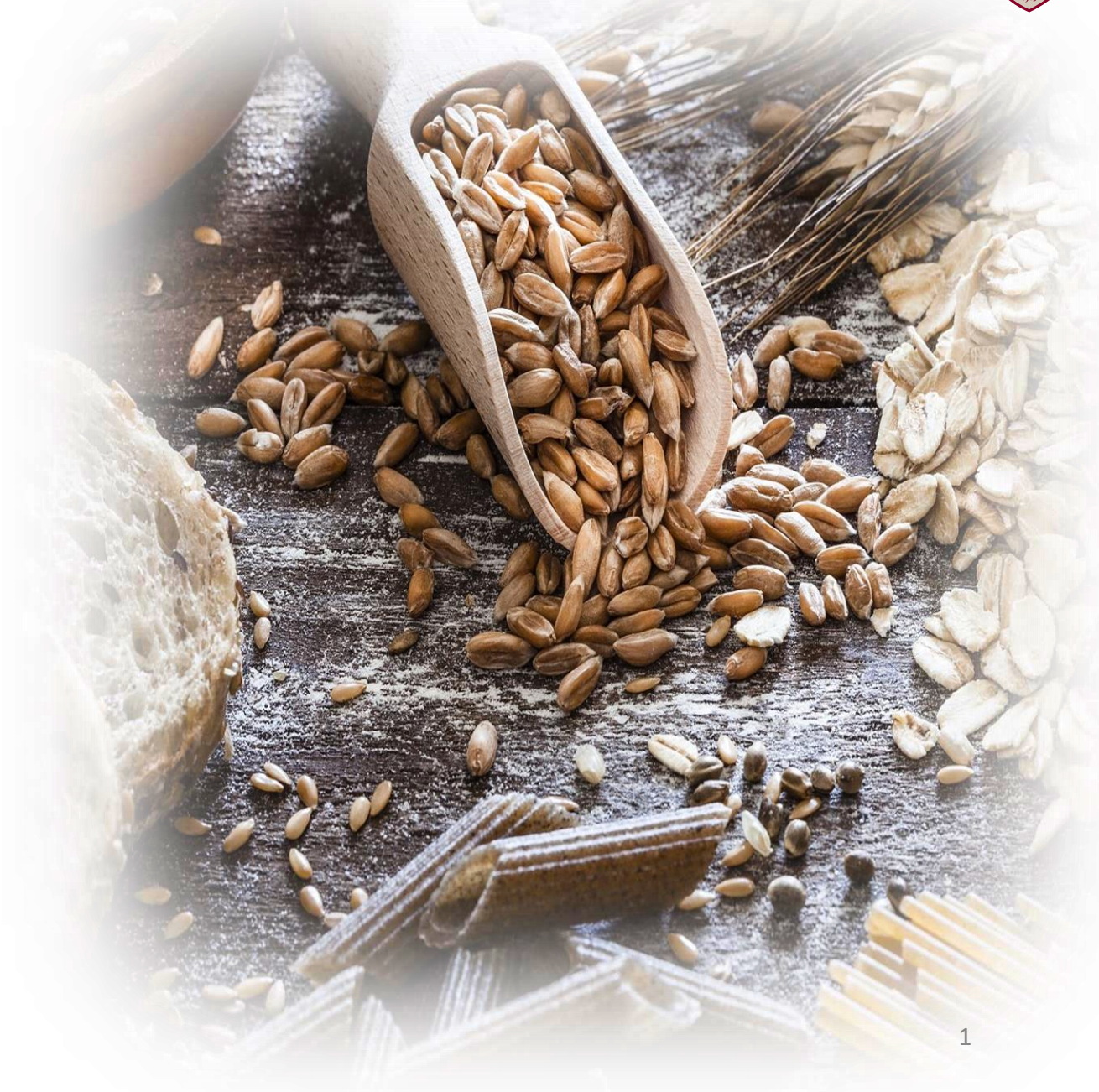
당뇨 중증도에 따른 군집에 영향을 미치는 요인 파악 및 중재 제시



2021. 01. 14

고려대학교 BA 신은영

목차

연구 및 데이터 소개

연구 주제 : 당뇨 중증도에 따른 군집에 영향을 미치는 요인 파악 및 중재 제시
- 심리적, 행동적, 사회 인구학적 요인을 중심으로

연구 소개

연구 목적

본 연구의 목적은 당뇨 중증도 지표들을 통해 일반인부터 전당뇨, 당뇨 환자들의 군집을 분류하고, 분류에 영향을 미치는 행동적, 심리적, 사회인구학적 요인을 파악한 뒤 문제 요인의 변화를 유도하기 위해 적절한 Intervention을 파악함에 있음

주제 선정 배경

- 개인의 고통(개인적 문제) + 19년 기준 2조 7000억의 당뇨병 진료비(사회적 문제)
- 당뇨병은 예방이 가장 중요
- 전당뇨를 바로잡는 것 중 가장 중요한 적은 적극적인 식사 및 운동요법과 같은 행동요인
- 사람들은 서로 다른 인구사회학적, 행동적, 심리적 특징을 지님
- 같은 것은 같게, 다른 것은 다르게 → 당뇨 예방에 있어서도 적용 필요

데이터 및 변수 소개

데이터 출처

- 국민건강영양조사 제8기 1차년도(2019), 질병관리청
- 조사 기간: 제 8기 1차년도(2019)
- 조사 대상: 목표 모집단인 대한민국에 거주하는 만 19세 이상 국민에 대하여 추출한 대표성 있는 표본
- 조사 내용: 가구원확인조사, 건강설문조사, 검진조사, 영양조사

당뇨 중증도 변수

▪ 추후 Segment 분류 시 사용

→ BMI, 체중, 허리둘레, 비만 유병 여부, 공복혈당, 당화혈색소, 인슐린, HDL_콜레스테롤, 중성지방 , 당뇨 여부

· **심리적, 행동적, 사회인구학적 변수:**

→ 가구 관련 10개 변수, 당뇨병 건강 설문 관련 7개 변수, 우울증 관련 4개 변수, 활동 제한 및 삶의 질 관련 13개 변수, 비만 및 체중조절 관련 12개 변수, 음주 관련 5개 변수, 음주 관련 5개 변수, 수면 및 정신건강 관련 5개 변수, 흡연 관련 6개 변수, 신체활동 관련 18개 변수, 식생활 조사 관련 10개 변수, 식품섭취조사 관련 3개 변수, 식품안전성조사 관련 4개 변수

변수 선택 기준 ► Appendix.1

자료 분석

(1) 데이터 전처리 및 요인 분석

DNN 기법으로 결측치 처리 및 파생 변수를 생성하고, 요인 분석의 필요성을 파악함.

1. DNN[1] 기법을 사용한 결측치 처리 및 파생변수 생성

전처리 과정:

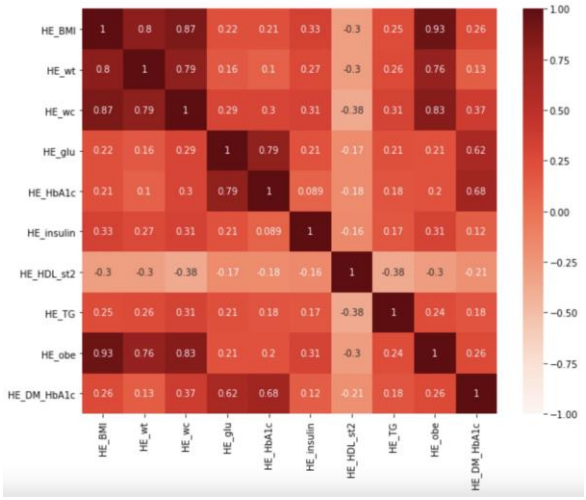
- 데이터에 28%이하의 무응답과 미응답 결측치들이 존재했는데, 이러한 Nan 값들에 대해 **DNN으로 Imputation**을 실시
- 아래와 같은 **파생변수(Feature generation) 생성**

1) 우울증을 앓은 기간 - (현재 나이 – 우울증 진단 시기)를 계산하여 우울증을 앓은 기간 변수 생성

2) 당뇨병을 앓은 기간 - (현재 나이 – 당뇨병 진단 시기)를 계산하여 당뇨병을 앓은 기간 변수 생성

2. 요인분석 필요성 확인

- 당뇨 중증도 변수들의 Correlation matrix



► 당뇨 중증도 지표들 간 **높은 상관성** 확인

- KMO 검정 결과 `0.8154984521858487`

► KMO 검정 결과 약 0.815의 KMO 값을 가지므로, 요인 분석에 적절함을 확인

- Bartlett 검정 결과 `(46296.98741731899, 0.0)`

► Bartlett 검정 결과 p-value가 0.0의 값을 보이므로 요인분석에 대한 적합성 확인

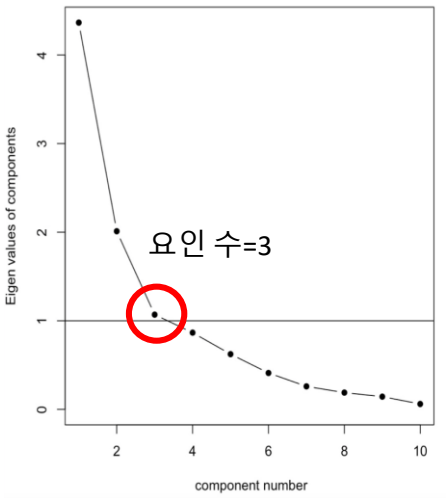
1) DNN: Deep Neural Network: 심층신경망. 본 연구에서는 심층 신경망을 통해 데이터에 존재하는 누락된 값(결측치)를 유추함

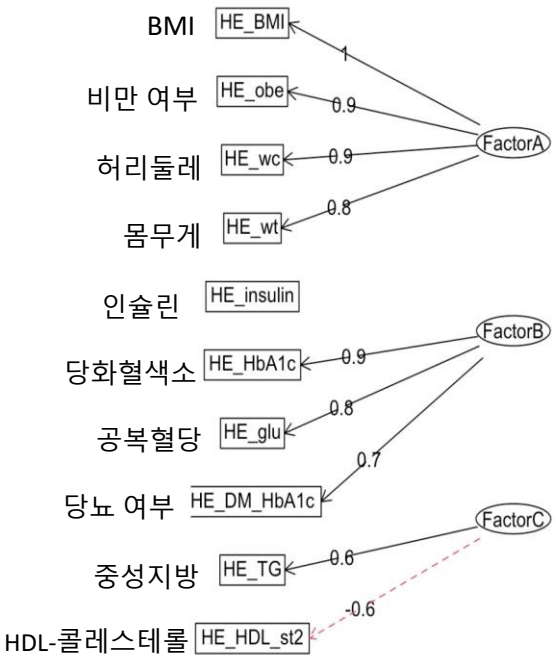
요인 3개로 요인 분석을 실시하고 결과를 해석함.

