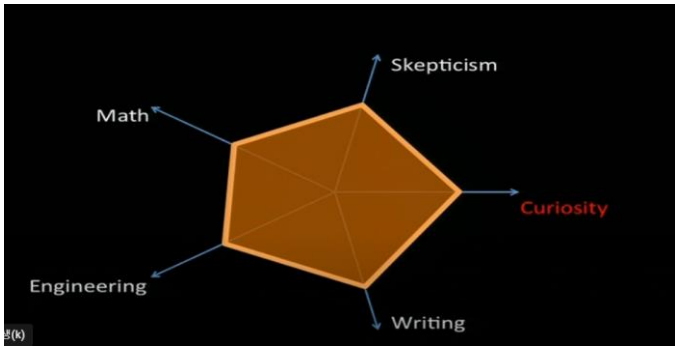
데이터 사이언티스트의 역량 6가지



John Rauser
(Data Scientist)
What is a Career in Big Data?

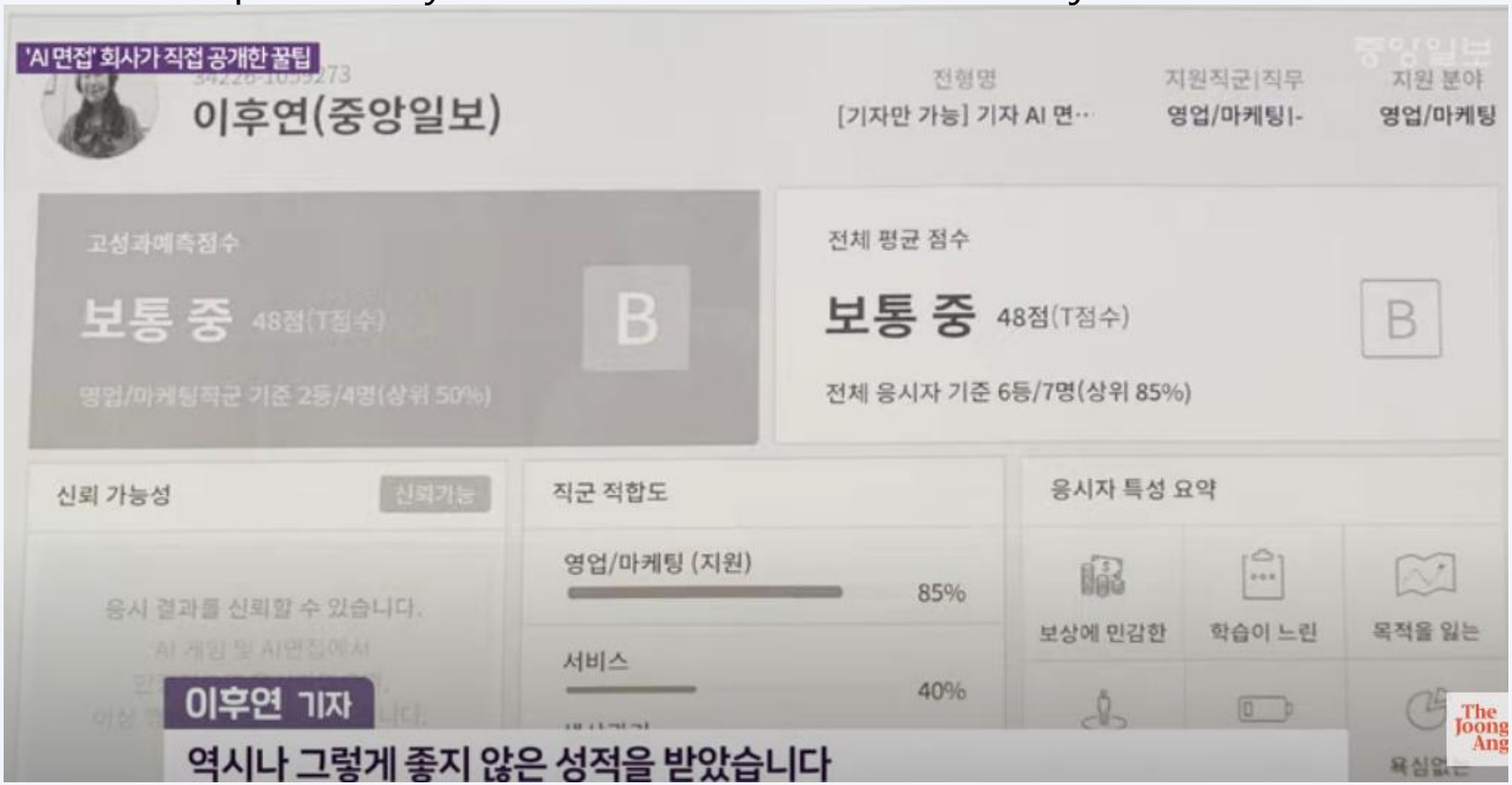
https://www.youtube.com/watch?v=0tuEEnL61HM

