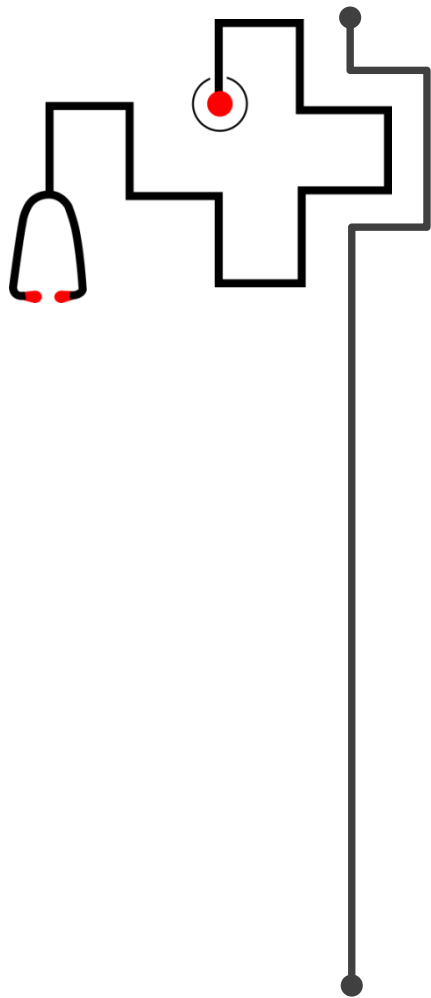


Personal Health Record

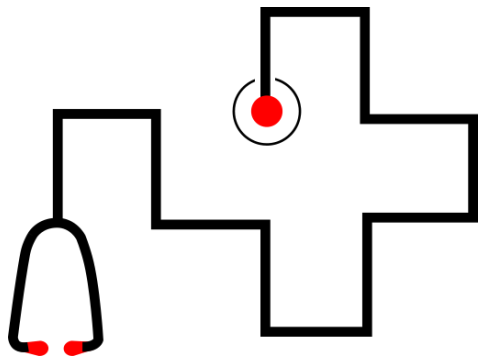
국민건강영양조사 데이터를 통한
당뇨 확률 예측

김윤승 연구원



Analysis Index

- 01 데이터셋 특징
- 02 전처리
- 03 모델링
- 04 중요 변수 분석
- 05 가능성 및 보완점



1. 데이터셋 특징

데이터셋

01

국민건강영양조사

질병관리본부 공공데이터

[링크](#)