3. 요인 분석 시행 결과

당뇨 중증도 지표 요인분석 결과 및 해석

당뇨 중증도 지표들의 Scree plot





- Factor A :

BMI, 비만 여부(수준에 따라 6단계 분류), 허리둘레, 몸무게 변수

→ Anthropometric(측정 지표)

- Factor B: 당화혈색소, 공복혈당, 당뇨 여부

→ Glucose related index (혈당연관지표)

- Factor C: 중성지방, HDL_콜레스테롤

→ Cholesterol index (콜레스테롤지표)

자료 분석

(2) 분석 방법

각 연구 목적에 부합하는 분석 방법을 다음과 같이 적용

STEP 1. 환자 군집 분류

요인 분석을 통해 도출한
3개의 당뇨 중증도 Factor들을 통해
환자 군집 분류를 진행하여
당뇨 중증도에 따른 군집 도출

요인 분석 후 3개 당뇨 중증도 Factor

↓

군집 분류

- **K-means Clustering**
- Hierarchical clustering
- DBCSAN clustering

↓

성능 평가

- 실루엣 계수
- 칼린스키 하라바츠 점수

↓

모델 채택

STEP2. 변수 중요도 파악

당뇨 중증도에 따른 환자 군집 분류에
영향을 미치는 변수들의 중요도 파악

Feature selection

- 전체 변수
- **Correlation Coefficient**
- Variance threshold
- RFECV

↓

머신러닝 알고리즘 활용 환자 군집 분류

- XGBoost
- LightGBM
- **CatBoost**

↓

AUC/Accuracy 통한 분류 성능 평가

↓

분류에 있어 변수 중요도 파악

STEP3. Intervention 제시

각 환자 군집 별 문제 요인 해결을 위한
Intervention 제시

베이지안 네트워크

높은 변수 중요도를 보인 상위 15개
변수들 중심으로 각 군집 내에서
Feature selection을 통해
선택된 변수들 간의 관계 파악

↓

SHAP*

각 군집으로의 분류에 해당 변수가
미치는 영향력의 크기와 방향(+/-)
파악

↓

각 군집 별 Intervention 제시

분석 결과

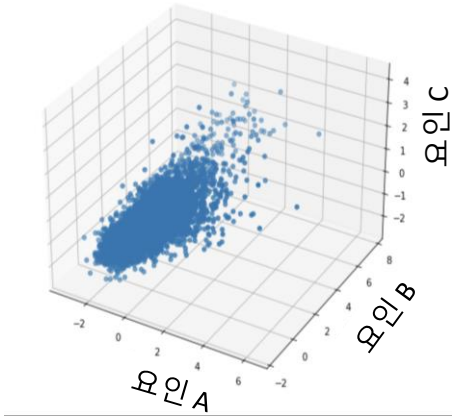
(1) 환자 군집 분류

군집 분석 기법을 선택하고 군집화 성능을 평가하여 모델을 채택함.

1. Clustering 기법 선택

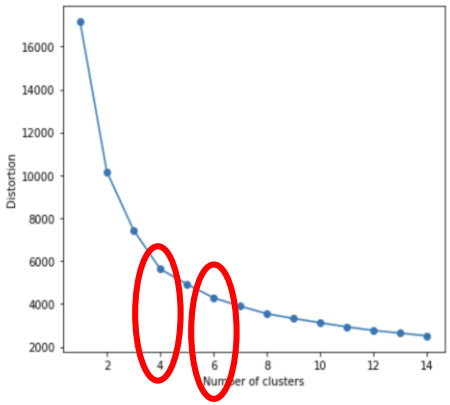
Clustering 기법	특징
K-means clustering	클러스터 수를 정하기만 하면 데이터 분포 형태를 파악하지 않아도 된다는 장점
Hierarchical clustering	군집 수 사전에 정하지 않아도 되는 트리 형태
DBSCAN clustering	다른 밀도 분포를 가진 데이터의 군집 분석에 적절하지 않으며, 데이터가 한 곳에 뭉쳐있는 형태로 해당 데이터에 적절하지 않음[3]

요인 분석 후 데이터 분포 형태



2. 군집화 성능 평가

K-means Clustering Scree Plot



	군집 수	성능 평가	
		실루엣 계수	칼린스키 하라바츠 점수
K-means clustering	4	0.5147	7821.40
	6	**0.6193**	**10125.06**
Hierarchical clustering	4	0.4564	5146.08
	6	0.5425	7761.27

Appendix.군집화 성능 평가 참고

분석 결과

(1) 환자 군집 분류

성능 평가로 채택된 K=6일 때 k-means clustering를 모델로 설정하고 환자 군집 분류 결과를 해석

K=6일 때 k-means clustering을 실행한 결과

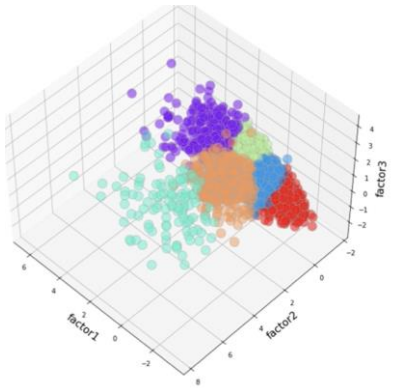
Cluster	신체측정지표 (Factor A)	혈당관련지표 (Factor B)	콜레스테롤 지표 (Factor C)	군집 명
Cluster 2	▲	▲▲	▲▲	당뇨 - 고위험군
Cluster 4	▶	▲	▲	당뇨 – 위험군(가)
Cluster 0	▲▲	▲▶	▲▲	위험군 (나)
Cluster 3	▲►	▼►	▼►	관리 대상
Cluster 1	▼	▼	▼	정상 (다)
Cluster 5	▼	▼	▼	정상 (라)

표 설명) (매우 높음) ▲▲ > ▲ > ▲▶ > ▼► > ▼ > ▼▼ (매우 낮음)

Clustering 결과 – cluster 별로 당뇨 중증도 변수 통계량 확인 및 해석

BMI 지수 / 몸무게 / 허리둘레 / 공복혈당 / 당화혈색소 / 인슐린 / HDL 콜레스테롤 / 중성지방 / 비만 상태(숫자가 커질수록 심각) / 당뇨 여부

Cluster_fa_kmeans_6	HE_BMI	HE_wt	HE_wc	HE_glu	HE_HbA1c	HE_insulin	HE_HDL_st2	HE_TG	HE_obe	HE_DM_HbA1c
0	31.445064	88.732925	103.782159	109.882814	6.029437	19.500160	43.199432	220.151897	5.0	2.133489
1	23.275614	62.035811	82.725901	97.316285	5.651183	7.758513	52.293152	119.632465	3.0	1.662724
2	26.038057	69.997368	92.038596	215.271930	9.971053	11.954386	44.684211	210.912281	4.0	3.000000
3	26.515973	73.518189	92.030050	98.833989	5.666428	11.272555	46.863024	165.452277	4.0	1.718642
4	24.591235	64.515702	89.327575	131.740011	7.034367	11.090686	47.781506	159.213131	4.0	2.954455
5	20.334879	52.787266	71.971883	90.776413	5.411617	5.714664	62.269284	81.463199	2.0	1.261862



Cluster 0 : BMI , 비만지수,허리둘레가 가장 높고, 현재 비만 전당뇨 -> 관리가 가장 필요한 그룹
Cluster 1 : BMI 가 정상이지만 낮은 비만 정도가 존재 - 전당뇨에 약간 가깝지만 정상임
Cluster2 : BMI , 비만지수,허리둘레가 높은편 이며, 현재 당뇨를 앓고 있음
Cluster 3 : BMI , 비만지수,허리둘레가 높은편 이고 전당뇨이기 때문에 관리를 해야함
Cluster 4 : BMI , 비만지수,허리둘레가 높은편 이고 당뇨를 앓고 있으나 당화혈색소와 공복혈당이 낮은 편
Cluster 5: 정상

→ **위험군(나)**
정상(다)
당뇨-고위험군
관리 대상
당뇨-위험군(가)
정상(라)

분석 결과

(2) 변수 중요도 파악

환자 군집 분류에 영향을 미치는 변수 중요도 파악 전, 먼저 분류 모델 성능을 높이기 위해 feature selection 선행

*** 데이터셋을 랜덤하게 분할하여 80%는 training data로, 20%는 test data로 구분**

① Feature Scaling 및 범주형 변수 처리

방법론	상세 설명
Min-max scaling	설명변수들에 대하여 Scaling 실시하지 않음 (Tree기반 분류 알고리즘을 사용하기 때문에 scaling 효과 없음)
label-encoding (적용)	범주형 데이터가 많은 현재 데이터 특성상 One-hot encoding시 데이터가 너무 sparse 해져 분류에 악영향