엔지니어링 | 수학/통계학 | 커뮤니케이션

비판적 사고 | 호기심과 탐구력 | 비즈니스 이해

https://www.youtube.com/watch?v=0tuEEnL61HM

- **AI면접**https://www.youtube.com/watch?v=8xuHFWZjcx4

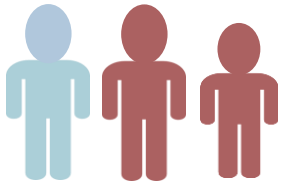


- **챗봇 올리비아 등** https://dataartist-kor.tistory.com/category/DT%20%26%20IT%20Evangelist%20%ED%99%9C%EB%8F%99%20%EC%A0%95%EB%A6%AC

비즈니스 상황

퇴사자 발생을 줄이는 것은 비즈니스적으로 매우 중요

퇴사자 1명 발생 → 채용+교육+생산성 저하 등 비용 발생 → 비즈니스 악영향



비즈니스 상황

Q.
당신은 반도체 대기업 A사의 채용팀으로, 직원을 채용하고 회사에 잘 적응하도록 교육하는 일을 맡고 있다. 하지만 최근 퇴사가 증가하여, 대신할 직원들을 다시 뽑고 교육하는 데 많은 비용이 들고 있다. 이를 해결하기 위해 직원의 퇴사 여부를 예측하여 알맞은 조치를 취하고자 한다.
*퇴사 여부를 예측하는 이진분류 모델을 만드시오. (데이터셋 별도 제공)

[공통사항]
- 데이터는 해당 회사의 직원관리를 위한 인사데이터로 총 1470 index로 구성
- 예측해야 할 변수 (Target)는 퇴사 여부 (Yes/No) 이며 알맞은 데이터 변환 필요
- Train/test 비율은 7:3, random_state = 1로 설정
- parameter 설정, 사용할 feature은 자유롭게 진행

데이터의 유형

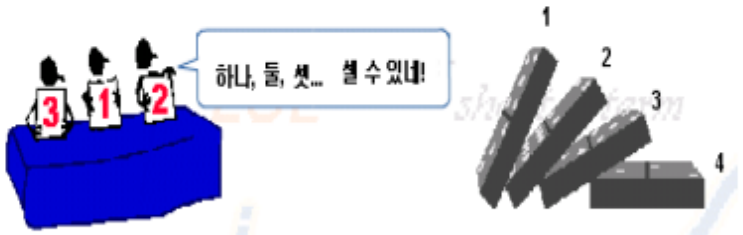
▪ 데이터 유형

연속형(Continuous)

- **나누어질 수 있고, 연속적으로 측정될 수 있는 것**
- 예 : 제품 중량(kg), 온도(℃), 강도(kg/㎟) 등 계량형 데이터

이산형(Discrete) , 범주형

- **나누어질 수 없고, 발생 빈도를 세어서 산출**
- 예 : 적합/부적합, 1등급, 2등급, 3등급 등



데이터의 유형

- 연속형 Data

등간척도 (Interval)

- 같은 간격을 가지지만 진정한 영점이 없는 척도로 수치의 비율 관계가 성립하지 않음.
- 관찰대상이 가지고 있는 속성 크기의 차이는 절대적 기준이 없어 상대적인 차이로만 나타남.
- 등간척도로 측정된 변수들간의 가감(+, -) 연산이 가능함
  예) 온도, 물가지수

비율척도 (Ratio)

- 등간척도에 비율의 개념이 추가된 척도로서 절대적 기준값이 존재하는 척도.
- 수치상 가감승제와 같은 모든 산술적인 사칙연산(+, -, ×, ÷)이 가능함.
  예) 제품 중량(kg), 강도(kg/mm^2) 등

▪ 이산형 Data

명목척도(Nominal)

- **관찰대상의 속성에 따라 관찰대상을 상호배타적이고, 포괄적인 범주로 구분하는 데이터.**
- **변수간의 사칙연산(+, -, ×,÷)은 의미가 없음.**

예) 성별(남,여), 품질(양품, 불량), 운동선수 등번호, 종교 등

추신수 **17**번

+

류현진 **99**번

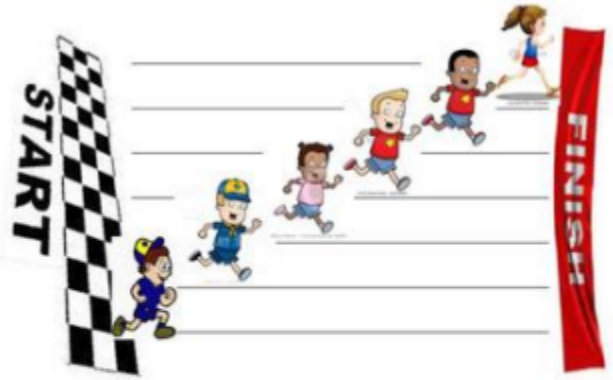
= 116

(아무런 의미 없음)

순위척도(Ordinal)

- **관찰대상이 가지고 있는 속성 크기에 따라 관찰대상의 순위, 서열을 부여하는 데이터.**

예) 만족도(1,2,3,4,5), 학교성적등급(1등, 2등, 3등), 크기(Small, Medium, Large) 등



용어 정리

머신-러닝 / 모델링

▪ 정의

: 대용량의 Data로부터 이들 Data 내에 존재하는 **관계, 패턴, 규칙** 등을 탐색하고 변수들간의 관련성을 찾아내어 **모형화**(수학적 함수, 논리적 구조)함으로써 유용한 지식을 추출하는 일련의 과정을 의미함.

