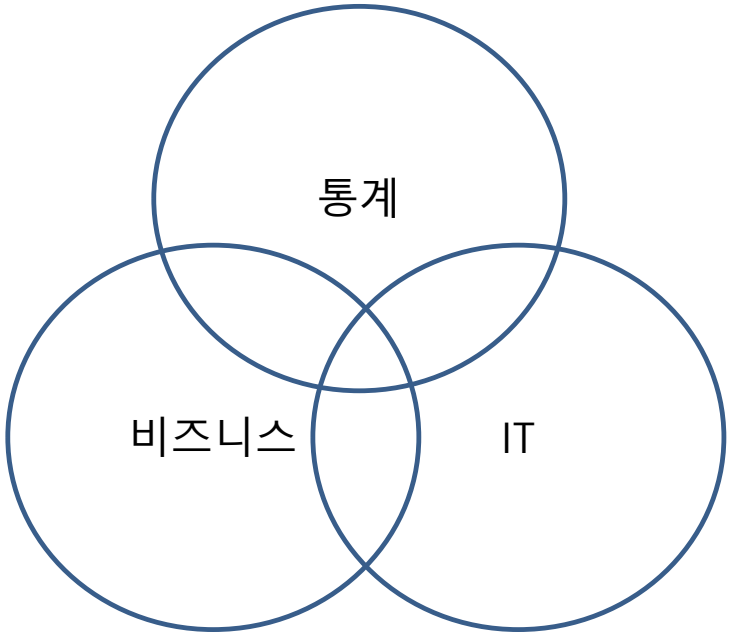
실무 데이터 분석

기업멤버십 SW 캠프

9시 5분에 시작하겠습니다.

Chapter 0. 데이터 분석의 트랜드 변화



0. 외부 전문가

-> 통계 + IT

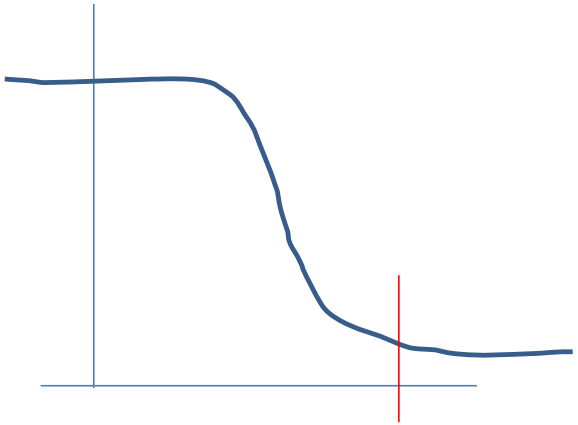
-> 모바일 Game N사 => 분석 Team

-> 과금 유도 / 잔존율

-> 보안 / 비용 …

1. 사내 교육

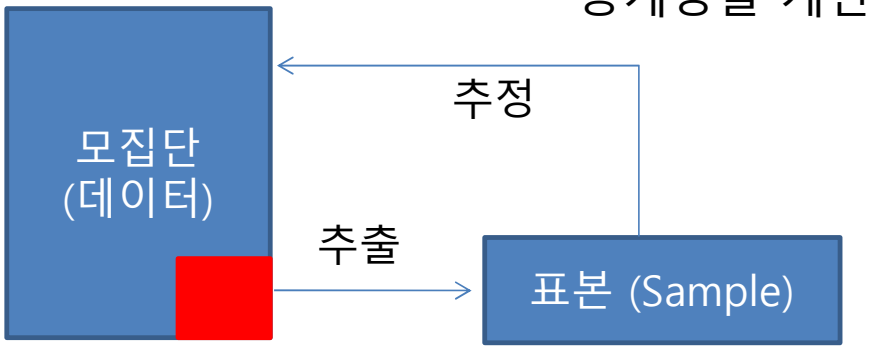
2. 채용



0. 데이터 분석의 트랜드 변화

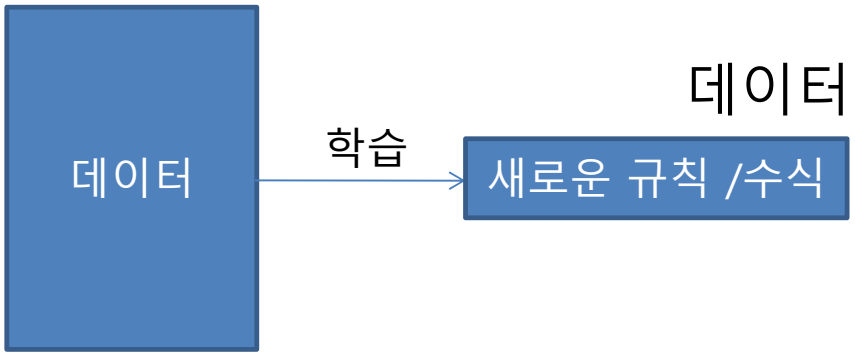
데이터 분석 : 정보, 자료들의 집합 (데이터)을 분석하여 Insight 도출

1. 통계학 (전통 통계) 파악하고자 하는 모집단에서 분석 가능한 일부를 추출하여, 통계량을 계산한 다음 , 모집단을 추정하는 형태



통계량 : 평균 / 최대 / 최소 / 합 …

2. 데이터 마이닝 (기계학습, Machine Learning) / 식스시그마



데이터 + 해답 -> **수식/규칙**

0. 데이터 분석의 트랜드 변화

데이터 분석 : 정보, 자료들의 집합 (데이터)을 분석하여 Insight 도출

3. 빅데이터

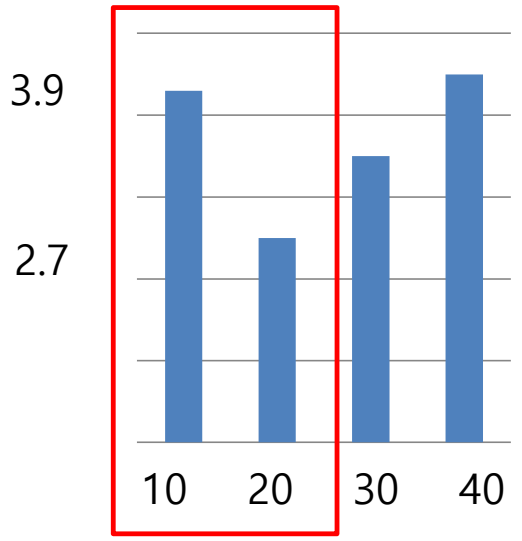
- 데이터 크기 : 여러대의 컴퓨터에 나누어 저장/처리/분석 하는 분산/병렬처리 기술

- 데이터 구조 : 정형 데이터 (Excel, CSV, … / RDBMS)
  비정형 데이터 (이미지, 소리, 영상, 자연어 …)

	이름	나이	주소	
0	홍길동	40	경기	
1	이몽룡	50	서울	
…	…	…	…	

Chapter 1. 통계적 데이터 분석 절차

1. 통계적 데이터 분석 절차



1. 기술적 데이터 분석 (DDA)

- 데이터를 수집 / 불러오는 단계
- **비즈니스 문제 정의**
- 데이터의 **구조와 타입**을 확인 (분석의 방법이 달라진다!)
- 목표변수(Y, Output, Label)와 설명변수(X, Feature, Input) 설정
- 기술 통계량 확인 (평균 / 분산 / 최빈값 / 결측값 ...)
- 데이터 전처리

2. 탐색적 데이터 분석 (EDA)

- 목표변수와 설명변수 간 관계/연관성/유사성 트랜드를 파악
- 데이터 시각화 / 데이터의 경향성을 파악하는 단계
- 데이터 시각화 -> 주관적 해석

3. 확증적 데이터 분석 (CDA)

- 앞서 확인한 트랜드을 **가설로 수립 (귀무/대립)** 하여 **객관적인 수치(P.value)로 검증**하는 작업
- 통계적 가설 검정
- 비즈니스 인사이트 도출 (insight) / Y 주당방문횟수 <-> X age / X 주소 ...)

4. 예측적 데이터 분석 (PDA)

- 수식화 작업을 수행 (Y = 100X1 + 200X2 +40)
- 새로운 X (설명변수) 데이터가 들어올 때, Y목표변수를 예측/대응
- 시스템화

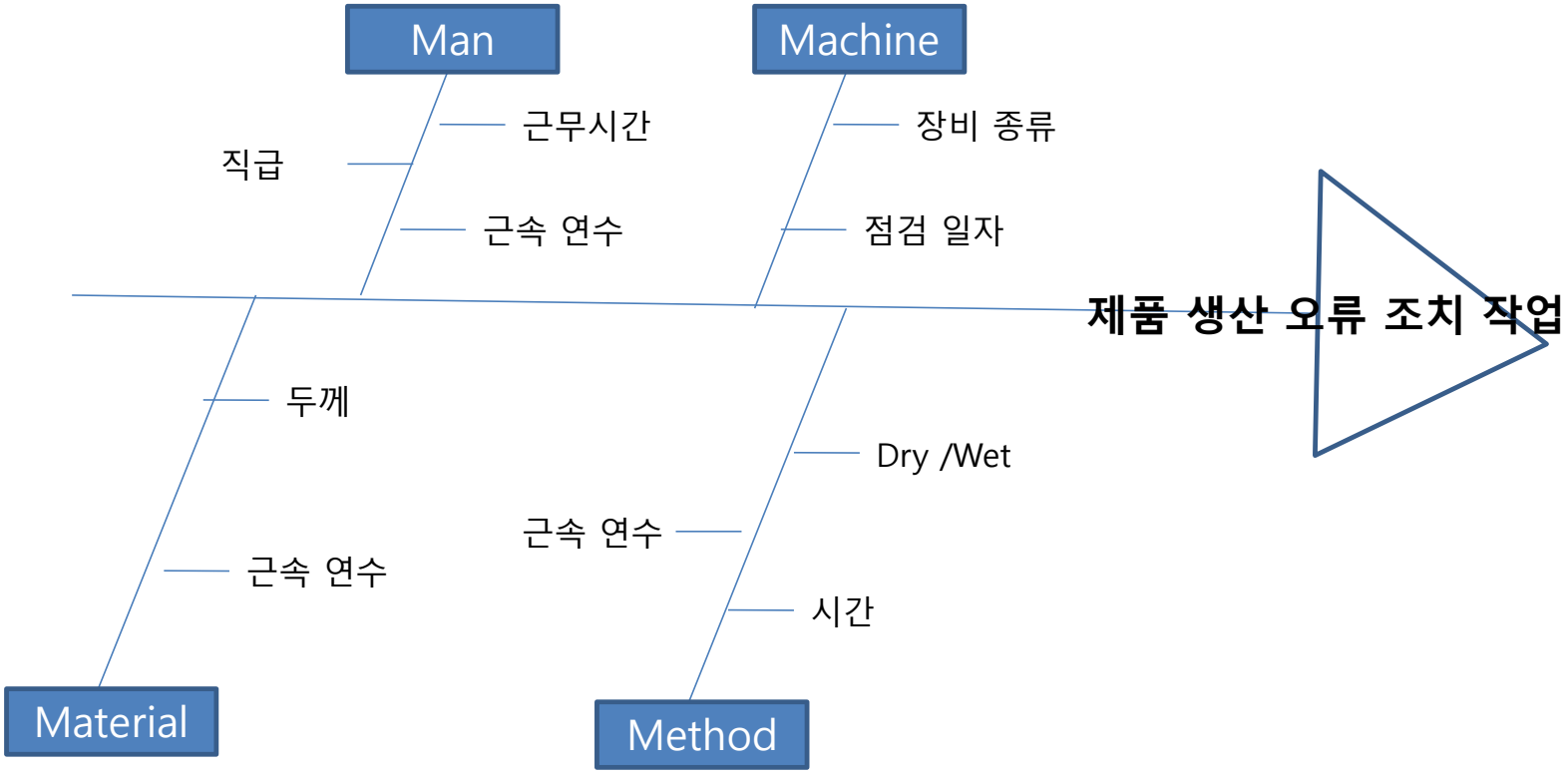
1. 통계적 데이터 분석 절차

1. 기술적 데이터 분석 (DDA)

- **데이터 수집**

특성 요인도 (Fish Bone Chart , 어골도) : 결과 (특성)에 어떤 요인(원인)이 있는지 그래프로 표현

-> QCD 제조/품질에서 QC7 (품질관리도구) 로 사용



1. 통계적 데이터 분석 절차

1. 기술적 데이터 분석 (DDA)

- **데이터 수집**

특성 요인도 (Fish Bone Chart , 어골도) : 결과 (특성)에 어떤 요인(원인)이 있는지 그래프로 표현

-> QCD 제조/품질에서 QC7 (품질관리도구) 로 사용

주요인자	수집가능성	중요도
근속연수	3	1
숙련도	1	9
근무시간	**9**	**9**
장비종류	**9**	**9**
수리일자	9	3
재료두께	3	3
...	...	...

항목명	데이터 타입	설명
근무시간	**연속형 (datetime)**	**작업자의 근무시간 (format)**
장비종류	**범주형**	
수리일자	**연속형 (datetime)**	
재료두께	연속형	
...		

1-1. 기술적 데이터 분석 (DDA)

DDA (기술적/묘사적 데이터 분석)

- 데이터를 불러오기 -> **데이터의 구조**와 **타입** 확인
- **정형데이터의 데이터 구조**
    1) Index (순서, 행, row) : 데이터의 개수 확인
    2) Column (항목, 열) : 데이터의 항목 확인 / 각 항목 별 **데이터 타입 (숫자/문자)**
    3) Value : 데이터 구성 값

index

	이름	성별	나이
0	홍길동	남성	30
1	이몽룡	남성	50
2	성춘향	여성	35
3	허준	남성	40
...	....		

Columns

Values

2022년9월5일

2022-06-05

- **데이터의 타입 (데이터 타입에 따라 분석의 방법과 방향성이 달라짐)**
    - 연속형 : 통계량 계산 (기술 통계량)
    - 범주형 : 항목 / 빈도수

- **결측치 (Missing Value) :** 데이터의 수집/저장/처리 과정에서 누락된 값

1-1. 기술적 데이터 분석 (DDA)

DDA (기술적/묘사적 데이터 분석)

- **기술 통계량 확인**

1) 연속형 :

- **대표 값 (중심위치) :** 해당 숫자 데이터을 대표할 수 있는 값 (평균 mean / 중앙값 median)

A : 2, 4, 1, 3, 5 / 평균 : 3

1, 2, 3, 4, 5 / 중앙값 : 3

B : 2, 4, 1, 3, 100000 / 평균 : 약 20000 (이상치 Outlier : 트랜드에 벗어난 값)

1, 2, 3, 4, 100000 / 중앙값 : 3

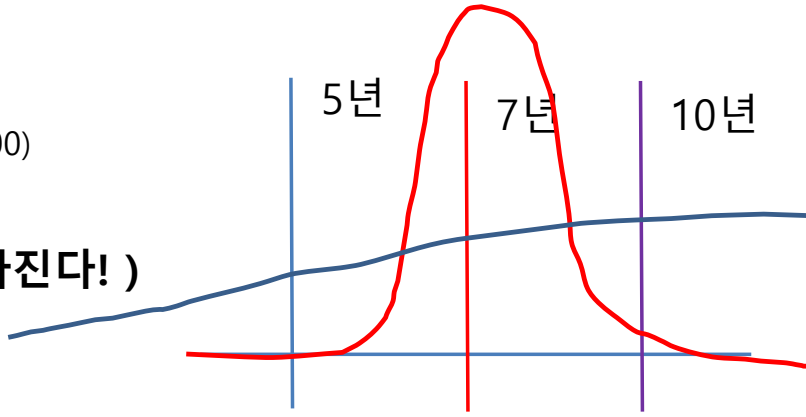
- **산포 :** 각 데이터들이 중심위치로 부터 얼마나 떨어져 있는가 척도 (분산 / 표준편차 / ...)