02

2007 – 2017
기간의 데이터

03

한 해당 약 10,000건

층화 추출을 통한 표집
대표성을 가진 데이터
한 해당 평균 10,000 건

04

약 1200 가지의 질의

다양한 병의 유무 확인 가능
암, 고혈압, 당뇨 등의 상태
과거 병력

데이터셋 특징

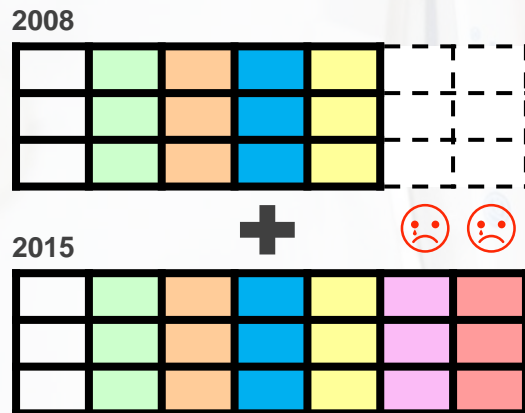
01 불균형 데이터

당뇨환자와 비당뇨환자 간 데이터가
약 10%로 매우 불균형



02 연도별 다른 조사항목 갯수


데이터셋 변수의 개수가 연도별 상이
데이터 통합에 어려움



데이터셋 특징

03 결측치 문제

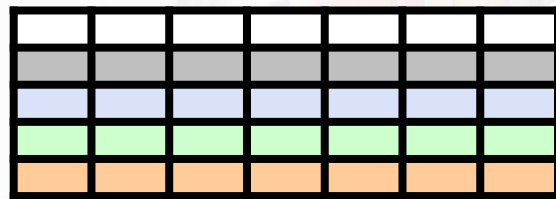
결측치가 많아 사용할 수 없는 변수 다수 존재



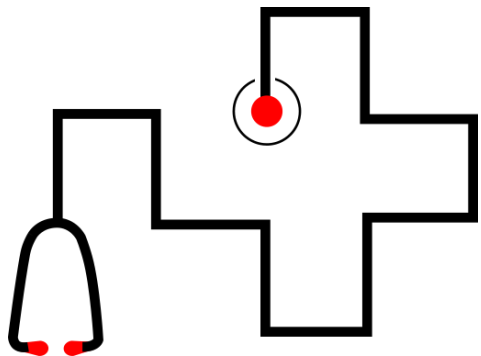
Red	Orange	White	Green	Green	Blue	Dark Blue
Red	Orange	Yellow	Green	Green	Blue	White
Red	White	Yellow	Green	White	Blue	White
Red	Orange	White	Green	Green	Blue	White

04 코호트자료가 아님

변화상태나 개선여부를 예측할 수 없음

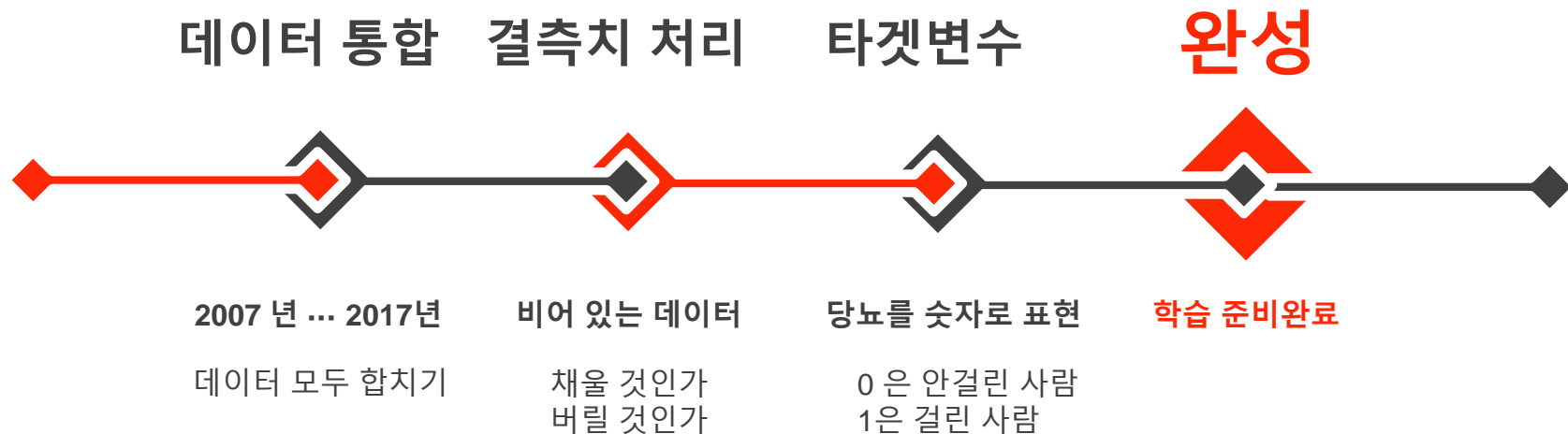


White	White	White	White	White	White	White
Grey	Grey	Grey	Grey	Grey	Grey	Grey
Light Blue	Light Blue	Light Blue	Light Blue	Light Blue	Light Blue	Light Blue
Light Green	Light Green	Light Green	Light Green	Light Green	Light Green	Light Green
Orange	Orange	Orange	Orange	Orange	Orange	Orange



2. 전처리

전처리



데이터 통합

2007년, 2008년, 2009년... 2017년 데이터



집합 데이터
(89630, 1884)

