EP

25F-

EP4-1

2025 대한의료정보학회 추계학술대회

Generative AI in Healthcare Systems: From Insight to Impact

한국인 성별 맞춤형 당뇨 예측 모델 개발

김정희*, 김유빈*, 서주희*, 정재효†, 박주용*,†

* 을지대학교 바이오융합대학 빅데이터의료융합학과

† 을지대학교 첨단학부 빅데이터인공지능전공

[배경]

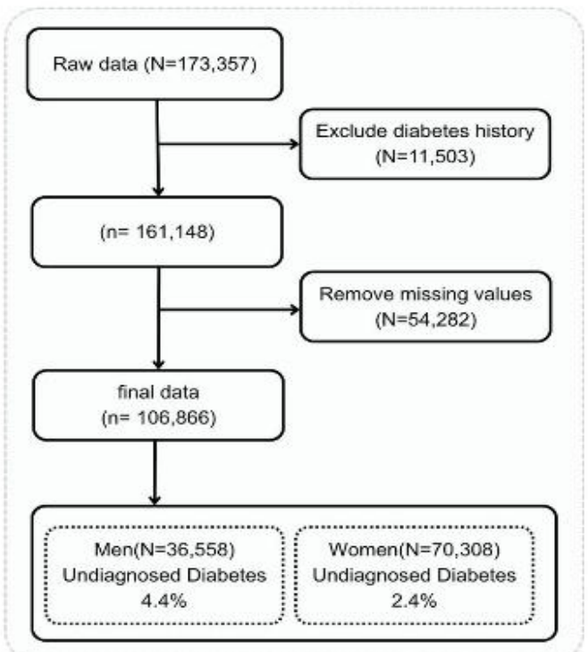
- 당뇨 예측 모델 개발을 위해 통계적 방법과 기계학습 방법을 활용한 다양한 연구들이 수행되어왔으나 기계학습 기반 모델이 항상 높은 성능을 보이는 것은 아니었음.
- 대부분의 연구는 성별을 단순 설명 변수로만 포함하여 성별 간 차이를 충분히 반영하지 못함.
- 남녀 간 당뇨병 유병률 양상과 위험요인, 생활습관의 차이가 뚜렷하게 나타나기 때문에, 성별 특성을 반영한 맞춤형 당뇨 예측 모델의 개발이 필요함.

[목적]

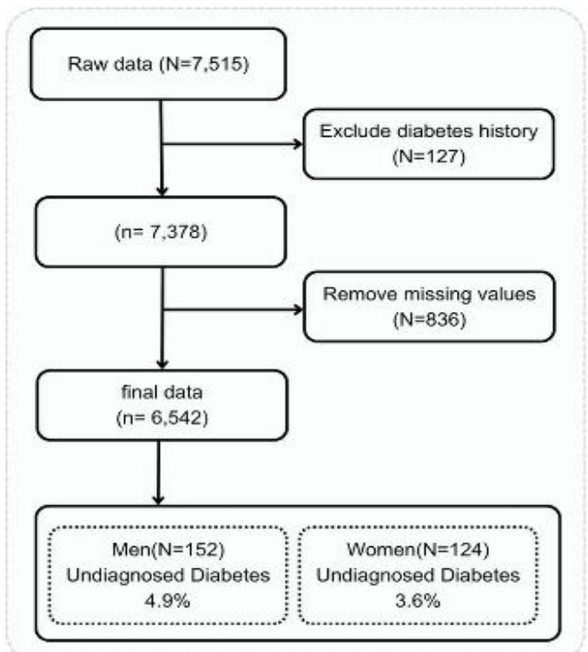
남성과 여성 각각의 특성을 반영하여 예측 성능 및 주요 예측 변수의 차이를 규명하고, 도시 기반 코호트에서 학습 후 지역사회 기반 코호트에서 검증함으로써 한국인 성별 맞춤형 당뇨 예측모델을 개발하고자 함.

