Pima Indians에 대한 데이터 분석연구

김연경

문제정의

- 사망자의 약 60%가 생활습관병이 원인이 되고 있다

- 당뇨병으로 인한 합병증으로 뇌경색이나 심근경색증 등이 발생 될 수 있다

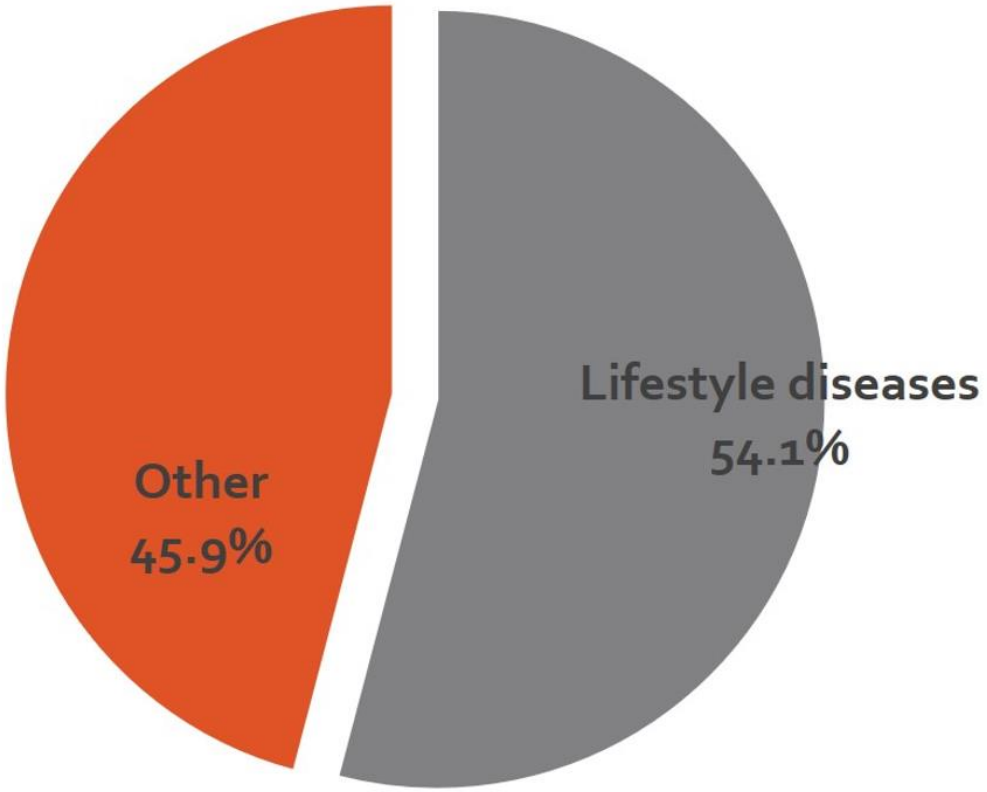


그림 1. 일본 국내 사망원인 비율

※후생노동청의 일본 인구 동태통계, 2015년

Pima Indians Diabetes Database

- Kaggle에 있는 오픈 데이터세트
- 21살 이상의 여성을 대상으로 한 768(500:정상인, 268:당뇨병 환자)명에 대한 당뇨병 진단 데이터
- 항목

