머신러닝을 이용한 교내 버스정류장 신설 위치 선정

키워드	CA	ML	PCA	R

📌 주요 내용

- 분석 배경
  - 캠퍼스 특성상 건물 이동에 있어 학교 순환 버스가 필수적이나 현재 노선의 불균형으로 많은 학우들이 불만을 나타내고 있음
- 분석 목적
  - 통계적 이론을 바탕으로 머신러닝을 이용하여 최적의 버스 정류장 신설 위치를 제안하고자 연구 진행
- 분석 내용
  - 설문조사를 통해 학생들이 자주 사용하는 건물과 희망 정류장 위치 조사
  - 학교 캠퍼스를 격자별로 나누고 격자별 시설의 빈도를 측정하여 데이터셋 구성
  - 주성분 분석 후 군집 분석 시행
  - 군집 분석 결과와 주성분 행렬도를 연결하여 분석
  - 분석 결과를 통한 버스 정류장 신설 위치 선정

분석 배경

- 캠퍼스 특성상 건물 이동에 있어 학교 순환 버스가 필수적이나 현재 노선의 불균형으로 많은 학우들이 불만을 나타내고 있음

분석 목적

- 통계적 이론을 바탕으로 머신러닝을 이용하여 최적의 버스 정류장 신설 위치를 제안하고자 연구 진행

데이터 수집

- 최적의 버스 정류장 신설 위치 검출을 위해 학교 학생들을 대상으로 구글폼을 활용하여 두 차례의 설문 진행
- 1차 설문조사 결과
  - 순환버스 만족도, 문제점, 해결방안 조사 / 총 127명의 학생이 설문에 응답하였음
  - 만족도 조사
    - 순환 버스에 어느 정도 만족하나요? 라는 질문에 불만족과 매우 불만족이 전체의 41.7%로 높은 비율을 차지함
  - 교내 버스 문제점 조사
    - 단답형 질문 형식을 이용함. 응답 결과를 수집하여 텍스트 파일로 변환시킨 후, 워드클라우드 분석 시행
      - R프로그램을 사용하여 응답 결과를 최소 단위의 형태소로 분리하여 워드 클라우드 생성
      - ‘노선’, ‘배차’, ‘운행’등의 키워들의 빈도수가 높게 나타남. 이를 통해 많은 학생들이 노선에 큰 불만을 가지고 있다는 근거를 확보함.

- 순환버스의 문제점 해결방안 조사
  - 단답형 질문 형식을 이용함. ‘이용 빈도를 기반으로 한 노선 개편 및 신설’, ‘노선의 이원화’, 배차간격 축소 및 정확한 시간표 게시’등의 의견이 다수를 이루었음.

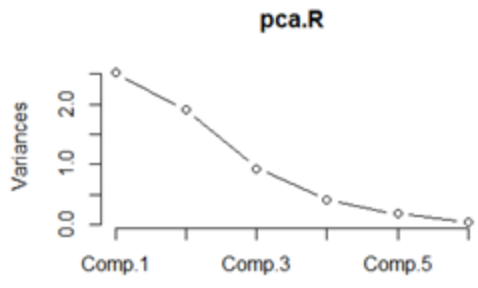
- 2차 설문조사 결과
  - 순환버스 노선에 대한 세부적인 만족도 조사
  - 소속과건물, 이용빈도 1순위 건물, 이용 빈도 2순위 건물, 거주지 인접 건물, 희망 정류장 위치에 대한 설문

- 데이터 생성
  - 학교 캠퍼스 지도를 적절한 크기의 격자로 나눈 후 도로가 있는 곳에 격자 번호를 부여
  - 격자에 위치한 시설들의 빈도를 측정하여 데이터셋 구성 (행 : 격자, 열 : 소속과건물, 이용빈도 1순위 건물, 이용빈도 2순위 건물 등, 값 : 빈도)

- 데이터 전처리
  - 변수 결합
    - 대형마트, 우체국, 스터디카페, 편의점, atm, 카페, 음식점, 학생시설의 변수들의 빈도가 다른 변수들에 비해 매우 작은 관계로 의미있는 결과를 도출해내기 어렵다고 판단. 따라서 이 변수들을 모두 하나의 변수(편의시설)로 결합함.
    - 최종적으로 희망정류장, 과건물, 이용빈도 1순위 건물, 이용빈도 2순위 건물, 편의시설, 거주지 변수 사용
  - 데이터 스케일링
    - z-점수 정규화는 min-max 정규화보다 이상치를 잘 처리하는 방법이기 때문에 해당 스케일링 방법 이용

