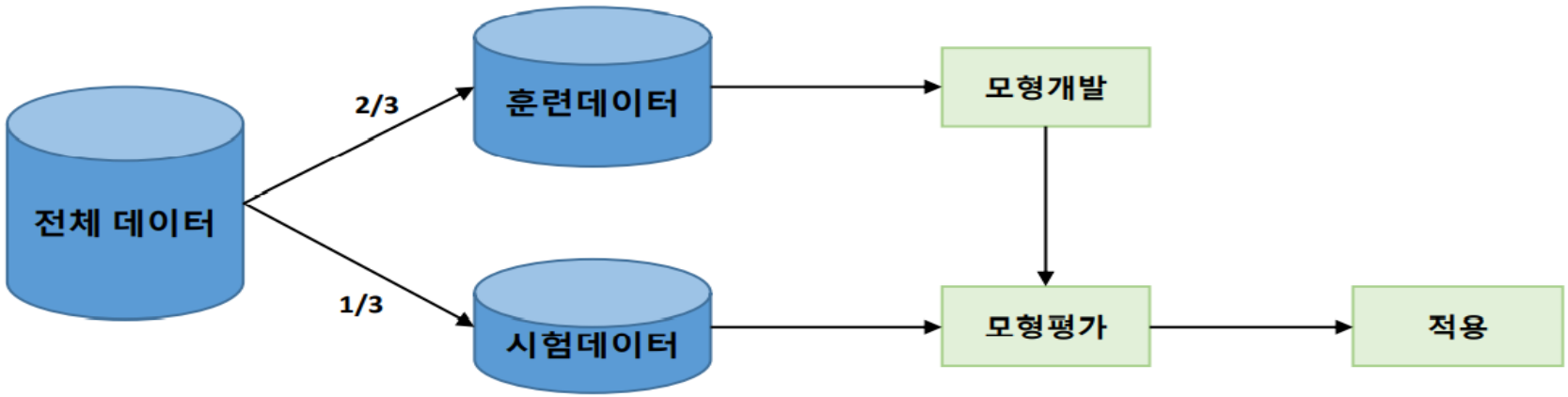
KNN: 최근접 이웃을 사용한 분류

숙명여자대학교 경영학부 오중산

머신러닝 소개

- 머신러닝(machine learning: ML)이란?
  - ◆ 전통적인 통계분석 방법에 기반한 데이터마이닝
  - ◆ ML의 4대 목표는 예측/분류/군집화/연관성 파악
  - ◆ 전통적인 통계분석방법과의 차이는 1)데이터 크기 증가 2)훈련(모형개발)과 검증/시험(모형평가)의 두 단계로 나누어 진행
    - ▪ 데이터가 커지면서 AI를 이용하게 됨에 따라 ML이라고 부르게 됨



