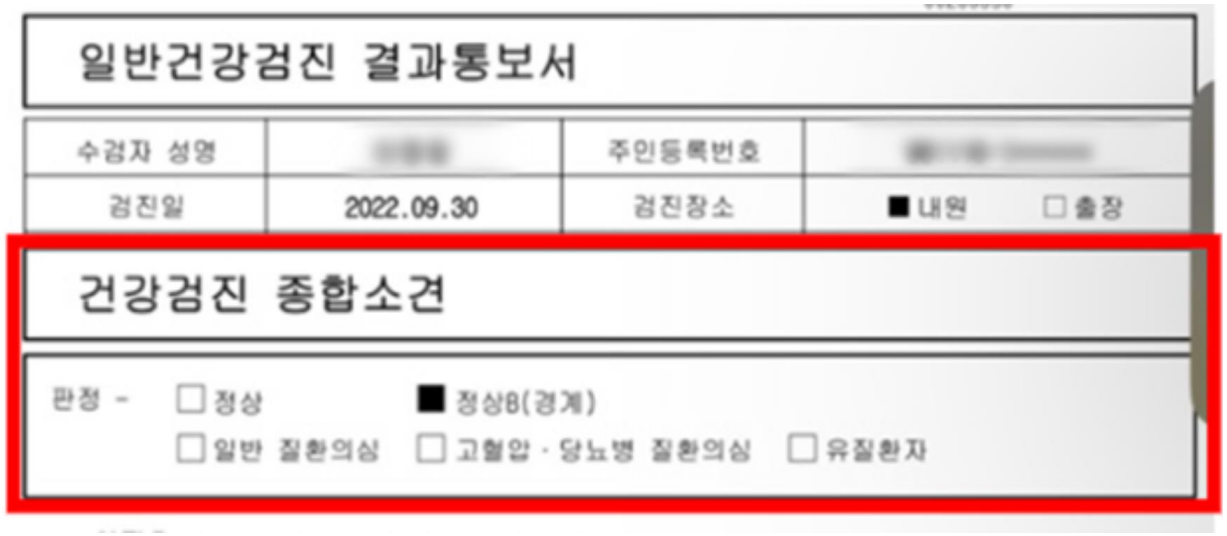
건강검진 이해도 향상을 위한

환자 집단별 특성 파악 모델 개발

2조 김정준, 박상은, 박민규, 박정빈, 이현재

분석 배경

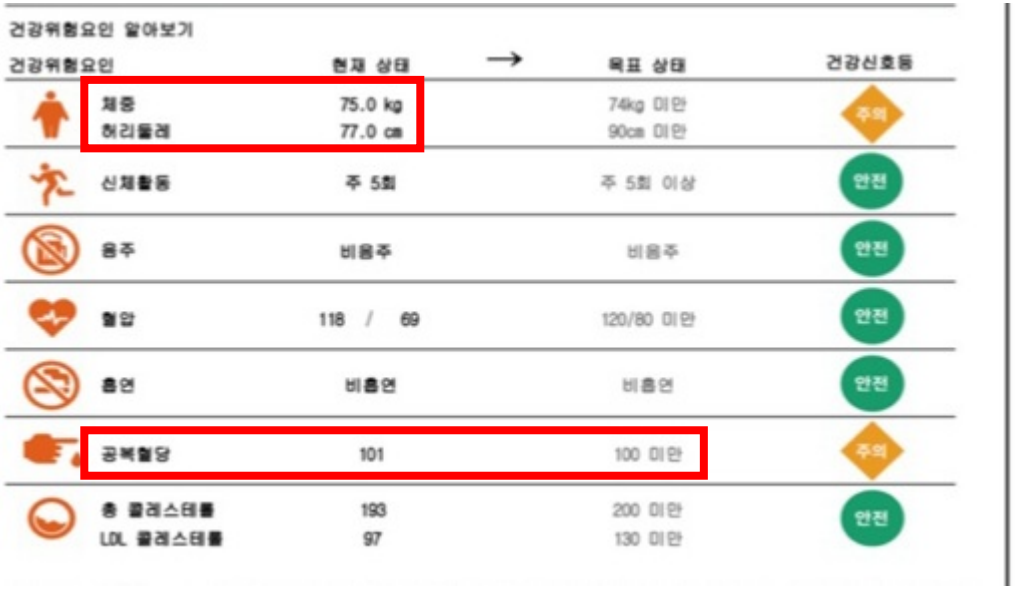


일반건강검진 결과통보서

수검자 성명		주민등록번호	
검진일	2022.09.30	검진장소	■ 내원 □ 출장

건강검진 종합소견

판정 - □ 정상 ■ 정상B(경계)
□ 일반 질환의심 □ 고혈압 · 당뇨병 질환의심 □ 유질환자



건강위험요인 알아보기

건강위험요인	현재 상태	→	목표 상태	건강신호등
체중	75.0 kg		74kg 미만	주의
허리둘레	77.0 cm		90cm 미만	
신체활동	주 5회		주 5회 이상	안전
음주	비음주		비음주	안전
혈압	118 / 69		120/80 미만	안전
흡연	비흡연		비흡연	안전
공복혈당	101		100 미만	주의
총 콜레스테롤	193		200 미만	안전
LDL 콜레스테롤	97		130 미만	

▲ 현 건강검진 결과통보서 형태

분석 배경

직장인 10명 중 7명, 건강검진 결과지 "이해 못했다" 1)

장진숙 기자 | 승인 2020.03.26 13:17 | 댓글 0

