104: Modeling

ML workflow

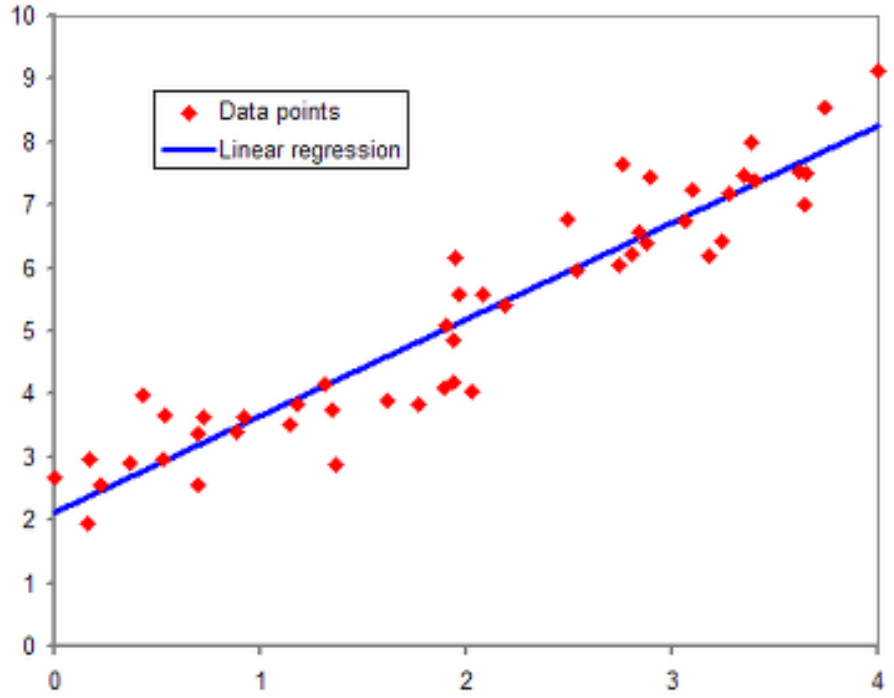
titanic_submission.csv (2.84 kB)