② Feature selection

방법론	상세 설명
전체 변수	**변수 전부를 선택하여** 모델 훈련에 사용
Correlation coefficient	Target variable(환자 군집 분류)과의 상관계수가 높은 변수들 선택하여 모델 훈련에 사용 → **상위 33개 선택**
Variance threshold	분산이 매우 적어 유용한 정보가 없는 변수들을 제거하여 모델 훈련에 사용 → **(Variance threshold=0.6) 35개 변수 선택**
RFECV[1)]	먼저 correlate feature들을 삭제한 뒤 RFECV 시행하여 모델 훈련에 사용 → **28개 변수 선택**

1) RFECV(Recursive Feature Elimination with Cross Validation) : 전체 변수를 포함한 모델에서 중요도가 가장 낮은 변수들을 제거해가면서 K-fold CV를 통해 모델의 성능을 계산하
이 과정을 recursive하게 수행하여 최적의 변수를 도출

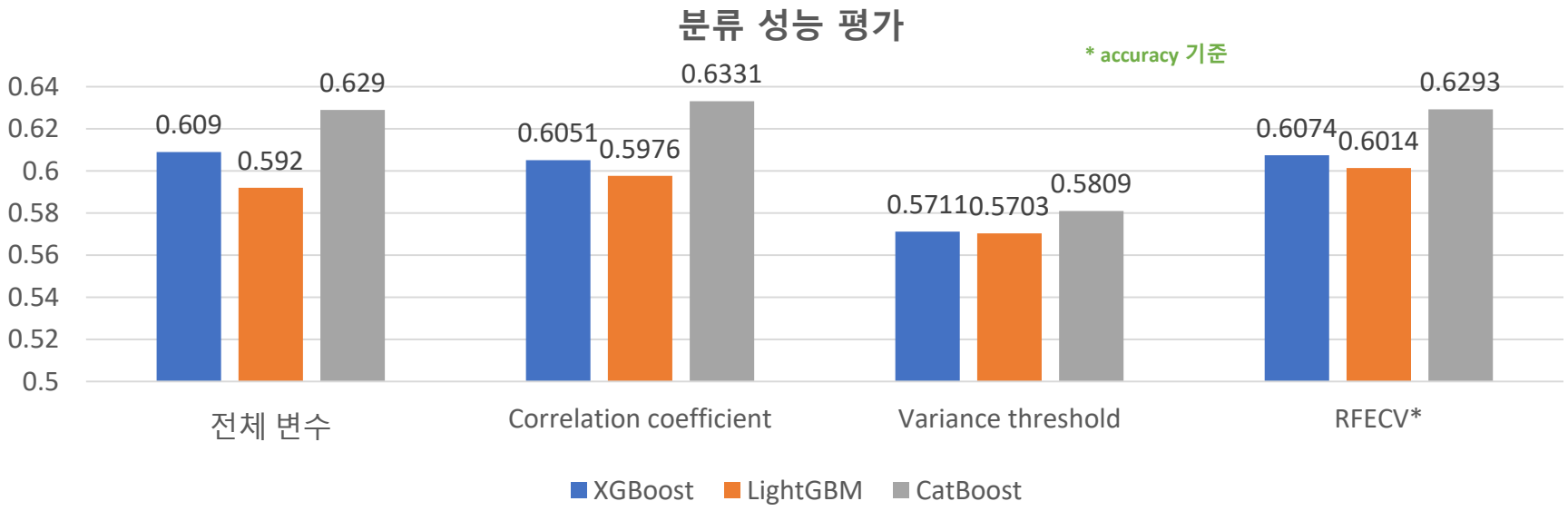
분석 결과

(2) 변수 중요도 파악

환자 군집 분류 성능을 높이기 위해 feature selection 이후 분류 알고리즘 적용함.

③ 알고리즘을 통한 분류 성능 평가

알고리즘	알고리즘 선택 이유
XGBoost	여러개의 decision tree를 조합해 사용하는 앙상블 알고리즘, 비교적 정확도가 낮은 모델들의 여러 개 조합하여 정확도를 높임
LightGBM	Leaf-wise 방식을 통해 더 손실을 줄여 사이즈가 큰 데이터를 빠르게 분석
CatBoost	순서에 따라 모델을 만들고 예측하는 ordered boosting 기법을 사용하며 범주형 데이터에 특화되어 있음



→ **상위 상관계수 feature selection 결과들을 투입하였을 때 CatBoost의 분류 성능이 가장 뛰어남**

분석 결과

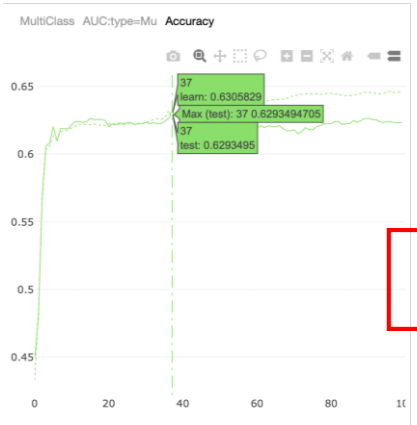
(2) 변수 중요도 파악

상위 상관계수 변수들을 투입해 하이퍼 파라미터 최적화를 거친 CatBoost 모델을 통해 중요 변수 추출함.

④ Feature importance 파악

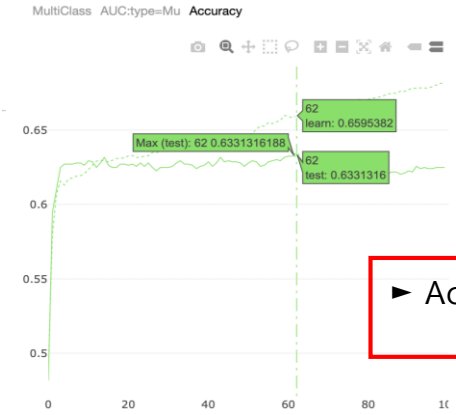
6개의 환자 군집 분류에 있어 최고의 분류 성능을 보인 모델인 상위 상관계수를 투입한 CatBoost 모델을 통해 Feature Importance 파악

최적화된 CatBoost model 검증 - **AUC**



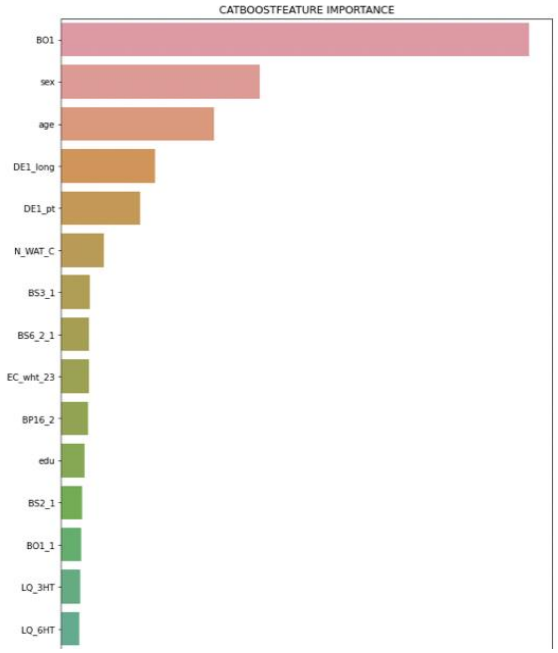
► AUC 0.893
(Test dataset 기준)

최적화된 CatBoost model 검증 - **Accuracy**



► Accuracy 0.6331
(Test dataset 기준)

CatBoost model feature importance



당뇨 중증도에 영향을 미치는 상위 15개 변수들

- 주관적 체형 인식
- 성별
- 나이
- 당뇨 진단 후 기간
- 당뇨 치료 경험 여부
- 물 섭취량
- 현재 흡연 여부
- 과거 흡연자 흡연 기간
- 주당 평균 근로시간
- 주말 하루 평균 수면 시간
- 교육 수준
- 흡연 시작 연령
- 1년간 체중 변화 여부
- 기운
- 기억

분석 결과

(3) 환자 군집 별 중재(Intervention) 제시 및 정책 제언

각 환자 군집 별로 베이지안 네트워크를 생성하여 변수 간 인과관계 파악 후 중재(Intervention)를 제시하고자 함.

베이지안 네트워크

각 변수들 간의 인과관계를 파악하기 위해서 인과성이 높은 관계를 네트워크 구조로 표현하고, 표현되지 않은 관계에 대해서는 상호 독립성을 가정하며, 직접적인 인과 관계에서의 조건부 확률만을 정의하여 확률 분포를 표현한 네트워크

1. 상위 상관계수를 통한 Feature selection에서 선택된 33개 변수를 투입하여 다음 조건에 따라 노드를 차단한 베이지안 네트워크

<조건>

a) 성별과 나이는 다른 변수들에 영향을 받을 수 없다.
b) 성별과 나이는 서로 영향을 주고 받을 수 없다.
c) 현재 흡연 여부는 과거 금연 기간과 과거 흡연자 흡연 기간에 영향을 줄 수 없다.
d) 당뇨 치료 여부가 당뇨 걸린 기간에 영향을 줄 수 없다.