01

Outer join

연도별로 다른 변수를 통합하기 위해 outer join 사용

02

상이한 변수 처리

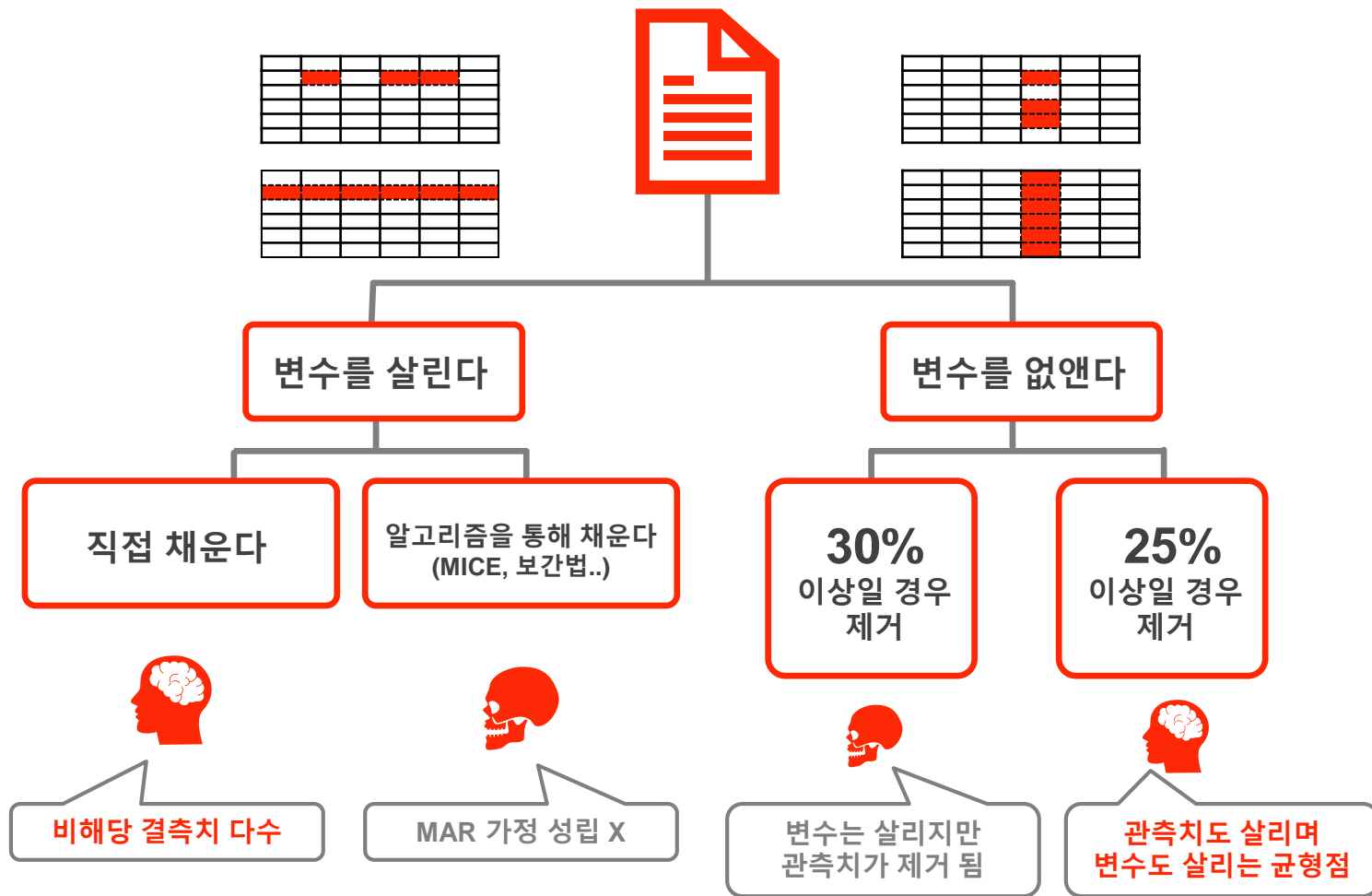
연도별 상이한 변수 통합

e.g.) 평균 수면시간
2007 - 2015 : BP8
2016, 2017 : Total_slp_wk

03

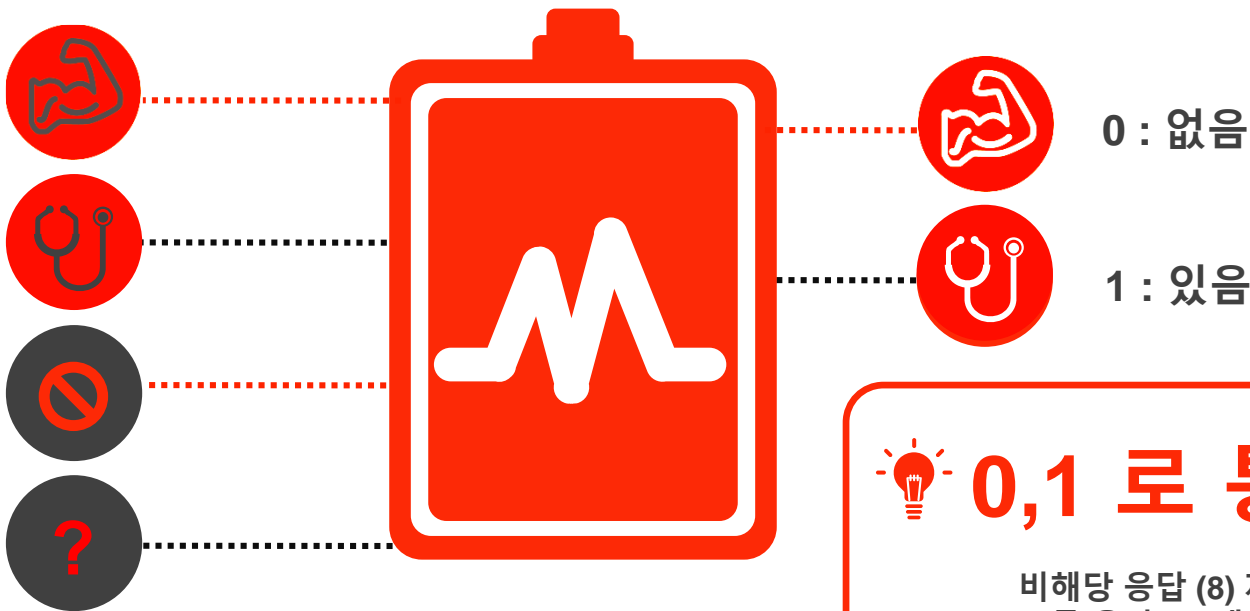
집합 데이터셋 사이즈

89630 건, 1884의 변수(질의)



타겟변수

당뇨 진단 변수 : **DE1_dg**



 **0,1 로 통일**

비해당 응답 (8) 제거
모름 응답 (9) 제거

최종 데이터셋



데이터셋 크기

24327 건의 관측치
321 개의 변수



당뇨관련 변수 제거

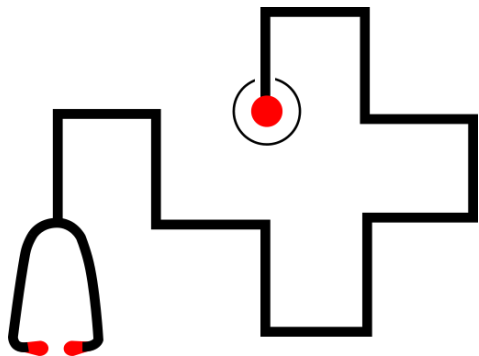
당뇨병 치료여부, 혈당관리치료, 인슐린, 약복용..