PassengerId	Survived
892	0
893	0
894	0
895	0
896	0

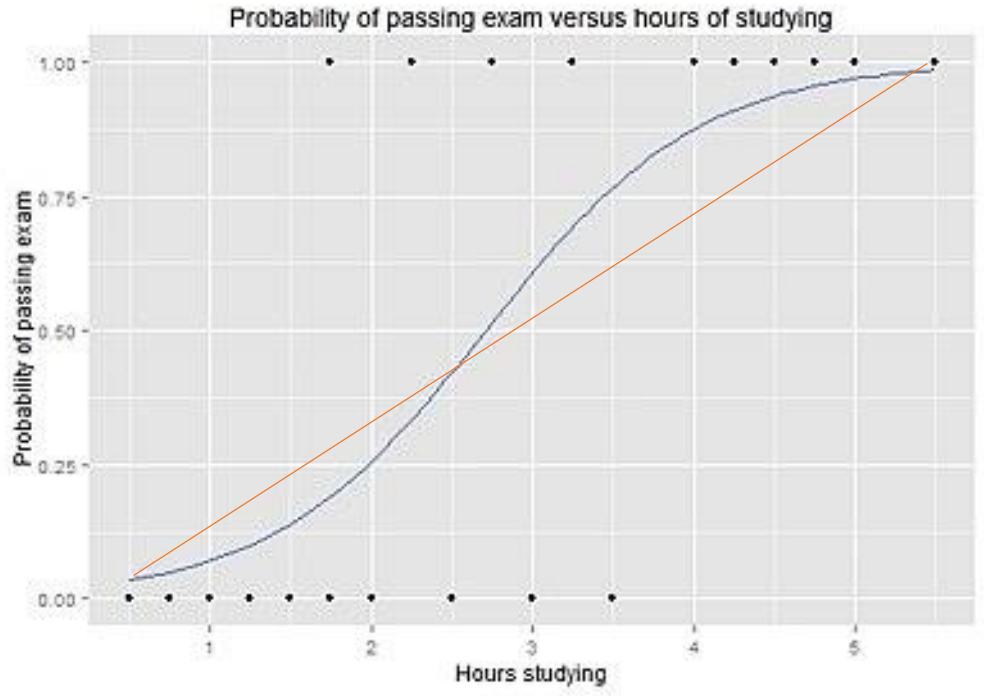
정답이 있으므로 **지도학습**

분류 결과는
생존과 사망 두가지: **이진분류**

예측과 분류



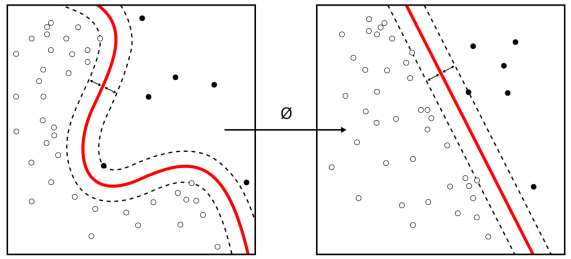
Linear Regression



Logistic Regression

분류 모델

