CSFSM7053-00
Business Analytics

College of Hotel & Tourism
Kyung Hee University

Lecture Note 10

나무 기반 모형 – 의사결정나무 & 랜덤포레스트

Dohyung Bang

Fall, 2021

Syllabus

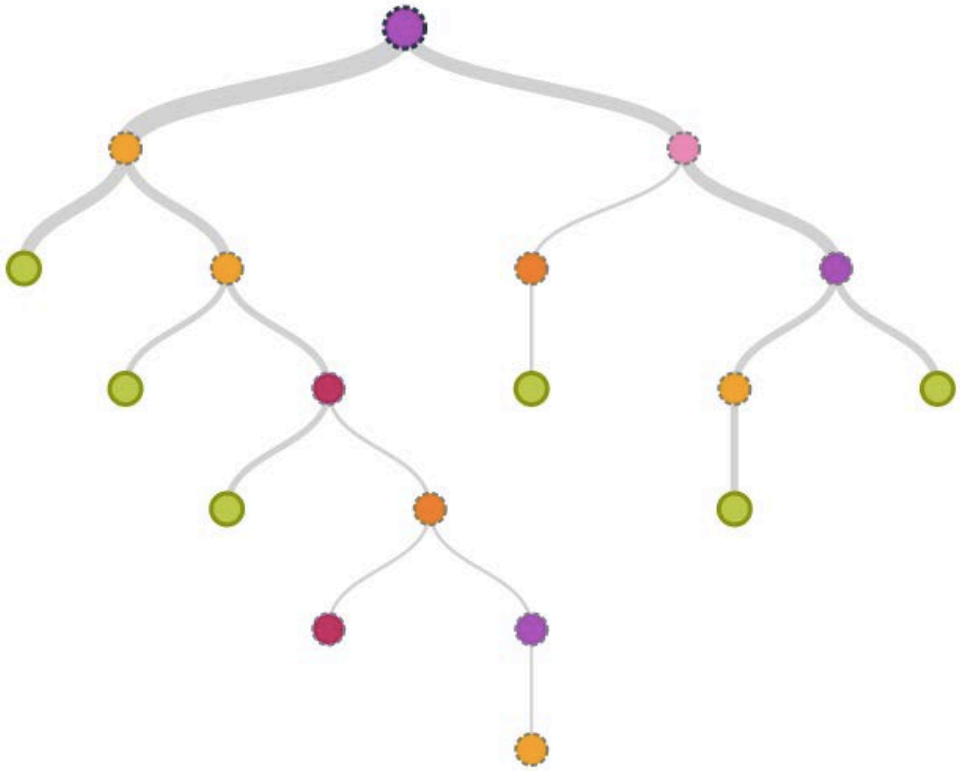
Week	Date	Topic	Note
1	9/6(월)	R Basic - R 기초 문법 학습	
2	9/13(월)	R Basic – Data Manipulation I	과제#1
3	9/20(월) (추석)	<추석> (보충영상) R Basic - Data Manipulation II	
4	9/27(월)	Descriptive Analytics I - 데이터 요약하기/상관관계/차이검증	~~과제#2~~
5	10/4(월) (대체공휴일)	<대체공휴일> (보충영상) Descriptive Analytics II - 데이터 시각화	과제#2
6	10/11(월) (대체공휴일)	<대체공휴일> (보충영상) Supplementary Topic I - 외부 데이터 수집 (정적 컨텐츠 수집)	~~과제#4~~ 과제#3
7	10/18(월)	Predictive Analytics I – Linear regression	
8	10/25(월)	Predictive Analytics II – Logistic Regression	시험 대체 수업
9	11/1(월)	Predictive Analytics III – Clustering & Latent Class Analysis	과제#4
10	**11/8(월)**	**Predictive Analytics IV – Tree-based Model and Bagging (Random Forest)**	
11	11/15(월)	Predictive Analytics V – Association Rules	
12	11/22(월)	Prescriptive Analytics I – Linear Programming	과제#5
13	11/29(월)	Prescriptive Analytics II – Data Envelopment Analysis (DEA)	
14	12/6(월)	Prescriptive Analytics III – Integer Programming	과제#6
15	12/13(월)	Prescriptive Analytics IV – Simulation	Quiz
16	**12/20(월)**	**Final Presentation**	

