데이터 분석 실습

김지선

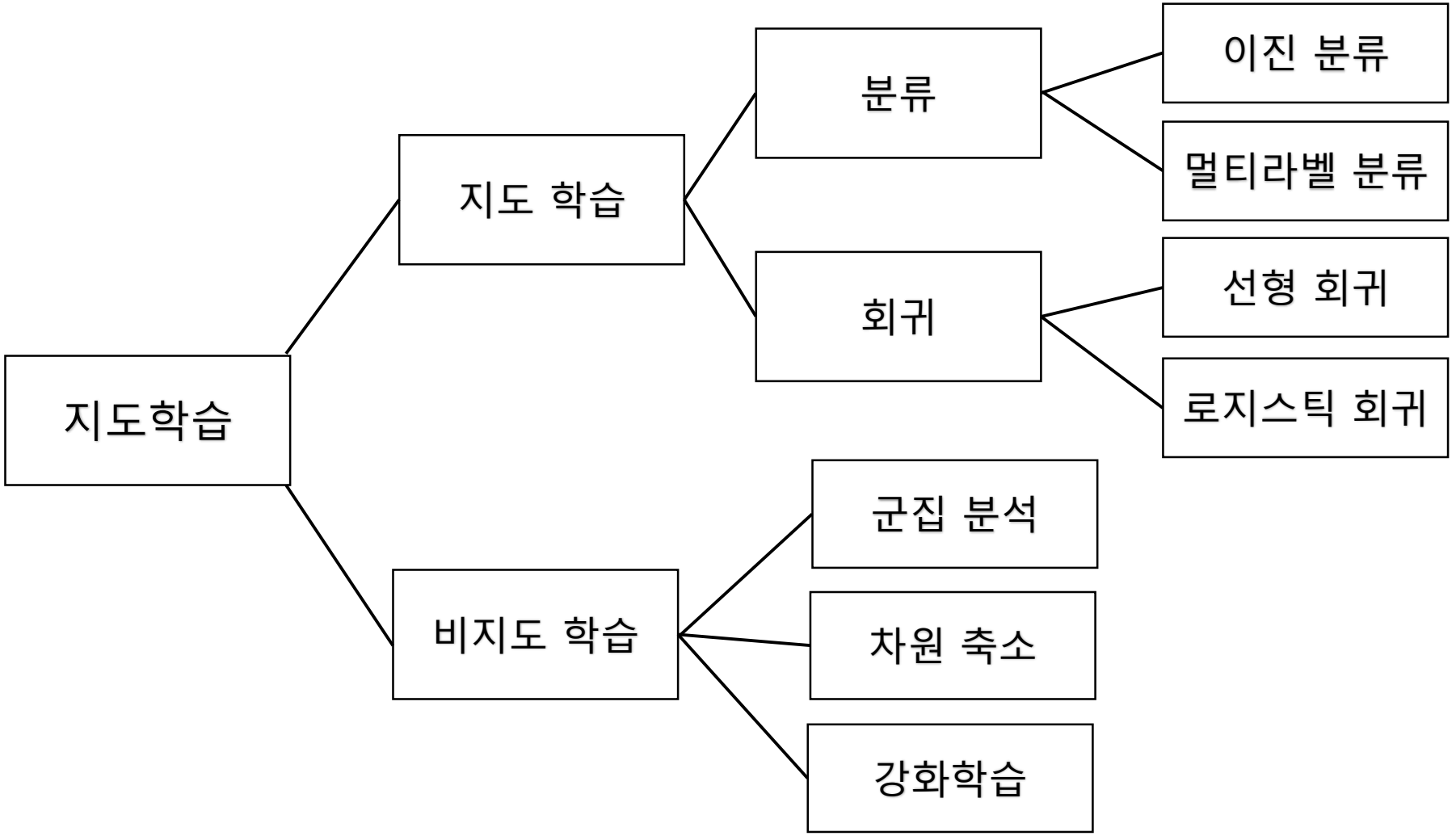
Contents

- 용어정리
- 판다스 결측치 처리
- 판다스 유사도 및 특성조합
- 데이터 분석 실습

머신러닝의 주요 패키지

넘파이 - 선형대수 처리에 유용(속도가 매우 빠름)
판다스 - 2D 배열의 데이터 분석에 주로 쓰임 /
사이킷런 - 여러 알고리즘 및 모델링에 유용한 메서드
맷플롯립 – 시각화 라이브러리