다차원, 초평면

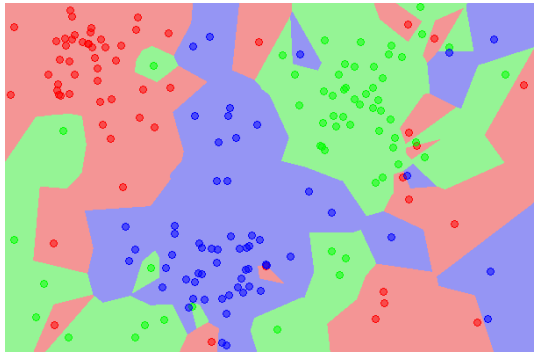


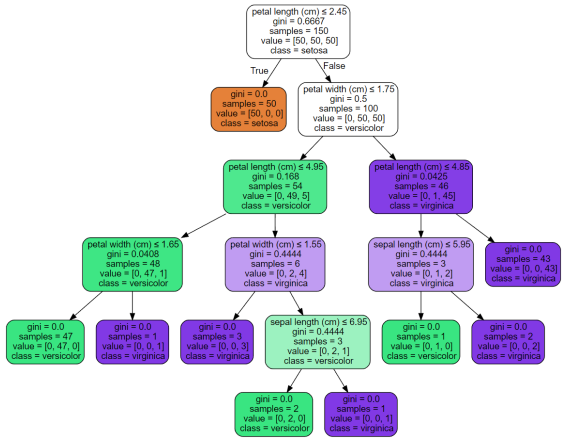
Logistic Regression
선형분리불가능 해결

Support Vector Machine

거리함수

k-Nearest Neighbor





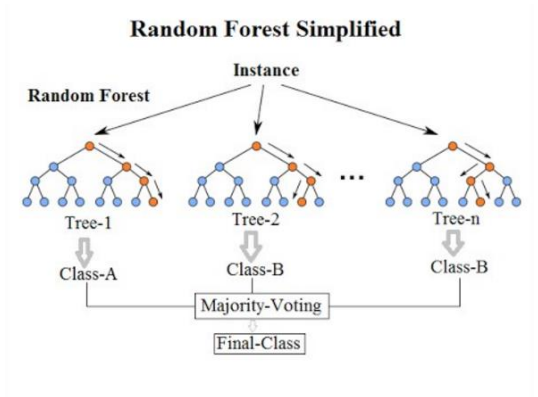
Decision Tree

불순도, 정보이득

