머신러닝 기반 당뇨병 예측 모델

Kaggle Playground Series S5E12 (Diabetes Prediction Challenge)

- 출처 Kaggle S5E12, ML분석.pdf, 통계분석.htm
- 데이터 합성 당뇨 예측 데이터 (Train/Test Dataset)
- 조원 김새한, 송미영, 이세미, 조가영
- 작성일 2026. 1. 5

CONTENTS

section1 Research Background & Topic Selection

연구 배경 및 주제 선정

* 연구배경
* 분석 목적 및 목표

01

연구 배경

당뇨병과관련된 된기초 대표적 기존연구

당뇨병은 단순한 혈당 상승이 아닌
복합적 상호작용의 산물

"제2형 당뇨병은 유전적 요인, 생활습관, 사회·환경적 요인, 신체·대사 상태가 복합적으로 작용하는 다인성 질환으로 알려져 있다" (Park, 2011)

"다수의 역학 연구에서 연령, 비만(BMI), 혈압, 지질 지표 등은 일관된 위험 요인으로 보고된다다" (Umesh Kumar Sharma, 2024)

"생활습관(신체활동, 식습관) 및 가족력·기저질환은 당뇨 발생 위험을 조절하는 주요 요인으로 제시된다다"(Arya P, 2023)

" 생활습관, 유전, 환경 등 다양한 영역의 변수를 조합해 당뇨병을 예측하는 모델 생성 "

분석 목적 및 목표

당뇨병 예측모델 데이터 선정의 목적과 최종 목표

분석 목적

"당뇨병은 만성질환이기에 '치료'보다 '관리와 예방'이 중요합니다."

당뇨병은 초기 자각 증상이 뚜렷하지 않아 골든타임을 놓치기 쉽습니다.
잠재적 고위험군을 사전에 식별하여, 의료 자원의 효율적 배분을 돕고
조기 개입을 유도하는 예방 의학적 가치를 실현하고자 합니다.

최종 목표

본 분석은 당뇨병에 유의미한 영향을 미치는 핵심 인자를 도출하고,
복잡한 현실 환경에서도 실제 환자를 놓치지 않는 정교한 스크리닝 모델을
구축하여 사용자에게 데이터 기반의 객관적인 예방 가이드를 제시하는 것이
최종 목표입니다.

section2 Data and Feature Design

데이터 및 변수 설계

* 변수 분류
* EDA(탐색적 데이터 분석)
* 전처리 전략

02

변수분류 및 요인 분석

독립변수를 성격에 따라 5개 범주로 분류

요인군	개수	주요 변수
배경변수	3	age, gender, ethnicity
사회경제/환경	4	education_level, income_level, employment_status, screen_time_hours_per_day
생활습관	5	smoking_status, alcohol_consumption_per_week, physical_activity, diet_score, sleep_hours_per_day
유전/병력	3	family_history_diabetes, hypertension_history, cardiovascular_history
신체/대사 지표	9	bmi, waist_to_hip_ratio, systolic_bp, diastolic_bp, heart_rate, cholesterol_total, triglycerides, hdl_cholesterol, ldl_cholesterol
제외/타겟 전용	6	glucose_fasting, glucose_postprandial, insulin_level, hba1c, diabetes_risk_score, diabetes_stage
타겟 변수	1	diagnosed_diabetes

출처 : Diabetes Health Indicators Dataset – A Comprehensive Dataset of 100,000 Patient Records for Diabetes Risk Analysis (Source: Kaggle)

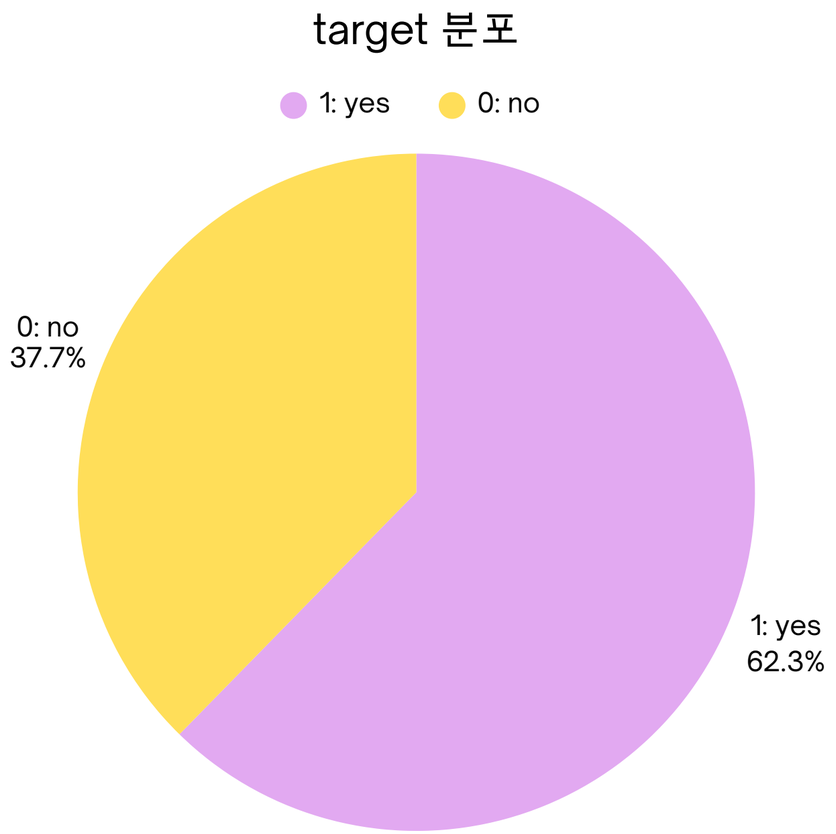
EDA (탐색적 데이터 분석)