2. 분류 성능이 가장 높아 데이터를 가장 잘 설명했던 CatBoost 모델에서 도출한 Feature importance에서 중요도가 높았던 상위 15개 변수를 중점으로 인과관계 파악 후 중재(Intervention)제시

<주요 상위 feature 15개>

주관적 체형 인식(B01), 성별(sex), 나이(age), 당뇨병 진단 후 기간(DE1_long), 당뇨병 치료 여부(DE1_pt), 물 섭취량(N_WAT_C), 현재 흡연 여부(BS3_1), 과거 흡연자 흡연 기간(BS6_2_1), 주당 평균 근로시간(EC_WHT_23), 주말 평균 수면 시간(BP16_2), 흡연 시작 연령(BS2_1), 1년간 체중 변화 여부(B01_1), 기운 정도(LQ_3HT), 기억력 정도(LQ_6HT)

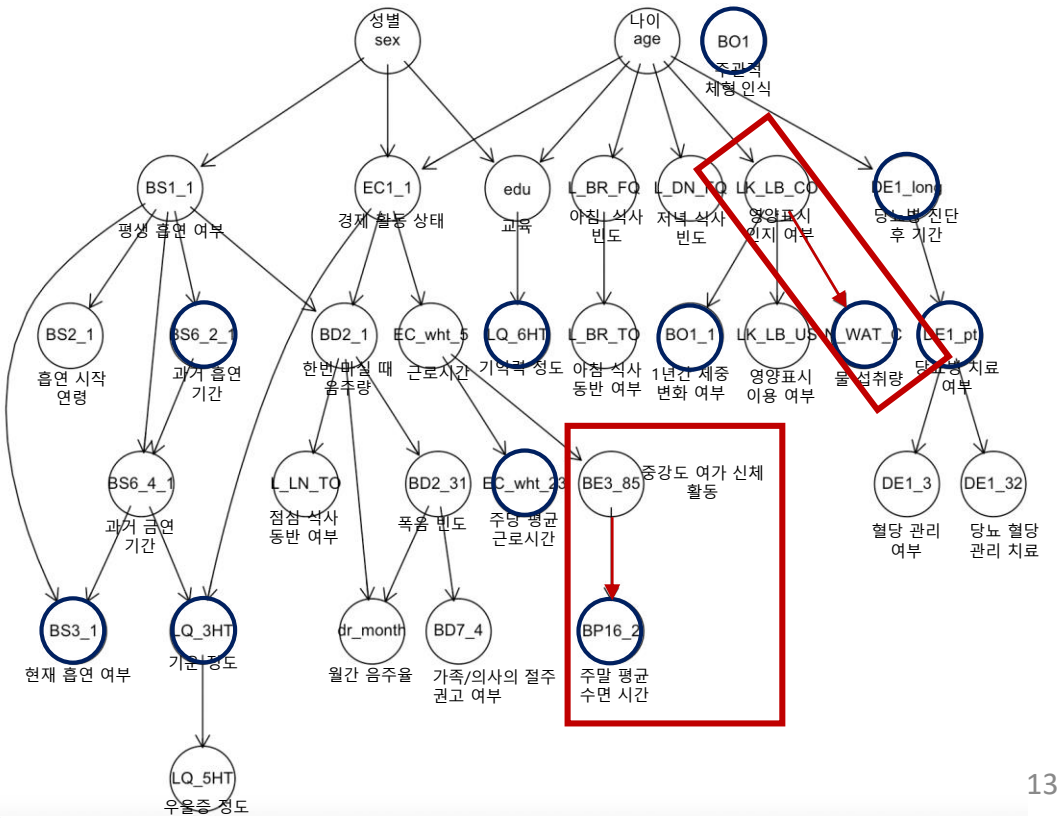
분석 결과

(3) 환자 군집 별 중재(Intervention) 제시 및 정책 제언

위험군(나)의 베이지안 네트워크를 생성함.

1. 위험군(나) – 전당뇨 군집
   - 해당 군집은 당뇨에 가까운 전당뇨
   - BMI, 비만 지수, 허리둘레가 가장 높은 편
   - 공복 혈당과 당화혈색소는 크게 높지 않으나, 인슐린과 중성 지방이 가장 높은 군집
   - HDL-콜레스테롤이 낮은 편으로 가장 비만 정도가 높은 군집

위험군(나) 베이지안 네트워크



: feature selection된 34개 변수

: feature importance 상위 15개 변수

: 노드

: intervention 제시

분석 결과

(3) 환자 군집 별 중재(Intervention) 제시 및 정책 제언

위험군(나)의 베이지안 네트워크를 통해 도출한 변수간 관계와 SHAP을 통해 위험군(나)로의 분류에 중요 변수가 미치는 +, -영향력의 정도를 파악하여 중재를 제시하고 정책을 제안함.

1. 위험군(나) – 전당뇨 군집

* 적정 수면 시간: 7~ 9시간 (세계보건기구(WHO))

중재 1) 주말 수면 시간에 영향을 미치는 중강도 여가 신체 활동을 적정 수준 유지할 것

- **SHAP[1] summary plot** 주말에 비정상적으로 적은 수면 시간을 유지할 때에 당뇨 위험군으로 분류될 확률이 높아짐을 확인
- **베이지안 네트워크** 중강도 여가 신체활동 → 주말 평균 수면 시간 (Direct Influence)

< 조건부 확률 테이블 >

		중강도 신체 여가 활동	
		안한다	한다
주말 평균 수면 시간	적정 수면 이하	**0.5888**	0.3113
	적정 수면	0.3022	**0.5755**
	적정 수면 이상	**0.1090**	0.1132

* 중강도 여가 신체활동 여부가 주말 평균 수면 시간에 영향을 미치는 정도

- 여가 신체 활동을 하지 않은 사람이 적정 수면 시간을 가지지 못할 확률 : 69.7%
- 여가 신체 활동을 하는 사람이 적정 수면 시간을 가질 확률 : 57.5%
- **즉, 중강도 여가 신체 활동을 할 수 록 적정 수면 시간을 가질 확률이 높음**

➡ **중강도 여가 신체 활동을 통해서 주말 수면 시간을 적정 시간으로 조절 가능, 당뇨 위험군으로의 분류 가능성을 낮추는 적절한 주말 수면 시간 유지를 위해 여가 시간에 중강도 신체 활동을 적정 수준 유지할 것**

정부

- 전당뇨 환자들을 위한 신체 활동 증진 보건 사업 고도화
- 지역 사회에 여가 시 즐길 수 있는 신체 활동 프로그램을 제공
- 개인 서비스, 홍보, 캠페인을 통해 여가 시 신체활동을 즐기는 환경 조성

개인

- 신체 활동을 통한 여가 시간을 보낼 필요
- 바쁜 생활 속에서도 사소한 움직임으로 신체 활동

1) SHAP Value : 전체 성과를 창출하는 데 각 Feature가 기여한 정도를 marginal 하게 계산한 값. Feature importance와 달리 음의 영향력까지 계산 가능

분석 결과

(3) 환자 군집 별 중재(Intervention) 제시 및 정책 제언

위험군(나)의 베이지안 네트워크를 통해 도출한 변수간 관계와 SHAP을 통해 위험군(나)로의 분류에 중요 변수가 미치는 +, -영향력의 정도를 파악하여 중재를 제시하고 정책을 제안함.

1. 위험군(나) – 전당뇨 군집

* 하루에 마셔야 할 물의 양을 최소 4컵, 적정 물 섭취량 8~ 10컵 (미국 의학협회 및 세계보건기구(WH

중재 2) 적정량의 물 섭취를 위해 영양 표시를 인지할 것

- **SHAP summary plot** 하루에 적정량의 물을 섭취할 때에 당뇨 위험군으로 분류될 확률이 낮아짐을 확인
- **베이지안 네트워크** 영양표시 인지여부 → 물 섭취량 (Direct Influence)

< 조건부 확률 테이블 >

		영양 표시 인지 여부	
		안한다	한다
물 섭취량	적정 섭취 이하	**0.4852**	0.2822
	적정 섭취	0.4554	**0.6749**
	적정 섭취 이상	**0.0594**	0.0429

* 영양 표시 인지 여부가 물 섭취량에 영향을 미치는 정보

- 영양 표시를 인지하는 사람이 적정량의 물을 마실 확률 : 67.5%
- 영양표시를 인지하지 않는 사람이 적정량의 물을 마시지 않을 확률 : 54.4%
- **즉, 영양 표시를 인지할 수록 적정량의 물을 마실 확률이 높음**
- 영양 표시를 인지할 수록 건강 문제에 관심이 높아 적정량의 물을 섭취하려고 노력하는 것으로 보임.

➡ **영양 표시 인지를 통해 물 섭취를 적정량으로 조절 가능, 당뇨 위험군으로의 분류 가능성을 낮추는 적정량의 물 섭취를 위해 평소에 영양 표시를 인지할 것**

정부

- 영양 표시 가이드라인 제시 및 조기 교육 실시
- 홍보, 캠페인, 환경 조성
- 영양 표시 인지를 용이하게 하기 위한 도안 변경, 기준 변경 등 수행

분석 결과

(3) 환자 군집 별 중재(Intervention) 제시 및 정책 제언

관리 대상의 베이지안 네트워크를 생성함.