▪ 종류

[Data 모델링]

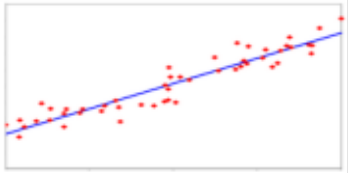
지도학습(Supervised Learning)

- 각각의 입력 x에 대해 레이블 y를 달아 놓은 Data를 이용해 모델링하는 방법

  * 회귀 : y 연속형, 분류 : y 이산형(범주형)

x:학습시간	y:점수
10	90
9	80
3	50
2	30

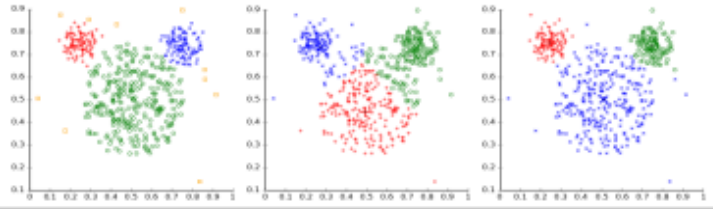
회귀 분석



비지도학습(Unsupervised Learning)

- 레이블 y 없이 Data x 만 이용하여 모델링하는 방법

  * 군집(Clustering), 연관분석



" 내가 지금 어떤 문제를 풀고 있는가? "

- **회귀** : 주어진 Data에 근거하여 모델을 만들고 이 모델을 이용하여 새로운 Case에 대하여 예측(강도, 온도)
- **분류** : 일련의 범주가 사전에 분류되어 있고 특정 Case가 어디에 속하는지를 결정(양품/불량, 스팸/정상)
- **군집** : 여러 속성의 Data를 비교하여 유사한 속성을 갖는 Data를 함께 그룹화시키는 것(고객 세분화)
- **연관** : 한 패턴의 출현이 다른 패턴의 출현을 암시하는 특성이나 항목간의 관계를 파악(장바구니 분석)

용어 정리

“ 내가 지금 어떤 문제를 풀고 있는가? “ - 회귀 vs 분류

“ 가지고 있는 데이터에 **독립변수**와 **종속변수**가 있고,
종속변수가 **숫자**일 때 **회귀**를 이용하면 됩니다.”

“ 가지고 있는 데이터에 **독립변수**와 **종속변수**가 있고,
종속변수가 **이름**일 때 **분류**를 이용하면 됩니다.”

용어 정리

- 데이터프레임, column, index, value

Column

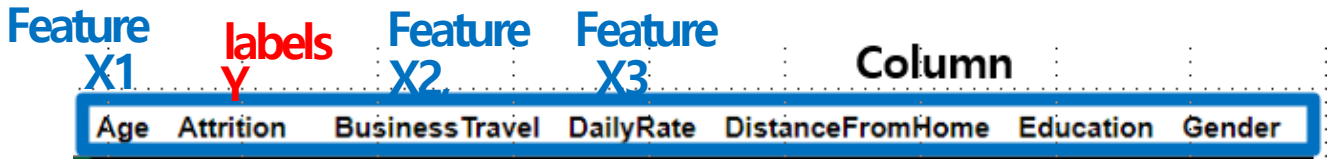
	Age	Attrition	BusinessTravel	DailyRate	DistanceFromHome	Education	Gender
0	41	Yes	Travel_Rarely	1102	1	2	Female
1	49	No	Travel_Frequently	279	8	1	Male
2	37	Yes	Travel_Rarely	1373	2	2	Male
3	33	No	Travel_Frequently	1392	3	4	Female
4	27	No	Travel_Rarely	591	2	1	Male
5	32	No	Travel_Frequently	1005	2	2	Male
6	59	No	Travel_Rarely	1324	3	3	Female

Index

Value

용어 정리

- **부르는 방법이 다양한 '변수'**



- **우리가 예측하고자 하는 column: (종속변수, label, y값, 목적변수…)**
  각 특성을 나타내는 column: (독립변수, Feature, X, 차원, 설명변수..)

퇴사 여부 예측 모델: (?)X1 + (?)X2 + (?)X3 + ... = Y

	Col1 나이	Col2 성별	Col3 출장빈도	Col4 통근거리	Col5 **퇴사**
사번1					퇴사
사번2					재직
사번3					퇴사
...					...

용어 정리

“ 내가 지금 어떤 문제를 풀고 있는가? “ - 회귀 vs 분류

“ 가지고 있는 데이터에 **독립변수**와 **종속변수**가 있고,
종속변수가 **숫자**일 때 **회귀**를 이용하면 됩니다.”

