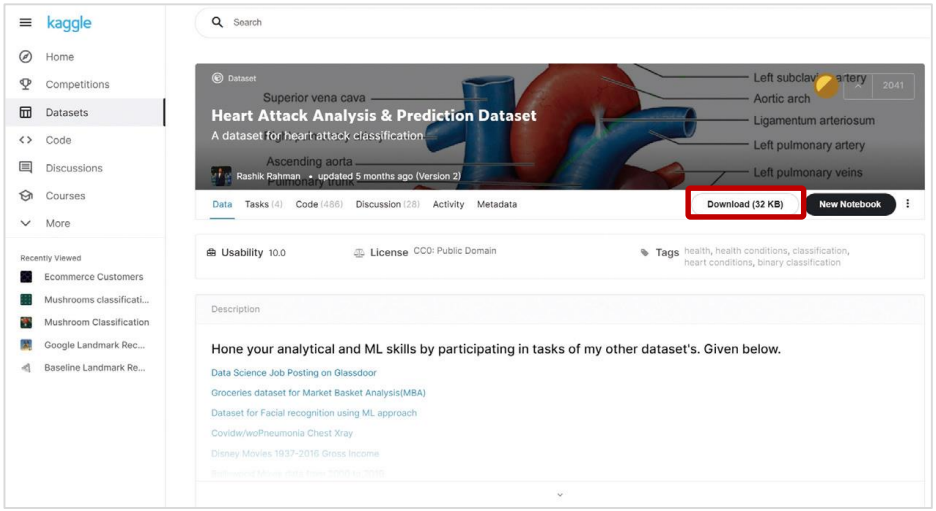
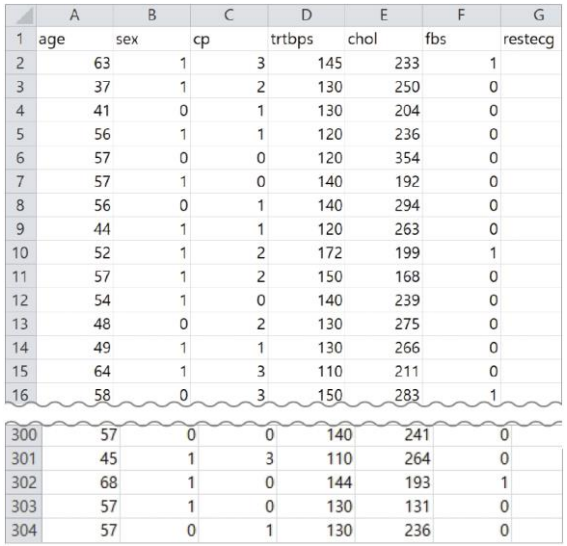


그림 3-1 캐글의 Heart Attack Analysis & Prediction 데이터

- 캐글(Kaggle)은 2010년에 설립된 데이터 분석 대회 플랫폼
- 캐글에서 'Heart Attack Analysis & Prediction'을 검색하여 데이터 다운로드

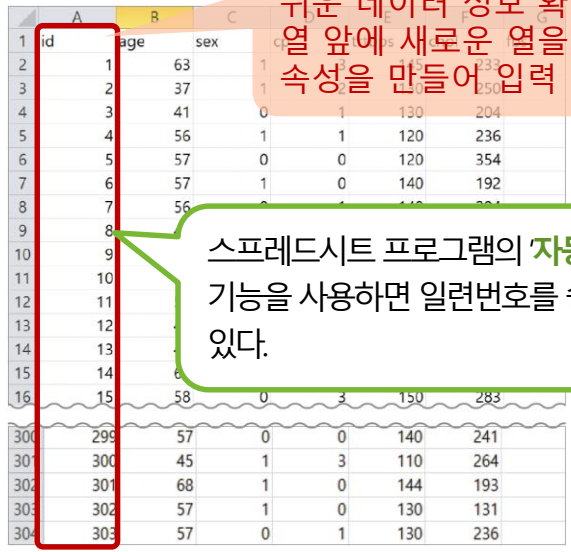
② 훈련 데이터와 테스트 데이터 나누기

- archive.zip 파일을 풀어 heart.csv 파일을 열고, 훈련 데이터와 테스트 데이터 파일로 나누어 저장



	A	B	C	D	E	F	G
1	age	sex	cp	trtbps	chol	fbs	restecg
2	63	1	3	145	233	1	
3	37	1	2	130	250	0	
4	41	0	1	130	204	0	
5	56	1	1	120	236	0	
6	57	0	0	120	354	0	
7	57	1	0	140	192	0	
8	56	0	1	140	294	0	
9	44	1	1	120	263	0	
10	52	1	2	172	199	1	
11	57	1	2	150	168	0	
12	54	1	0	140	239	0	
13	48	0	2	130	275	0	
14	49	1	1	130	266	0	
15	64	1	3	110	211	0	
16	58	0	3	150	283	1	
300	57	0	0	140	241	0	
301	45	1	3	110	264	0	
302	68	1	0	144	193	1	
303	57	1	0	130	131	0	
304	57	0	1	130	236	0	

그림 3-2 id 속성 추가하기 전



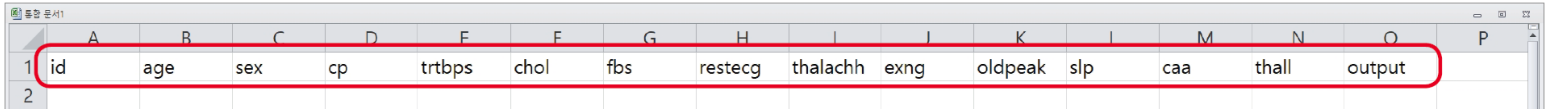
	A	B	C	D	E	F	G
1	id	age	sex				
2	1	63					
3	2	37					
4	3	41	0	1	130	204	
5	4	56	1	1	120	236	
6	5	57	0	0	120	354	
7	6	57	1	0	140	192	
8	7	56					
9	8						
10	9						
11	10						
12	11						
13	12						
14	13						
15	14						
16	15	58	0	3	150	283	
300	299	57	0	0	140	241	
301	300	45	1	3	110	264	
302	301	68	1	0	144	193	
303	302	57	1	0	130	131	
304	303	57	0	1	130	236	

쉬운 데이터 정보 확인을 위해 A 열 앞에 새로운 열을 삽입 후, id 속성을 만들어 입력

스프레드시트 프로그램의 '**자동 채우기 핸들**' 기능을 사용하면 일련번호를 쉽게 입력할 수 있다.

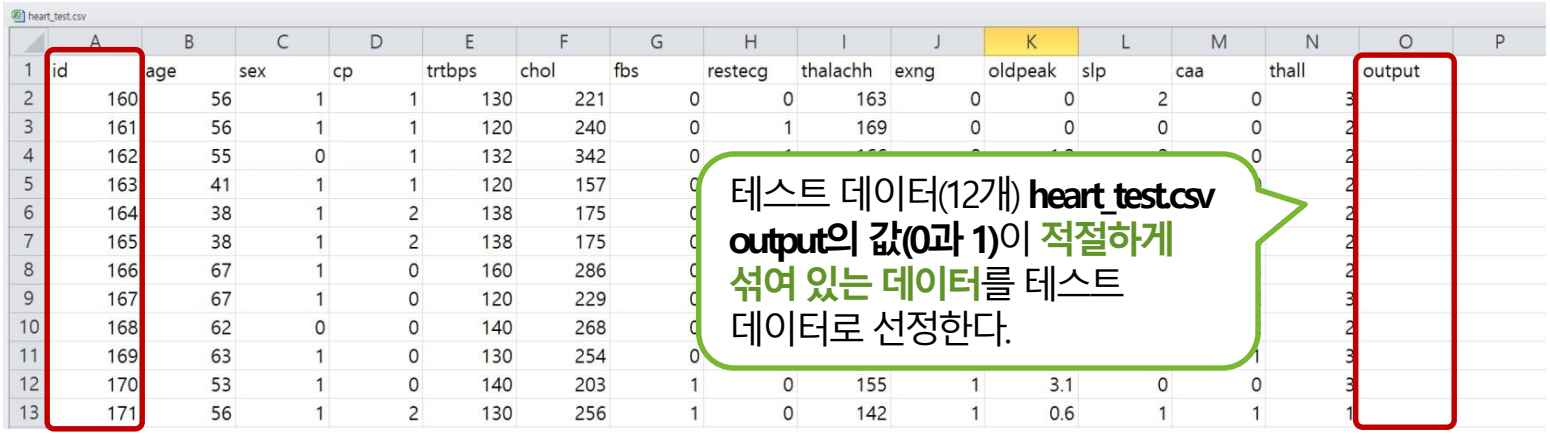
그림 3-3 id 속성 추가한 후

- 새 파일을 열어 **heart.csv** 파일 1행의 속성명을 복사하여 붙여넣기



	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P
1	id	age	sex	cp	trtbps	chol	fbs	restecg	thalachh	exng	oldpeak	slp	caa	thall	output	
2																

