N / A / N / O / D / E / G / R / E / E

지도학습 알고리즘

12. Decision Tree

Decision tree

- 특징
  - Feature 공간을 여러 개의 단순한 영역으로 분할
  - Feature 공간을 분할하는 일련의 규칙들을 tree의 형태로 표현
- 장점
  - 데이터나 모델에 대한 전제, 가정이 없음
  - 간단하기 때문에 이해나 해석이 용이함
  - Classification과 regression에 모두 사용이 가능
  - Numeric feature와 categorical feature 모두 처리가 가능
  - 스케일링이나 중앙화 같은 데이터에 대한 전처리가 거의 필요 없음
  - 탐색적 데이터 분석(exploratory data analysis, EDA)에서 유용

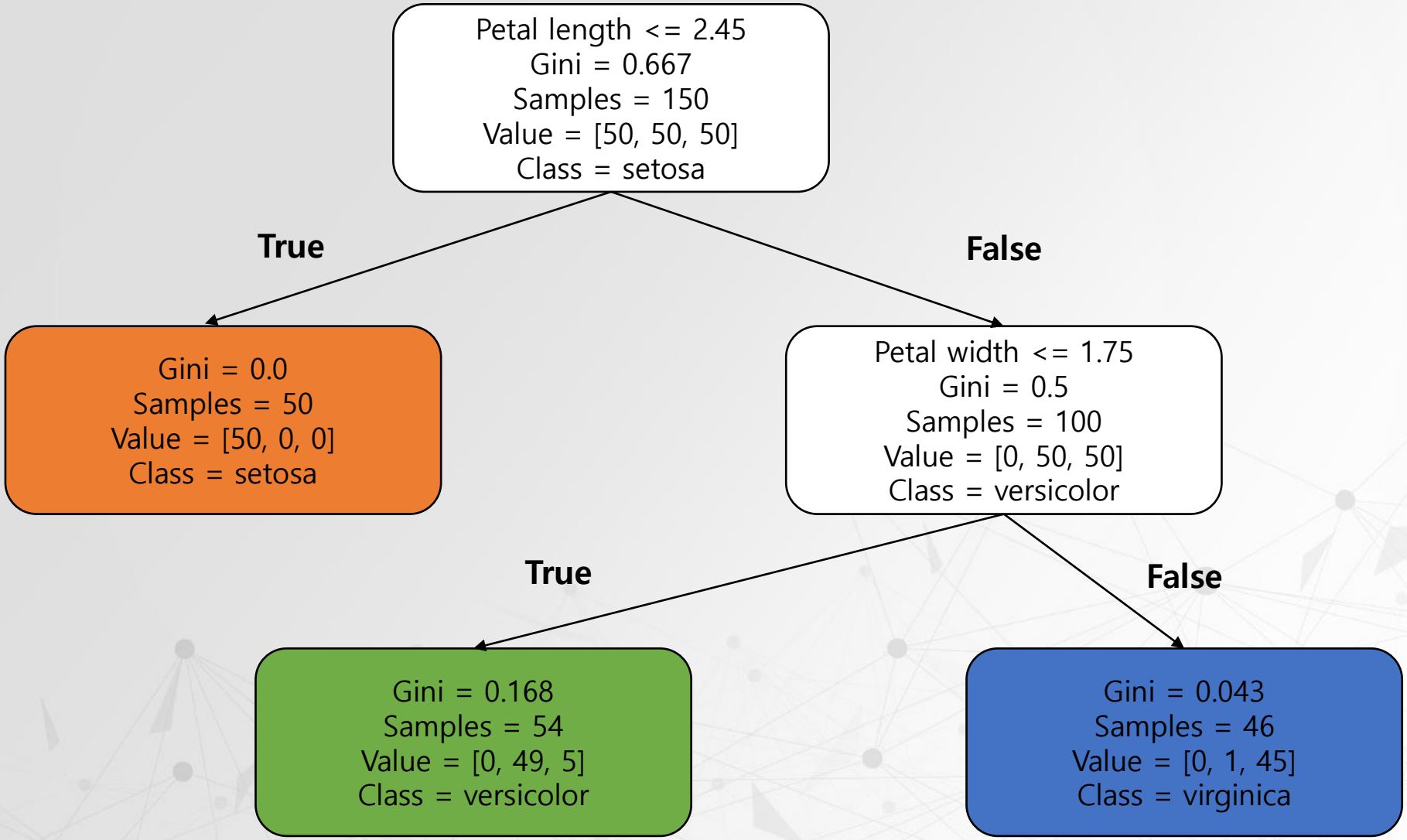
Decision tree

- 단점
  - 데이터의 회전이나 작은 변화에 매우 민감
  - Tree growing에 제약을 주지 않으면 쉽게 과적합
  - 예측 정확도 측면에서 다른 머신러닝 알고리즘보다 떨어지는 경우가 많음
    = bagging, boosting, random forest 등을 사용해 예측 정확도를 향상

Training and visualizing a decision tree: iris dataset

- Iris dataset
  - 3개의 클래스: iris setosa, iris versicolor, iris virginica
  - 각 클래스별로 50개의 샘플
  - 4개의 feature
- Terminology
  - Row: observation (sample, example, instance, record)
  - Column except the last: feature
  - Last column: response (target, outcome, label)