전체데이터의 개략적 분포와 target데이터의 분포

전체 데이터 / target

Shape
train : (700000, 26) test : (300000, 25)
Data type
int64 : 13 / float64 : 5 object : 6

target
diagnosed_diabetes (0: NO, 1: Yes)

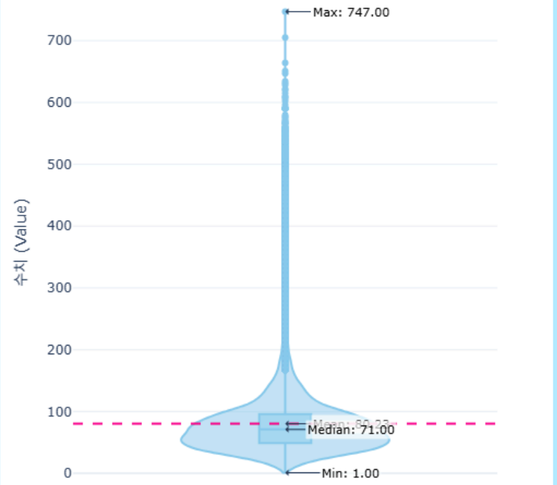


출처 : Diabetes Health Indicators Dataset – A Comprehensive Dataset of 100,000 Patient Records for Diabetes Risk Analysis (Source: Kaggle)

EDA (탐색적 데이터 분석)

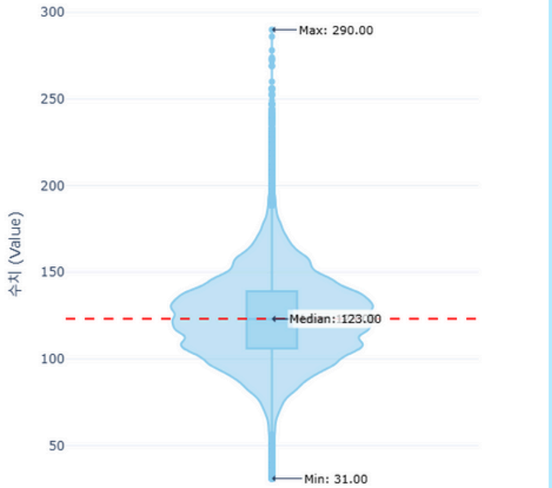
수치형 데이터 중 왜도, 이상치가 많은 변수의 시각화와 전처리 방법

수치형 데이터



physical_activity(운동량)

대부분의 데이터가 왼쪽에 쏠려 있고 일부가 아주 큰 값을 가진 'Long-tail' 분포



triglycerides(중성지방)