반면 직장인들의 건강검진 결과 이해도는 낮은 수준이었다. 결과지 내용을 충분히 이해하지 못한 응답자가 71%를 차지했다. 이해하지 못한 이유로 '내 수치에 대한 자세한 설명'(44%) 응답이 가장 많았고 '어려운 용어'(40%), '복잡한 항목'(15%) 등이 뒤를 이었다.

1)

"건강검진 이상소견 발견해도 치료 연계 느슨...사후관리 강화 필요" 2)

건보공단 국가검진 효율화 방안 연구, 의료진 수가·검진항목 조정 제안

기사입력시간 21-01-16 09:50

건강검진을 통해 만성질환을 조기에 발견하고 사망률까지 감소시킬 수 있으나, 이상소견이 있다고 해도 의료기관 방문 등 치료 연계가 잘 이뤄지지 않았다.

16일 국민건강보험공단은 국가건강검진의 효율적 실시를 위한 심층분석 및 개선방안 연구(연구책임자 울산의대 기초의학교실 조민우 교수)를 시행한 결과에 따르면, 건강검진의 사후관리체계 구축의 필요성이 대두됐다.

1) 장진숙, '직장인 10명 중 7명, 건강검진 결과지 "이해 못했다"', 데이터솜, 2020.03.26
2) 정승원, "건강검진 후 이상소견 발견해도 치료 연계 느슨...사후관리 강화 필요", MEDI:GATE NEWS, 2021.01.16

