

# 데이터 분석 개요

# CONTENTS

---

<b>1. 데이터 분석 개요</b>	03
1) 데이터 분석 개념	
2) 데이터를 분석을 위한 역량	
3) 데이터 분석 - 머신 러닝	
<b>2. 데이터 형식</b>	07
1) 데이터 양식	
2) 데이터 모양	
3) 데이터 형태	
4) 데이터 처리	
<b>3. Machine Learning</b>	12
1) 수치 예측 vs. 분류	
2) 수치 예측 모델 성능 평가	
3) 분류 모델 성능 평가	

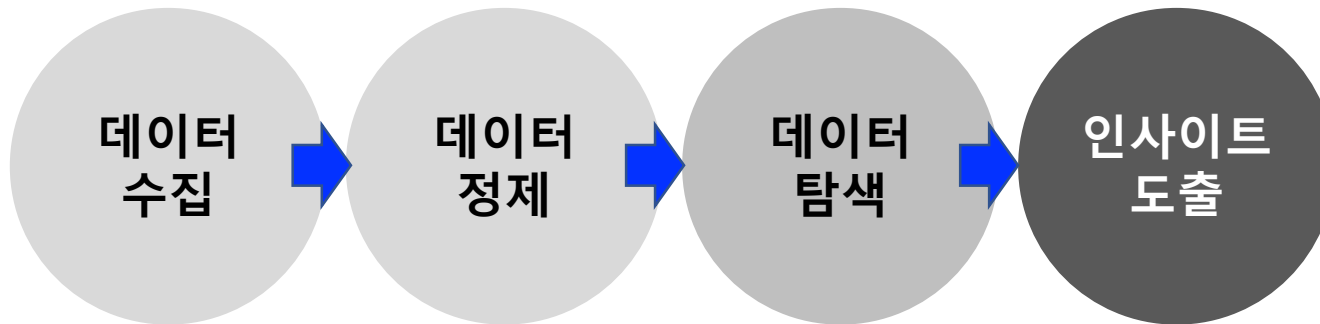
## 1. 데이터 분석 개요

## 1. 데이터 분석 개요

### 1-1. 데이터 분석 개념

---

- 데이터간의 관계를 파악하고 의미 있는 형태로 가공한다.
- 가공된 데이터를 기반으로 의미 있는 정보를 추출한다.
- 추출한 정보를 토대로 의사 결정을 수행한다.
- 차트, 대시 보드 등을 통한 시각화 구성한다.

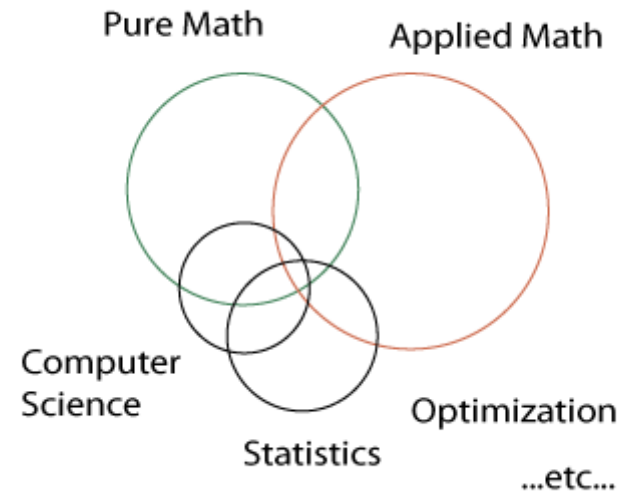
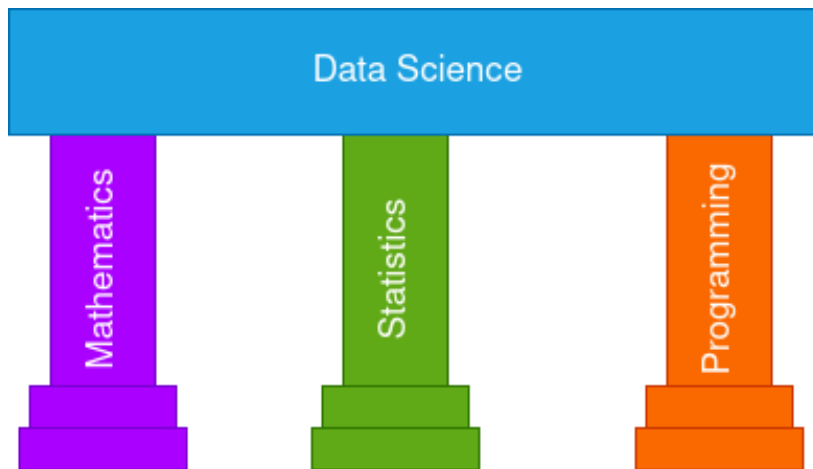