극단적인 이상치
Right-skewed 형태
표준편차가 24.7로 상당히 큼

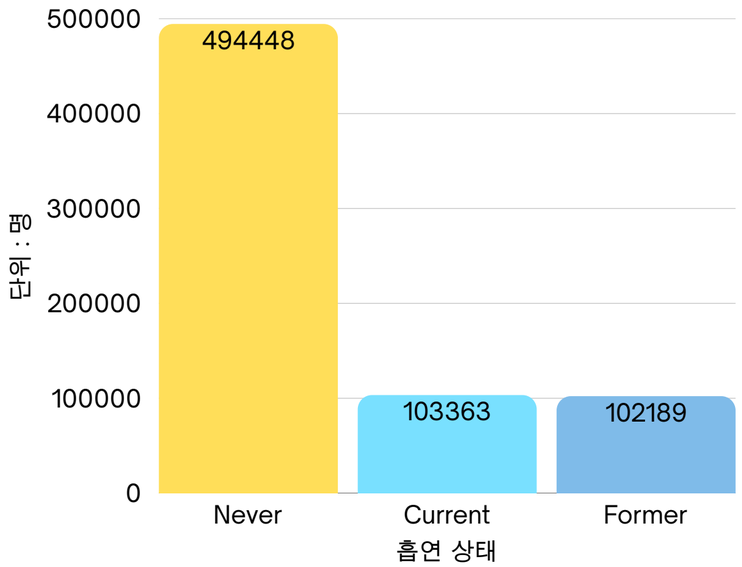
💡분석 point

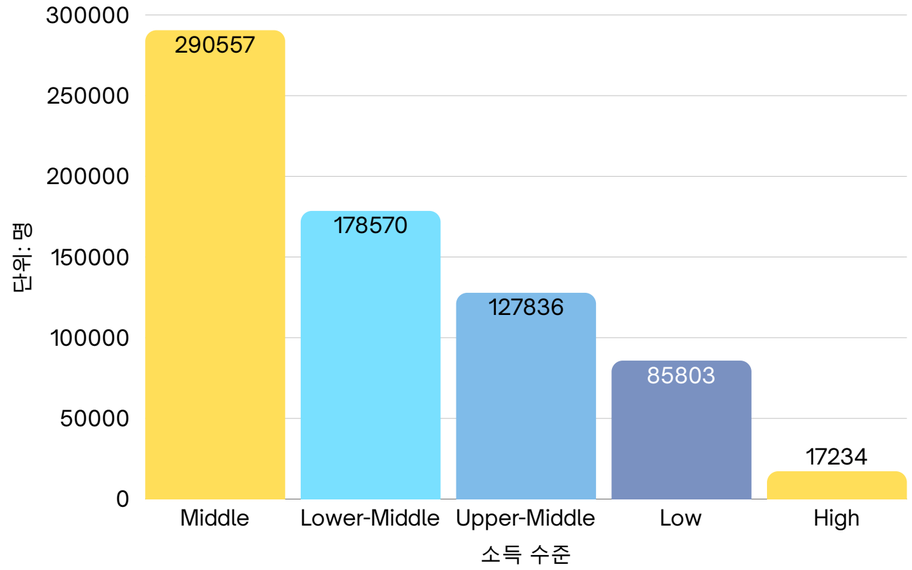
- 각 요소별 큰 범위 차이 : Scaling
- physical_activity, triglycerides 등 치우친 분포 : Log 변환

EDA (탐색적 데이터 분석)

범주형 데이터 중 클래스 불균형이 있는 데이터 시각화 및 전처리 방법

범주형 데이터





💡분석 point

- 클래스 불균형 : 클래스 병합(흡연/비흡연), 샘플링 등 클래스 불균형 완화 필요

EDA (탐색적 데이터 분석)

독립변수 사이의 상관관계로 모델에 부정적 영향을 주는 다중공선성 확인

다중공선성 확인

변수명 (Feature)	VIF 지수	비고
waist_to_hip_ratio	**830.99**	BMI와 높은 상관관계
cholesterol_total	**620.44**	LDL/HDL 수치와 중복
log_triglycerides	**517.57**	대사 지표 간 중복
bmi	**186.46**	신체지표 핵심 변수
ldl_cholesterol	**173.86**	총 콜레스테롤과 중복
systolic_bp	**154.52**	이완기 혈압과 중복
diastolic_bp	**123.91**	수축기 혈압과 중복

💡분석 point

- **높은 VIF**

특정 지표의 영향력이
과대, 과소 평가되는 왜곡 발생

→ 중복되는 지표를 하나로 통합
→ 대표 변수 하나만 선택

EDA 결과 분석

EDA결과로 도출한 전처리 전략 및 고려사항

전처리 전략

01

독립 변수

StandardScaler
log변환

OneHotEncoder
OrdinalEncoder
class 통합

02

target

모델 생성 시
클래스 불균형 옵션 적용
(class_weight="balanced"
scale_pos_weight=ratio)