“ 가지고 있는 데이터에 **독립변수**와 **종속변수**가 있고,
종속변수가 **이름**일 때 **분류**를 이용하면 됩니다.”

용어 정리

“ 내가 지금 어떤 문제를 풀고 있는가? “

독립변수	종속변수	학습시킬 데이터를 만드는 방법
공부시간	시험점수 (10점, 20점)	사람들의 공부시간을 입력받고 점수를 확인한다.
온도	레모네이드 판매량	온도와 그날의 판매량을 기록한다.
역세권, 조망 등	집 값	집과 역까지의 거리, 수치화된 조망의 평점 등을 집 값과 함께 기록한다

용어 정리

머신-러닝 / 모델링

" 내가 지금 어떤 문제를 풀고 있는가? "

독립변수	종속변수	학습시킬 데이터를 만드는 방법
공부시간	합격 여부 (합격/불합격)	사람들의 공부시간을 입력받고, 최종 합격여부를 확인한다.
X-ray 사진과 영상 속 종양의 크기, 두께	악성 종양 여부 (양성/음성)	의학적으로 양성과 음성이 확인된 사진과 영상 데이터를 모은다.
품종, 산도, 당도, 지역, 연도	와인의 등급	소믈리에를 통해서 등급이 확인된 와인을 가지고 품종, 산도 등의 독립변수를 정하고 기록한다.

용어 정리

머신-러닝 / 모델링

“ 내가 지금 어떤 문제를 풀고 있는가? “

독립변수	종속변수	학습시킬 데이터를 만드는 방법
공부시간	합격 여부 (합격/불합격)	사람들의 공부시간을 입력받고, 최종 합격여부를 확인한다.
X-ray 사진과 영상 속 종양의 크기, 두께	악성 종양 여부 (양성/음성)	의학적으로 양성과 음성이 확인된 사진과 영상 데이터를 모은다.
품종, 산도, 당도, 지역, 연도	와인의 등급	소믈리에를 통해서 등급이 확인된 와인을 가지고 품종, 산도 등의 독립변수를 정하고 기록한다.
나이, 경력, 성별, 소속 팀, 미혼/기혼 여부...	퇴사 여부 (퇴사/재직)	인사관리 데이터를 활용한다.

용어 정리

▪ 정의

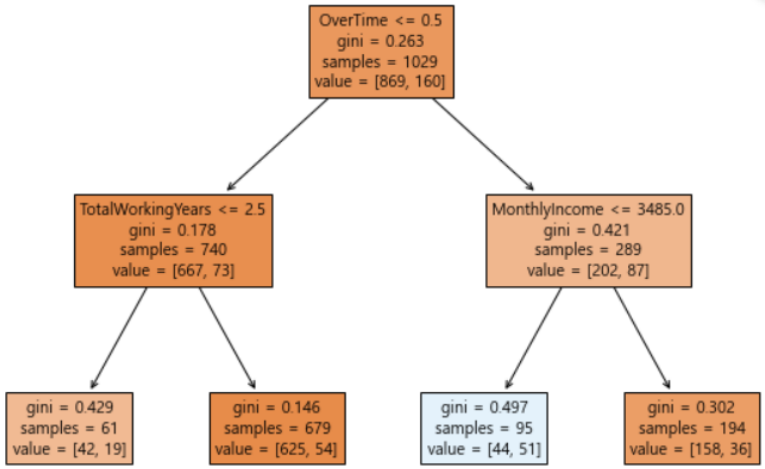
: 대용량의 Data로부터 특정 조건에 해당하는 데이터가 어디에 속하는 지를 결정하는 분석방법

▪ 종류

- 로지스틱 회귀분석(Logistic Regression)
- 의사결정나무(Decision Tree)
- 앙상블(Ensemble)
  - Random Forest
  - Gradient Boosting
  - Xgboost
- 지지벡터 머신(Support Vector Machine)
- K-Nearest Neighbors
- Naïve Bayesian Classification

용어 정리

의사결정나무



- **Decision Tree Classifier (의사결정 나무 분류모델)**
  - ✓ 설명변수들의 규칙, 관계, 패턴 등으로 관심대상인 목표변수를 분류하는 나무 구조의 모델을 만들고, 설명변수의 값을 생성된 모델에 입력하여 목표변수를 예측하는 지도학습 기법
  - ✓ 목표변수에 영향을 주는 설명변수를 탐색하고 해당 설명변수의 최적 분리기준을 제시
- **활용 용도**

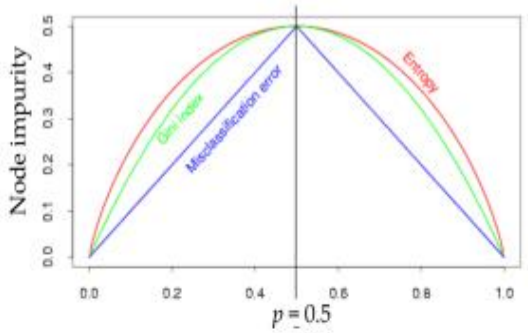
  다양한 분류(Classification) 분석에 활용되고 있음
- **장점**

  구조가 단순, 정규성/등분산성 등 가정이 불필요한 비 모수적 모형이다.