## 1. 데이터 분석 개요

### 1-2. 데이터 분석을 위한 역량

---

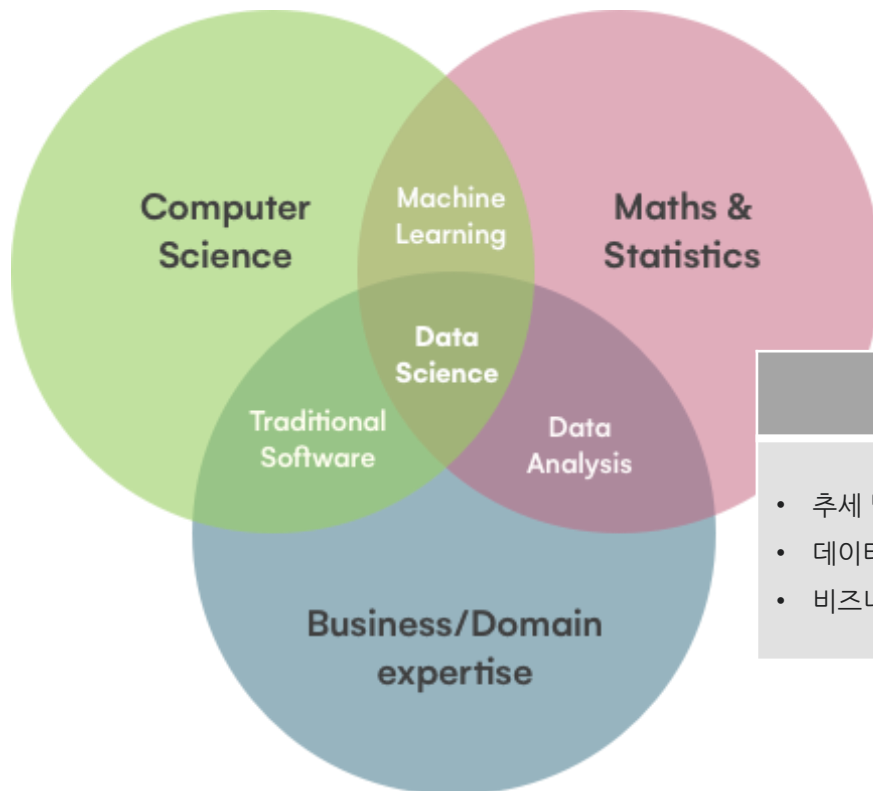
- 데이터 분석 및 과학은 **수학과 통계, 프로그래밍** 등 다양한 분야의 지식을 필요로 한다.
- 선형 대수학은 선형 방정식과 선형 방정식 그래프에 대한 연구로 통계 그래프를 이해하기 위한 기초이며
- 통계는 데이터를 이해하고 해석하고 제시하기 위한 기초 지식이다.



## 1. 데이터 분석 개요

### 1-3. 데이터 분석 - 머신 러닝

- 데이터 분석은 데이터 탐색 및 가공을 통하여 의미 있는 정보를 발굴한다.
- 머신 러닝의 역할은 알고리즘을 통해 기계가 의사 결정을 수행하도록 구성한다.



데이터 분석	데이터 과학
<ul style="list-style-type: none"><li>• 추세 및 메트릭 탐색</li><li>• 데이터 시각화</li><li>• 비즈니스 지식 및 의사 결정 능력</li></ul>	<ul style="list-style-type: none"><li>• 수학적 알고리즘 설계 및 구현</li><li>• 기계 학습에 대한 지식</li><li>• 비정형 데이터로 작업하는 경향</li></ul>

## 2. 데이터 형식

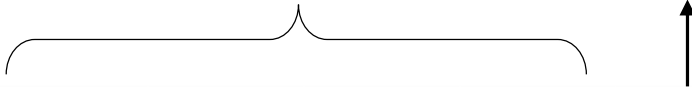
## 2. 데이터 형식

### 2-1. 데이터 양식

---

- 독립 변수 ( $X$ )
- 종속 변수 ( $y$ )

