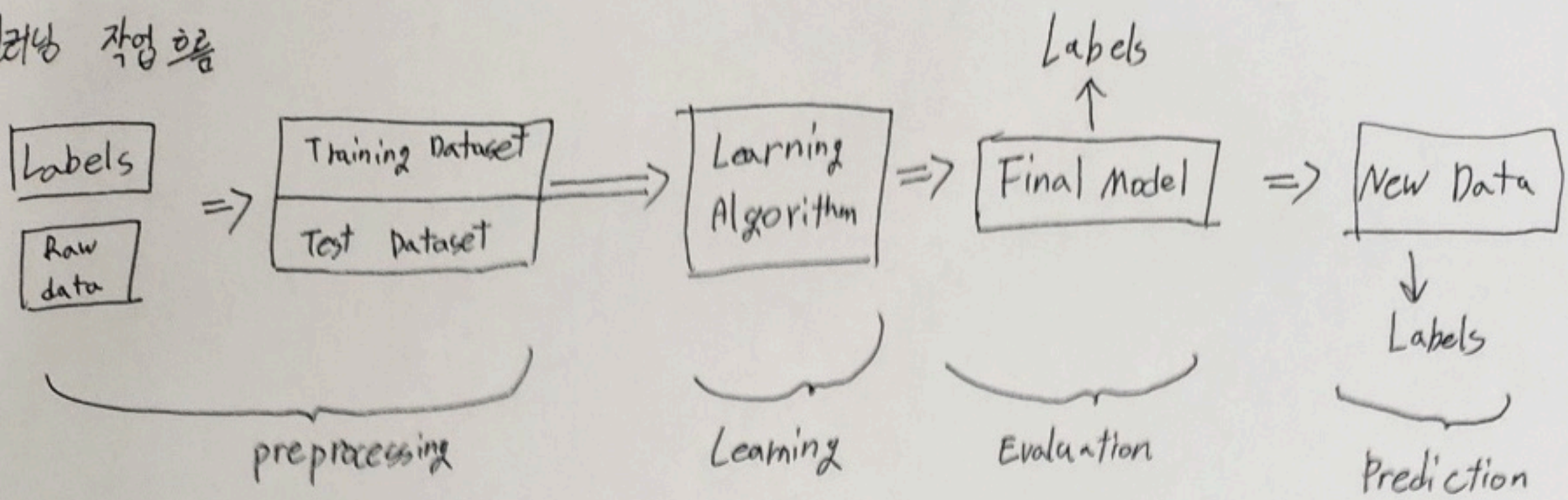
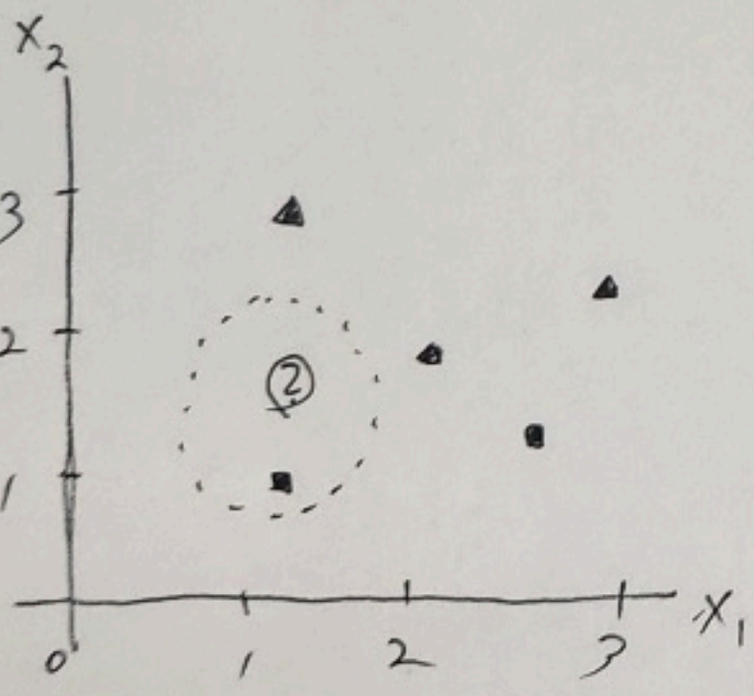
머신러닝 작업흐름

