머신러닝 모델 학습절차 분석

202476624 권호근

Ⅰ. 서론

1. 연구 배경

- 당뇨병은 현대 사회에서 가장 흔한 만성 질환 중 하나로, 조기 예측 및 예방이 매우 중요
  특히 혈당, 체질량지수(BMI), 연령 등 기본적인 건강 데이터만으로도 당뇨병 발병 위험을
  어느 정도 예측할 수 있어 머신러닝 기반의 분석이 활발히 이루어지고 있음
- 본 연구에서는 Kaggle의 Pima Indians Diabetes 공개 의료 데이터를 활용하여 여러
  머신러닝 모델을 비교·분석하고, 당뇨병 여부 예측의 정확도와 중요 변수를 도출하는
  것을 목표로 함

2. 연구 목적

- 건강 지표를 기반으로 당뇨병 발병 여부를 예측하는 머신러닝 모델 구축
- 모델별 성능 비교(Logistic Regression, Random Forest, SVM)
- 데이터 분석을 통한 당뇨병 위험 요인 도출
- 실험 결과를 기반으로 의료적 시사점과 모델 개선 방향 제안

Ⅱ. 연구 방법

1. 데이터셋 개요

- 데이터 출처: Kaggle – Pima Indians Diabetes Database
- 전체 샘플 수: 768개
- 특성(Feature): 8개
- 타깃(Target): Outcome (0 = 비당뇨, 1 = 당뇨)
- 구성 변수

변수명	설명
Pregnancies	임신 횟수
Glucose	포도당 수치
BloodPressure	이완기 혈압
SkinThickness	피부 두겹 두께
Insulin	혈중 인슐린
BMI	체질량지수
DiabetesPedigreeFunction	가계력 기반 당뇨 지수
Age	나이
Outcome	당뇨 여부

2. 데이터셋 전처리

- 0 값 처리: 생리적으로 0이 될 수 없는 항목(Glucose 등)은 결측치로 간주하고
  중간값(median)으로 대체
- 정규화(Scaling): StandardScaler를 이용해 모든 변수 스케일 조정
- 훈련/테스트 분리

· Train:Test = 80:20

· stratify=Outcome 적용하여 클래스 비율 유지

3. 탐색적 데이터 분석(EDA)

EDA를 활용한 당뇨/비당뇨 그룹 간 차이 분석

(1) 변수별 평균 비교

- Glucose

· 비당뇨: 109.9

· 당뇨: 140.1

- BMI

· 비당뇨: 30.4

· 당뇨: 35.2

- Age

· 비당뇨: 31.2

· 당뇨: 37.1

→ 당뇨 그룹이 모든 주요 변수에서 높은 값을 보임

(2) Outcome과의 상관관계

변수명	상관계수
Glucose	0.49
BMI	0.29
Age	0.23
Pregnancies	0.22

→ 포도당 수치가 Outcome에 가장 큰 영향을 미침

Ⅲ. 모델 학습 및 평가

1. 실험 활용 모델

- Logistic Regression
- Random Forest Classifier
- Support Vector Machine (RBF Kernel)

Ⅳ. 실험 결과

1. 1차 모델 성능 비교 (기본 모델)

모델	Accuracy	Precision	Recall	F1-score
Logistic Regression	0.803	0.693	0.667	0.680
Random Forest	0.779	0.653	0.612	0.632
SVM (RBF)	0.753	0.624	0.600	0.611

2. 해석

- 정확도(Accuracy): Logistic Regression이 가장 높음
- 재현율(Recall): 당뇨 환자를 놓치지 않는 비율

· Logistic Regression이 가장 우수(0.667)

- 의료 분야에서는 환자를 놓치지 않는 것이 중요 → Logistic Regression 우수

Ⅴ. 모델 최적화(Hyperparameter Tuning)

1. 최적화 대상 모델

- Random Forest
- GridSearchCV로 n_estimators, max_depth, min_samples_split 튜닝

2. 최적화 결과

- Best Params : n_estimators=50, max_depth=5, min_samples_split=5

3. 성능

변수명	상관계수
Accuracy	0.784
Precision	0.678
Recall	0.612
F1-score	0.643

4. 결론

- 최적화 후에도 Logistic Regression보다 성능 열세
- 특히 Recall이 낮아 의료적 활용성 떨어짐

→ 최종 모델은 Logistic Regression 선정

Ⅵ. 최종 분석: 중요 변수 해석

1. Logistic Regression 계수 분석

- 영향도 순서:

1) Glucose(포도당)
2) BMI(비만도)
3) Age(나이)
4) Pregnancies(임신 횟수)

→ Glucose는 가장 강력한 위험 인자

→ BMI와 Age 또한 당뇨병 판단에서 중요한 역할

2. 의료적 의미

- 혈당 관리가 가장 중요
- 비만과 고령이 당뇨 위험을 크게 증가
- 임신 횟수 증가도 일정 위험을 갖는다