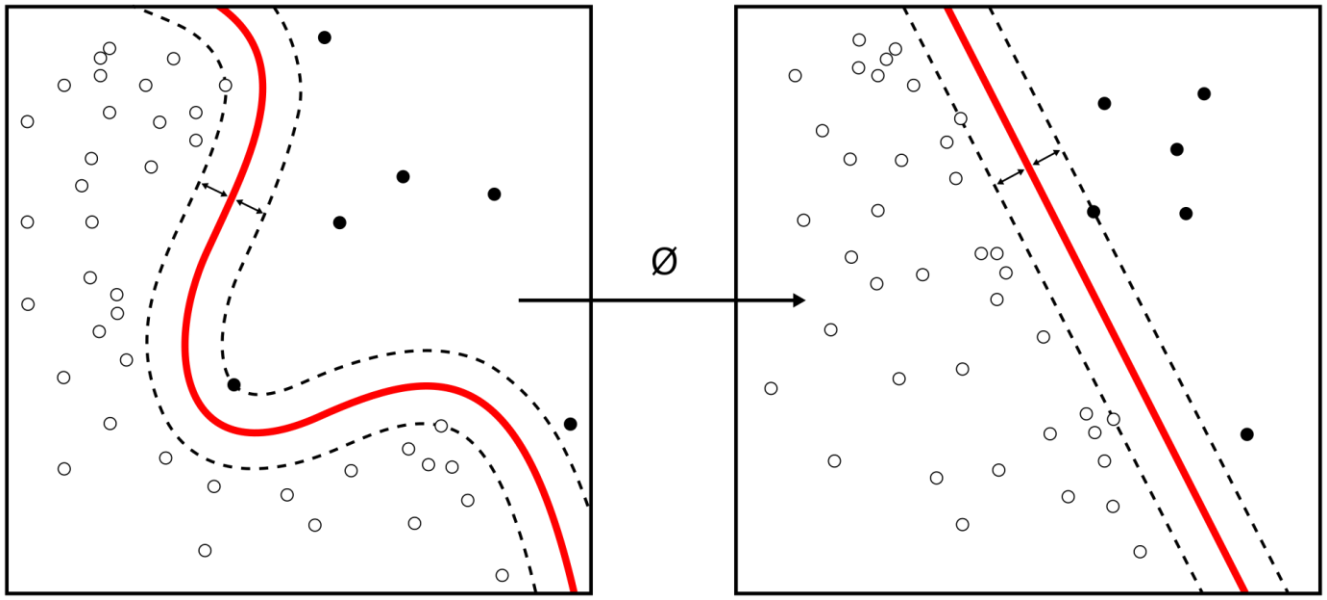
overfitting ↓

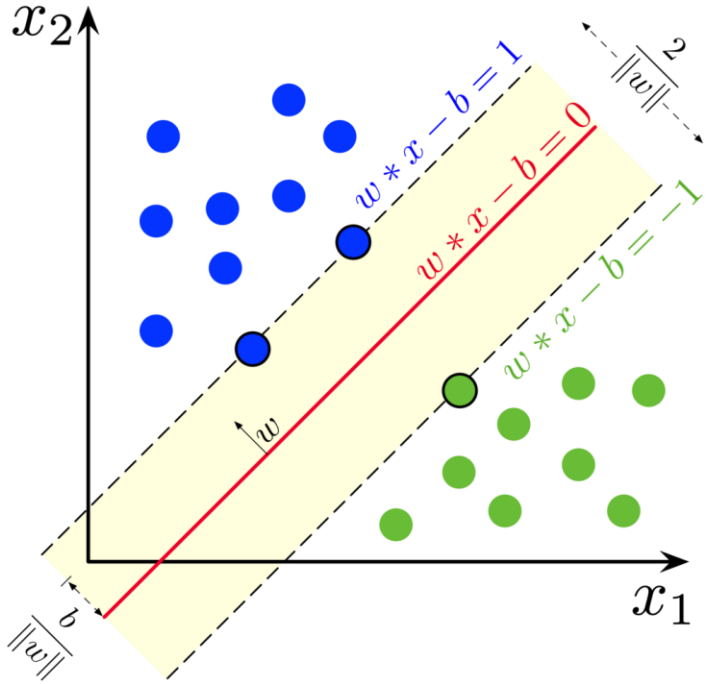
Random Forest

ensemble, bagging

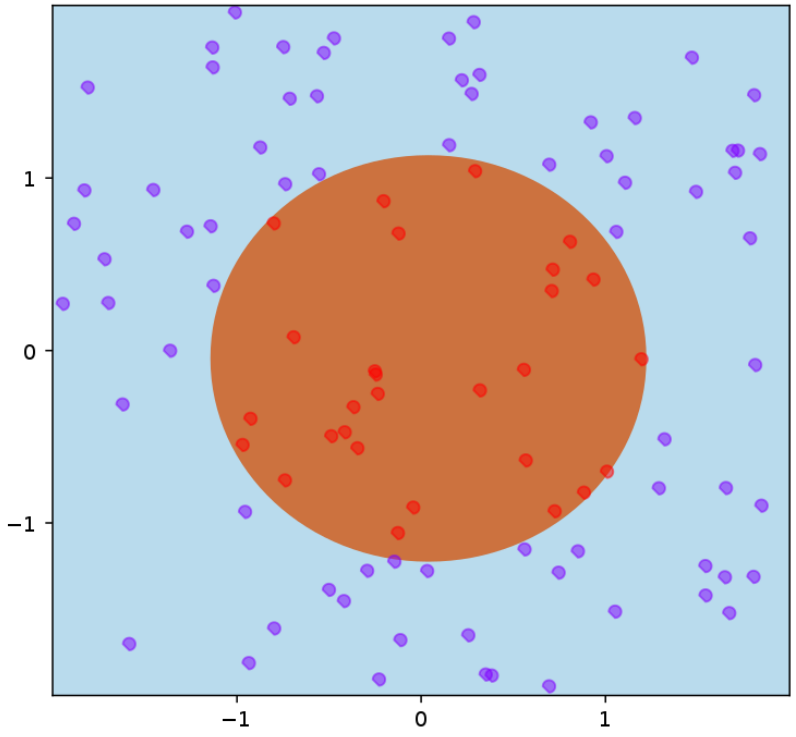


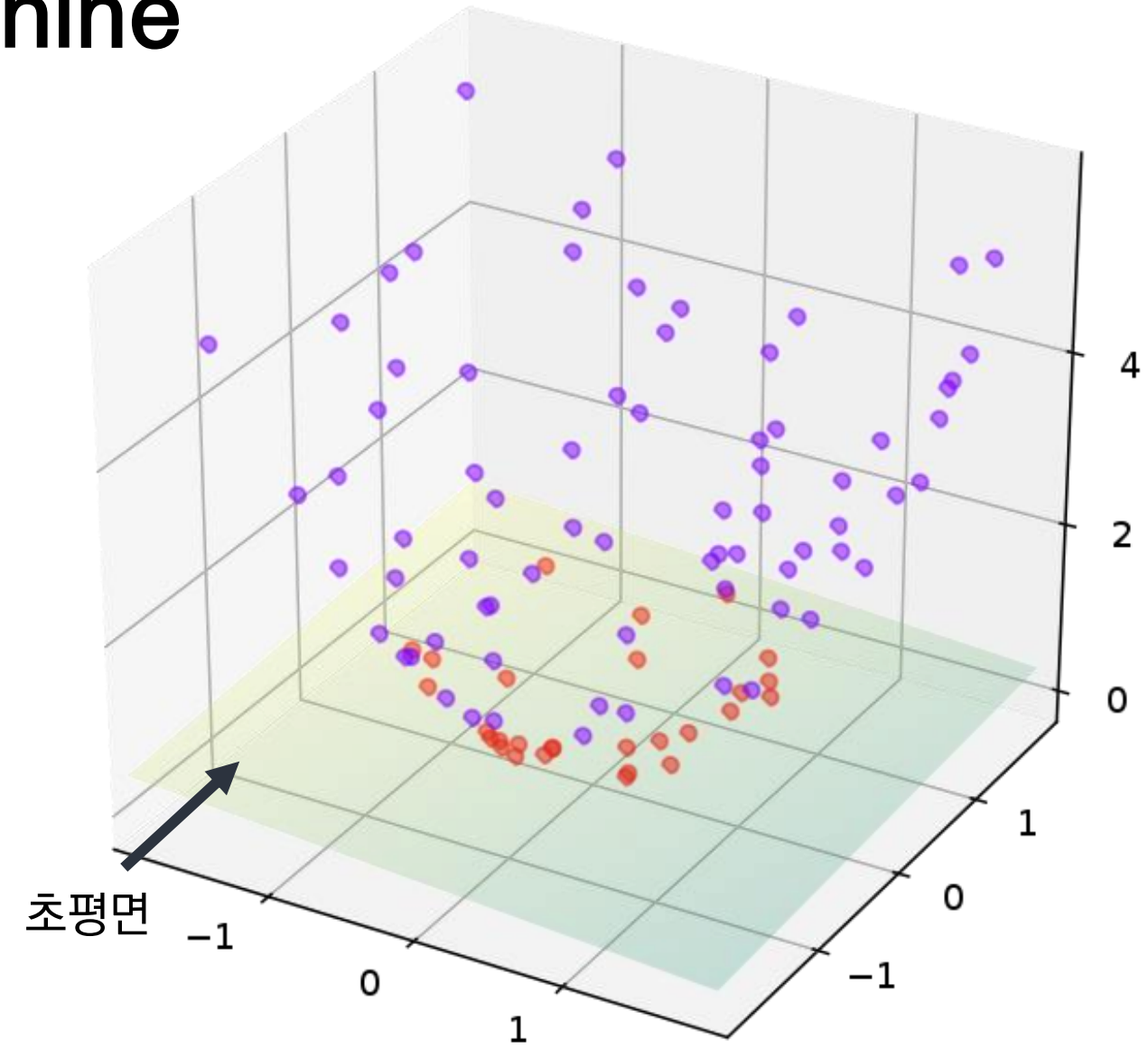
Support Vector Machine