2. 관리대상 – 전당뇨 군집

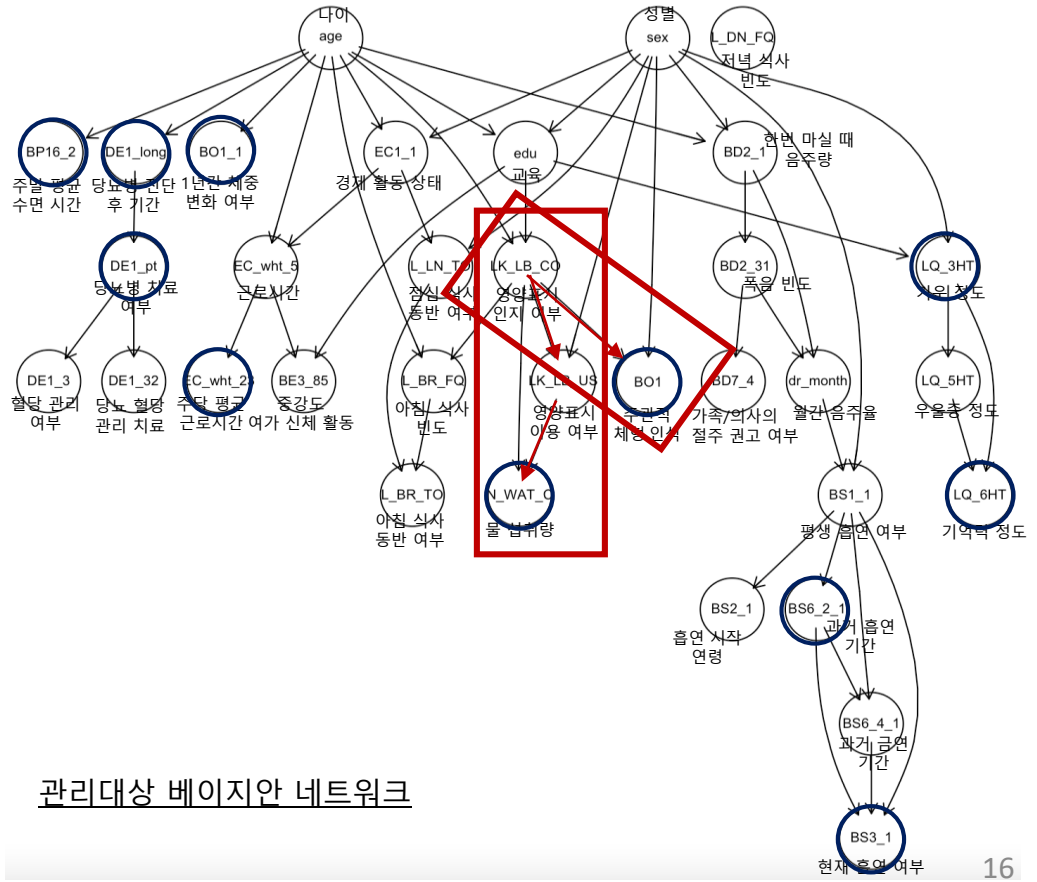
- 해당 군집은 당뇨에 가까운 전당뇨
- BMI, 비만 지수, 허리둘레가 가장 높은 편
- 공복 혈당과 당화혈색소는 크게 높지 않으나, HDL-콜레스테롤이 낮고 인슐린이 높은 군집

: feature selection된 34개 변수

: feature importance 상위 15개 변수

: 노드

: intervention 제시



관리대상 베이지안 네트워크

분석 결과

(3) 환자 군집 별 중재(Intervention) 제시 및 정책 제언

관리 대상의 베이지안 네트워크를 통해 도출한 변수간 관계와 SHAP을 통해 관리대상으로의 분류에 중요 변수가 미치는 +, -영향력의 정도를 파악하여 중재를 제시하고 정책을 제안함.

2. 관리대상 – 전당뇨 군집

중재 1) 객관적인 본인 체형 인식을 위해 영양 표시를 인지할 것

- **베이지안 네트워크 :** 영양 표시 인지 여부 → 주관적 체형 인식 (Direct Influence)

< 조건부 확률 테이블 >

영양표시 인지여부	주관적 체형 인식	성별	
		남자	여자
안한다	마름	0.0330	0.0606
	적정	**0.5047**	**0.3152**
	비만	0.4623	0.6242
한다	마름	0.0127	0.0028
	적정	**0.2746**	**0.0534**
	비만	0.7127	0.9438

* 영양 표시 인지 여부와 성별이 주관적 체형 인식에 영향을 미치는 정도

- 영양 표시를 인지하는 사람 중 정상이라고 생각: 여 5%, 남 27.4%
- 영양 표시를 인지하지 않는 사람 중 정상이라고 생각하는 사람: 남 50.5%, 여 31.5%
- **즉, 영양 표시를 인지하지 않는 사람이 영양 표시를 인지하는 사람보다 본인의 체형이 정상이라고 생각할 확률이 높음**

➡ **실제 관리대상 전당뇨 군집은 정상 군집보다 비만 지수, BMI, 허리 둘레가 비정상적으로 높은 군집이나 이들 중 영양표시를 인지하지 않는 사람들은 본인의 체형에 대해 객관적이지 못한 인식을 가졌다고 할 수 있음,**
몸 관리의 필요성을 느끼게 하는 객관적 체형 인식을 위해 영양 표시를 인지할 것

개인

- 체형 인식에 있어 **객관성이 선행되었을** 때 전당뇨에 대한 심각성을 깨닫고 당뇨 예방 행동에 적극 참여할 수 있을 것
- 영양 표시 인지를 통해 객관적인 체형 인식을 통해 당뇨 예방 행동에 적극 참여할 필요

분석 결과

(3) 환자 군집 별 중재(Intervention) 제시 및 정책 제언

당뇨:고위험군의 베이지안 네트워크를 생성함.

3.당뇨:고위험군 – 당뇨 군집

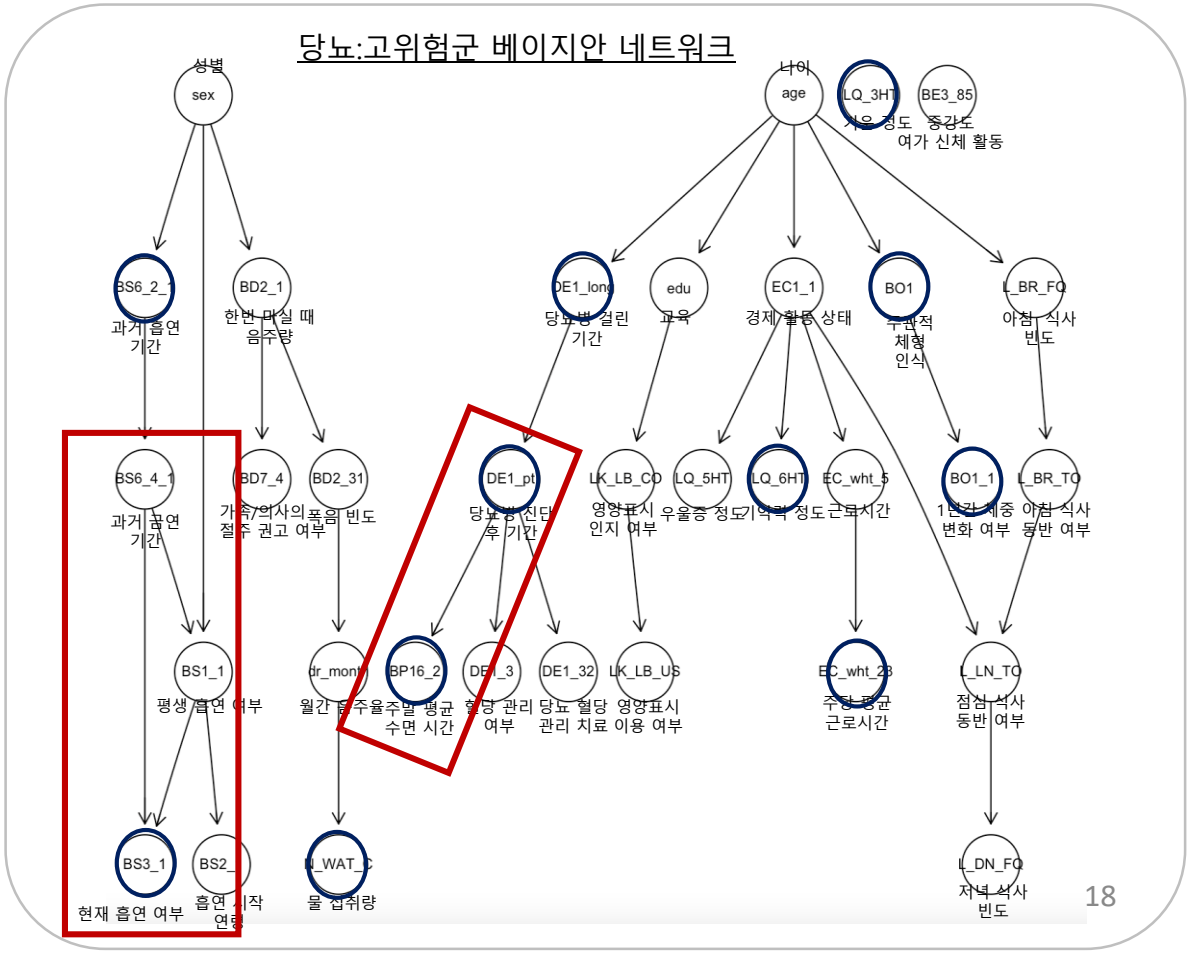
- 해당 군집은 당뇨
- BMI, 비만 지수, 허리 둘레가 높은 편이며,
- 당화혈색소, 공복혈당이 가장 높은 군집
- 인슐린, 중성지방이 높고 HDL-콜레스테롤이 낮은 군집

: feature selection된 34개 변수

: feature importance 상위 15개 변수

: 노드

: intervention 제시



분석 결과

(3) 환자 군집 별 중재(Intervention) 제시 및 정책 제언

고위험군의 베이지안 네트워크를 생성하고 SHAP을 통해 고위험군으로의 분류에 변수가 미치는 영향력의 정도를 파악하여 중재를 제시하고 정책을 제안함.