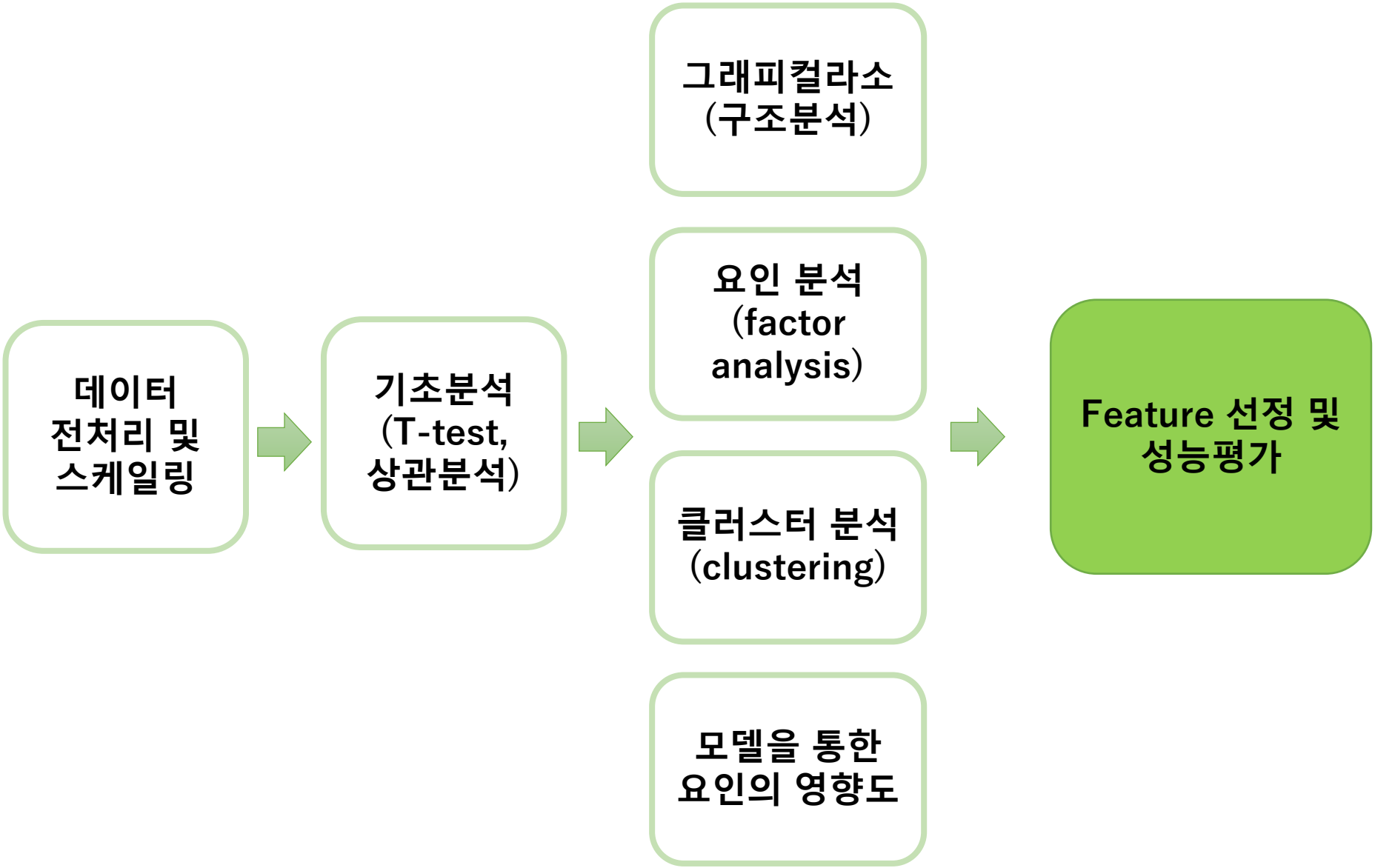
(1) Pregnancies (2) Glucose

(3) Blood Pressure (4) Skin Thickness (5) Insulin

(6) BMI (7) Diabetes Pedigree Function (8) Age

Database : 당뇨병 진단 데이터, Kaggle
https://www.kaggle.com/

데이터 분석의 흐름


데이터 전처리 및 스케일링
기초분석 (T-test, 상관분석)
그래피컬라소 (구조분석)
요인 분석 (factor analysis)
클러스터 분석 (clustering)
모델을 통한 요인의 영향도
Feature 선정 및 성능평가

데이터 전처리

- 일반적인 수치에 대한 검증을 위해 1-class SVM을 이용해서 이상치를 제거
- 이상치의 10%(nu = 0.9)를 제외한 데이터에 대해서 분석
- 그 결과, 정상인 452명, 당뇨병 환자 243명에 대한 분석을 진행

