데이터 분석 및 시각화 과제 보고서 작성

A4팀(김현주, 김승현, 정성수, 임성구)
B4팀(양정연, 신정훈, 김용현, 조성운)
C4팀(조아진, 신유라, 이해동, 권호영)

팀별로 다음 과제를 수행하고 결과와 R code를 포함한 과제보고서를 제출하시오.

(시계열분석/시각화)

1. USD/KRW 환율 데이터 (https://kr.investing.com/currencies/usd-krw-historical-data)에 있는 일별 환율데이터(2021년 11월 15일 ~ 2022년 11월 14일 대상)를 기반으로 회귀분석 또는 시계열분석을 이용하여 2022년 12월 우리나라 미국달러대비 원화 환율을 예측하고 시각화하시오.

참고사항:
* 회귀분석 실시 시 기본가정 충족 여부 확인 포함.
* 기본가정 불충족 시에는 충족시킬 수 있는 방법을 제시하고 실행.
* 제시한 방법으로도 기본가정 불충족 시에는 회귀선을 구하여 예측하고 comment 하시오.

배점: 20점
난이도: 4

(분류분석/시각화)

2. 위스콘신 유방암 데이터셋을 대상으로 분류기법 2개를 적용하여 기법별 결과를 비교하고 시각화하시오. (R과 python 버전으로 모두 실행)
-종속변수는diagnosis: Benign(양성), Malignancy(악성)

배점: 20점
난이도: 4

(예측기법/시각화)

3. mlbench패키지 내 BostonHousing 데이터셋을 대상으로 예측기법 2개를 적용하여 기법별 결과를 비교하고 시각화하시오. (R과 python 버전으로 모두 실행)

-종속변수는MEDV 또는CMEDV를사용

배점: 20점
난이도: 4

(군집분석/시각화)

4. 아래의 조건을 고려하여 군집분석을 실행하시오.

(1) 데이터: ggplot2 패키지 내 diamonds 데이터
(2) philentropy::distance() 함수 내 다양한 거리 계산 방법 중 Euclidian거리를 제외한 3개를 이용하여 거리 계산 및 사용된 거리에 대한 설명
(3) 탐색적 목적의 계층적 군집분석 실행
(4) 군집수 결정 및 결정 사유 설명
(5) k-means clustering 실행
(6) 시각화
(7) 거리 계산 방법에 따른 결과 차이 비교

배점: 10점
난이도: 3

(텍스트분석/시각화)

5. 제공된 데이터를 대상으로 텍스트 분석을 실행하시오.

데이터:

(1) 제공된 데이터를 이용하여 토픽 분석을 실시하여 단어구름으로 시각화 하고 단어 출현 빈도수를 기반하여 어떤 단어들이 주요 단어인지 설명하시오
(2) 제공된 데이터를 이용하여 연관어 분석을 실시하여 연관어를 시각화 하고 시각화 결과에 대해 설명하시오

배점: 10점
난이도: 3

(데이터 시각화)

6. R의 ggplot2 패키지 내 함수와 python의 matplotlib 패키지 내 함수를 사용하여 막대 차트(가로, 세로), 누적막대 차트, 점 차트, 원형 차트, 상자 그래프, 히스토그램, 산점도, 중첩자료 시각화, 변수간의 비교 시각화, 밀도그래프를 수업자료pdf 내 데이터를 이용하여 각각 시각화하고 비교하시오.

배점: 20점
난이도: 4

제출기한: 2022년 12월 2일(금) 14:30
제출처: 카페 내 과제제출 게시판
제출물: R코드, python 코드 가 포함된 과제보고서
제출양식: 자유(워드, 한글, PPT 등)
발표는 없음