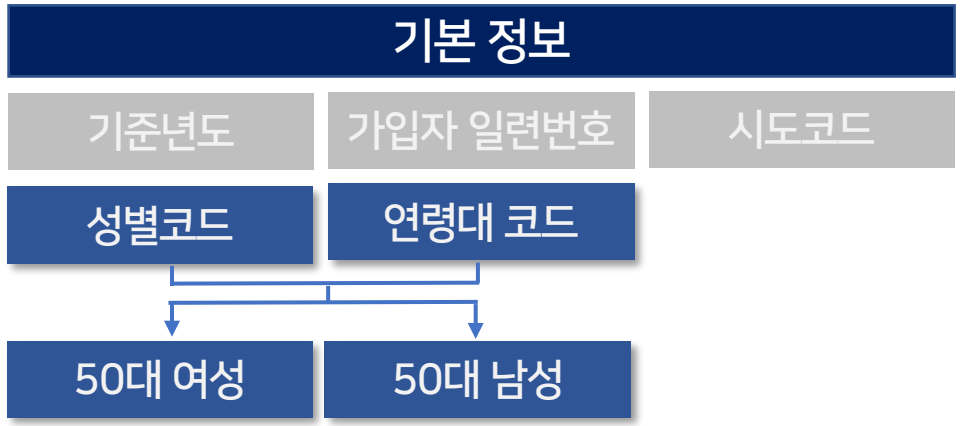
문제 정의

건강검진 데이터 군집화를 통해 그룹별 특성 파악

- **기존 건강검진 결과지의 문제점**
  - **모호함**
    - 정상과 경계(정상b)를 나누는 기준이 명확하지 않음
    - 판정 결과가 구체적이지 않음
    - 자신의 건강상태가 어느 위치에 속하는지 알 수 없음
  - **후속조치가 없음**
    - 건강검진 후 어떠한 진료를 받아야 하는지 알 수 없음
- **건강검진 결과 해석이 중요한 부모님 세대(50대)를 타겟으로 하여 분석 진행**

사용 데이터

- 공공데이터 포털의 **국민건강보험공단 건강검진정보 데이터** 사용
  - 40세 이상의 각 연도별 진료 및 건강검진 수진 환자 중100만명을 무작위 추출한 데이터
  - 기본정보(성, 연령대, 시도코드 등)과 검진내역(신장, 체중, 총콜레스테롤, 혈색소 등)의 30개의 feature로 구성됨
  - 성별과 연령에 따라 데이터 분할