3.당뇨:고위험군 – 당뇨 군집

중재 1) 3년 초과의 금연 기간 유지를 통한 금연

- **SHAP summary plot : 당뇨-고위험군의 다수가 과거 흡연자였음을 확인, 흡연이 당뇨에 좋지 않은 영향을 미치므로 당뇨 진단 이후 건강을 이유로 금연했을 확률이 높음**
- **베이지안 네트워크 : 과거 흡연 여부 & 평생 흡연 여부** → 현재 흡연 여부 (Direct Influence)

< 조건부 확률 테이블 >

평생 흡연 여부	현재 흡연 여부	과거 금연 기간	
		3년 이하	3년 초과
5갑 미만	매일 흡연	0	**1**
	가끔 흡연	0	0
	과거O현재X	0	0
	안 피움	0	0
5갑 이상	매일 흡연	**0.6944**	0
	가끔 흡연	0.0833	0
	과거O현재X	0.2222	**1**
	안 피움	0	0

* 과거 금연 기간과 평생 흡연 여부가 현재 흡연 여부에 영향을 미치는 정도

** 평생 흡연 하지 않는 사람을 제외

- 5갑 이상의 흡연을 한 사람 중 과거의 금연 기간의 3년 초과인 사람이 흡연하지 않을 확률 : 100%
- 평생 5갑 이상의 흡연을 한 사람 중에서 금연 기간이 3년 이하인 사람은 매일 흡연을 하는 확률 : 69.4%
- 즉 금연 기간이 3년을 초과하면 완전 금연(현재 금연)에 성공할 가능성이 높아짐

금연 기간이 3년이 초과될 수록 현재 담배를 끊었을 확률이 높음, 금연 초기에 3년의 금연을 목표로 계획을 세워 효과적인 금연으로 이어질 수 있도록 함

개인

- 3년 초과의 금연 기간을 유지하여 금연을 할 필요

당뇨:위험군(가)의 베이지안 네트워크를 생성함.

4.당뇨:위험군(가) – 당뇨 군집

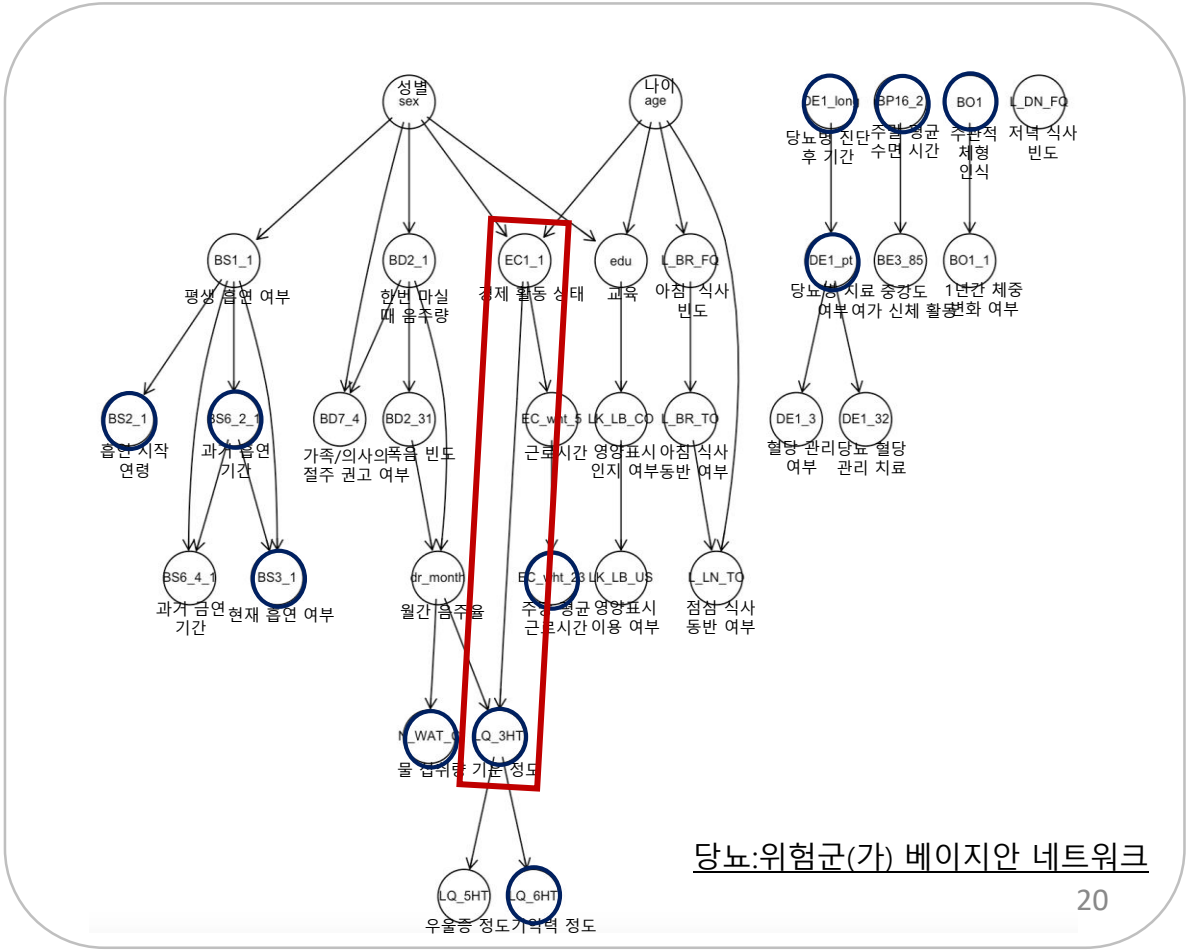
- 해당 군집은 당뇨
- BMI, 비만 지수, 허리 둘레가 평균
- 당화혈색소, 공복혈당이 높은 군집
- 인슐린, 중성지방이 높고 HDL-콜레스테롤이 약간 낮은 군집

: feature selection된 34개 변수

: feature importance 상위 15개 변수

: 노드

: intervention 제시



당뇨:위험군(가) 베이지안 네트워크

분석 결과

(3) 환자 군집 별 중재(Intervention) 제시 및 정책 제언

당뇨:위험군(가)의 베이지안 네트워크를 생성하고 SHAP을 통해 위험군(가)로의 분류에 변수가 미치는 영향력의 정도를 파악하여 중재를 제시하고 정책을 제안함.

4.당뇨:위험군(가) – 당뇨 군집

중재 1) 경제 활동을 통해 무기력감을 감소시킬 것

- 무기력감은 환자의 당뇨 관리에 치명적인 영향을 미침
- **베이지안 네트워크** 경제활동 상태 → 기운 정도 (Direct Influence)

< 조건부 확률 테이블 >

		경제활동 상태	
		안한다	한다
기운 정도	주로 기운 있다	0.4275	**0.5935**
	주로 기운 없다	**0.5725**	0.4065

* 경제활동 상태가 가운 정도에 영향을 미치는 정도

- 경제 활동을 하지 않는 사람이 기운이 없을 확률 : 57.2%
- 경제 활동을 하는 사람이 기운이 있을 확률 : 59.3%
- **즉, 경제 활동을 하지 않는 사람이 기운이 없다고 대답할 확률이 높음**

➡ **당뇨 관리에 악영향을 주는 무기력감 감소를 위해 경제활동 필요**

그러나 오래동안 경제 활동을 하지 않았거나 한번도 경제 활동을 수행하지 않은 사람에게는 어려운 intervention, 당뇨병을 앓고 있는 지원자들은 취업에 큰 불이익을 받을 가능성이 높음.

정부

- 국가는 당뇨병을 앓고 있는 취업 지원자들을 지원하고 불이익을 받지 않도록 정책을 제시하는 것이 필요
- 당뇨병 유병률이 높은 노인 인구를 위한 경제활동 참가 기회 확대 정책 필요

개인

- 국민 취업 지원 제도, 취업 지원 프로그램, 청년들을 위한 온라인 청년 센터, 고용노동부의 취업 지원 등에 참여 가능

분석 결과

(3) 환자 군집 별 중재(Intervention) 제시 및 정책 제언

정상(다) 군집과 콜레스테롤 지표에 있어 더 긍정적인 정상(라) 군집의 특성을 집단 간 평균 비교를 통해 파

정상 군집 특징 파악 – 정상(다), 정상(라)

변수 / 군집	**정상(라)**	**정상(다)**	당뇨 :고위험군	당뇨 :위험군(가)	위험군(나)	관리대상
적정 근로 시간 (40시간 이하)	**81.2%**	**77.4%**	70.2%	74%	65.5%	68.7%
현재 금연 상태	**75.3%**	**52%**	42.1%	51%	44.4%	45.1%
주관적 체형 인식 (비만이 아니라는 인식)	**87%**	**67.7%**	42.1%	56.8%	5.4%	27.7%
과거 흡연자 흡연 기간 (5년 이하)	**93.8%**	**83.9%**	74.5%	73.4%	74.9%	74.8%
폭음 빈도 (전혀 없거나 월 1회 미만)	**76.9%**	**76.1%**	74.6%	72.7%	63.9%	61.9%
영양 표시 인지	**84.3%**	**76%**	65.8%	55.2%	76.3%	73.8%
영양 표시 사용	**35.2%**	**23.6%**	14%	12.7%	21%	20%
아침 식사 시 사람 동반	**68.9%**	**60%**	57.9%	51.9%	64.2%	58.4%

당뇨에 있어 정상 집단들의 특징

- 적정 근로시간을 준수
- 현재 금연
- 비만이 아니라고 스스로 인지
- 과거 흡연을 했더라도 그 기간이 비교적 짧음
- 폭음 빈도가 전혀 없거나 월 1회 미만
- 영양표시를 인지하고 사용
- 아침 식사 시 사람을 동반

분석 결과

(3) 환자 군집 별 중재(Intervention) 제시 및 정책 제언

정상(다)과 정상(라)의 특성과 두 정상 군집 간의 주요 차이점을 SHAP summary plot을 통해서도 확인함.