데이터표준화

빠른 학습과 과적합 방지



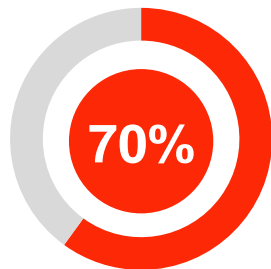


3. 모델링

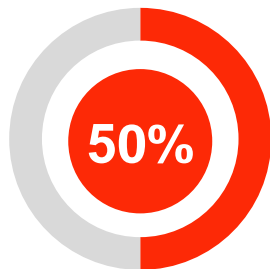
앙상블기법

$$P(y | X) ?$$

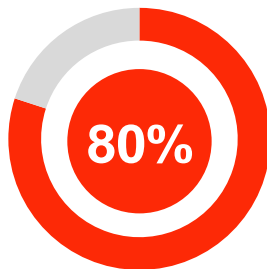
Model 1



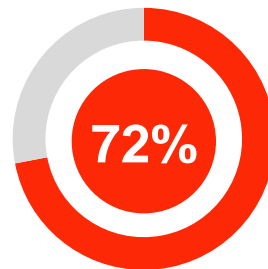
Model 2



Model 3



Model 4



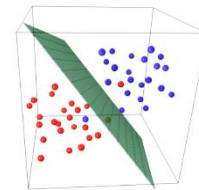
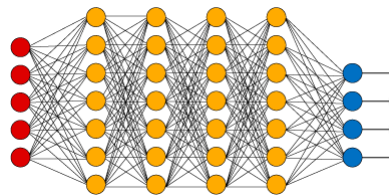
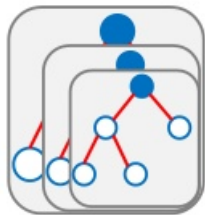
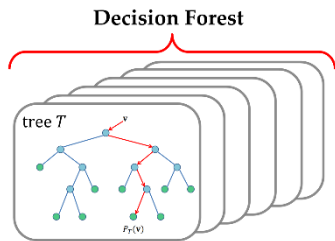
다양한 모델의 예측을
종합하여 결론 도출

