건강검진 항목 및 진료 정보를 활용한 당뇨병 예측 모델 개발

심민경, 김상대

순천향대학교 의료IT공학과

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

Diabetes Prediction Using Health Checkups and Outpatient Records

Minkyoung Shim, Sangdae Kim

Dept. of Medical IT Engineering, Soonchunhyang University

요 약

당뇨병은 인슐린 분비 이상으로 발생하는 대사 질환으로 초기 증상이 거의 없어 진단이 지연되는 경우가 많다. 본 논문은 국가건강검진과 진료데이터를 사용하여 당뇨병 예측 모델을 구축하였다. 데이터 불균형 문제는 언더샘플링을 하여 XGBoost기반 모델에 하이퍼 파라미터 튜닝을 적용하여 성능을 개선하였다. 제안된 모델은 향후 모바일 또는 웹 기반 서비스를 확장하여 당뇨병 조기 발견과 예방적 관리에 활용될 수 있을 것으로 기대된다.

Ⅰ. 서론

당뇨병은 인슐린 분비 이상으로 발생하는 대사질환으로, 고령, 비만, 스트레스 등이 주요 요인이다. 과거 중장년층 질환으로 여겨졌으나, 최근에는 젊은 층에서도 발병률이 증가하고 있다. 대한당뇨병학회 2024에 따르면[1], 국내 19~39세 청년 약 2.2%(30만 8천 명)가 당뇨병 환자로 추산되며, 이는 고칼로리 식습관, 비만, 운동 부족, 스트레스, 불규칙한 수면 등 생활습관 및 환경적 요인과 관련이 있다. 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 항목에도 포함되어 있으며, 이를 활용하면 당뇨병의 조기 발견과 예측에 유용할 수 있다.

이에 본 연구는 국가건강검진 데이터[2]와 진료 코드[3]를 활용하여 음주, 흡연 등과 같이 환경요인을 함께 고려하여 당뇨병 예측 모델을 개발하고자 한다. 이를 통해 당뇨병의 조기 발견과 예방적 관리에 기여하며, 나아가 젊은 층의 건강 개선과 합병증 위험 감소를 도모하고자 한다. 특히, 제안된 모델을 활용하여 고위험군을 미리 찾아내어 정기적으로 관리하고, 적절한 치료 계획을 세우는 데 도움이 될 것으로 기대된다.

Ⅱ. 본론

Ⅱ-Ⅰ.시스템 구성요소

본 연구에서 제안하는 시스템에서는

○개발환경 : python

○데이터 관리 환경 : oracle sqldeveloper

○데이터 출처 : 국민건강보험공단_건강검진,진료내역정보(2023)

Ⅱ-Ⅱ. 시스템 데이터 관리

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

Ⅱ-Ⅲ. 데이터 불균형

본 연구에서는 건강검진 데이터를 기반으로 한 당뇨병 예측 모델을 구축하는데, 심각한 클래스 불균형 문제가 확인되었다. 실제로 전체 데이터 중 당뇨병 환자의 수는 87명에 불과했지만, 비당뇨 환자는 약 999,913명에 달하였다. 이러한 불균형은 모델이 비 당뇨에 치우친 예측을 하게 만들 수 있으므로, 데이터의 대표성을 확보하기 위한 사전 처리가 필요하였다.

이에 본 연구에서는 언더샘플링 기법을 적용하여 소수 클래스인 당뇨 환자 데이터를 모두 유지 후, 다수 클래스인 비당뇨 환자 중 동일한 수만큼을 무작위로 추출하여 그림1과 같이 샘플링 하였다.

```
언더샘플링 후 데이터 수:
label
1    87
0    87
Name: count, dtype: int64
```

그림 1. 언더샘플링 후 데이터 수

최종적으로 구성된 언더샘플링 데이터셋은 총 N=2n 개로, 당뇨 환자 n명과 무작위로 선택된 비당뇨 환자 n명으로 구성된다.

