



# Decision Tree

## R로 쉽게 이해하기!

조인식 조교



## 금요일의 즐거운 시험

예상시험시간 : 1시간 ~ 1시간반

시험내용 : Classification 알고리즘 활용/Clustering 알고리즘 활용

시험방법 : R studio를 이용한 실습시험 및 hwp 시험지파일로 결과제출

오전수업 이후 1시부터 2시반까지 개인스터디 후 2시반에 시험시작

시험을 다본 뒤 시험지파일(.hwp) + 이번주간 머신러닝실습파일(.R) 파일 압축 후

**(X조)학생이름.zip 파일로 제출**

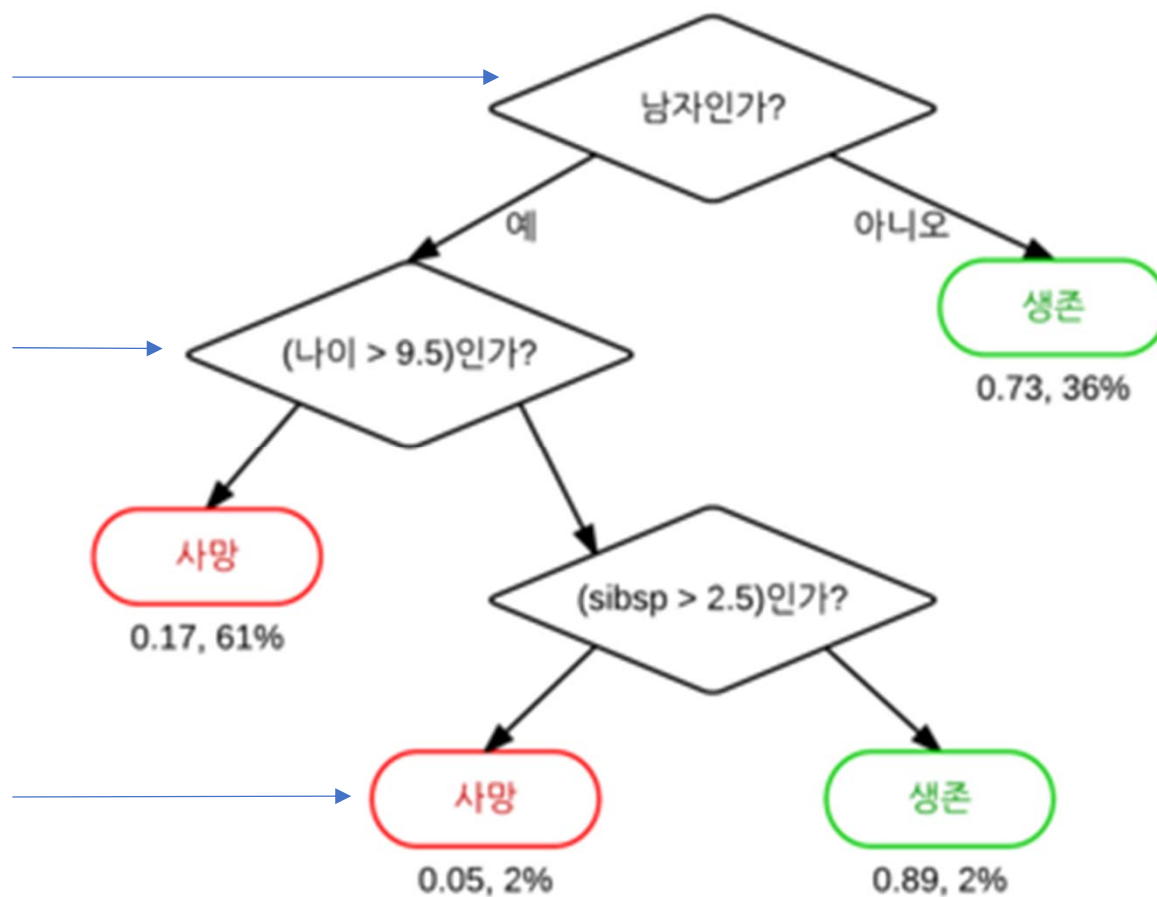
후 귀가

\* 인터넷/교재/카톡 사용 불가.