Lecture 9-1

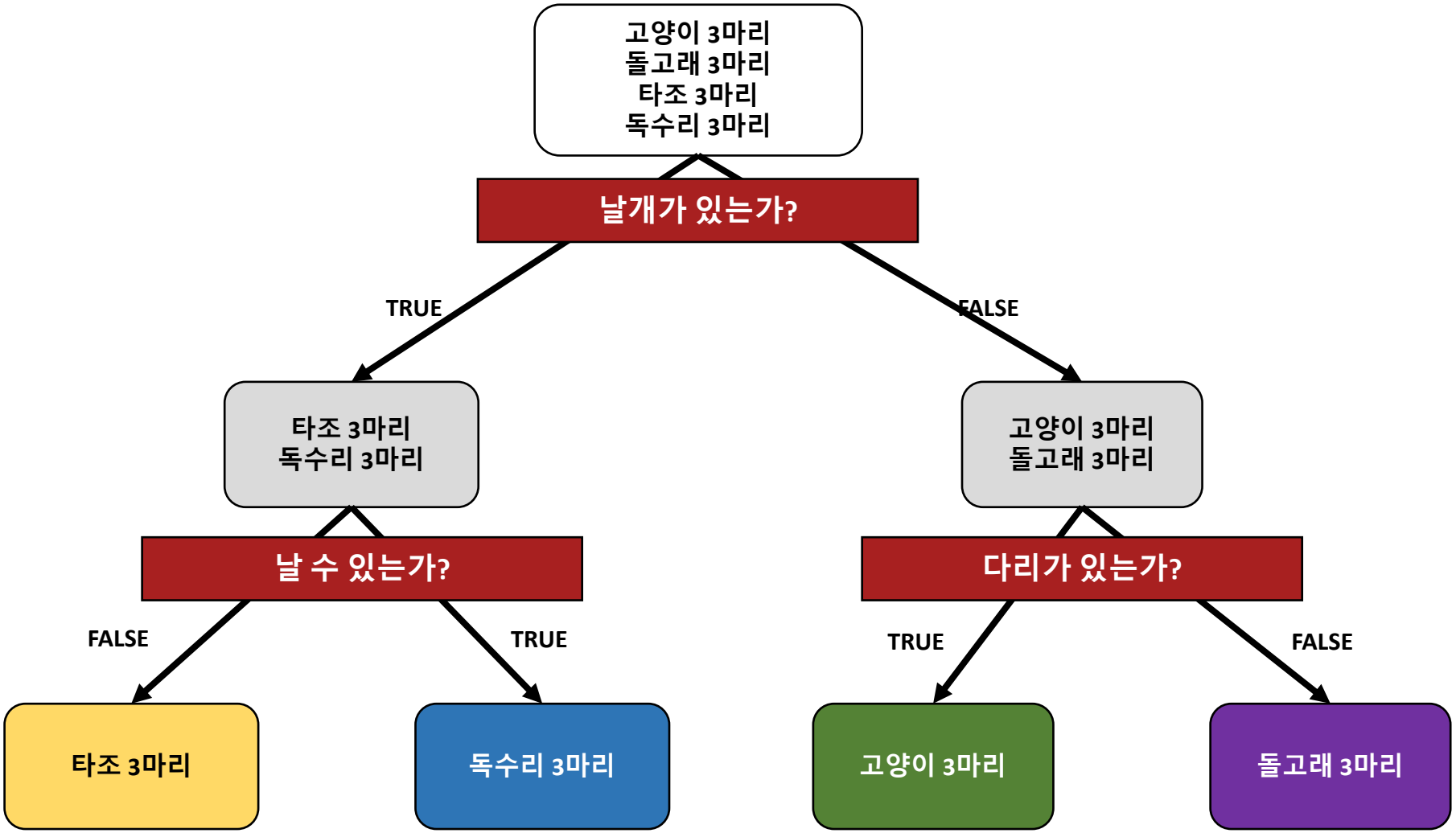
의사결정나무
(Decision Tree)

의사결정나무(Decision Tree) 란?

데이터들 사이에 존재하는 일련의 의사결정 규칙(Decision Rule)을 나무형태로 분류해 나가는 분석 기법으로, 그 결과를 나무형태의 그래프로 표현할 수 있다는 점에서 이름이 유래됨.



의사결정나무의 원리

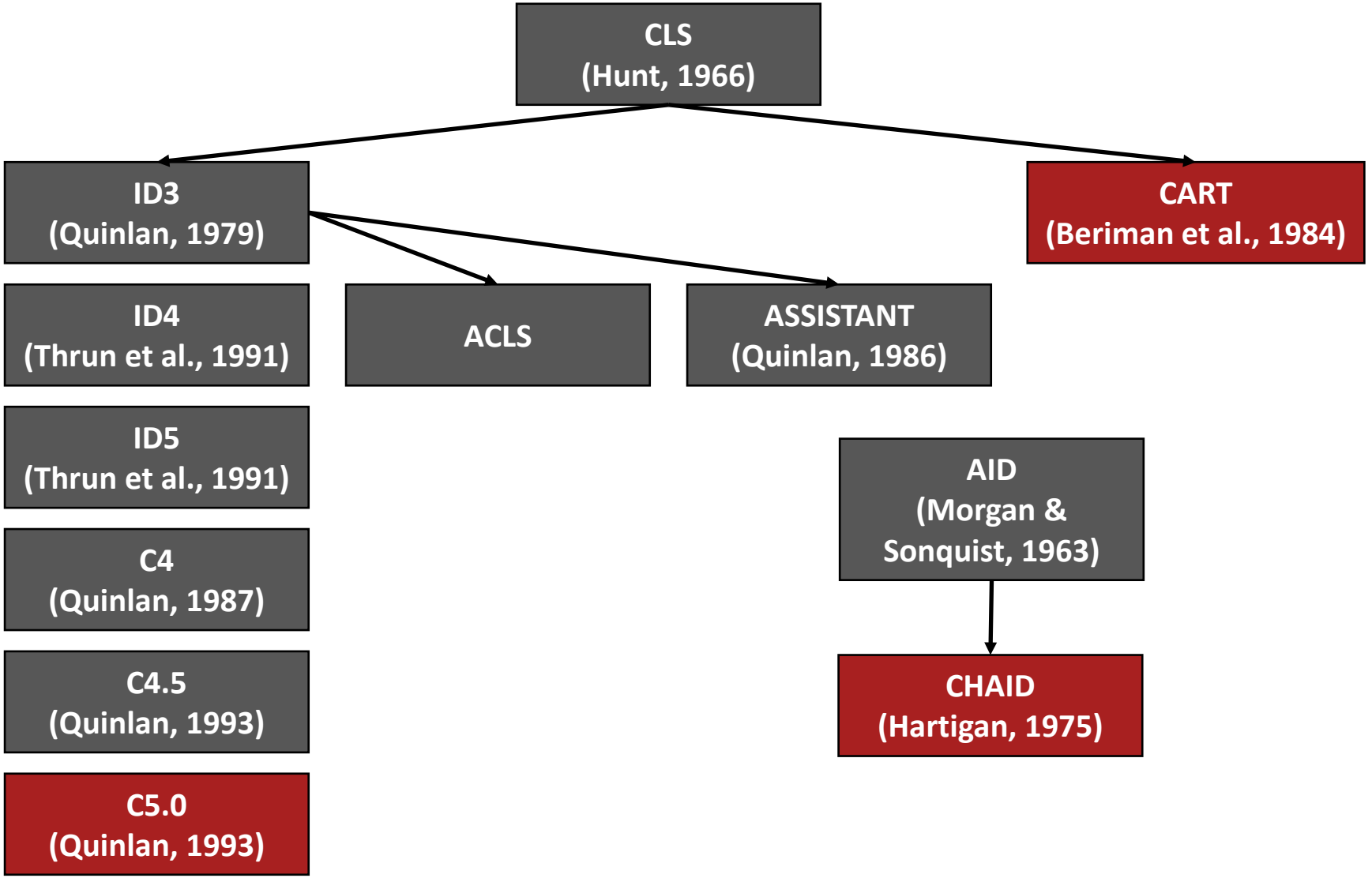

고양이 3마리
돌고래 3마리
타조 3마리
독수리 3마리
날개가 있는가?
TRUE
FALSE
타조 3마리
독수리 3마리
고양이 3마리
돌고래 3마리
날 수 있는가?
다리가 있는가?
FALSE
TRUE
TRUE
FALSE
타조 3마리
독수리 3마리
고양이 3마리
돌고래 3마리