-> 중심위치로부터 얼마나 정확한가 / 신뢰성

Ex) 자동차 생산 -> **신형 G 차량 (품질 보증 기간 5년)**

- 타이어 휠 하청 -> **A (200,000) / B (200,000)** / C (500,000)
- 타이어 휠의 평균 수명 **-> A 7년 / B 10년**

- **분포의 모양 : 정규 분포 (분포모양에 따라 분석의 방법이 달라진다!)**



2) 범주형 : 항목 / 빈도수

1-1. 기술적 데이터 분석 (DDA)

- **기술 통계량 확인**

- **산포의 개념**

편차 (Deviation) = 개별값 – 평균

편차 합 (Sum of Deviation) = ∑ (개별값 – 평균)

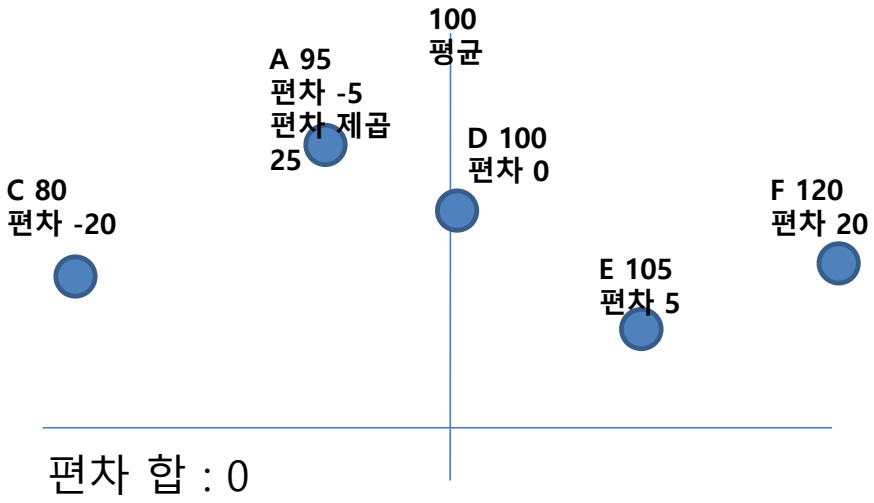
편차 제곱 합 (Sum of Square Deviation) = ∑ (개별값 – 평균)²

분산 (Variance) = ∑ (개별값 – 평균)² / n

표준편차 (Standard Deviation) = (∑ (개별값 – 평균)² / n)½

사분 범위 (Inter Quantile Range, IQR)
전체 데이터의 50% 분포해있는 구간

-> Box Plot



1-2. 탐색적 데이터 분석 (EDA)

1) 단일 변수

- 연속형 : 연속형 데이터의 분포 확인 (정규분포)

- 범주형 : 항목 / 빈도수 확인

2) 다 변수

- X: 범주형 / Y: 연속형 : 집단 간 통계량 비교

- X: 연속형 / Y: 연속형 : 두 데이터 간 상관성 확인

- X: 순서형(날짜) / Y: 연속형 : 시간(순서)에 따른 데이터의 추이

1-2. 탐색적 데이터 분석 (EDA)

1) 단일 변수

- 연속형 : 연속형 데이터의 분포 확인 (정규분포)

  Histogram : 숫자데이터의 분포를 막대 그래프로 표현

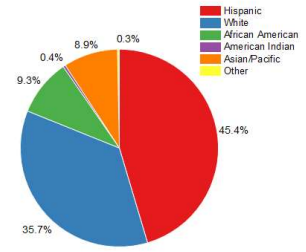
  (x: 해당 숫자 데이터 / y: 빈도수)

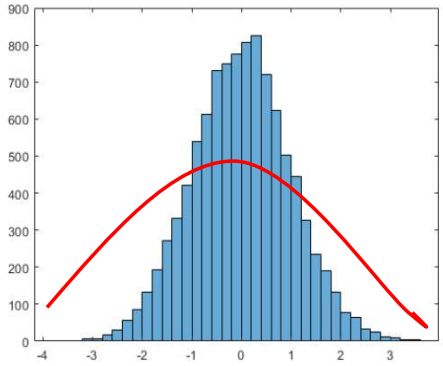
  Kernel Density Estimator (확률밀도함수, KDE)

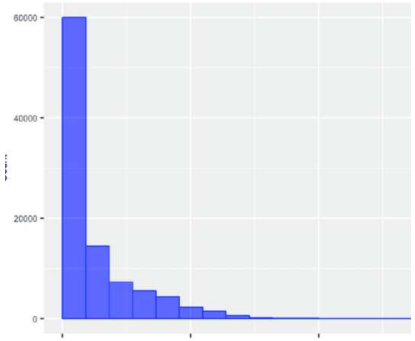
  Box Plot : 데이터의 분포를 사분범위(IQR)값을 이용해 표현

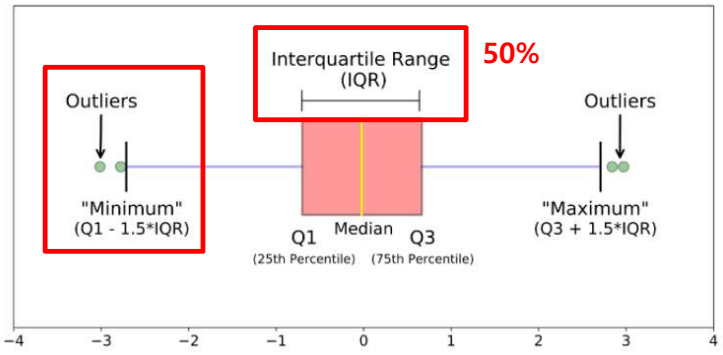
- 범주형 : 항목 / 빈도수 확인

  Bar Chart









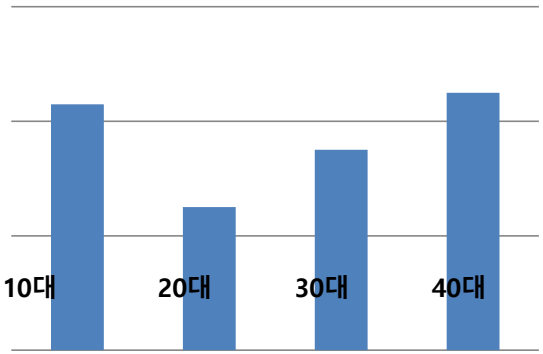
1-2. 탐색적 데이터 분석 (EDA)

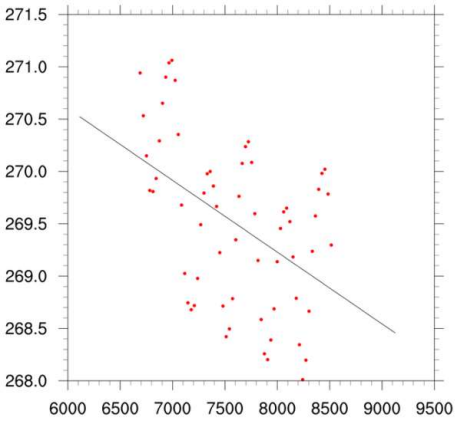
2) 다 변수

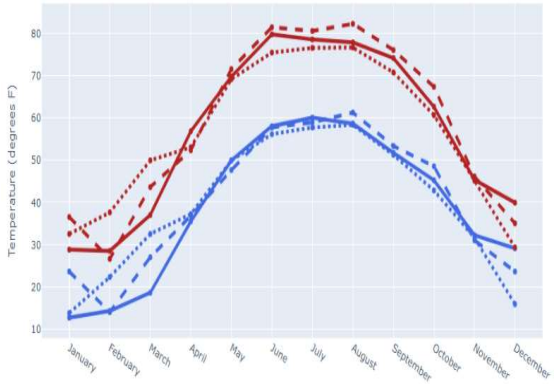
- X: 범주형 / Y: 연속형 : 집단 간 통계량 비교
  -> Bar Chart / Box Plot

- X: 연속형 / Y: 연속형 : 두 데이터 간 상관성 확인
  -> Scatter Plot (산점도)

- X: 순서형(날짜) / Y: 연속형 : 시간(순서)에 따른 데이터의 추이
  -> Line Plot (선그래프)







1-3. 확증적 데이터 분석 (CDA)

통계적 가설 검정 : 규명하고자 하는 바를 **가설**로 수립하여,

객관적인 통계 값 (P.value)를 이용해, 가설이 참인지 거짓인 판별하는 분석 기법

가설

- 귀무 가설 : 기각 시킬 목적으로 수립하는 가설 (**보통 가설**)

  (-> 평균의 차이가 없다 / 연관성이 없다 / 독립적이다 / 상관성이 없다 ...)

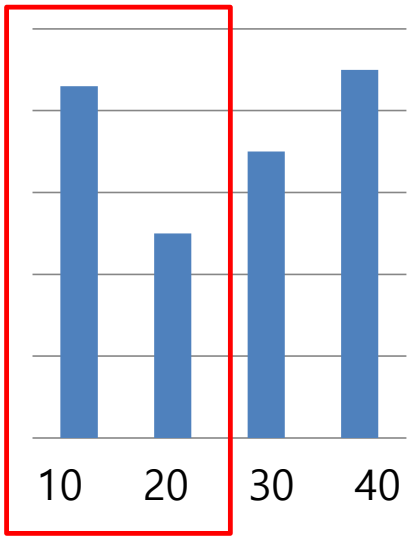
- 대립 가설 : 채택을 목적으로 수립하는 가설

  (-> 평균의 차이가 있다 / 연관성이 있다 / 상관성이 있다 ...)

- **P - value (확률 값) : 귀무 가설이 참일 확률 (0%~100%)**
- **유의 수준 (5%, 0.05)**

  -> P . Value > 0.05 : 귀무가설 참 (귀무가설기각실패)

  -> P . Value < 0.05 : 대립가설 참 (귀무가설기각)



Ex) 연령대 간의 주당 방문횟수의 차이가 있는가?

귀무 가설 : 10대와 20대의 주당방문횟수의 평균의 차이가 없다.

대립 가설 : 10대와 20대의 주당방문횟수의 평균의 차이가 있다. -> T test P. value 0.13 (13%)

1-3. 확증적 데이터 분석 (CDA)

통계적 가설 검정

1) 단일 변수

- **정규성 검정 (연속형) Normal Test**
    - 귀무 가설 : 해당 숫자데이터의 분포가 정규분포를 따른다. (중심극한정리)
    - 대립 가설 : 해당 숫자데이터의 분포가 정규분포를 따르지 않는다.
- 정규분포인지 아닌지 따라 분석 방법이 달라지기 때문

2) 다 변수

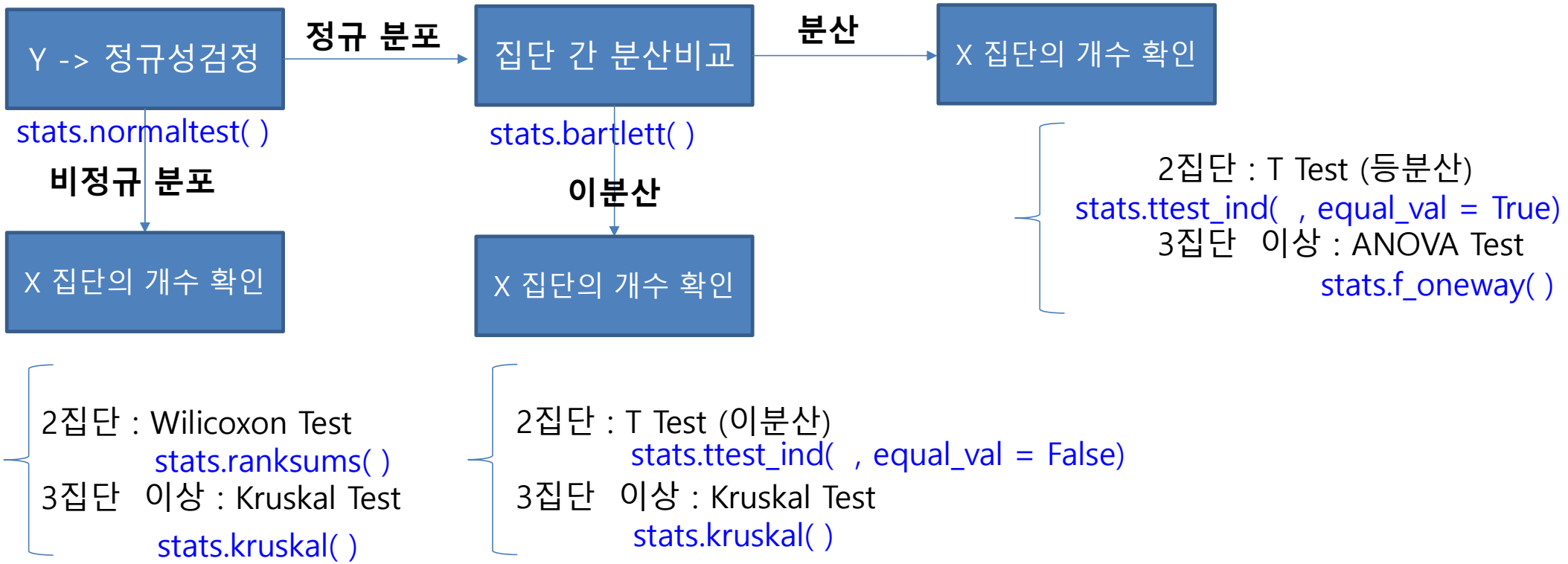
- X : 범주형 / Y : 연속형 – 집단 간 평균/분산을 비교하는 경우
    - 정규 분포 :
    - 비정규 분포
- **X : 연속형 / Y : 연속형 – 두 숫자데이터의 상관성이 있는지 확인하는 경우**
    - 정규 분포 : Pearson Test (**stats.pearsonr( )**)
    - 비정규 분포 : Spearman Test (**stats.spearmanr( )**)
- **X : 범주형 / Y : 범주형 – 두 항목이 서로 독립/연관 있는지 확인하는 경우**
    - Chi^2 Test / **stats.chi2_contingency( )**

1-3. 확증적 데이터 분석 (CDA)

통계적 가설 검정

2) 다 변수

- X : 범주형 / Y : 연속형 – 집단 간 **평균**을 비교하는 경우



- **가설 검정**

1) 데이터 타입 -> 숫자 / 문자

2) 숫자가 포함되어 있다 **->** 숫자 데이터에 대해 **정규성 검정 (stats.normaltest( ))**

- 숫자가 없다 (문자 vs 문자 -> 두 항목간 독립성 검정 **) :**

df_c = pd.crosstab(df1[범주형1] , df1[범주형2])

stats.chi2_contiengency(df_c)

3) 정규분포가 아닌 경우 (p.value < 0.05)

-> 집단 간 평균 비교 ? (나머지 변수가 문자)

-> 집단 개수 2 : stats.ranksums( ) / 집단 개수 3 : stats.Kruskal( )

-> 두 숫자의 상관성을 비교 ? (나머지 변수도 숫자)

-> stats.spearmanr( )

4) 정규분포 (p.value > 0.05)

-> 집단 간 평균 비교 ? (나머지 변수가 문자) / 등분산 검정 stats.levene( )

-> 분산이 같을 때 (P.value >0.05) 2 / stats.ttest_ind(, equal_var =True) | 3 / stats.f_oneway( )

-> 분산이 다르게 (P.value < 0.05) 2 / stats.ttest_ind(, equal_var=False) | 3 / stats.Kruskal( )

-> 두 숫자의 상관성을 비교 ? (나머지 변수도 숫자)

-> stats.pearsonr( )

1-4. 예측적 데이터 분석 (PDA)

수식화 (회귀분석 -> 기계학습(데이터마이닝))

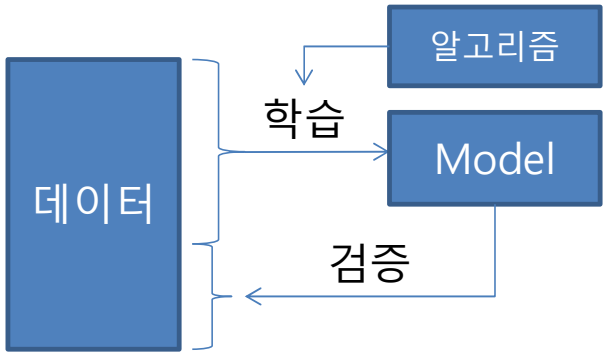
- **기계학습** : 데이터 간 연관성/관계/수식등을 컴퓨터가 학습을 통해 도출해내는 작업
- **실무적 Point :**
  - 학습 능력 : 데이터로부터 적절한 규칙을 찾아내는 능력
  - 일반화 능력 : 학습으로부터 얻은 수식에 새로운 데이터가 들어 올 때, 잘 예측 하는 능력

- 기계학습의 핵심 3요소 :

1) 데이터 (교과서) : 학습의 목적에 맞게
데이터를 깔끔하게 다듬는 작업
-> 특성 공학 (Feature Engineering)

2) 알고리즘 (선생님) : 학습의 목적에 맞게 / 데이터 맞게 적절한 알고리즘을 선택
-> 회귀분석 / 결정나무 / SVM / 앙상블 /신경망 ...