기초분석1

T-test

두 집단의 평균이 유의미 한지 파악

표 1. pima indians 대상으로 한 t 검정결과

항목	t-value	p-value
Pregnancies	-6.00	0.00
Glucose	-14.37	0.00
Blood Pressure	-2.75	0.01
Skin Thickness	-1.17	0.2
Insulin	-2.72	0.00
BMI	-7.88	0.00
Diabetes Pedigree Function	-5.25	0.00
Age	-7.04	0.00

- Skin Thickness의 p-value는 0.05보다 크므로 해당 feature는 유의미하지 않음

기초분석2

상관분석

상관분석을 통해 두 피쳐 (feature) 가 서로 상관이 있는지에 대해서 검정

- 당뇨병 환자들은 피부 두께에 따라 인슐린양이 달라지며 체형이 마른 환자에게도 피부 두께가 두꺼울 가능성이 있다.

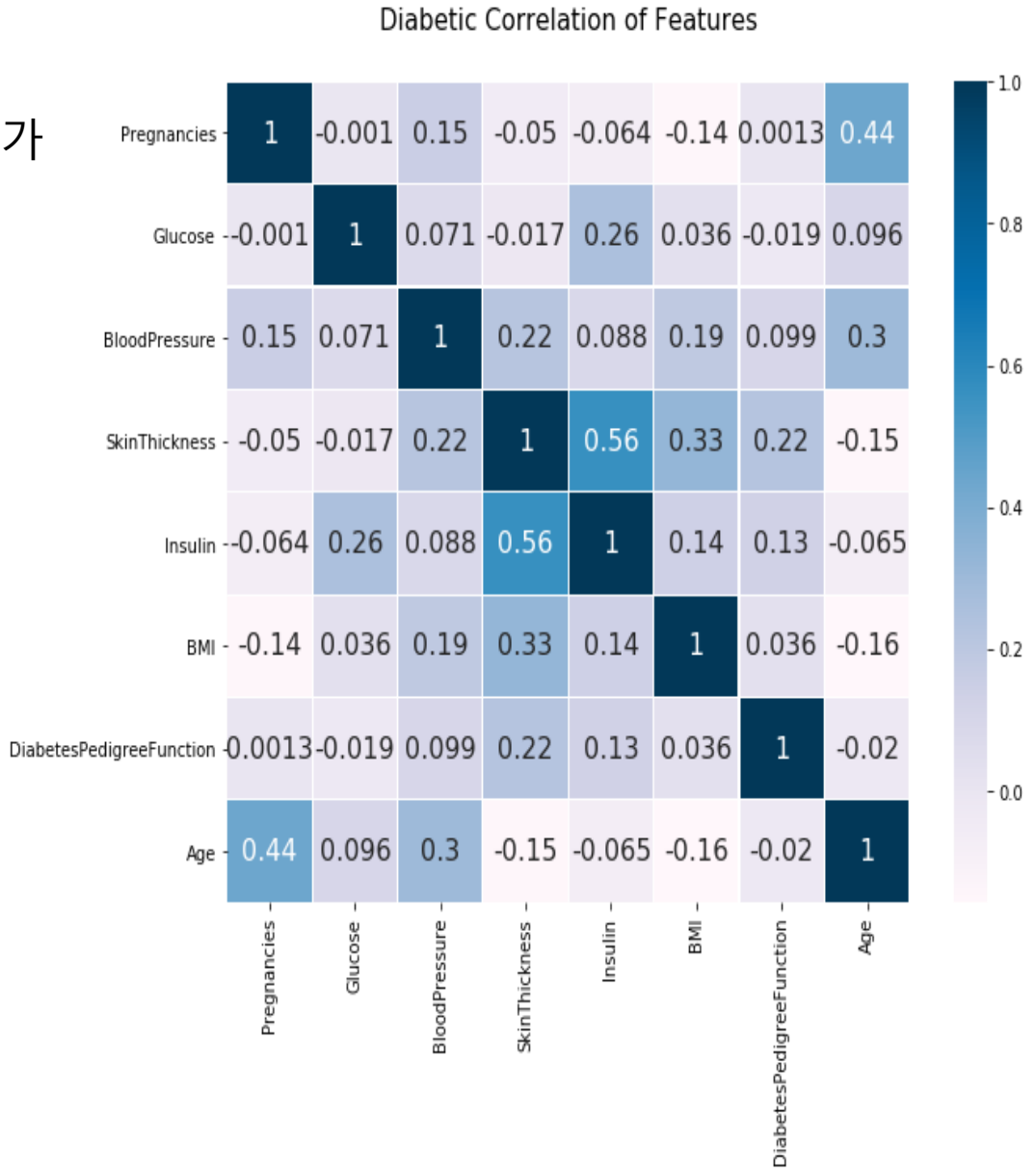


그림 2. 당뇨병 환자에 대한 상관분석 결과

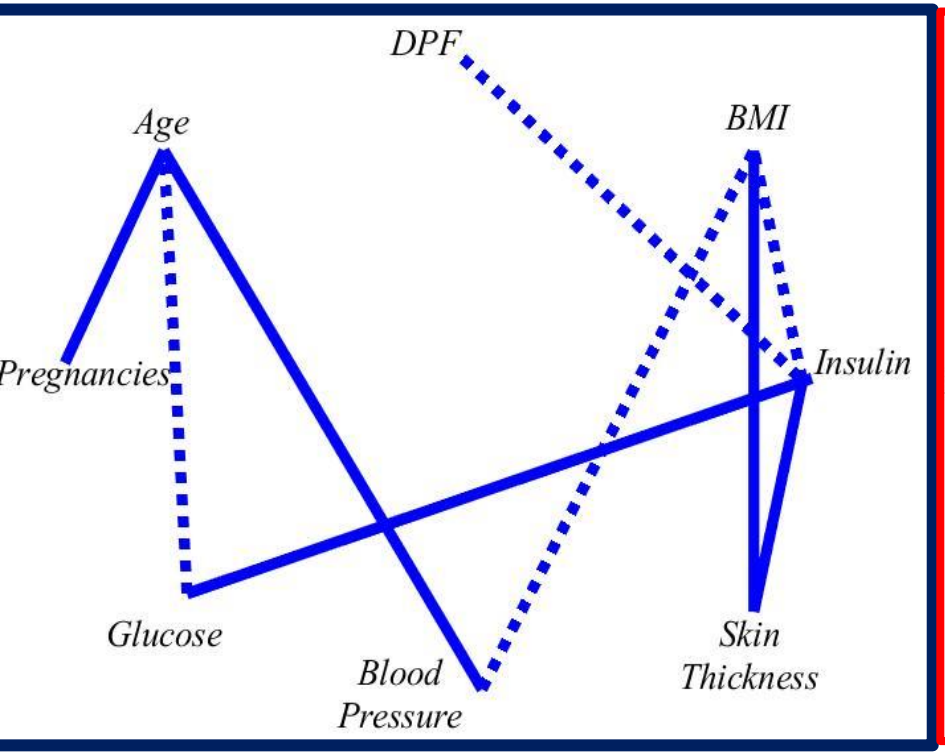
그래피컬라소（Graphical Lasso）

- 당뇨병 환자와 정상인 사이의 피쳐 간의 상관관계를 파악하기 위해 시각화 함
- 다변량 정규 분포의 복잡한 구조인 파라미터를 그래프 구조로서 변형시켜 이상감지를 추출하기 쉬움
- 두 피쳐 간의 상관 관계를 당뇨병 환자와 정상인 간의 비교가 가능
- 정상인에서 당뇨병 환자에 있어서 피쳐 간의 변화를 수치화 함