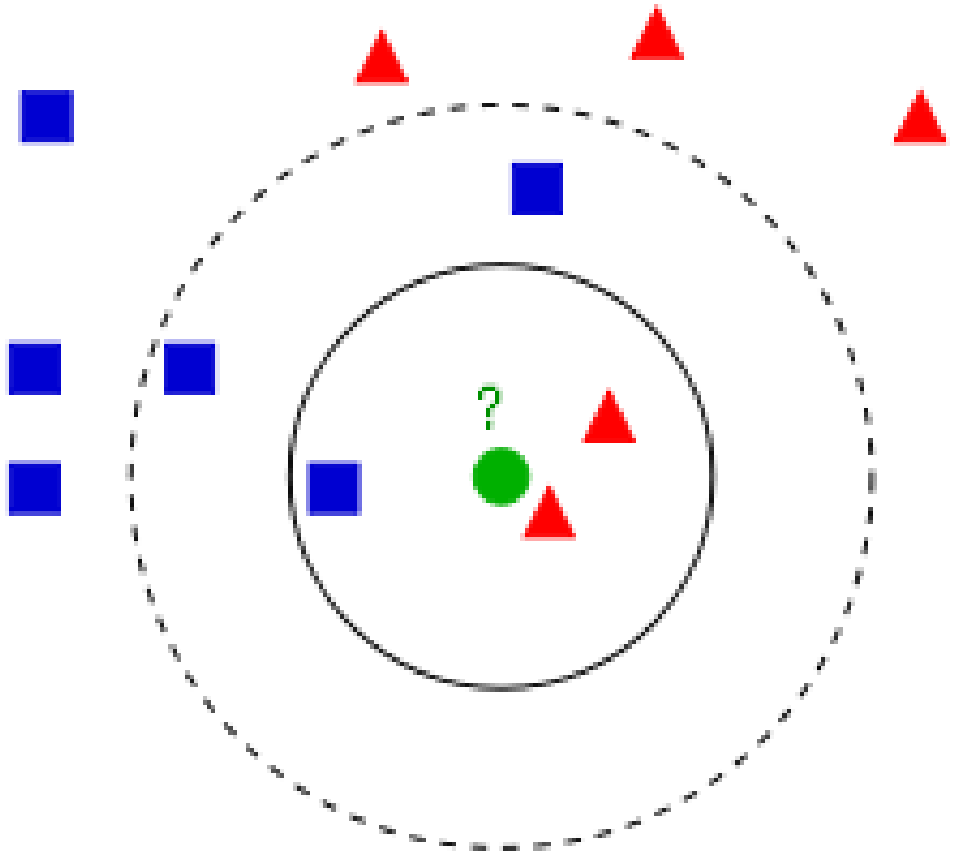


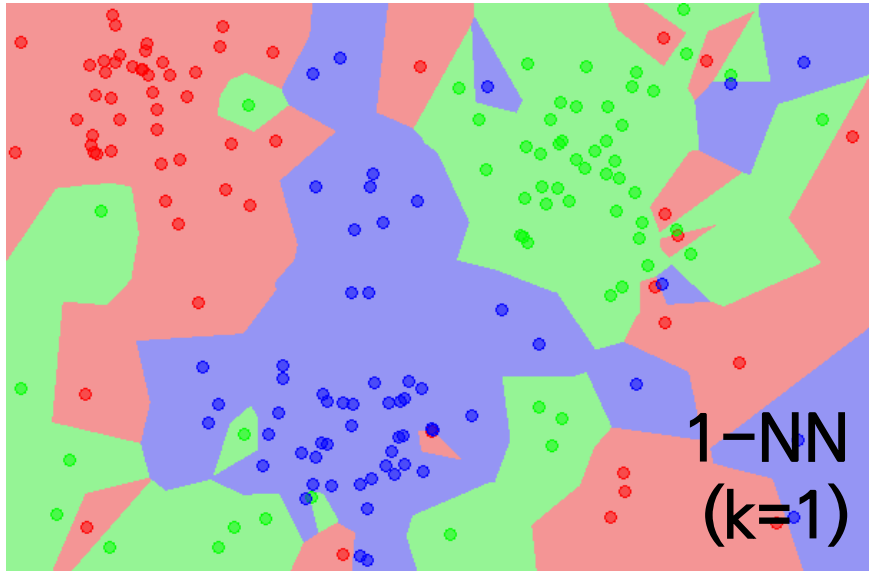
Support Vector Machine

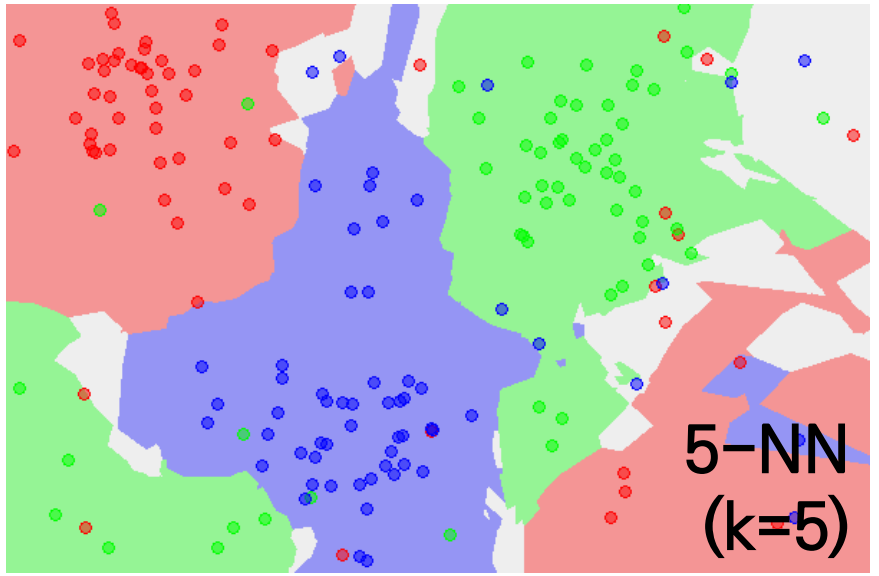




KNN







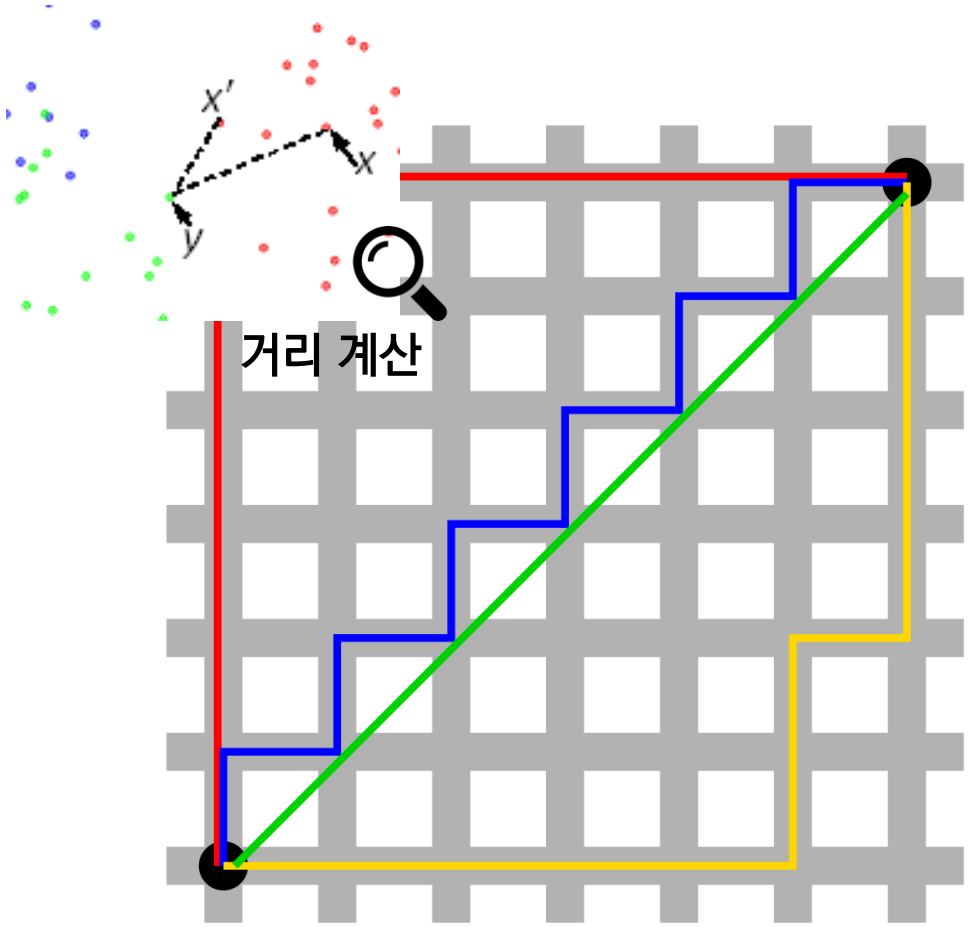
거리 함수

Euclidean distance (L2)