의사결정나무의 형성과정

의사결정나무(Decision Tree)의 형성과정

성장(Growing); 분리기준과 정지규칙 결정	가지치기 (Pruning)	모형 평가	해석 및 예측
• 각 노드에서 최적의 분리규칙(Splitting Rule)을 찾아 나무를 성장(자식노드 혹은 하위노드 생성)시킴 • 정지규칙(Stopping rule)을 충족하면 성장을 중지시킴 • 목표변수가 연속형이면 "회귀나무“, 이산형이면 “분류나무 " 로 정의됨	• 처음 가지가 증가할 때 오분류가 감소 하다가 일정수준이상 가지가 증가하면 오분류율이 증가함 • 이때, 오분류율를 크게 할 위험이 높거나 부적절한 추론 규칙을 갖고 있는 가지 또는 불필요한 가지를 제거	• 혼동행렬(Confusion matrix)를 통해 모형 성능 평가	• 구축된 Tree Model을 해석하고 예측모형을 설정한 후 예측에 적용함

의사결정나무 알고리즘 종류



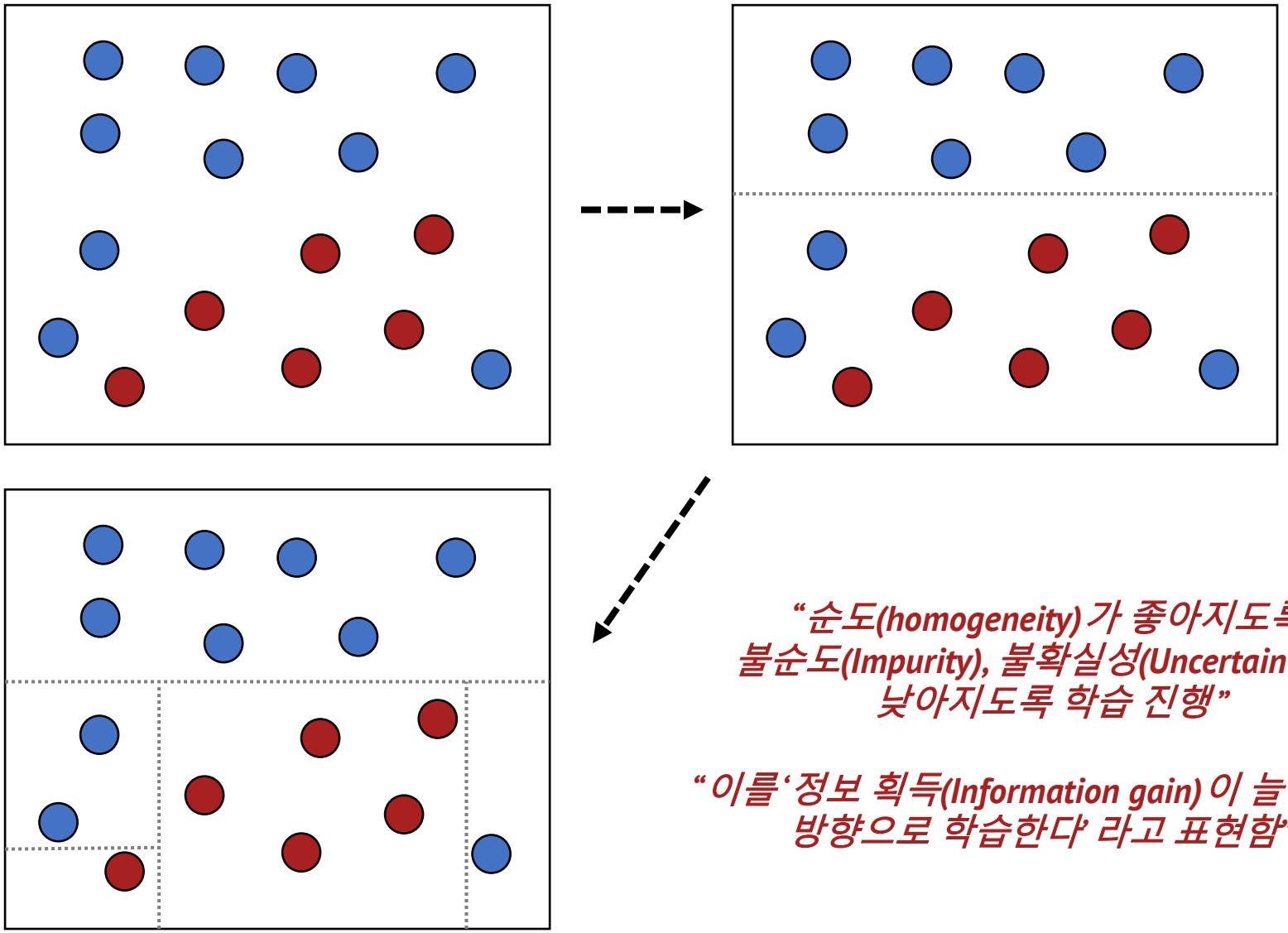
Source : http://www.birc.co.kr/2017/01/11/%EC%9D%98%EC%82%AC%EA%B2%B0%EC%A0%95%EB%82%98%EB%AC%B4decision-tree/

의사결정나무 알고리즘 종류

가장 널리 쓰이는 방법론으로 CART, CHAID, C4.5 or C5.0이 있으며, 이들은 나무의 성장과정에서 분할(Partitioning) 방법에 차이가 있음.

