ANALYSIS REPORT

머신러닝 기반 당뇨병 진단 예측 모델

Kaggle Playground Series S5E12 (Diabetes Prediction Challenge)

- **데이터** 합성 당뇨 예측 데이터 (Train/Test Dataset)
- **목표** 진단 확률 예측 및 위험 요인에 대한 통계적 가설 검정
- **작성일** 2025년 12월 31일
- **출처** Kaggle S5E12, ML분석.pdf, 통계분석.html
- **조원** 김새한, 송미영, 이세미, 조가영

CONTENTS

SECTION 01 Research Background & Topic Selection

연구 배경 및 주제 선정

- 프로젝트 선정 이유
- 분석 목적 및 기존 연구 소개

01

연구 배경 및 주제 선정

분석목적

분석목적

본 분석은 Kaggle의 Diabetes Health Indicators (BRFSS) 데이터셋을 활용하여, 단순한 예측 모델링을 넘어 건강 지표들 간의 상호작용과 당뇨병 발병의 인과 경로를 규명하는 것을 목적으로 합니다.

주제선정 이유

당뇨병은 완치보다 관리와 예방이 중요한 질환입니다.

분석을 통해 고위험군을 선별해낼 수 있다면, 의료자원을 효율적으로 배분하고 환자 스스로 조심하게 만드는 예방 의학적 가치가 크다고 생각하여 당뇨병 진단 데이터를 선택했습니다

연구 배경 및 주제 선정 - 기존연구 소개

당뇨병은 단순한 혈당 상승이 아닌 복합적 상호작용의 산물입니다.

*"제2형 당뇨병은 **유전적 요인, 생활습관, 사회·환경적 요인, 신체·대사 상태**가 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·기저질환**은 당뇨 발생 위험을 조절하는 주요 요인으로 제시된다다"(Arya P, 2023)*

기존 연구의 한계를 극복하는 **복합 요인 상호작용 분석 기반의 당뇨병 예방 전략**을 세우고자 합니다.

SECTION 02 Data and Feature Design

데이터 및 변수 설계

- 데이터 개요
- 변수 분류 및 선정 기준
- 탐색적 데이터 분석(EDA)

02

데이터 개요

Diabetes Dataset

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

데이터 분석 목적

- 본 연구는 임상 데이터셋을 활용하여 당뇨병 진단 여부와 관련된 다양한 위험 요인들을 탐색하고, 각 요인들이 당뇨병 발생과 어떤 통계적 관련성을 가지는지 검증하는 것을 목적으로 한다.

데이터 규모

- 관측치: 100,000명
- 변수: 31개

분석 대상 변수 선정 기준

- 이론적 / 도메인 기반 변수 선정
  - 앞서 제시한 기존 연구의 이론적 분류를 기준으로 분석 대상 변수를 1차적으로 선별
  - 당뇨병의 원인 또는 위험 요인에 해당하는 변수만을 분석 대상으로 포함

- 진단 지표 변수의 제외 (정보 누수 방지)
  - 기존연구(So-RaKim et al., 2014)에서는 진단 기준 또는 질병 상태를 직접 반영하는 임상 지표로 혈당, HbA1c, 인슐린 수치 등을 삼음

    *The participants were classified as having DM if they met one of the following conditions: 1) fasting plasma glucose 126 mg/dL or higher***(glucose_fasting)***, 2) medical diagnosis of DM by a trained medical professional***(insulin_level)***, or 3) treatment with oral hypoglycemic agents or insulin injections. The control status of DM was evaluated by HbA1c levels***(HbA1c)***, with less than 7% being regarded as the optimal level.*

- 요인군별 구조적 분석 설계
  - 각 요인군 내 변수들은 당뇨 진단 여부(diagnosed_diabetes)와의 관련성을 요인군 단위로 개별적으로 검토할 예정
  - 변수 간 성격 차이와 해석 가능성을 고려해 요인군 단위로 분석을 설계
    - 사회경제/생활환경 요인
    - 생활습관 요인
    - 유전/병력 요인
    - 신체/대사 상태 지표

