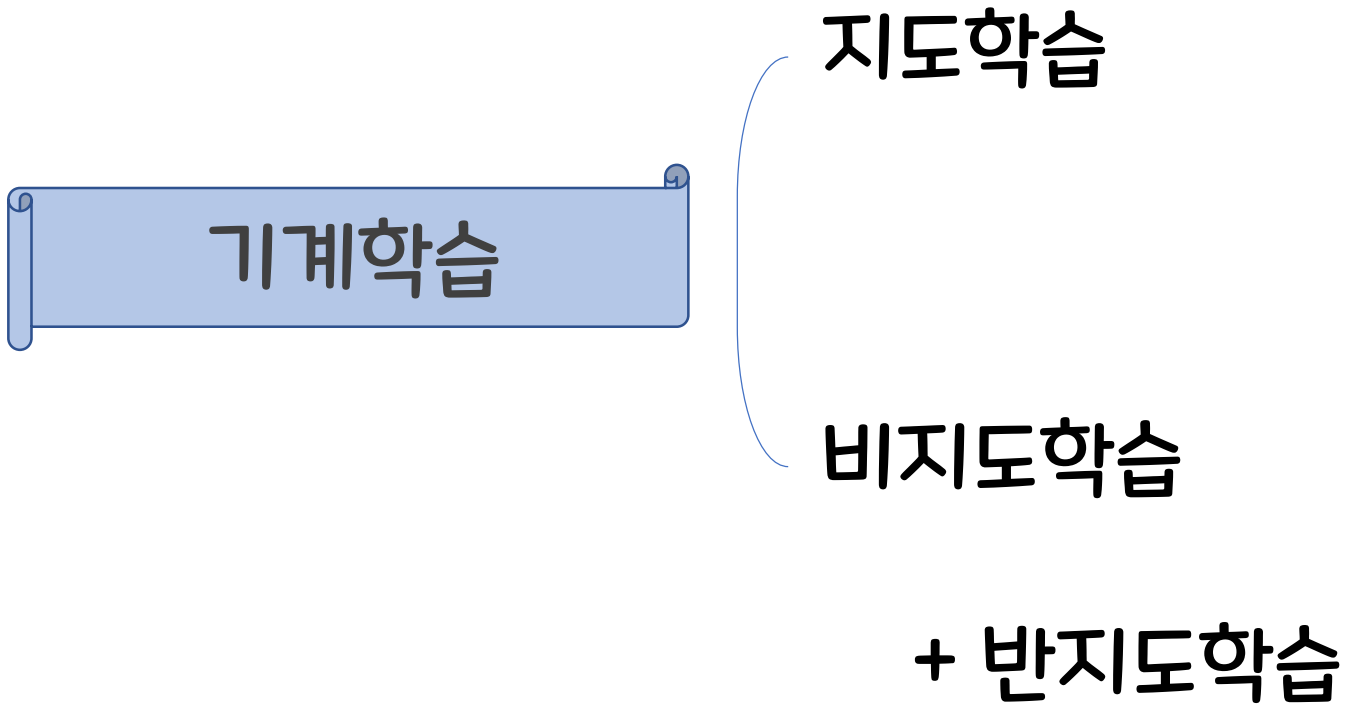
k-NN(k-Nearest Neigborhood) R로 쉽게 이해하기

조인식 조교

기계학습이 뭐지?

- **데이터에 근거하여 컴퓨터가 학습하거나 성능을 개선할 수 있는 방법을 조사**하는 이론
- 데이터에 근거하여 복잡한 패턴을 자동으로 인식하고 지능적인 의사결정을 할 수 있도록 자동으로 학습하는 분야


기계학습
지도학습
비지도학습
+ 반지도학습

기계학습

지도학습

지도할 데이터(Label, Class)가 있는지의 유무로 구분

비지도학습

지도학습

연령대코드	시도코드	신장(5Cm	체중(5Kg단	허리둘레	시력(좌)	시력(우)
10	27	175	75	86	1	1.2
11	41	160	65	85	1.5	1.2
11	43	150	55	80	1.5	1.5
12	48	175	70	85	1.2	1.2
6	11	160	50	67	0.4	0.5
9	46	170	55	64	1.1	1.2
10	31	175	70	90	1.5	1.5
16	48	155	60	81	0.7	0.5
16	27	165	60	74	0.4	0.6
13	48	160	60	80	0.8	0.2
7	46	170	80	86	1.5	1
6	11	155	50	66	0.9	0.6
7	41	170	80	92	0.6	0.4
15	11	155	65	81	0.8	0.5
15	28	160	65	79	0.7	0.2
11	48	165	80	93	0.7	0.8
9	41	170	75	81	1.2	1.2

독립변수(x)

구강검진 수검0
0
0
1
1
0
1
0
1
1
0
0
0
1
1
0
1
0

or

감마지티피
12
9
19
8
17
13
12
10
13
14
12
12
24
12
15
12
8

종속변수(y)