KNN 소개

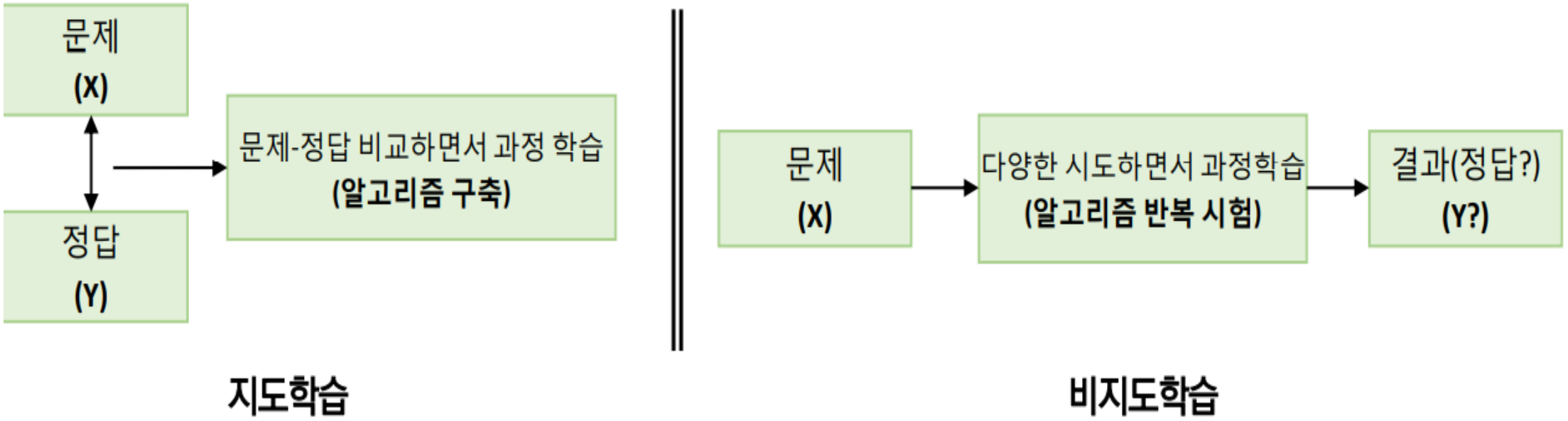
- KNN(K-Nearest Neighbors) 분석이란?
  - 인접한 K개 기존 사례와의 유사성을 기준으로 새로운 사례를 분류하는 분석방법
    - 기존 사례는 계량형 척도로 측정된 IV를 기준으로 여러 집단(class) 중 하나에 속해 있음
      - 여러 집단은 비계량형 척도로 측정된 DV(혹은 class 변수) 결과값
      - 예: 직원 이탈(종속변수: 이탈/미이탈)과 관련된 IV(만족도, 근무시간, 고과평가 등)
    - 새로운 사례와 기존 사례와의 유사성은 Euclidean distance를 기준으로 측정
      - 새로운 사례는 Euclidean distance를 기준으로 가까운 사례가 많이 속해 있는 집단으로 분류됨

KNN 소개

- KNN의 장·단점
  - ◆ KNN의 장점
    - 단순하고 효율적이며, 훈련단계가 빠름
    - 데이터의 분포에 대한 가정이 필요 없음
  - ◆ KNN의 단점
    - 추상화(별도 모델 수립)가 없어서 집단과 변수 간의 인과관계 파악이 어려움
    - K를 적절히 선택하지 않으면 결과가 제대로 나오지 않거나 왜곡됨