3) 하드웨어 (학생) : 비용 (Cost)



Chapter 2. 기계학습 (Machine Learninig)

2-1. 기계 학습 개요 (Machine Learning Intro)

- 수식화 (회귀분석 -> 기계학습(데이터마이닝))
- **기계학습** : 데이터 간 연관성/관계/수식등을 컴퓨터가 학습을 통해 도출해내는 작업

기계 학습의 종류

- 1) 지도 학습 : Y(목표변수, Output, Label)와 X(설명변수, Feature, input) 간 관계를 규명하여 수식을 도출, 새로운 X 왔을 때, Y 예측 하는 학습 방식
  - 회귀 (Regression) : Y (연속형)
  - 분류 (Classification) : Y (범주형)

- 2) 비지도 학습 : X(설명변수)끼리 연관성/수학적 거리/ 군집화 등 여러 기법을 이용하여 비슷한 데이터를 묶거나 찾는 학습방식
  - 연관분석 : 장바구니 분석 / 추천시스템
  - 군집분석

- 3) 강화 학습 : 컴퓨터가 시뮬레이션을 통해 주어진 상황에 대해서 사용자 제공하는 적절한 보상을 획득하는 방향으로 학습하는 방식

2-1. 기계 학습 개요 (Machine Learning Intro)

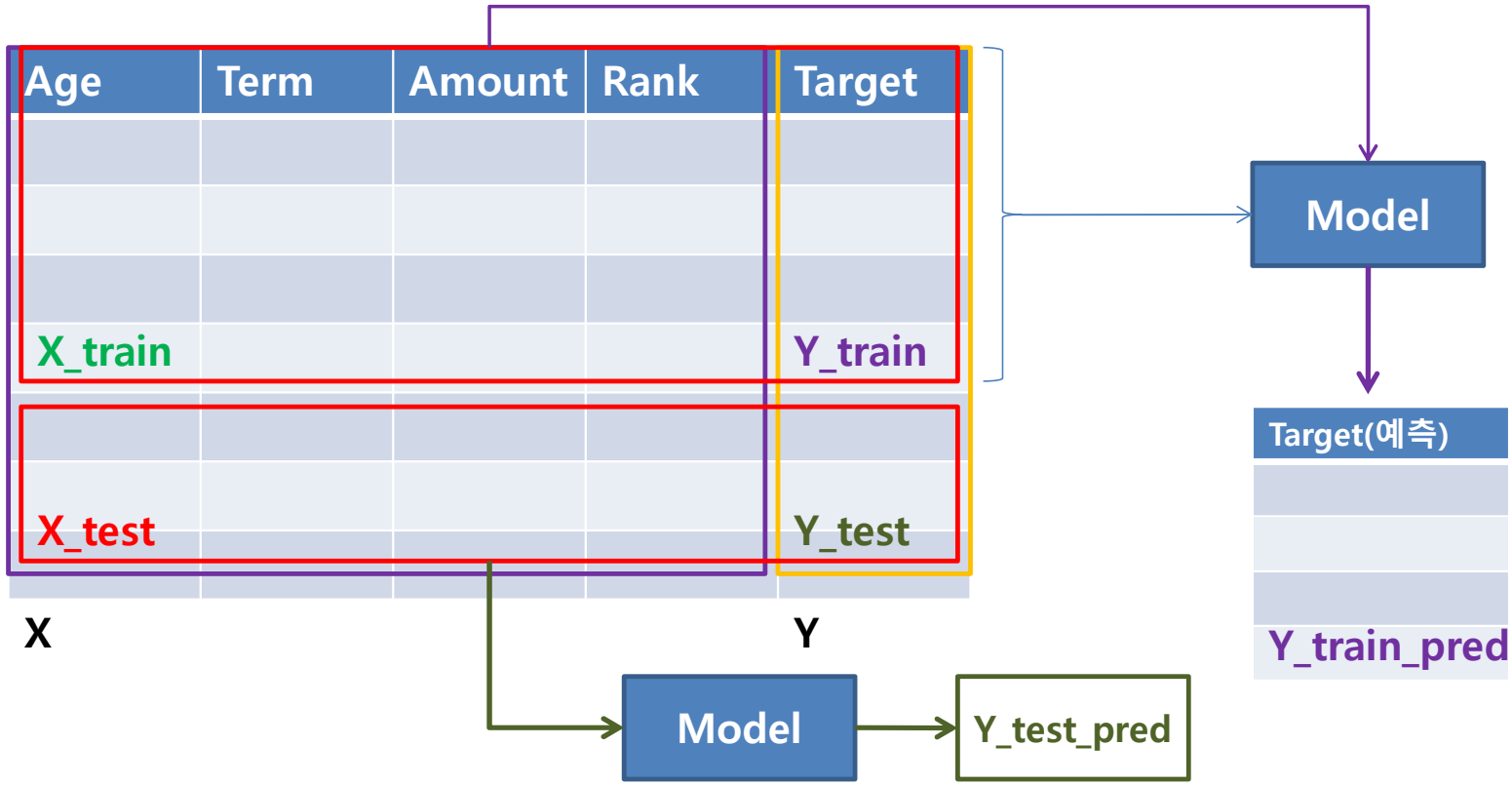
- 수식화 (회귀분석 -> 기계학습(데이터마이닝))
- **기계학습** : 데이터 간 연관성/관계/수식등을 컴퓨터가 학습을 통해 도출해내는 작업

지도학습 (분류모델) 절차

1. 데이터 전처리 (파생변수 / 이상치 처리 / 결측값 처리)
2. 목표변수 (Y) 와 설명변수 (X)를 선언
3. 학습 데이터(Train Set / 학습 능력)와 검증 데이터(Test Set / 일반화 능력)를 분할
4. 학습 진행
   - 특성공학 (Feature Engineering)
   - 알고리즘에 의한 학습
5. 평가
6. 적용

2-1. 기계 학습 개요 (Machine Learning Intro)

- 수식화 (회귀분석 -> 기계학습(데이터마이닝))
- **기계학습** : 데이터 간 연관성/관계/수식등을 컴퓨터가 학습을 통해 도출해내는 작업



- **학습 능력 평가 :** Y_train (책에 있는 정답) – Y_train_pred (공식으로 푼 정답)
- 일반화 능력 평가 : Y_test (새로운 책의 정답) – Y_test_pred (공식으로 푼 정답)

2-2. 기계 학습 평가 (Model Evaluation)

모델 평가 기법 (Evaluation)

1) 분류에서의 평가

$$\text{정확도 Accuracy} = \frac{\text{정확하게 분류한 데이터 수}}{\text{전체 데이터 수}}$$

Confusion Matrix (CM)

	Real Negative	Real Positive
predict Negative	True Negative (TN)	False Negative (FN)
Predict Positive	False Positive (FP)	True Positive (TP)

오차 행렬 Confusion Matrix

$$\text{정밀도 Precision} = \frac{TP}{FP + TP \text{ (Predict Positive)}}$$

$$\text{재현율 Recall} = \frac{TP}{FN + TP \text{ (Real Positive)}}$$

$$\text{F1 Score} = \frac{2 \times (\text{정밀도} \times \text{재현율})}{\text{정밀도} + \text{재현율}}$$

Ex) 불량 제품을 분류하는 시스템 구축

-> 불량판별 데이터 1000명 (학습)

-> 정상 (Negative) : 950 / 불량 (Positive) : 50명

-> Model : 대부분의 제품을 불량으로 분류

정확도 : 950/1000 = 95%

정밀도 : 10 / 20 = 50% // 재현율 = 10/50 = 20%

F1 = 28.5%

	실제 정상	실제 불량
예측 정상	940	40
예측 불량	10	10
	950	50

2-2. 기계 학습 평가 (Model Evaluation)

모델 평가 기법 (Evaluation)

2) 회귀 모델에서 평가

- **R^2 (결정 계수) : 회귀 식 얼마나 데이터를 잘 대변하는가 (0 ~ 1)**

R^2 = 회귀 변동 / 총 변동 = (총 변동 – 오차 변동) / 총변동

총 변동 : $\sum$ (실제값 – 평균)^2 = 회귀 변동 + 오차 변동

회귀 변동 : $\sum$ (예측값 – 평균)^2

오차 변동 : $\sum$ (실제값 – 예측값)^2

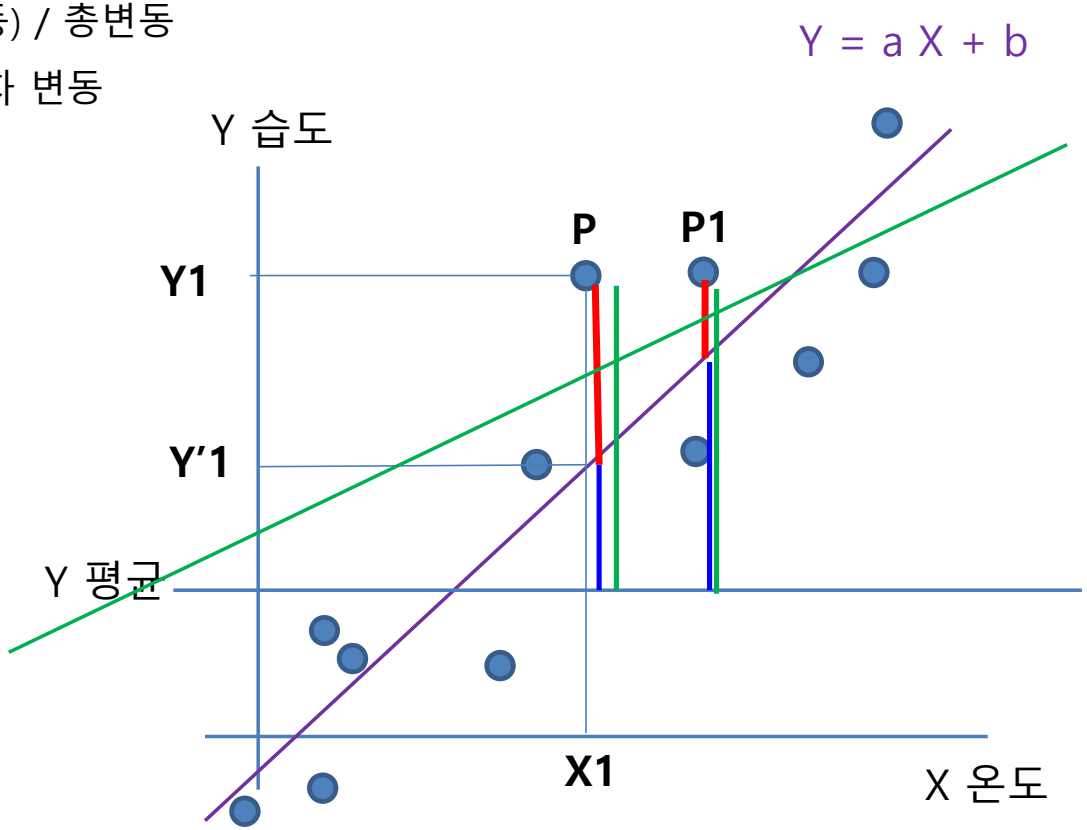
- **Mean Square Error (평균 제곱 오차)**

  $\sum$ (실제값 – 예측값)^2 / (데이터수)

- **Root Mean Square Error**

- **Mean Absolute Error (평균 절대 오차)**

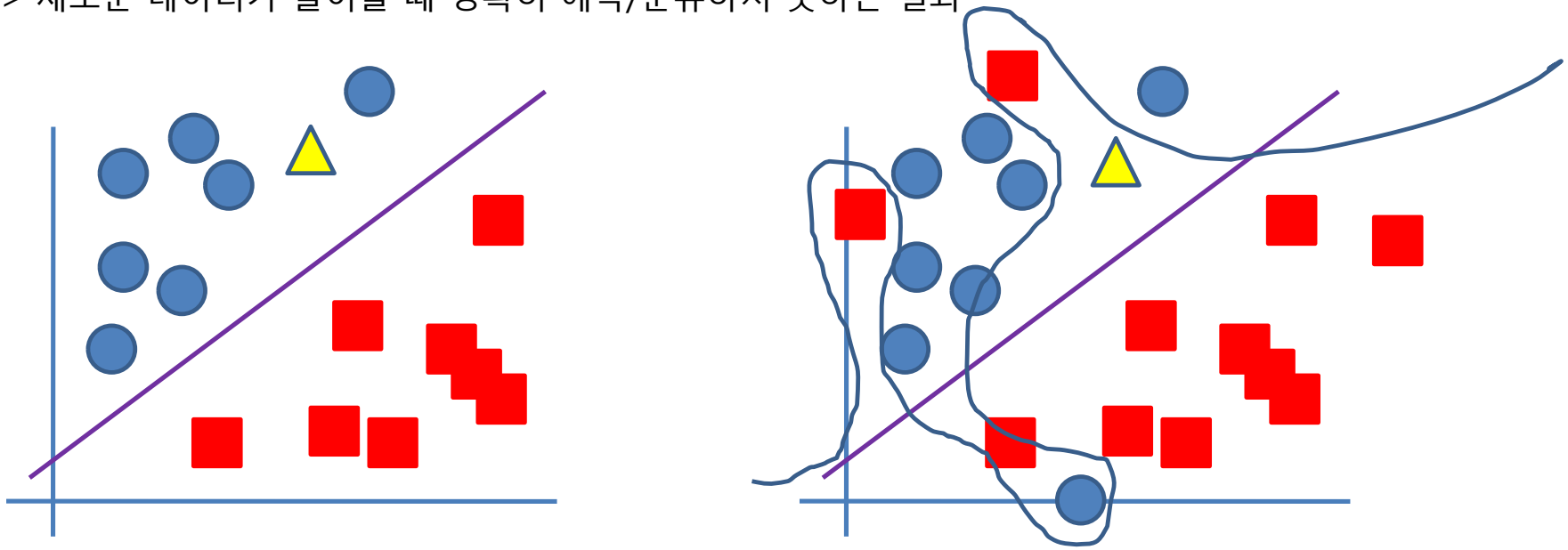
  $\sum$ |실제값 – 예측값 | / (데이터수)



2-2. 기계 학습 평가 (Model Evaluation)

- **과적합 (Overfitting) :** 학습 데이터의 모델에 의한 평가 결과가 검증 데이터의 평가 결과에 비해 매우 높게 나오는 현상

-> 새로운 데이터가 들어올 때 정확히 예측/분류하지 못하는 결과



-> 해결

- 최대한 학습이 잘 되도록 데이터를 깔끔하게 다듬는다. -> **Feature Engineering**
- 알고리즘이 학습과정에서 "적당히" 학습할 수 있도록 통제 -> **최적화 (Hyper Parameter Tuning)**

2-3. 특성공학 (Feature Engineering)

수식화 (회귀분석 -> 기계학습(데이터마이닝))

- **특성공학 (Feature Engineering)**
  학습의 목적에 맞게 데이터를 깔끔하게 다듬는 작업

1. **Scaling & Encoding :** 숫자데이터의 스케일을 맞추거나, 문자데이터를 숫자로 변환하여 학습에 사용

2. **Imputation :** 결측 값 (Missing Value)을 대치하여 학습
   새로운 데이터 올 때 결측 값이 있더라도 예측 또는 분류

3. **Cross Validation :** 학습 데이터를 여러 단계로 나누어 분할하여 학습

4. **Hyper Parameter Tuning :** 알고리즘이 내 수학적 구조나 학습에 발생하는 구조, 함수들을 통제

5. **Imbalanced Data Sampling (분류) :** 데이터의 비율이 깨진 경우, 한쪽의 데이터를 줄이거나, 다른 한쪽의 데이터를 생성하여 비율을 맞춰주는 작업

2-3. 특성공학 (Feature Engineering)

수식화 (회귀분석 -> 기계학습(데이터마이닝))

- **특성공학 (Feature Engineering)**

  학습의 목적에 맞게 데이터를 깔끔하게 다듬는 작업

주소	주소
경기	1
경기	1
제주	2
서울	3
강원	4
경기	1
…	…

1. **Scaling & Encoding :** 숫자데이터의 스케일을 맞추거나, 문자데이터를 숫자로 변환하여 학습에 사용

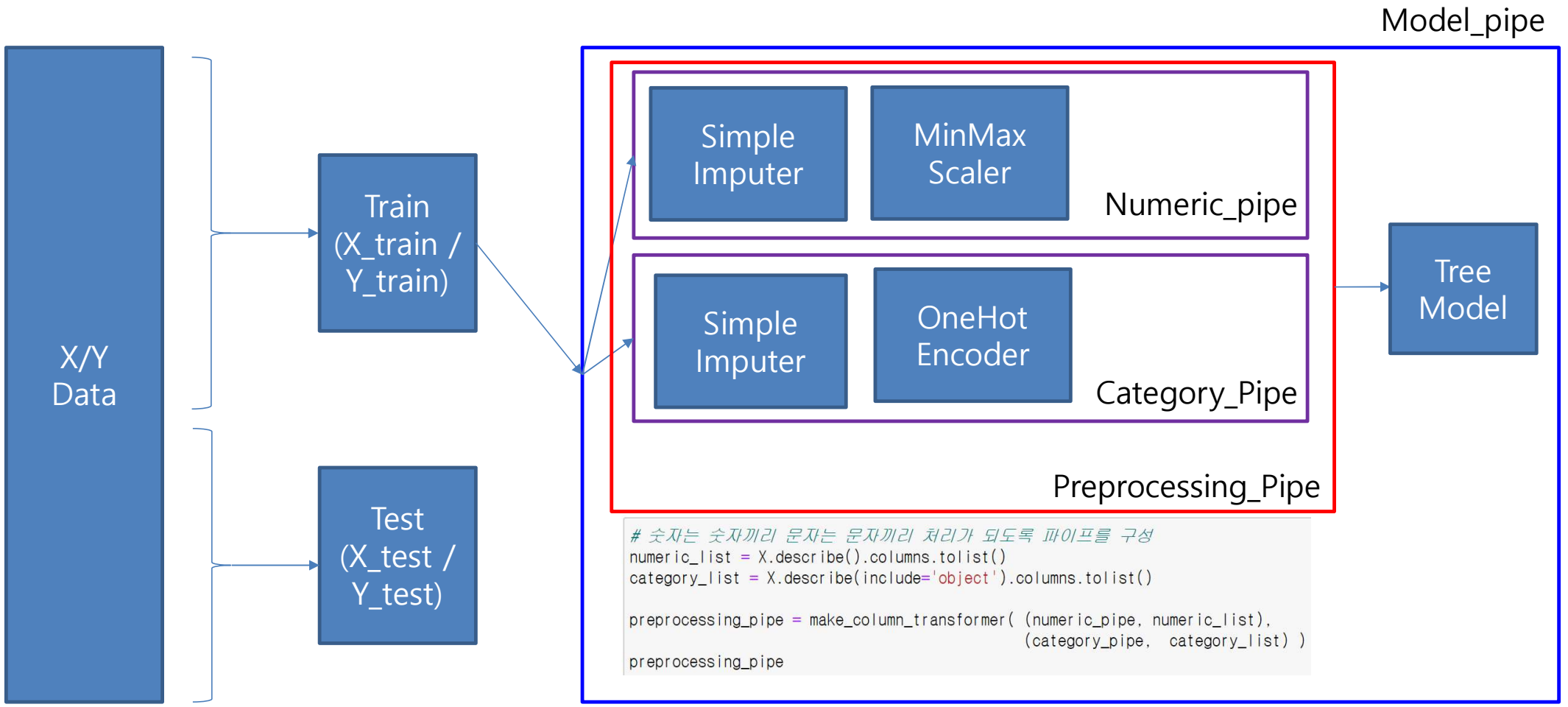
   Scaling : 숫자 데이터의 스케일을 맞춰주는 작업

   - **Standard Scaler :** 평균 0 / 표준편차 1 (선형 회귀 기반 / 연속형 기반의 데이터)
   - **Min Max Scaler :** 최소값 0 / 최대값 1 (범주형 데이터 / 비정형 데이터)
   - **Robust Scaler :** 중앙값 0 / IQR(사분범위) 1