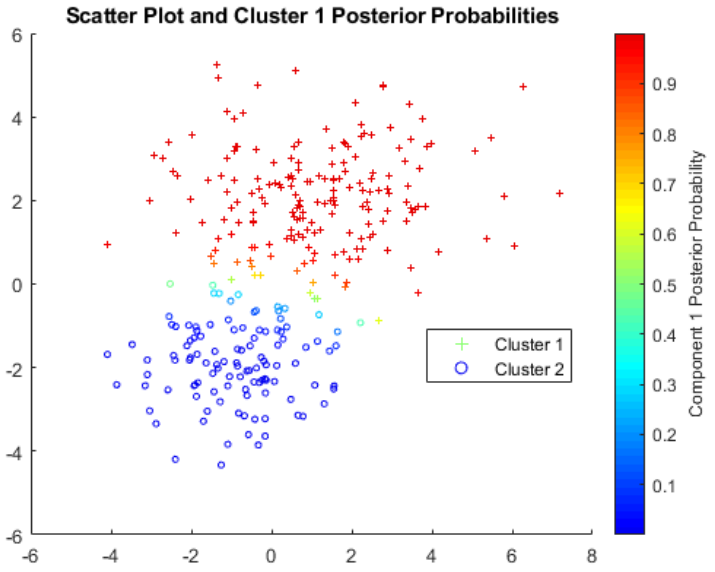
분석하고자 하는

우리!

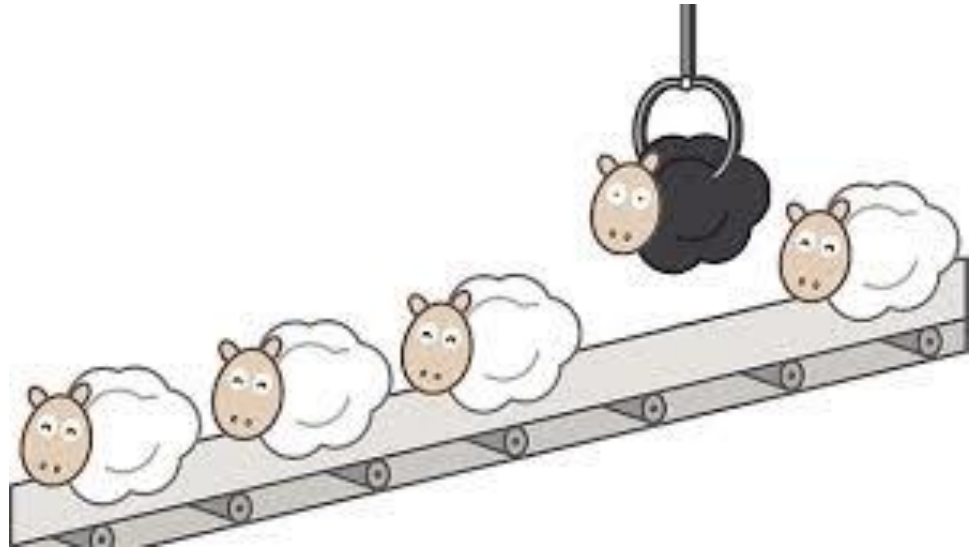
비지도학습

연령대코드	시도코드	신장(5Cm	체중(5Kg단	허리둘레	시력(좌)	시력(우)	구강검진 수검0
10	27	175	75	86	1	1.2	0
11	41	160	65	85	1.5	1.2	0
11	43	150	55	80	1.5	1.5	1
12	48	175	70	85	1.2	1.2	1
6	11	160	50	67	0.4	0.5	0
9	46	170	55	64	1.1	1.2	1
10	31	175	70	90	1.5	1.5	0
16	48	155	60	81	0.7	0.5	1
16	27	165	60	74	0.4	0.6	1
13	48	160	60	80	0.8	0.2	0
7	46	170	80	86	1.5	1	0
6	11	155	50	66	0.9	0.6	0
7	41	170	80	92	0.6	0.4	1
15	11	155	65	81	0.8	0.5	1
15	28	160	65	79	0.7	0.2	0
11	48	165	80	93	0.7	0.8	1
9	41	170	75	81	1.2	1.2	0

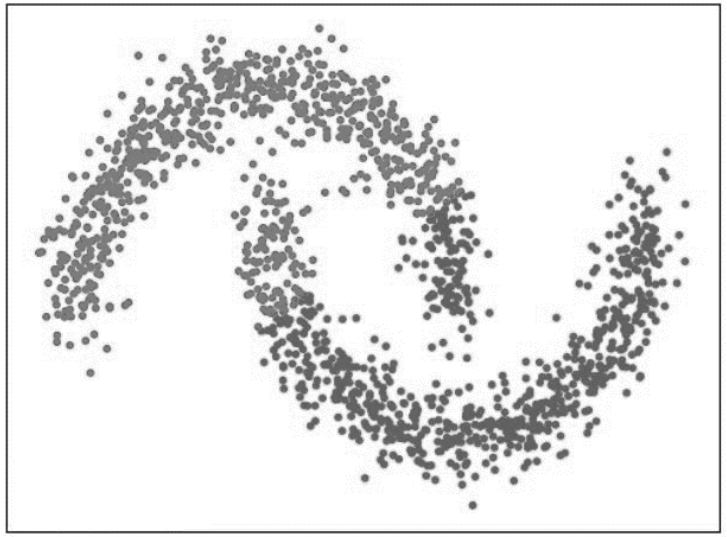
독립변수(x)



