개인과제

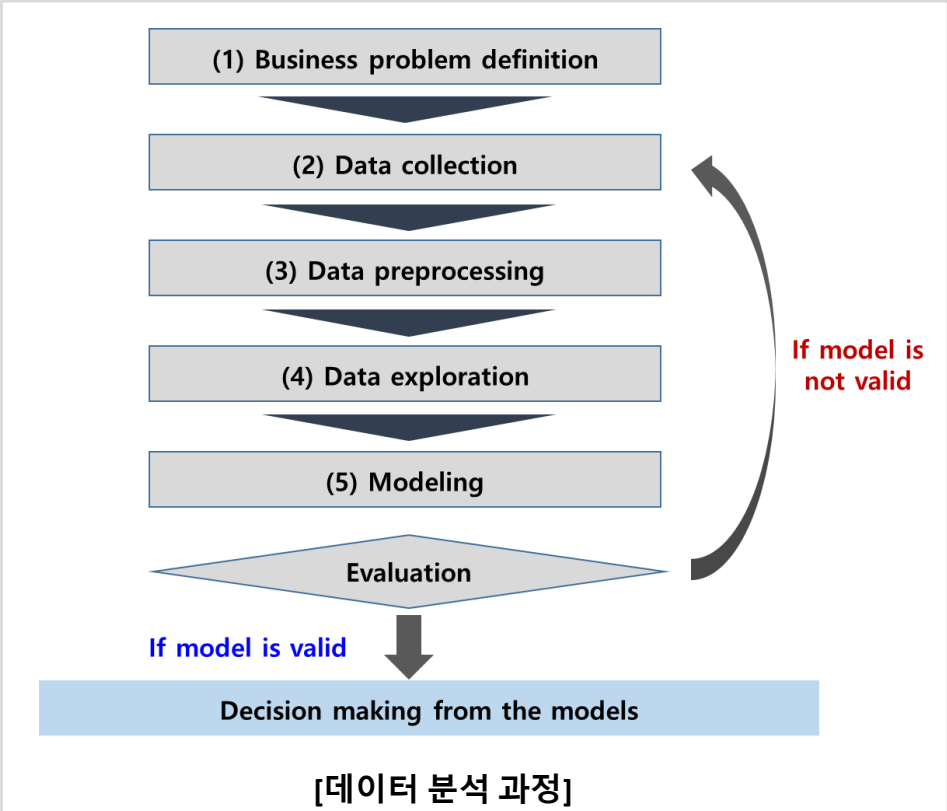
School of Industrial and Management Engineering, Korea University

Jieun Son

문제

- scikit-learn에서 제공하는 데이터 셋 2개 中 택 1
  - california-housing-dataset
  - breast-cancer-wisconsin-diagnostic-dataset

- 2강에서 강의한 '데이터 분석과정'에 맞도록 문제를 정의하고 선택한 데이터를 분석 (파이썬)



[데이터 분석 과정]

문제

- 수업시간에 강의한 내용과 데이터마이닝 알고리즘을 자유롭게 선택하여 활용
  - 데이터 전처리
  - 데이터 탐색
  - 분류
  - 군집
  - 연관규칙
  - 차원축소

- 반드시 포함 할 내용
  - 데이터 분석 과정 (문제 정의, 데이터 설명, 데이터 전처리, 데이터 탐색, 모델, 평가)

- 자율선택
  - 파이썬 패키지 및 모듈 종류

데이터 셋 1

- **california-housing-dataset**

  https://scikit-learn.org/stable/datasets/real_world.html#california-housing-dataset

```
import pandas as pd
from sklearn import datasets
load_df=datasets.fetch_california_housing()

data= pd.DataFrame(load_df.data)
feature= pd.DataFrame(load_df.feature_names)
data.columns = feature[0]
target=pd.DataFrame(load_df.target)
target.columns=['target']
df= pd.concat([data,target], axis=1)
print(df.shape)
df.head()
```

```
(20640, 9)
```

	MedInc	HouseAge	AveRooms	AveBedrms	Population	AveOccup	Latitude	Longitude	target
0	8.3252	41.0	6.984127	1.023810	322.0	2.555556	37.88	-122.23	4.526
1	8.3014	21.0	6.238137	0.971880	2401.0	2.109842	37.86	-122.22	3.585
2	7.2574	52.0	8.288136	1.073446	496.0	2.802260	37.85	-122.24	3.521
3	5.6431	52.0	5.817352	1.073059	558.0	2.547945	37.85	-122.25	3.413
4	3.8462	52.0	6.281853	1.081081	565.0	2.181467	37.85	-122.25	3.422

데이터 셋 2

- **breast-cancer-wisconsin-diagnostic-dataset**

  https://scikit-learn.org/stable/datasets/toy_dataset.html#breast-cancer-wisconsin-diagnostic-dataset

```
import pandas as pd
from sklearn import datasets
load_df=datasets.load_breast_cancer()

data= pd.DataFrame(load_df.data)
feature= pd.DataFrame(load_df.feature_names)
data.columns = feature[0]
target=pd.DataFrame(load_df.target)
target.columns=['target']
df= pd.concat([data,target], axis=1)
print(df.shape)
df.head()
```

```
(569, 31)
```

	mean radius	mean texture	mean perimeter	mean area	mean smoothness	mean compactness	mean concavity	mean concave points	mean symmetry	mean fractal dimension	...	worst texture	worst perimeter	worst area	worst smoothness	con
0	17.99	10.38	122.80	1001.0	0.11840	0.27760	0.3001	0.14710	0.2419	0.07871	...	17.33	184.60	2019.0	0.1622	
1	20.57	17.77	132.90	1326.0	0.08474	0.07864	0.0869	0.07017	0.1812	0.05667	...	23.41	158.80	1956.0	0.1238	
2	19.69	21.25	130.00	1203.0	0.10960	0.15990	0.1974	0.12790	0.2069	0.05999	...	25.53	152.50	1709.0	0.1444	
3	11.42	20.38	77.58	386.1	0.14250	0.28390	0.2414	0.10520	0.2597	0.09744	...	26.50	98.87	567.7	0.2098	
4	20.29	14.34	135.10	1297.0	0.10030	0.13280	0.1980	0.10430	0.1809	0.05883	...	16.67	152.20	1575.0	0.1374	

제출 기한 및 제출 방법

- **제출 기한**
  - **4월 29일 수업시간 종료 전**
  - **대면 제출을 원칙으로 하며, 비대면 참가자는 메일로 발송**
  - **비대면 참가자: jieunson@korea.ac.kr 주소로 4월 29일 19시 15분까지 제출**
  - **기한 이후 제출은 인정되지 않음**

- **제출 방법**
  - **양식 따로 없음 (자율 양식, 분량 제한 없음)**
  - **대면 제출: 프린팅 된 하드카피 양식으로 제출**
  - **대면 제출 시 맨 앞장에는 [데이터마이닝] 개인과제_이름_학번 기재**
  - **비대면 제출: 파일 첨부하여 메일로 제출**
  - **비대면 제출시 메일 제목은 [데이터마이닝] 개인과제_이름_학번**