- **heart.csv** 파일에서 id 160~171까지의 데이터를 복사하여 새 파일에 붙여 넣은 후, **output의 내용을 삭제**
- 새 파일의 이름을 **heart_test**로 설정하고 csv 파일로 저장



heart_test.csv

	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P
1	id	age	sex	cp	trtbps	chol	fbs	restecg	thalachh	exng	oldpeak	slp	caa	thall	output	
2	160	56	1	1	130	221	0	0	163	0	0	2	0	3		
3	161	56	1	1	120	240	0	1	169	0	0	0	0	2		
4	162	55	0	1	132	342	0						0	2		
5	163	41	1	1	120	157	0							2		
6	164	38	1	2	138	175	0							2		
7	165	38	1	2	138	175	0							2		
8	166	67	1	0	160	286	0							2		
9	167	67	1	0	120	229	0							3		
10	168	62	0	0	140	268	0							2		
11	169	63	1	0	130	254	0						1	3		
12	170	53	1	0	140	203	1	0	155	1	3.1	0	0	3		
13	171	56	1	2	130	256	1	0	142	1	0.6	1	1	1		

테스트 데이터(12개) **heart_test.csv output의 값(0과 1)**이 **적절하게 섞여 있는 데이터**를 테스트 데이터로 선정한다.

- **heart.csv** 파일에서 id 160~171까지의 데이터를 **삭제**한 후, 파일명을 **heart_train**으로 바꾸고 **csv** 파일로 저장

heart_train.csv

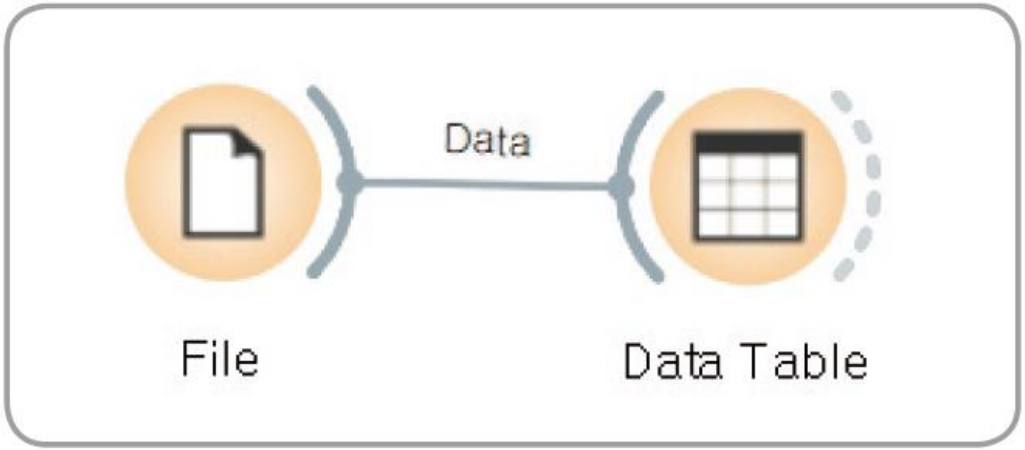
	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P
1	id	age	sex	cp	trtbps	chol	fbs	restecg	thalachh	exng	oldpeak	slp	caa	thall	output	
2	1	63	1	3	145	233	1	0	150	0	2.3	0	0	1	1	
3	2	37	1	2	130	250	0	1	187	0	3.5	0	0	2	1	
4	3	41	0	1	130	204	0	0	172	0	1.4	2	0	2	1	
5	4	56	1	1	120	236	0	1	178	0	0.8	2	0	2	1	
6	5	57	0	0	120	354	0	1	163	1	0.6	2	0	2	1	
7	6	57	1	0	140	192	0	1	148	0	0.4	1	0	1	1	
8	7	56	0	1	140	294	0	0	153	0	1.3	1	0	2	1	
9	8	44	1	1	120	263	0	1	173	0	0	2	0	3	1	
10	9	52	1	2	172	199	1	1	162	0	0.5	2	0	3	1	

훈련 데이터(291개)
heart_train.csv

2 데이터 불러오기

- Data 카테고리 - [File] 위젯을 가져와서 더블 클릭 후, 미리 준비한 데이터 (**heart_train.csv**) 불러오기

- Data 카테고리 – [Data Table]위젯을 가져와서 [File]위젯과 연결



- [Data Table] 위젯을 통해 전체 데이터를 한눈에 살펴보고, Info 탭에서 데이터 정보 확인

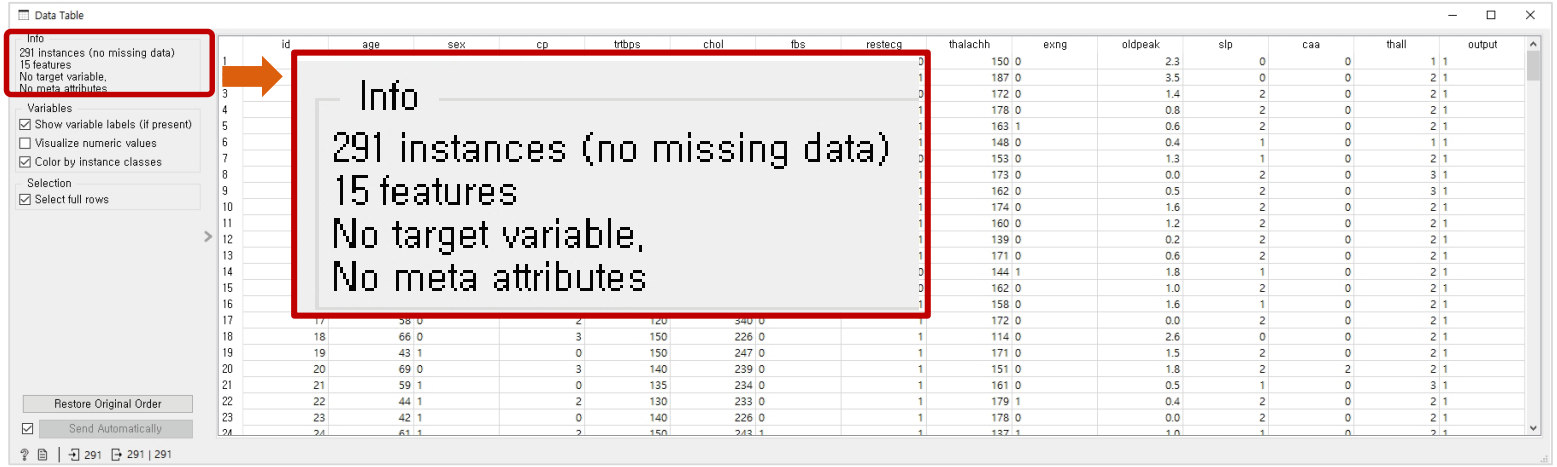


그림 3-4 Data Table 창에서 본 데이터

3 데이터 속성 정보 확인하기

◆ 심장 마비 데이터: 291명에 대한 15개 속성 정보

의료 정보에 대한 데이터 분석을 시행할 때는 의료 데이터에 대한 의학적 지식이 풍부한 의료 전문가의 도움을 받는 것이 좋다.

속성명	속성 정보
id	일련번호
age	환자 나이
sex	성별 • 0: 여성 • 1: 남성
cp	가슴 통증 유형 • 0: 무증상 • 1: 전형적 협심증 • 2: 비정형적 협심증 • 3: 비협심증이나 통증 있음.
trtbps	안정 시 혈압(mmHg)
chol	혈청 콜레스테롤
fbs	공복 혈당 • 0: 120미만 • 1: 120 이상
restecg	휴식 중 심전도 결과 • 0: 정상 • 1: ST-T파 이상 • 2: 비대

속성명	속성 정보
thalachh	최대 심박수
exng	운동으로 인한 협심증 • 0: 무증상 • 1: 증상
oldpeak	심전도 결과 ST 분절 하강 정도
slp	심전도 결과 최고 ST 분절 기울기 • 0: 하향 경사 • 1: 평평함 • 2: 상승 경사
caa	플루로 스코피(투시 조영)로 인해 착색된 주요 혈관 수(0-3)
thall	탈륨 스트레스 검사 결과 • 1: 비가역적 결함 • 2: 정상 • 3: 가역적 결함
output	심장 마비 가능성 • 0: 가능성 낮음 • 1: 가능성 높음