68%

평균값

단일 모델보다 안정적인 예측이 가능

모델



Random Forest

XGBoost

DNN

SVM

ACC

93%

94%

94%

92%

F1 score

71

75

75

73

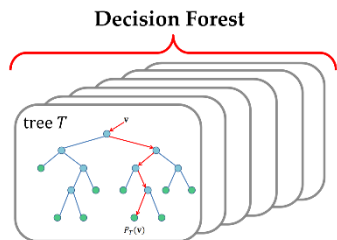
Entropy / Bagging

Entropy / Boosting

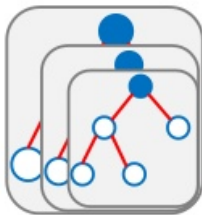
Neural network / ADAM

Margin / Rbf

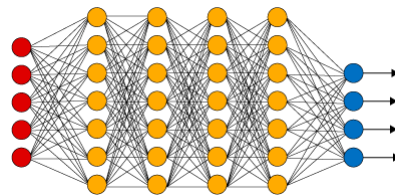
정확도



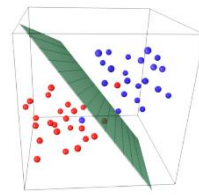
Random Forest



XGBoost



DNN

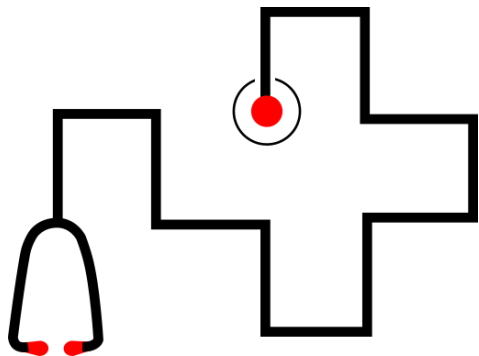


SVM



최종 정확도 **95%**

F1 score **77**



4. 중요 변수 분석



DL1_dg

아토피피부염 의사진단 여부



DJ4_dg