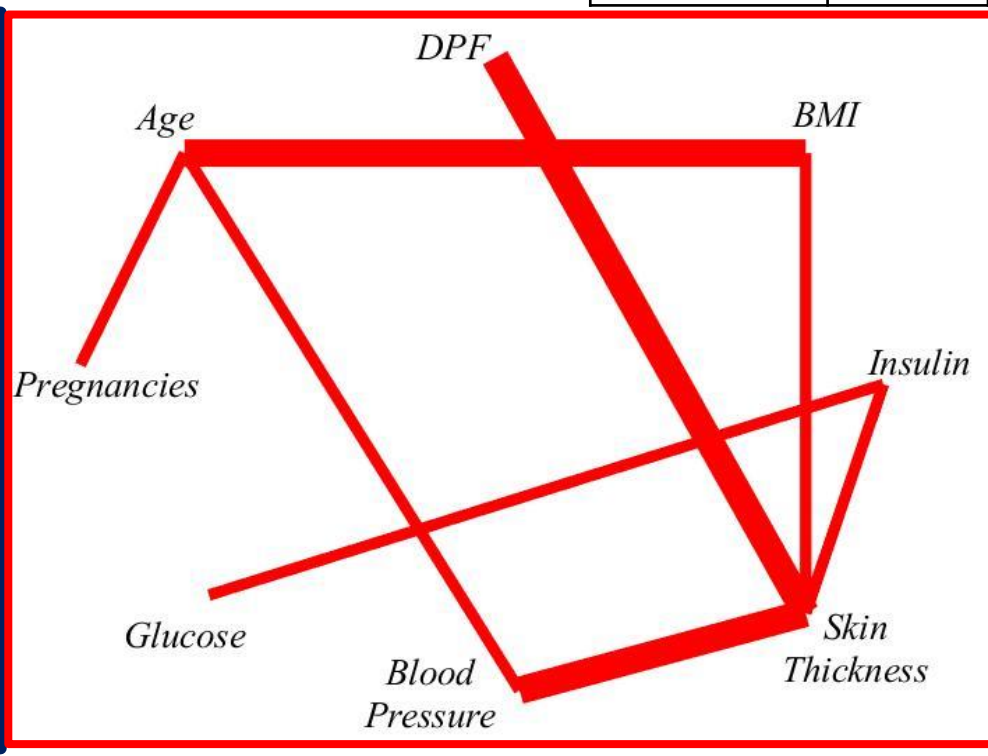
정상인과 당뇨병 환자의 그래피컬라소

- BMI와 Age 간의 의존 관계도가 높아지고, 변화도도 가장 크다

	변화값
Pregnancies	0.041
Glucose	0.023
Blood Pressure	0.028
Skin Thickness	0.013
Insulin	0.015
BMI	0.064
DPF	0.012
Age	0.037



(a) 정상인



(b) 당뇨병 환자

그림 3. 그래피컬라소의 그래프 ($\rho = 0.2$)