정상(다) - 정상 군집	• BO1(주관적 체형인식)과 관련하여 자신이 날씬하다고 혹은 정상 체형이라고 생각한다. • LK_LB_CO(영양표시 인지여부)과 관련하여 영양표시를 인지한다. • EDU(교육수준)에 있어 학력이 높다.
정상(라) - 정상 군집 (정상(다)군에 비해 더 긍정적인 콜레스테롤 factor를 지님)	• BO1(주관적 체형인식)과 관련하여 자신이 날씬하다고 혹은 정상 체형이라고 생각한다. • **BS3_1(현재 흡연 여부)에 있어 흡연하지 않는다. - 정상(다)군과의 차이** • LK_LB_CO(영양표시 인지여부)과 관련하여 영양표시를 인지한다. • BO1_1(1년간 체중변화 여부)에 있어 체중증가가 없다. • **BS1_1(평생흡연여부)에 있어 흡연한 적이 아예 없다. - 정상(다)군과의 차이** • EDU(교육수준)에 있어 학력이 높다. • BP16_2(주말 하루 평균 수면 시간)에 있어 적절한 수면 시간을 지킨다. • L_LN_TO(최근 1년동안 점심식사 가족 및 가족 외 사람과 동반 여부)에 있어 사람을 동반한다. • LK_LB_US(영양표시 이용여부)에 있어 영양표시를 이용한다.

< 정상 군집 내에서도 보다 더 긍정적인 콜레스테롤 factor를 이끄는 요인 >

현재 흡연하지 않고 흡연 경험이 없을 것이 당뇨에 있어 정상인 사람들 사이에서도 긍정적인 콜레스테롤 factor를 이끄는 주요 요

기대 효과 및 연구 함의

문제 요인 중심의 환자 군집 생성 – 문제 해결 유도 Intervention 제시

[기대 효과] 당뇨 중증도에 영향을 주는 심리적, 인구 사회학적, 행동적 요인을 탐색하여, 중증도에 따라 각 군집에 맞게 맞춤형 중재를 제시하면서 전당뇨 환자의 당뇨병 예방에 유용한 자료를 제공할 수 있음.

[연구 의의] 해당 연구는 당뇨 중증도에 영향을 주는 다차원적인 요소를 포괄적으로 다루어, 이에 따라 각 군집에 맞춤형 중재를 제시하여 보다 세부적인 모형을 제공함.

Appendix. 군집화 성능 평가 : Hierarchical clustering

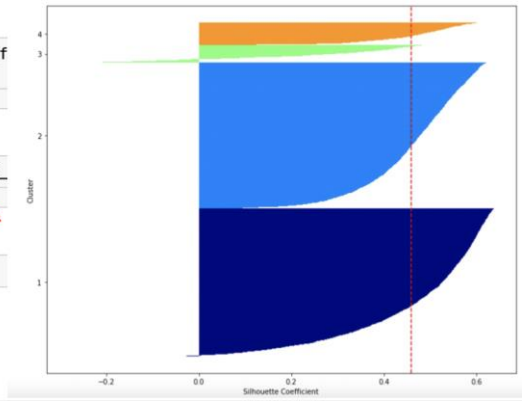
· 전체 실루엣 계수의 평균값이 1에 가까울수록 좋으며, 개별 군집의 평균값 편차가 작아야 좋은 군집화임을 고려하여 평가를 실시

· Variance ratio 개념인 칼린스키 하라바츠 점수는 값이 클수록 좋은 군집화

군집 수가 4일 때 Hierchical clustering 성능 평가 결과

```
sklearn.metrics.silhouette_score(x_after_f
executed in 499ms, finished 20:12:25 2021-11-08
0.4564323349216127 : 실루엣 계수

sklearn.metrics.calinski_harabasz_score(x_
executed in 11ms, finished 20:12:27 2021-11-08
5146.084356850881 : 칼린스키 하라바츠 점수
```

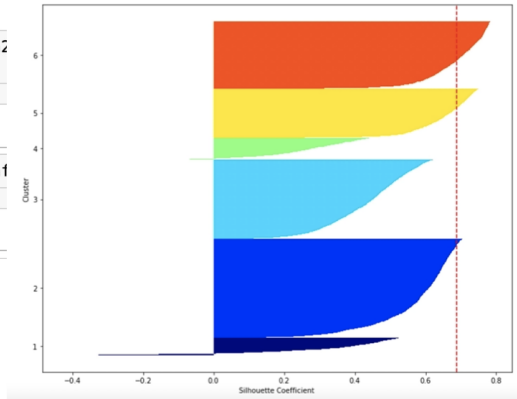


k=4
- 평균 실루엣 계수는 약 0.456
- K-means에서 k=4,6일 때와 Hierarchical에서 k=6일 때보다 훨씬 낮은 칼린스키 하라바츠 점수

군집 수가 6일 때 Hierchical clustering 성능 평가 결과

```
sklearn.metrics.silhouette_score(x_after_fa2
executed in 489ms, finished 22:27:28 2021-11-08
0.5425428335464779 : 실루엣 계수

sklearn.metrics.calinski_harabasz_score(x_af
executed in 13ms, finished 22:27:28 2021-11-08
7761.2656303879585 : 칼린스키 하라바츠 점수
```



k=6
- 평균 실루엣 계수는 약 0.543
- 군집 간 평균이 들쑥날쑥한 모습 보임

결과 종합 → Hierarchical clustering을 시행하였을 때보다 k-means clustering을 시행하였을 때보다 전체 실루엣 계수의 평균과 칼린스키 하라바츠 점수가 낮기 때문에 k-means clustering에 집중하기로 결정.

Appendix. 군집화 성능 평가 : K-means Clustering

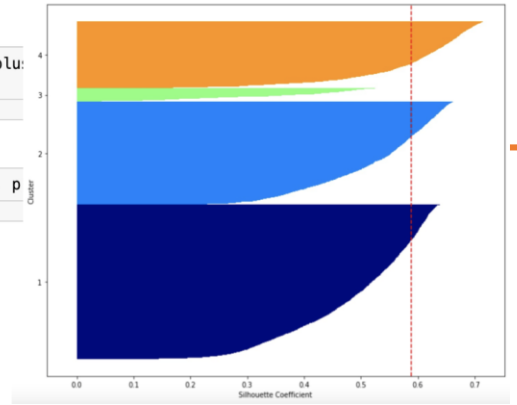
· 전체 실루엣 계수의 평균값이 1에 가까울수록 좋으며, 개별 군집의 평균값 편차가 작아야 좋은 군집화임을 고려하여 평가를 실시

· Variance ratio 개념인 칼린스키 하라바츠 점수는 값이 클수록 좋은 군집화

K=4일 때 K-means clustering 성능 평가 결과

```
sklearn.metrics.silhouette_score(x_after_fa2, pred_plu
executed in 485ms, finished 20:09:21 2021-11-08
0.5147060779176008  : 실루엣 계수

sklearn.metrics.calinski_harabasz_score(x_after_fa2, p
executed in 13ms, finished 20:09:24 2021-11-08
7821.4004807909705  : 칼린스키 하라바츠 점수
```

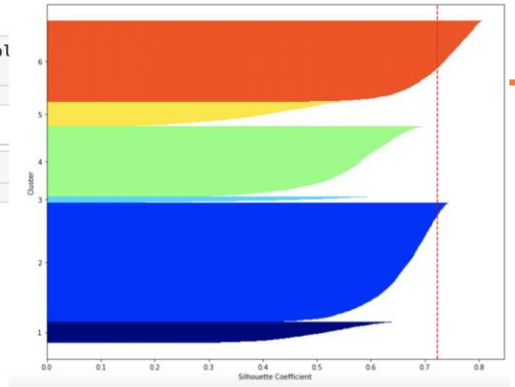


k=4
- 평균 실루엣 계수는 약 0.515
- 한 군집이 평균보다 낮은 실루엣 계수 지님

K=6일 때 K-means clustering 성능 평가 결과

```
sklearn.metrics.silhouette_score(x_after_fa2, pred_pl
executed in 484ms, finished 22:09:09 2021-11-08
0.6192792567626305 : 실루엣 계수

sklearn.metrics.calinski_harabasz_score(x_after_fa2,
executed in 13ms, finished 22:09:10 2021-11-08
10125.057640252939 : 칼린스키 하라바츠 점수
```

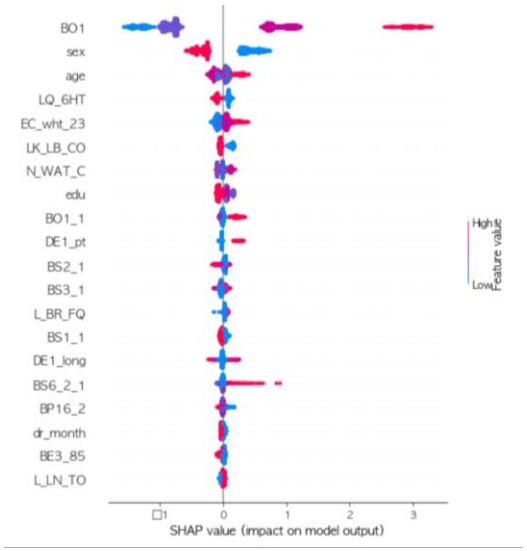


k=6
- 평균 실루엣 계수는 약 0.619
- 비교적 매우 높은 칼린스키 하라바츠 점수

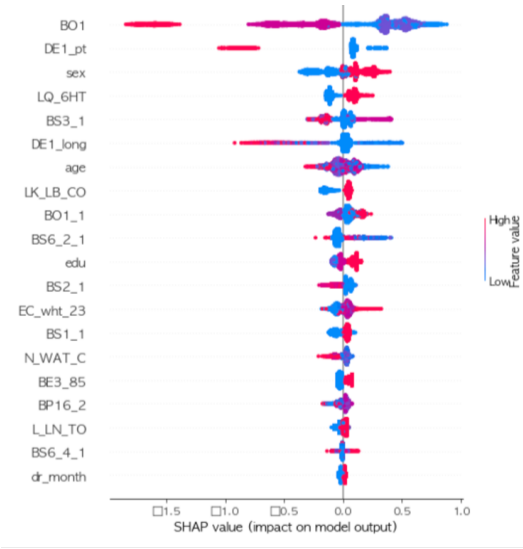
결과 종합 → k=6일 때 군집 간 실루엣 계수의 편차가 더 크기는 하지만, **실루엣 계수가 0.1이상 높고 칼린스키 하라바츠 점수가 k=4일때보다 높음, 또한 해석 면에서 단순한 Segmentation 결과를 벗어날 수 있기 때문에 k=6 모델로 결정**