	분할방법	이산형 목표변수 분할변수 선택기준	연속형 목표변수 분할변수 선택기준
CHAID	다지분할	카이제곱 통계량	F-통계량 (ANOVA)
C4.5 or C5.0	다지분할(범주형) 2진분할(연속형)	엔트로피지수	•
CART (Classification and Regression Tree)	2진(Binary) 분할	지니(Gini) 계수 혹은 지니불순도	분산감소량

어떤 기준으로 분할하는가 ?

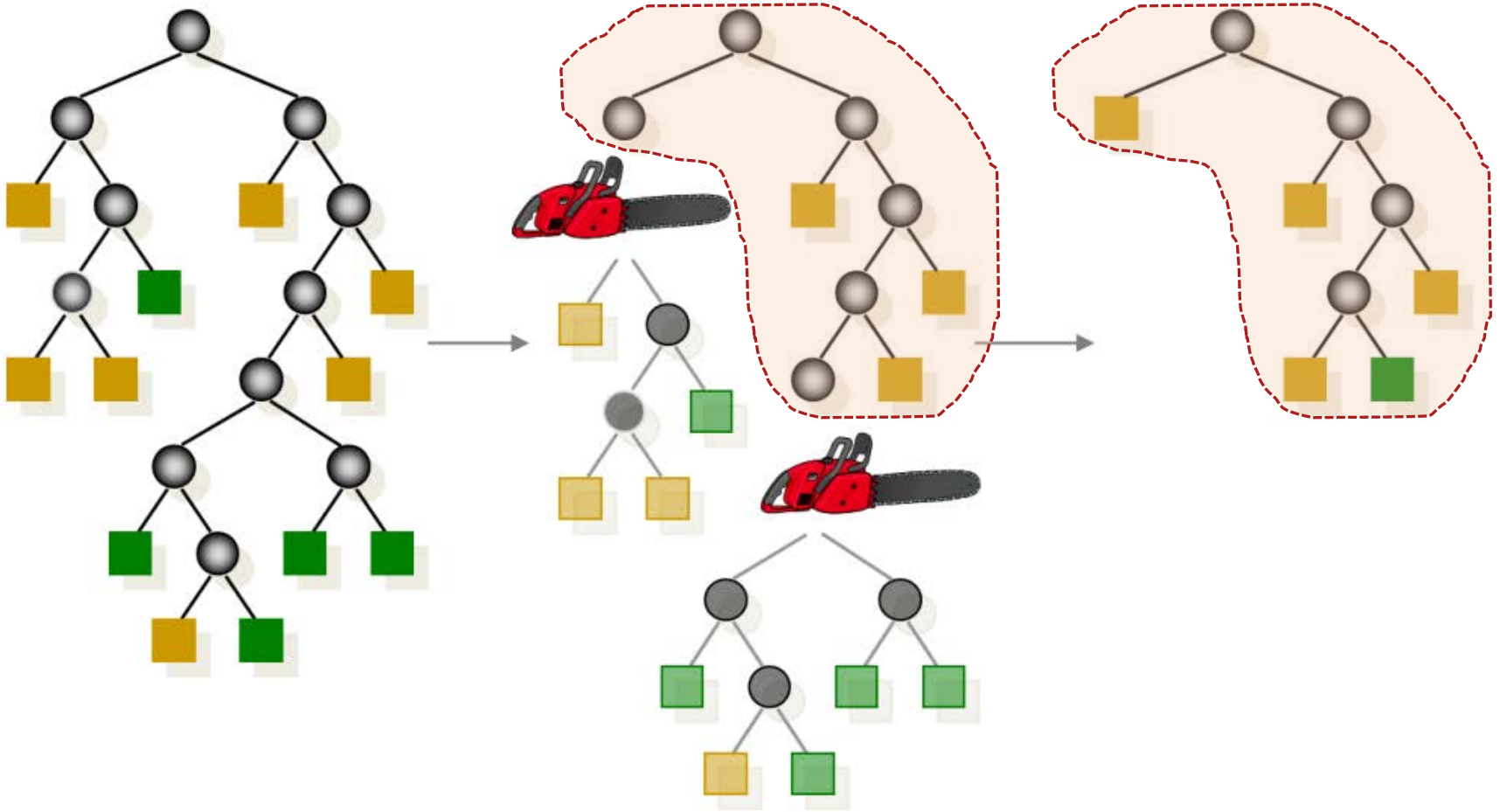


"순도(homogeneity)가 좋아지도록,
불순도(Impurity), 불확실성(Uncertainty)이
낮아지도록 학습 진행"

"이를'정보 획득(Information gain)이 늘어나는
방향으로 학습한다' 라고 표현함"

가지치기(Pruning)

재귀적 분할 과정에서 모든 Terminal node가 순도 100%가 되는(Full tree) 경우, 과적합이 발생해 오히려 오분류율이 높아짐. 가지치기(Pruning)는 적절히 잔가지를 잘라내 분기를 합치는 과정임



Source : https://ratsgo.github.io/machine%20learning/2017/03/26/tree/

의사결정나무의 특징 – 높은 해석력

- 일반적으로 최종 모형의 예측력(Prediction)과 해석력(Interpretation) 모두 중요하나 상황에 따라 둘 중 하나를 더 중시하기도 함
- 가령, 신용평가에서 심사 결과 부적격 반정이 나온 경우, 적격 혹은 부적격이 나왔다는 예측사실(예측력)보다 **왜 부적격이 나왔는지에 대해 고객에게 잘 설명(해석력)하는 것이 더 중요함**