※MATLAB R2019a를 이용해서 실행함

요인 분석

요인 추출방법으로 주성분 분석을 활용

- 제1성분 : Skin Thickness, BMI, Diabetes Pedigree Function

  제2성분 : Age, Pregnancies, Blood Pressure

  제3성분 : Glucose, Insulin

제1성분은 피부두께로 관련된 요인,
제2성분은 연령에 관련된 요인,
제3성분은 혈당에 관련된 요인으로
볼 수 있다

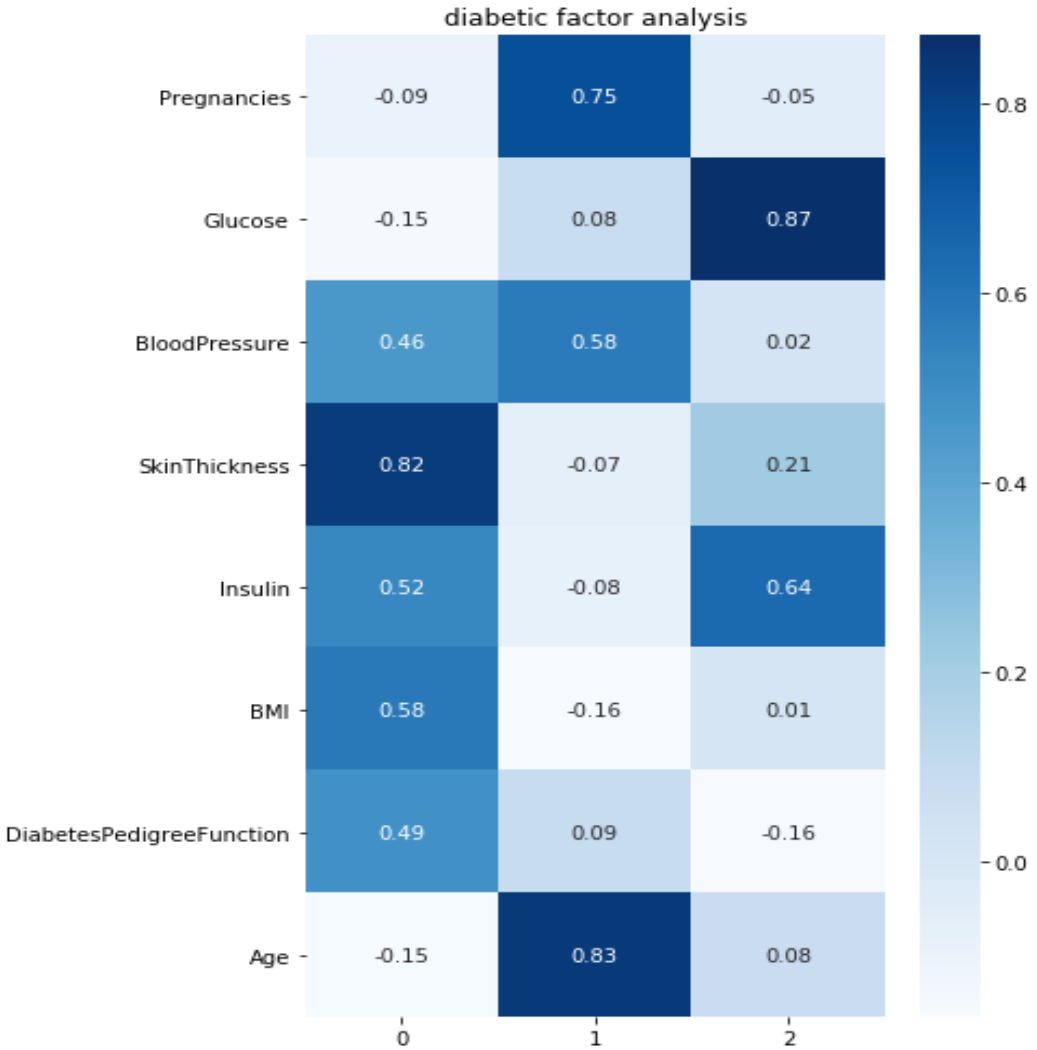
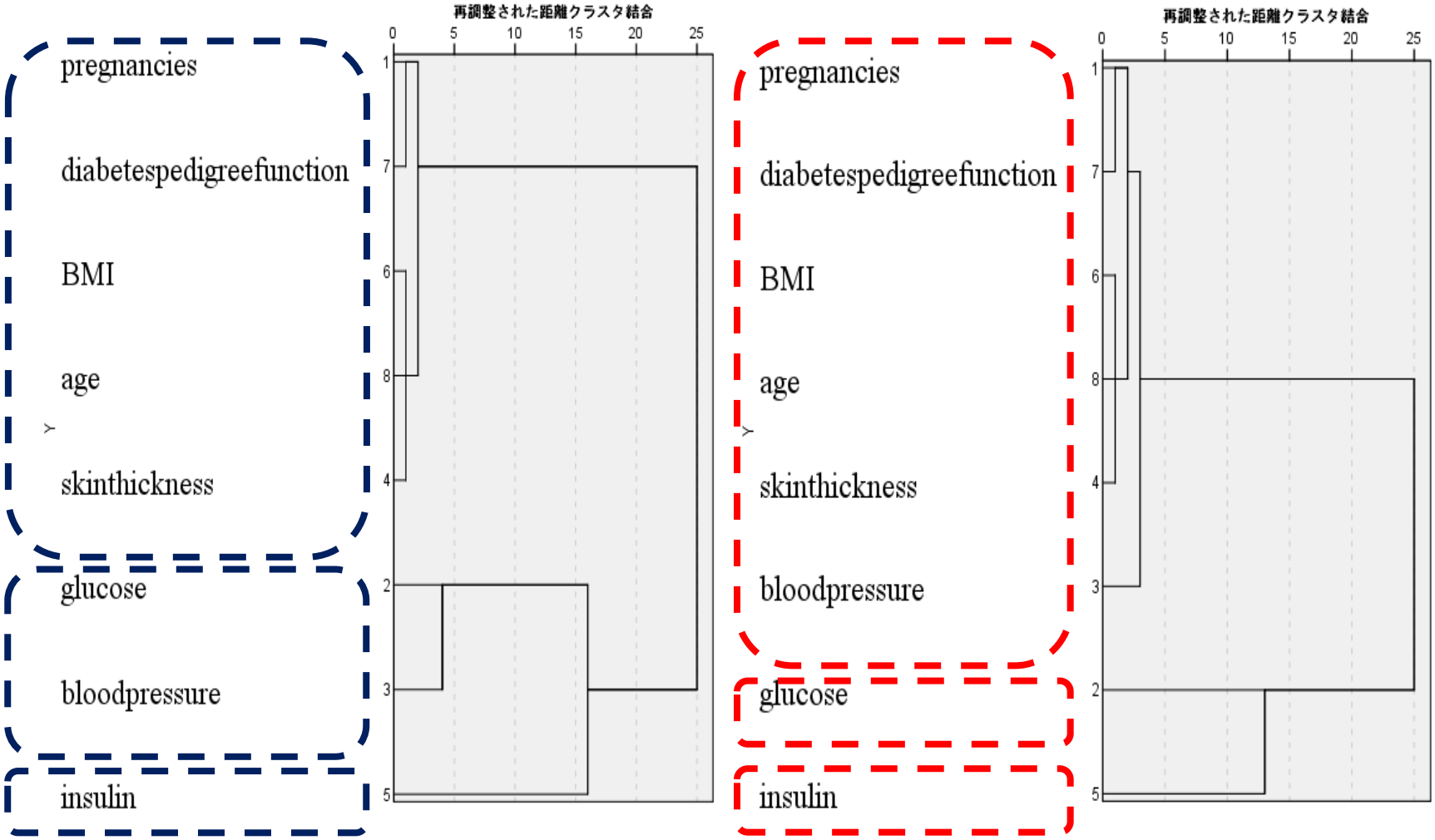