   Encoding : 문자 데이터를 숫자 데이터로 변환

   - Label Encoding : 각 범주형 항목을 정수형태로 변환 (데이터 간 서열이 발생)
   - One Hot Encoding : 각 범주형 항목의 존재여부를 새로운 항목으로 만들어 1,0 값으로 처리 (Dummy)

경기	서울	제주	강원
1	0	0	0
1	0	0	0
0	0	1	0
0	1	0	0
0	0	0	1
1	0	0	0




```
# 숫자는 숫자끼리 문자는 문자끼리 처리가 되도록 파이프를 구성
numeric_list = X.describe().columns.tolist()
category_list = X.describe(include='object').columns.tolist()

preprocessing_pipe = make_column_transformer( (numeric_pipe, numeric_list),
                                              (category_pipe,  category_list) )
preprocessing_pipe
```

```
X_train, X_test, Y_train, Y_test = train_test_split( X,Y, test_size=0.3,
                                                     random_state=1234 )
```

2-3. 특성공학 (Feature Engineering)

- 수식화 (회귀분석 -> 기계학습(데이터마이닝))
- **특성공학 (Feature Engineering)**

2. Imputation

- 결측값을 다른 값으로 대치하여 데이터 공백 없이 학습을 수행
- **결측값 (Missing Value) :** NaN / NA / None / 공백 ...
- **제거 : df1.dropna( )**
- **대치 : df1['col'].fillna( )**
- **보간 : df1['col'].interploate(' ')**

- **from sklearn.impute**
  - SimpleImputer : 단순 특정 통계량 / 원하는 값으로 대치
  - KNNImputer : KNN(K 최근접 이웃)알고리즘을 이용해 결측치를 대치

2-3. 특성공학 (Feature Engineering)

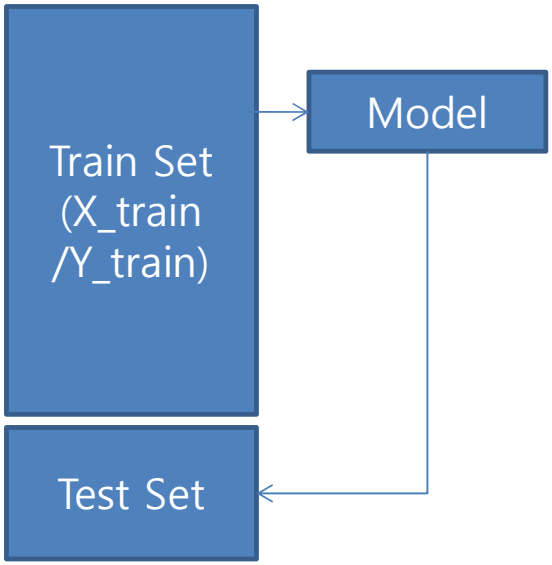
- 수식화 (회귀분석 -> 기계학습(데이터마이닝))
- **특성공학 (Feature Engineering)**

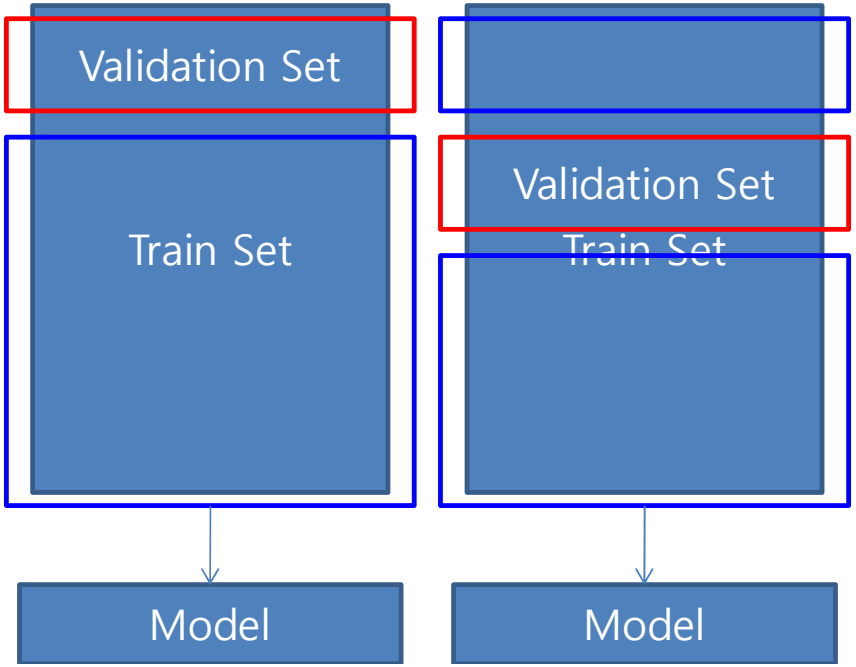
3. Cross Validation (교차검증)

- K – Fold : 특정 K 개수 만큼 데이터를 나누어 교차로 모델을 생성해 학습

- Stratified K Fold : (분류) 특정 클래스의 비율을 유지하며, K 개수 만큼 교차로 모델을 생성해 학습

- 데이터에서 특정 개수 만큼 검증 데이터를 나누어 교차로 모델을 생성해 학습
- Train Set - > 학습 / 학습 능력 평가 (Train / Validation Set)
- Test Set -> 일반화 성능 평가





2-3. 특성공학 (Feature Engineering)

- 수식화 (회귀분석 -> 기계학습(데이터마이닝))
- **특성공학 (Feature Engineering)**

4. Hyper Parameter Tuning (매개변수튜닝)

- Hyper Parameter : 알고리즘이 학습을 수행하면서 갖는 수학적인 구조
  알고리즘 내 구성되어 있는 구조들

- Random Search : 알고리즘 내 있는 파라미터 값을 무작위로 부여하여 가장 적절한 파라미터를 찾는 기법
- Grid Search **:** 사용자들이 파라미터 값의 범위를 지정하여, 특정 범위에 대해서 파라미터를 부여하여 적절한 모델을 찾는 기법

2-3. 특성공학 (Feature Engineering)

- 수식화 (회귀분석 -> 기계학습(데이터마이닝))
- **특성공학 (Feature Engineering)**

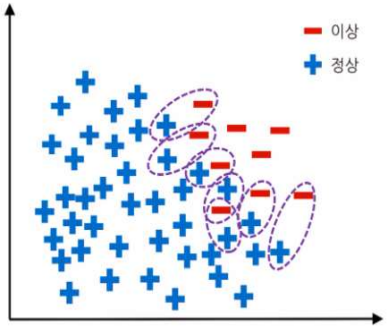
5. Imbalanced Data Sampling

- 비율이 다른 범주형 Y에 대해 분류를 진행 할 때, 데이터의 비율을 맞춰주는 작업
- 학습하는 과정에서 학습데이터에 적용하여 사용 -> Pipe Line (Sklearn)

1) Under Sampling : 데이터의 비율이 적은 쪽으로 데이터를 맞춰주는 작업

- Random Under Sampling : 데이터의 비율이 많은 쪽에 값을 무작위로 줄이는 방법
- Tomek's Link : 인접합 Class(서로 다른 범주형 항목)들을 묶어, 밀집되어 있는 부분의 데이터를 제거
- CNN (Condensed Nearest Neighbor) :

  비율이 많은 쪽에 데이터에서 밀도가 높은 부분의 데이터를 제거
- One Sided Selection : Tomek' Line + CNN



◀ **Tomek Link 기법**

인접한 데이터 대해서는 Tomek Link 기법으로 데이터를 지우고,
밀집된 다수의 데이터클래스에 데이터를 삭제

2-3. 특성공학 (Feature Engineering)

- 수식화 (회귀분석 -> 기계학습(데이터마이닝))
- **특성공학 (Feature Engineering)**

5. Imbalanced Data Sampling

- 비율이 다른 범주형 Y에 대해 분류를 진행 할 때, 데이터의 비율을 맞춰주는 작업
- 학습하는 과정에서 학습데이터에 적용하여 사용 -> Pipe Line (Sklearn)

2) Over Sampling : 데이터의 비율이 큰 쪽으로 데이터를 생성

- Random Over Sampling : 무작위하게 데이터의 비율이 많은 수에 맞춰 생성
- SMOTE (Synthetic Minority Over Sampling Technique) :

  KNN 기법을 활용해, 비율이 적은 쪽 데이터를 K개의 인접데이터 수 만큼 묶어,
  묶인 데이터 내 중심위치를 찾아 새로운 데이터를 생성
- ADASYN (Adaptive Synthetic Sampling) :

  SMOTE 기법을 이용해 데이터를 생성하는 단계에서
  임의의 작은 값을(감마 확률 밀도 함수) 더하여 사실적인 데이터로 생성

2-4. 학습 알고리즘

- **수식화** (회귀분석 -> 기계학습(데이터마이닝))
- **최적화 (Optimization) :** 문제 상황에서 여러 해결방안 중 가장 최적의 해결방안을 찾는 방법
- 기본적으로 머신러닝(기계학습)에서 데이터로부터 학습을 통해 수식화를 진행 할 때, 목표 값을 잘 예측/분류하는 수식을 만드는 과정에서 사용 **-> Best 수식을 찾자!**

- **수학적 접근 :**

  특정 함수의 값을 **최소화(또는 최대화)**시키는 최적의 수식 값(최적의 파라미터)의 조합을 찾는 문제

  최적화 문제는 기본적으로 최소화(Minimization)와 최대화(Maximization)으로 나눠 볼 수 있다.

  - 최소화(Minimization) : 함수(수식)의 **목표변수(Y, Label, Output)**를 **최소가** 되게끔 파라미터(계수와 절편, Weight 가중치)값을 찾는 문제

    -> **오류 / 오차 / 비용 / 손실** ...

  - 최대화(Maximization) : 함수(수식)의 **목표변수(Y, Label, Output)**를 **최대가** 되게끔 파라미터(계수와 절편, Weight 가중치)값을 찾는 문제

    -> **이윤 / 점수 / ...**

Y(습도) = 100 * X(온도) + 20

2-4. 학습 알고리즘

- **수식화** (회귀분석 -> 기계학습(데이터마이닝))
- **최적화 (Optimization) :** 문제 상황에서 여러 해결방안 중 가장 최적의 해결방안을 찾는 방법
- 기본적으로 머신러닝(기계학습)에서 데이터로부터 학습을 통해 수식화를 진행 할 때, 목표 값을 잘 예측/분류하는 수식을 만드는 과정에서 사용 **-> Best 수식을 찾자!**

- **데이터 분석 :**
  - 머신 러닝에서 **Model(수식, 함수)**를 구축하여 분류/예측을 진행 할 때
  - Model을 생성 할 적절한 알고리즘을 선택하는 문제

  - Ex) 공장 / Y 온도 <-> X1(두께) / X2(습도) / X3(강도)

  1) 어떤 알고리즘으로 학습을 수행 해야할지 결정 (회귀분석 / 결정나무 / 앙상블 / 신경망 ...)
  2) 해당 알고리즘으로 도출해 낼 수 있는 가장 적절한 Model 수식 함수를 도출
  3) Y 값을 가장 낮게끔 하는 X를 도출

2-4. 학습 알고리즘

- **수식화** (회귀분석 -> 기계학습(데이터마이닝))
- **최적화 (Optimization) :** 문제 상황에서 여러 해결방안 중 가장 최적의 해결방안을 찾는 방법
- 기본적으로 머신러닝(기계학습)에서 데이터로부터 학습을 통해 수식화를 진행 할 때, 목표 값을 잘 예측/분류하는 수식을 만드는 과정에서 사용 **-> Best 수식을 찾자!**

- **데이터 분석 :**
  - Model을 생성 할 적절한 알고리즘을 선택하는 문제 (Y = **100** X **+ 20** **)**

최적화의 다양한 기법

- **Least Square Method (최소제곱법)**
- **Gradient Decent Method (경사하강법)**
- Newton's Method
- Gauss Newton's Method
- Bayesian Method
- Markov Bayesian Method

2-4. 학습 알고리즘

- **수식화** (회귀분석 -> 기계학습(데이터마이닝))
- **Least Square Method (최소 제곱 법, OLS)**
- 데이터를 대표하는 회귀선을 찾을 때, 회귀선이 예측한 예측값과 실제 데이터에 있는 실제값의 차이(잔차, Residual, Error) 제곱의 합 또는 평균이 최소화(Minimization)되는 방향으로 파라미터를 결정하는 방법

- 실제 값 : 데이터로 부터 수집된 Y (Y_train, Y_test)
- 예측 값 : 모델에의해 계산된 값 Y′ (Y_train_pred, Y_test_pred)
- 잔차 : $Y - Y'$
- RSS (Residual Sum of Square) $= \sum (Y - Y')^2$ -> 최소화 0

$= \sum (Y - (\mathbf{a} X + \mathbf{b}))^2$

-> $Y' = a X + b$

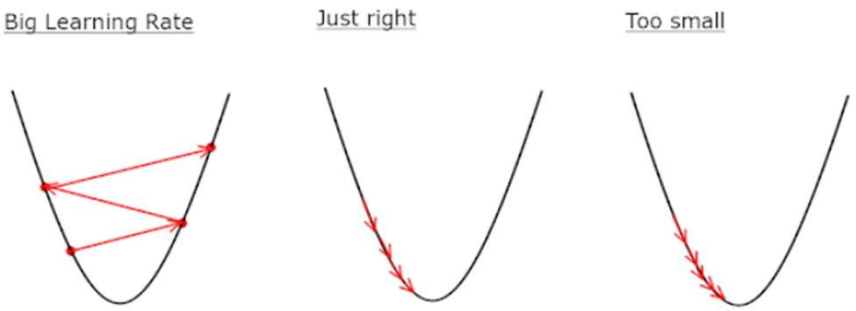
- 대수적 기법 : 오차항이 최소가 되는 지점에서의 a,b를 찾기 위해, 오차항을 각각의 a,b에 대해 편미분 연립방정식을 통해 a,b를 찾는 방법

  Minimum $= \sum (Y - Y')^2$ -> 미분 값 0
- **해석학적 기법 : 오차항을 행렬로 표현하여 행렬의 유사역행렬(Pseudo Inverse)을 이용해 계산**

2-4. 학습 알고리즘

- **수식화** (회귀분석 -> 기계학습(데이터마이닝))
- **Gradient Descent Method (경사 하강 법)**
- **점진적인 반복 계산을** 통해 함수가 최소가 되는 파라미터를 찾아주는 방법
- **그래디언트(Gradient) :** 함수가 증가하는 방향과 크기를 표현 (점진적인 반복)
- Y = a X + b / 적절한 a, b 찾는 것
- 초기에 임의의 a,b값을 세팅 -> Gradient 가 감소하는 방향으로 점진적 반복계산을 통해 a,b를 갱신
- 실제값 – 예측값 = 최소 가 될 때 까지 반복하여 계산

- **Learning Rate (학습율) :** 얼마나 점진적으로 파라미터(a,b)를 변화해 가면서 계산할 지 곱해주는 값
- **적절한 학습율을 찾아야만 실제값과 예측값이 최소가 되는 a,b 도출**



2-4. 학습 알고리즘

- **수식화** (회귀분석 -> 기계학습(데이터마이닝))
- **회귀 알고리즘**
- **Linear Regression (선형 회귀 알고리즘) : Y= A X + B** / 최소 제곱 법 + 경사 하강 법
  - 회귀 : 특정 객체의 값이 집단의 평균과 같은 일정한 값으로 돌아가려는 경향
  - 회귀 계수 (Regression Coefficient) 찾는 것이 목표 (A, B)

    A -> 회귀 계수 (Coefficient) / B -> 절편 (intercept) -> 가중치 Weight

- 일반 선형 회귀 (Ordinary Linear Regression) : 예측값과 실제값의 차이인 RSS 가 최소가 되는 방향으로 가중치를 찾는 기법 (최소 제곱 법 + 경사 하강 법)
- 규제선형회귀 (Regularization Regression) : 과적합(Overfitting) 현상을 방지하여 **규제 항**을 추가해 가중치를 찾는 기법
  - Ridge : 특정 항목의 Weight값을 적절하게 낮추어 규제 (상대적으로 큰 회귀 계수를 통제)
  - Lasso : 예측 영향력이 적은 Weight을 제거하여 규제 (변수 선택법 / Column이 매우 많은 데이터)
  - Elastic Net : 상대적으로 큰 회귀계수는 낮추고, 영향력이 매우 적은 항목은 제거

Y습도 = 100000 x 온도 + 200 x 재료두께 + 150 x 강도 + 0.0001 x 회전수 + ..