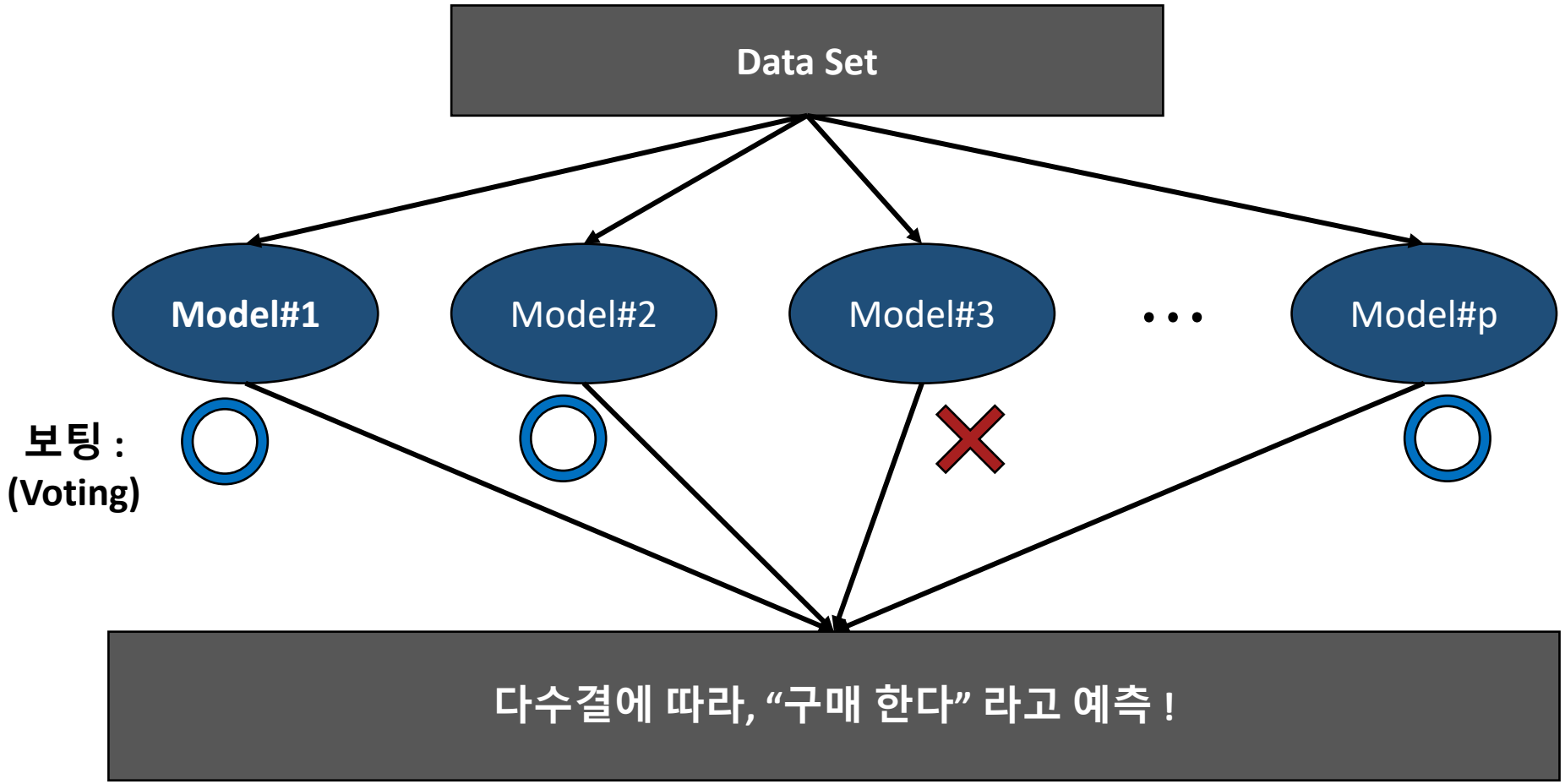
  ※ 로지스틱이나 판별분석은 변수 간 상대적 영향력 크기만 나타내지, 개별 데이터가 왜 해당 Class로 분류되었는지 명확하게 설명하는 것은 제한됨
- 의사결정나무는 상대적으로 다른 분류 지도학습(Supervised Learning)에 비해 예측성능이 다소 떨어지나 예측된 결과에 대한 **해석력이 매우 우수**하며, 결과를 쉽게 해석할 수 있다는 점에서 많이 활용됨

Lecture 9-2

랜덤 포레스트
(Random Forest)

앙상블 기법의 철학

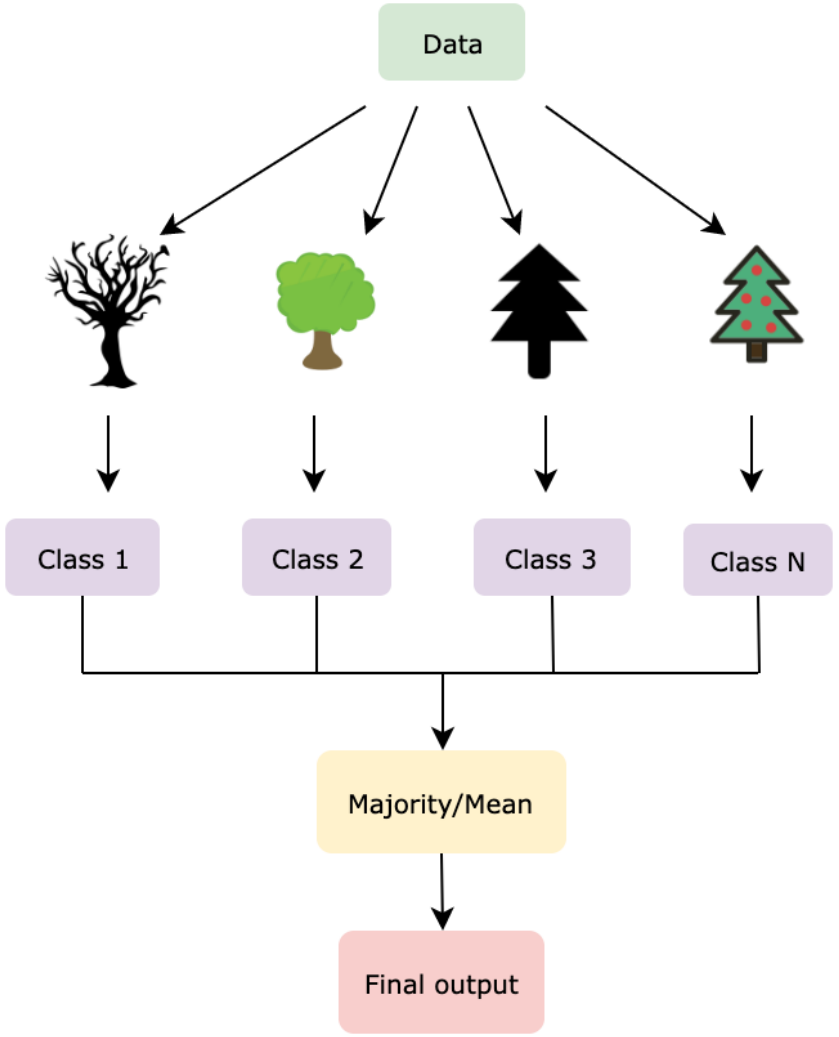
Q: 쿠폰을 발송하면, A 소비자는 구매할까 ? 구매하지 않을까?