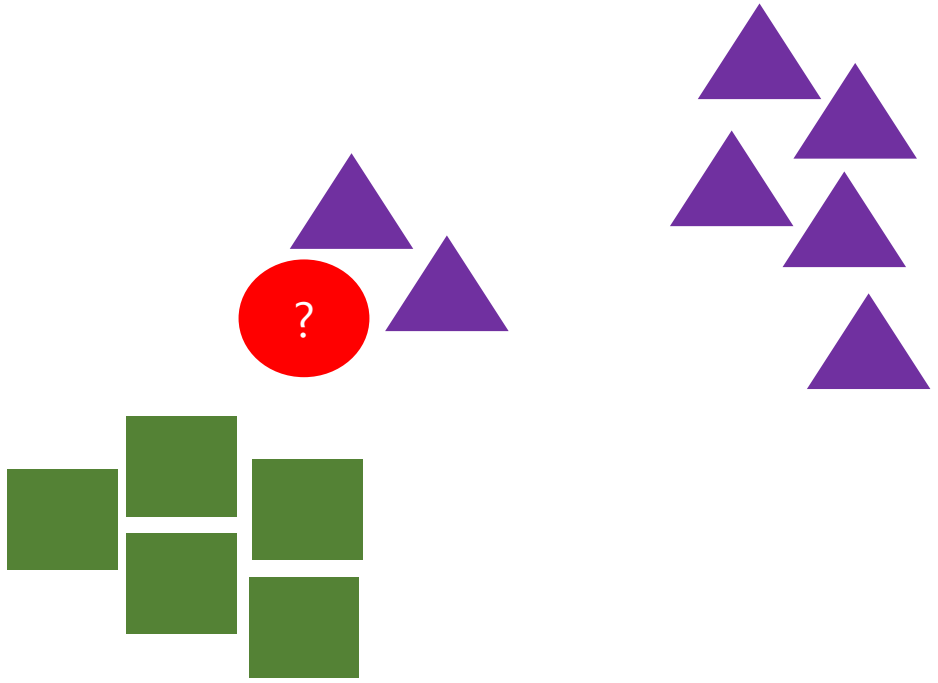
KNN과 머신러닝

- KNN과 머신러닝의 관계
  - ◆ KNN은 분류를 목적으로 하는 ML 지도학습의 한 유형



KNN 개요

- K를 얼마로 해야 하는가?
  - K값에 따라 분류 정확성에 편차가 발생함
    - 그림에서 K=3, K=4, K=5인 경우 빨간 동그라미는 어느 집단으로 분류 되는가?
    - K가 작아지면 이상치에 영향을 받고, K가 커지면 큰 집단으로 분류되는 편향이 발생함



암진단 사례

- 데이터 소개
  - Wisconsin university에서 기증한 데이터이며, $n = 569$
    - 보다 자세한 정보는 http://archive.ics.uci.edu/ml 참고
  - 32개 변수로 구성됨
    - id: 환자를 식별하기 위한 변수
    - diagnosis: class 변수로서 양성(benign, B) 혹은 악성(malignant, M) 두 가지 결과
    - IV 혹은 feature: 세포핵의 10개 특성에 대한 평균, 표준오차, 최대값

암진단 사례

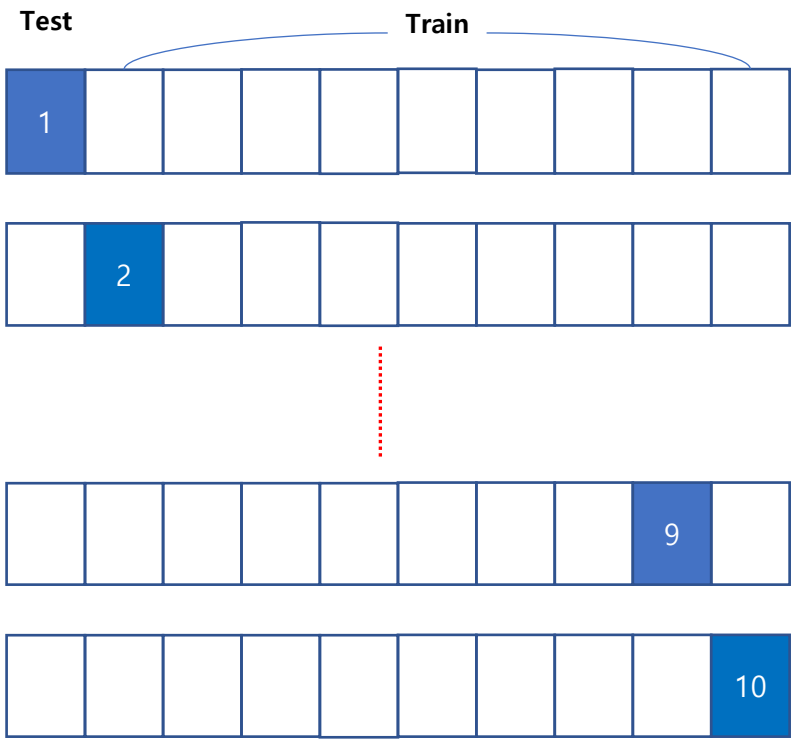
- STEP1: 데이터 프레임 준비 및 전처리
  - cancer.csv 파일을 불러와서 데이터프레임(cancer) 구성
  - str과 summary를 활용한 데이터 검토
  - class 변수의 척도를 범주형으로 변경
    - 측정값이 수치이고 계량형 척도로 되어 있으면 KNN이 아니라 logit이 실행될 수 있음
  - 불필요한 변수인 id 제거
  - IV에 대한 표준화 및 새로운 데이터 프레임(cancer_z) 구성