[방법]

Flow-chart 1. Health Examinees Study, HEXA



Flow-chart 2. Community-based cohorts



Internal training / validation / test (6 : 2 : 2) → External validation

- 도시 기반 코호트 기반조사 173,357명과 지역사회 기반 코호트 3기 7,515명을 사용함.
- 두 코호트에서 당뇨병 과거력이 있는 대상자를 제외한 후, 공통으로 수집된 17개의 비침습적 변수 중 결측값이 존재하는 대상을 제거함. 최종적으로 도시기반 코호트 106,866명, 지역사회기반 코호트는 6,542명을 사용함.
- 도시 기반 코호트에서 남성은 비당뇨 34,951명(95.6%), 당뇨 1,607명(4.4%)으로 나타났으며, 여성은 비당뇨 68,631명(97.6%), 당뇨 1,677명(2.4%)으로 나타남.
- 지역사회 기반 코호트에서 남성은 비당뇨 2,978명(95.1%), 당뇨 152명(4.9%)으로 나타났으며, 여성은 비당뇨 3,287명(96.4%), 당뇨 124명(3.6%)으로 나타남.
- 도시기반 코호트에서 모델 학습과 검증, 내부 테스트 과정을 모두 마친 후 지역사회 기반 코호트 데이터셋을 이용해 외부 검증을 수행하여 모델의 실제 일반화 성능과 신뢰도를 최종적으로 평가함.
- 통계 유의성 검토한결과 남녀 하나씩 제외됨. (남성: 결혼상태, 여성: 흡연여부)
- 총 16개의 변수 중 stepwise를 활용해 변수를 선택한 결과, 남성 집단에서 9개, 여성 집단에서 7개의 변수가 선택됨.
- Feature Importance(FI)는 각 변수의 예측 기여도를 평가하고 Recursive Feature Elimination(RFE)는 중요도가 낮은 변수를 반복적으로 제거하여 최적의 변수 조합을 찾는 방식으로, 머신러닝에서는 FI와 RFE 두 가지 변수 선택 방법을 적용함.
- RFE은 반복적인 변수 제거 과정을 거쳐 15개와 10개 변수 조합을 비교한 결과, 예측 성능이 우수한 10개 변수를 선택함.
- 머신러닝 기반 모델의 FI는 stepwise를 통해 남성 집단에서 10개, 여성집단에서 9개의 변수가 선택됨.

[결과]

Table 1. Variable Selection by Gender and Method

Men	통계	머신러닝					Women	통계	머신러닝				
Select	Stepwise	FI	FI	RFE	RFE	FI	Select	Stepwise	FI	RFE	RFE	RFE	FI
Rank	Logistic Regression	Random Forest	Logistic Regression	AdaBoost	XGBoost	LightGBM	Rank	Logistic Regression	Random Forest	Logistic Regression	AdaBoost	XGBoost	LightGBM
1	복부비만율	체지방율	복부비만율	복부비만율	복부비만율	복부비만율	1	복부비만율	복부비만율	체지방율	복부비만율	복부비만율	복부비만율
2	고혈압과거력	체질량지수	체지방율	나이	체지방율	체지방율	2	체지방율	체지방량	복부비만율	체지방율	체지방율	엉덩이둘레
3	흡연유무	맥박수	엉덩이둘레	체지방율	당뇨병가족력	나이	3	당뇨병가족력	체질량지수	엉덩이둘레	체질량지수	수축기혈압	체질량지수
4	당뇨병가족력	이완기혈압	수축기혈압	체질량지수	체질량지수	수축기혈압	4	체칠량지수	근육량	나이	엉덩이둘레	당뇨병가족력	나이
5	주당운동시간	복부비만율	맥박수	이완기혈압	엉덩이둘레	이완기혈압	5	맥박수	나이	수축기혈압	근육량	나이	수축기혈압
6	체지방율	근육량	주당운동시간	수축기혈압	나이	맥박수	6	수축기혈압	수축기혈압	수축기혈압	나이	체질량지수	맥박수
7	나이	나이	당뇨병가족력	맥박수	결혼유무	주당운동시간	7	나이	엉덩이둘레	체질량지수	수축기혈압	음주유무	고혈압과거력
8	수축기혈압	엉덩이둘레	나이	흡연유무	맥박수	흡연유무	8		이완기혈압	고혈압과거력	맥박수	소득수준	흡연유무
9	맥박수	수축기혈압	체질량지수	당뇨병가족력	소득수준	고혈압과거력	9		맥박수	맥박수	고혈압과거력	교육수준	당뇨병가족력
10		당뇨병가족력	고혈압과거력	고혈압과거력	주당운동시간	당뇨병가족력	10			당뇨병가족력	당뇨병가족력	결혼유무	