앙상블 기법의 종류

단순집계 앙상블(Ensemble)	배깅(Bagging)	부스팅(Boosting)	스태킹(Stacking)
• 동일 알고리즘의 반복 혹은 여러 알고리즘의 예측 결과를 단순 집계함으로써 얻는 앙상블 모형 • 분류모형의 경우, 앙상블 결과의 Voting 방법을 통해 최종 예측 결과 선정 • 예측의 경우, 각 모형이 예측한 결과의 취합 방법으로 모형 평가	• Bootstrapping을 통해 여러 샘플을 생성한 후, 각 샘플을 단일 알고리즘으로 학습해 여러 버전의 모형을 만들어 개별 예측 • 예측된 결과 최종 투표(Voting)하여 예측결과 선정 • **"모든 문제를 동일한 가중치로 풀자"**	• 배깅과 거의 동일하나, 배깅의 경우 부스트랩 샘플에 대해 독립적인 여러 모형이 생성되는 데 반해, 부스팅은 순차적으로 학습시키면서 가중치를 조절하는 모형 • 즉, 먼저 생성된 모델을 개선해 나가면서 학습 진행 • **"틀린 문제를 더 유심히 보자"**	• 여러 독립적인 분류 모형들이 예측한 결과들을 모아 다시 해당 결과를 학습데이터로 정의함 • 상위의 학습모형이 이렇게 생성된 데이터를 재학습해서 최종 예측결과를 도출함 • 과적합 가능성 높음 • 계산복잡도가 매우 높아짐
여러모형 결합	**랜덤포레스트(Random forest)**	**AdaBoost, GradientBoost(GBM), XGBoost**	**여러 모형 결합**

Random forest – 배깅 방법



Source : https://blog.paperspace.com/random-forests/

- 랜덤 포레스트는 **<u>두 가지 방법</u>**을 이용해 트리를 생성함
  - ① 의사결정나무를 만들 때, 데이터의 일부를 복원추출로 꺼내고, 해당 데이터에 대해서만 의사결정을 만드는 방식 => **<u>부스트래핑(Boostrapping)</u>**
  - ② 노드 내 데이터를 하위 노드로 나누는 기준을 정할 때, 일부 변수만 대상으로 가지를 나눌 기준을 찾는 방법 => **<u>Feature Split</u>**
- 새로운 데이터에 대한 예측을 수행할 때, 여러 개의 의사결정 나무가 내놓은 예측 결과를 투표(Voting) 방식으로 합해 최종적으로 예측결과를 내놓음

여러 나무를 만든다; 부스트래핑(Boostrapping)

나이	소득	학생여부	신용 평가	구매여부
23	High	No	Fair	No
24	High	No	Excellent	No
35	High	No	Fair	Yes
42	Medium	No	Fair	Yes
46	Low	Yes	Fair	Yes
53	Low	Yes	Excellent	No
38	Low	Yes	Excellent	Yes
26	Medium	No	Fair	No
25	Low	Yes	Fair	Yes
42	Medium	Yes	Fair	Yes
21	Medium	Yes	Excellent	Yes
44	Medium	No	Excellent	Yes
40	High	Yes	Fair	Yes
56	Medium	No	Excellent	No

추출하고,

나이	소득	학생여부	신용 평가	구매여부
23	High	No	Fair	No
53	Low	Yes	Excellent	No
25	Low	Yes	Fair	Yes
42	Medium	Yes	Fair	Yes
56	Medium	No	Excellent	No

다시 넣고

추출하고,

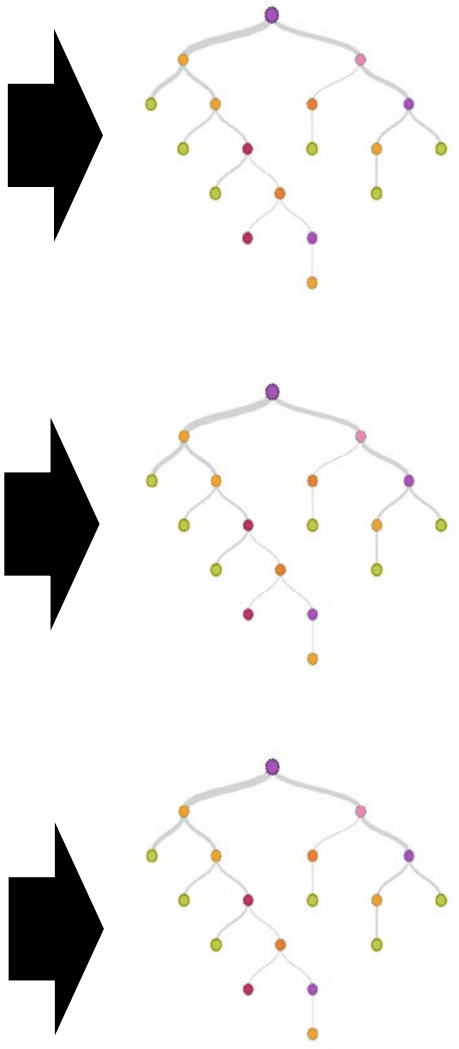
나이	소득	학생여부	신용 평가	구매여부
24	High	No	Excellent	No
35	High	No	Fair	Yes
53	Low	Yes	Excellent	No
38	Low	Yes	Excellent	Yes
42	Medium	Yes	Fair	Yes

