Diabetes Prediction

1. Early detection of type 2 diabetes mellitus using machine learning-based prediction models (https://www.nature.com/articles/s41598-020-68771-z)

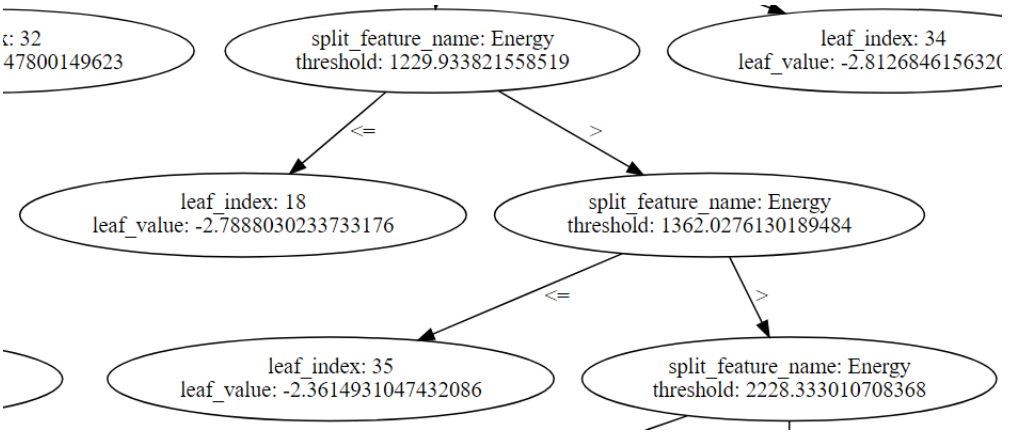
- object : 2형 당뇨병의 조기검출이 목적이며 그 도구로서 머신러닝을 활용 하겠다.

- data : 당뇨병 환자 데이터 (T2DM 사전 진단이 없는 27,050명의 성인 개인의 EHR)

- feature : 생물학적 및 임상적 특징

- methods
  - 결측치 및 이상치 처리
    - 평균에서 +-3표준편차 만큼의 경우 결측치로 대체 결측치가 50% 이상인 변수는 제거
    - 수치형 변수 결측치는 Bayesian linear regression으로 대체
    - 2개 class 범주형 변수는 logistic regression, 3개 이상의 class 범주형 변수는 polytomous regression으로 대체
    - 최종적으로 3723명의 환자 데이터를 쓰게 됨
  - 차원축소 : 111개 변수 -> 58개 변수
    - PCA등의 방식으로 줄이는게 아니라 결측치 처리하면서 제거된 것을 차원축소 라고 표현(?)
    - 결론에 보면 모델별로 변수 중요도에 차이가 나는것으로 확인 되는데, 변수간 상관관계 확인 후에 계수가 높은 변수들 확인되면 PCA등의 방식으로 변수 줄이고 모델 학습 시키는 과정 추가하는 것도 고려해 봄직 함
  - 사용 모델 : Glmnet, RF, XGBoost, LightGBM 을 일반적인 회귀분석 모델과 비교 검증
  - 평가지표 : 정확도, 민감도, 특이도, RMSE
    - 100회의 bootstrap을 통해 검증을 하며 검증 set은 각 bootstrap에서 선택되지 않은 샘플을 사용
    - 100회에 대한 95% 신뢰구간 구해서 성능지표로 사용
  - feature importance 제공

- 논문에서는 decision tree based model 들의 importance를 설명 할 때 모델에 대한 기여도의 순위로 제공이 되고 있는 것으로 보이는데, tree를 가져와서 node에 매겨지는 weight parameter 값을 가져와서 보여주고 그 다음에 기여도를 보여 주는 것이 현실에 반영해서 보여주기 더 좋지 않을까?



- 의견
  - Glmnet, RF, XGBoost, LightGBM 예측모델을 이렇게 4개 써서 비교 검증 했을 것으로 보이는데, 정형 데이터를 활용한 예측모델 만들 때 많이 쓰이는 ML 들입니다. 좀 더 나아가면 ensemble 혹은 voting 기법으로 더 나은 결과를 뽑을 수 있지 않았나 싶음
  - LightGBM 모델의 경우 학습 과정에서 선택되지 않는 node에 대한 weight를 버리는 특징이 있는것으로 알고 있는데, 이는 적은 data set(feature수가 많다면 아무리 row가 많은 data여도 상대적으로 적은 data set일 수 있습니다.)에 대해 over fitting 덜 일어나게 됨 또한 이러한 학습 방식 덕에 학습에 소요되는 computing power 가 매우 낮기 때문에 다양한 hyper parapmeter를 사용해서 가장 잘 맞춰주는 parapmeter를 찾기가 수월하기에 이런 결과가 나오지 않았을까 하는 의문
  - EHR 데이터를 추가로 활용 하는 부분의 의미 : 기존에 유사한 연구가 있을 때 차별화를 둘 수 있는 데이터를 추가하는 방식으로 연구 인정이 되는 케이스