03

다중공선성

카테고리 대표변수 선정
파생변수 생성

04

파생변수

TG/HDL Ratio
두 변수의 비율이
인슐린 저항성을 잘 설명

Age x Family
연령 증가 시 유전요인 발현
가능성이 높아지는
비선형적 특성을 반영

section3 Statistic Analysis

통계분석

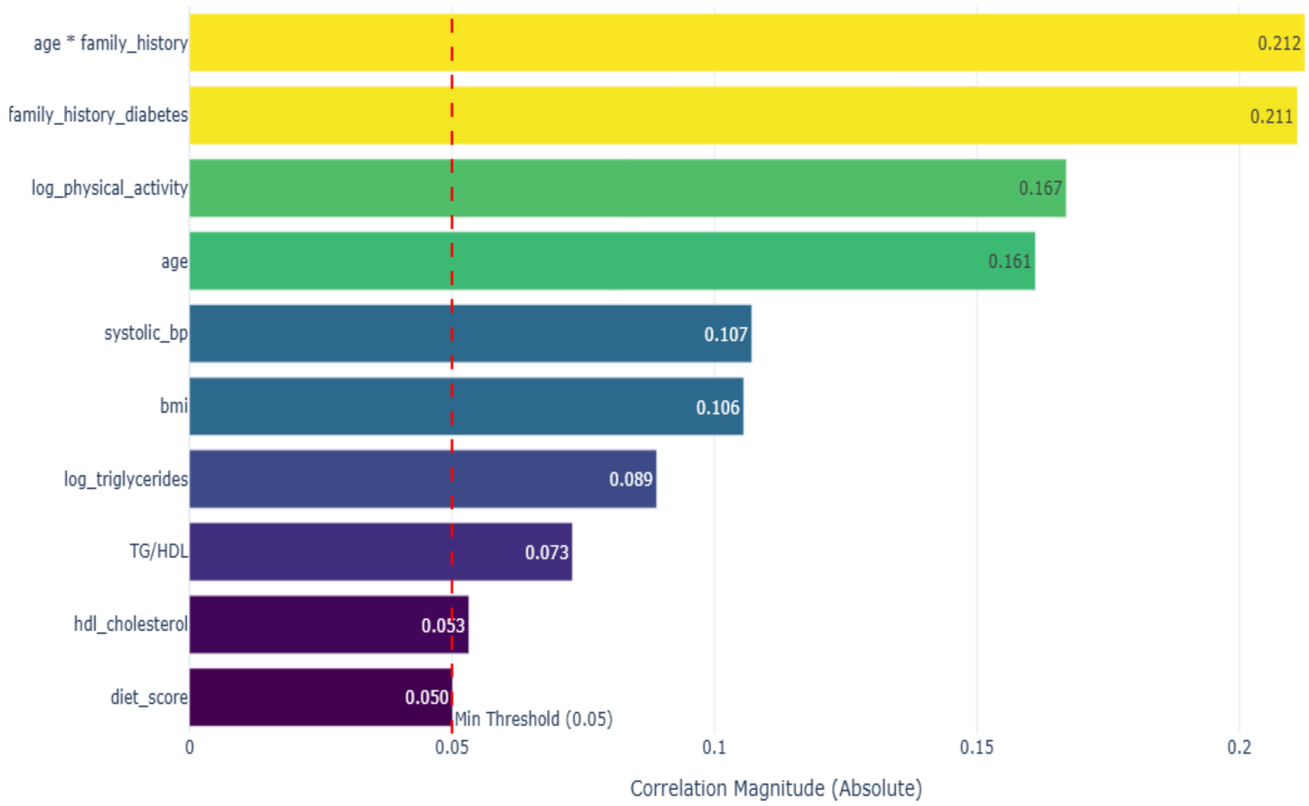
* 상관분석
* 단변량분석
* 다변량분석
* 최종 변수 선택

03

상관계수 분석

target과의 correlation이 0.05이상인 변수

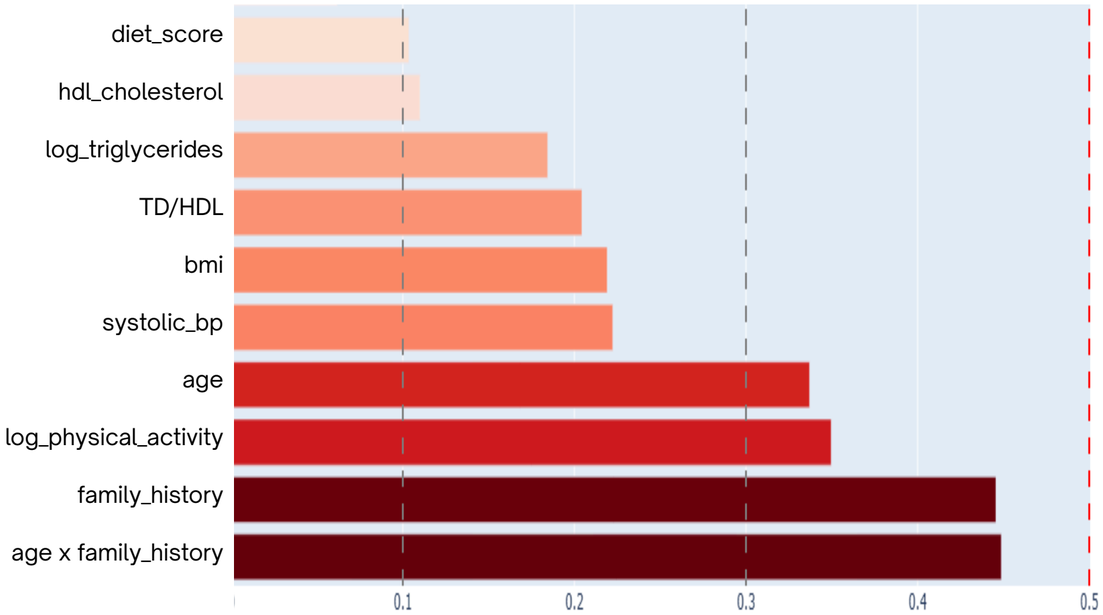
요인군	cohen's D > 0.3	cohen's D > 0.1
유전/병력	family_history	
파생변수	age*family_history	TG/HDL
배경변수	age	
생활습관	physical_activity	diet_score
신체/대사	systolic_bp, bmi,	triglycerides, hdl_cholesterol



효과크기 분석 – 수치형 변수

효과크기(cohen's D) >0.1 수치형 변수

요인군	cohen's D > 0.3	cohen's D > 0.1
배경변수	age	
생활습관	physical_activity	diet_score
유전/병력	family_history_diabetes	
파생변수	age*family_history	TG/HDL
신체/대사		systolic_bp, bmi, triglycerides, hdl_cholesterol