Root node

child node

Leaf node





depth

Titanic dataset을 이용한 decision tree




## Decision Tree의 활용목적

- Classification  
데이터들의 규칙을 찾아내 최적의 트리를 수축한 뒤 미래에 들어오는 데이터를 모델에 의해 분류하기 위하여 사용
  - Clustering(Segmentation)  
데이터들을 특정 그룹에 의하여 분류
  - Dimension reduction / Feature selection  
예측변수(독립변수)가 매우 많을 경우 이 중 목표변수(종속변수)에 영향을 미치며 작용하는 최적의 변수들을 찾아내 결과적으로 데이터의 차원을 축소시킴
- 



## 어떤 모델을 사용하는게 좋을가에 대한 고찰

- 통계 ↔ 기계학습  
통계는 데이터를 설명하고 규칙을 이해하기 위하여 데이터를 분석한다면  
기계학습은 데이터간의 규칙을 발견하여 미래의 데이터에 예측하는데에  
보다 초점을 둔다.
  - 기계학습에서의 설명력  
기계학습을 통해 구축된 모델이 어떠한 근거로 예측할 데이터(또는 test  
데이터)를 분류했는지를 분석가(또는 사용자)에게 설명해주는 능력
  - 기계학습에서의 예측력  
기계학습을 통해 구축된 모델이 TEST 데이터를 예측한 성능의  
척도(Accuracy, Recall 등)가 높음
  - 통상적으로 예측력이 높은 모델은 설명력이 낮고 설명력이 높은 모델은  
예측력이 낮다(= 완벽한 모델은 없다)
  - Decision Tree는 보통 설명력이 높고 예측력이 상대적으로 낮은 모델에  
해당함
- 

## 어떤 모델을 사용하는게 좋을가에 대한 고찰

- Decision Tree가 설명력은 높지만 예측력이 낮다면 좋은 알고리즘이 아닐까?

**No!** 비즈니스 분야에 따라서 데이터의 예측력보다 설명력을 더 요구하는 분야도 있다.

- **의료분야에서 암환자를 예측한다고 가정하자.**  
모델이 암환자인지 아닌지 예측할 때 정확도가 99% 일지라도 1%의 오차는 치명적으로 작용할 수 있다. 이럴때는 정확도는 80%일지라도 모델이 이를 의사에게 잘 설명해줄 경우 이는 진단에 더 효과적으로 사용될 수 있다.

## 어떤 모델을 사용하는게 좋을가에 대한 고찰

성능척도에 대한 고찰		실제 결과	
		정상	암환자
예측결과	정상	50	15
	암환자	5	30

$$\text{Accuracy} = (50+30) / (50+5+15+30) = 80\%$$

$$\text{Recall} = 50 / (50+15) = 91\%$$

$$\text{Precision} = 50 / (50+5) = 71\%$$

의료분야에서는 모델이 정상인을 암환자로 예측한 것 보다 암환자를 정상이라고 예측한 것이 더 치명적일 수 있다.

(정상인을 암환자로 예측한 경우 재검사를 하면 되지만 암환자를 정상인이라고 예측할 경우 암환자가 정상이라 판단되어 검사를 실시하지 않을 수 있으므로)

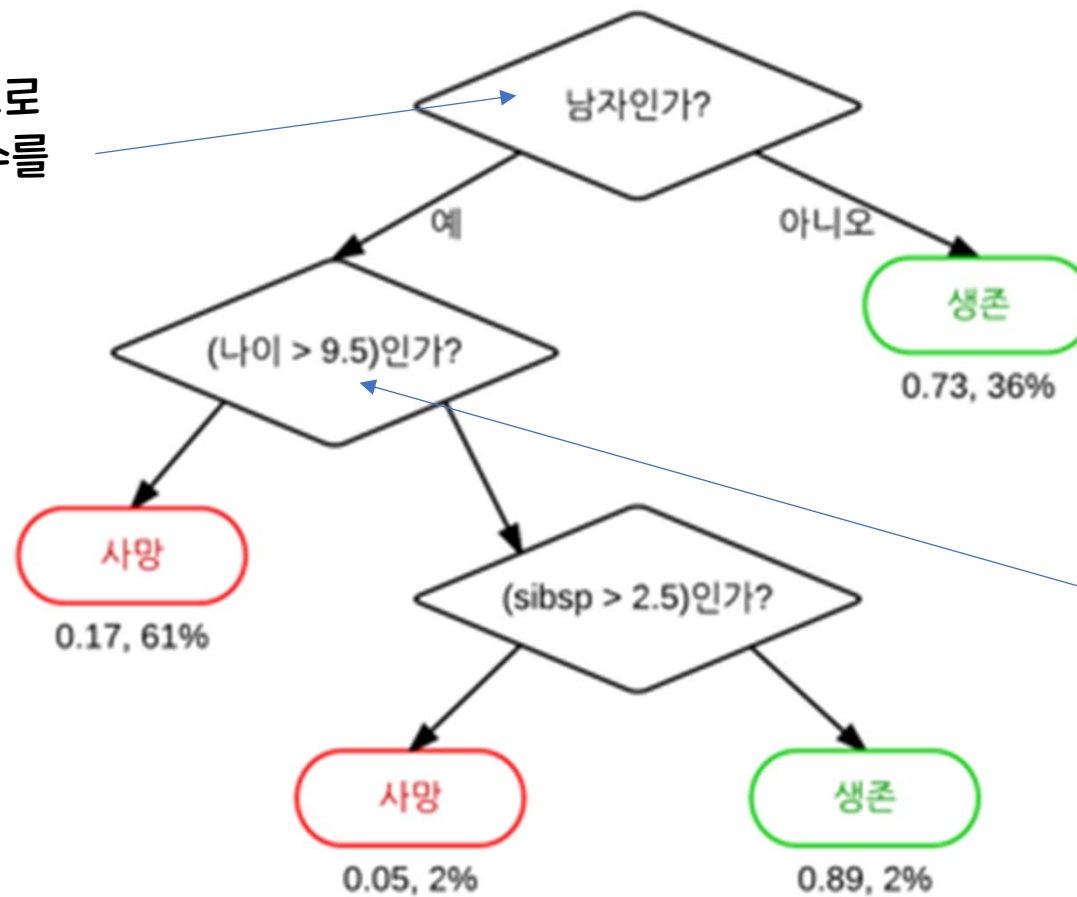
Accuracy 는 가장 공평한 평가방법일까 ? ... No!