2. Prediction of metabolic syndrome: A machine learning approach to help primary prevention (https://pubmed.ncbi.nlm.nih.gov/36029889/)

- object : 대사 증후군(Metabolic Syndrome)의 예측을 위해 기계 학습을 사용 하겠다.

- data : 17,182 adults attending a checkup program sequentially (37,999 visit pairs)
- feature : 사회 인구학적 특성, 임상, 실험실 및 라이프스타일 특성에 대한 변수
- methods
  - 결측치 및 이상치 처리
  - 차원축소
  - 사용 모델 : logistic regression, linear discriminant analysis, k-nearest neighbors, decision trees, Light Gradient Boosting Machine (LGBM), Extreme Gradient Boosting
  - 평가지표 : 민감도, 특이도, ROC-AUC

- 의견
  - 앞의 논문과 진행 방식이 매우 유사하여 ref에 앞의 논문이 있는지 찾아봤는데 없음.
  - 질병예측 분야에서는 이런 식의 목표 수행을 자주 할 것으로 보임

3. Development of Various Diabetes Prediction Models Using Machine Learning Techniques (https://pubmed.ncbi.nlm.nih.gov/35272434/)

- object : 건강 검진 데이터 활용해서 ML로 당뇨병 예측을 하겠다.

- data : 2009년부터 2018년까지 서울 성모 병원 건강 증진 센터 전자 의무 기록에서 추출, 최소 2회 이상 종합검진을 받은 대상자를 포함
  - 3,952명의 당뇨병 환자와 134,691명의 비당뇨병 환자 (이하, 3952, 134691)
- feature : 병원 검사 결과 62종, 국가 정기 건강 검진 변수 27종

- methods
  - Gradient Boosting, Random Forest
  - 10-cross-validation
  - 62개 변수 쓸 때 성능이 좋게 나왔고 단일 변수로 공복 혈당 변수가 매우 좋다.
  - 당뇨병 환자 정의 (y-value)
    - 자가 보고 당뇨병
    - 포도당 저하제 복용

        - 공복 혈당 수치 ≥126mg/dL 또는 당화혈색소(HbA1c) ≥6.5%. 전당뇨병은 공복 혈당 수치가 100~125mg/dL이거나 HbA1c가 5.7~6.4%인 상태
    - 4개의 예측모델 생성
        - 모델 1 : 당뇨병이 없는 피험자에서 1년후의 DM을 예측 (752, 26175) GB
        - 모델 2 : 당뇨병이 없는 피험자에서 2년후의 DM을 예측 (641, 33380) GB
        - 모델 3 : 당뇨병 대상자의 1년후 DN 예측 (519, 6345) GB
        - 모델 4 : 당뇨병 진단 1년전과 2년전의 차이를 학습하여 당요병 전단계인 피험자의 1년후 DM예측 (281, 3814) RF
    - 하이퍼 파라미터 랜덤으로 여러개 시도(?) -> 데이터 특성을 고려해서 일정 범위를 지정이라도 해주는게 좋을것으로 보임
    - 18~30개월은 2년, 8~16개월은 1년, 2회 이상의 경우 가장 최근 값을 사용
    - 성능과 변수 중요도를 보여줌
    - 변수간의 상관관계와 변수 중요도, 모델의 성능을 의학적 사실과 와 연결시켜 설명
- 의견
    - 데이터가 불균형 인게 문제일 것 같은데 5년 전쯤부터 smote로 oversampling 하는 방법을 GAN으로 해결 하는 방식이 흥행 하였음. 딱 올해 difusion model이 흥하고 있는데 이것도 생성 모델이라 마찬가지로 불균형 데이터 문제 해결할 수 있을거 같긴 한데 찾아본 결과 시도한 경우들이 종종 보임
        - https://arxiv.org/abs/1903.09730
        - https://arxiv.org/abs/2008.09202
        - https://koreascience.kr/article/CFKO201835372171472.pdf

1. Novel subgroups of adult-onset diabetes and their association with outcomes: a data-driven cluster analysis of six variables (https://www.thelancet.com/article/S2213-8587(18)30051-2/fulltext)