다시 넣고

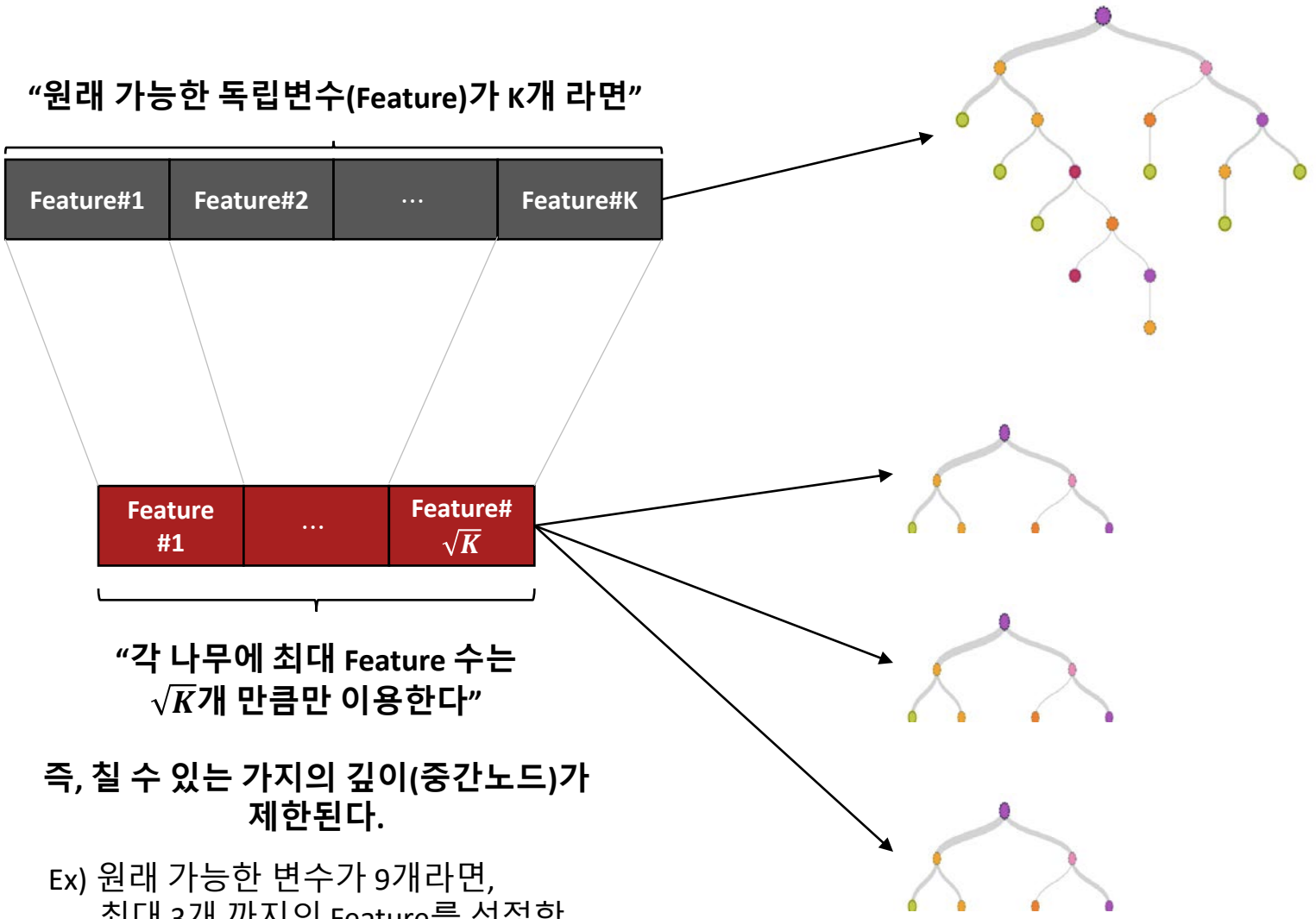
추출하고,

나이	소득	학생여부	신용 평가	구매여부
24	High	No	Excellent	No
35	High	No	Fair	Yes
46	Low	Yes	Fair	Yes
53	Low	Yes	Excellent	No
21	Medium	Yes	Excellent	Yes

다시 넣고



성장에 필요한 양분을 제한한다; Feature Split

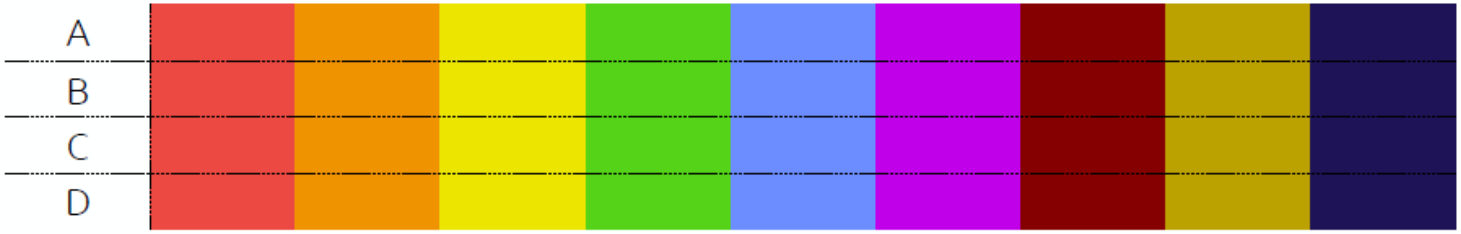


"각 나무에 최대 Feature 수는 $\sqrt{K}$개 만큼만 이용한다"

즉, 칠 수 있는 가지의 깊이(중간노드)가 제한된다.