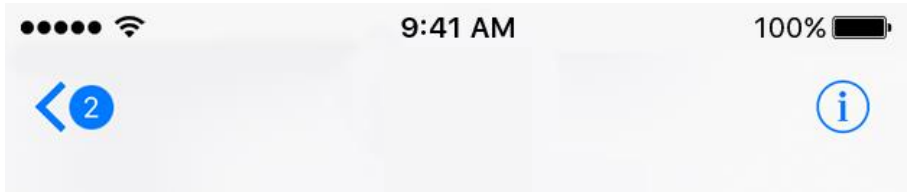
$$d(A,B) = \sqrt{(x2-x1)^2 + (y2-y1)^2}$$

Manhattan distance (L1)

$$d(A,B) = |x1-x2| + |y1-y2|$$



Decision Tree



영화 OOO 봤어?

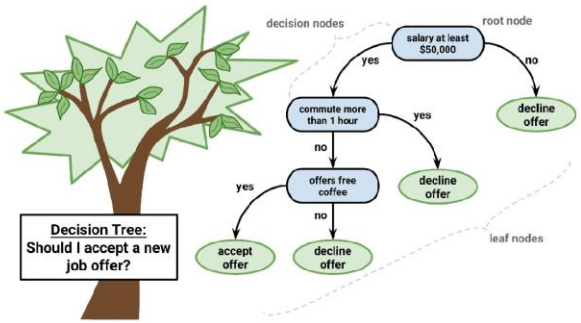
네가 좋아할 거 같더라

정말?

왜?

음……

대부분의 모델은 작동 과정을 알 수 없는 블랙박스



결정트리는 설명이 가능한 스무고개

Decision Tree

불순도 impurity

지니 Gini 불순도	I_G	• 한 노드의 모든 데이터가 같은 클래스라면 0 클래스가 균등하게 분포되어 있다면 최대 1
엔트로피 entropy	I_H	• *클래스가 섞인 정도가 클수록 1에 가깝다* • 트리의 **상호의존정보** 최대화
분류오차 classification error	I_E	• *클래스가 섞인 정도가 클수록 1에 가깝다* • 노드의 클래스 **확률 변화에 둔감** (권장 X)

정보이득 IG

$$IG(D_p, f) = I(D_p) - \sum_{j=1}^{m} \frac{N_j}{N_p} I(D_j)$$

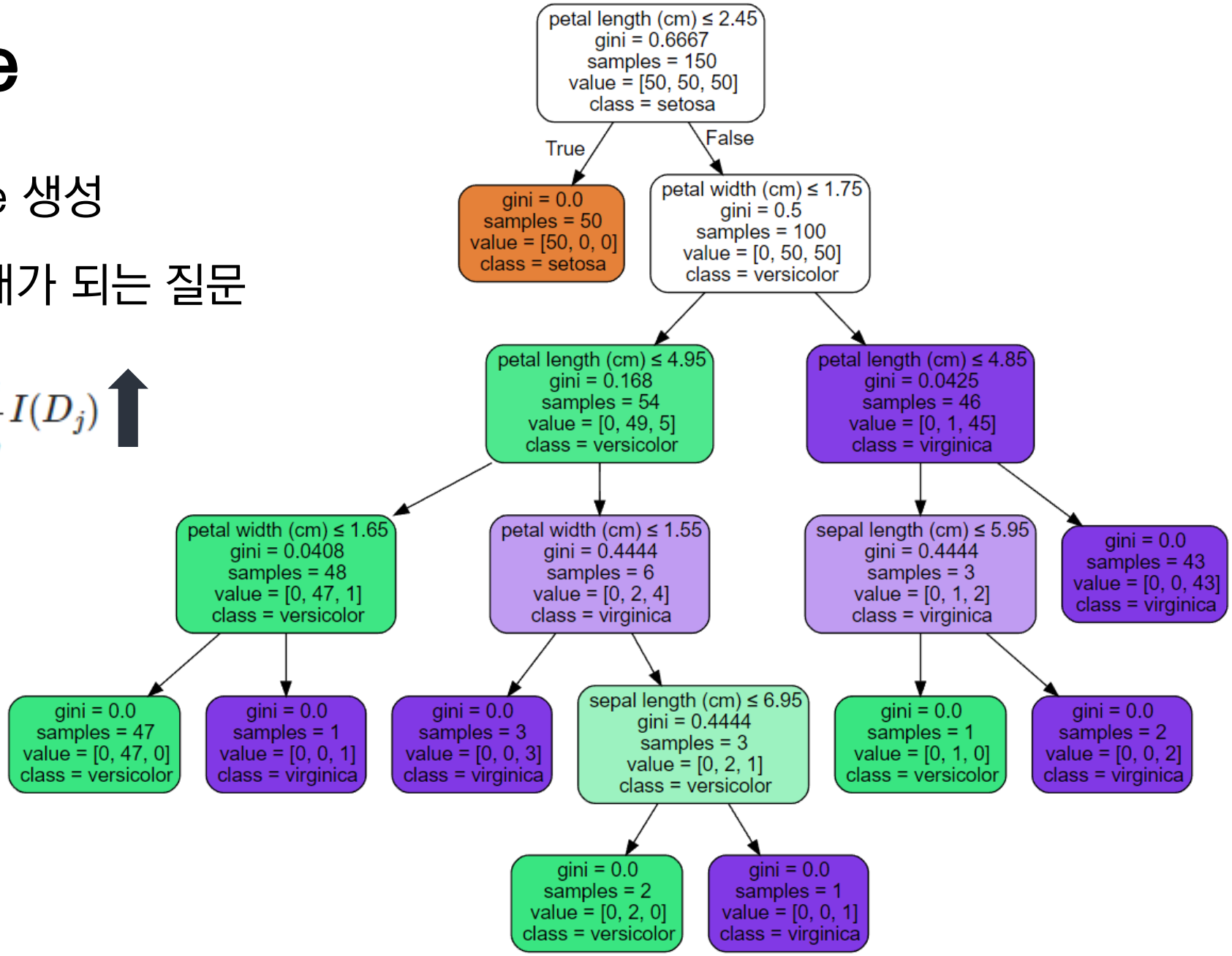
f: 분할 기법
I : 불순도 지표
D : 데이터셋
N : 노드 데이터 개수