① 속성 형식 및 역할 변경하기

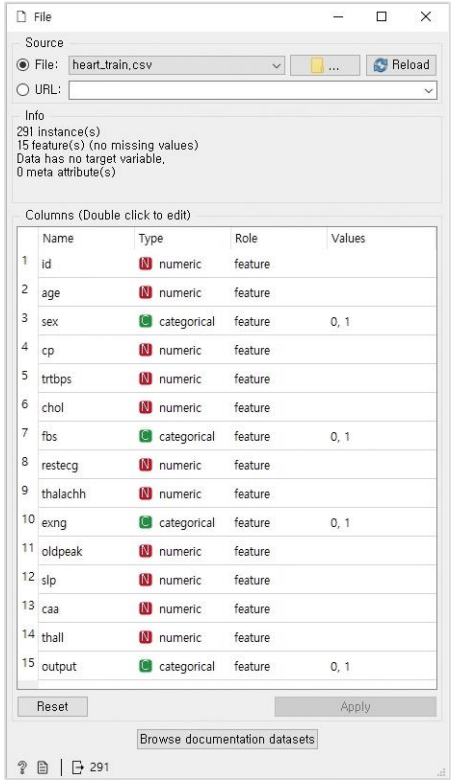


그림 3-5 속성의 형식과 역할 변경 전

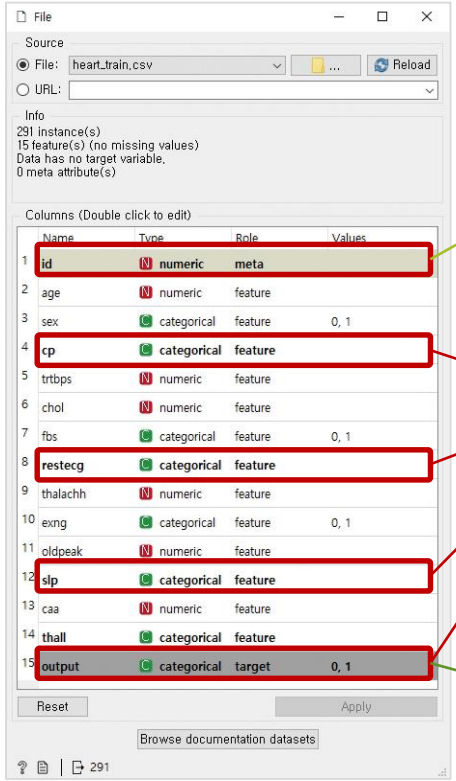


그림 3-6 속성의 형식과 역할 변경 후

훈련 및 예측에서 제외하고, 데이터 내용만 참조하므로 **meta**로 변경

범주형 데이터이므로 **categorical**로 변경

건강 정보 데이터를 바탕으로 심장마비를 예측하므로 **target**으로 변경

AI랑 친해지기

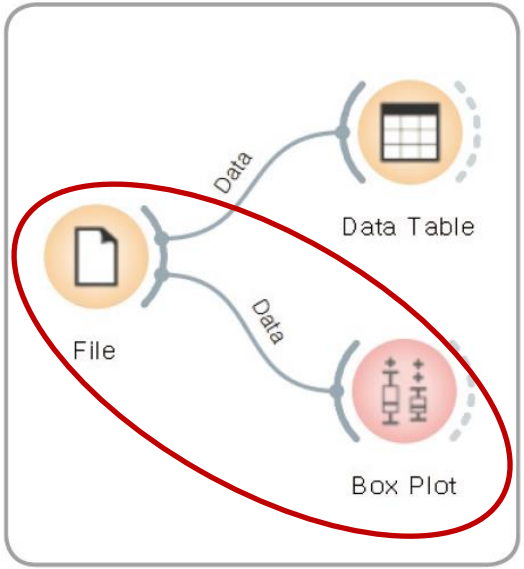
속성의 역할(Role)

- **feature** : 학습에 영향을 미치는 속성(독립 변수)
  target : 예측하고자 하는 속성(종속 변수)
  meta : 훈련 및 예측에서 제외하고 데이터의 내용을 참고하는 데 사용될 속성
  skip : 분석 시 무시하고 싶은 속성

4 데이터 시각화하기

① 값의 분포 확인하기

- Visualize 카테고리의 [Box Plot] 위젯을 가져와서 [File] 위젯에 연결



- [Box Plot] 위젯을 더블 클릭하여 선택한 속성(Variable)의 값의 분포를 확인
- 하위 그룹(Subgroups)으로 나누어서 그룹별로 속성의 차이값 확인

