공공 건강 데이터 기반 당뇨병 예측 모델

심민경, 이석현, 김상대

순천향대학교 의료IT공학과

minkyoung.shim03@gmail.com, shlee3930@naver.com, sdkim.mie@sch.ac.kr

Health Data-Based Diabetes Prediction Model

Minkyoung Shim, Seokhyeon Lee, Sangdae Kim

Dept. of Medical IT Engineering, Soonchunhyang University

요 약

본 연구는 국민건강보험공단의 건강검진 및 진료내역 데이터를 활용하여 당뇨병 예측 모델을 제시하였다. 결측치는 Iterative Imputer로 보정하고 BMI, 맥압, TG/HDL, LDL/HDL 등의 파생 변수를 추가하였다. 데이터 불균형 문제는 SMOTETomek 기법으로 개선하였으며, 일부 정보가 부족한 경우에도 예측이 가능하도록 설계하여 사용자 편의를 반영하였다. CatBoost, LightGBM, XGBoost를 기반으로 Stacking 앙상블을 적용한 결과, 당뇨 환자군의 재현율을 높이며 전체적으로 균형 잡힌 성능을 확보할 수 있었다.

Ⅰ. 서 론

질병관리청에 따르면 우리나라 30세 이상 성인의 당뇨병 유병률은 2021년 16.3%로 약 600만 명이 앓고 있는 것으로 추정된다. 당뇨병 전(前)단계까지 포함하면 성인의 절반 이상이 당뇨병에 대한 관리가 필요하다. 문제는 초기 증상이 거의 없어 진단이 늦어지는 경우가 많다는 점이며, 젊은 층에서 치료가 지연될수록 질환 기간이 길어지고 합병증 위험도 커질 수 있다.

대한당뇨병학회에 의하면 혈당검사, 표준포도당 부하검사, 당화혈색소, 흡연, 음주 등의 요소가 당뇨병에 영향을 미치는 주요 요인으로 알려져 있다. 이와 같이 당뇨병에 영향을 미치는 요인들은 건강검진 항목[1]에도 포함되어 있으며, 이를 활용하면 당뇨병의 조기 발견과 예측에 유용할 수 있다[2]. 그러나 혈액검사와 같은 세부 검진 항목은 개인이 일상적으로 직접 알기 어려운 경우가 많아, 이를 고려한 예측 모델의 설계가 필요하다.

본 논문에서는 모든 건강정보를 완전하게 수집하기 어렵다는 점을 고려하여, 일부 정보가 부족하더라도 예측이 가능하도록 모델을 설계하였으며, 이를 통해 젊은 층의 건강 관리와 합병증 위험 완화에 기여하고, 고위험군을 조기에 선별하는 데 도움이 될 것으로 기대된다.

Ⅱ. 본론

Ⅱ-1.시스템 구성요소

○개발환경 : python

○데이터 관리 환경 : oracle sqldeveloper

○데이터 출처 : 국민건강보험공단_건강검진, 진료내역정보(2023)

Ⅱ-2. 시스템 데이터 관리

본 논문에서는 당뇨병 예측 모델을 구축하기 위해 대용량 데이터를 처리하는 데 적합한 오라클 데이터베이스를 활용하여 두 개의 주요 테이블을 구성하였다. 첫 번째 테이블은 건강검진[3]테이블이며, 두 번째 테이블은 진료내역[4] 테이블이다. 각 테이블은 가입자 일련번호를 기준으로 관리되며, 두 테이블은 이 가입자 일련번호를 통해 서로 통신하여 연결된다.

Ⅱ-3. 데이터 불균형

원본 데이터에서 당뇨병 환자(양성 클래스)의 비율은 약 6:1로 불균형한 분포를 보였다. 이러한 불균형은 소수 클래스의 예측 성능 저하, 특히 정밀도(precision)와 재현율(recall)의 불안정을 초래할 수 있다. 이를 해결하기 위해 본 연구에서는 SMOTETomek 기법을 적용하였다.

이 기법을 통하여 소수 클래스의 새로운 합성 샘플을 생성하여 데이터 불균형을 완화하는 기법이며, 클래스 간 경계에 위치한 중복 혹은 불확실한 샘플을 제거하여 데이터의 분리를 명확히 하는 언더샘플링 기법이다.

원본 데이터에서 비당뇨군(0)과 당뇨군(1)의 비율은 약 6:1로 불균형한 분포를 보였다. 학습 데이터에서도 유사한 비율이 관찰되었으나, SMOTETomek 기법을 적용한 이후 두 클래스가 각각 144,534건으로 [그림1]과 같이 균형을 이루게 되었다

데이터 구분	비당뇨(0)	당뇨(1)	비율(0:1)
원본 데이터	180693	30403	180693 : 30403
이전 데이터	144554	24322	144554 : 24322
샘플링 후 데이터	144534	144534	144534 : 144534

그림 1. 불균형 처리 데이터

```
imputer = IterativeImputer(random_state=42)
X_imputed = imputer.fit_transform(X)
X = pd.DataFrame(X_imputed, columns=X.columns)
X_train, X_test, y_train, y_test = train_test_split(
    X, y, test_size=0.2, stratify=y, random_state=42
)
```

그림 2. Iterative Imputer 기법적용

Ⅱ-4. 결측치 처리

결측치 처리는 Iterative Imputer 기법을 적용하여 다변량 회귀 기반으로 누락값을 보정하였다. 이후 데이터는 8:2 비율의 Stratified split을 통해 학습용과 평가용으로 분할하였다. 학습 데이터의 클래스 불균형 문제를 해소하기 위해 SMOTETomek 기법을 적용하여 과 샘플링과 언더 샘플링 [그림2]와 같이 동시에 수행하였다.

