WISET 동남권역 R-WeSET사업단

산학연계 여대생 팀프로젝트 결과발표회

당뇨병 발생 예측을 위한 다층 스태킹 앙상블 모델 구축 기법

성아영, 윤소현, 강수연, 김건우†
경상국립대학교 컴퓨터과학부

■ INTRODUCTION

- 대한당뇨병학회가 발표한 2022년 한국 당뇨병 팩트시트(Diabetes Fact Sheet in Korea 2022)에 따르면 30세 이상 성인 약 10명 중 4명이 당뇨병 전단계에 해당함
- 따라서 당뇨병 발생을 예측하는 모델을 개발하고자 하는 연구가 늘어나고 있음
- 다양한 특성으로 이루어져 있고, 이들 간의 복잡한 관계가 형성되는 의료 데이터의 특성상 단일 모델만으로는 데이터의 패턴을 충분히 학습하기 어려움
- 본 연구는 데이터에 맞춰 자동으로 다층 스태킹 앙상블 모델을 구성하는 알고리즘을 제안하며 이를 기반으로 스태킹 앙상블 모델을 구축함

■ METHOD

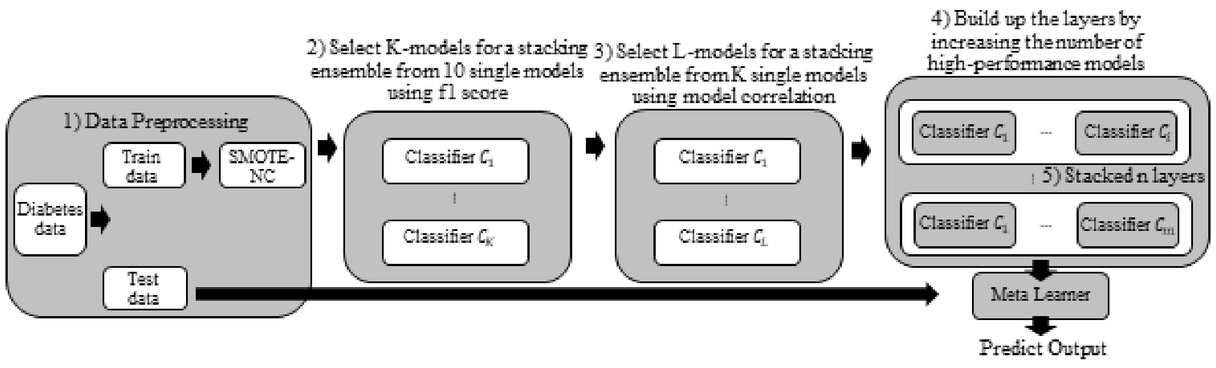

Fig. 1 Propoesd Process

Data Preprocessing

- 결측값 처리를 위해 KNN 알고리즘과 평균화 기법을 활용
- 결측값이 50% 이상일 때 해당 열 삭제
- Step-wise Feature Selection을 이용해 사용할 변수 축소
- 최종적으로 총 27가지 변수가 학습에 사용
- SMOTE-NC 알고리즘을 이용해 학습 데이터 증강

Build Automated Model

- 자주 사용되는 분류 모델 중 10가지(Naïve Bayes, Logistic regression, SGD Classifier, KNN, SVM, Decision tree, Random forest, lightGBM, XGBoost, Catboost)를 선정
- 10개의 모델 중 K-fold 교차 검증을 진행해 얻은 F1-score를 기준으로 성능 비교 후 성능이 낮은 모델을 일차적으로 제외
- 남은 모델들의 예측값을 비교해 모델들 간의 상관관계를 분석하고, 0.7 이상의 높은 상관관계를 보이는 모델 쌍에서 성능이 낮은 모델을 추가로 제외
- 최종적으로 남은 모델을 성능 순으로 정렬한 후, 모델의 개수를 점진적으로 하나씩 늘려가며 최적의 성능을 가지는 조합을 찾음
- 조합 탐색 과정에서 모델의 수를 추가해도 성능 향상이 없다면 반복을 중단하고 해당 조합을 최적의 단일 층으로 선택
- 앞 단계를 반복하며 최적의 성능을 가지는 조합을 찾고 층을 늘려나감
- 이전 층과 현재 층의 모델 구성이 같다면 반복을 종료
- 모델의 복잡도가 상당히 증가할 것이 예상되기 때문에 과적합 되는 것을 방지하고자 메타 학습기로 Logistic regression을 선택