# Decision Tree

Anyway...

Decision Tree에도 다양한 방법의 개선알고리즘이 존재한다.

어떠한 기준으로  
분기하는 변수를  
선택할텐가?



어떠한 기준으로  
해당 변수의  
분기점을 잡을텐가?



## Decision Tree

### 변수들의 선택의 기준

- 어떤 변수로 구분했을 때 가장 Class Label이 구분이 되는 점을 찾는다.

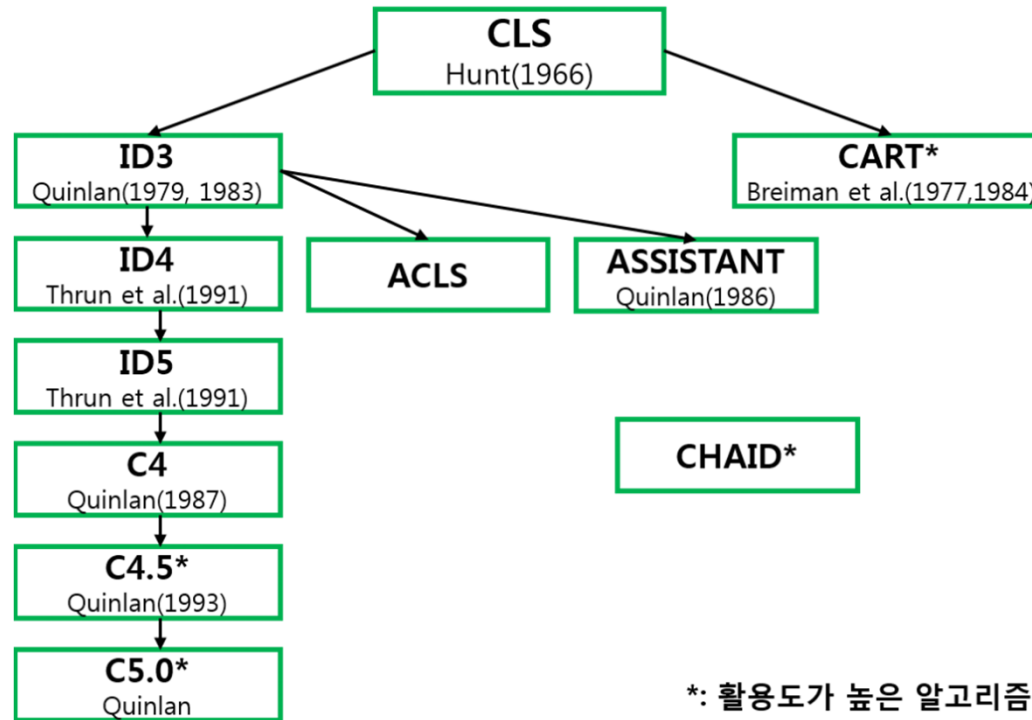
### 변수의 분기포인트의 기준

- 불순도가 가장 낮아지는 지점을 찾는다( = 순도가 가장 높아지는 지점을 찾는다)
- > 순도가 증가하고 불순도가 감소하는 방향으로 모델이 학습