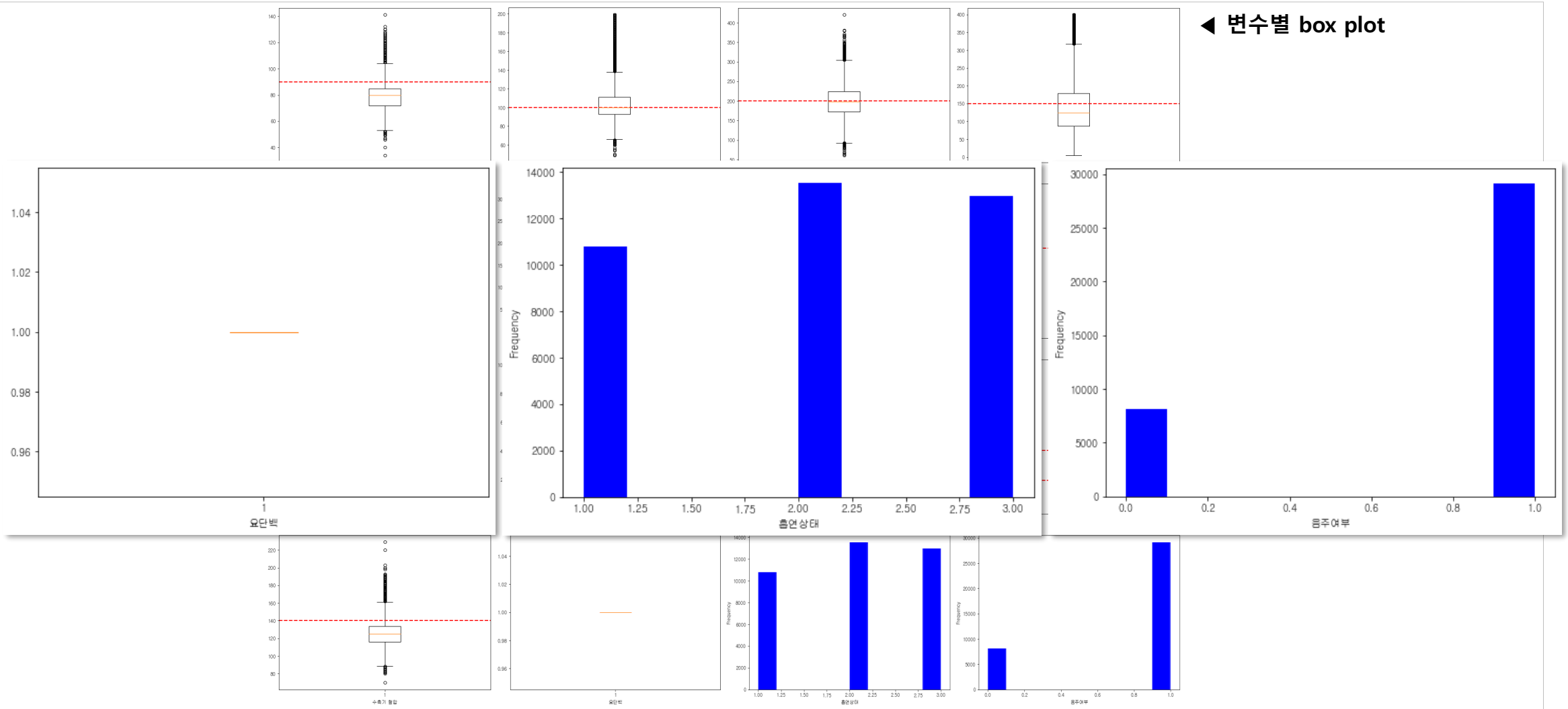


사용 데이터 상세 설명


목표 질환
검진 내역
시각이상
시력(좌)
시력(우)
간장질환
(혈청지오티)AST
(혈청지오티)ALT
감마지티피
이상지질 혈증
총 콜레스테롤
트리글리세라이드
HDL 콜레스테롤
LDL 콜레스테롤
심뇌혈관 질환
흡연여부
음주여부
이완기 혈압
수축기 혈압
구강관련
치석
구강검진 수검여부
치아우식증유무
비만
체중
신장
허리둘레
BMI
당뇨병
식전혈당
요검사
요단백
신장질환
혈청크레아티닌
빈혈
혈색소