▪ 분리기준 : 불순도 함수 (impurity function)

▶ 지니 지수 $G(p_1, \dots, p_J) = 1 - \sum_{j=1}^{J} p_j^2$

▶ 엔트로피 지수 $E(p_1, \dots, p_J) = -\sum_{j=1}^{J} p_j log_2 p_j$



▪ 지니 지수

지니 지수
(Gini index)

지니 지수 감소량이 최대가 되는 분리에 의해 **자식 노드 형성** (CART에 적용)

*CART : 지니지수를 활용하여 Classification And Regression하는 의사결정 알고리즘

* 지니지수 : $1 - 2(3/8)^2 - 2(1/8)^2 = 0.69$

* 지니지수 : $1 - (6/7)^2 - (1/7)^2 = 0.24$

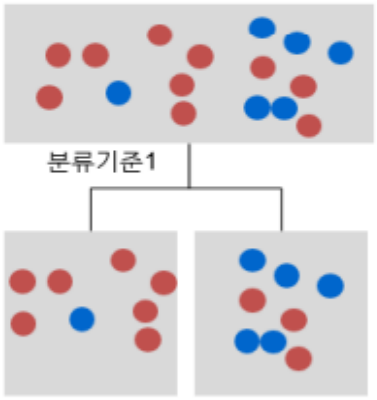
▪ 엔트로피 지수

엔트로피
(Entropy index)

엔트로피 지수 감소량이 최대가 되는 분리에 의해 **자식 노드 형성** (C5.0에 적용)

*C5.0 : 엔트로피를 활용하여 다중 분류/예측하는 의사결정 알고리즘

$$Entropy = \sum p_i \log_2 \frac{1}{p_i}$$



$$Entropy(A) = -\frac{10}{16}\log_2\left(\frac{10}{16}\right) - \frac{6}{16}\log_2\left(\frac{6}{16}\right) \approx 0.95$$

$$Entropy(B) = 0.5 \times \left(-\frac{7}{8}\log_2\left(\frac{7}{8}\right) - \frac{1}{8}\log_2\left(\frac{1}{8}\right)\right) + 0.5 \times \left(-\frac{3}{8}\log_2\left(\frac{3}{8}\right) - \frac{5}{8}\log_2\left(\frac{5}{8}\right)\right) \approx 0.75$$

▪ 과적합 방지를 위한 가지치기 (분기 정지 기준 설정)

▶ **더 이상 분기가 일어나지 않고, 현재의 마디가 끝마디가 되도록 하는 규칙 설정**

① 의사결정나무의 깊이(depth)를 지정 ② 끝마디의 데이터 수의 최소 개수 지정

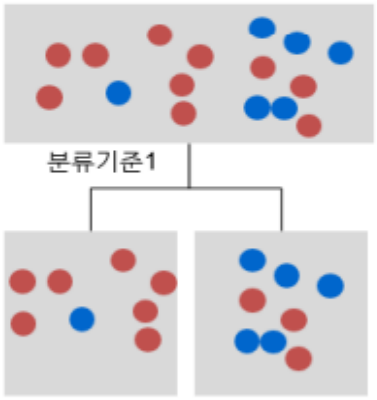
▪ 엔트로피 지수

엔트로피
(Entropy index)

엔트로피 지수 감소량이 최대가 되는 분리에 의해 **자식 노드 형성** (C5.0에 적용)

*C5.0 : 엔트로피를 활용하여 다중 분류/예측하는 의사결정 알고리즘

$$Entropy = \sum p_i \log_2 \frac{1}{p_i}$$



$$Entropy(A) = -\frac{10}{16}\log_2\left(\frac{10}{16}\right) - \frac{6}{16}\log_2\left(\frac{6}{16}\right) \approx 0.95$$

$$Entropy(B) = 0.5 \times \left(-\frac{7}{8}\log_2\left(\frac{7}{8}\right) - \frac{1}{8}\log_2\left(\frac{1}{8}\right)\right) + 0.5 \times \left(-\frac{3}{8}\log_2\left(\frac{3}{8}\right) - \frac{5}{8}\log_2\left(\frac{5}{8}\right)\right) \approx 0.75$$

▪ 과적합 방지를 위한 가지치기 (분기 정지 기준 설정)

▶ **더 이상 분기가 일어나지 않고, 현재의 마디가 끝마디가 되도록 하는 규칙 설정**

① 의사결정나무의 깊이(depth)를 지정 ② 끝마디의 데이터 수의 최소 개수 지정

- 주요 Parameter

▪ 잎사귀 노드 최소 자료 수(Leaf Size)

: 잎사귀의 최소 자료 수 지정.

: 최소 자료 수를 증가시키면 과대적합이 방지됨

: min_samples_leaf - 분리된 노드의 최소 자료 수. 이상치 영향, 과대적합 방지를 위해 적정 자료 수 지정

▪ 분리 노드의 최소 자료 수(Split Size)

: 분리 노드의 최소 자료 수 지정.