Decision Tree

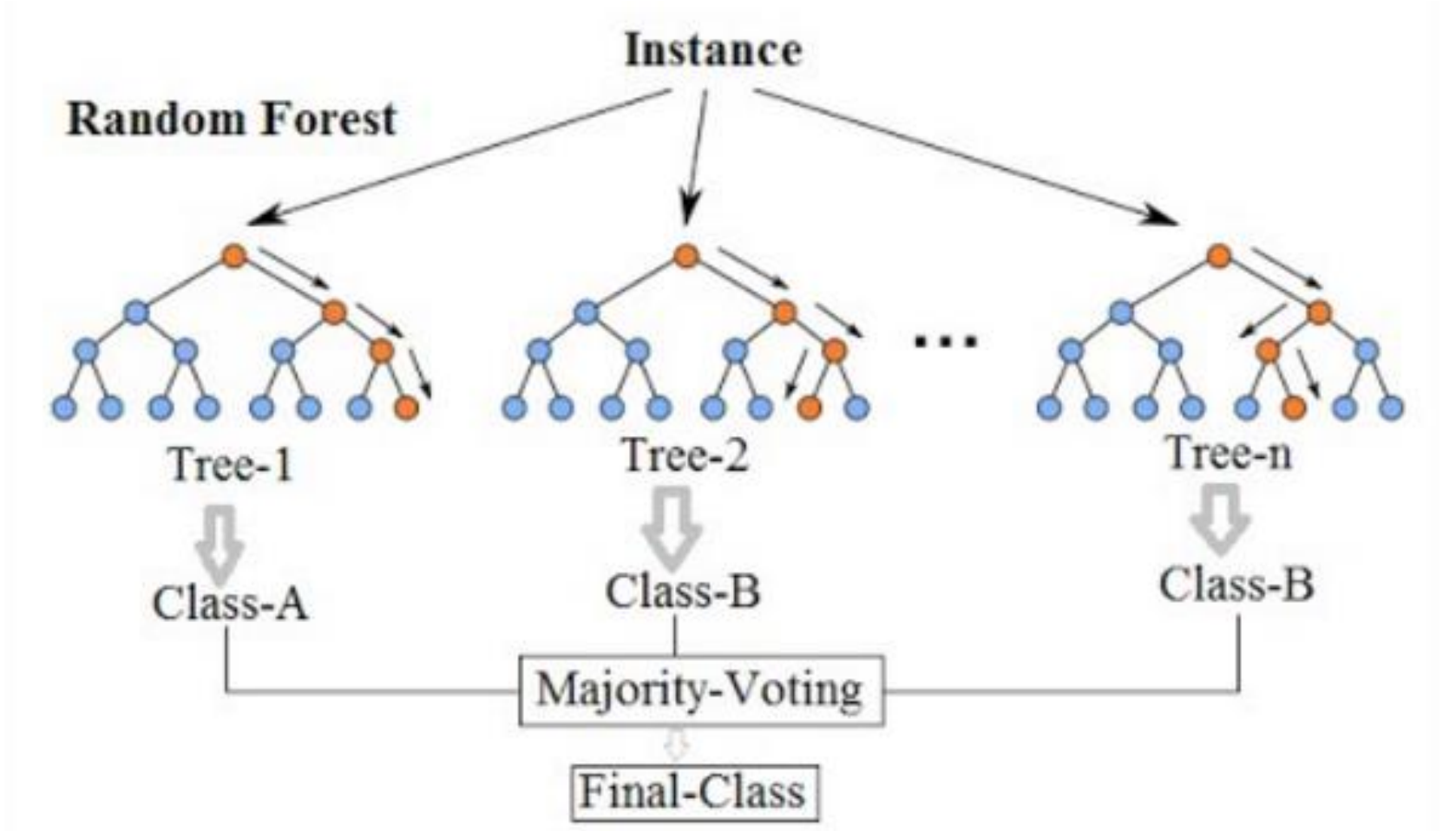
목표: 순수한 Leaf Node 생성

방법: 정보이득 IG가 최대가 되는 질문

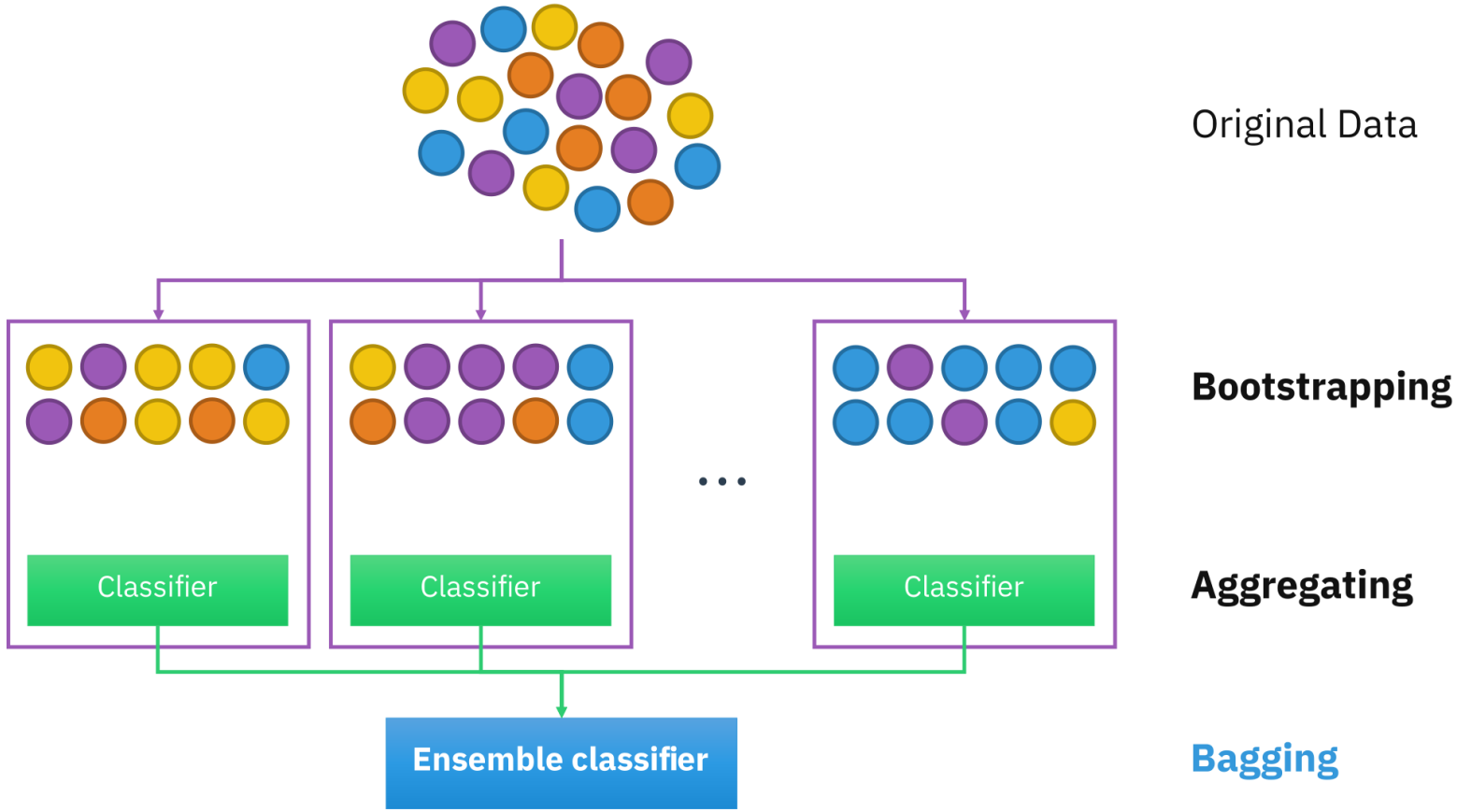
$$IG(D_p, f) = I(D_p) - \sum_{j=1}^{m} \frac{N_j}{N_p} I(D_j)$$



Random Forest



Random Forest



Original Data

Bootstrapping

Aggregating

Bagging

혼동행렬

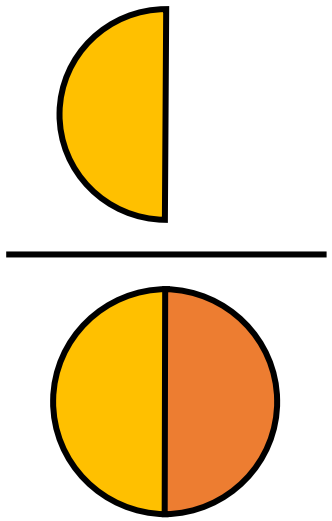
		실제	
		Positive	Negative
예측	Positive	True Positive	False Positive
	Negative	True Negative	False Negative

평가지표

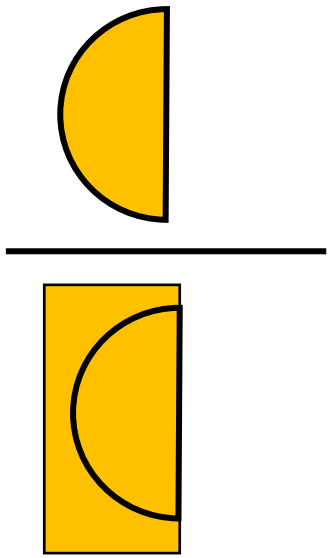
$$\frac{TruePositives + TrueNegatives}{TruePositives + TrueNegatives + FalsePositives + FalseNegatives}$$

정확도 Accuracy

평가지표



Recall
재현율



Precision
정밀도