### 순도와 불순도의 지표

- 카이제곱 통계량 : 데이터의 분포, 가정된 분포 사이의 차이를 나타내는 측정값
- 지니지수 : 순수도, 1에 가까울수록 데이터가 순수한 상태
- 엔트로피 지수 : 혼란도, 높을수록 데이터가 순수하지 않은 상태

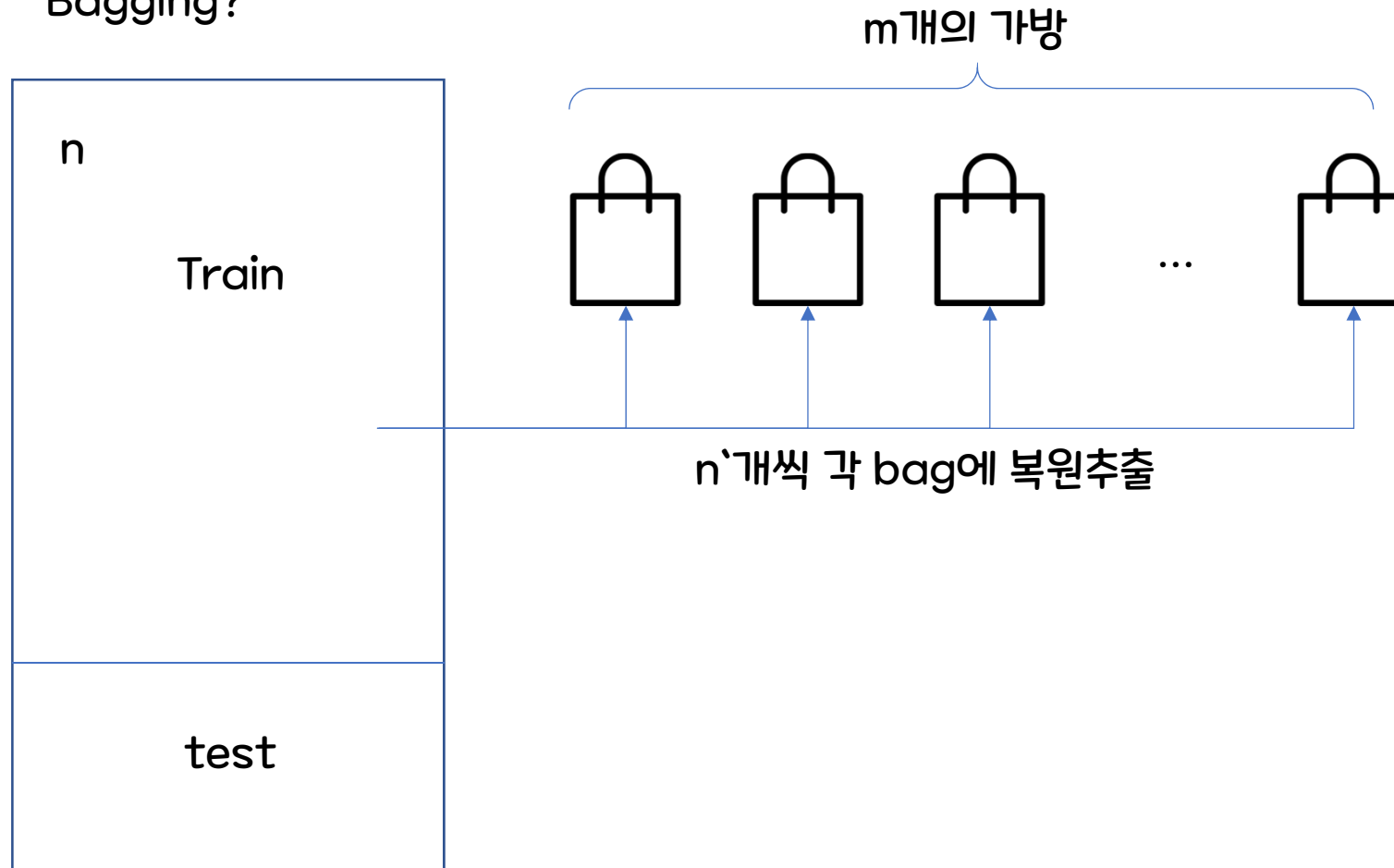
## Decision Tree의 계보



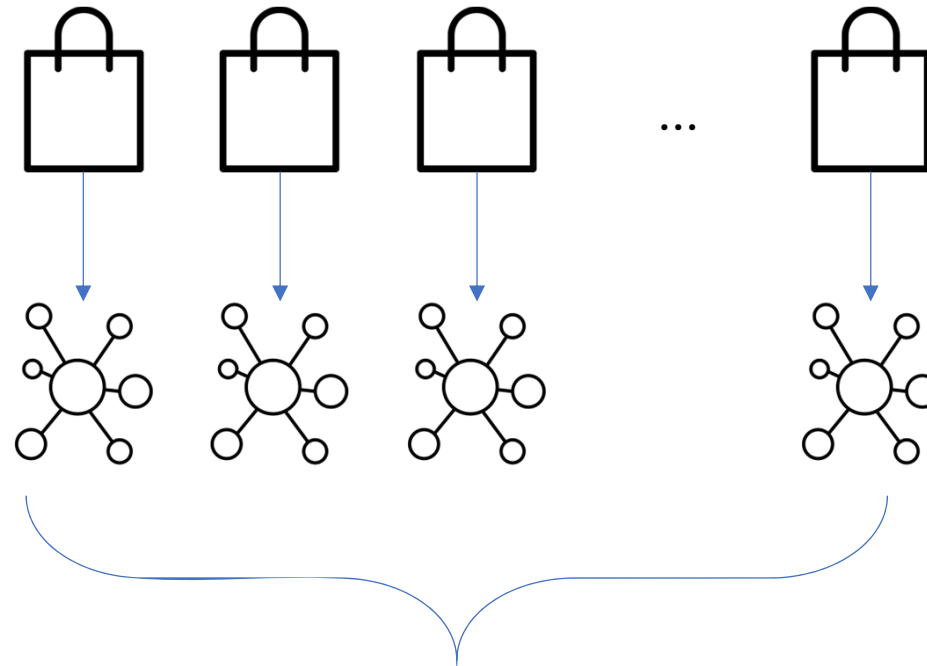
## Decision Tree의 낮은 예측력을 앙상블기법으로 보완한 Random Forest

앙상블 기법중 bagging(Bootstrap aggregating)기법을 활용한 대표적인 기법

Bagging?



## Decision Tree의 낮은 예측력을 앙상블기법으로 보완한 Random Forest



Y = vote result of sub models



# Random Forest

## Decision Tree

특정 질문에 대한 응답을 따라가는 방식으로 데이터를 분류해가는 알고리즘

## Forest

이러한 Decision Tree로 구성된 Forest(숲)을 구축하는 알고리즘으로 여러 Tree들에 대하여 다수결의 원칙에 따르듯 숲의 구성원들의 투표를 통해 의사결정을 내린다.

## Random

숲을 구성하는 각 트리를 구축할 때 각 트리를 구성하는 변수(열)을 무작위성에 근거하여 구성한다. 즉 각 트리는 모든 요소들에 대한 트리로 구성하지 않고 일부만을 각각 선별하여 작은 트리들을 만들어낸다.

