# Data Mining Final HW

Due: 2022.06.08 24:00

1. R 패키지 'arules' 안에 있는 'Income'자료에 대한 연관규칙 분석을 수행하여라. (apriori 알고리즘 이용)

   (a) 'Income' data 불러오기

   (b) 'Income' data 변수 확인

   (c) 고소득자(income="$40,000+") 그룹에 대한 itemFrequencyPlot를 그리고 설명하여라.

   (d) 연관규칙분석 : (rhs) 고소득자에 대한 연관규칙을 신뢰도 기준 상위 5개 추출하고 설명하여라. (단 최소지지도 0.1, 최소신뢰도 0.8, 향상도 1.0 적용)

2. 'sample_DT.csv'는 제품의 불량 여부(DEFECT_TYPE)를 분류하기 위한 데이터이다. 각 변수들은 제품을 생성하는 공정에서 관측된 값들이다. 이를 분류하기 위한 모형 적합을 하여라.

   (a) 모형 검증을 위해 training data 와 testing data로 나누시오.
   (단, training data : testing data = 7:3, seed=1234)

   (b) 의사결정나무(가지치기 실행)

   (c) 배깅

   (d) 부스팅

   (e) 랜덤포레스트

   (f) 위에서 적합한 모형 중 가장 좋은 모형을 선택하시오.

3. 'Wholesale customers data.csv'는 다양한 상품군에 대한 고객의 구매이력에 대한 자료이다. 다음에 대하여 다양한 방법으로 군집분석을 시행하여라.

   - 변수설명

     1) FRESH: annual spending on fresh products (Continuous)

     2) MILK: annual spending on milk products (Continuous)

     3) GROCERY: annual spending on grocery products (Continuous)

     4) FROZEN: annual spending on frozen products (Continuous)

     5) DETERGENTS_PAPER: annual spending on detergents and paper products (Continuous)

     6) DELICATESSEN: annual spending on and delicatessen products (Continuous)

- 정답은 없습니다. 다양한 군집 분석 알고리즘을 적용해보고, 데이터 분석을 해보는 것이 목표입니다.

  데이터에는 이상치가 존재합니다. 가능하면 이상치를 제거하는 단계를 포함시키는 것이 필요합니다. 이상치가 있는 경우와 없는 경우의 분석 결과 차이를 비교해 보는 것도 좋을 것 같습니다.

  실제 변수는 8개입니다. 위에 제시한 6개의 변수만을 이용하여 분석합니다.