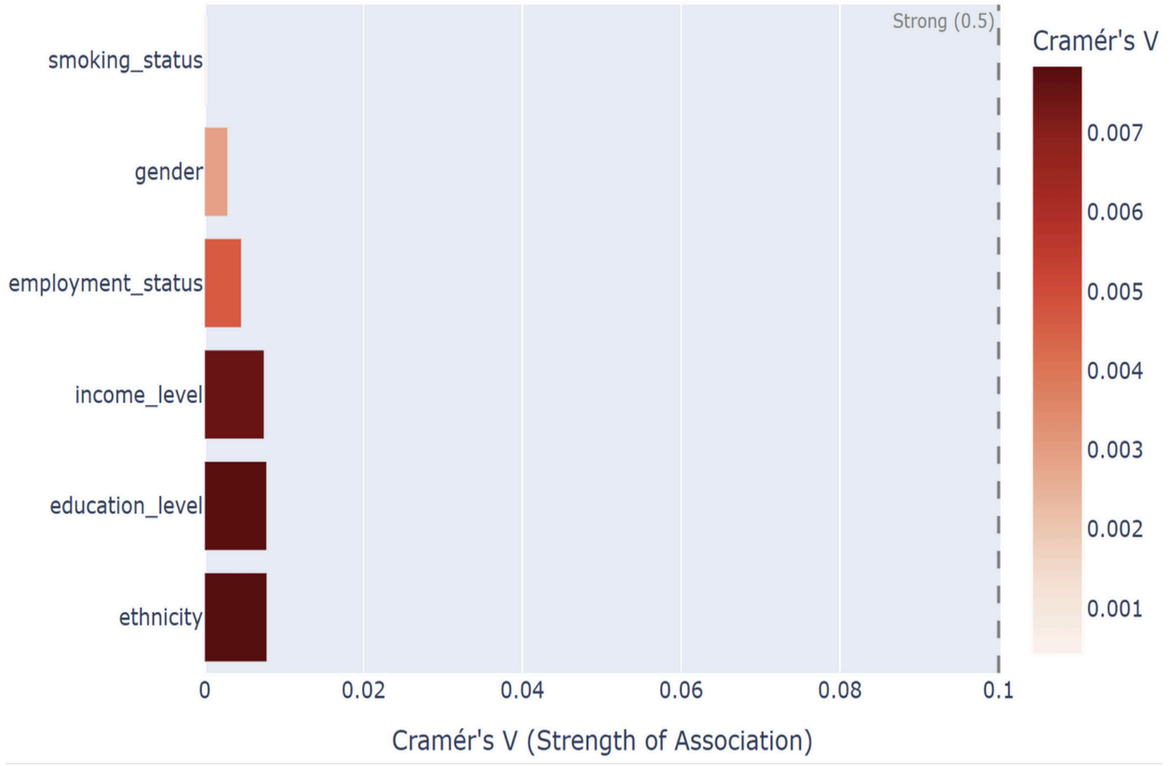


0.5 큰효과	0.3 중간효과	0.1 작은효과

Cohen, J. (1988). Statistical Power Analysis for the Behavioral Sciences (2nd ed.). Hillsdale, NJ: Lawrence Erlbaum Associates.

효과크기 분석 – 범주형 변수

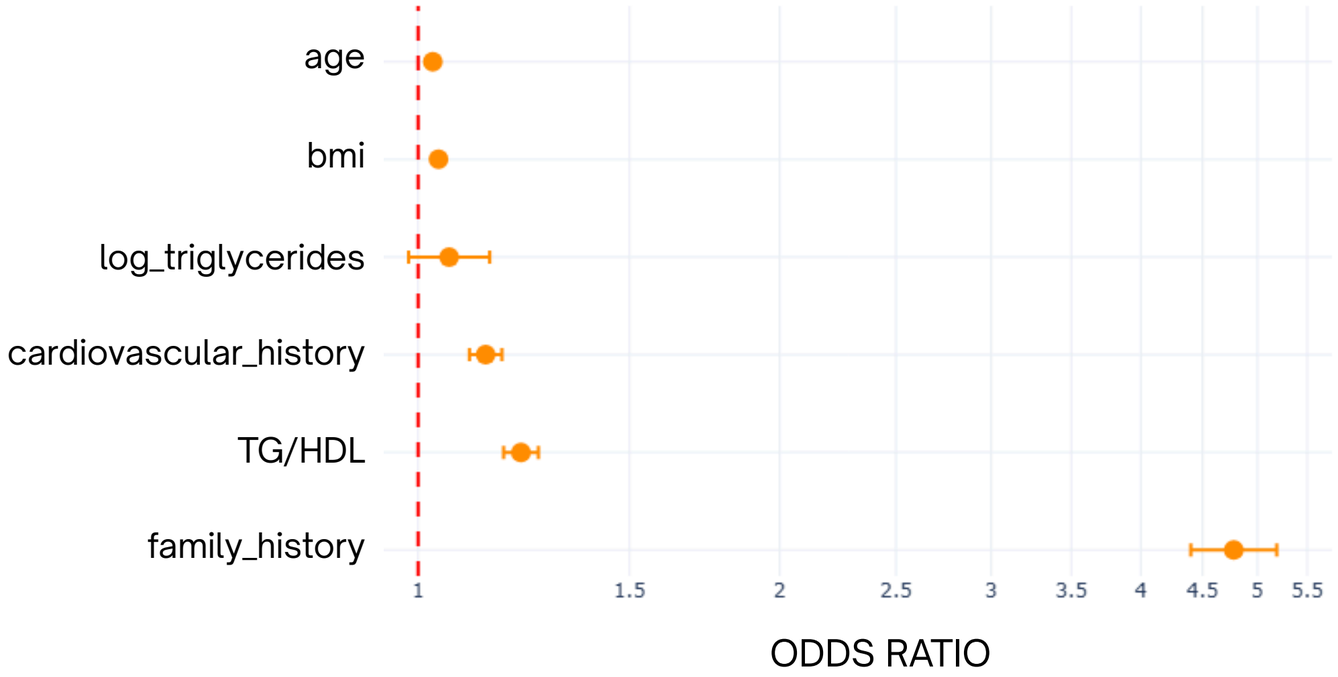
효과크기(cramer's V) 상 유의한 변수 미확인



Cohen, J. (1988). Statistical Power Analysis for the Behavioral Sciences (2nd ed.). Hillsdale, NJ: Lawrence Erlbaum Associates.