그림 4. 당뇨병 환자에 대한 요인 분석

클러스터 분석

➢ 정상인과 당뇨병 환자에 대해 피쳐들이 어떻게 군집을 이루는지를 확인



(a) 정상인 (b) 당뇨병 환자

그림 5. 정상인과 당뇨병 환자의 덴드로그램

※IBM SPSS Statistics을 이용해서 실행함

SVM을 이용한 피쳐의 영향도①

SVM을 이용해 각 피쳐를 1개씩 제거하여 나머지 7개에 대한 정확도를 측정

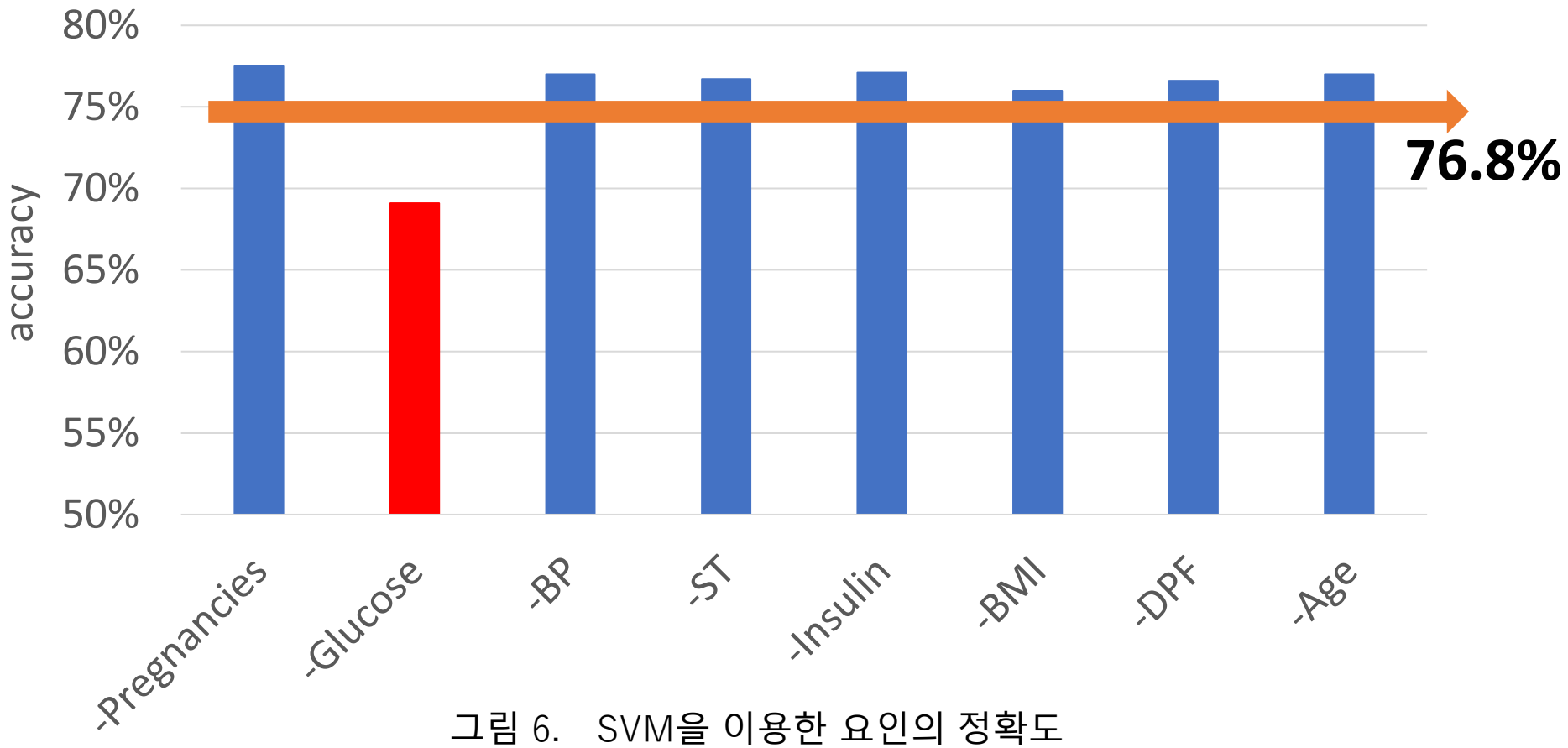


그림 6. SVM을 이용한 요인의 정확도

- 피쳐의 영향도가 클수록 정확도는 낮아지므로, Glucose의 정확도가 현저하게 낮다

※MATLAB R2019a를 이용해서 실행함

결정 트리를 이용한 피쳐의 영향도②

요인 8개가 트리를 만드는데 얼마나 중요한지를 평가

표 2. 결정 트리를 이용한 feature importance

항목	Feature importance
Pregnancies	0
Glucose	0.76
Blood Pressure	0
Skin Thickness	0
Insulin	0
BMI	0.09
Diabetes Pedigree Function	0.09
Age	0.06

- Glucose, BMI, Diabetes Pedigree Function, Age의 중요도가 높게 측정됨

Feature 선정

그래피컬라소 → **BMI, Age**
요인분석 → **Glucose, BMI, Age**
클러스터 분석 → **BMI, Age**
SVM을 이용한 영향도 → **Glucose**
결정 트리를 이용한 영향도 → **Glucose, BMI, Age**
Diabetes Pedigree Function

① Glucose ② BMI ③ Age

Feature 조합별 모델별 성능비교

➢ 데이터를 랜덤으로 학습용과 검증용으로 7:3으로 설정

표 3. 각 모델별에 따른 accuracy (%)

모델	적용 여부		Glucose BMI Age
주성분 적용	안함	함	
선형 판별 분석	77.70	74.07	75.72
이차 판별 분석	75.97	75.79	74.08
로지스틱 회귀 분석	76.69	78.10	75.72
결정 트리	73.81	73.06	75.72
K-NN	73.96	74.21	73.04
SVM	74.24	75.51	75.92

다른 조합들과 비교해
3 feature만으로도 충분한 accuracy