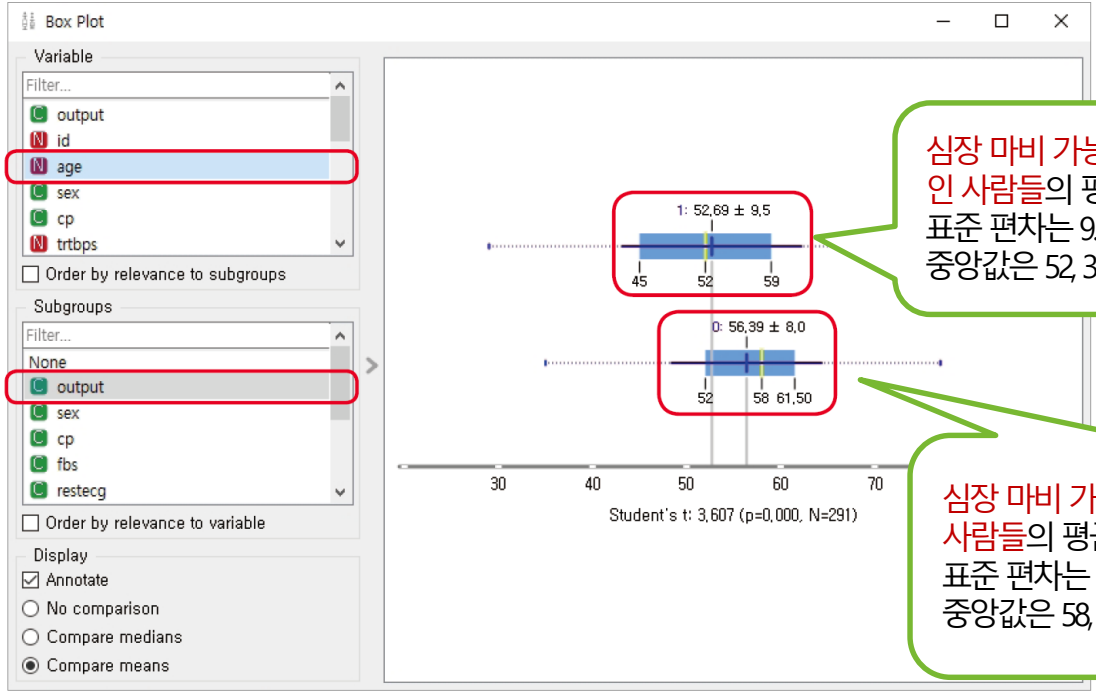


그림 3-7 나이(age) 속성을 심장 마비 가능성(output) 그룹으로 나누어 시각화한 결과

- 심장 마비 데이터의 각 속성들의 Box Plot을 확인하여 이상치와 통계적 특성을 파악

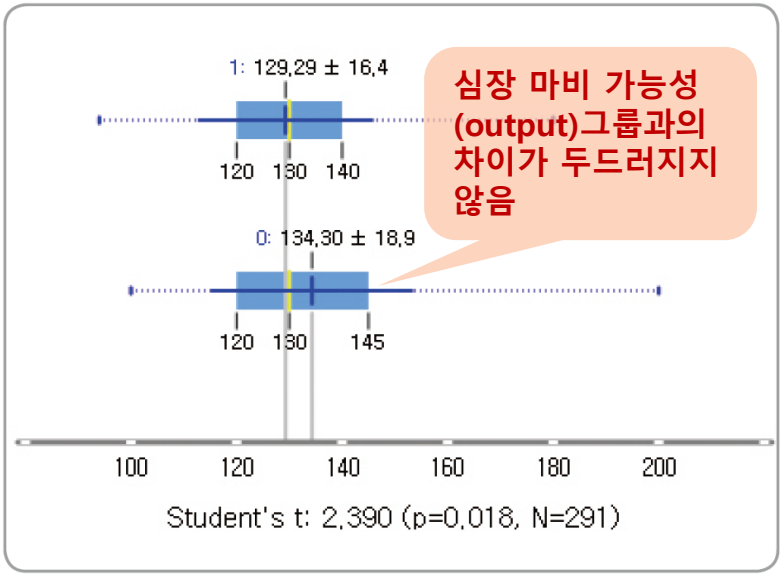


그림 3-8 안정 시 혈압(trtbps) 속성을 심장 마비 가능성(output) 그룹으로 나누어 시각화한 결과

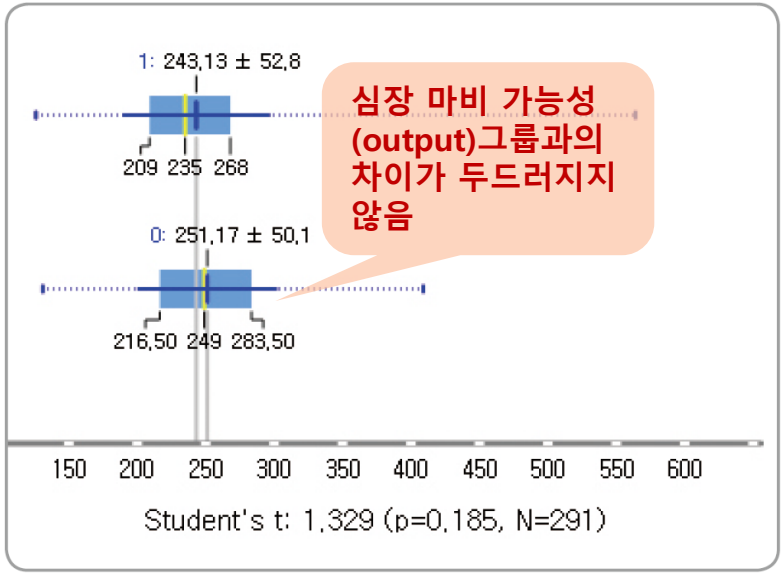


그림 3-9 혈청 콜레스테롤(chol) 속성을 심장 마비 가능성(output) 그룹으로 나누어 시각화한 결과

- 심장 마비 데이터의 각 속성들의 Box Plot을 확인하여 이상치와 통계적 특성을 파악

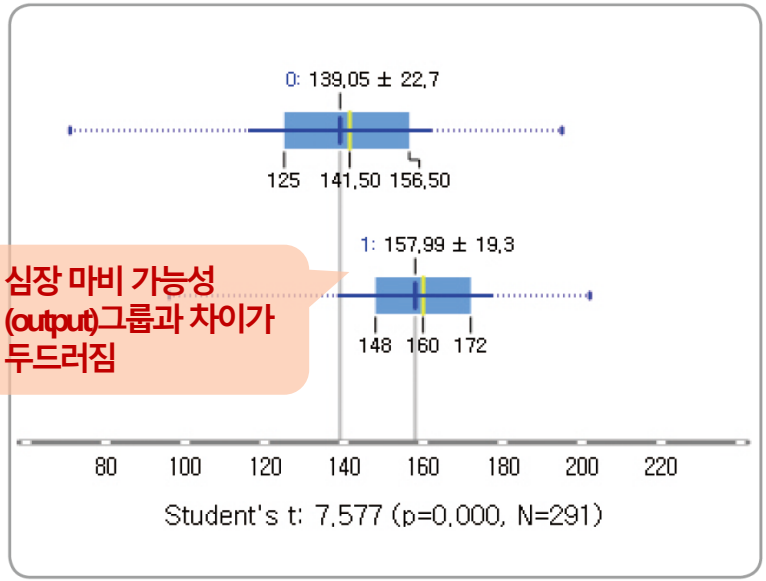


그림 3-10 최대 심박수(thalachh) 속성을 심장 마비 가능성(output) 그룹으로 나누어 시각화한 결과

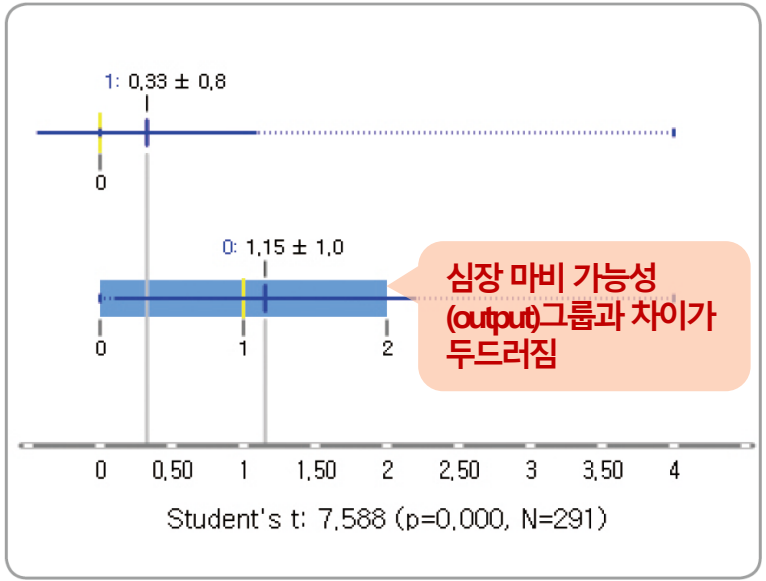


그림 3-11 착색된 주요 혈관 수(caa)속성을 심장 마비 가능성(output) 그룹으로 나누어 시각화한 결과

- 박스 플롯을 통해 데이터 분석에서 속성의 통계적 특성과 그룹별 차이를 확인

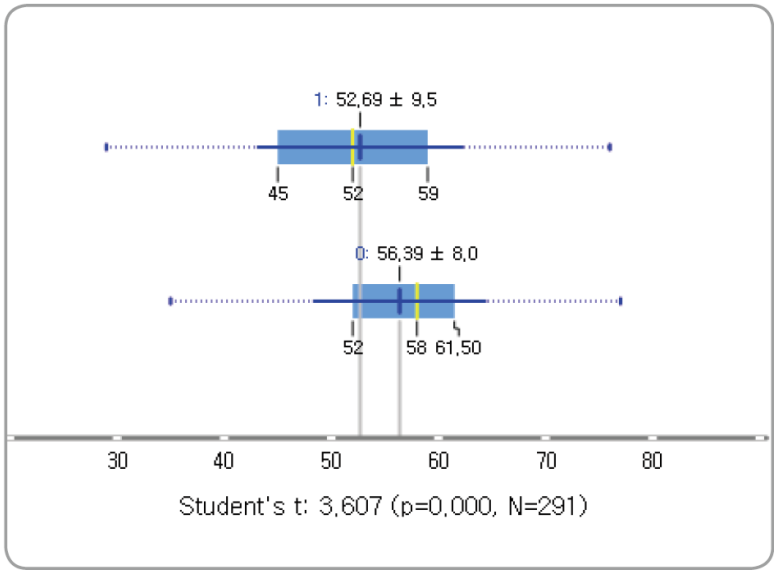


그림 3-12 나이(age) 속성을 심장 마비 가능성(output) 그룹으로 나누어 시각화한 결과

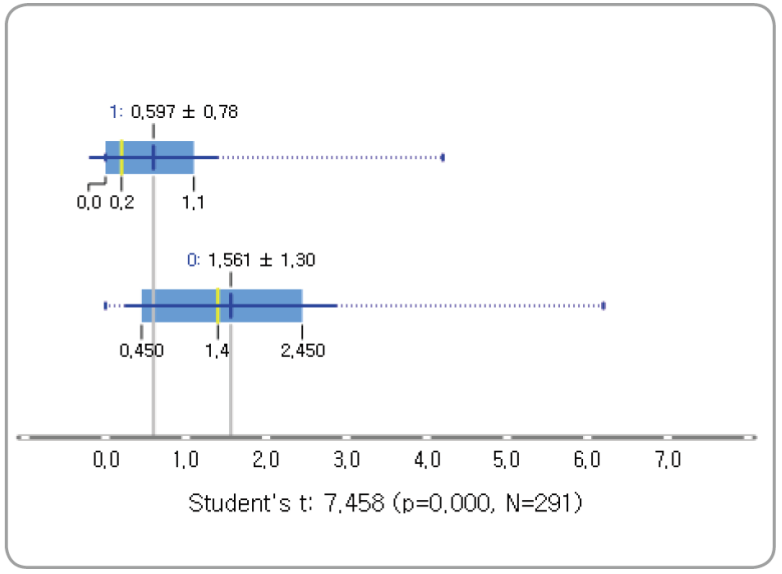


그림 3-13 심전도 결과(oldpeak) 속성을 속성을 심장 마비 가능성(output) 그룹으로 나누어 시각화한 결과

AI랑 친해지기

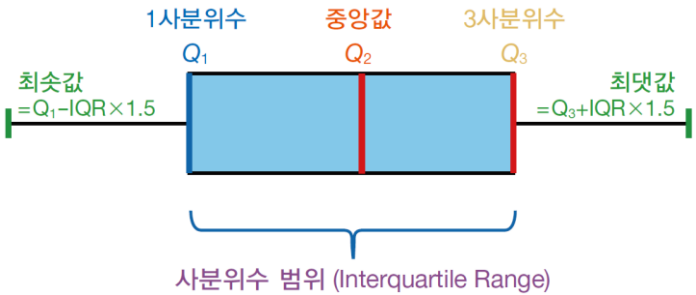
박스 플롯(Box Plot)과 박스 플롯값의 의미

박스 플롯은 데이터의 대략적인 분포와 개별적인 이상치를 동시에 보여 주며, 서로 다른 데이터 뭉치를 쉽게 비교할 수 있는 시각화 기법이다.