다변량 분석 – 로지스틱 회귀

odds ratio가 유의한(1과 가장 먼)변수 분석



분석 point

- **cardiovascular_history**

  단변량분석과 다르게 유의한 결과
  심혈관 질환자의 비율이 적기 때문

  ➡️ 고위험군 식별에 유의

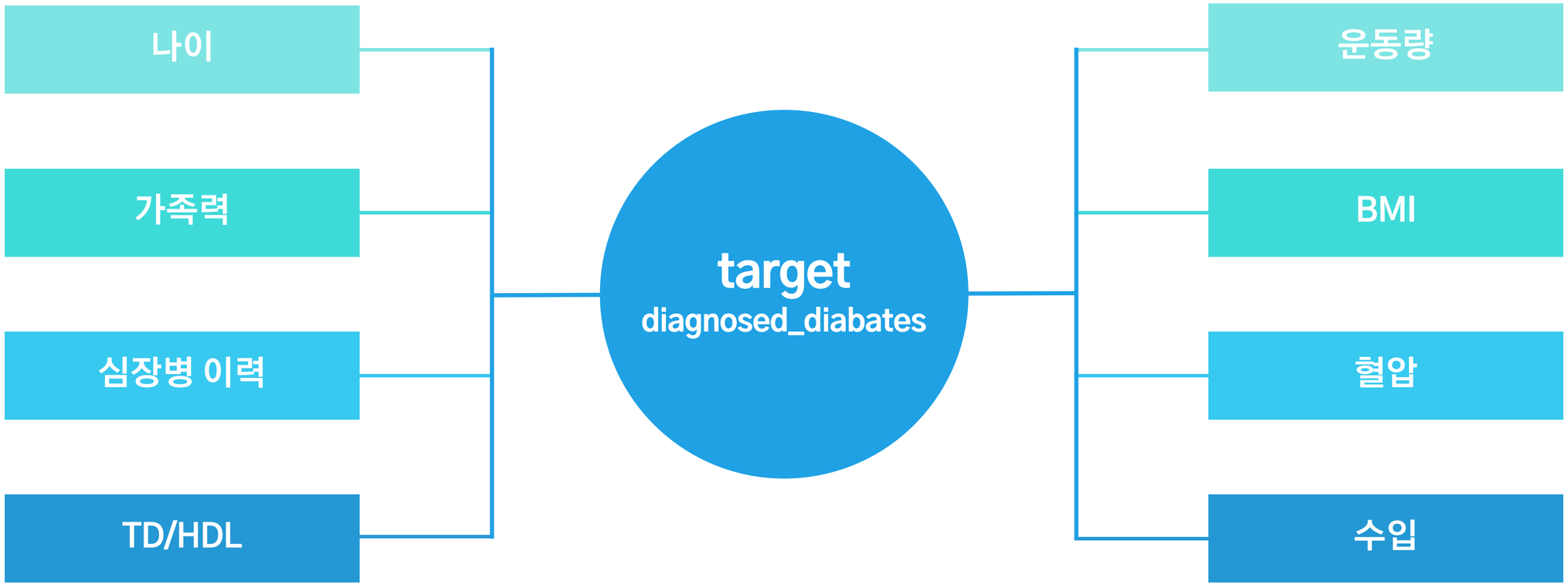
- **age x family_history**

  개별 변수가 효과를 다 흡수해 유의X

  ➡️ 변수 제외 고려

최종 변수 선정

통계분석을 통해 도출해낸 최종 당뇨병 모델의 독립변수 8개



최종 변수 선정

다변량, 단변량, 상관계수 분석, 배경지식을 종합한 변수선정 근거

요인군	변수명	상관분석 (correlation)	단변량분석 (효과크기)	다변량분석 (odds ratio)	선정근거
배경변수	age	🟢	🟢	🟢	연령에 따른 생물학적 위험요인 통제
생활습관	physical_activity	🟢	🟢	🟢	발병 위험을 낮추는 핵심 방어 기제
유전/병력	family_history_diabetes	🟢	🟢	🟢	가장 강력한 예측 지표 / 유전요인 대표
	cardiovascular_history	🟠	🟡	🟡	심혈관 질환 고위험군(Red Flag) 식별
신체/대사 지표	bmi	🟢	🟢	🟢	비만도 및 신체 대사 상태 대표
	systolic_bp	🟡	🟢	🟢	혈관 건강 및 합병증 위험도 반영
파생변수	TG/HDL	🟡	🟢	🟢	인슐린 저항성을 나타내는 핵심 대사 지표
사회경제/환경	income_level	🟠	🟡	🟡	사회경제적 배경 / 의료 접근성 일반화

section4 Machine Learnig

머신러닝

* 모델링 전략 및 실험 설계
* 모델 성능 비교
* 최종 모델 선택

04

모델링 전략 및 실험 설계

데이터 파이프라인 최적화 및 계층적 검증 기반의 다각적 모델 벤치마킹

예측 모델 및 전처리 (Models & Preprocessing)

Category	Details	
Baseline Models (기본성능 확인)	📈Linear	Logistic Regression
	🍏Tree	Decision Tree Classifier
Ensemble Models (고성능 부스팅 알고리즘)	🚀Boost	XGBoost
	🚀Boost	LightGBM
Preprocessing (Column Transforme)	📊수치형	StandardScaler (표준화)
	📋범주형	OneHotEncoder / OrdinalEncoder