Ⅱ-Ⅳ. 모델선택

예측 모델로는 트리 기반 앙상블 모델인 XGBoost (Extreme Gradient Boosting를 선정하였다. XGBoost는 경사 하강 방식의 부스팅 알고리즘으로, 오차를 보완하며 학습해가는 특성이 있어 불균형 데이터에 상대적으로 강건한 성능을 나타낸다. 또한, 결측치 처리 및 특성 중요도 해석이 용이하다는 점에서 의료 데이터를 분석하는 데에 적합하다고 생각하여 선택하였다.

Ⅱ-Ⅴ. 하이퍼 파라미터 튜닝

본 논문에서는 하이퍼 파라미터 최적화를 통해 분류 성능을 향상하고자 하였다. 그림2에서와같이 분류기는 XGBClassifier 클래스를 기반으로 구현하였으며, 목적 함수는 binary:logistic, 평가지표는 logloss로 설정하였다. 모델 학습의 일관성을 위해 random_state=42를 지정하였다.

또한, 하이퍼 파라미터 탐색은 GridSearchCV를 활용하여 3겹 교차 검증을 기반으로 수행하였고, 평가지표는 불균형 데이터에서의 성능을 반영할 수 있도록 F1-score로 설정하였다. 탐색 대상 파라미터는 그림2와 같이 다섯 가지이며, 총 32개 조합에 대해 병렬로 탐색하였다.

```
# 하이퍼파라미터 설정
param_options = {
    'n_estimators': [100, 200],
    'max_depth': [3, 5],
    'learning_rate': [0.01, 0.1],
    'subsample': [0.8, 1.0],
    'colsample_bytree': [0.8, 1.0]
}

# GridSearchCV를 통한 성능 최적화
search = GridSearchCV(
    estimator=model,
    param_grid=param_options,
    cv=3,
    scoring='f1',
    verbose=1,
    n_jobs=-1
)
```

그림 2. 하이퍼 파라미터 설정코드

Ⅲ. 실험 결과

혼동행렬:

```
[[15  3]
 [ 1 16]]
```

분류:

	precision	recall	f1-score	support
비당뇨	0.94	0.83	0.88	18
당뇨	0.84	0.94	0.89	17
accuracy			0.89	35
macro avg	0.89	0.89	0.89	35
weighted avg	0.89	0.89	0.89	35

성능:

	정밀도	재현율	F1 점수	샘플 수
비당뇨	0.938	0.833	0.882	18.000
당뇨	0.842	0.941	0.889	17.000
accuracy	0.886	0.886	0.886	0.886
macro avg	0.890	0.887	0.886	35.000
weighted avg	0.891	0.886	0.886	35.000

그림 3 모델 테스트 결과

위 결과와 같이 최적화된 모델은 테스트 세트에 대해 정확도 88.6%, 비당뇨군 F1-score 0.882, 당뇨군 F1-score 0.889를 기록하였다. 특히 당뇨군에 대한 재현율은 0.941로 매우 높게 나타나, 실제 당뇨 환자 분류 성능이 우수함을 보여준다. 이로써 하이퍼 파라미터 최적화가 모델의 전반적인 균형성과 민감도를 향상하는 데 효과적임을 확인할 수 있었다.

Ⅳ. 결론

본 논문에서는 국가건강검진 데이터와 진료 기록, 그리고 생활습관 정보를 기반으로 당뇨병을 조기 예측할 수 있는 모델을 구축하였다. 극단적인 클래스 불균형 문제는 언더샘플링 기법을 통해 효과적으로 조정하였고, 예측 모델로는 트리 기반 앙상블 기법인 XGBoost를 활용하였다. 또한, GridSearchCV를 통한 하이퍼 파라미터 최적화를 통해 모델의 성능을 극대화하였다.

향후 제안 모델 기반으로, 개인이 건강검진 결과 정보를 입력하면 당뇨병 고위험 여부를 예측해주는 모바일 앱 또는 웹 기반 서비스로 확장하는 방안을 고려할 수 있다. 이를 통해 사용자는 보다 쉽게 자신의 건강 위험을 인지하고, 조기 진단 및 예방하고 조치할 수 있을 것으로 기대된다.

ACKNOWLEDGMENT

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

참 고 문 헌

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

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

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