변수 분류 및 요인 분류

개수	주요 변수	변수 유형
3	age, gender, ethnicity	연속형 / 범주형
4	education_level, income_level, employment_status, screen_time_hours_per_day	연속형 / 범주형
5	smoking_status, alcohol_consumption_per_week, physical_activity, diet_score, sleep_hours_per_day	연속형 / 범주형
3	family_history_diabetes, hypertension_history, cardiovascular_history	이진 범주형
9	bmi, waist_to_hip_ratio, systolic_bp, diastolic_bp, heart_rate, cholesterol_total, triglycerides, hdl_cholesterol, ldl_cholesterol	연속형
6	glucose_fasting, glucose_postprandial, insulin_level, hba1c, diabetes_risk_score, diabetes_stage	연속형
1	diagnosed_diabetes	이진 범주형

EDA(탐색적 데이터 분석)

대상: Kaggle - Diabetes Prediction Challenge 의 train, test data set

Shape

train : (700000, 26)
test : (300000, 25)

Target

'diagnosed_diabetes' : (0: NO, 1: Yes)

null / NA / Duplicated

X

age	int64
family_history_diabetes	int64
triglycerides	int64
ldl_cholesterol	int64
hdl_cholesterol	int64
hypertension_history	int64
heart_rate	int64
diastolic_bp	int64
cholesterol_total	int64
physical_activity_minutes_per_week	int64
alcohol_consumption_per_week	int64
systolic_bp	int64
cardiovascular_history	int64
waist_to_hip_ratio	float64
bmi	float64
screen_time_hours_per_day	float64
sleep_hours_per_day	float64
diet_score	float64
gender	object
ethnicity	object
education_level	object
income_level	object
smoking_status	object
employment_status	object

Data types

int64 : 13
float64 : 5
object : 6
(encoding 필요)

EDA(탐색적 데이터 분석) - 기초통계량 (1)

대상: Kaggle - Diabetes Prediction Challenge 의 train, test data set

	age	alcohol_consumption_per_week	physical_activity_minutes_per_week	diet_score	sleep_hours_per_day	screen_time_hours_per_day	bmi	waist_to_hip_ratio	systolic_bp
mean	50.36	2.07	80.23	5.96	7.0	6.01	25.87	0.86	116.29
std	11.66	1.05	51.20	1.46	0.9	2.02	2.86	0.04	11.01
min	19.00	1.00	1.00	0.10	3.1	0.60	15.10	0.68	91.00
25%	42.00	1.00	49.00	5.00	6.4	4.60	23.90	0.83	108.00
50%	50.00	2.00	71.00	6.00	7.0	6.00	25.90	0.86	116.00
75%	58.00	3.00	96.00	7.00	7.6	7.40	27.80	0.88	124.00
max	89.00	9.00	747.00	9.90	9.9	16.50	38.40	1.05	163.00