2-5. 회귀 알고리즘 (Regression Model)

- **회귀 알고리즘 종류**

1) **Linear Regression :** 연속형 / 정규분포 / 균일하고 일정한 연속형

2) **Regularization Regression :** 연속형 / 정규분포 / 균일하고 일정한 연속형

3) **Decision Tree Regreesor :** 분포 모양에 제약 X / 연속형-범주형 / **과적합**

4) **Bagging Ensemble** (Random Forest)

   - Ensemble : 여러가지 알고리즘을 결합 모델

5) **Boosting Ensemble** (GB / AdaB / XGB / LightGB)

6) **Support Vector Machine Regressor (SVM, SVR) :** 항목 수 많으나, 데이터 개수가 적은 경우 / 연속형

2-5. 분류 알고리즘 (Classification Model)

- **분류 알고리즘 종류**

1) **Decision Tree :** 분포 모양에 제약 X / 연속형-범주형 / **과적합**

2) **Bagging Ensemble** (Random Forest)

3) **Boosting Ensemble** (GB / AdaB / XGB / LightGB)

4) **Support Vector Machine (SVM)**

5) **KNN :** 분류 / 밀도 기반 / 학습 X

6) **Voting**

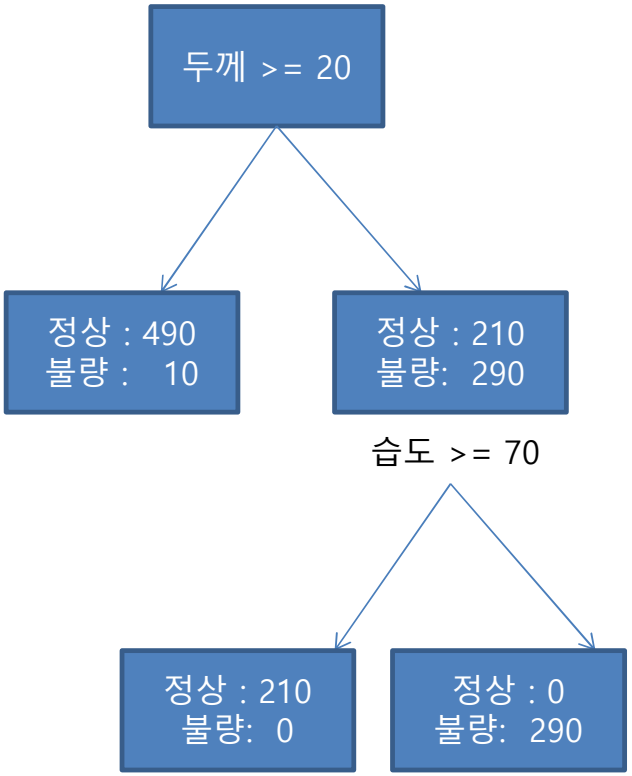
Etc) Naïve Bayes / Logistic Regression / DNN ...

2-5. 분류 알고리즘 (Classification Model)

- **Decision Tree Model (의사결정나무모델)**

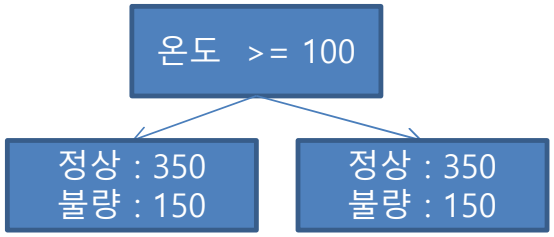
설명변수(X)들의 규칙,관계,패턴을 파악해 목표변수를 분류하는 나무 구조의 모델을 생성

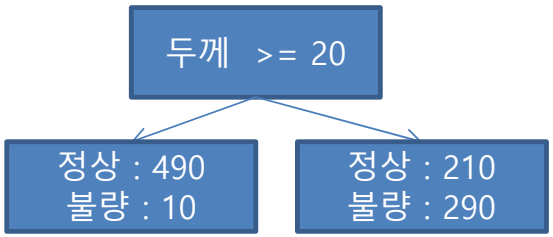
- **장점 :**
  - 대용량 데이터 잘 작동
  - 데이터 전처리에 대한 영향이 적다
  (이상치가 많거나 , 비모수적 데이터에 대해 잘 작동)
  - 결과 해석이 쉽다
- **단점 :**
  - 과적합 (Overfitting) 잘 발생
  - 분류경계에서 오류가 발생할 가능성이 높다



Ex) 공정 불량 여부 (Y) <-> 온도/습도/두께/ ... (X)

정상 700 / 불량 300 -> 1000개





2-5. 분류 알고리즘 (Classification Model)

- **Decision Tree Model (의사결정나무모델)**

설명변수(X)들의 규칙,관계,패턴을 파악해 목표변수를 분류하는 나무 구조의 모델을 생성

- **분류 기준 :**

gini : 불순도 지표를 계산하여 해당 값이 낮아지는 방향으로 학습

entropy : 질서정연한 방향으로 학습 (깔끔히 데이터를 분류할 수 있는 방향)

- **Tree 구조 :**

Node : 데이터가 나눠지는 기점

Root Node : 나무구조가 시작되는 기점

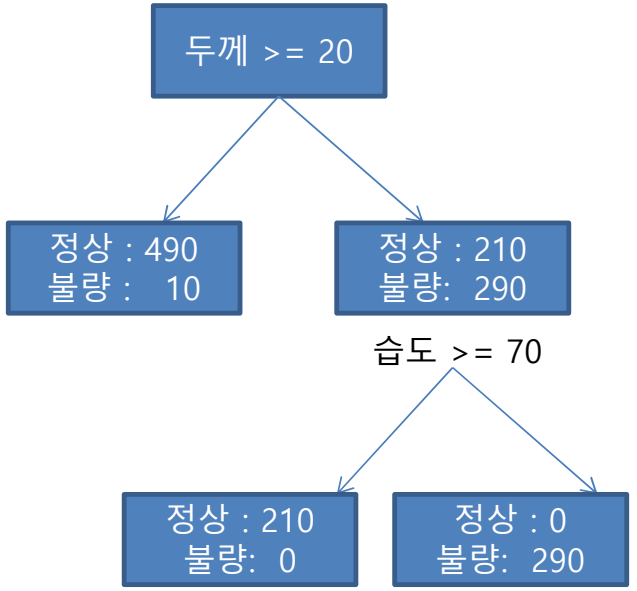
Child Node : 상위 노드로부터 파생되어 나온 기점

Parent Node : 자식노도의 상위 노드

Depth : 노드가 분할되어 내려가며 발생한 나무 층

Leaf (Terminal Node) : 가장 끝부분에 위치한 기점

Branch : 최상위 기점부터 맨 끝 기점까지 연결된 일련의 노드

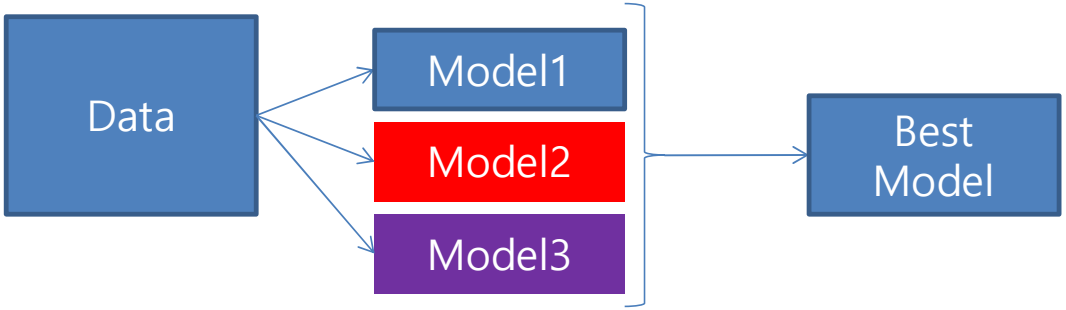


2-5. 분류 알고리즘 (Classification Model)

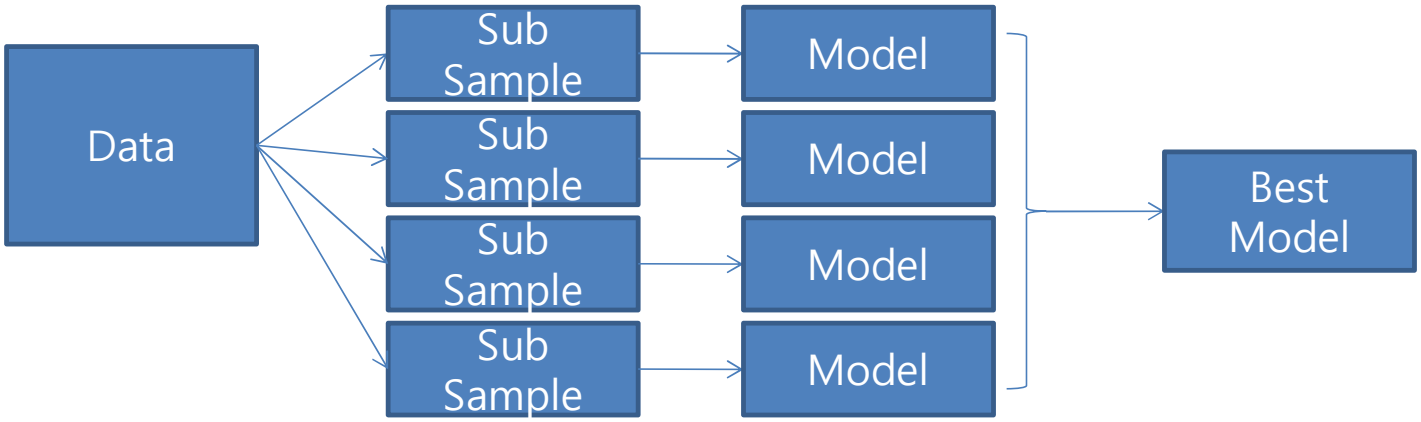
- **Ensemble Model**

기존에 존재하는 여러 가지 알고리즘을 결합 또는 조합하여 새로운 강력한 하나의 모델 생성

1) Voting Model : 서로 다른 알고리즘을 결합하는 방법 / 가장 좋은 성능의 알고리즘을 찾아 사용



2) Bagging Model : 학습 데이터에서 서로 다른 복원추출(Sub Sample)된 데이터를 학습하여 결합/투표하여 학습하는 방법 (Random Forest)

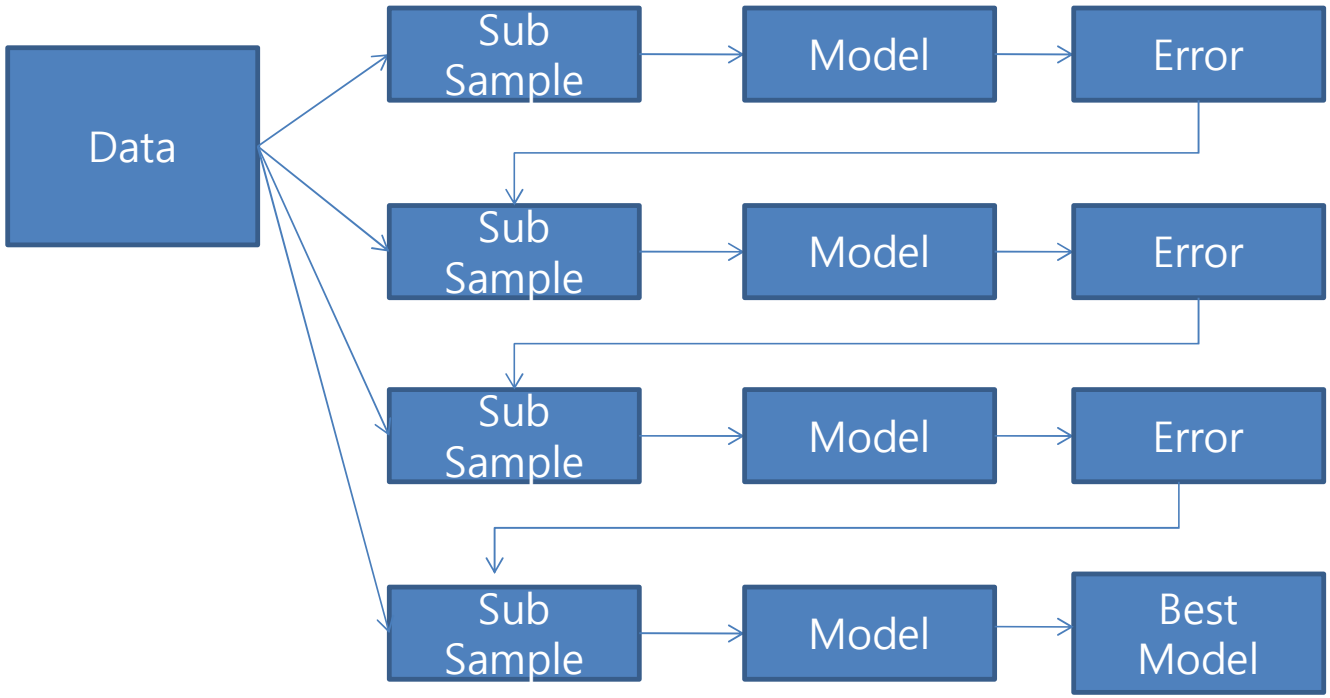


2-5. 분류 알고리즘 (Classification Model)

- **Ensemble Model**

기존에 존재하는 여러 가지 알고리즘을 결합 또는 조합하여 새로운 강력한 하나의 모델 생성

3) Boosting Model : 알고리즘을 구성할 때 마다 오차를 줄이는 방향으로 복원추출학습을 수행하여 모델을 향상시키는 알고리즘 (Gradient Boosting / Ada / CatBoost …)



2-5. 분류 알고리즘 (Classification Model)

- **Ensemble Model**

기존에 존재하는 여러 가지 알고리즘을 결합 또는 조합하여 새로운 강력한 하나의 모델 생성

3) Boosting Model : 알고리즘을 구성할 때 마다 오차를 줄이는 방향으로 복원추출학습을 수행하여 모델을 향상시키는 알고리즘 (Gradient Boosting / Ada / CatBoost …)

1. Ada Boosting (Adaptive Boosting) : 이전에 학습의 결과에서 잘 반영되지 않은 데이터에 대해 Weight(가중치)를 부여하여 복원 추출을 수행

- 앞서 분류를 적절히 수행하지 못한 데이터에 대해 계속 가중을 주며 학습의 성능을 향상
- 데이터에 이상치가 존재하는 경우, 이상치에 대해 계속 Weight(가중치)를 부여하여 모델이 구성 -> 사전에 구성한 데이터가 깔끔하지 않다면, 지속적 학습에 대해 성능이 개선되지 않을 가능성이 높다.
- 앞서 분류된 결과에 높은 Weight가 부여되는 경우엔 Weight가 낮게 부여된 데이터에 대해 오분류 할 가능성이 높다

2-5. 분류 알고리즘 (Classification Model)

- **Ensemble Model**

기존에 존재하는 여러 가지 알고리즘을 결합 또는 조합하여 새로운 강력한 하나의 모델 생성

3) Boosting Model : 알고리즘을 구성할 때 마다 오차를 줄이는 방향으로 복원추출학습을 수행하여 모델을 향상시키는 알고리즘 (Gradient Boosting / Ada / CatBoost …)

2. Gradient Boosting : Ada boosting 모델에서 이상치가 있거나 너무 높게 부여된 Weight에 대해 주변데이터가 오분류될 가능성을 극복하고자 분류 결과에 Weight(가중치)를 부여할 때 마다 Weight에 의한 모델의 오차가 최소가되는 방향으로 (Gradient Descent, 경사하강법)을 이용하여 오분류 값을 최소화하는 방식으로 학습

- 앞서 사용한 Ada Boosting 모델보다 오차에 대해 더욱 민감한 모델을 구성
- 복원추출된 데이터를 학습한 Model에 오차를 수식으로 계산하여 갱신 -> 시간이 더 많이 소요
- 순차적으로 Model이 업데이트 되기 때문에 나중에 학습된 모델에 대해서는 과적합이 발생가능이 높다

2-5. 분류 알고리즘 (Classification Model)

- **Ensemble Model**

기존에 존재하는 여러 가지 알고리즘을 결합 또는 조합하여 새로운 강력한 하나의 모델 생성

3) Boosting Model : 알고리즘을 구성할 때 마다 오차를 줄이는 방향으로 복원추출학습을 수행하여 모델을 향상시키는 알고리즘 (Gradient Boosting / Ada / CatBoost …)

3. XGboosting : Gradient Boosting 모델에서 발생하는 Overfitting 과적합 현상을 방지하기 위해, 규제항(Regularization) 을 추가하여 학습을 수행

- 복원 추출한 데이터의 오차가 줄어들게 끔 학습을 하다보면 각 Tree 모델에 대해 분류 구조가 복잡해 질 수 있음
- 규제항을 추가하여 과적합 방지
- 오차를 계산하는 함수를 다양하게 적용 -> 하이퍼 파라미터 튜닝

4. Light Boosting : 복원추출된 데이터의 양을 조절하는 알고리즘을 활용해 시간과 자원의 소요를 효과적으로 줄여줌

- 대용량 데이터에 대해 절약된 시간과 자원으로 학습이 가능
- 복원 추출된 데이터를 근사치를 이용해 데이터의 양을 조절
- 자원 절약과 동시에 Overfitting 해결 (하이퍼 파라미터)