박스 플롯 작성법

사분위수 범위(1사분위수 - 3사분위수)를 구하고, 사분위범위로부터 1.5배 떨어진 점을 안울타리로 설정한다. 안울타리는 데이터의 이상치를 판단하는 기준이다.

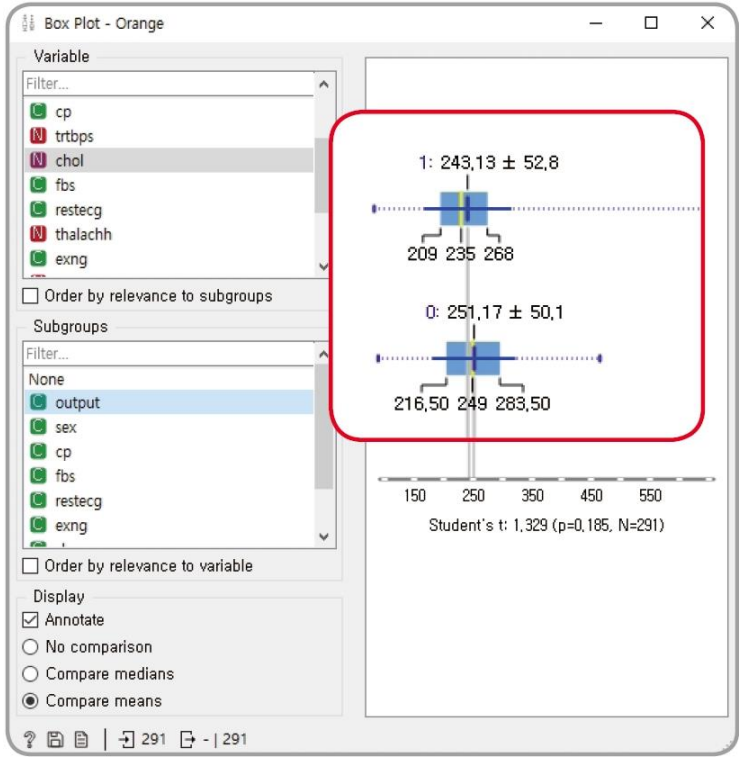
IQR = Q3 – Q1



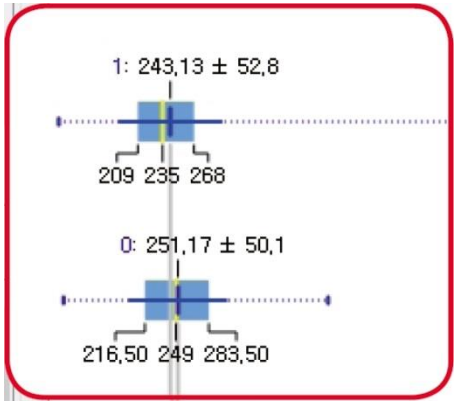
교재 60페이지 참조

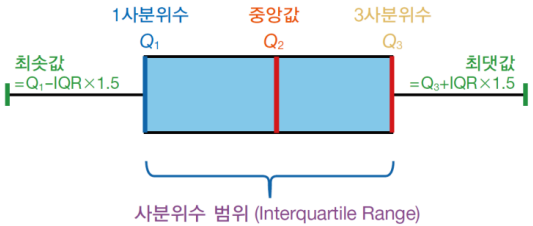
② 이상치 유무 확인하기

- 이상치를 가진 데이터가 있는지 확인하기 위해 [Box Plot] 위젯을 더블 클릭하여 Box Plot 창을 띄운 후 Variable과 Subgroups를 다음과 같이 설정하기



- 박스 플롯을 분석하여 데이터 이상치의 유무 확인하기





- **output이 1인 경우**

IQR = Q3 – Q1		사분위수 범위			
Q1	Q3	IQR	IQRx1.5	안	바깥
209	268	59	88.5	120.5	356.5

⇨ output이 1인 데이터 중 chol값이 120.5보다 작거나 356.5보다 큰 값이 있으면 이상치로 파악한다.

- **output이 0인 경우**

IQR = Q3 – Q1		사분위수 범위			
Q1	Q3	IQR	IQRx1.5	안	바깥
216.50	283.50	67	100.5	116	384

⇨ output이 0인 데이터 중 chol값이 116보다 작거나 384보다 큰 값이 있으면 이상치로 파악한다.

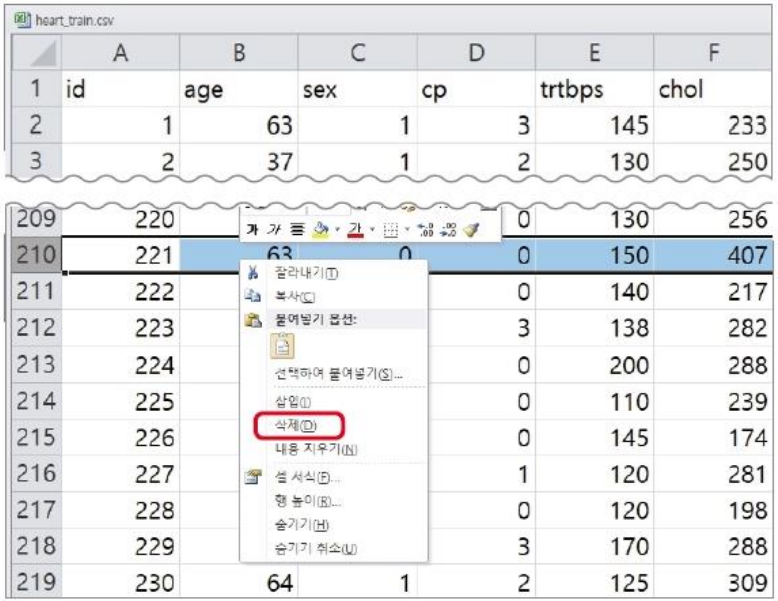
- 속성별 데이터의 이상치 유무 확인 결과

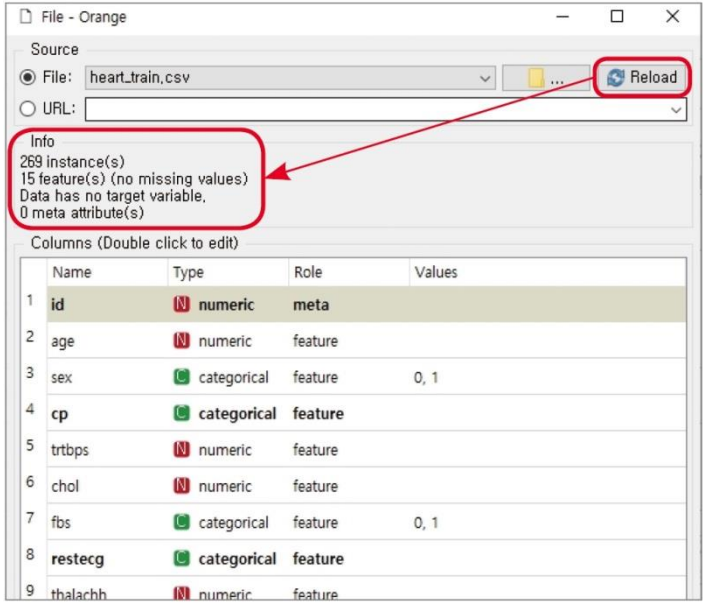
속성	이상치가 있는 id	
	output: 0	Output: 1
trtbps	224, 249	9, 102, 111
chol	221, 247	29, 40, 86, 97
thalachh	273	96, 137, 140
oldpeak	205, 222	2, 43, 102
caa	252	93, 159
합계	22	

caa는 박스 플롯으로 확인했을 때에는 이상치가 없지만 0~3이 아닌 데이터를 이상치로 파악하였다.

이상치가 있는 23개의 데이터 중에서, id 102번은 중복되는 id이므로 이상치가 있는id의 합계는 총 22개이다.