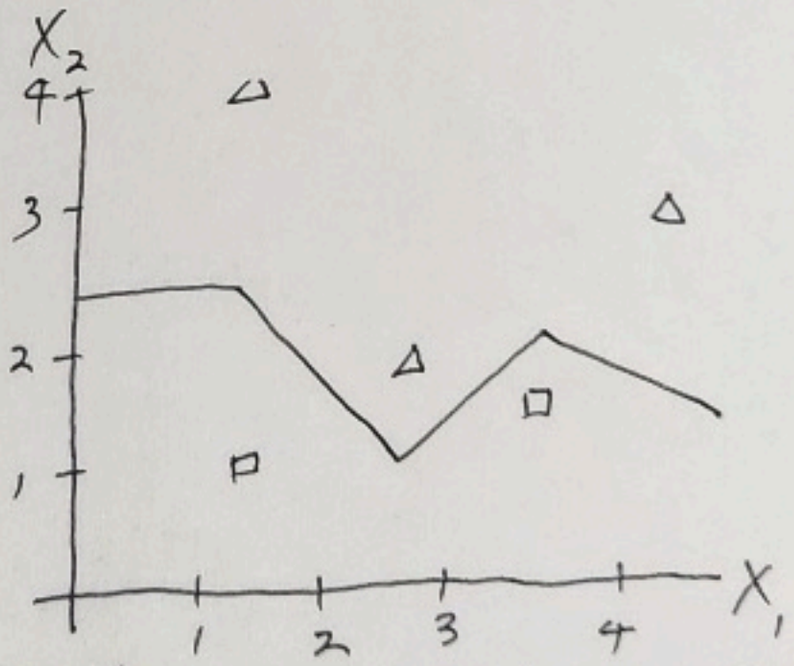


- 데이터 세트 읽기
  - CSV 파일: 표형태의 데이터를 저장하는 파일
  - Pandas : 파이썬 라이브러리, 데이터를 쉽게 다룰수 있게 도와준다.
    - lamda 함수: 이름없이 정의, 주로 간단한 연산을 수행할 때 사용
    - map : 각 요소들에게 특정한 함수를 적용시킬때 사용
    - head : 상위 5줄의 데이터 출력
    - tail : 하위 5줄의 데이터 출력
    - values : 딕셔너리의 values 값을 리턴
- 데이터 과학을 갖춘 라이브러리
  - EDA: 수집한 데이터를 다양한 각도에서 이해하고 과학적수 있게 도와준다.

- Nearest Neighbor Methods

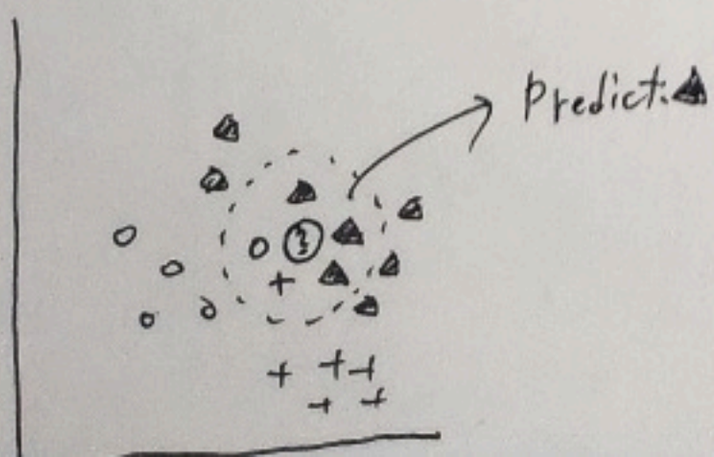


* 1-NN ; 입력값과 가장 유사한 값 1개를 결과로 선정



* Decision Boundary 1NN: 포인트간 거리를 기반으로 가장 가까운 이웃데이터를 찾아 분류

- K-Nearest Neighbors



* Hyperparameters
  ↳ k=1, k=2, k3 ...

- 기계학습 라이브러리

  - Scikit-learn ; 다양한 머신러닝 모델 전처리도구 등 다양한 기능 제공

    - 3-Nearest Neighbor classifier ; 분류문제에서 사용되는 간단한 지도학습 알고리즘
      포인트 간의 거리를 기반으로 수행된다.

    - Stratified Splits ; 각 분할에서 클래스 레이블의 분포를 유지하려는 데이터 분할 방법

    - 정규화 ; 각 데이터 포인트의 상대적 크기 보존

$$x_{norm}^{[ij]} = \frac{x^{[ij]} - x_{min}}{x_{max} - x_{min}}$$