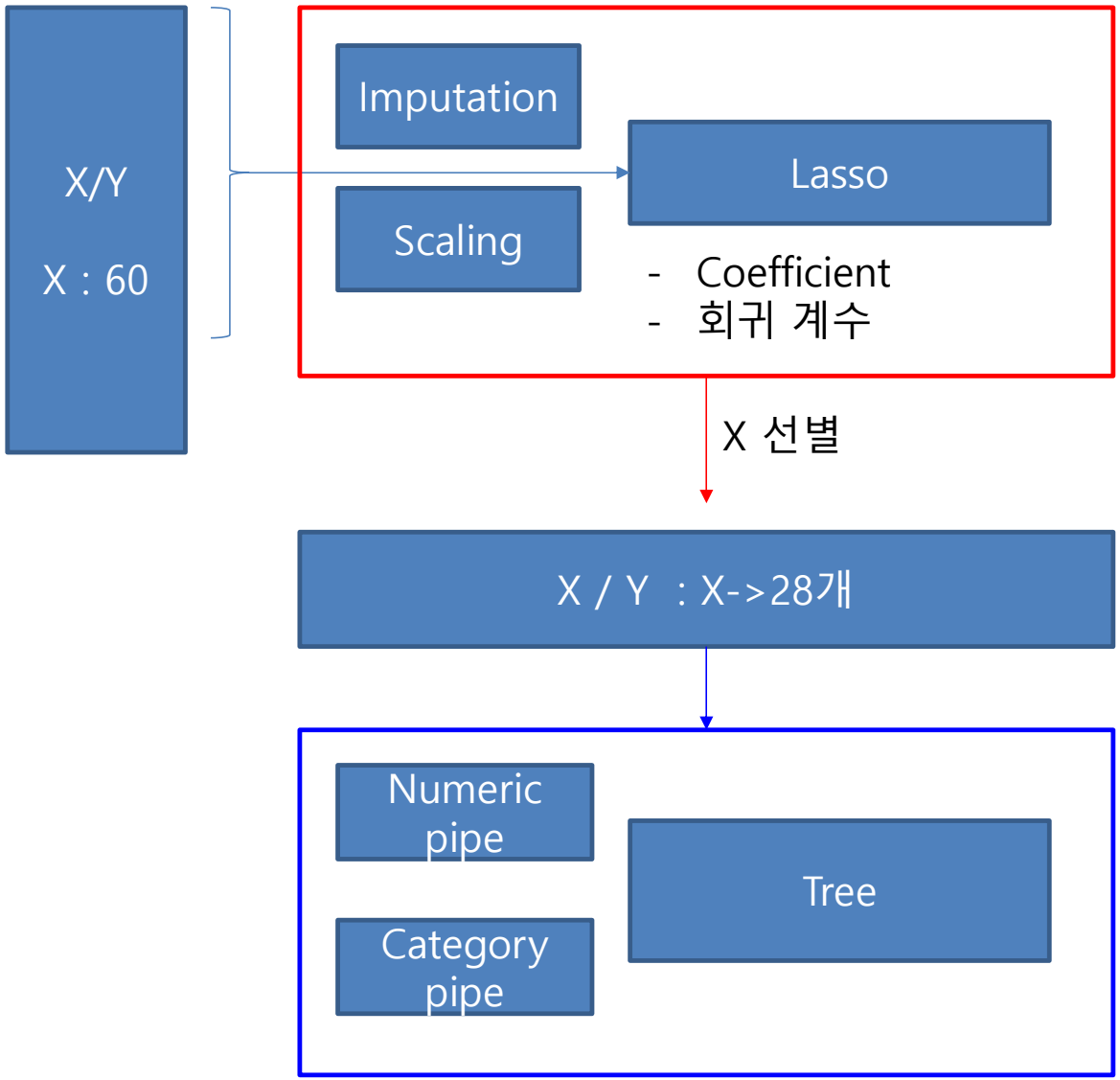
2-5. 분류 알고리즘 (Classification Model)

- **Ensemble Model**

기존에 존재하는 여러 가지 알고리즘을 결합 또는 조합하여 새로운 강력한 하나의 모델 생성

3) Boosting Model : 알고리즘을 구성할 때 마다 오차를 줄이는 방향으로 복원추출학습을 수행하여 모델을 향상시키는 알고리즘 (Gradient Boosting / Ada / CatBoost …)

5. CatBoosting (Categorical Boosting) : 범주형 데이터를 One Hot Encoding을 통해 변수를 구성하여 처리, 이 One Hot 값을 잘 학습 시켜주기 위해 X데이터들에 대해 clustering 작업을 별도로 수행하며 복원추출 및 학습을 수행


X/Y
X : 60
Imputation
Scaling
Lasso
- Coefficient
- 회귀 계수
X 선별
X / Y : X->28개
Numeric pipe
Category pipe
Tree

2-5. 회귀 알고리즘 (Regression Model)

- **회귀 알고리즘**

- **Linear Regression**

회귀 : 특정 값이 집단의 평균과 같은 일정한 값으로 돌아가려는 경향

여러 개의 설명변수(X)와 하나의 목표변수(Y, 연속형)간 상관 관계를 모델링 하는 기법

- 적절한 Y를 예측하기 위한 X앞의 **회귀계수(Regression Coefficient)**를 찾는 것

$$\mathbf{Y' = a\,X1 + b} \;\rightarrow\; \text{minimum} \sum (Y - Y')^2$$

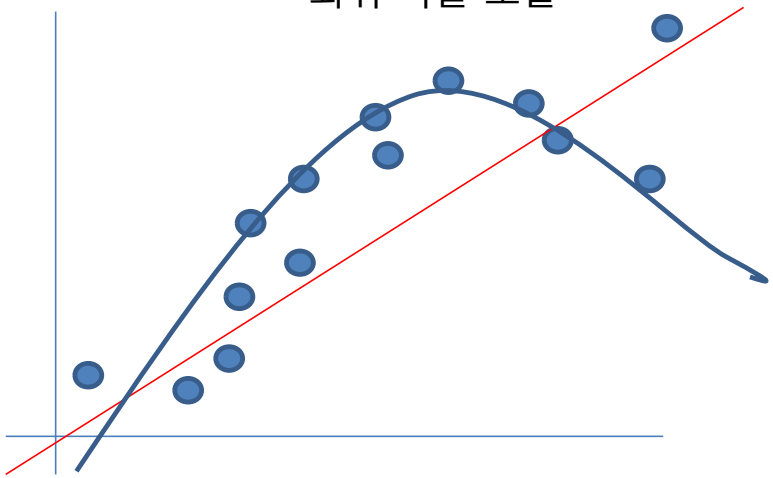
- 학습에 있어서 차수가 높아지면 Overfitting -> 규제 (Regularization)
- 최소제곱법을 이용해 회귀계수를 계산 (전통 통계)
- 컴퓨터 알고리즘으로는 **경사하강법을 이용해** 회귀 계수(Weight)를 계산

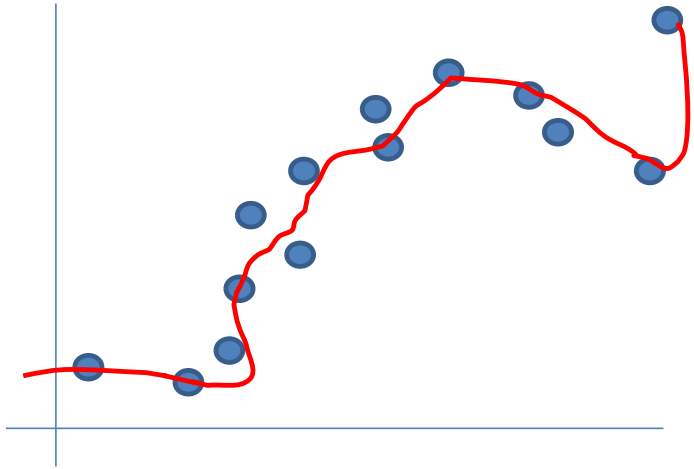
2-5. 회귀 알고리즘 (Regression Model)

- **회귀 알고리즘**

- **Regularization Regression**

규제 선형 회귀 : 선형 회귀 모델에 규제를 추가하여 회귀 계수를 조절해 과적합을 방지하여 회귀 식을 도출





Lasso 회귀 : 예측 영향력이 적은 X (Feature)에 대해 회귀 계수를 0으로 만들어 예측 시 해당 X가 선택되지 않게 규제 **(변수선택법)** Y = 0.00000023 X1 + 123 X2 + 104 X3 + …

Ridge 회귀 : 상대적으로 큰 회귀계수를 조절하여, 회귀계수 값을 줄여, 규제하는 방법 **(Overfitting 방지)**

Y = 20000 X1 + 123 X2 + 104 X3 + …

Elastic Net 회귀 : X 설명변수가 많은 데이터 셋에 대해, 영향이 적은 설명변수를 줄이며, 회귀 계수 값 통제

2-5. 회귀 알고리즘 (Regression Model)

- **Time Series (시계열 분석)**
- 예측의 여러 기법

1. 정성적 기법 (주관적 예측법)

- 시장조사법 : 제품의 서비스 출시 전 소비자의 의견조사/시장의견조사를 통해 수요를 예측하는 방법

- 한정된 표본 / 통계적인 지표들을 이용해 단기 예측 (기술통계량)
- 거시적 환경 분석 / 산업 환경 분석 / 기업 환경 분석 / 소비자 환경 분석

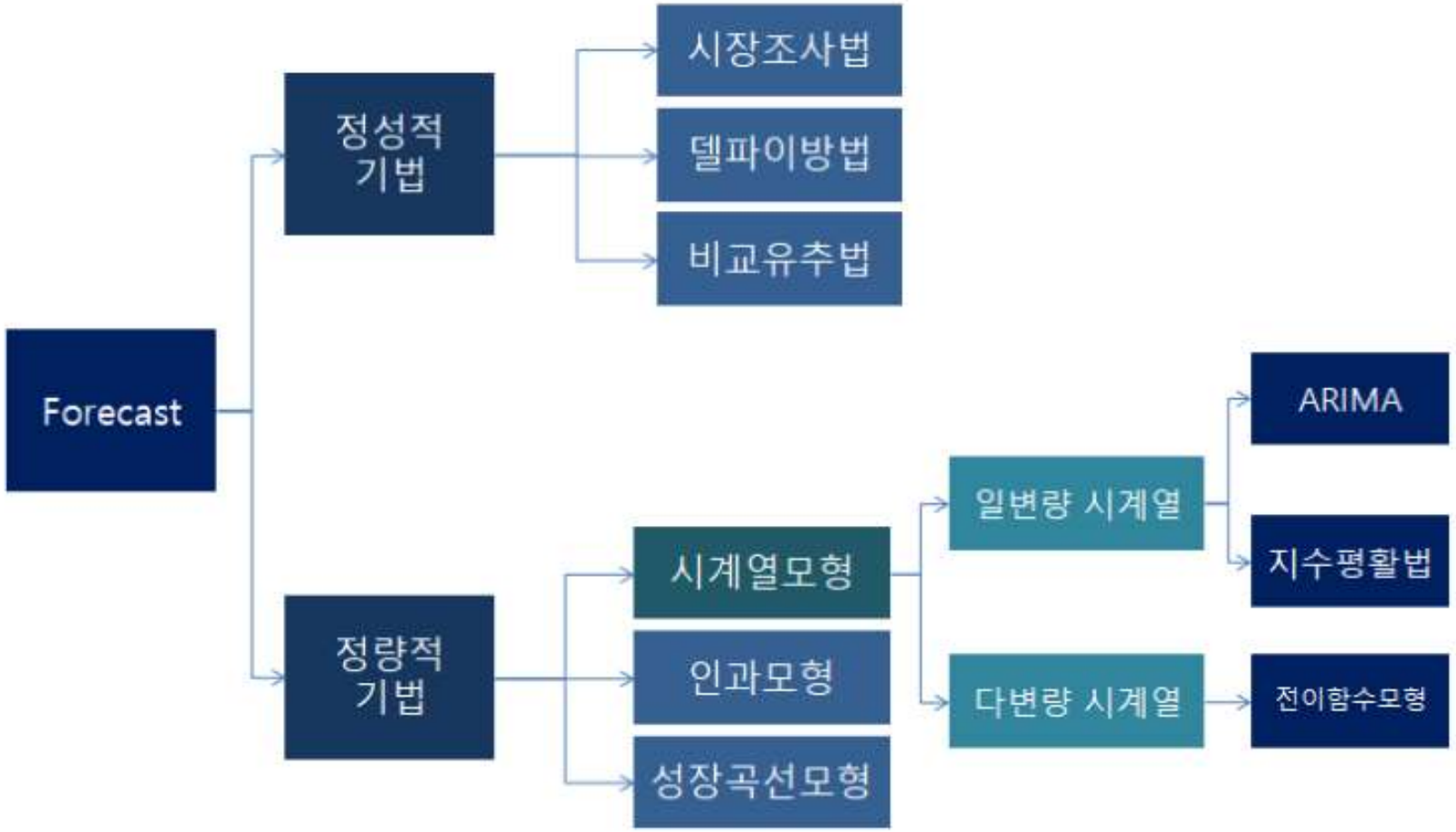
- 델파이법 : 수요예측을 전문가의 직관을 이용해 수행
  - 기존의 자료가 없는 경우 / 매우 주관적 / 장기 예측
- 비교유추법 : 수요예측을 전문가의 직관을 이용해 수행
  - 기존의 자료를 바탕으로 판단 / 단기 예측

2. 정량적 기법 (객관적 예측법)

- 시계열 모형
- 인과 모형 : 목표 변수에(수요량 / 판매량) 대한 직,간접적인 영향을 미치는 설명변수를 파악하는 모형 (회귀분석 -> 회귀계수 / 시각화 / 가설검정)
- 성장 곡선 모형 : 특정 시간이 지남에 따라 특정 변수의 변화량을 측정한 데이터의 패턴을 분석하는 모형

2-5. 회귀 알고리즘 (Regression Model)

- **Time Series (시계열 분석)**



2-5. 회귀 알고리즘 (Regression Model)

- **Time Series (시계열 분석)**
- 시간에 따른 연속형 변수의 예측 및 Trend 파악
- 특정 시간 간격을 가진 주기 : Lag
- 각 Lag에서 Data Point를 찾는다

- **시계열 패턴**

1. **추세 (Trend) :** 데이터가 장기적으로 증가하거나 감소할 때 발생하는 일정한 패턴
   (전체적인 데이터에 걸쳐서 발생)
2. **계절성 (Seasonality) :** 1년 중 특정한 때, 주중 특정 요일에 발생하는 특정 요인이 시계열에 영향을 줄 때 발생하는 패턴
3. **주기 (Cycle) :** 빈도가 정해지지 않은 형태로 데이터가 증가하거나 감소할 때 발생하는 패턴
4. **잡음 (Noise) :** 시간에 따라 독립적인 데이터
   White Noise : 통계적, 기술적 분석이 가능한 정도의 Noise

2-6. System 구축

- **System 구축**
- **Dash Library를 활용한 Web Application 구성**

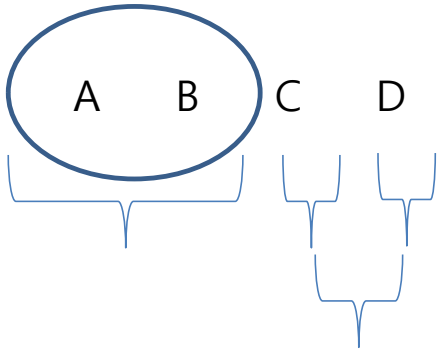
- Web Framework : Web 서비스 제작할 때 사용하는 도구
- Python 에서 Web 제작할 때 사용하는 여러가지 Web Framework 을 제공 (라이브러리)
- Web : MVC 패터닝 (Model (Data) / View (UX/) / Control)
- Python 대표적인 Web Framework :
  1. Django : 고 수준의 웹을 개발 / 각종 웹을 구성하고 있는 구성요소에 대한 모듈이 함수로 제공
     - 풀 프레임 워크 / 정교하게 웹 제작 / 대규모 프로젝트
  2. Flask : 마이크로 웹 프레임워크 / 간단한 Web 개발 할 수 있도록
     - 직관적 / 제작속도가 빠르다
  3. FastAPI : 매우 빠르고 간단한 형태로 Web 제작 프레임워크
     - 문서자동생성 / 직관 / 속도가 빠르다
  4. Dash : Flask 기반으로 제작 / 데이터 분석에 관련한 웹 프레임워크
     - 데이터 분석과 관련한 대시보드 / 분석 보고서 / 인터렉티브한 형태의 데이터 표현을 쉽고 직관적으로 구성
     - Dash : Plotly + React + Flask 구성 되어 있음
     - Callbcak 기능을 이용해 dynamic

2-6. System 구축

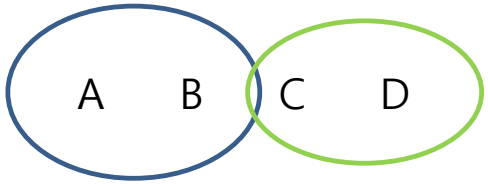
- **System 구축**
- **Dash Library를 활용한 Web Application 구성**