암진단 사례

- STEP2: train 데이터셋과 test 데이터셋 구성
  - ◆ cancer_z를 7:3의 비율로 두 개 데이터 프레임으로 구분
    - 전자(70% 비율)는 train 데이터 프레임(cancer_train)으로 구성
    - 후자(30% 비율)는 test 데이터 프레임(cancer_test)으로 구성
  - ◆ sample 함수 사용한 객체 형성 및 객체를 활용한 데이터셋 구분
    - sample 함수를 쓸 때 결과가 달라지지 않도록 set.seed() 함수를 먼저 사용해야 함
    - 기본 형식: ind <- sample(추출 대상, 추출 개수, replace = F/T, prob = c(p_1, p_2, p_3,…))
      - ❖ 추출대상으로부터 추출 개수만큼 반복(비반복)해서 추출하되 p_i의 비율로 추출
  - ◆ diagnosis 측정값 비율 비교
    - 세 가지 데이터셋(cancer_z, cancer_train, cancer_test)에서 악성(M) 측정값 비율이 비슷해야 함

암진단 사례

- STEP3: train 데이터 프레임을 이용한 훈련
  - ◆ expand.grid 함수를 통한 k값 범위(3~10) 설정
  - ◆ caret패키지에 있는 trainControl 함수로 학습방법 채택
    - repeated *K*-fold cross validation 시행(대개 $K = 10$)
  - ◆ caret 패키지에 있는 train 함수를 이용한 최적의 K값 선정
    - set.seed() 함수로 고정시킨 후 실행
    - 기본 형식: train(변수설정, 데이터, ML방법, 학습방법, k범위)
    - Accuracy(0~1)와 Kappa(-1~1) 평균값이 제일 클 때 최적의 K값
  - ◆ varImp() 함수를 이용하여 IV의 중요도 확인



암진단 사례

- 성능 측정 지표
  - Accuracy (= $\frac{TP+TN}{(TP+FP+TN+FN)}$)
    - 실제 분류 결과와 예측 분류 결과의 정확도(일치도)
    - 1에 가까울수록 우수함

예측

실제	Positive(+)	Negative(-)
Positive(+)	TP	FN
Negative(-)	FP	TN

  - Kappa (= $\frac{Accuracy\ -\ p_e}{1\ -p_e}$, p_e: 기대 정확도)
    - 기대 정확도: P로 예측했는데 실제 P인 확률 혹은 N으로 예측했는데 실제 N인 확률
      - $p_e = \frac{TP+FP}{total} \times \frac{TP+FN}{total} + \frac{FN+TN}{total} \times \frac{FP+TN}{total}$
    - Kappa는 Accuracy와 상관관계가 없고, -1에서 1의 값을 갖는데 1에 가까울수록 우수함
      - $0.6 < Kappa \leq 0.8 \rightarrow$ Good
      - $0.8 < Kappa \leq 1 \rightarrow$ Very Good

암진단 사례

- Kappa의 중요성
  - ◆ Case 1 vs. Case 2: accuracy는 같지만 kappa는 큰 차이
    - Case 2처럼 Pr(P)가 높은 경우, 모든 예측을 P로 하면 accuracy를 높일 수 있음
    - 이러한 극단적 예측방법보다 실제 모형이 우수한지 판단하려면 kappa값을 함께 평가해야 함
  - ◆ Case 2 vs. Case 3: accuracy는 비슷하지만 kappa가 증가함
    - Case 2와 Case 3은 Pr(P)가 동일하게 높음
    - 이럴 경우 빈도가 낮은(N) 결과를 정확히 예측하는 것(kappa값 개선)이 중요함