- object : 2형 당뇨병의 세분화된 분류를 통해 치료 방법을 제안 할 수 있는 도구를 만들자

- data : 스웨덴 당뇨병 환자 코호트에서 당뇨병 진단을 받은 환자 8,980명

- feature : 6개 변수(glutamate decarboxylase antibodies, age at diagnosis, BMI, HbA1c, and homoeostatic model assessment 2 estimates of β-cell function and insulin resistance)를 기반으로 한 합병증 발생 및 약물 처방에 대한 환자 기록 데이터
- methods
  - 같은 방식을 3개의 독립적인 코호트 Scania Diabetes Registry (n=1466), All New Diabetics in Uppsala (n=844), and Diabetes Registry Vaasa (n=3485)에 대해서 수행
  - COX regression, Logistic regression 를 써서 약물 치료 까지의 시간, 치료 목표에 도달하는 시간, 당뇨병 합병증 위험 및 유전적 연관성 비교를 수행
    - COX regression : 합병증 위험 계산
    - mle method : 군집과 유전자 간의 연관성 분석
    - OR(odds ratio) : 군집별 특성을 확인하는 작업
  - k-means clustering으로 군집화

- 결과
  - 환자 특성과 당뇨병 합병증 위험이 크게 다른 5개의 군집 추출
    - 군집3(인슐린 저항성이 가장 높은 환자군)은 군집4, 5 에 속한 환자보다 신장 질환의 위험이 높다
    - 군집2는 망막병증 위험이 높다
    - 각 군집의 특징을 뒷받침 하는 유전적 연관성은 2형 당뇨병 만으로는 확인할 수 없는 특징 이다. 따라서, 군집이 필요함
    - 군집화는 남성과 여성을 구분해서 수행 하였음 -> 남성, 여성을 구분해서 해야 할 정도로 명확한 의학적 근거가 있는지?
    - 유전 변이와의 연관성을 도출
- 의견
  - 신체, 생활, 가족력 등의 특징에 따라 4~n개 까지의 군집으로 나눠서 각 집단별로 어떤 요인에 의해 당뇨병이 발생하는지 찾아보고 집단별 learning을 수행 해보는것 고려

Cox AI

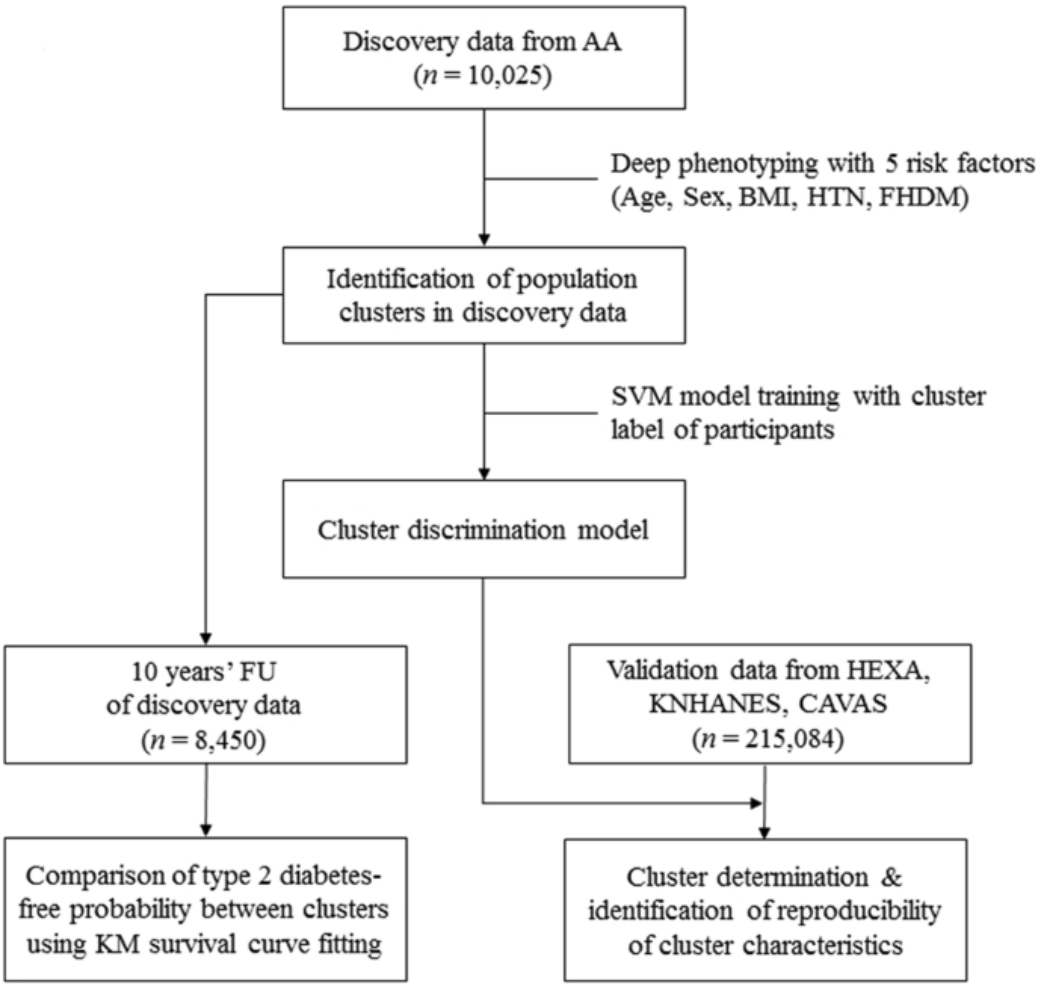
1. 다중 모드 데이터를 사용한 폐암 생존분석 검토 (https://manuscriptlink-society-file.s3-ap-northeast-1.amazonaws.com/kips/conference/2020fall/presentation/KIPS_C2020B0311.pdf)
   - image data 활용