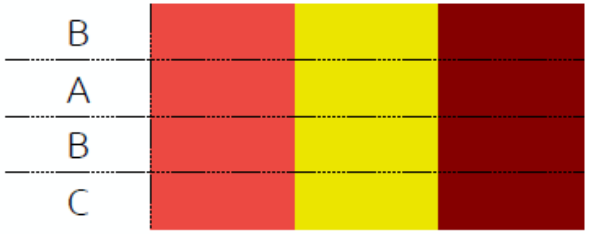
Ex) 원래 가능한 변수가 9개라면,
최대 3개 까지의 Feature를 선정함

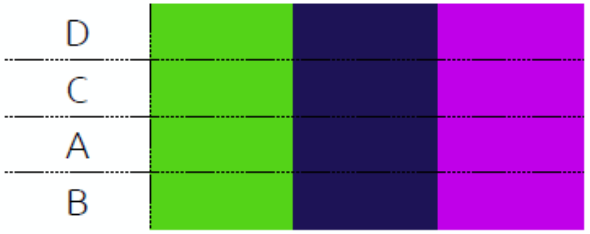
Bootstrapping 과 Feature Split

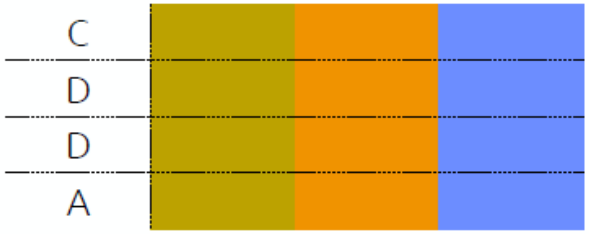


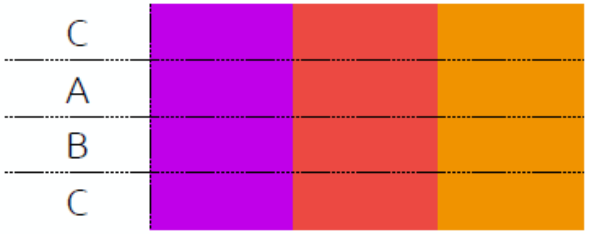
Feature Split

Bootstrapping

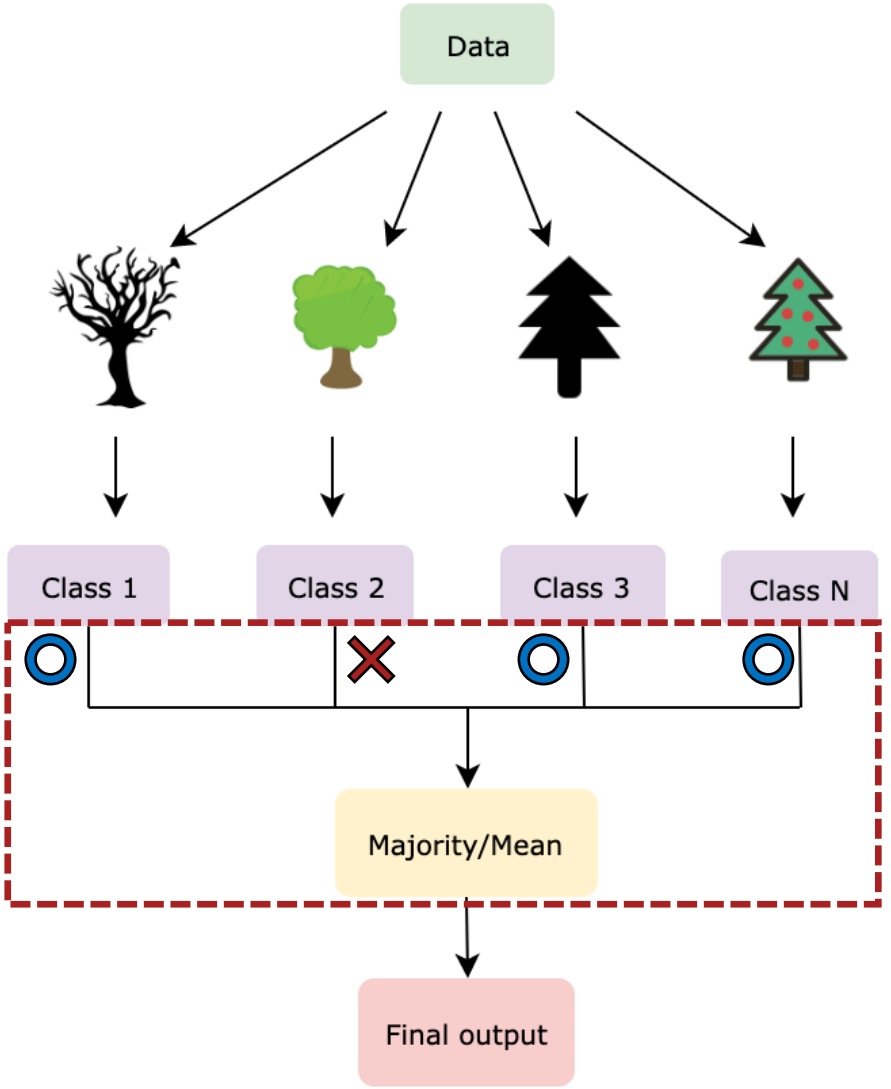








Voting을 통해 최종 예측



"다수결의 원칙"

✓ 각각의 나무가 예측한 결과를 바탕으로 분류 모형의 경우, 가장 많은 빈도를 차지한 클래스를 예측(Majority), 예측 모형의 경우 중심 경향성을 고려한 평균값(Mean)을 예측값을 제안함

랜덤포레스트의 장단점

장점

- 의사결정나무보다 예측력이 높음
- 분석하는 사람이 특별히 파라미터를 튜닝하지 않아도 높은 정확도를 보임
- 과적합 가능성을 낮춰줌 -> 배깅(Bagging)의 장점
- 중요 변수 선정 가능 -> 많은 나무에서 반복적으로 등장하는 변수

단점

- 의사결정나무와 비교했을 때, 예측과정을 시각적으로 표현 불가
- 데이터 셋이 커지면 학습(Training)과 예측(Testing) 모두 느려짐
  -> 정보 획득 계산 부담이 데이터 셋의 크기와 비례적으로 커짐
- 차원이 높은 데이터 즉, Feature 수가 많은 데이터는 예측 정확도 떨어짐
- 블랙박스 모델(모델의 설명력이 낮음)