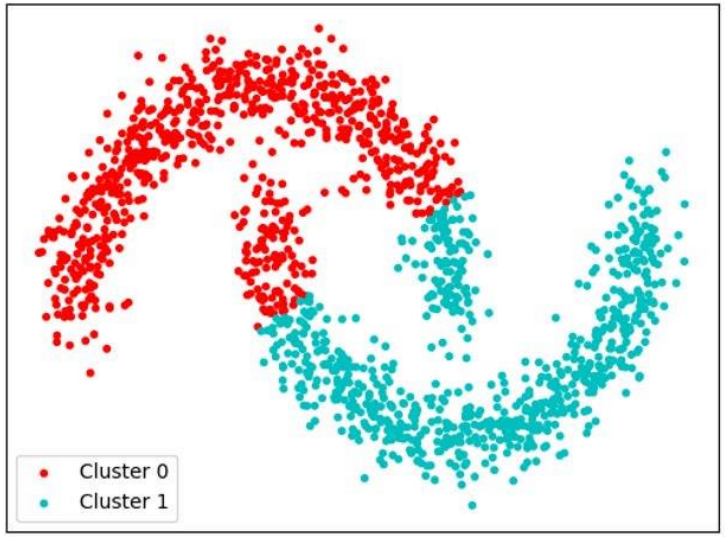
지도학습



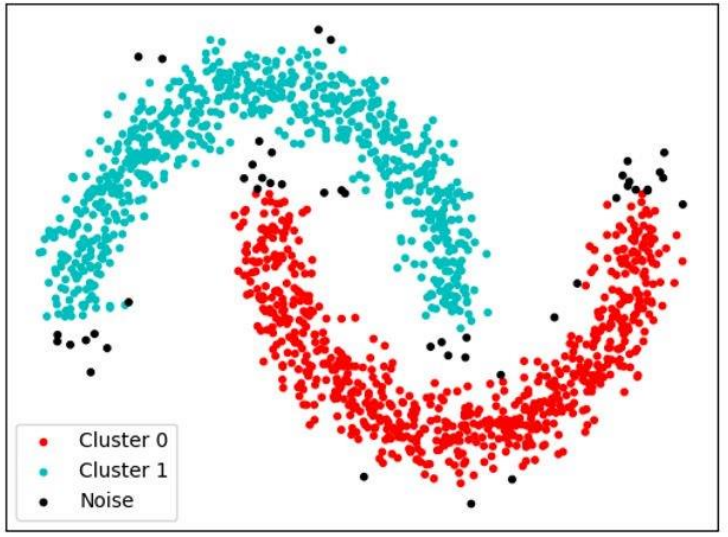
비지도학습



K-means



K-means



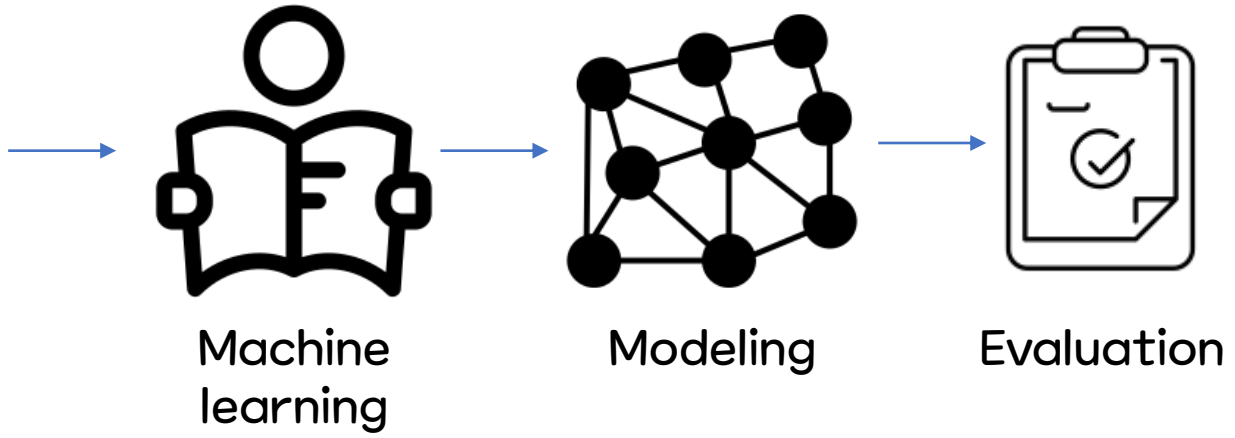
OPTICS

우리가 오늘 배울 것은 **지도학습!**

연령대코드	시도코드	신장(5Cm	체중(5Kg	허리둘레	시력(좌)	시력(우)	구강검진 수검
10	27	175	75	86	1	1.2	0
11	41	160	65	85	1.5	1.2	0
11	43	150	55	80	1.5	1.5	1
12	48	175	70	85	1.2	1.2	1
6	11	160	50	67	0.4	0.5	0
9	46	170	55	64	1.1	1.2	1
10	31	175	70	90	1.5	1.5	0
16	48	155	60	81	0.7	0.5	1
16	27	165	60	74	0.4	0.6	1
13	48	160	60	80	0.8	0.2	0
7	46	170	80	86	1.5	1	0
6	11	155	50	66	0.9	0.6	0
7	41	170	80	92	0.6	0.4	1
15	11	155	65	81	0.8	0.5	1
15	28	160	65	79	0.7	0.2	0
11	48	165	80	93	0.7	0.8	1
9	41	170	75	81	1.2	1.2	0