사용 데이터 상세 설명

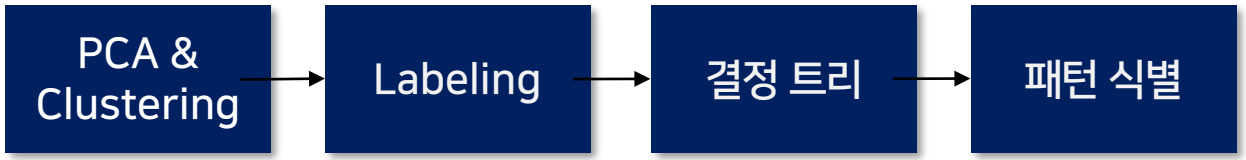
◀ 변수별 box plot



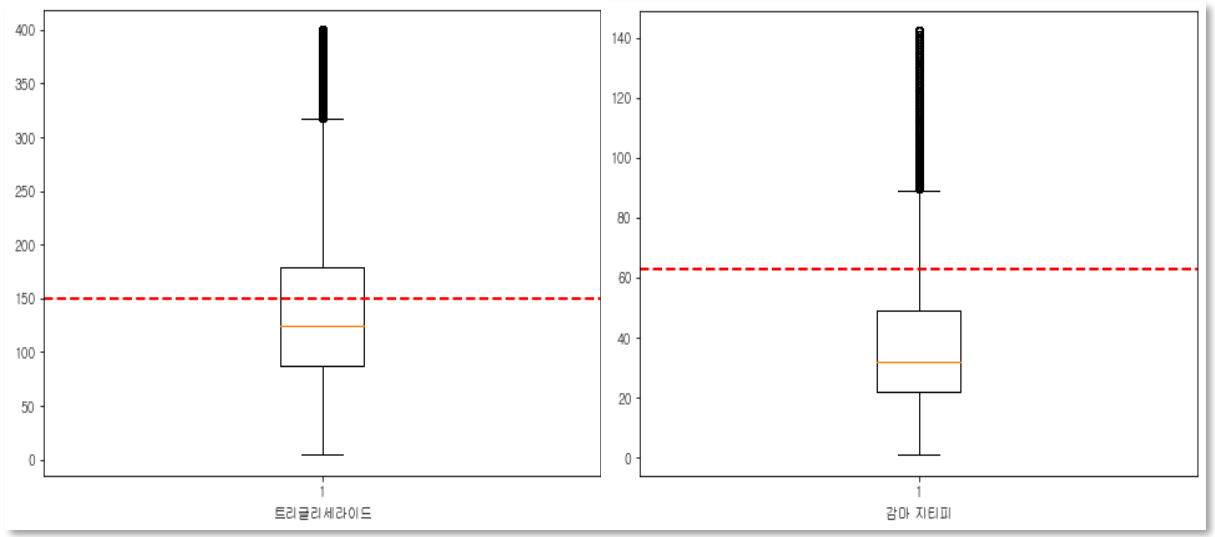
분석 과정

1. 중요 변수 식별

2. 중요 변수들만을 사용해서 그룹별 패턴 식별



01. 데이터 전처리



3) 사분위수 범위(interquartile range, IQR)를 활용한 이상치 탐색은 상자그림에서 사분위수 범위의 1.5배를 초과하는 관측치는 이상치, 3배를 초과하는 관측치는 극단적 이상치로 정의하는 방법이다. 상자그림은 최솟값, 최댓값, 제1사분위수(Q_1), 제2사분위수(Q_2), 제3사분위수(Q_3)를 활용하여 자료를 시각적으로 요약한 그래프이다. 상자그림에서 표현되는 최솟값과 최댓값은 이상치를 제외한 자료의 최솟값과 최댓값을 의미하고, 사분위수 범위는 제3사분위수와 제1사분위수의 차이를 말한다[그림 3]. 기존 사분위수 범위를 일반화한 수정된 사분위수 범위와 일부 사분위수 범위를 변형한 준사분위수 범위도 이상치 탐색 방법으로 사용된다.

3) 김진휘. "이상치 탐색을 위한 통계적 방법." HIRA 정책동향 14.1 (2020): 49-56.

02.차원 축소

Index	설명가능한 분산 비율(고윳값)	기여율	누적기여율
pca1	2.54965	0.196122	0.196122
pca2	2.00411	0.154158	0.35028
pca3	1.68946	0.129955	0.480235
pca4	1.38285	0.10637	0.586606
pca5	0.998701	0.0768212	0.663427

4)

있다. 이러한 5개 공통 요인에 의해 설명되는 설명력은 전체 분산의 50.45%를 설명하고 있다. 5개 요인별 크론바하 알파계수를 산출한 결과 0.5 이상으로 나타나 각 요인들의 신뢰도는 양호하다.

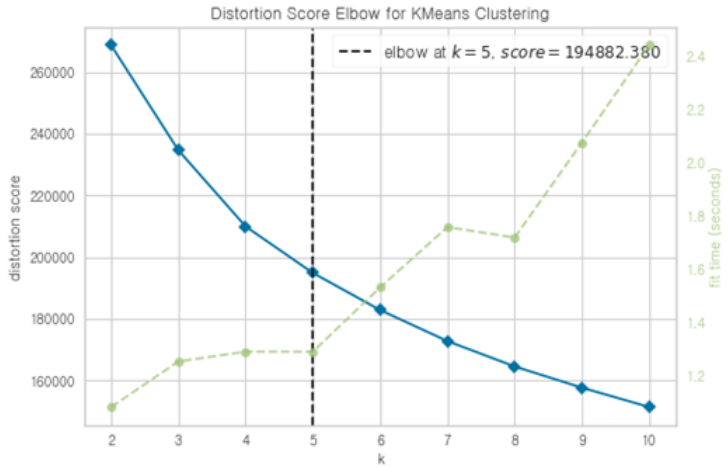
식품소비 라이프스타일에 따라 유형을 구분하기 위해 요인분석 결과 도출된 5개 요인점수를 이용하여 K-평균 군집분석을 실시한 결과, 5개 유형의 군집으로 결정하였다. 각 군집별 성향을 파악하여 명명하기 앞서, 각 군집유형별로 식품소비 라이프스타일 5개 요인의 평균이 차이가 있는지를 파악하기 위해 분산분석[2]을 수행하였다. 군집 1에