Table 2. Male - Results of Variable Selection Methods and Performance Metrics by Model

	FI vs RFE	Recall	Accuracy	F1-Score
Logistic Regression(Stats)	-	0.70	0.60	0.14
Random Forest	FI	0.63	0.57	0.12
Logistic Regression(ML)	FI	0.73	0.61	0.15
XGBoost	RFE	0.67	0.54	0.12
LightGBM	FI	0.61	0.61	0.13
AdaBoost	RFE	0.58	0.58	0.12

Table 3. Female - Results of Variable Selection Methods and Performance Metrics by Model

	FI vs RFE	Recall	Accuracy	F1-Score
Logistic Regression(Stats)	-	0.84	0.58	0.13
Random Forest	FI	0.88	0.61	0.14
Logistic Regression(ML)	FI	0.84	0.67	0.15
XGBoost	RFE	0.64	0.56	0.10
LightGBM	FI	0.64	0.62	0.11
AdaBoost	RFE	0.64	0.67	0.12

Figure 1. Male - AUC curves

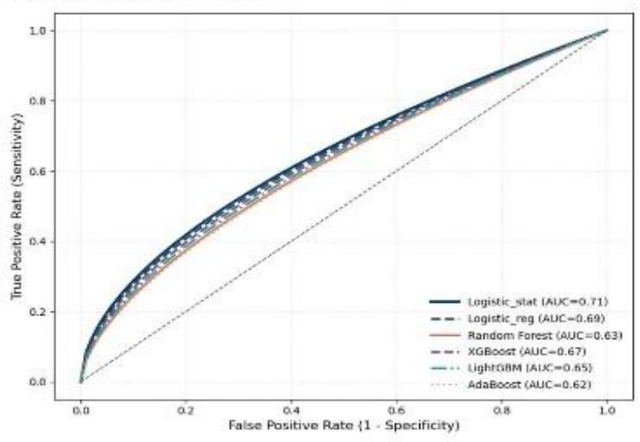


Figure 3 Male - SHAP Summary Plot for Logistic Regression

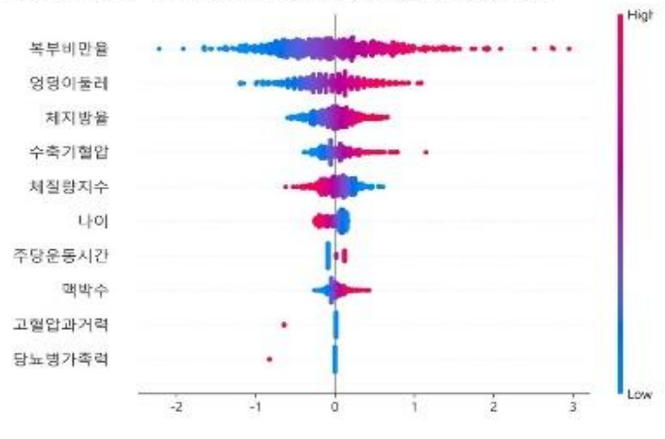


Figure 2. Female - AUC curves

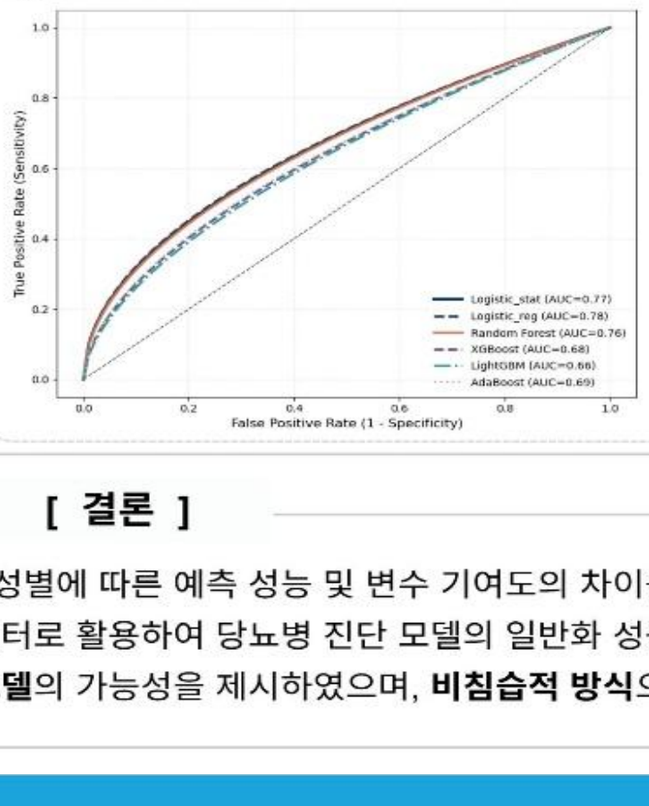
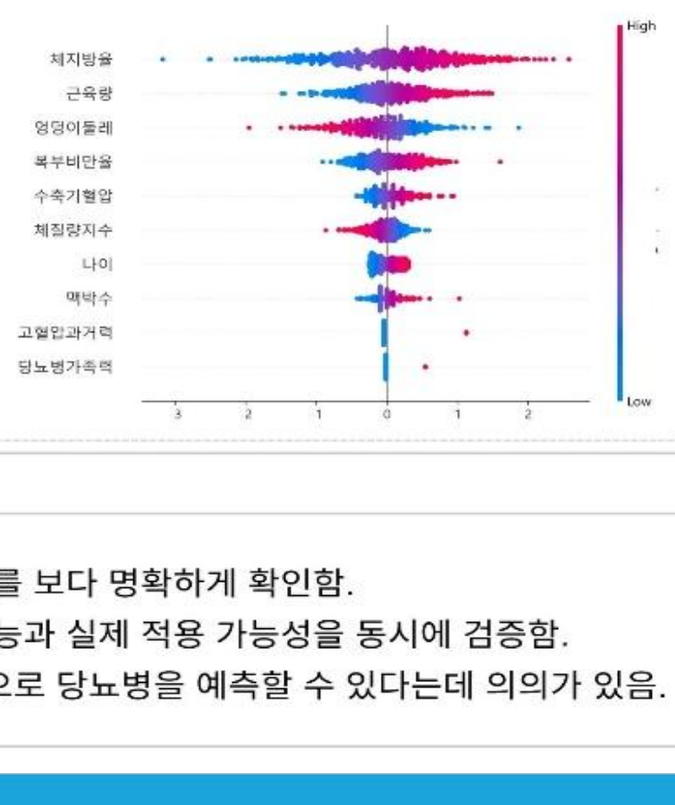


Figure 4. Female - SHAP Summary Plot for Logistic Regression



[결론]

- 기존 연구들과 달리 성별을 분리하여 각각 독립적으로 분석함으로써 성별에 따른 예측 성능 및 변수 기여도의 차이를 보다 명확하게 확인함.
- 도시기반 코호트에서 학습 후, 지역사회기반 코호트를 외부 검증 데이터로 활용하여 당뇨병 진단 모델의 일반화 성능과 실제 적용 가능성을 동시에 검증함.
- 기계학습과 통계적 방법의 성능을 비교하여 **성별 맞춤형 당뇨 예측 모델**의 가능성을 제시하였으며, **비침습적 방식**으로 당뇨병을 예측할 수 있다는데 의의가 있음.