- heart_train.csv 파일을 열어서 22개의 해당 id의 행을 모두 삭제한 후,
  캔버스에 있는 [File] 위젯을 더블 클릭하여 창을 열고 파일을 Reload

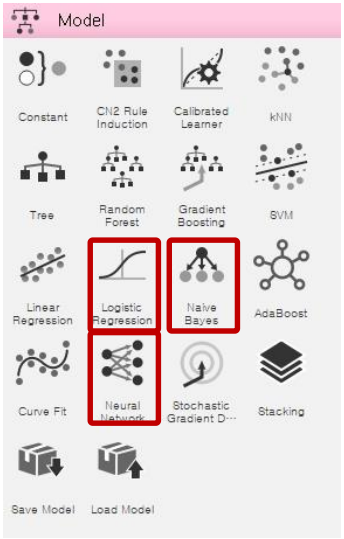


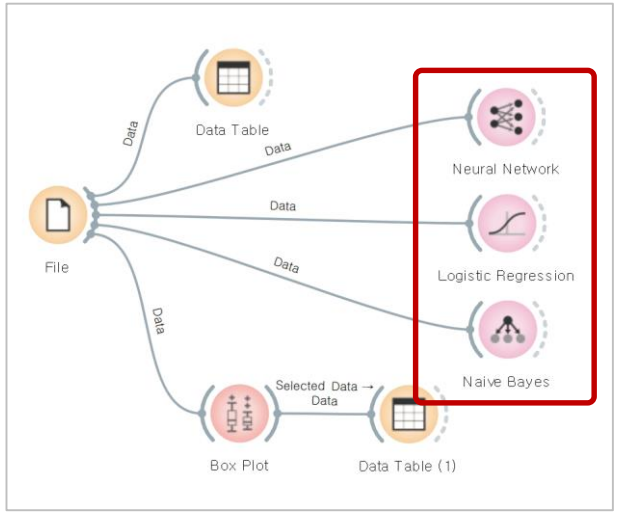


3 어떤 모델을 선택하고 학습시킬까?

1 학습 모델 선택하기

- Model 카테고리의 [Neural Network]위젯과 [Logistic Regression] 위젯, [Naive Bayes]위젯을 [File]위젯과 연결





2 학습시키기

- 별도의 실행 명령을 주지 않아도 위젯을 연결하면 모델 위젯이 자동으로 실행되어 각 모델이 데이터를 학습

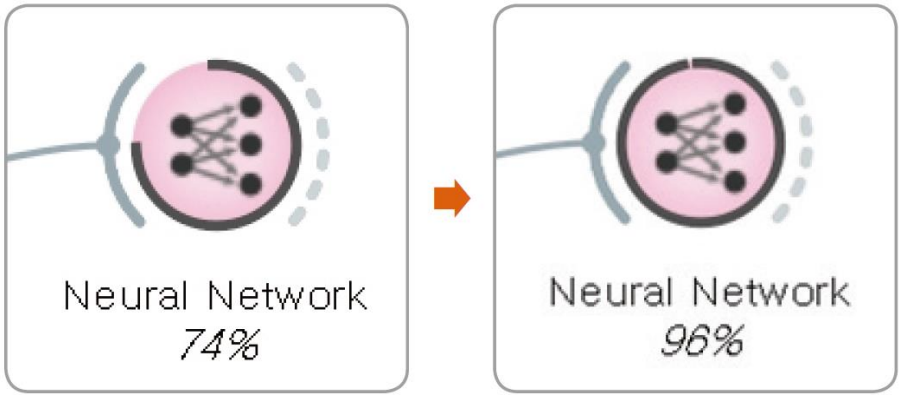


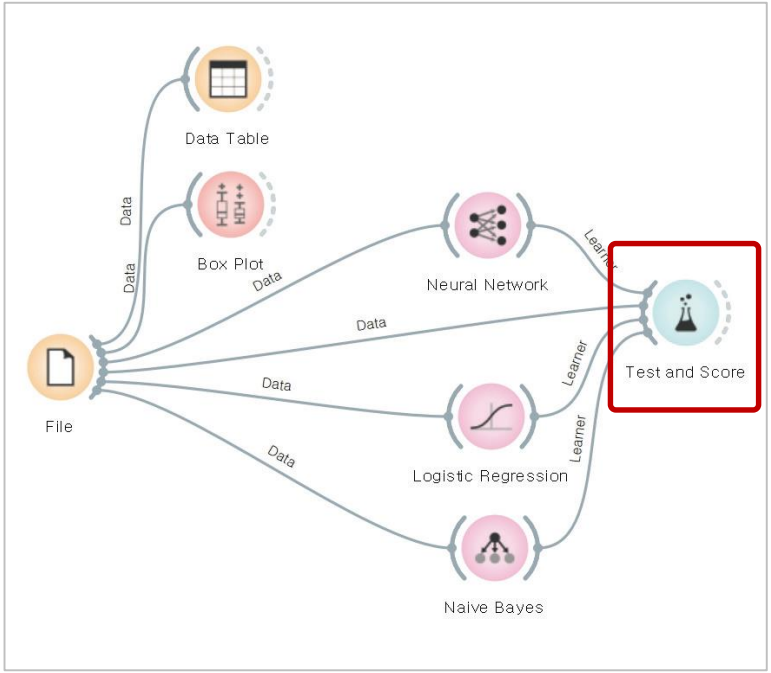
그림 3-14 Orange3에서 모델이 데이터를 학습하는 과정

4 모델의 성능을 확인해 보자!

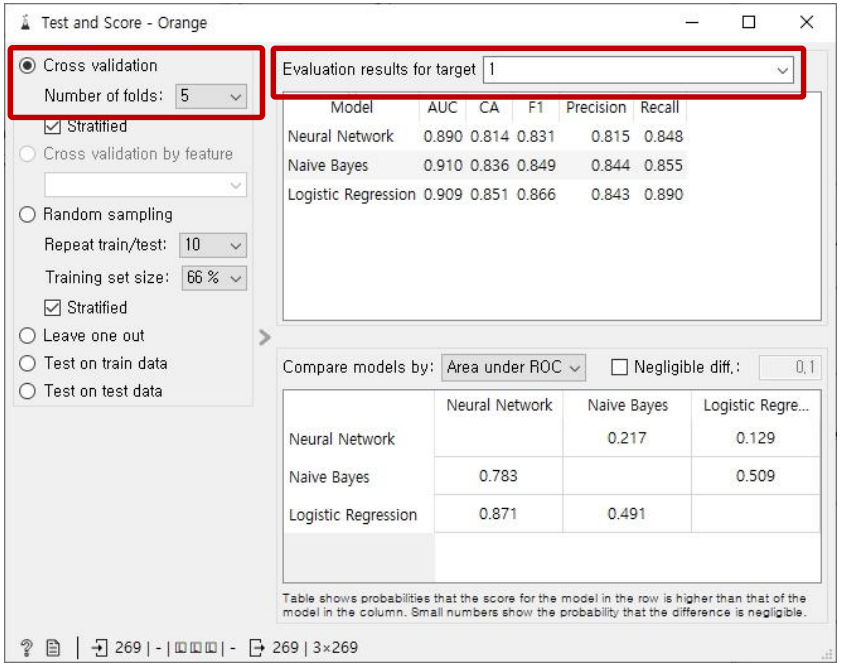
1 학습 결과 확인하기

① 성능 확인하기

- Evaluate 카테고리의 [Test and Score] 위젯을 가져온 후, 각 모델 위젯과 [File] 위젯에 연결



- [Test and Score] 위젯을 더블 클릭하여 각 모델의 성능을 확인
- 데이터 샘플링 방법 : Cross validation(교차 검증)
- 폴드의 수(Number of folds) : 5
- Target Class : 1



AI랑 친해지기

교차 검증(cross validation)

- 모델 학습 과정에서 훈련 데이터와 테스트 데이터를 나눌 때 단순히 한 번만 나누어서 검증하는 것이 아니라 k번 나누고 각각의 학습 모델의 성능을 비교하여 평균값으로 성능을 표시하는 방법이다.
- k가 5인 경우 데이터를 5등분 한 후 1/5을 검증 데이터로 사용하고 나머지 4/5는 훈련 데이터로 사용한다. 이것을 각 등분마다 돌아가면서 5번 시행하고 성능의 평균을 계산해서 모델의 성능을 표시한다.