독립변수(x)

종속변수(y)



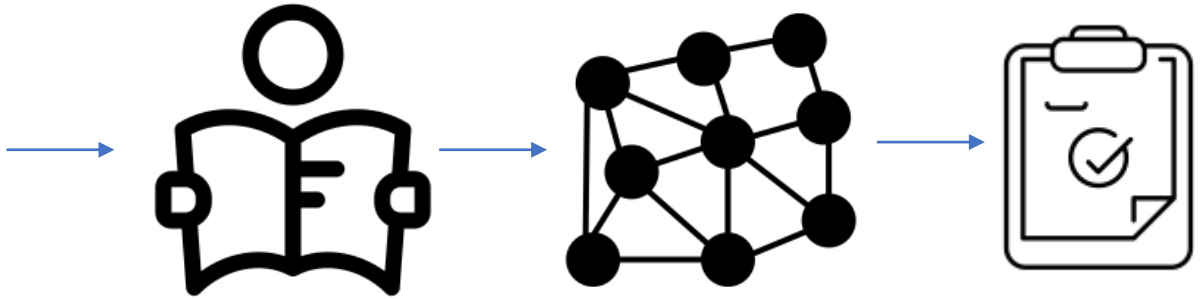
고등학교 시절 중간고사를 볼 때를 생각해봅시다…

연령대코드	시도코드	신장(5Cm단	체중(5Kg단	허리둘레	시력(좌)	시력(우)
10	27	175	75	86	1	1.2
11	41	160	65	85	1.5	1.2
11	43	150	55	80	1.5	1.5
12	48	175	70	85	1.2	1.2
6	11	160	50	67	0.4	0.5
9	46	170	55	64	1.1	1.2
10	31	175	70	90	1.5	1.5
16	48	155	60	81	0.7	0.5
16	27	165	60	74	0.4	0.6
13	48	160	60	80	0.8	0.2
7	46	170	80	86	1.5	1
6	11	155	50	66	0.9	0.6
7	41	170	80	92	0.6	0.4
15	11	155	65	81	0.8	0.5
15	28	160	65	79	0.7	0.2
11	48	165	80	93	0.7	0.8
9	41	170	75	81	1.2	1.2

연습문제들

구강검진 수검0
0
0
1
1
0
1
0
1
1
0
0
0
1
1
0
1
0

각 문제들의 답안들



중간고사에 필요한 것

여러분이 컴퓨터란 학생에게 공부시키고 시험을 보게하는 교수가 되었다 생각해봅시다

시험공부할 연습문제

연습문제의 답안지

중간고사 시험지

중간고사 시험답안지

…노력?

데이터?

연습문제

연령대코드	시도코드	신장(5Cm	체중(5Kg단	허리둘레	시력(좌)	시력(우)
10	27	175	75	86	1	1.2
11	41	160	65	85	1.5	1.2
11	43	150	55	80	1.5	1.5
12	48	175	70	85	1.2	1.2
6	11	160	50	67	0.4	0.5
9	46	170	55	64	1.1	1.2
10	31	175	70	90	1.5	1.5
16	48	155	60	81	0.7	0.5
16	27	165	60	74	0.4	0.6
13	48	160	60	80	0.8	0.2
7	46	170	80	86	1.5	1
6	11	155	50	66	0.9	0.6
7	41	170	80	92	0.6	0.4
15	11	155	65	81	0.8	0.5
15	28	160	65	79	0.7	0.2
11	48	165	80	93	0.7	0.8
9	41	170	75	81	1.2	1.2

구강검진 수검0
0
0
1
1
0
1
0
1
1
0
0
0
1
1
0
1
0

연습문제 답안지

중간고사 시험지

중간고사 시험답안

독립변수(x)

종속변수(y)

Train data

Test data

연령대코드	시도코드	신장(5Cm	체중(5Kg단	허리둘레	시력(좌)	시력(우)
10	27	175	75	86	1	1.2
11	41	160	65	85	1.5	1.2
11	43	150	55	80	1.5	1.5
12	48	175	70	85	1.2	1.2
6	11	160	50	67	0.4	0.5
9	46	170	55	64	1.1	1.2
10	31	175	70	90	1.5	1.5
16	48	155	60	81	0.7	0.5
16	27	165	60	74	0.4	0.6
13	48	160	60	80	0.8	0.2
7	46	170	80	86	1.5	1
6	11	155	50	66	0.9	0.6
7	41	170	80	92	0.6	0.4
15	11	155	65	81	0.8	0.5
15	28	160	65	79	0.7	0.2
11	48	165	80	93	0.7	0.8
9	41	170	75	81	1.2	1.2