- Predictor variables(예측변수)
- Input variables(입력변수)
- Independent(독립변수)
- Target variables(타겟변수)
- Output variables(출력변수)
- Dependent variables(종속변수)



id	$X_1$	$X_2$	...	$X_p$	$Y$
1	$x_{11}$	$x_{12}$	...	$x_{1,p}$	$y_1$
2	$x_{21}$	$x_{22}$	...	$x_{2,p}$	$y_2$
...	...	...	...	...	...
$n$	$x_{n,1}$	$x_{n,2}$	...	$x_{n,p}$	$y_n$



## 2. 데이터 형식

### 2-2. 데이터 모양

---

- Wide Data Format
- Long Data Format

**Wide Format**

Team	Points	Assists	Rebounds
A	88	12	22
B	91	17	28
C	99	24	30
D	94	28	31

**Long Format**


Team	Variable	Value
A	Points	88
A	Assists	12
A	Rebounds	22
B	Points	91
B	Assists	17
B	Rebounds	28
C	Points	99
C	Assists	24
C	Rebounds	30
D	Points	94
D	Assists	28
D	Rebounds	31

## 2. 데이터 형식

### 2-3. 데이터 형태 (수치형 / 범주형)

- 명목 변수와 서열 변수는 **범주형** 이고 등간 변수와 비율 변수는 **수치형**이다.
- **범주형** 데이터 보다 **수치형** 데이터에 대해 더 많은 통계 테스트를 해볼 수 있다.
- 등간 (예: 온도) 및 비율(예: 거리) 척도는 모두 동일하게 '간격'이라는 특성을 갖지만 비율 척도에만 절대 '0'이 있다.

	명목 척도	서열 척도	등간 척도	비율 척도
	Nominal	Ordinal	Interval	Ratio
Categories	●	●	●	●
Rank order		●	●	●
Equal spacing			●	●
True zero				●

 **The 4 levels of measurement**

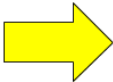
## 2. 데이터 형식

### 2-4. 데이터 처리

데이터 분석의 안정적인 결과와 성능 향상을 위해서 주어진 데이터를 분석에 적합하게 가공하는 작업이다.  
대표적인 작업으로는 필터링, 클리닝, 결측치 처리, 이상치 처리, 데이터 형태 변경 등이 있다.

- 범주형 데이터 인코딩 : 레이블 인코딩 (Label Encoding) & 원핫 인코딩 (One-hot Encoding)
- 수치형 데이터 스케일링 : 표준화 (Normalization) & 정규화 (Standardization)
- Filtering / Cleaning / Missing Value / Outlier
- Data Shape : Long Data, Wide Data

One-hot Encoding

Color		Red	Yellow	Green
Red		1	0	0
Red		1	0	0
Yellow		0	1	0
Green		0	0	1
Yellow		0	0	1

Feature Scaling

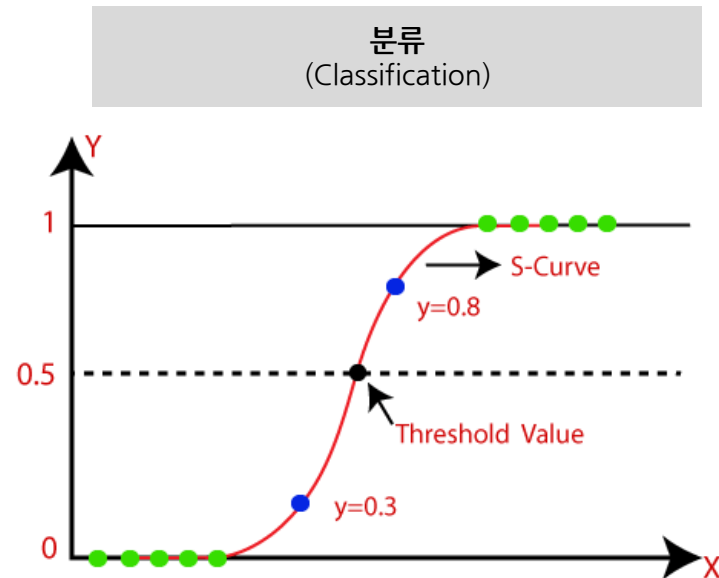
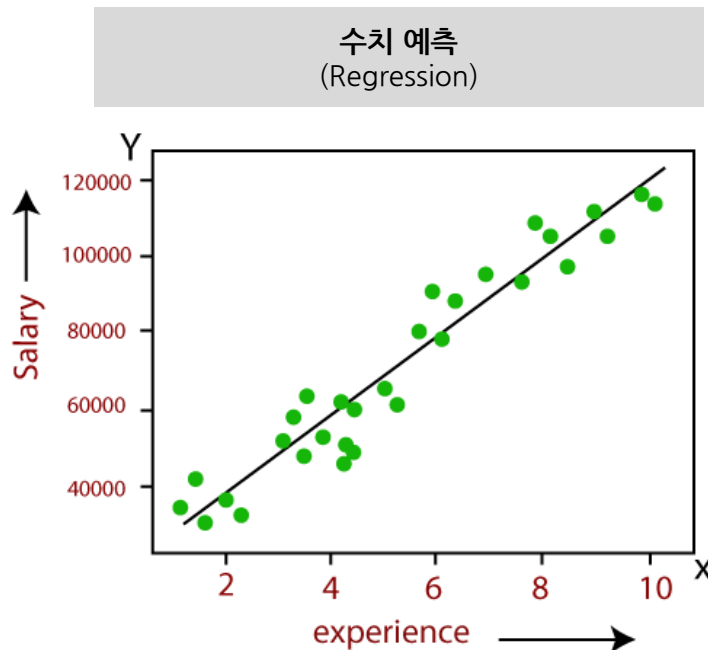
Standardisation	Normalisation
$x_{\text{stand}} = \frac{x - \text{mean}(x)}{\text{standard deviation}(x)}$	$x_{\text{norm}} = \frac{x - \min(x)}{\max(x) - \min(x)}$