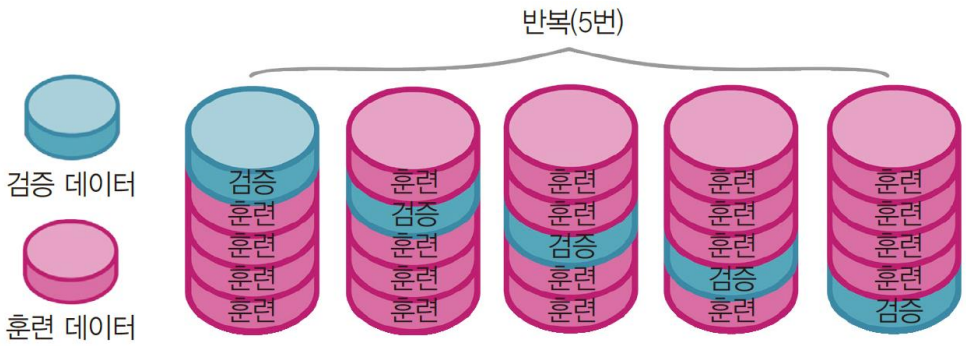
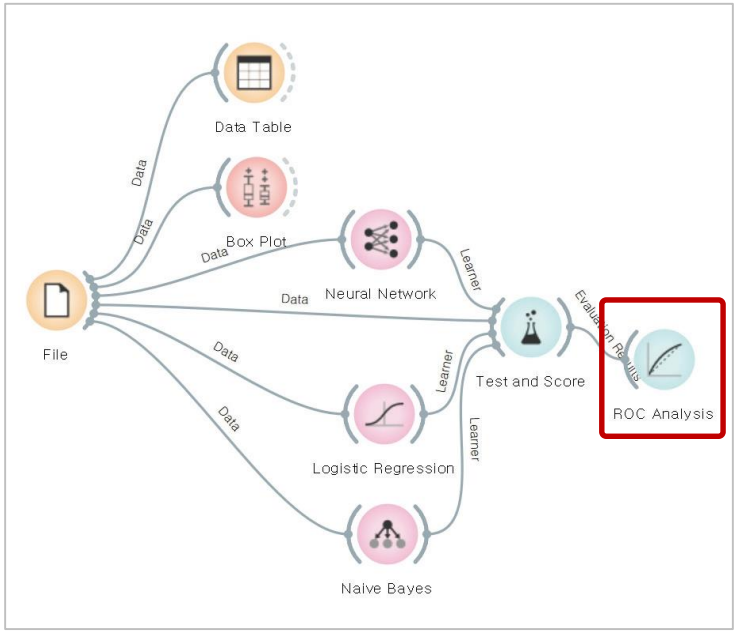


그림 3-15 k=5일 때 교차 검증 과정

② ROC 커브로 결과 분석하기

- Evaluate 카테고리의 [ROC Analysis] 위젯을 가져와서 [Test and Score] 위젯에 연결한 후, 3개 모델의 ROC 그래프를 살펴보기



② ROC 커브로 결과 분석하기

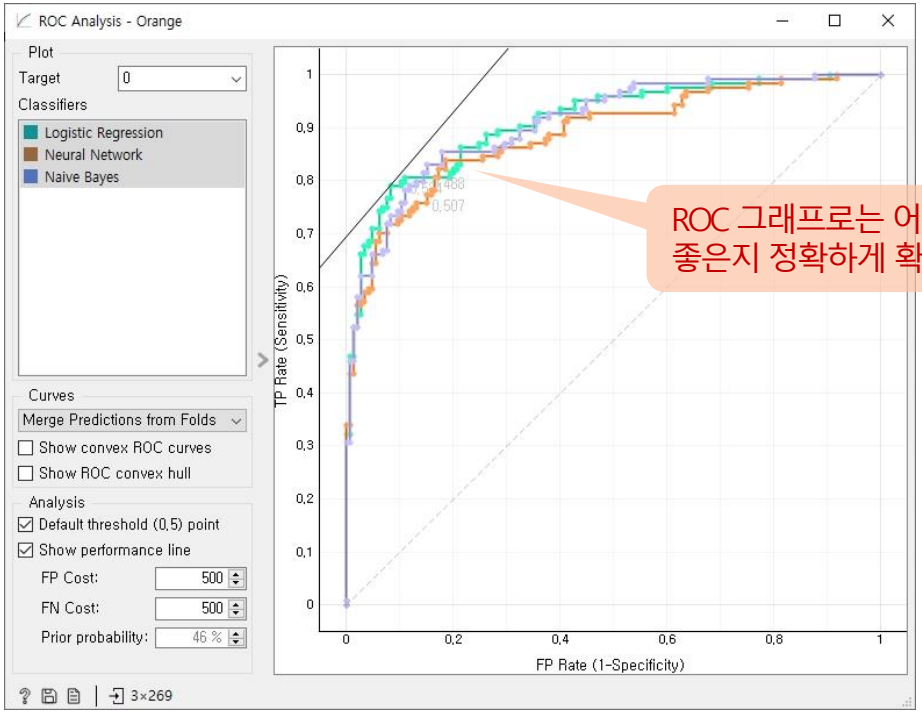


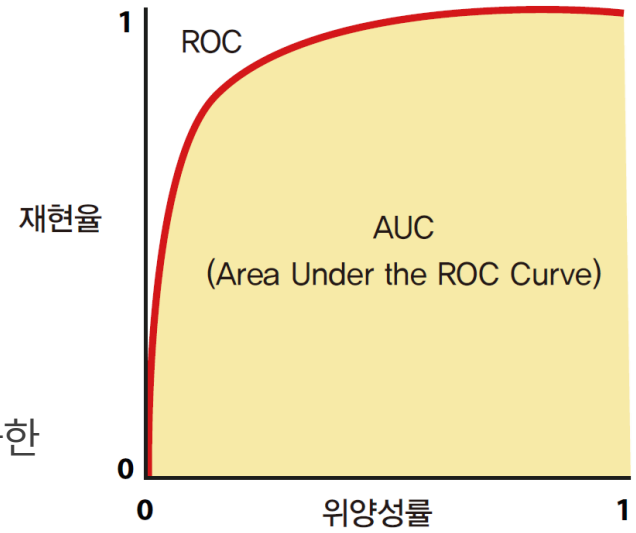
그림 3-16 3개 모델의 ROC 그래프

AI랑 친해지기

ROC 커브

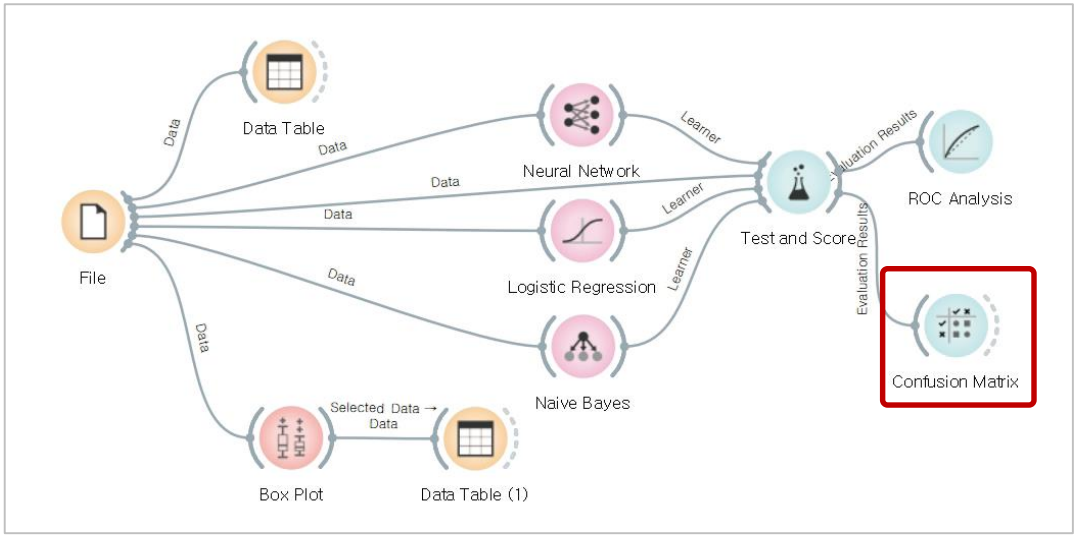
ROC 커브(Receiver Operating Characteristic Curve)는 여러 임계치들을 기준으로 **실제 True인 것 중에서 모델이 True라고 예측한 재현율(Recall; True Positive Rate)의 비율**과 **실제 False인 데이터 중에서 모델이 True라고 예측한 위양성률(Fall-out; False Positive Rate)의 비율**의 변화를 시각화한 그래프이다.

x축(위양성률)에 대한 y축(재현율)의 변화를 그래프로 표현하며 모델의 정확도를 판단한다. **AUC(Area under the ROC Curve)는 ROC 커브 아래쪽의 면적을 말하는데 1에 가까울수록 즉, AUC가 클수록 모델의 성능이 우수**하다.



③ 혼동 행렬로 결과 분석하기

- Evaluate 카테고리의 [Confusion Matrix] 위젯을 가져온 후, [Test and Score] 위젯에 연결
- [Confusion Matrix] 위젯을 더블 클릭하여 3개 모델의 혼동 행렬 비교



정확한 예측 측면에서 Logistic Regression 모델이 269건 중 229건으로 가장 우수하다. 하지만 심장 마비가 발생하지 않을 환자를 발생 가능성이 있다고 진단하는 경우(Actual: 0, Predicted: 1)도 환자에게는 부담스러운 일이다. 이 관점에서 보면 심장마비가 발생하지 않은 환자를 가장 잘 예측한 모델은 Naive Bayes이다.