분석 내용

- 주성분 분석을 통해 얻어진 결과를 군집분석에 활용
- 주성분 분석
  - 주성분 개수 선택
    - 첫 번째 주성분에 의한 설명력은 약 50.312%, 두 번째 주성분에 의한 설명력은 약 27.042%
    - 두개의 주성분에 의한 설명 비율의 합이 70%이상이므로 2개의 주성분만을 이용해도 원자료의 정보의 손실이 많지 않게 되며 원 변수들을 2개의 주성분으로 대신할 수 있음.
    - 스크리 그림에서 팔꿈치가 3에서 발생하였기 때문에 주성분을 2개 내지는 3개 정도 선택할 수 있음.



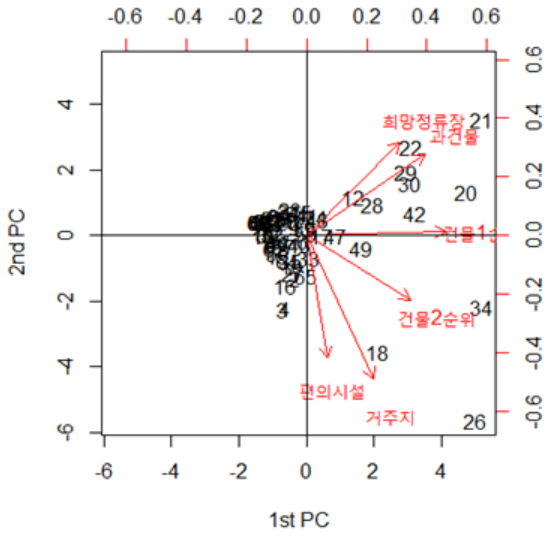
  - 주성분 계수 해석
    - 주성분 계수는 해당 원 변수가 주성분을 만드는 데에 기여한 정도를 나타냄.
    - 제 1 주성분
      - 과건물, 이용빈도 1순위 건물, 이용빈도 2순위 건물의 주성분 계수가 양(+)이며 양의 방향으로 큰 값을 가짐. 따라서 제 1주성분은 학교 내 시설을 나타내는 주성분

        - 제 2주성분
            - 거주지, 편의 시설의 주성분 계수가 음(-)이며 음의 방향으로 큰 값을 가짐. 따라서 제 2 주성분은 교외 시설을 나타내는 주성분

- 군집 분석
    - K-means 방법을 이용한 군집 분류 결과를 PCA의 biplot(주성분 행렬도)과 연결하여 분석
    - K-means를 이용한 군집분석 결과

	군집1	군집2	군집3
격자	18,26,34	12,20,21,22, 29,30,42	나머지

    - PCA의 주성분 행렬도



    - 첫번째 군집에 해당하는 개체들은 행렬도 상에서 편의시설과 거주지 변수의 화살표 방향에 위치하므로 첫번째 군집은 거주지와 편의시설의 값이 크다는 특징을 가진다.
    - 두번째 군집에 해당하는 개체들은 행렬도 상에서 과건물, 건물 이용빈도 1순위, 2순위 변수의 화살표 방향에 위치하므로 두번째 군집은 과건물, 건물 이용빈도 1순위, 2순위의 값이 크다는 특징을 가진다.
    - 마지막으로 세번째 군집은 어떠한 변수의 값도 크거나 작게 나타나지 않아 특징적인 값을 갖지 않는다.
    - 따라서 세번째 군집에 해당하는 격자에는 정류장 설치를 고려하지 않아도 된다는 결론을 내릴 수 있다.

결론

- 먼저 학생들의 거주지와 편의시설에 대한 접근성을 고려한다면, 첫번째 군집이 가리키고 있듯이 순환버스가 도서관과 여자 기숙사 사이를 경유하도록 정류장을 설치하는 것이 합당하다고 판단됨
- 다음으로 학생들의 과건물과 자주 이용하는 건물에 대한 접근성을 고려한다면, 두번째 군집이 가리키고 있듯이 공학관, 인문관 근처에 순환버스 정류장을 추가 설치하는 것이 합당하다고 판단된다.