Index	0	1	2	3
수축기 혈압	0.358132	-0.0471559	-0.452415	0.344841
이완기 혈압	0.353679	-0.0159983	-0.440532	0.37125
식전혈당(공복혈당)	0.244904	-0.0784759	-0.0914996	-0.151341
총 콜레스테롤	0.112772	0.685982	0.0182545	-0.0268606
트리글리세라이드	0.31162	0.0324829	-0.082183	-0.501414
HDL 콜레스테롤	-0.185395	0.19941	0.0973624	0.564059
LDL 콜레스테롤	0.105363	0.657726	0.0049412	-0.107693
혈색소	0.242028	0.108754	0.0471881	0.111976
혈청크레아티닌	0.0221894	0.0899994	0.0921038	-0.0063217
(혈청지오티)AST	0.268531	-0.0610755	0.52345	0.265913
(혈청지오티)ALT	0.40113	-0.118917	0.456578	0.103804
감마 지티피	0.357709	-0.0318857	0.252763	-0.0373257
BMI	0.333673	-0.0924781	-0.138663	-0.207364

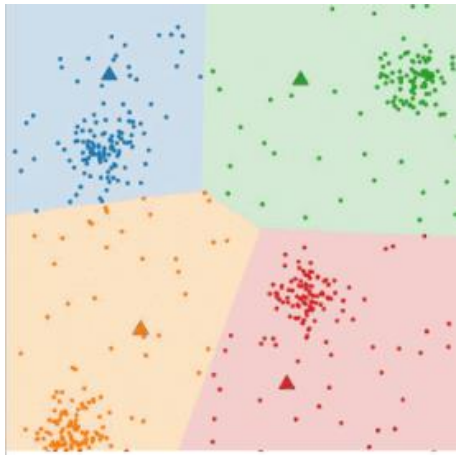
4) 박미성 and 안병일. (2014). 식품소비 라이프스타일이 가공식품 지출에 미치는 효과 분석: 군집분석과 매칭 기법을 이용하여. 농촌경제, 37(3), 25-58.

03. 1차 클러스터링

- **K-평균 알고리즘**
  - 클러스터의 중심과 데이터 사이의 평균 거리를 기반으로 K개의 클러스터 구성



◀ 클러스터 개수 결정



5)