심장 마비가 발생하지 않은 환자를 올바르게 예측한 경우 (Actual: 0, Predicted: 0)

Logistic Regression	Neural Network	Naive Bayes
100건	96건	101건

심장 마비가 발생한 환자를 올바르게 예측한 경우(Actual: 1, Predicted: 1)

Logistic Regression	Neural Network	Naive Bayes
129건	123건	124건

심장 마비가 발생한 환자를 발생하지 않을 것으로 예측한 경우 (Actual: 1, Predicted: 0)

Logistic Regression	Neural Network	Naive Bayes
16건	22건	21건

심장 마비가 발생하지 않은 환자를 발생할 것으로 예측한 경우 (Actual: 0, Predicted: 1)

Logistic Regression	Neural Network	Naive Bayes
24건	28건	23건

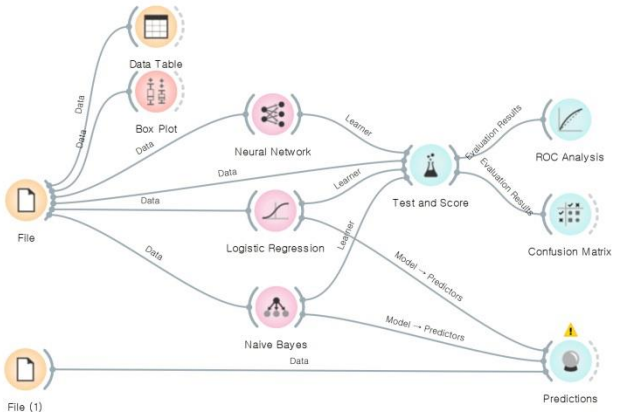
2 성능 결과 확인하기

① 테스트 데이터 불러오기

- Data 카테고리 - [File] 위젯을 가져온 후, 테스트 데이터(heart_test.csv) 파일 열기
- **형식 변경하기** : 'sex', 'cp', 'fps', 'restecg', 'exng', 'slp', 'thall' 속성의 형식을 categorical로 변경
- **역할 변경하기** : 'id'의 역할을 meta로, 'output'의 역할을 target으로 변경

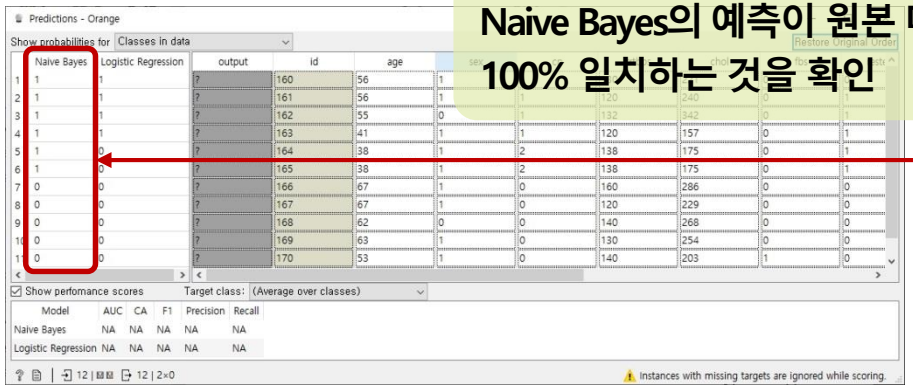
② 예측하기

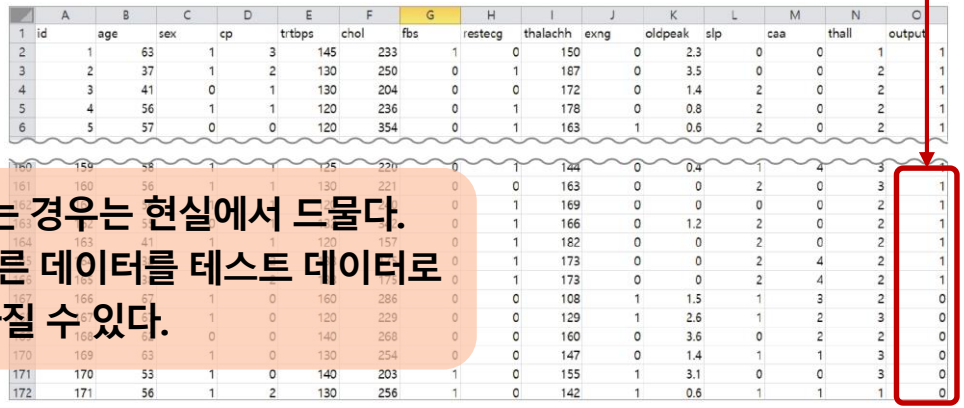
- Evaluate 카테고리의 [Predictions] 위젯을 [Logistic Regression] 위젯과 [Naive Bayes] 위젯에 연결하여 예측 결과 확인



② 예측하기

Naive Bayes의 예측이 원본 데이터(heart.csv)와 100% 일치하는 것을 확인





	id	age	sex	cp	trtbps	chol	fbs	restecg	thalachh	exng	oldpeak	slp	caa	thall	output
2	1	63	1	3	145	233	1	0	150	0	2.3	0	0	1	1
3	2	37	1	2	130	250	0	1	187	0	3.5	0	0	2	1
4	3	41	0	1	130	204	0	0	172	0	1.4	2	0	2	1
5	4	56	1	1	120	236	0	1	178	0	0.8	2	0	2	1
6	5	57	0	0	120	354	0	1	163	1	0.6	2	0	2	1
161	160	56	1	1	130	221	0	0	163	0	0	2	0	3	1
162	161	56	1	1	120	240	0	1	169	0	0	0	0	2	1
163	162	55	0	2	132	342	0	1	166	0	1.2	2	0	2	1
164	163	41	1	1	120	157	0	1	182	0	0	2	0	2	1
165	164	38	1	2	138	175	0	1	173	0	0	2	4	2	1
166	165	38	1	2	138	175	0	1	173	0	0	2	4	2	1
167	166	67	1	0	160	286	0	0	108	1	1.5	1	3	2	0
168	167	67	1	0	120	229	0	0	129	1	2.6	1	2	3	0
169	168	62	0	0	140	268	0	0	160	0	3.6	0	2	2	0
170	169	63	1	0	130	254	0	0	147	0	1.4	1	1	3	0
171	170	53	1	0	140	203	1	0	155	1	3.1	0	0	3	0
172	171	56	1	2	130	256	1	0	142	1	0.6	1	1	1	0

모델 예측 결과가 100% 일치하는 경우는 현실에서 드물다.
id가 160~171인 데이터 외에 다른 데이터를 테스트 데이터로 선정하는 경우 예측 결과가 달라질 수 있다.

그림 3-17 heart.csv 파일의 id 160~171번까지의 데이터

AI 전문가 되기

나이브 베이즈(Naive Bayes)

나이브 베이즈는 베이즈의 정리(Bayes' theorem)에 기반한 통계적 분류 모델이다.
데이터가 각 dass에 속할 확률을 계산하는 조건부 확률을 기반으로 분류를 시행하며
스팸 필터링, 비정상적인 상황 감지, 의학적 질병 진단 등에 활용된다.

나이브 베이즈는 간단하고 빠르며 효율적인 모델로 훈련할 때 훈련 데이터의
크기와 상관 없이 잘 동작할 뿐만 아니라 예측을 위한 추정 확률을 쉽게
얻을 수 있다는 장점이 있다.

베이즈의 정리(Bayes' theorem)

$$P(H \mid \mathrm{E}) = \frac{P(\mathrm{E} \mid H)P(H)}{P(\mathrm{E})}$$

- P(H|E): 새로운 정보(E)를 바탕으로 사전 확률P(H)을 수정한 사후 확률로 조건부 확률
- P(H): 어떤 사건 발생 주장에 대한 신뢰도인 사전 확률
- P(E): 증거(Evidence)
- P(E|H): 가능도(Likelihood)

교재 67페이지 참조

정리하기

지금까지 Orange3를 이용하여 환자의 건강 정보 데이터를 분석하고, 심장 마비가 발병할 가능성이 있는 환자를 예측할 수 있는 인공지능 모델을 만들어 보았다.
이 활동에서는 모델을 예측할 때 12개의 데이터로 테스트하여 100% 일치하는 결과를 얻었지만 데이터 셋의 크기를 다르게 하면 어떤 결과를 얻을 수 있을지 확인해 보는 것도 좋을 것이다.

이러한 예측 모델을 통해 건강 정보를 입력하면 심장 마비 발병에 대해 미리 알고, 이를 바탕으로 조기에 심장 마비 발병을 예방할 수 있는 의료 시스템을 구축할 수 있을 것이다. 또한, 한국인의 건강 정보 데이터를 수집하여 빅 데이터를 만듦으로써 한국형 심혈관 질환 예측 인공지능 모델을 만들 수 있을 것이다.

Q & A