Case 1

실제 \ 예측	P	N
P	89	8
N	6	97

accuracy	0.930
P(e)	0.501
kappa	0.860

Case 2

실제 \ 예측	P	N
P	186	0
N	14	0

accuracy	0.930
P(e)	0.930
kappa	0.000

Case 3

실제 \ 예측	P	N
P	184	2
N	2	12

accuracy	0.980
P(e)	0.870
kappa	0.846

암진단 사례

		예측: Positive(+)	예측: Negative(-)
실제	Positive(+)	TP	FN
제	Negative(-)	FP	TN

- 기타 성능 측정 지표
  - ◆ Precision(정밀도 = $\frac{TP}{(TP+FP)}$): Positive로 예측했는데, 실제 Positive인 비율
    - caret 패키지에서 Pos Pred Value
    - 암진단에서 오진을 막기 위한 중요한 지표
  - ◆ Sensitivity(민감도 = $\frac{TP}{(TP+FN)}$): 실제 Positive 중에서 Positive로 예측된 비율
    - Recall(재현율)이라고도 하며, 전염병(코로나19) 검사에서 확진자 선별을 위해 중요한 지표
  - ◆ Specificity(특이도 = $\frac{TN}{(FP+TN)}$): 실제 Negative 중에서 Negative로 예측된 비율

암진단 사례

- Precision and Sensitivity
  - ◆ Precision과 Sensitivity는 한계가 있으므로 반드시 Accuracy나 Kappa를 함께 고려해야 함
  - ◆ 이상적으로는 네 가지 지표가 모두 높은 모형이 적합하나 특성을 고려해야 함