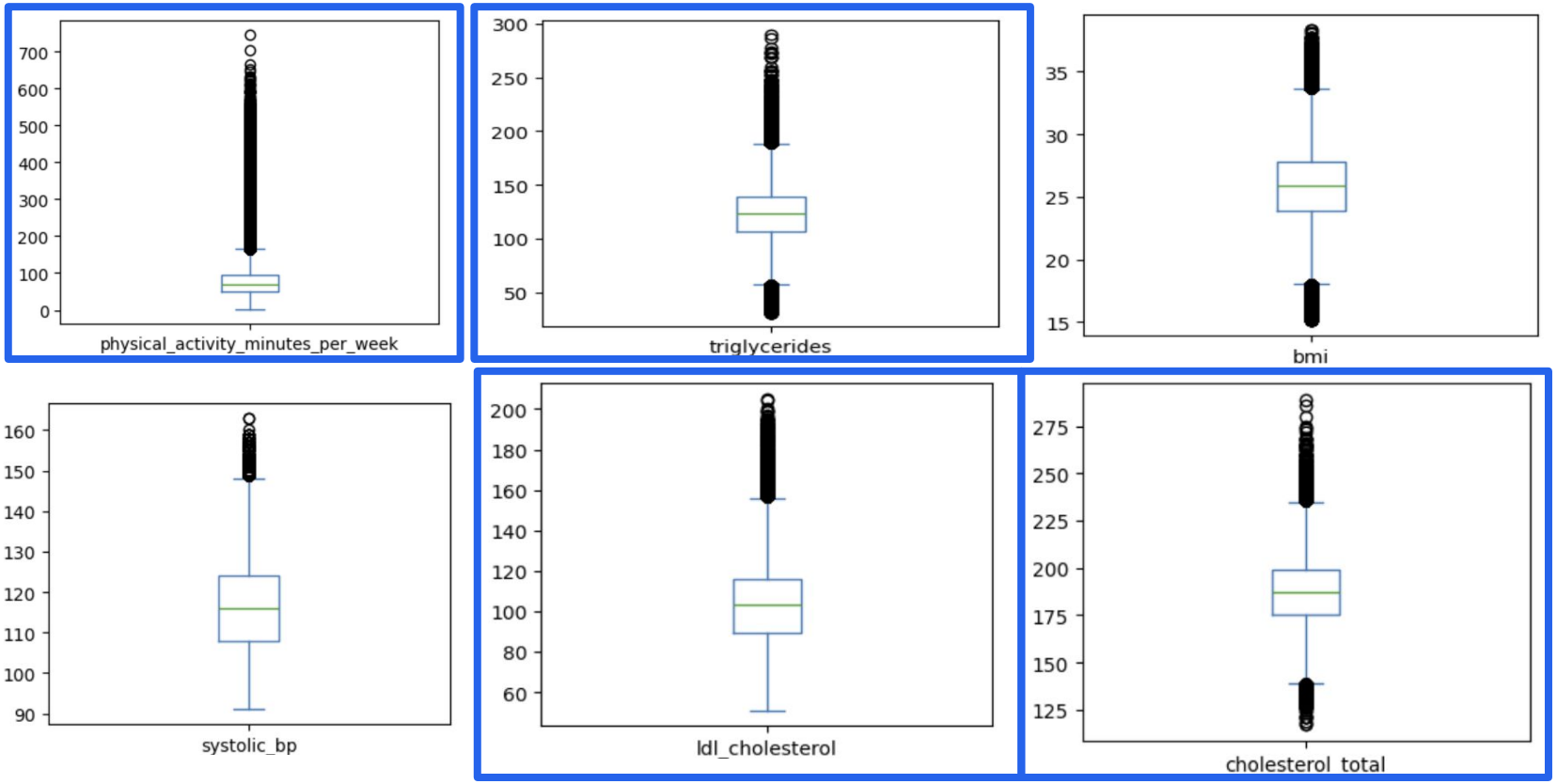
EDA(탐색적 데이터 분석) - 기초통계량 (2)

대상: Kaggle - Diabetes Prediction Challenge 의 train, test data set

	diastolic_bp	heart_rate	cholesterol_total	hdl_cholesterol	ldl_cholesterol	triglycerides	family_history_diabetes	hypertension_history	cardiovascular_history
mean	75.44	70.17	186.82	53.82	102.91	123.08	0.15	0.18	0.03
std	6.83	6.94	16.73	8.27	19.02	24.74	0.36	0.39	0.17
min	51.00	42.00	117.00	21.00	51.00	31.00	0.00	0.00	0.00
25%	71.00	65.00	175.00	48.00	89.00	106.00	0.00	0.00	0.00
50%	75.00	70.00	187.00	54.00	103.00	123.00	0.00	0.00	0.00
75%	80.00	75.00	199.00	59.00	116.00	139.00	0.00	0.00	0.00
max	104.00	101.00	289.00	90.00	205.00	290.00	1.00	1.00	1.00