Appendix. 각 군집 별 도출된 SHAP – summary plot

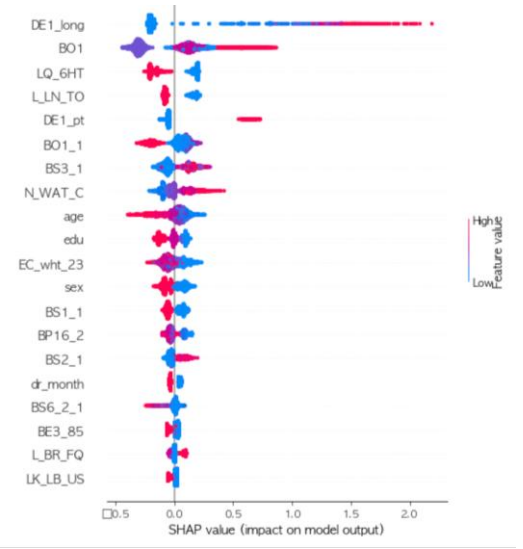
1) Cluster 0 (위험군)



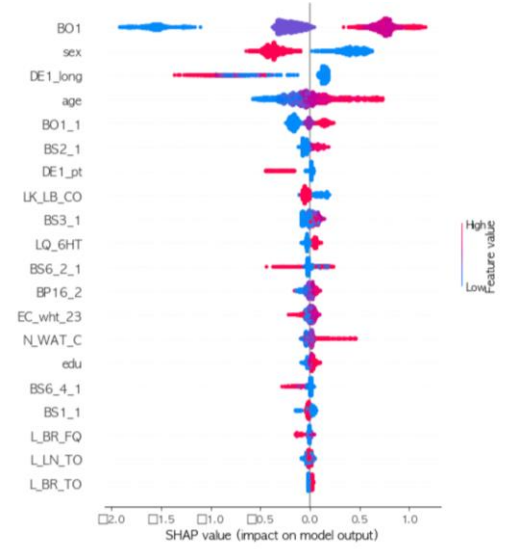
2) Cluster 1 (정상-다)



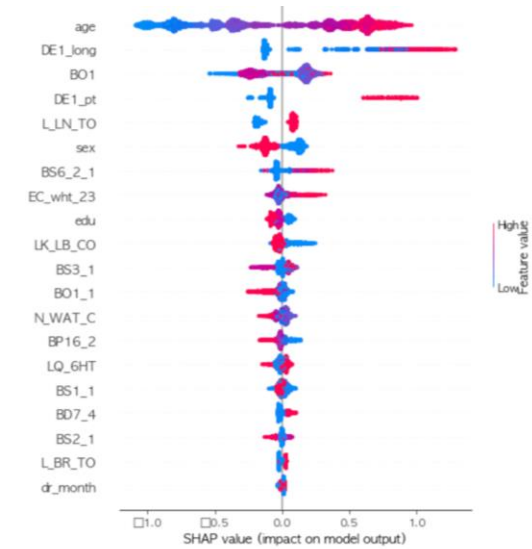
3) Cluster 2 (당뇨 - 고위험군)



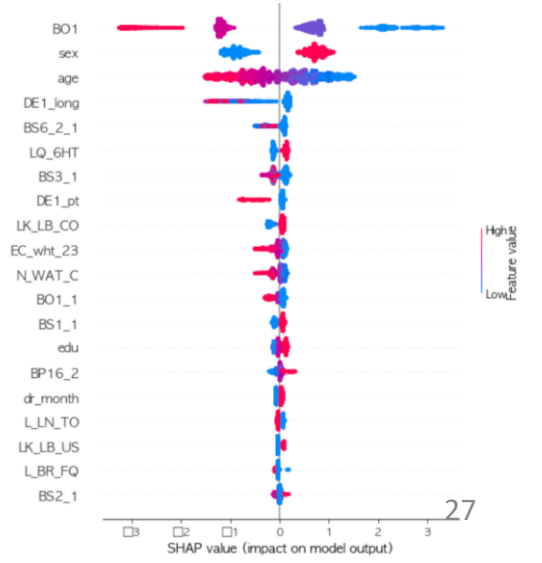
4) Cluster 3 (관리 대상)



5) Cluster 4 (당뇨 - 위험군)



6) Cluster 5 (정상-라)



Appendix. 위험군(나) 중재(Intervention) 추가 자료

1. 위험군(나) – 전당뇨 군집

중재 1) 적정 수면 시간에 따른 중강도 여가 신체 활동의 필요성

주말 수면 시간		
적정 수면 시간 보다 적게 자는 사람	적절하게 수면하는 사람	적정 수면 시간보다 많이 자는 사람
52%	37%	11%

중강도 여가 신체 활동	
한다	안한다
75%	25%

중재 2) 물 섭취량에 따른 영양 표시 인지의 필요성

물 섭취량		
적정량의 물을 마시는 사람	최소한의 물의 양보다 적게 섭취하는 사람	적정량보다 많이 마시는 사람
62.3%	33%	4.7%

Appendix. 관리대상 중재(Intervention) 추가 자료

2. 관리대상 – 전당뇨 군집

중재 1) 객관적인 본인 체형 인식을 위한 영양 표시 인지의 필요성

주관적 체형 인식		
본인 체형이 말랐다고 생각하는 사람	정상이라고 생각하는 사람	비만이라고 생각하는 사람
2%	:26%	72%

영양 표시 인지 여부	
한다	안한다
74%	26%

중재 2) 물 섭취량에 따른 영양 표시 인지의 필요성

물 섭취량		
적정량의 물을 마시는 사람	최소한의 물의 양보다 적게 섭취하는 사람	적정량보다 많이 마시는 사람
62%	37%	1%

Appendix. 당뇨:고위험군 중재(Intervention) 추가 자료

3.당뇨:고위험군 – 당뇨 군집

중재 1) 3년 초과의 금연 기간 유지를 통한 금연의 필요성

흡연 여부			
매일 흡연하는 사람	가끔 피우는 사람	과거에 피웠으나 현재는 피우지 않는 사람	한번도 안 피운 사람
22%	3%	33%	42%

중재 2) 당뇨병 진단 후 1년 이하의 환자들의 당뇨병 치료 필요성

당뇨병 발병 후 기간			
1년 이하의 당뇨병 발병 환자	1년 초과 5년 이하	5년 초과 10년 이하 환자	10년 초과
30%	14%	14%	42%

Appendix. 당뇨:고위험군 중재(Intervention) 추가 자료

3.당뇨:고위험군 – 당뇨 군집

중재 1) 3년 초과의 금연 기간 유지를 통한 금연의 필요성

흡연 여부			
매일 흡연하는 사람	가끔 피우는 사람	과거에 피웠으나 현재는 피우지 않는 사람	한번도 안 피운 사람
22%	3%	33%	42%

중재 2) 당뇨병 진단 후 1년 이하의 환자들의 당뇨병 치료 필요성

당뇨병 발병 후 기간			
1년 이하의 당뇨병 발병 환자	1년 초과 5년 이하	5년 초과 10년 이하 환자	10년 초과
30%	14%	14%	42%

Appendix. 당뇨:위험군(가) 중재(Intervention) 추가 자료

4. 당뇨:위험군(가) – 당뇨 군집

중재 1) 경제 활동을 통한 무기력감 감소의 필요성

경제활동 여부	
한다	안한다
55.6%	44.4%

기운 여부	
주로 있다	주로 없다
61.6%	38.4%

Appendix. 당뇨:고위험군 중재 1 조건부 확률 테이블

중재 1) 3년 초과의 금연 기간 유지를 통한 금연의 필요성

< 조건부 확률 테이블 >

평생 흡연 여부	현재 흡연 여부	과거 금연 기간	
		3년 이하	3년 초과
5갑 미만	매일 흡연	**0**	0
	가끔 흡연	0	**0**
	과거O현재X	**0**	0
	안피움	**0**	1
5갑 이상	매일 흡연	0	0.3162
	가끔 흡연	0	**0.5842**
	과거O현재X	0	0.0996
	안피움		
안 피움	매일 흡연	0	0
	가끔 흡연	0	0
	과거O현재X	0	0
	안피움	1	0

Appendix. 당뇨:고위험군 중재 1 조건부 확률 테이블

중재 1) 3년 초과의 금연 기간 유지를 통한 금연의 필요성

< 조건부 확률 테이블 >

평생 흡연 여부	현재 흡연 여부	과거 금연 기간	
		3년 이하	3년 초과
5갑 미만	매일 흡연	**0**	0
	가끔 흡연	0	**0**
	과거O현재X	**0**	0
	안피움	**0**	1
5갑 이상	매일 흡연	0	0.3162
	가끔 흡연	0	**0.5842**
	과거O현재X	0	0.0996
	안피움		
안 피움	매일 흡연	0	0
	가끔 흡연	0	0
	과거O현재X	0	0
	안피움	1	0