: 최소 자료 수를 증가시키면 과대적합이 방지됨

: min_samples_split – 분리가 되기 위한 현재 노드(상위 또는 부모 노드)의 최소 자료 수

▪ 최대 깊이(Maximum Depth)

: 분리 최대 깊이 지정

: 최대 깊이를 감소시키면 깊이 제약으로 과대적합 방지

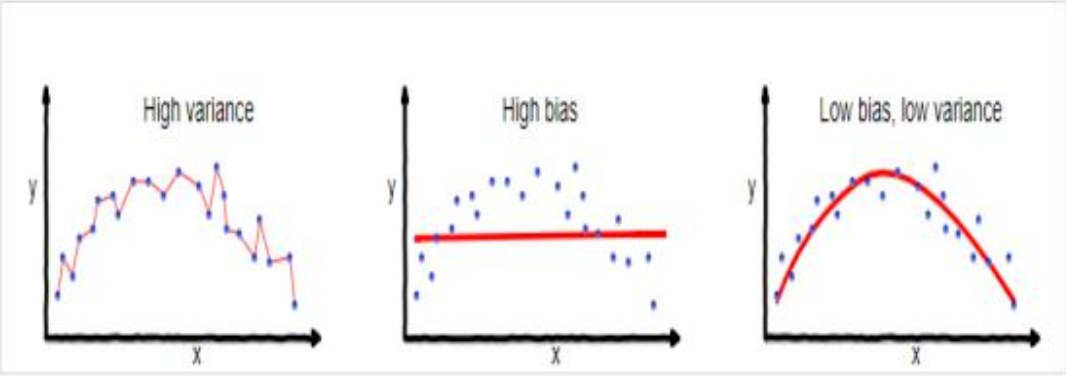
: max_depth – 분리되는 노드들의 최대 깊이 지정

용어 정리

과적합

▪ **정의**

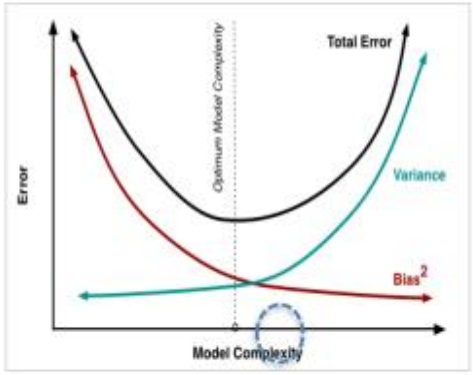
: 기계학습(machine learning)에서 학습 데이터를 **과하게 학습**(overfitting)하여 학습데이터에서는 성능이 좋으나, 평가 데이터에 대해서는 성능이 나쁜 경우를 의미.



Overfitting
(과대적합)
Low Bias, High Variance

Underrfitting
(과소적합)
High Bias, Low Variance

Ideal fitting
(최적 모델)
Low Bias, Low Variance



모델의 최적화

용어 정리

모델링 데이터 구조

▪ **정의**

: 모델을 만들기 위한 학습용 데이터와 모델링 결과를 평가하기 위한 평가용 데이터를 구분하는 것을 의미.

▪ **Hold Out**

① 데이터를 랜덤하게 Train과 Test 데이터로 나누며, Train 데이터는 분석모델을 만들고, Test 데이터로 모델을 검증함.

② 분석 모델을 적용하였을 때 오차를 최소로 하는 모델을 선택함.

▪ **k-Fold CV(Cross Validation)**

① 전체 데이터를 **K개**의 부분집합$\{D_1, D_2, \cdots, D_K\}$으로 나눔.

② 데이터 $\{D_1, D_2, \cdots, D_{K-1}\}$를 학습용 데이터로 사용하여 분석모델을 만들고 데이터 $\{D_K\}$로 검증을 함.

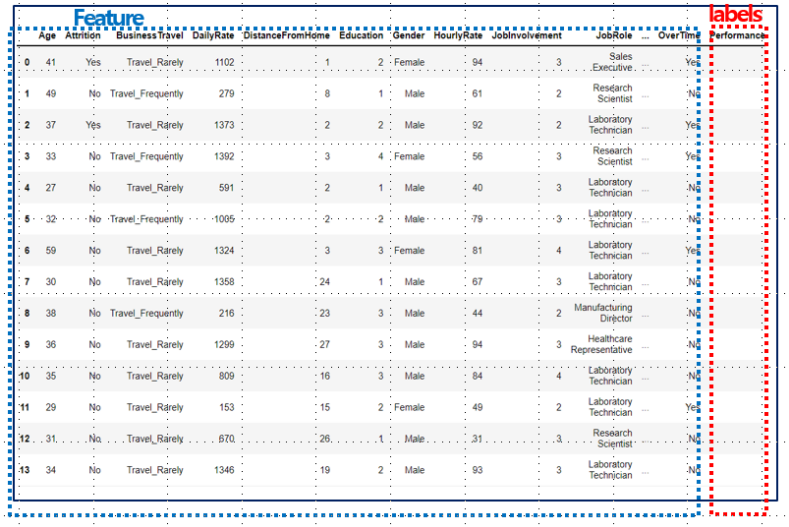
③ 데이터 $\{D_1, D_2, \cdots, D_{K-2}, D_K\}$를 학습용 데이터로 사용하여 분석모델을 만들고 데이터 $\{D_{K-1}\}$로 검증을 함.