머신러닝의 전반적인 종류



+ 추천 시스템, 텍스트(자연어) 처리

용어정리

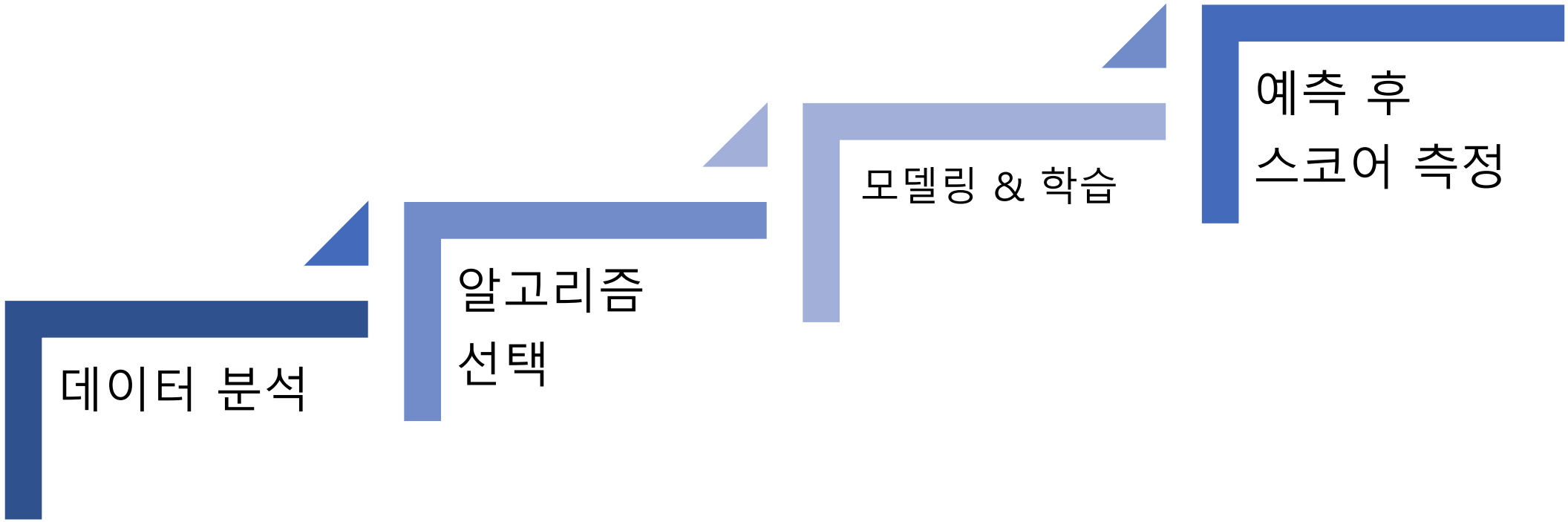
- Target – 예측하려는 목표(정답)
- Feature - 정답(타겟)을 제외한 나머지 칼럼들

Ex) 당뇨병 걸린지 예측 ⇨ 당뇨 여부가 타겟

	나이	체중(kg)	유전병 여부	당뇨 여부
0	22	54	0	?
1	43	68	1	?
2	34	80	1	?
3	64	43	1	?

Machine Learning Workflow

- 현재 데이터 분석 단계에 있고 다음 주차부터 여러 알고리즘 알아볼 예정



NumPy reshape에서 -1은?

- -1은 반대편을 고정시킨 형태로 만들어라
  ⇨ 객체.reshape(-1,n): n개의 칼럼을 고정하여 변형
  ⇨ 객체.reshape(n,-1): n개의 행을 고정하여 변형)

- reshape(-1,1) // 열벡터
- reshape(1,-1) // 행벡터

Pandas 결측치 처리 방식

- 결측 값을 특정 값으로 채우기
  ⇨ df.fillna(n)

- 결측 값을 앞 방향 혹은 뒷방향으로 채우기
  ⇨ df.fillna(method='ffill/pad') //앞방향
  ⇨ df.fillna(method='bfill/backfill') //뒷방향

- 결측 값을 변수 별 평균으로 대체하기
  ⇨ df.fillna(df.mean()[col])

Pandas 유사도 및 특성 조합 생성

- 유사도 구하기
  ⇨ 변수=df.corr()
  ⇨ 변수[칼럼]으로 해당 칼럼 기준으로 유사도 구함

- 칼럼 조합해서 새 칼럼 만들기
  ⇨ df[‘new_col’]=df[‘col1’]+df[‘col2’]

실습

- 사이킷런 내장 데이터 사용

  ⇨ from sklearn.datasets import [데이터명]

  - 아이리스 붓꽃 데이터(load_iris)
  - 당뇨병 환자 데이터(load_diabetes)
  - 보스톤 집값 데이터(load_boston)
  - 와인 데이터(load_wine)
  - 위스콘신 유방암 환자 데이터(load_breast_cancer)

Thanks