## Decision Tree vs. Random Forest



Vs.



# Random Forest

## Vs. Decision Tree

- 예측성능 향상
- Bias (데이터 편향문제 해결)
- 모델의 신뢰도 향상
- 모델의 시각화가 어려워짐(각 tree들을 모두 시각화할 수 없으므로 설명력감소)
- 트리를 여러가지 구성하여야 하므로 속도적 성능 하락
- 데이터사이언티스트들이 좋아하는 모델  
HPO(Hyper Parameter Optimization)이 상대적으로 쉬움  
예측성능이 훌륭하게 나타남(“일단 랜포에 돌려보자”라는 말도 있음)  
연속형/명목형 변수를 둘다 사용가능
- 사장님이 싫어하는 모델  
분석업무를 맡긴 직원은 고성능을 얘기하지만 어떻게 그렇게 분류했는지가 명확하게 드러나지 않음(설명력 부족)
- 최근 각 관측치를 모델이 왜 그렇게 판단했는지를 알려주는 알고리즘도 개발  
(LIME PACKAGE)  
Why should I Trust you? 논문 참고추천



## My Little Project

- 건강보험자료의 국민의 혈압혈당 데이터를 토대로 고혈압을 진단하는 모델을 만들어보자.
- 데이터 : 국민건강보험공단(nhiss.nhis.or.kr) > 통계 > 국가건강검진데이터 > 혈압혈당데이터 에서 수령가능
- 클라이언트의 의뢰 : **혈압혈당데이터를 토대로 고혈압을 예측하는 모델을 만들어주세요.**
- 실습한 Decision Tree를 활용해보고 RandomForest도 활용해보자. 또한 이전에 실습한 kNN알고리즘도 활용해보고 각 알고리즘별로 성능이 어떠한지 판단한다.