◀ K-평균 알고리즘

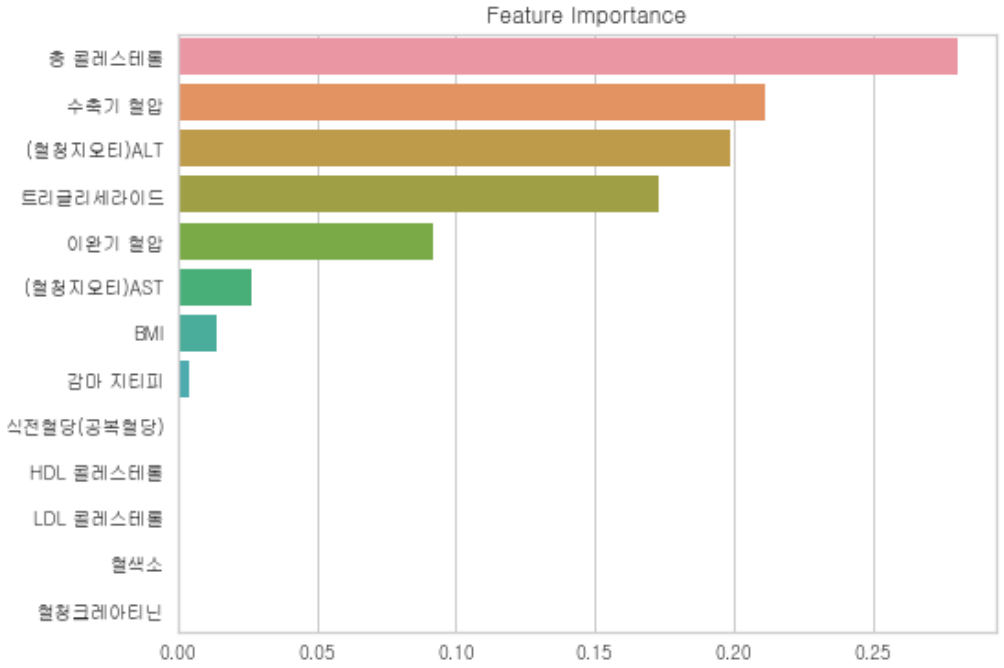
- 차원 축소된 데이터에 K-평균 알고리즘 적용

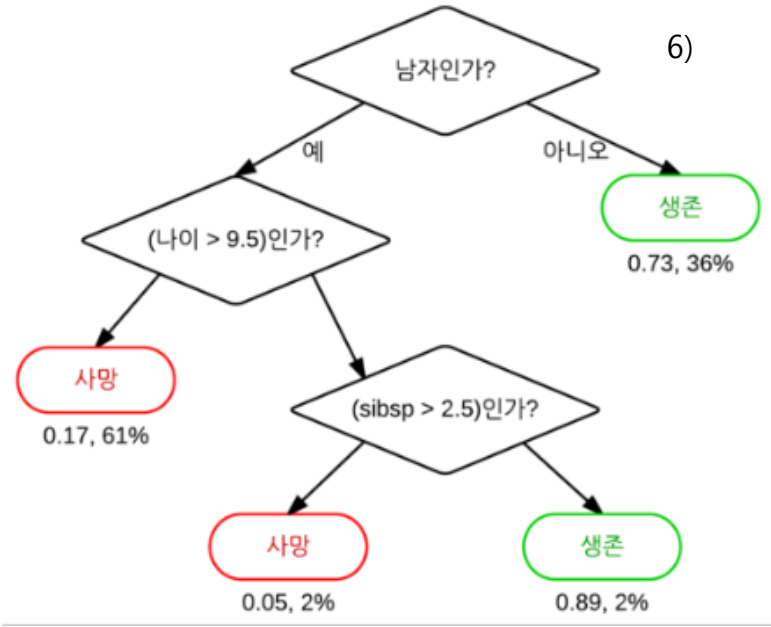
-> **그룹별 특성 파악, 레이블 지정**

5) https://upload.wikimedia.org/wikipedia/commons/thumb/7/7b/Kmeans_animation_withoutWatermark.gif/330px-Kmeans_animation_withoutWatermark.gif

04. 중요 변수 식별

- **결정 트리 학습법**
  - 일련의 분류 규칙을 통해 데이터를 분류하는 지도 학습 모델
  - 불순도를 이용해 학습하고 **중요 변수 식별 가능**