- Web Framework : Web 서비스 제작할 때 사용하는 도구
- Python 에서 Web 제작할 때 사용하는 여러가지 Web Framework 을 제공 (라이브러리)
- Web : MVC 패터닝 (Model (Data) / View (UX/) / Control)
- Python 대표적인 Web Framework :
  1. Django : 고 수준의 웹을 개발 / 각종 웹을 구성하고 있는 구성요소에 대한 모듈이 함수로 제공
     - 풀 프레임 워크 / 정교하게 웹 제작 / 대규모 프로젝트
  2. Flask : 마이크로 웹 프레임워크 / 간단한 Web 개발 할 수 있도록
     - 직관적 / 제작속도가 빠르다
  3. FastAPI : 매우 빠르고 간단한 형태로 Web 제작 프레임워크
     - 문서자동생성 / 직관 / 속도가 빠르다
  4. Dash : Flask 기반으로 제작 / 데이터 분석에 관련한 웹 프레임워크
     - 데이터 분석과 관련한 대시보드 / 분석 보고서 / 인터렉티브한 형태의 데이터 표현을 쉽고 직관적으로 구성
     - Dash : Plotly + React + Flask 구성 되어 있음
     - Callbcak 기능을 이용해 dynamic

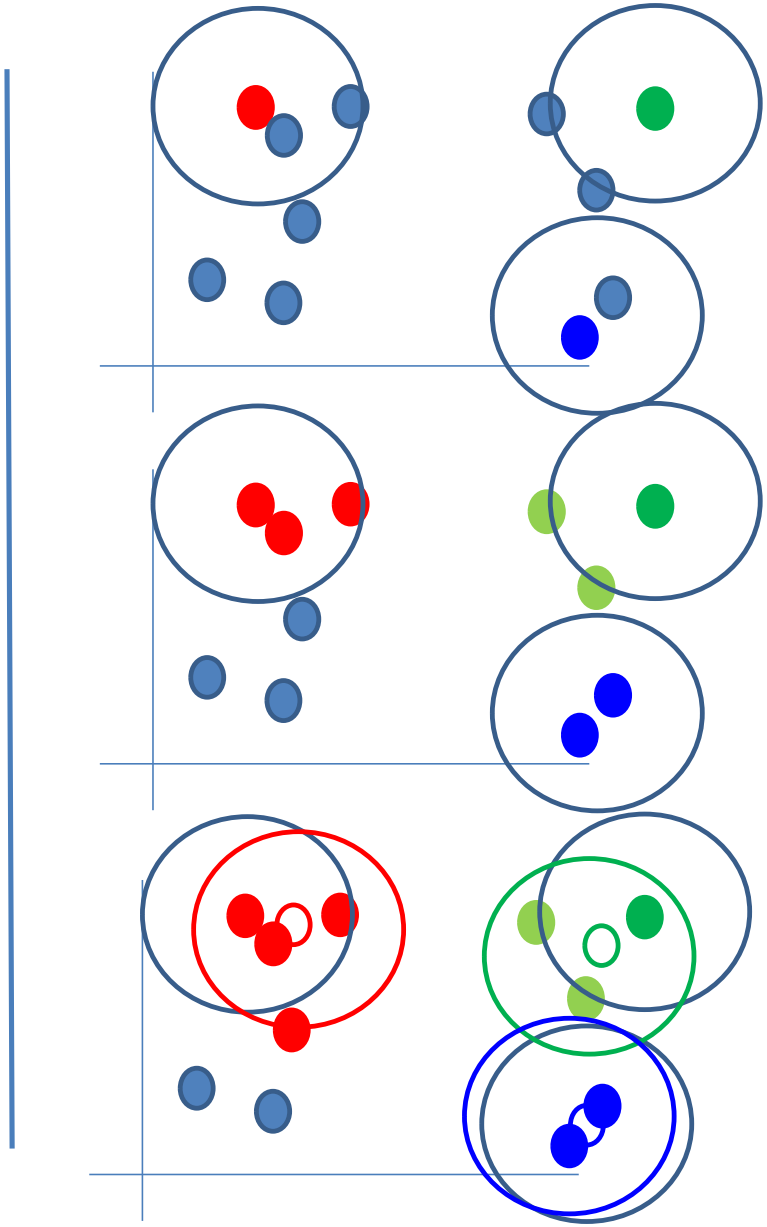
	A	B	C	D
A	0	5	10	20
B		0	30	20
C			0	10
D				0



	AB	C	D
AB	0	20	20
C		0	10
D			0



	AB	CD
AB	0	20
CD		0



Chapter 3. Python 프로그래밍 언어

3. Python 프로그래밍 언어

- **변수 :** 데이터를 담고 다니는 공간 (변수를 선언한다)
- **자료 구조 :** 데이터를 들고 다니는 상자들의 종류
    - List [ ] : 데이터의 추가/삭제/변경 가능 / 데이터 순서/중복 허용
    - Tuple ( ) : 데이터의 추가/삭제/변경 불가능 / 데이터 순서/중복 허용
    - Set { } : 데이터의 추가/삭제/변경 가능 / 데이터 순서/중복 불가능
    - Dictionary {key : value} : 키와 값의 매칭형태로 사용 -> 정형데이터 기본 틀이되는 기초 구조

-**데이터 타입 :**

- 연속형(숫자) : int / float
- 범주형(문자) : str / object
- 논리형(참/거짓) : bool (True/ False)

3. Python 프로그래밍 언어

구문 (Statement) :

-**조건문 :**

```
if ( 조건식) :
        (종속문장)
elif (조건식2) :
        (종속문장)
else :
        (종속문장)
```

-**반복문 (While / For)**

```
While (조건):
        (종속문장)
For (반복):
        (종속문장)
```

- **함수 :** 특정 기능을 수행하는 코드
- **라이브러리 :** 특정 목적을 달성하기 위한 함수들의 집합

3. Python 프로그래밍 언어

- 라이브러리 종류

Numpy (Numeric + Python) : 모든 숫자데이터의 통계/수학/과학 연산에 관련된 함수들의 집합

Pandas (Panel + Data Set) : 정형데이터의 통계/처리 관련 함수들의 집합

Matplot (Matlab + Plot) : 모든 숫자데이터의 시각화 관련된 함수들의 집합

Seaborn : 정형데이터의 통계 관련 시각화 함수들의 집합

Scipy (Science + Python) : 숫자데이터의 응용 통계 연산 (가설검정)과 관련된 함수들의 집합

Scikit Learn (Science Python Tool Kit Learning):

정형데이터의 기계학습 (Machine Learning)과 관련된 함수들의 집합

데이터 수집 : bs4 / selenium / request ...

비정형 데이터 분석 : Tensorflow (Tensor) / Keras / Pytorch ...

3. Python 프로그래밍 언어

- Pandas 정형 데이터 처리관련 함수

- **데이터 불러오기** pd.read_csv(' ') / pd.read_excel(' ') / pd.read_csv(' ' ,encoding='cp949')
- **데이터 저장** df1.to_csv(' ') / df1.to_excel(' ')
- **데이터 추출**
    - 행단위 추출 df1.head( ) / df1.tail( ) / df1.iloc[n : m]
    - 열단위 추출 df1['Col1'] / df1[['Col1', 'Col2']]
- **데이터 정렬** df1.sort_values(by='col1' , ascending = True/False)
- **데이터 필터**
    - df1.loc[(cond1) & (cond2)] / df1.loc[(cond1) | (cond2)]
- **데이터 요약**
    - df1.pivot_table(index='범주형', values='연속형' , aggfunc='통계량')
- **데이터 병합**
    - df1.merge(df1, df2, on='key col', how = 'inner / outer / left / right')
- **데이터 재구조화**
    - df1.melt(id_vars='key col')

3. Python 프로그래밍 언어

- 데이터 병합

이름	나이	성별
홍길동	30	남
이몽룡	36	남
성춘향	24	여

이름	부서	주소
홍길동	A	서울
이몽룡	B	경기
허준	A	제주
변사또	A	경기

Inner Join

이름	나이	성별	부서	주소
홍길동	30	남	A	서울
이몽룡	36	남	B	경기

Outer Join

이름	나이	성별	부서	주소
홍길동	30	남	A	서울
이몽룡	36	남	B	경기
성춘향	24	여		
변사또			A	제주
허준			A	경기

Left Join

이름	나이	성별	부서	주소
홍길동	30	남	A	서울
이몽룡	36	남	B	경기
성춘향	24	여		

ID	A	B
001A	1000	2000
001B	30	10
001C	20	20

	분류	값
001A	A	1000
	B	2000
001B	A	30
	B	10
0001C	A	20
	B	20

3. Python 프로그래밍 언어

- Pandas 정형 데이터 처리관련 함수

-**날짜 데이터 처리 :** pd.to_datetime(df1['날짜'], format='%Y%m%d')

연도 추출 : df1['날짜타입'].dt.year

월 추출 : df1['날짜타입'].dt.month

요일 추출 : df1['날짜타입'].dt.day_name()

- 결측치 처리 :

결측치 확인 : df1.isnull( ).sum( )

결측치 제거 : df1.dropna( )

결측치 대치 : df1.fillna(특정값)

3. Python 프로그래밍 언어

- 시각화 라이브러리 관련 함수

import seaborn as sns

import matplotlib as mpl

import matplotlib.pyplot as plt

그래프 옵션

그래프 저장 : plt.savefig('image1.png')

그래프 사이즈 : plt.figure(figsize= [10,5])

그래프 축 설정 : plt.ylim([n : m]) / 축을 n에서 m사이의 범위로 설정

그래프 이름 : plt.title( ) / 그래프 이름 설정

3. Python 프로그래밍 언어

- 시각화 라이브러리 관련 함수

한글 글꼴 설정

Colab

```
# 글꼴 설치 및 설정
!sudo apt-get install -y fonts-nanum
!sudo fc-cache -fv
!rm ~/.cache/matplotlib -rf  # 이후 런타임 재실행
mpl.pyplot.rc('font',family='NanumBarunGothic')
```

Jupyter

Window : mpl.rc('font', family= 'Malgun Gothic')

Mac : mpl.rc('font', family= 'AppleGothic')

3. Python 프로그래밍 언어

- 시각화 라이브러리 관련 함수

단일변수

빈도수 확인 : sns.countplot(data= df1, x='Col1')

분포 확인 : sns.histplot(data= df1, x='Col1')

sns.displot(data= df1, x='Col1' , kde=True)

sns.displot(data=df1, x='Col1', kind='kde')

상자그림 (사분범위) : sns.boxplot(data= df1, x='Col1')

다 변수

X : 범주형 / Y : 연속형 : sns.barplot(data=df1, x='범주형', y='연속형')

X : 연속형 / Y : 연속형 : sns.scatterplot(data=df1, x='연속형', y='연속형')

sns.lmplot(data=df1, x='연속형', y='연속형')

sns.pairplor(data=df1)

X : 순서형(시간) / Y : 연속형 : sns.lineplot(data=df1, x='순서형', y='연속형')

sns.pointplot(data=df1, x='순서형', y='연속형')

3. Python 프로그래밍 언어

scikit Learn

학습 데이터 검증 데이터 분할

```
from sklearn.model_selection import train_test_split
```

파이프 라인 구축

```
from sklearn.pipeline import PipeLine
```

스케일링

```
from sklearn.preprocessing import StandardScaler # 평균 0 / 표준편차 1 스케일링
from sklearn.preprocessing import MinMaxScaler # 최소값 0 / 최대값 1 스케일링
from sklearn.preprocessing import RobustScaler # 중앙값 0 / IQR 1 스케일링
```

결측 대치

```
from sklearn.impute import SimpleImputer  # 평균 대치
from sklearn.impute import KNNImputer  # 모델 대치
```

교차검증 & 하이퍼파라미터 튜닝

```
from sklearn.model_selection import GridSearchCV
```

Imbalanced Data Sampling

```
from imblearn.under_sampling import RandomUnderSampler
from imblearn.over_sampling import RandomOverSampler
```

학습 모델 저장 / 불러오기

```
import pickle
pickle.dump( best_model, open('model.sav', 'wb'))
pickle.load( open('model.sav', 'rb'))
```

Chapter 4. 제조 공정/품질 데이터 분석

4. 제조 공정/품질 데이터 분석

- **공정 데이터 분석 -> QCD (Quality / Cost / Delivery)**
- **품질 (Quality) :** 제품 생산 및 품질 검사에서 나오는 데이터를 활용해 품질을 향상시키기 위한 개선 작업
- **QC7 : 적정수준의 품질을 관리하기 위해 데이터로부터 확인하는 7가지**

1) 특성요인도 (Fish Bone Chart) : 결과(특성)에 어떤 원인(요인)이 있는지 그래프로 표현
(4M 주 요인 4가지 : Man / Material / Method / Machine)

2) 파레토도 (Pareto Plot) : 현장에서 발생하는 주 핵심 지표(결점수/불량수/VOC..)
그래프로 표현 / 막대그래프(빈도수) + 선 그래프(누적확률)
- 전체 80% 결과는 20% 원인에 의해 발생한다.

3) 관리 항목 (Check Sheet) : 핵심 지표들을 범주형 항목별로 구분하여 정리한 표 (점검용/기록용)

4) 산점도 (Scatter Plot) : 공정에서 발생하는 주요 인자를 좌표평면위에 점 형태로 표현하여
변수 별 상관관계 연관성 등을 파악

5) 히스토그램 (Histogram) : 공정작업 / 품질검사 단계에서 발생하는 연속형 자료들의 분포를 확인

6) 층별화 (Stratification) : 데이터를 특성(범주형)에 따라 분류하여 확인

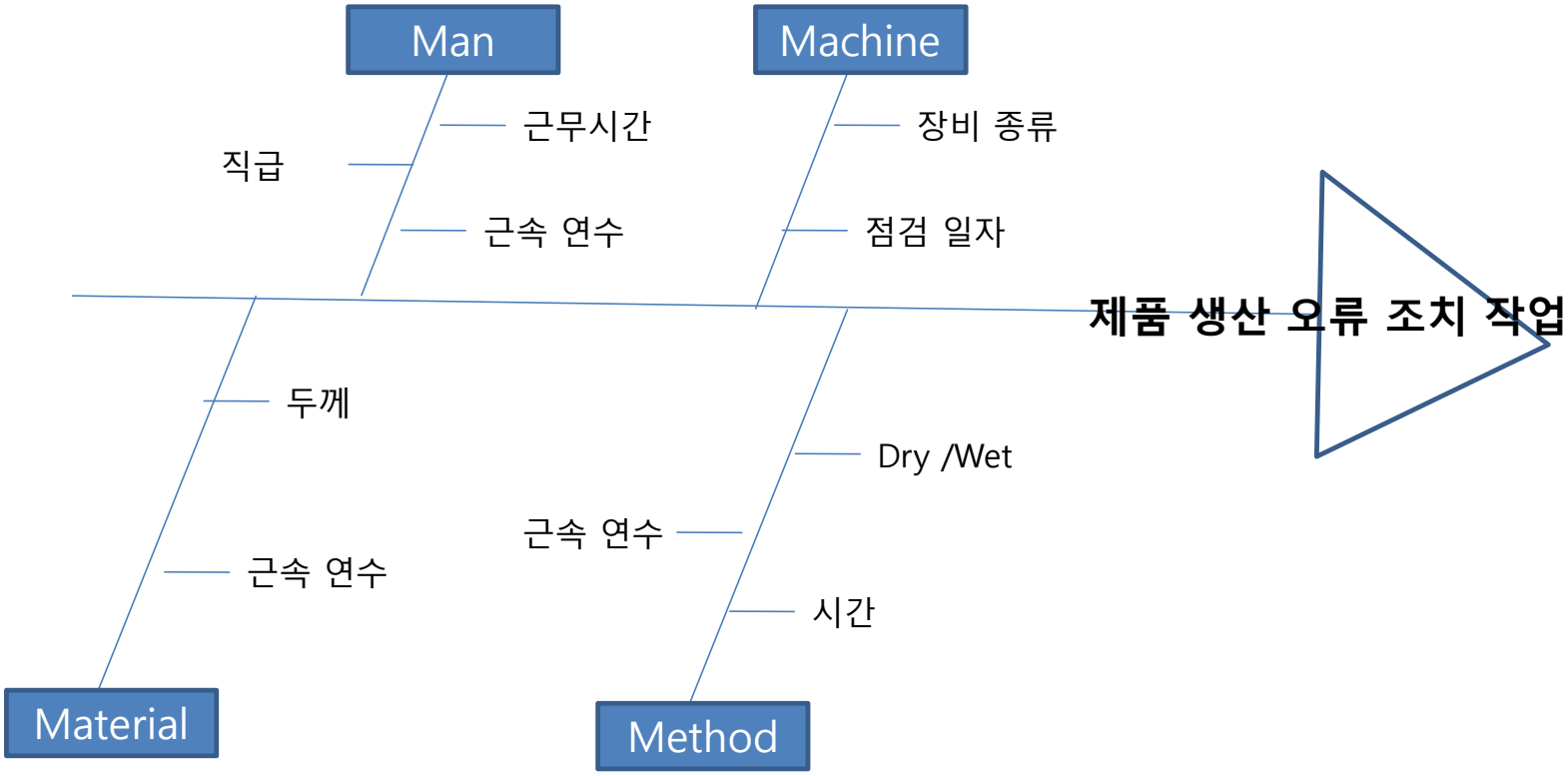
7) 관리도 (Control Chart) : 공정이나 품질에서 발생하는 데이터의 이상치를 확인하고
빠르게 조치하기 위해 사용하는 시각화 기법

4. 제조 공정/품질 데이터 분석

- **공정 데이터 분석 -> QCD (Quality / Cost / Delivery)**

1) 특성요인도 (Fish Bone Chart) : 결과(특성)에 어떤 원인(요인)이 있는지 그래프로 표현

(4M 주 요인 4가지 : Man / Material / Method / Machine)