Ⅱ-5. 변수 선택

변수 선택 과정에서는 혈압, 혈당, 콜레스테롤 등 핵심 임상 지표를 포함하였으며, 사용자가 비교적 쉽게 파악할 수 있는 기존 변수를 활용하여 체질량지수(BMI), TG/HDL, LDL/HDL과 같은 파생 변수를 생성하였다. 상관계수 분석 결과, 식전혈당은 당뇨 여부와 가장 강한 양의 상관관계를 보였으며(0.73), 허리둘레, 혈압 및 콜레스테롤 관련 지표 또한 [그림2]와 같이 확인되었다.

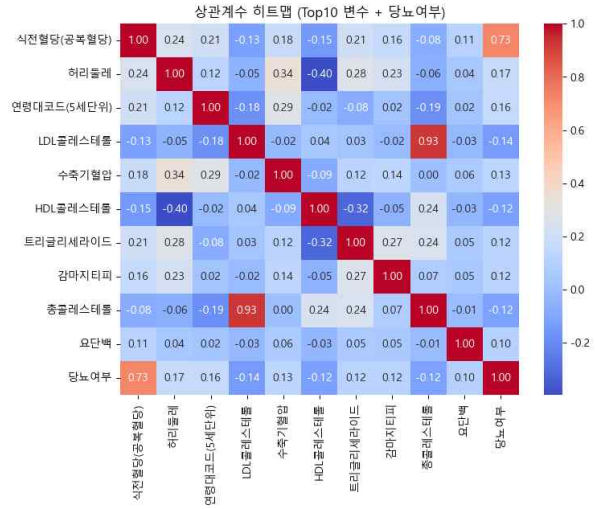

그림 3. 상관관계 히트맵

Ⅱ-6. 모델 학습

모델 학습 단계에서는 CatBoost, LightGBM, XGBoost를 기본 분류기(base learner)로 설정하였으며, 각 모델에는 클래스 가중치(class weight)를 적용하여 불균형 문제를 반영하였다. 이후 세 가지 모델을 기반으로 한 Stacking 앙상블을 구축하고, 최종 메타 모델로는 CatBoost를 활용하였다.

평가 단계에서는 모델이 산출한 예측 확률을 바탕으로 Precision-Recall curve를 작성하고, Precision·Recall·F1-score의 평균값이 최대가 되는 지점을 최적 임계값으로 설정하였다. 최종적으로 이 임계값을 기준으로 classification_report를 산출하여 모델의 성능을 평가하였다.

Ⅲ. 실험 결과

비당뇨군(클래스 0)은 정밀도 0.96, 재현율 0.95, F1-score 0.95로 안정적인 결과를 보였으며, 당뇨군(클래스1)은 정밀도 0.70, 재현율 0.76, F1-score 0.73으로 나타나 불균형 데이터 특성을 고려했을 때 비교적 양호한 수준으로 전체 정확도는 0.92였고, macro 평균 F1-score는 0.84로 클래스 간 균형 있는 예측 성능을 [그림4]와 같이 확인할 수 있었다. 특히 소수 클래스인 당뇨군에서 재현율이 0.76로 확보되어 실제 진단 과정에서 환자를 놓칠 가능성을 보여주었다.

	precision	recall	f1-score	support
0	0.96	0.95	0.95	36139
1	0.70	0.76	0.73	6081
accuracy			0.92	42220
macro avg	0.83	0.85	0.84	42220
weighted avg	0.92	0.92	0.92	42220

그림 4. 성능지표

Ⅳ. 결론

본 연구에서는 국가건강검진 및 진료내역 데이터를 활용하여 당뇨병 예측 모델을 제시하였다. SMOTETomek을 이용한 불균형 보정과 Stacking 앙상블 기법을 적용하여 성능의 균형을 확보하였고, 당뇨 환자군의 재현율을 높여 고위험군 조기 선별에 기여할 가능성을 제시하였다. 또한 모든 건강정보를 완전하게 수집하기 어려운 현실을 고려하여, 일부 정보가 부족한 상황에서도 예측이 가능하도록 모델을 설계함으로써 사용자 편의를 반영하였다. 다만 데이터 범위와 변수 구성이 제한적이어서 성능 개선에는 제약이 있었으며, 향후 더 다양한 변수와 최신 알고리즘을 반영하고 외부 데이터 검증을 수행한다면 모델의 정확도와 일반화 수준을 한층 높일 수 있을 것으로 기대된다.

ACKNOWLEDGMENT

"본 연구는 2025년 과학기술정보통신부 및 정보통신기획평가원의 SW중심대학사업의 연구 결과로 수행되었음"(2021-0-01399)

참 고 문 헌

[1]곽찬희, 김가현, and 양희림. "건강검진항목 및 진료정보를 활용한 총 처방일수 예측모델 개발." 한국컴퓨터정보학회 학술발표논문집 32.2 (2024): 943-944.

[2] 젊어지는 당뇨병 환자…20~30대 30만명·전단계 300만명 https://n.news.naver.com/mnews/article/001/0015035714?sid=102

[3] 국민건강보험공단_건강검진정보 https://www.data.go.kr/data/15007122/fileData.do

[4] 국민건강보험공단_진료내역정보 https://www.data.go.kr/data/15007115/fileData.do