국립암센터

EMR 질환 예측 모델 개발 Docs

개발기간: 2023.03 - 2024.04 (1년 1개월) | 역할: 과제연구원 | 기여도: 50% (데이터 정형화, 분석, 모델링 전담)

Python 3.8+ Scikit-learn Latest XGBoost 1.6+ SHAP Explainable AI

프로젝트 개요

목적	**비정형 EMR 데이터를 활용한 고성능 질환 예측 모델 개발로 의료진 진단 지원 시스템 구축**
핵심 도전	수기 입력으로 인한 심각한 데이터 품질 문제 + 기존 연구 대비 예측 성능 한계 돌파

프로젝트 성과

정규표현식과 사분위수 이상치 탐지로 3,000명 비정형 EMR 데이터 100% 정형화 완료	의료진 협업 기반 BMI, CCI 파생 변수 생성으로 기존 논문 대비 45% 성능 향상 (ROC-AUC 0.61→0.87)
Logistic Regression + class_weight로 3:7 클래스 불균형 해결	ETL 파이프라인으로 반복 실험 80% 자동화 및 SHAP 기반 모델 해석성 확보

정량적 성과

지표	달성값	측정 방법
예측성능 향상	ROC-AUC 0..87	기존 동일 주제 논문 대비 45% 향상 (0.61 → 0.87)
데이터 정형화	약 3,000명 환자	비정형 EMR 데이터 정형화
자동화 효율성	약 80% 시간 단축	ETL 파이프라인으로 반복 실험 효율화 (20분 → 3분)

문제 해결 과정

1) 비정형 EMR 데이터 품질 문제

상황: 수기 입력으로 인한 심각한 데이터 품질 문제 (키 250cm, 체중 1130kg 등)
제약: 1. 의료 데이터 특성상 높은 정확도와 신뢰성을 요구
2. 의료 및 주제에 따른 도메인 지식이 필수
해결 과정:
1. 체계적 정규표현식 설계 → 15개 이상의 정형화 모듈로 다양한 단위 표현을 표준화, 정규표현식으로 숫자 추출
2. 통계적 이상치 탐지 → IQR 방법으로 논리적 오류 식별 및 제거
3. 의료진 협업 검증 → 2달 / 1회 대면 미팅 + 메일로 컨택하여 이중 검증 수행
결과: **15개 이상의 정형화 Script 구축** 및 3,000명 환자 데이터 정형화 완료

2) 예측 성능 제한 문제 (기존 논문 ROC-AUC 0.61)

상황: 기존 동일 주제 모델 보다 **개선된 예측 성능을 가진 모델 개발**
제약: 의학적 타당성과 통계적 유의성을 동시에 만족하는 변수와 개선된 모델 성능 필요
해결 과정:
1. 의학 문헌 조사 → BMI, CCI, 복합 위험인자 등 표준 의료 지표 및 논문 검토
2. 의료진 협업 변수 설계 → 연령에 대한 동반질환, 흡연 당뇨 등 임상적 의미있는 파생 변수 생성
3. 통계적 유의성 검증 → t-test, Mann-Whitney U test, Shapiro-Wilk test 모든 변수 $p<0.05$ 수준 변수 선택
4. Feature Selection 적용 → 통계적 분석과 ML 모델 비교
5. 정형화 과정과 모델 학습, 검증 Pipeline 구성 → 일관성 확보 및 오류를 감소하기 위한 작업으로 **처리시간 80% 단축**
6. 데이터 불균형 해결을 위한 Classweight 및 GridSearch 적용 → 7:3 데이터 불균형 해소와 모델 성능 향상
결과: LogisticRegression 모델, **ROC-AUC 0.61 → 0.87 (45% 향상)**. 결과 논문 2편 작성

3) 모델 해석성 확보

상황: 의료 분야 특성상 예측 근거 설명이 반드시 필요
제약: 1. 의료진이 이해할 수 있는 수준의 설명 필요
2. 통계분석과 ML 모델간 일관성 있는 해석 필요
해결 과정:
1. **SHAP 기반 모델 해석** 진행
1) 모든 머신러닝 모델 학습 및 검증에 SHAP 기반 모델 해석 추가
2) 각 결과 Graph를 참고하여 모델 해석 가능성 확보
2. 통계 분석과 ML 결과 비교
1) 동일한 전처리 과정으로 정형화된 데이터 활용
2) LogisticRegression 통계 분석 (statsmodel)과 비교
결과: **기존 변수 및 파생 변수에 대한 신뢰성 확보**