### 3. Machine Learning

### 3. Machine Learning

#### 3-1. 수치 예측 vs. 분류

- 머신 러닝의 지도 학습은 수치를 예측하는 수치 예측(Regression)과 분류(Classification)로 나눌 수 있다.
- 수치 예측(Regression)은 y값의 데이터 형태가 수치형 데이터 일때 사용 할 수 있는 알고리즘이다.
- 분류(Classification)는 y값의 데이터 형태가 범부형 데이터 일때 사용 할 수 있는 알고리즘이다.



### 3. Machine Learning

#### 3-2. 수치 예측 모델 (Regression) 성능 평가

평가 지표	설명	수식
MAE	Mean Absolute Error이며 실제 값과 예측 값의 차이를 절대값으로 변환해 평균한 것	$\frac{1}{n} \sum_{i=1}^n  y_i - \hat{y}_i $
MSE	Mean Squared Error이며 실제 값과 예측 값의 차이를 제곱해 평균한 것 *MAE값이 같은데 MSE가 클 경우 편차가 더 큼을 나타낸다.	$\frac{1}{n} \sum_{i=1}^n (y_i - \hat{y}_i)^2$
RMSE	MSE 값은 오류의 제곱을 구하므로 실제 오류 평균보다 더 커지는 특성이 있으므로 MSE에 루트를 씌운 것이 RMSE(Root Mean Squared Error)다.	$\sqrt{\frac{1}{n} \sum_{i=1}^n (y_i - \hat{y}_i)^2}$
R <sup>2</sup>	분산 기반으로 예측 성능을 평가합니다. 실제 값의 분산 대비 예측값의 분산비율을 지표로 하며, 1에 가까울수록 예측 정확도가 높다. *R <sup>2</sup> = 0.91인 경우, 전체 데이터 변동성의 91%를 선형회귀 모델이 설명	$\frac{\sum(\hat{y}_i - \bar{y})^2}{\sum(y_i - \bar{y})^2}$

### 3. Machine Learning

#### 3-3. 분류 모델(Classification) 성능 평가

- 오차 행렬(또는 혼동 행렬)은 분류 모델의 성능을 평가하기 위해 실제 값(Actual Values)와 예측 값(Predictive Value)을 비교 하는 표이다.

		PREDICTIVE VALUES		
		POSITIVE (1)	NEGATIVE (0)	
ACTUAL VALUES	POSITIVE (1)	TP = 3	FN = 1	4
	NEGATIVE (0)	FP = 2	TN = 4	6
		5	5	

PRECISION (green box around TP and FP)

RECALL (red box around TP and FN)

정확도  $Accuracy = \frac{TP + TN}{TP + TN + FP + FN}$

재현도  $Recall / Sensitivity = \frac{TP}{TP + FN}$

특이성  $Specificity = \frac{TN}{TN + FP}$

정밀도  $Precision = \frac{TP}{TP + FP}$

$$F1-Score = 2 * \frac{Precision * Recall}{Precision + Recall}$$

End of Document