검증전략

Stratified K-Fold CV

타겟 클래스 비율을 유지하며 5-Fold 교차 검증 수행
데이터 불균형을 고려한 안정적인 평가

PRIMARY METRIC

ROC - AUC score

보조지표: Accuracy, Precision, Recall, F1

최적화(optimization)

RandomizedSearchCV

광범위한 파라미터 공간을 효율적으로 탐색.
Iter: 10회 무작위 샘플링
Scoring: 'roc_auc' 기준 최적화
Params: 학습률, 트리 깊이, 정규화 계수 등

모델 성능 비교 평가 (Validation)

4개의 모델을 비교해 최적의 모델 찾기

model	ROC_AUC⭐	Accuracy	PRECISION	Recall	F1 score
Logistic Regression	0.691523	0.621686	0.75648	0.579629	0.65635
Decision Tree Classifier	0.697476	0.632057	0.757621	0.6024	0.671152
XGBoost	0.717675	0.647843	0.770098	0.62014	0.687031
LightGBM⭐	**0.71997**	**0.649736**	**0.770609**	**0.623704**	**0.689417**

최종 선정 모델: LIGHTGBM

모델 중 모든 지표에서 우수한 점수를 보였음.
복잡한 데이터 분포를 효율적으로 학습할 수 있고,
교차검증 성능과 Validation 성능 간 차이가 크지 않아
비교적 안정적인 성능을 보임.
또한 이상치와 치우친 데이터에 강한 장점이 있음

성능 분석 인사이트

선형 모델과 tree는 전반적으로 낮은 성능을 보였습니다.
트리 기반 앙상블 모델인 XGBoost와 LightGBM이
비선형 패턴을 더 잘 학습해 Accuracy와 F1 Score가
크게 향상되었습니다.

section5 Results Analysis & Interpretation

분석 결과 및 해석

* 모델링 전략 및 실험 설계
* 모델 성능 비교
* 최종 모델 선택

05

캐글 제출 및 최종 점수 (Submission & Final Score)

제출 프로세스 (Submission Pipeline)

LightGBM Model
최적화된 모델 로드

Predict Probability
예측 확률 산출 (Not Label)

Submission.csv
ID, diagnosed_diabetes

평가 지표 (Evaluation Metric)

ROC-AUC (Area Under the Curve)

이 대회는 이진 분류(Binary Classification) 문제로, 모델이 양성 클래스(당뇨병 발병)를 얼마나 잘 구분하는지를 평가합니다. 단순 정확도(Accuracy)보다 불균형 데이터 및 확률 예측 성능 평가에 적합합니다.

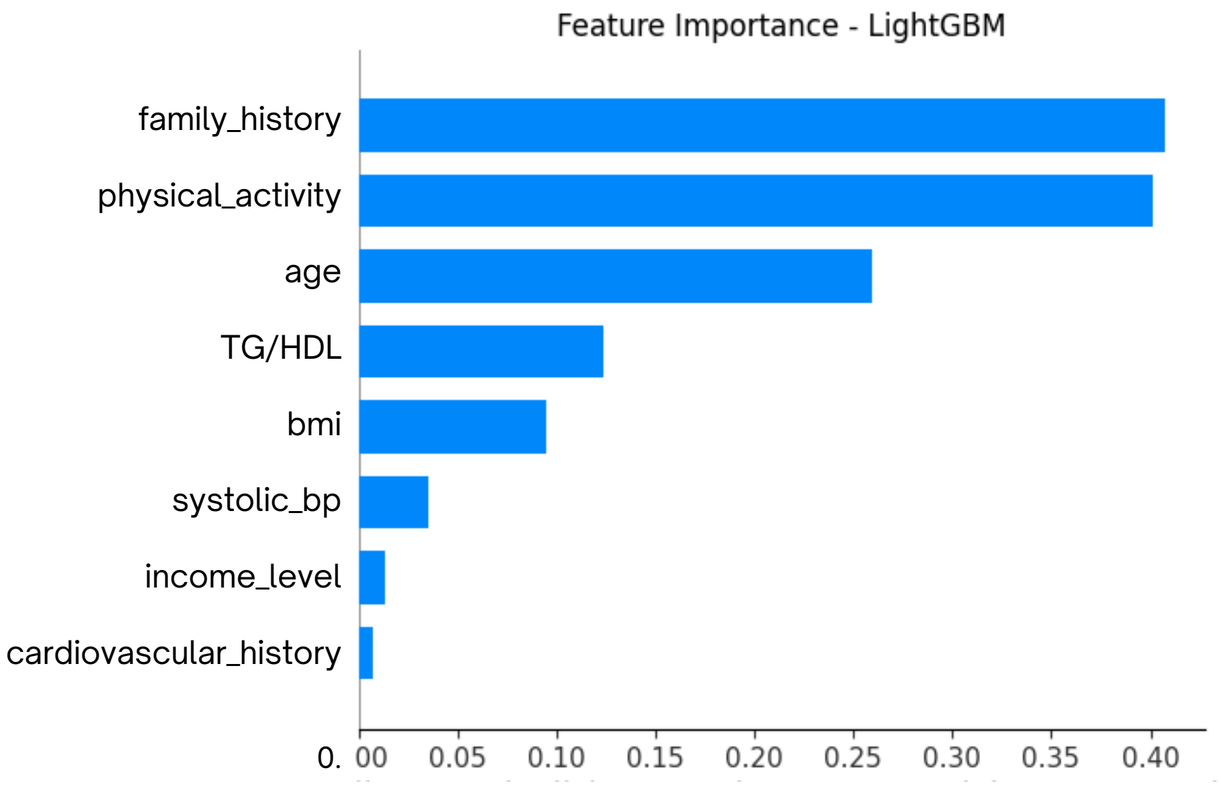
리더보드 점수 (Scores)