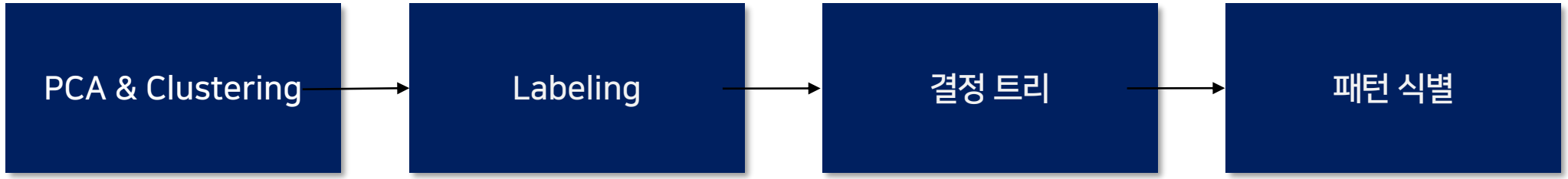


6) ◀ **결정 트리 학습법**

6) https://upload.wikimedia.org/wikipedia/commons/thumb/f/fe/CART_tree_titanic_survivors_KOR.png/525px-CART_tree_titanic_survivors_KOR.png

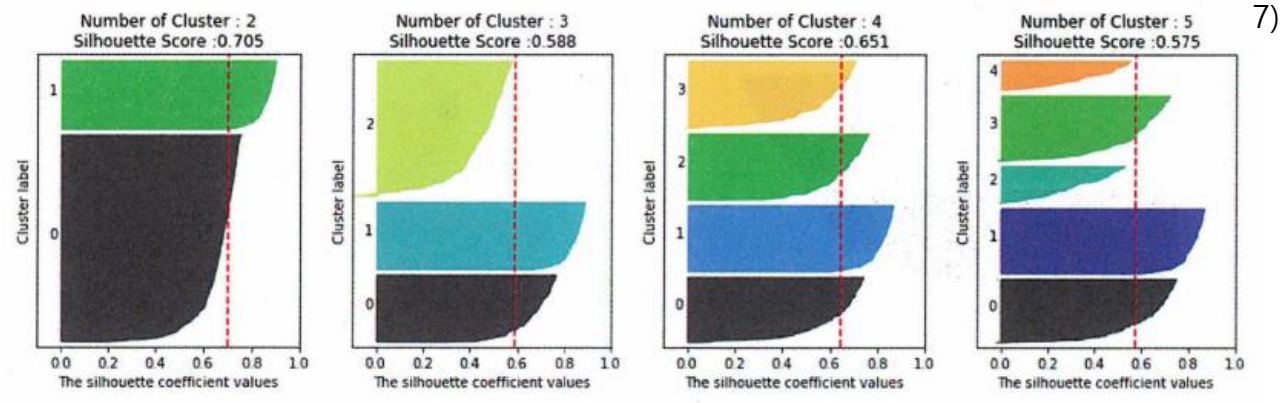
05. 2차 클러스터링 & 학습

- 중요 변수들만을 사용해서 그룹별 패턴 식별



06.검정

- 실루엣 계수
  - 해당 데이터가 같은 군집 내의 데이터와 얼마나 가깝게 군집화돼 있고, 다른 군집에 있는 데이터와는 얼마나 멀리 분리돼 있는지 나타내는 지표



7)

- 분산분석(ANOVA)
  - 평균 값과 분산을 이용하여 다수 집단의 평균 값에 차이가 있는지를 가설검정

7) 파이썬 머신러닝 완벽 가이드, 권철민

기대효과

- **기존 건강검진 결과지의 문제점**
  - **모호함**
    - 정상과 경계(정상b)를 나누는 기준이 명확하지 않음
    - 판정 결과가 구체적이지 않음 -> **자신이 어떤 그룹에 속하는지 파악 가능**
    - 자신의 건강상태가 어느 위치에 속하는지 알 수 없음
  - **후속조치가 없음**
    - 건강검진 후 어떠한 진료를 받아야 하는지 알 수 없음 -> **그룹별 특성을 통해 맞춤 진료 제시**

검진자의 건강상태 진단 후
맞춤 진료를 제시하여 건강검진 이해도 및 활용성 향상