실제

예측	Positive(+)	Negative(-)
Positive(+)	10	0
Negative(-)	80	10

Precision = 1 그러나 Accuracy = 0.2

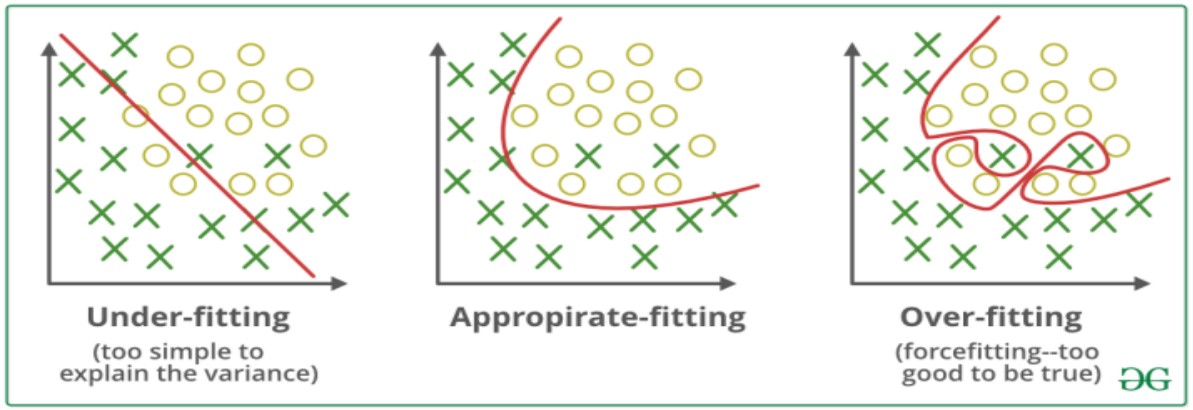
실제

예측	Positive(+)	Negative(-)
Positive(+)	10	80
Negative(-)	0	10

Sensitivity = 1 그러나 Accuracy = 0.2

암진단 사례

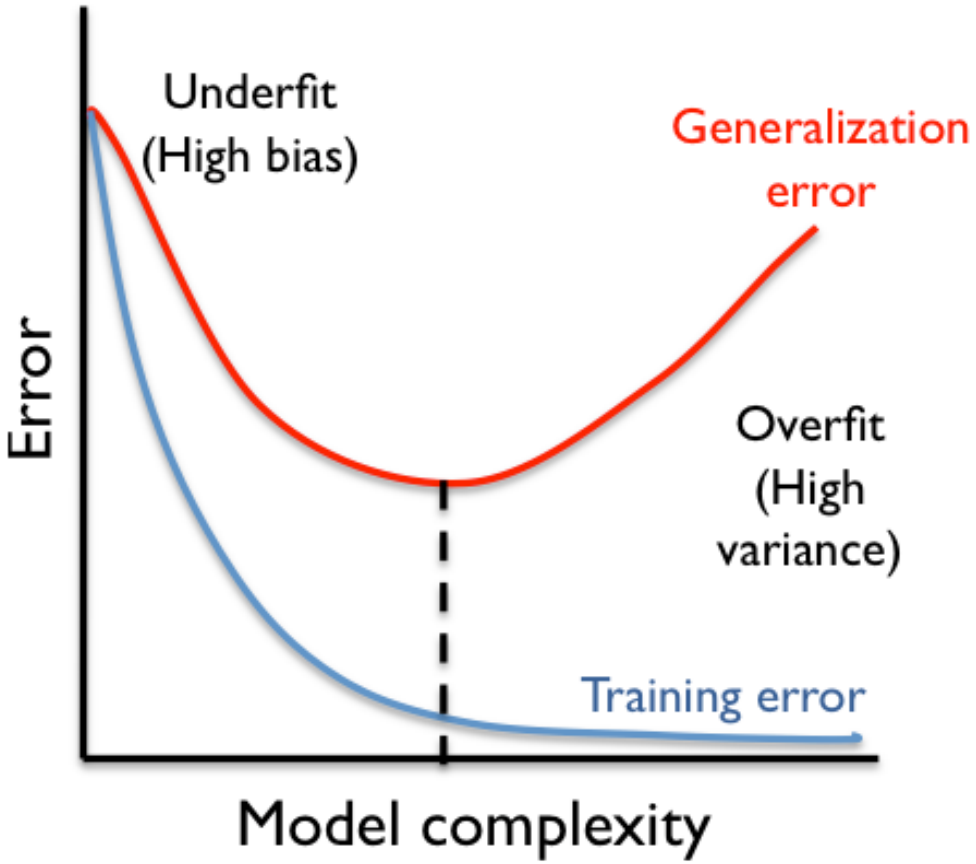
- STEP4: test 데이터 프레임을 이용한 성능 평가
  - ◆ 내장함수인 predict() 함수 사용하여 KNN 모델을 test 데이터 프레임에 적용
  - ◆ confusionMatrix() 함수를 사용하여 Accuracy & Kappa 도출
    - ▪ STEP3와 STEP4의 Accuracy & Kappa 비교
      - ❖ train 데이터 기반 Accuracy와 Kappa가 크면, 과대적합(overfitting) 문제를 의심할 수 있음
      - ❖ 과대적합 문제가 발생하면 STEP5에서 성능을 개선해야 함



출처: https://towardsdatascience.com/

암진단 사례

- 편향(bias)과 편차(variance)
  - ◆ 편향: train 데이터에서 모델 오류(=1-accuracy)
    - ▪ 모형이 단순해서 편향이 크면 과소적합 문제 발생
  - ◆ 편차: train 데이터와 test 데이터 간의 모델 오류 차이(혹은 정확도 차이)
    - ▪ 모형이 복잡하면 편향은 줄지만, 편차가 커져서 과대적합 문제 발생



출처: https://stickie.tistory.com/47

암진단 사례

- STEP5: 성능 개선
  - ◆ kknn 패키지의 train.kknn() 함수 사용
    - 가까운 사례에 더 많은 가중치를 적용한 KNN 방법
    - kmax를 지정하면 최적의 k 선정
    - 거리: 1(Manhattan distance) & 2(Euclidean distance)
    - kernel = c(가중치를 부여하는 다양한 방법 지정)
  - ◆ Accuracy와 Kappa 개선 확인

[다양한 성능개선 방법]

방법	효과	비용	난이도
데이터 추가 - 행	2	2	1
데이터 추가 - 변수 수집	3	3	3
데이터 추가 - 파생 변수 만들기	2	1	1
하이퍼파라미터 변경	1	1	1
클래스 균형 맞추기	1	1	1
방법론 변경	3	1	2

- STEP6: 예측
  - ◆ 양성/악성에 대한 판단이 없는 새로운 열 개 사례 진단 예측