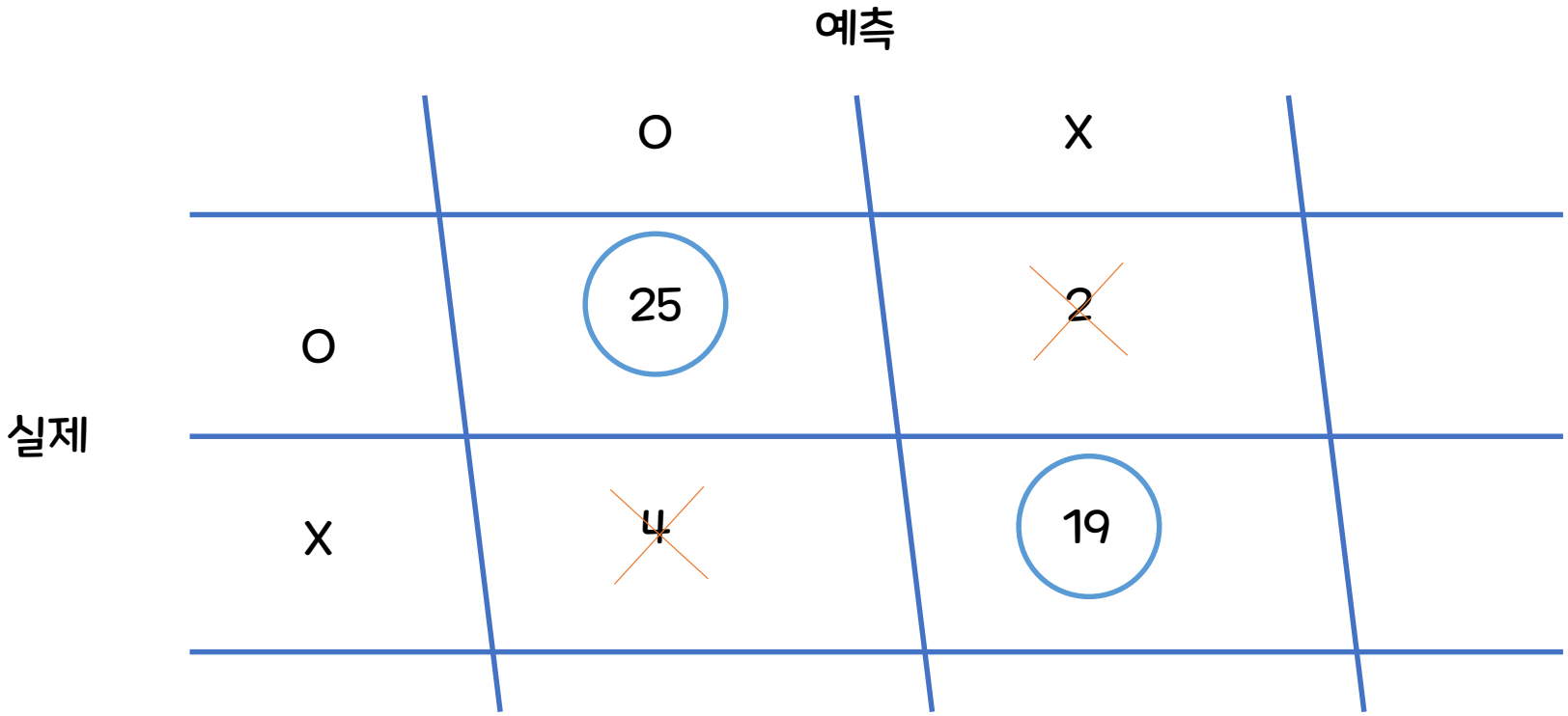
독립변수(x)

구강검진 수검0
0
0
1
1
0
1
0
1
1
0
0
0
1
1
0
1
0

종속변수(y)

