[지도교수] 한창희 교수님

공공의료 데이터 기반 고위험 의료비 예측 및 시각화 시스템

2024270655 이유라 & 2024270636 황유림

BACKGROUND

공공의료 통계만으로는 고비용 위험군의 신속한 식별과 해석이 어려움

연령·성별·행위코드 기반 위험 예측과 시각화를 결합한 분석 시스템 구현

DEFFERENT

공공의료 통계를 단순 조회가 아닌 예측 가능한 위험 도 분석 문제로 확장

상위 20% 비용 구간을 고위험군으로 직접 정의해 해석 가능한 기준 제시

분석 결과를 대시보드와 예측 화면으로 연결해 실제 활용 흐름까지 구현

BUILD

데이터 및 전처리

HIRA 공공의료 CSV 4종을 수집하고 분석용 데이터셋으로 통합

결측치 제거, 건당 평균 진료비 계산, 상위 1% 이상치 제거 수행

상위 20% 비용 구간을 high_risk로 정의해 학습용 라벨 구성

모델 및 시스템

One-Hot Encoding과 Logistic Regression으로 분류 모델 학습

비용 정답 컬럼은 입력에서 제외해 라벨 누수를 방지

Streamlit 기반 화면에서 위험 확률을 바로 확인할 수 있게 구성

RESULT

성능 요약

Accuracy	Precision	Recall	ROC-AUC
0.957	0.847	0.961	0.990

분석 데이터 228,520건 | 고위험 비율 20.1% | 기준 진료비 573,226원

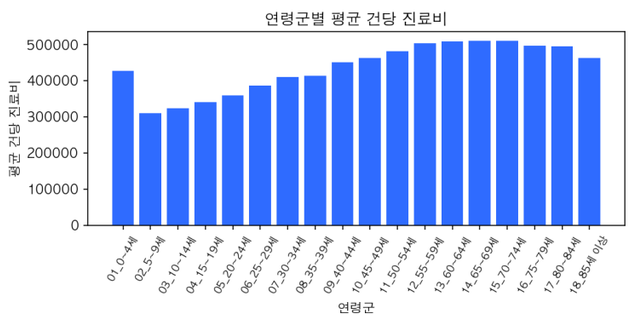
핵심 해석

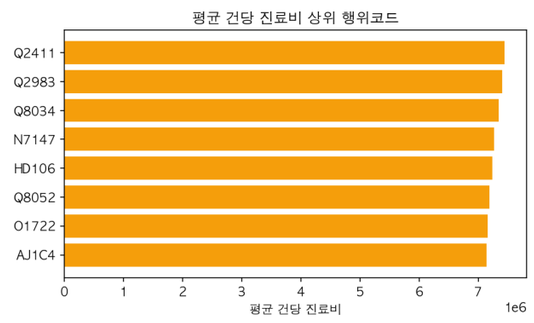
고위험군 비율은 전체의 20.1% 수준으로 나타남

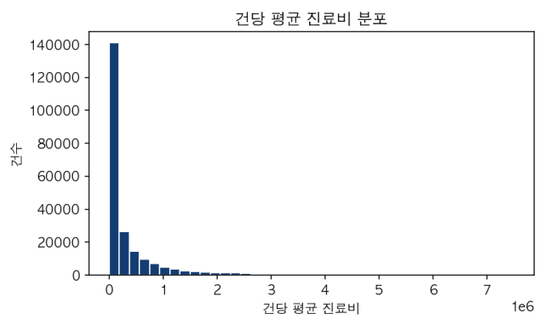
65세 이상 연령군에서 평균 진료비가 상대적으로 높게 관찰됨

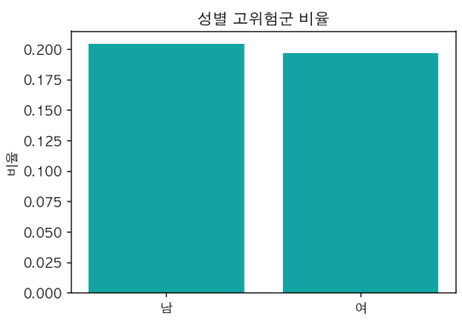
특정 행위코드는 건당 평균 진료비가 700만원 이상으로 집중됨

주요 시각화









SCREEN

EXPECTATION

활용 및 의의

공공의료 데이터의 복잡한 비용 패턴을 직관적으로 요약하고 비교 가능

고위험 의료비 발생 가능성을 정량적으로 예측해 분석 근거를 제공

후속 연구에서 질병군, 지역, 의료기관 유형 변수까지 확장 가능한 기반 마련

요약

문제정의 – 공공의료 데이터는 많지만 직관적 해석 도구가 부족함

차별점 – 비용 위험 라벨을 직접 정의하고 예측까지 연결함