1. DeepSurv: personalized treatment recommender system using a Cox proportional hazards deep neural network (https://bmcmedresmethodol.biomedcentral.com/articles/10.1186/s12874-018-0482-1)
    - 생존 분석에서 DeepSurv가 Cox보다 좋은 성능을 보이는 것을 증명
        - 공변량의 선형 및 비선형 효과 모두에서 생존 데이터에 대해 다른 생존 분석 방법과 같거나 더 나은 성능을 보인다.
        - 네트워크가 개인의 공변량과 치료 효과 간의 복잡한 관계를 학습할 수 있는 방법을 설명하기 위해 환자의 치료 그룹을 나타내는 추가 범주형 변수를 포함
        - 환자의 관찰된 특징에 맞는 치료 권장 사항을 제공
        - 환자에 대한 중간 생존 시간을 잠재적으로 증가시킬 수 있는 개인화된 치료 권장 사항을 제공
    - 데이터 : 10개의 feature를 쓰는데 균등 분포에서 추출한 값을 사용??

<결괏값에 영향을 미치는 경우가 다른 특정 집단 추출>

title : Identification of novel population clusters with different susceptibilities to type 2 diabetes and their impact on the prediction of diabetes

- data : KoGES(한국유전체역학연구) 안성/안산 코호트 데이터 (N=10,025) , 그 외 검증용 코호트 데이터 (N=215,084)
- feature : 연령, 성별, 체질량 지수, 고혈압 및 당뇨병 가족력
- intro
    - 제2형 당뇨병에서 이질성이 발견 됨
        - 포도당 또는 인슐린 프로필이 제2형 당뇨병 환자에게서는 예후가 다르게 나타날 수 있음
        - 그 외에 성별, 체질량 지수(BMI) 또는 인종 그룹 또한 이질적인 연관성을 보임
- methods

- 

  - 제2형 당뇨병 유병률의 이질성이 최대화될 때까지 여러 그룹으로 묶음
    - how? Gower's distance 값이 최대가 되게 하는 그룹이 나오도록 반복 수행 (https://elecs.tistory.com/381)
      - risk factor는 가중치를 더 줌
      - 그냥 상수배의 가중치를 주면 가까운 경우(0에 가까운)에도 값이 기존의 값보다 커지게 되는데 가까울수록 더 가깝고 멀수록 더 멀게 해주는 가중치는 뭐가 있을까?
        - 0~1 사이 값을 넣으면서 -무한대~무한대 값을 출력 해주고 출력 값에 limit을 걸면 되지 않을까 해서 생각난게 지수 함수, 탄젠트 함수 등..
  - 2개의 검증 데이터 세트로 검증
    - 위 결과로 SVM학습 : Y=군집, X=5개 risk factor
    - SVM으로 군집 달고 검증 (이렇게 하는게 맞냐..?)

- 결과
  -
- 의견

- 새로운 기술을 보이는 것도 좋지만 논문의 퀄리티를 올리기 위해서는 기존의 방식과 성능을 비교하는 과정이 필요하다고 느껴지며 이러한 문제를 해결하기 위한 기존의 기술에 대한 공부가 도음이 될 것 이라는 생각이 들었음.