빅데이터와 금융자료분석 기말대체과제

1. prob1_bank.csv 자료는 어느 포르투갈 은행의 정기예금 프로모션 전화 데이터이다. 이 데이터는 고객의 특징을 나타내는 특성 변수들과 고객이 정기예금에 가입했는지 여부를 나타내는 목표 변수로 구성되어 있다.

<특성변수> age : 나이 job : 직업의 형태 marital : 결혼 상태 education : 학력 default : 신용 불이행 여부 balance : 은행 잔고 housing : 부동산 대출 여부 loan : 개인 대출 여부 contact : 연락 수단 month : 마지막으로 연락한 달 <목표변수> y : 고객이 정기 예금에 가입했는지 여부

(1) 주어진 자료 중 범주형 변수 각각에 대해 적절한 전처리를 선택하고 진행하여라.

(2) 주어진 자료 중 수치형 변수 각각에 대해 적절한 전처리를 선택하고 진행하여라.

(3) 주어진 자료에 클래스 불균형이 있는지 확인한 뒤, 이에 대한 적절한 전처리 방법을 선택하여 진행하여라.

2. prob2_card.csv 자료는 어느 신용카드 회사의 고객 데이터로, 신용카드 사용 형태를 나타내는 여러 특성 변수들로 구성되어 있다.

CUST_ID : 신용카드 사용자 ID BALANCE : 구매 계좌 잔액 BALANCE_FREQUENCY : 구매 계좌 잔액이 업데이트 되는 빈도 지수로, 0(자주 업데이트 되지 않음)~1(자주 업데이트 됨) 사이의 값을 가짐. PURCHASES : 구매 계좌로부터의 구매액 PURCHASES_FREQUENCY : 구매 빈도 지수로, 0(자주 구매하지 않음)~1(자주 구매함) 사이의 값을 가짐. PURCHASES_TRX : 구매 거래 건수

(1) 주어진 자료에 K평균 Clustering 알고리즘을 적용하여, 적절한 군집을 생성하여라.

(2) 주어진 자료에 DBSCAN Clustering 알고리즘을 적용하여, 적절한 군집을 생성하여라.

(3) (1)과 (2)의 두 군집 분석 결과를 비교하고, 더 타당한 모델을 선택하여라.

(4) (3)에서 선택된 최종 모델로 생성한 군집들의 고객 특성을 분석하여라.

(5) t-SNE 알고리즘을 적용하여 주어진 자료를 2차원으로 축소하여라. 그 결과를, (3)에서 선택한 모델의 군집 레이블에 따라 점의 색상이 다르게 표현된 2차원 산점도로 시각화하여라.

3. 다음 데이터에서 특성변수인 X는 방의 개수, 목표변수인 Y는 주택 가격을 나타낸다.

ID	X	Y
1	3	1.25
2	1	1.2
3	2	1.3
4	4	1.5
5	?	1.4
6	?	1.3

(1) XGBoost 알고리즘을 적용하여 트리를 생성한다고 할 때, 첫번째 트리의 첫 마디에서 최적의 분리기준이 무엇인지를 구하여라. 단, 결측이 아닌 4개의 관찰치(ID 1~ID 4)만 이용할 것. 제곱오차 손실함수를 적용하며, 모델 초기값 $\hat{f}_0$는 0.5로 두고, 규제 하이퍼 파라미터 λ는 0으로 설정할 것. 또한 계산 과정을 상세하게 서술할 것.

(2) XGBoost 알고리즘을 적용하여 트리를 생성한다고 할 때, 첫번째 트리의 첫 마디에서 X의 값이 결측인 경우는 왼쪽과 오른쪽 자식마디 중 어느 쪽으로 보내야 할지를 결정하여라.