④ 데이터 $\{D_2, \cdots, D_K\}$를 학습용 데이터로 사용하여 분석모델을 만들고 데이터 $\{D_1\}$로 검증을 함.

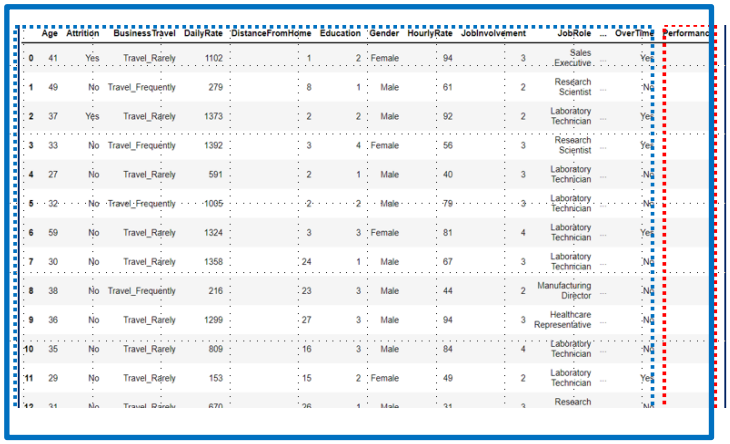
⑤ 이렇게 하면 총 K개의 모형과 검증결과가 나오며, **K개의 검증 성능**을 **평균**하여 최종 성능을 계산함.

용어 정리

모델링 데이터 구조

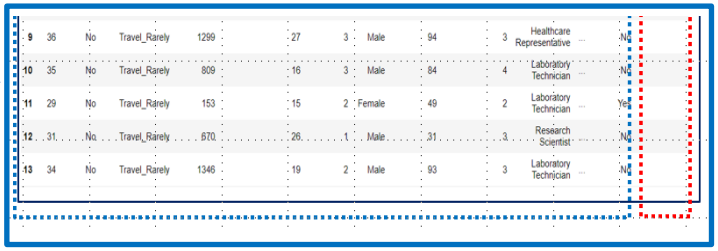


	Age	Attrition	BusinessTravel	DailyRate	DistanceFromHome	Education	Gender	HourlyRate	JobInvolvement	JobRole	...	OverTime	Performance
0	41	Yes	Travel_Rarely	1102	1	2	Female	94	3	Sales Executive	...	Yes	
1	49	No	Travel_Frequently	279	8	1	Male	61	2	Research Scientist	...	No	
2	37	Yes	Travel_Rarely	1373	2	2	Male	92	2	Laboratory Technician	...	Yes	
3	33	No	Travel_Frequently	1392	3	4	Female	56	3	Research Scientist	...	Yes	
4	27	No	Travel_Rarely	591	2	1	Male	40	3	Laboratory Technician	...	No	
5	32	No	Travel_Frequently	1005	2	2	Male	79	3	Laboratory Technician	...	No	
6	59	No	Travel_Rarely	1324	3	3	Female	81	4	Laboratory Technician	...	Yes	
7	30	No	Travel_Rarely	1358	24	1	Male	67	3	Laboratory Technician	...	No	
8	38	No	Travel_Frequently	216	23	3	Male	44	2	Manufacturing Director	...	No	
9	36	No	Travel_Rarely	1299	27	3	Male	94	3	Healthcare Representative	...	No	
10	35	No	Travel_Rarely	809	16	3	Male	84	4	Laboratory Technician	...	No	
11	29	No	Travel_Rarely	153	15	2	Female	49	2	Laboratory Technician	...	Yes	
12	31	No	Travel_Rarely	670	26	1	Male	31	3	Research Scientist	...	No	
13	34	No	Travel_Rarely	1346	19	2	Male	93	3	Laboratory Technician	...	No	