Decision tree for iris dataset



Decision tree for iris dataset

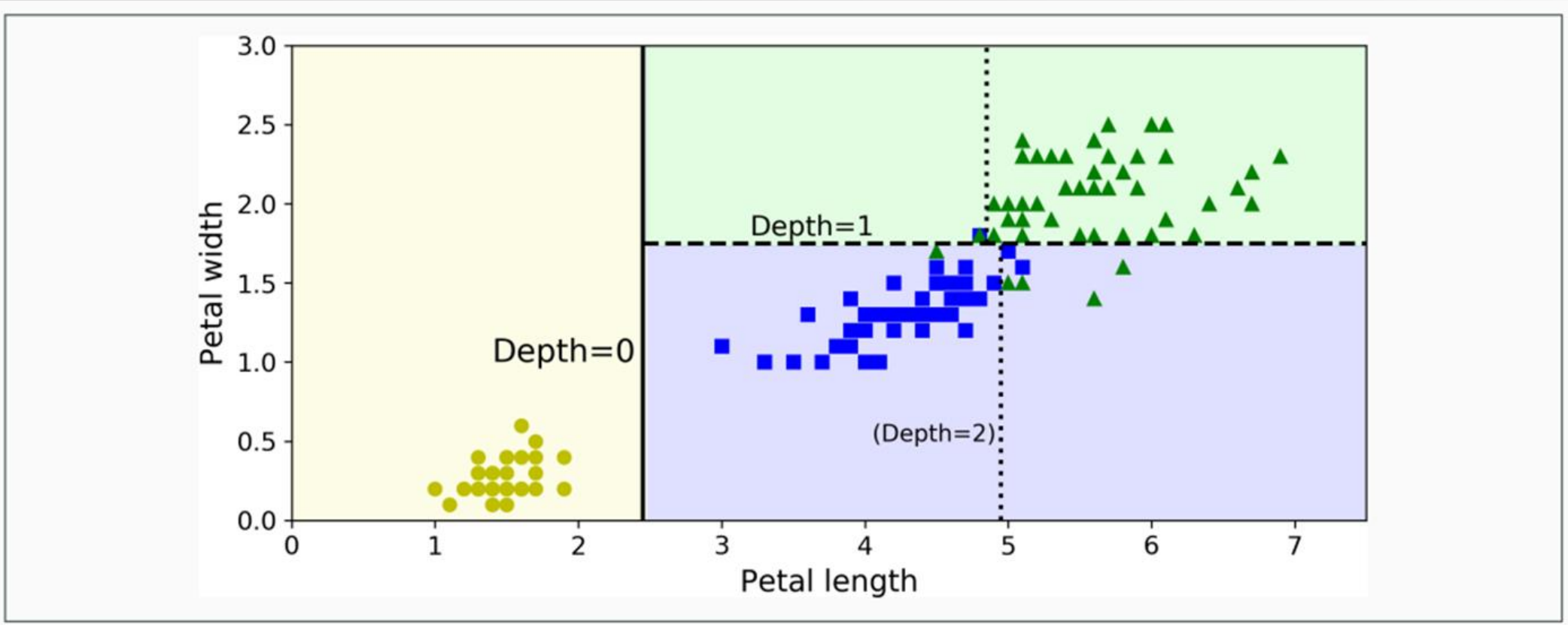


Figure 6-2. Decision Tree decision boundaries

[출처] A. Geron, Hands-On Machine Learning with Scikit-Learn and TensorFlow

Decision tree for regression

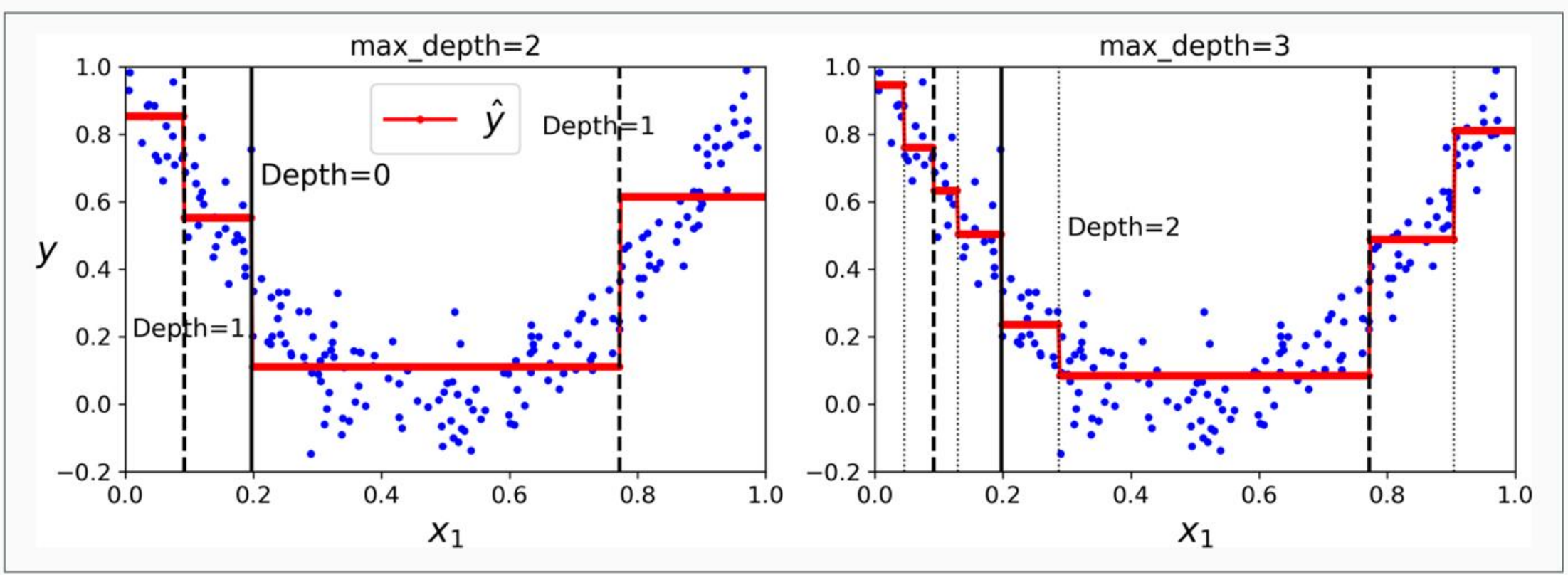


Figure 6-5. Predictions of two Decision Tree regression models

[출처] A. Geron, Hands-On Machine Learning with Scikit-Learn and TensorFlow

Gini impurity measure

Gini impurity measure

- 임의의 node i에서 impurity G_i는 다음과 같이 정의

$$G_i = 1 - \sum_k p_{i,k}^2$$

- $p_{i,k}$: node i에서 class k에 속하는 instance의 비율
- e.g.,

$$G = 1 - \left(\frac{49}{54}\right)^2 - \left(\frac{5}{54}\right)^2 \approx 0.168$$

Gini = 0.168
Samples = 54
Value = [0, 49, 5]
Class = versicolor

Information gain measure

- Information gain (entropy)
  - 임의의 node i에서 엔트로피

$$H_i = -\sum_k p_{i,k} \log p_{i,k}$$

  - $p_{i,k}$: node i에서 class k에 속하는 instance의 비율
  - e.g.,

$$H = -\frac{49}{54} \log \frac{49}{54} - \frac{5}{54} \log \frac{5}{54} \approx 0.31$$

Gini = 0.168
Samples = 54
Value = [0, 49, 5]
Class = versicolor

Regularization hyperparameters

- 과적합을 방지하기 위한 regularization
  - Tree growing에 있어 자유도에 제한을 가하는 것
  - Regularization 방법은 알고리즘에 따라 다름

- Regularization hyperparameters
  - max_depth: 트리의 최대 깊이
  - min_samples_split: 내부 node를 split하기 위한 최소의 샘플 수
  - max_leaf_nodes: leaf node 최대 수
  - max_feature: 각 node에서 split을 위해 계산하는 최대 feature의 수