Final score
0.69116
5-Fold Stratified CV Average

Public LB	0.69116	Pending
Private LB	0.68834	Hidden

Feature Importance (SHAP Analysis)

SHAP 분석을 통해 예측 모델이판단 시 중요하게 고려한 지표의 순위를 도출



분석 Point

1. **가족력 (family_history) - 0.407**
   당뇨 발생을 예측하는 가장 압도적인 지표

2. **신체 활동량 (physical_activity) - 0.401**
   활동량이 적을수록 발병 위험이 급격히 상승함.
   활동량 부족은 인슐린 저항성과 직접적인 인과 경로를 형성

3. **연령 (age) - 0.259**
   생물학적 노화에 따른 기본적인 발병 위험 기여도

4. **인슐린 저항성 지표 (TG/HDL) - 0.123**
   중성지방과 HDL의 비율이 단순 신체 지표보다 당뇨 예측에 더 민감하게 작용

5. **체질량 지수 (bmi) - 0.095**
   비만도가 물리적 위험 인자로서 유의미한 설명력을 가짐

결론 및 기대효과

🎯 프로젝트 성과 및 결론

1. **예방 의학적 가치 실현**
   - 방대한 데이터를 바탕으로 고위험군을 사전에 선별할 수 있는 지표를 구축하여 조기 관리를 돕는 예방 의학적 모델을 구현했습니다.

2. **데이터 기반의 정교한 변수 선별**
   - VIF 분석을 통해 다중공선성을 유발하는 중복 변수를 정제하고, 통계적으로 견고한 8개의 핵심 변수 모델을 완성했습니다.

3. 분석의 투명성 및 맞춤형 가이드라인 제시
   - SHAP 분석을 도입하여 각 지표가 당뇨 발병에 미치는 기여도를 정량화함으로써, 단순 예측을 넘어 **'설명 가능한 의료 AI'** 모델을 구축하고 개인별 맞춤형 관리 가이드를 제공합니다.

💡 기대 효과 및 활용 방안

1. **개인 맞춤형 위험 알림**
   - SHAP 분석 결과 기여도가 높았던 **신체 활동**과 **지질 수치(TG/HDL)** 등을 실시간 모니터링하여, 개인화된 건강 관리 가이드를 제공할 수 있습니다.

2. 디**지털 스크리닝 시스템 (Digital Screening)**
   - 병원 방문 전 단계에서 간단한 신체 지표 입력만으로 당뇨 위험도를 예측합니다. 이는 증상이 나타나기 전 **조기 발견 및 치료**를 유도하는 1차 스크리닝 도구로 활용 가능합니다.

3. 의**료 자원 효율화**
   - 예측 모델을 통해 선별된 고위험군에게 의료 서비스를 집중함으로써, 사회적 비용을 절감하고 공공보건의 효율성을 높일 수 있습니다.

참고 자료

References

학술적 근거 및 문헌 검토

- Albers, J. D., et al. (2025). Socioeconomic position and type 2 diabetes: Examining the mediating role of social cohesion. Social Science & Medicine, 376, 118046.
- Duan, M.-J. F., et al. (2022). Effects of Education and Income on Incident Type 2 Diabetes and Cardiovascular Diseases. Journal of General Internal Medicine, 37(15), 3907–3916.
- Hill-Briggs, F., et al. (2021). Social Determinants of Health and Diabetes: A Scientific Review. Diabetes Care, 44(1), 258–279.
- Ismail, L., et al. (2021). Association of risk factors with type 2 diabetes: A systematic review. Computational and Structural Biotechnology Journal, 19, 1759–1785.
- Kim, H.-J., et al. (2023). Lifestyle Factors Affecting Blood Sugar Control by Workers with Type 2 Diabetes. Journal of the Korea Academia-Industrial cooperation Society, 24(6), 105–115.
- Kim, S.-R., et al. (2015). Age- and Sex-Specific Relationships between Household Income, Education, and Diabetes Mellitus. PLOS ONE, 10(1), e0117034.
- Sharma, U. K., et al. (2024). Type-II-Diabetes Mellitus- Etiology, Epidemiology, Risk Factors and Diagnosis. Int. Journal of Health Sciences and Research, 14(1), 283–290.
- Song, Z., et al. (2021). Healthy lifestyle including a healthy sleep pattern and incident type 2 diabetes. Cardiovascular Diabetology, 20(1), 239.
- Wang, P., et al. (2024). Socioeconomic Status, Diet, and Behavioral Factors and Cardiometabolic Diseases. JAMA Network Open, 7(12), e2451837.
- Zhou, P., et al. (2025). Socioeconomic disparities on the association of physical behavior with incident type 2 diabetes. BMC Public Health, 25(1), 3579.