	Age	Attrition	BusinessTravel	DailyRate	DistanceFromHome	Education	Gender	HourlyRate	JobInvolvement	JobRole	...	OverTime	Performanc
0	41	Yes	Travel_Rarely	1102	1	2	Female	94	3	Sales Executive	...	Yes	
1	49	No	Travel_Frequently	279	8	1	Male	61	2	Research Scientist	...	No	
2	37	Yes	Travel_Rarely	1373	2	2	Male	92	2	Laboratory Technician	...	Yes	
3	33	No	Travel_Frequently	1392	3	4	Female	56	3	Research Scientist	...	Yes	
4	27	No	Travel_Rarely	591	2	1	Male	40	3	Laboratory Technician	...	No	
5	32	No	Travel_Frequently	1005	2	2	Male	79	3	Laboratory Technician	...	No	
6	59	No	Travel_Rarely	1324	3	3	Female	81	4	Laboratory Technician	...	Yes	
7	30	No	Travel_Rarely	1358	24	1	Male	67	3	Laboratory Technician	...	No	
8	38	No	Travel_Frequently	216	23	3	Male	44	2	Manufacturing Director	...	No	
9	36	No	Travel_Rarely	1299	27	3	Male	94	3	Healthcare Representative	...	No	
10	35	No	Travel_Rarely	809	16	3	Male	84	4	Laboratory Technician	...	No	
11	29	No	Travel_Rarely	153	15	2	Female	49	2	Laboratory Technician	...	Yes	
12	31	No	Travel_Rarely	670	26	1	Male	31	3	Research Scientist	...	No	
13	34	No	Travel_Rarely	1346	19	2	Male	93	3	Laboratory Technician	...	No	

Train: 70%



	Age	Attrition	BusinessTravel	DailyRate	DistanceFromHome	Education	Gender	HourlyRate	JobInvolvement	JobRole	...	OverTime	Performance
9	36	No	Travel_Rarely	1299	27	3	Male	94	3	Healthcare Representative	...	No	
10	35	No	Travel_Rarely	809	16	3	Male	84	4	Laboratory Technician	...	No	
11	29	No	Travel_Rarely	153	15	2	Female	49	2	Laboratory Technician	...	Yes	
12	31	No	Travel_Rarely	670	26	1	Male	31	3	Research Scientist	...	No	
13	34	No	Travel_Rarely	1346	19	2	Male	93	3	Laboratory Technician	...	No	

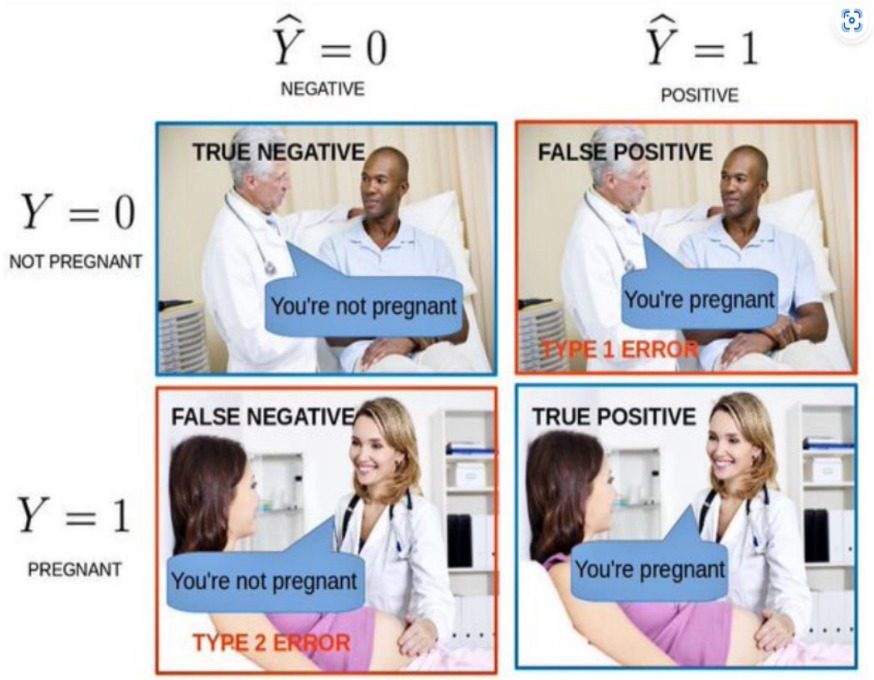
Test: 30%

- **분류모델의 평가**

  모델의 분류 레이블과 실제 레이블간의 정/오분류를 계산하여 모델의 분류 성능을 평가
  다양한 평가 기준 (정확도(accuracy), f1-score등)

Confusion Matrix

	Actually Positive (1)	Actually Negative (0)
Predicted Positive (1)	True Positives (TPs)	False Positives (FPs)
Predicted Negative (0)	False Negatives (FNs)	True Negatives (TNs)



머신러닝의 과정

***우리의 목적은 11월 수능에서 상위 90% 성적 내는 사람을 만드는 것**

수능 스타일의 문제를 **잘 풀 것 같은, 머리 좋은 사람**을 선택한다..	**목적에 맞는 모델** 선택 (목적: 퇴사자 분류 – 분류모델 선택)
수능 대비 **문제집**을 많이 사 놓는다 . 가능하면 이 사람의 공부 스타일에 맞도록.	**데이터를 준비**하고 미리 **가공** (결측치 확인, 데이터 타입 변경 등)
6월 9월 **모의고사**를 본다	**모델 평가** (score 확인)
90% 한참 밑이면 다른 사람으로 바꾸거나 **점수가 상승되도록 돕는다.**	모델간 **비교 / 파라미터 튜닝**

*** 수능 때 90% 성적을 낼 수 있을 것이다.**
(실무에 적용되었을 때 같은 성능이 나오도록 한다.)