4. 제조 공정/품질 데이터 분석

- **공정 데이터 분석 -> QCD (Quality / Cost / Delivery)**

1) 특성요인도 (Fish Bone Chart)

- **변수 중요도** (1/3/9 척도로 실무자와 분석가들이 스코어링 -> 종합하여 평가)

주요인자	수집가능성	중요도
근속연수	3	1
숙련도	1	9
근무시간	**9**	**9**
장비종류	**9**	**9**
수리일자	9	3
재료두께	3	3
…	…	…

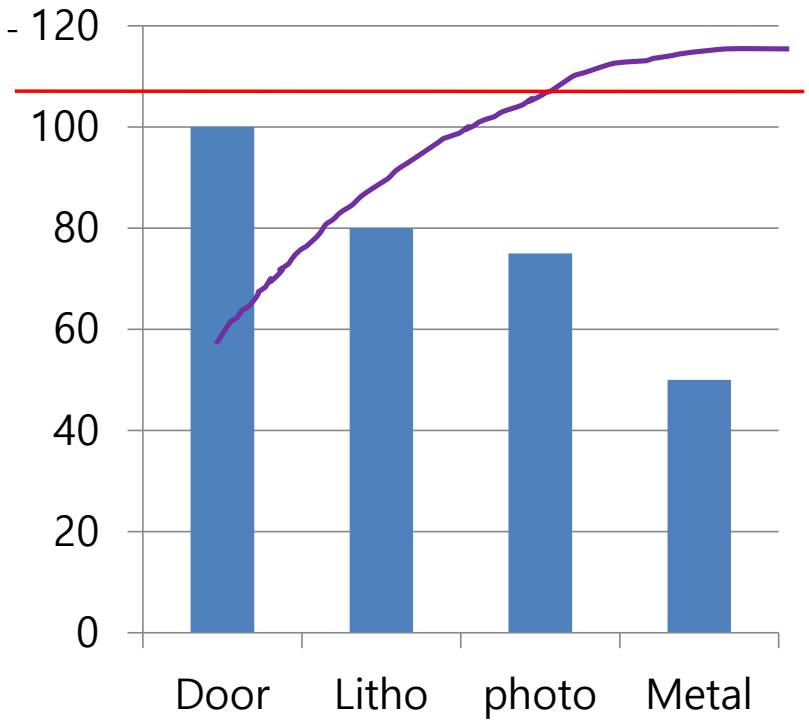
- **변수 정의서** (변수 중요도에서 선택된 인자를 정리)

항목명	데이터 타입	설명
근무시간	**연속형 (datetime)**	**작업자의 근무시간 (format)**
장비종류	**범주형**	
수리일자	**연속형 (datetime)**	
재료두께	연속형	
…		

4. 제조 공정/품질 데이터 분석

- **공정 데이터 분석 -> QCD (Quality / Cost / Delivery)**

2) 파레토도 (Pareto Plot) : 막대그래프(빈도수) + 선 그래프(누적확률)



불량요인	빈도수	비율	누적확률
Door	100	45%	45%
Lithograpy	80	20%	60%
Photo	76	15%	75%
Metal	50	10%	85%
…	…	…	…
			100%

4. 제조 공정/품질 데이터 분석

- **공정 데이터 분석 -> QCD (Quality / Cost / Delivery)**

7) 관리도 (Control Chart) : 공정이나 품질에서 발생하는 데이터의 이상치를 확인하고 빠르게 조치하기 위해 사용하는 시각화 기법

- 품질 변동 원인 (품질 불량 결과에 대해)

- 우연 원인 : 피치 못할 원인
- 이상 원인 : 통제 가능한 (피할 수 있는) 원인

통제 가능한 원인들 중, 이상상태 (Out of Control)를 찾아내는 목적

- 관리도의 종류
  - 계량치 관리도 : Xbar –R : 평균과 범위(최소-최대)를 이용하여 시각화 한 관리도
    - **Xbar –σ :** 평균과 표준편차를 이용하여 시각화 한 관리도

  (계량치 : 실수형태로 표현 가능 한 연속형 숫자 / 온도,강도,습도 ...)

  - 계수치 관리도 : **p 관리도 (불량률 관리도)**

    pn 관리도 (불량 개수 관리도)

    c 관리도 (결점수 관리도)

    u 관리도 (단위 당 결점수 관리도)

  (계수치 : 정수형태로 표현될 수 있는 연속형 데이터 / 불량 개수, 결점 수, ...)

Chapter 5. 데이터 분석 과제 수행

0. Team Building

1. 팀장 및 팀 이름 정하기 (응원가, 팀 구호, ...)
2. 서로의 MBTI 물어보기 (취미, 좋아하는 음식, ...)
3. 팀 채널 구성 (카카오톡 / 슬랙 / 등...)
4. 선호 과제 2순위 까지 설정

A팀	B팀	C팀	D팀
이준엽	김영우	남정윤	이용석
박상범	정연재	구본하	안주강
박태윤	양승호	변지은	정예은
박은영	이다련	김현영	전준우

김예슬

팀장의 역할 :

- 타임 키퍼 (PPT나 과제 제출 시간 엄수)
- 역할 배분 및 회의 의견 조율

주제

유아용품(CRM) / 간편식공정(제조) / 포장회사(유통) / 척추병원(의료)

0. Team Building

DNA : 이용석 안주강 정예은 전준우 / 팀장 : 정예은

- 과제 : 척추의료

기억해'조' : 박상범 이준엽 박태윤 박은영 / 팀장 : 박상범

- 과제 : 포장

타코야끼 : 남정윤 구본하 변지은 김현영 김예슬 / 팀장 : 남정윤

- 과제 : 간편식

영우형"해조" : 김영우 정연재 양승호 이다련 / 팀장 : 김영우

- 과제 : 유아용품

0. 평가 항목

- 프로젝트 결과물

1. 개인 보고서 (프로젝트 흐름에 따른 자유형식)

- 개인 분석 Code + 정리 PPT (100%)

2. 팀 발표 PPT (발표에 사용되는 정형 형식)

- 분석흐름의 논리성
- 분석기법의 다양성
- 개선안 독창성
- 발표 및 보고서 전달성
- 팀 협동성

3. 팀 활동 (팀 별 활동비 100,000) : 인증 사진

- 팀 협동성

1. 과제 정의

1. **“과제명” 설정**
2. **추진 배경 수립**
3. **현상 파악 및 목표 설정**
4. **잠재 인자 도출**

-> 10월 13일 오전 10시 변수정의서 제공

-> 10월 13일 오후 14:00 발표

-> 발표 자료는 12:00까지 송부 0001jmp@gmail.com (송부 후 수정가능)

-> 발표는 15분 이내

1. 과제 정의

1. “과제명” 설정

“대상 + 개선방법 + 효과(구체적 목표)”

【 Best 사례 】

- ✓ **프레스 속도 실시간 제어기술 개발을 통한 Loss Time Zero화**
- ✓ **적조발생 및 확산 예측시스템 구축을 통한 수산피해 최소화**
- ✓ **이천시 CCTV 사각지대 실시간 모니터링 및 사전 예측을 통한 범죄예방 강화**
- ✓ **장애인 콜택시 배차 최적화 알고리즘 구축을 통한 대기시간 단축**

【 Worst 사례 】

- ✓ 에너지설비 전력 원가절감
- ✓ 배기시스템 효율 향상 기술 개발
- ✓ 강남구 소매사업 상권분석
- ✓ 대형 컨테이너선용 에너지 절감 기술개발
- ✓ 오피스텔 임대 서비스 현황 분석

1. 과제 정의

2. 추진 배경 수립

-> 이 과제를 “왜” 수행하는가 에 대한 근거자료

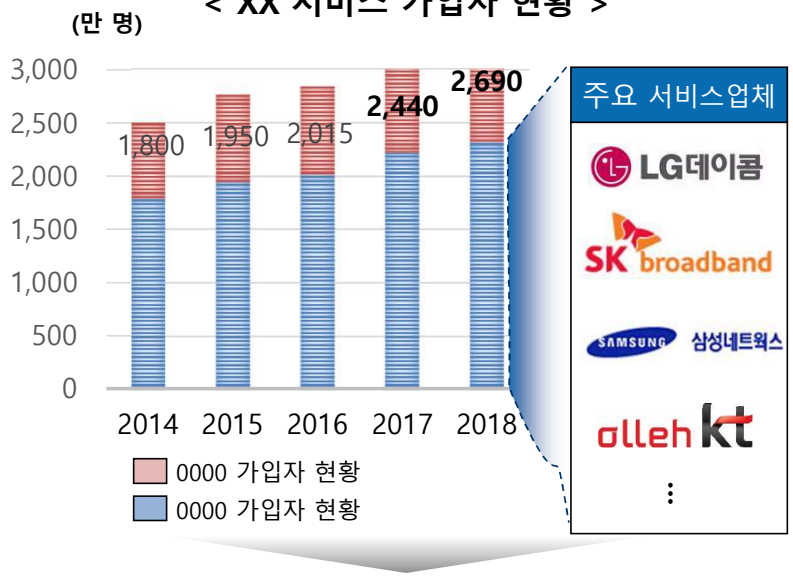
-> 목표 전략 / 시장 / 고객 니즈 등을 연계하여 작성

-> 비즈니스 시나리오에 대한 **도메인(“Domain”)** 공부

〈추진배경 사례〉

Mass 고객에 대한 무선통신서비스의 경쟁심화 및 정부의 통신요금 인하 의지 등은 우리회사의 성장을 위협하고 있음

< XX 서비스 가입자 현황 >



당사가 Monopolize 해왔던 XX 시장은 정체기에 들어있으며 이를 대체하여 급성장 중인 XX 시장은 타 업체들과 공유하고 있는 상황임

Source: 방송통신위원회, '유·무선 통신서비스 가입자 현황', 미래기획위원회와 방송통신위원회 공동 주최 '이동통신 요금정책 세미나 '

< 통신 서비스 관련 정부 정책 >

통신요금 인하

"이동통신 품질을 유지해 IT강국의 면모를 유지하는 동시에 현 정부의 통신요금 20% 인하 공약을 실현하겠다"

- XXX 방통위원장

"서민들이 체감할 수 있는 요금인하 정책을 마련하겠다"

- XXX 미래기획위원장

가상이동 통신망 사업자

"MVNO 도입을 통한 경쟁 활성화, 보조금의 요금인하 전환, 결합상품 활성화, 무선데이터 요금 인하 등을 중심으로 요금인하를 추진하겠다"

- XXX 방통위원장

향후, ~~~~~ 를 통해 ~~ 를 하고 ~~~~~~~ 필요성 있음

1. 과제 정의

3. 현상 파악 및 목표 설정

- 과제를 통해 해결 해야 할 현상과 문제를 구체적으로 기술

- 목적과 목표 설정

목적 : 분석의 방향성

목표 : 목적을 달성하기 위한 구체적인 수치

○ 가중치 : 지표가 여러 개일 경우 합이 100%가 되도록 작성
○ 현수준 : 과제 착수 前 12개월 ~ 36개월을 기준으로 산정
○ 목　표 : 점진적 증가가 아닌, 과거 개선율을 월등하게 뛰어 넘는 도전적인 목표 설정

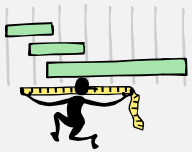
측정지표 (KPI)	가중치	현수준	목표수준		
			'18년	'19년	'20년
지역별 분리구매 대비 할인구매율 (%)	40%	**2.0**	2.7	3.3	3.8
해외법인 본사 통합구매 비율(%)	40%	**3**	42	46	50
외주운영 선진화 지표(점)	20%	-	85	90	95

<현상파악 사례>

○ 출력단 회로 오차 증가 시 시스템 정밀도가 감소하며, 차량 적용 평가 시 고객 불만족 증대 예상

자사

시험평가팀과
Argue 발생에 따른
검증 일정 지연



문제

* XXX 출력단 오차가 비교적 큼

- 현 국산화 H/W 오차 : 13%
- 해외경쟁사 H/W 오차 : 6%

* 문제 발생 시 개선 범위가 커서
즉시 해결 불가능 예상

고객 (OEM)

법규 확인 시험
Fail 가능성
➔ **적용 불가**

○ 전체 불량률 중 고객불량률 25%를 차지하며 셔틀라인의 불량률이 전체의 68.4%로 집중 개선 필요

201X년 공정별 고객불량 현황

(20XX. 01 ~ 20XX. 12)

구 분	불량수량(EA)	점유율(%)
셔틀라인	201	68.4
용접라인	38	12.9
도장라인	7	2.4
기 타	48	16.3
TOTAL	294	100

불량개선	• 불량 수량 or 건수 기준 WORST 불량 선정, 집중개선
품질의식	• 현장 품질 문제 발생시 대응PROCESS 정립 및 품질의식 향상

▶ 개선 목표 : XXX 출력단 오차 13% → 5% | 고객불량률 25% → 10%

1. 과제 정의

4. 잠재인자 도출

- 특성 요인도 (Fish Bone Chart)
- 잠재원인 우선순위화
- **변수 정의서**

잠재원인	중요도	분석 가능성
X1	9	2
X2	2	2
X3	7	7
X4	2	7
X5	8	9
X6	3	3

과제명 : ~

○ ~~~~~~~~~~~ 핵심 메시지

그래프, 그림, 도형, 표,설명자료

14 pt, 12 pt

그래프, 그림, 도형, 표,설명자료

14 pt, 12 pt

○ ~~~~~~~~~~~ 핵심 메시지

그래프, 그림, 도형, 표,설명자료

14 pt, 12 pt

그래프, 그림, 도형, 표,설명자료

14 pt, 12 pt

추진방향 또는 목적 기술 과제수행의 필요성 강조 *16 pt 볼드체, 컬러 강조

현황 *개선기회로 기술가능

○ ~~~~~~~~~~ 헤드라인 16 pt

그래프, 그림, 도형, 표,설명자료

14 pt, 12 pt

그래프, 그림, 도형, 표,설명자료

14 pt, 12 pt

○ ~~~~~~~~~~ 헤드라인 16 pt

그래프, 그림, 도형, 표,설명자료

14 pt, 12 pt

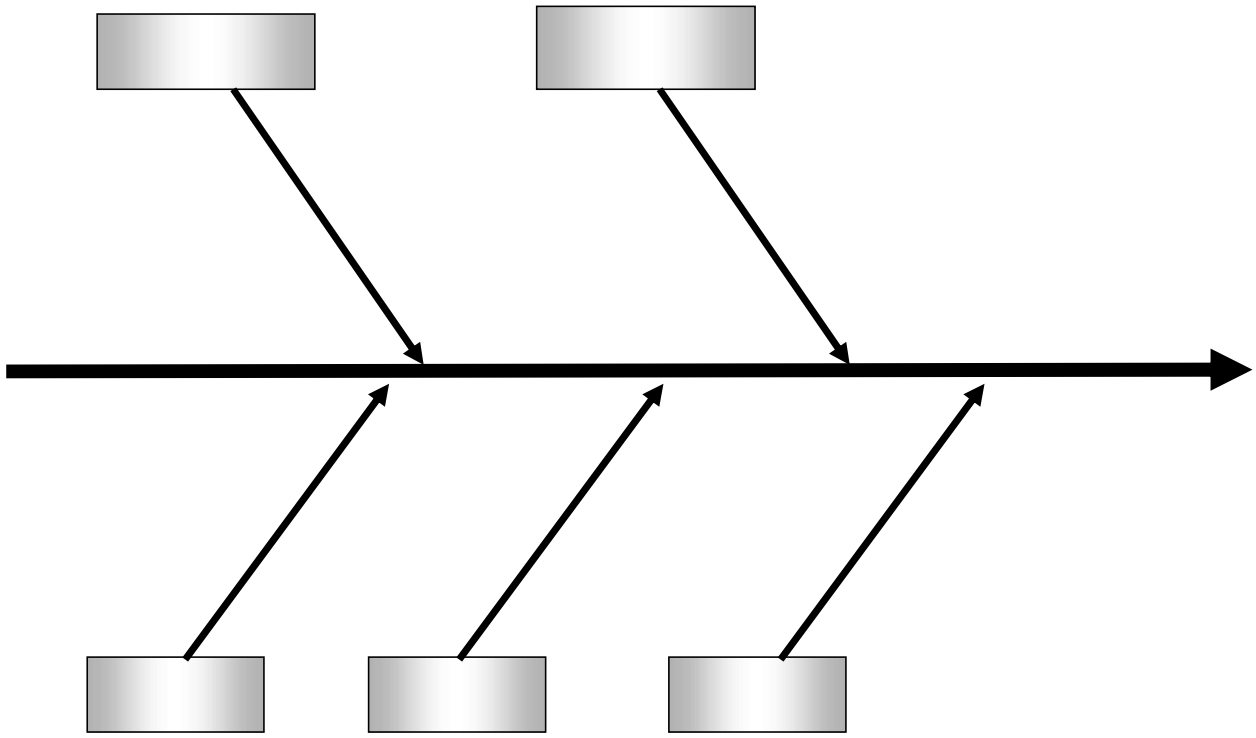
그래프, 그림, 도형, 표,설명자료

14 pt, 12 pt

지표명과 개선목표를 설정 *16 pt 볼드체, 컬러 강조

잠재원인 도출

○ ~~~~~~~~~~~ **헤드라인 16 pt**



잠재원인우선순위화

[실습 템플릿]

○ ~~~~~~~~~~~ 헤드라인 16 pt

잠재원인	중요도	분석가능성	합계	선정