Instability of decision tree

- Decision boundary가 항상 좌표축에 수직: 데이터셋의 회전에 민감

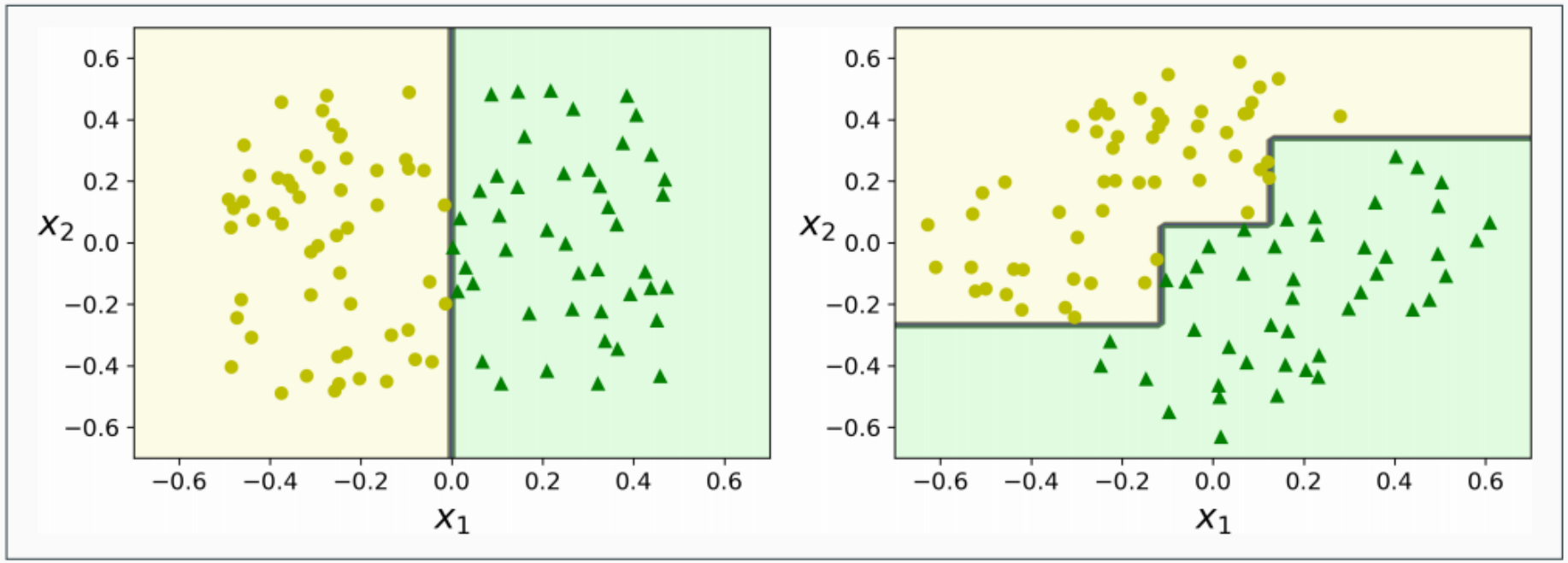
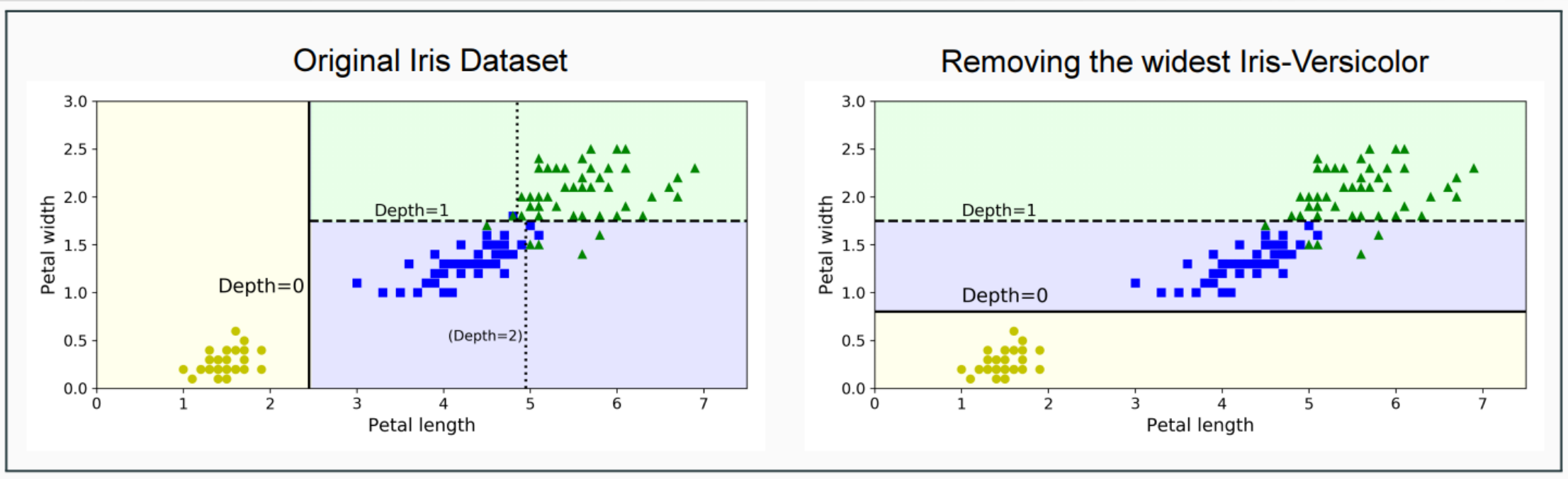


Figure 6-7. Sensitivity to training set rotation

[출처] A. Geron, Hands-On Machine Learning with Scikit-Learn and TensorFlow

Instability of decision tree

- 데이터의 작은 변화에도 매우 민감



[출처] A. Geron, Hands-On Machine Learning with Scikit-Learn and TensorFlow