■ EXPERIMENTS

Dataset

- 국내 코호트 데이터인 한국인유전체역학조사 자료를 활용
- 데이터의 클래스 분포를 고려하며 7:3의 비율로 나누어 실험 진행

Table. 1 Summary of Train and Test sets

Class	Train set	Test set	Total
정상	6004	2573	8577
당뇨	1017	436	1453

Proposed Model

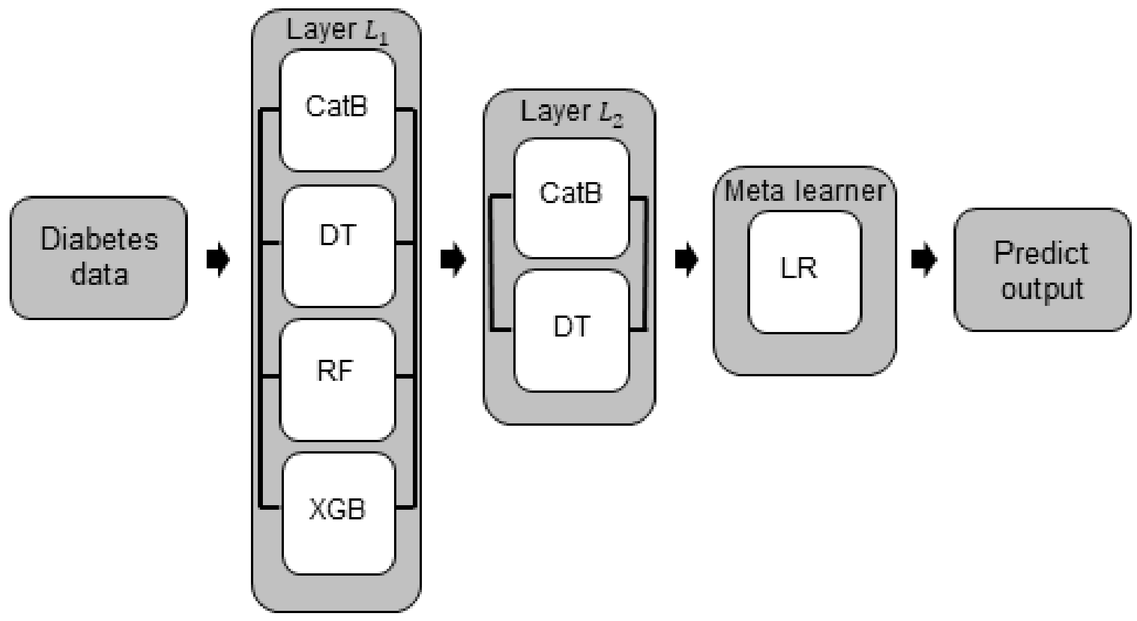

Fig. 2 stacking ensemble model using proposed method

- 첫 번째 층: CatBoost, Decision tree, Random forest, XGBoost
- 두 번째 층: CatBoost, Decision tree
- 메타 학습기: Logistic regression

Results

- 성능 비교 실험에는 자동 기계학습 라이브러리 중 TPOT, Pycaret, Autogluon 세 가지를 사용해 진행
- 표 2는 각 라이브러리에서 도출된 최적의 조합 및 성능 비교 결과

Table. 2 Performance comparison with other methods

Methods	Accuracy	F1-score	ROC AUC
TPOT (ExtraTree)	0.752	0.416	0.698
Pycaret (CatB, XGB, LightGBM)	0.863	0.3439	0.609
Autogluon (WeightedEnsemble_L2)	**0.867**	0.3789	0.625
Proposed Model	0.851	**0.4728**	**0.700**

■ CONCLUSION

- 본 연구에서는 당뇨병 발생을 예측하기 위해 다층 스태킹 앙상블 모델을 제안 했음
- 모델은 다른 라이브러리에서 구축한 모델과 비교해 F1 score 기준으로 최대 12.89%p의 성능 향상을 보였음
- 다른 autoML 라이브러리와 다르게 하이퍼 파라미터 튜닝을 진행하지 않았음에도 다른 방법과 비교해 비슷하거나 높은 성능을 보였음
- 후속 연구로 하이퍼 파라미터 튜닝을 수행하는 알고리즘을 추가한다면 성능을 더 높일 수 있을 것으로 기대됨