Label of Train data

Label of Train data


예측
O
X
실제
O
X
25
2
4
19

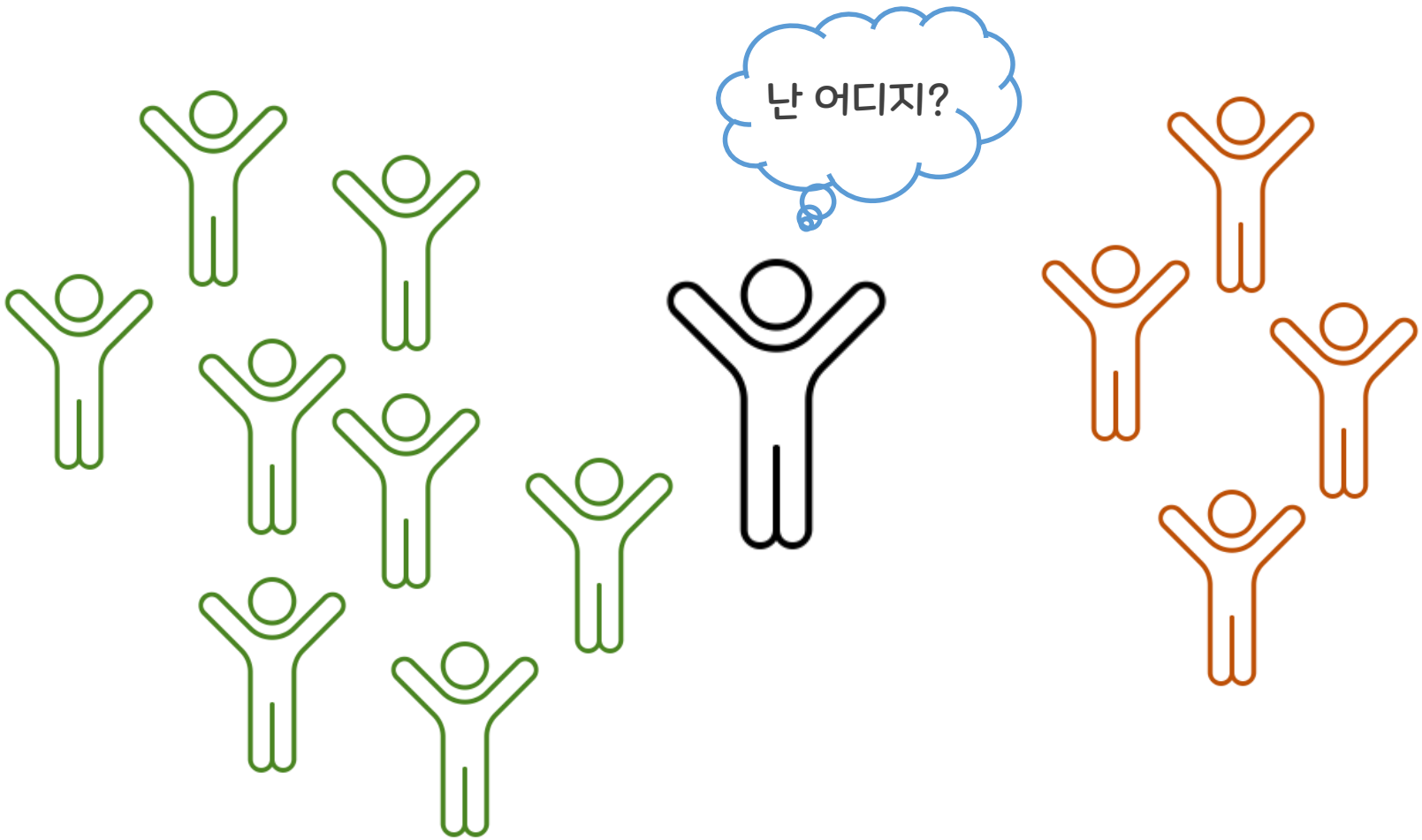
n=165	**Predicted: NO**	**Predicted: YES**	
Actual: NO	TN = 50	FP = 10	60
Actual: YES	FN = 5	TP = 100	105
	55	110	

Confusion Matrix

여기서 평가 지표를 만들어 내 모델이 얼마나 좋은지를 판단할 수 있다!
:Recall, Precision, Accuracy, F-socre, etc…:

k-NN은 뭘까?

kNN의 아이디어!



Green Label

Orange Label

kNN의 아이디어!
1. K개의 최근접한 이웃들의 label을 찾아본다. K는 사용자가 지정.

Green Label

Orange Label

kNN의 아이디어!

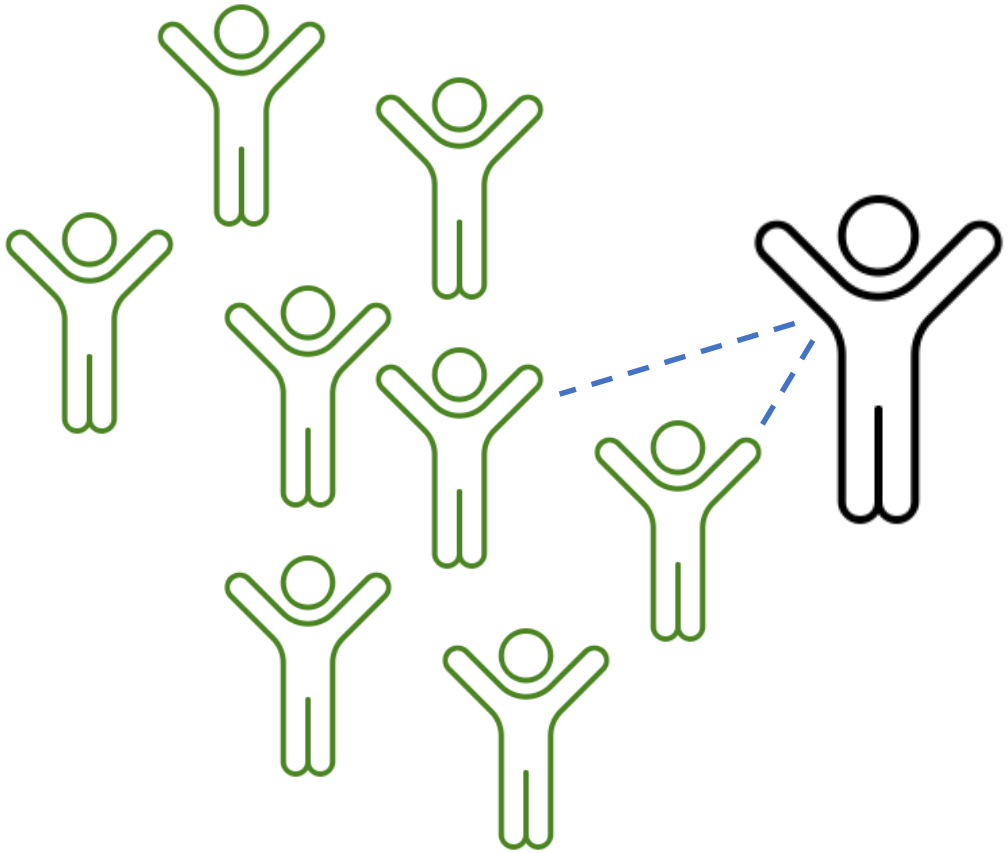
1. K개의 최근접한 이웃들의 label을 찾아본다. K는 사용자가 지정.
2. 인자들과의 거리계산

Green Label

Orange Label

kNN의 아이디어!