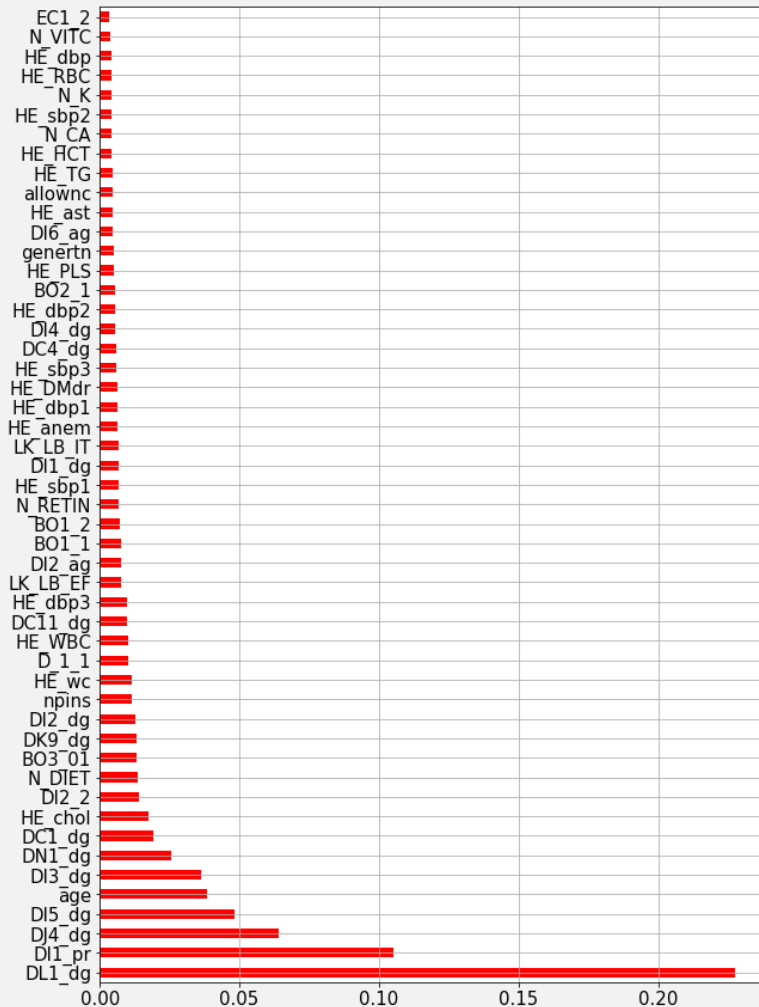
천식 의사진단 여부



HE_dbp3

이완기 혈압

Feature importances Top 50



HE_chol

총 콜레스테롤



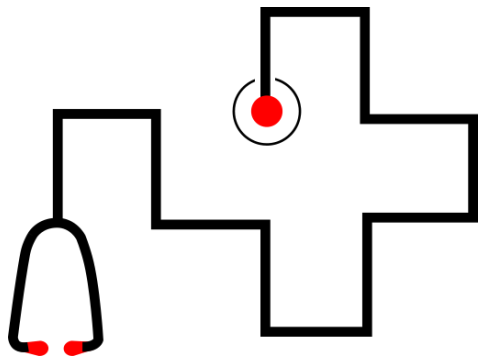
DI2_2

이상지질혈증 약복용주기



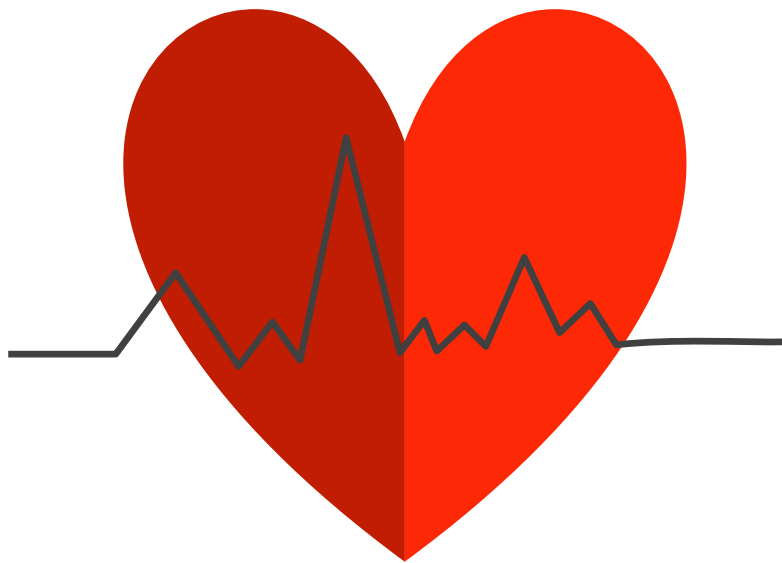
N_DIET , BO3_01

식사조절요법,
운동을 통한 체중감소



5. 가능성 및 보완점

가능성 및 보완점



01

예측 정확도 상승

모델 최적화를 통해 정확도 상승 가능성

02

더미데이터 생성

더미데이터를 생성하여 데이터 불균형문제를 해결할 가능성이 있음

03

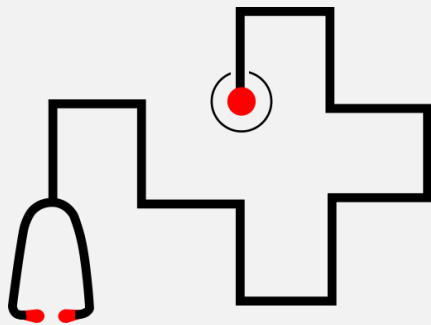
데이터 수집에 어려움

예측을 위해 많은 질의(변수)를 필요로 함

04

2종오류 가능성

당뇨 비해당자를 예측하는 정확도 보다
당뇨 해당자를 예측하는 정확도가 낮음



Thank you

감사합니다