1. K개의 최근접한 이웃들의 label을 찾아본다. K는 사용자가 지정.
2. 인자들과의 거리계산

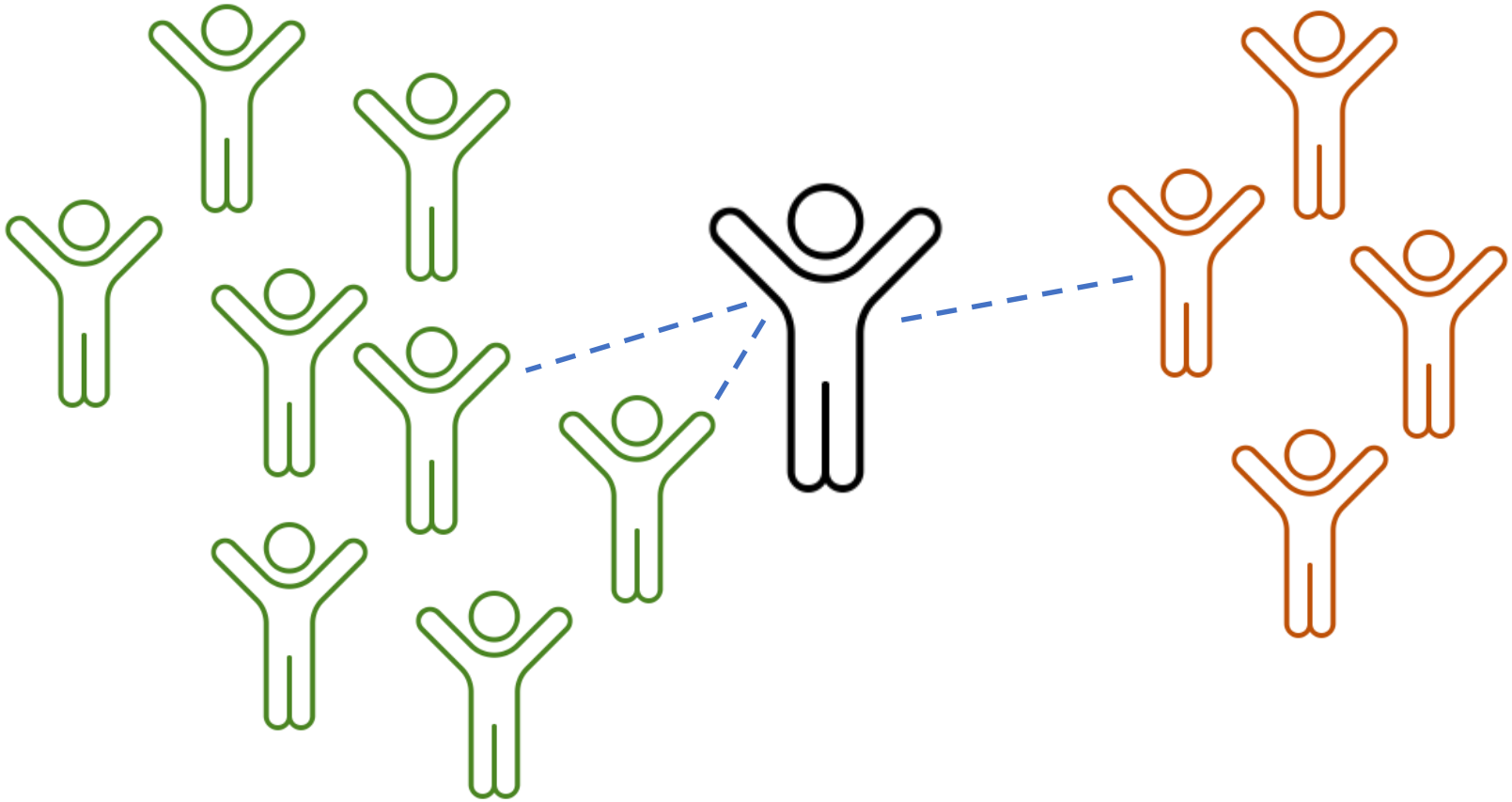


Green Label

Orange Label

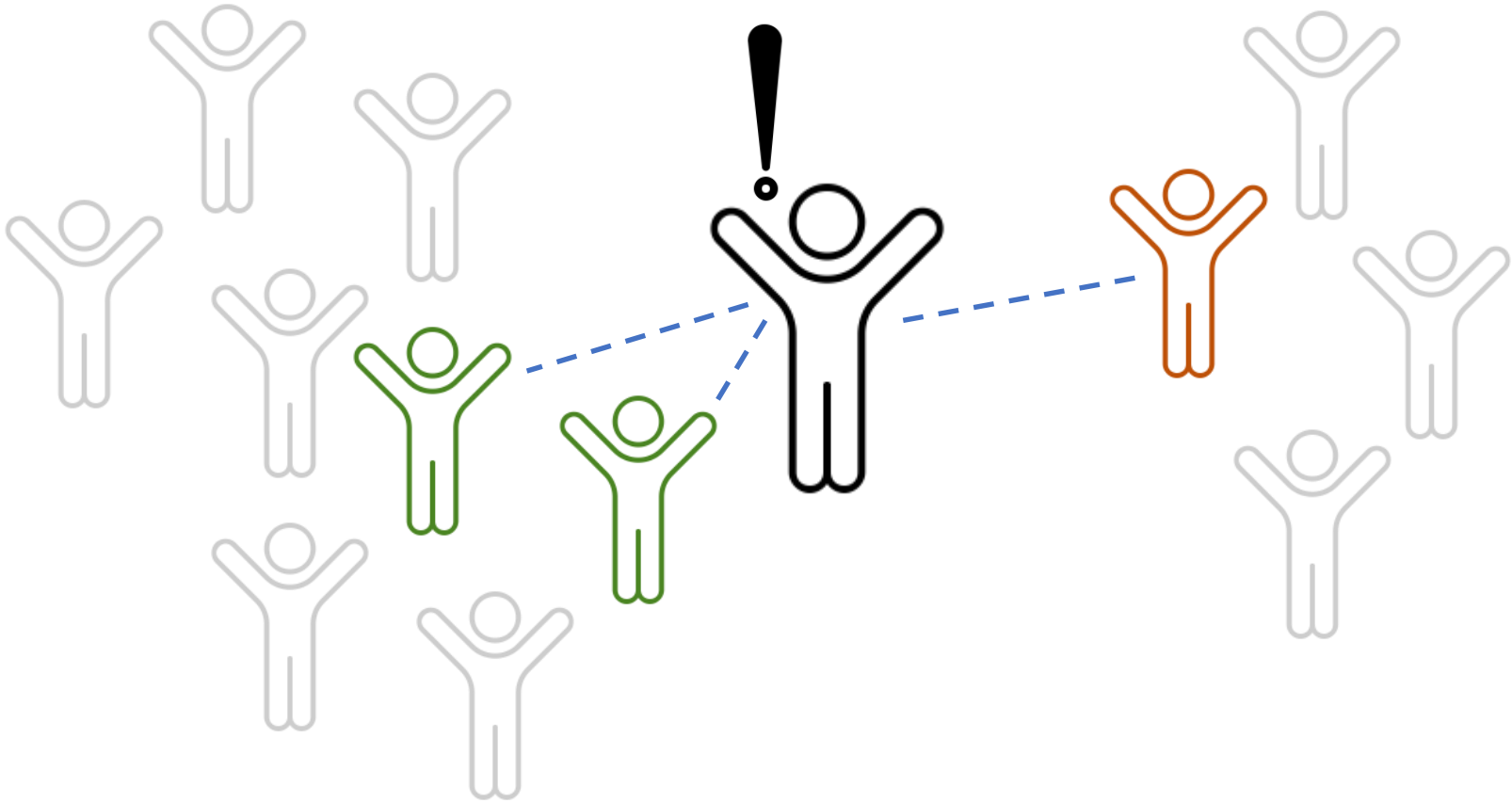
kNN의 아이디어!

1. K개의 최근접한 이웃들의 label을 찾아본다. K는 사용자가 지정.
2. 인자들과의 거리계산



kNN의 아이디어!

1. K개의 최근접한 이웃들의 label을 찾아본다. K는 사용자가 지정.
2. 인자들과의 거리계산
3. 최근접 이웃들의 label을 토대로 과반수인 label을 본인의 label로 선정



Green Label

Orange Label

kNN의 아이디어!

1. K개의 최근접한 이웃들의 label을 찾아본다. K는 사용자가 지정.
2. 인자들과의 거리계산
3. 최근접 이웃들의 label을 토대로 과반수인 label을 본인의 label로 선정

Green Label

Orange Label

kNN의 아이디어!

1. K개의 최근접한 이웃들의 label을 찾아본다. K는 사용자가 지정.
2. 인자들과의 거리계산
3. 최근접 이웃들의 label을 토대로 과반수인 label을 본인의 label로 선정

Green Label

Orange Label

이해가 됐다면
실습으로 알아가봅시다!

MY LITTLE PROJECT

현재 국내에 큰 유명세를 타고 있는 MOBA(Multiplayer Online Battle Arena) 게임중 하나인 League Of Legends(LOL)은 다양한 API와 데이터셋을 공개하고 있다.

우리는 LOL에서 개방한 데이터셋을 통하여 "게임 내에서 17분간의 획득한 골드를 토대로 게임의 승패를 예측할 수 있는가" 라는 가설을 토대로 kNN 모델을 구축한다.

1. 데이터셋을 구해야 한다. 데이터셋을 구하기 위하여 google에 [league of legends match dataset]을 검색한 뒤 필요한 데이터를 선정하고, 이를 알맞게 전처리한다.

2. 데이터를 train과 test로 나눈다. 총 데이터 7620건 중 5000건을 train으로 하고 나머지를 test로 설정한다.

3. kNN 모델을 구축한다. K값은 과제자가 최적이라 판단되는 값으로 선정한다.