EDA(탐색적 데이터 분석) - Boxplot

대상: Kaggle - Diabetes Prediction Challenge 의 train, test data set



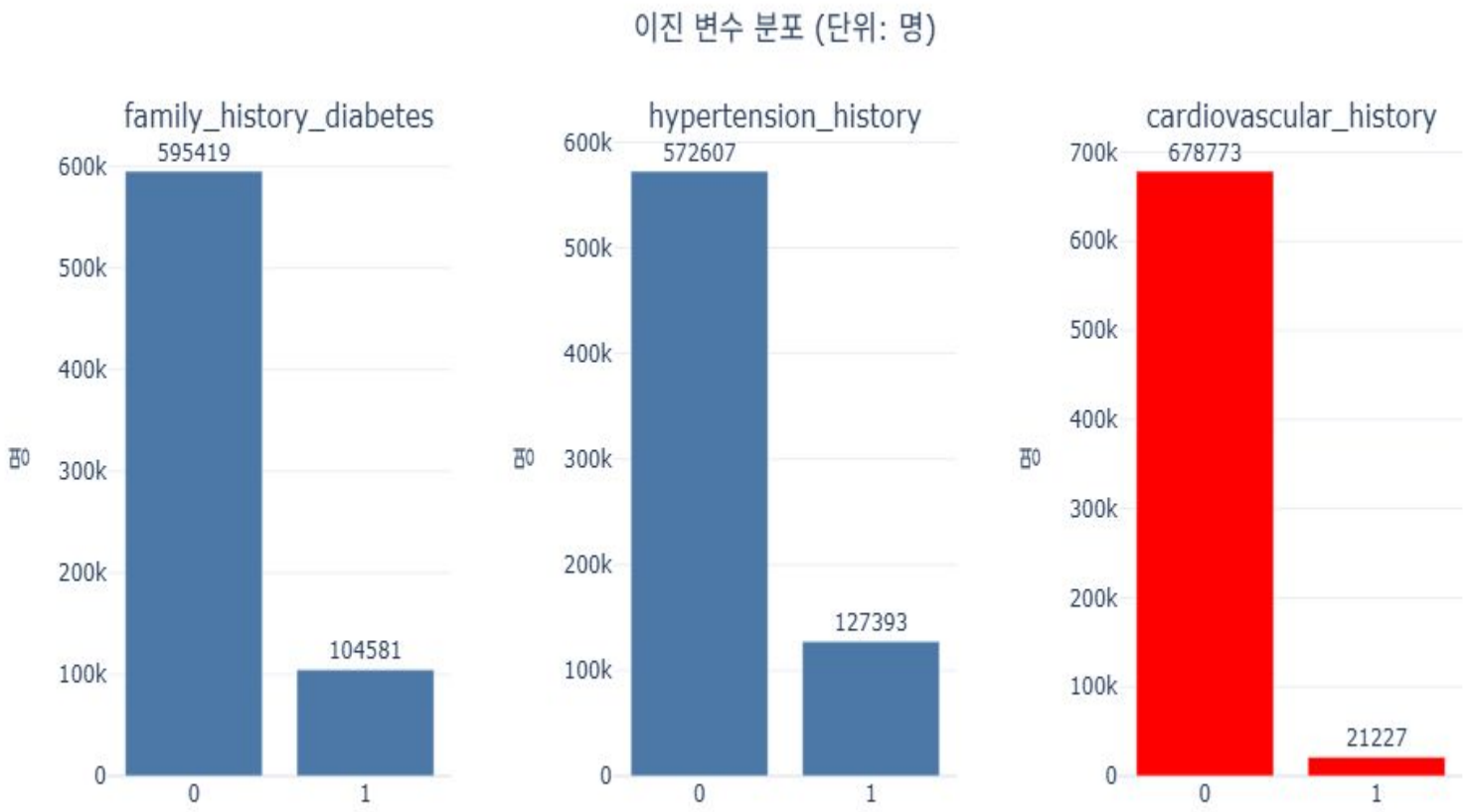
이상치 많고 치우친 분포

➡️ log변환 고려

➡️log변환 필요

EDA(탐색적 데이터 분석) - 이진변수

대상: Kaggle - Diabetes Prediction Challenge 의 train, test data set

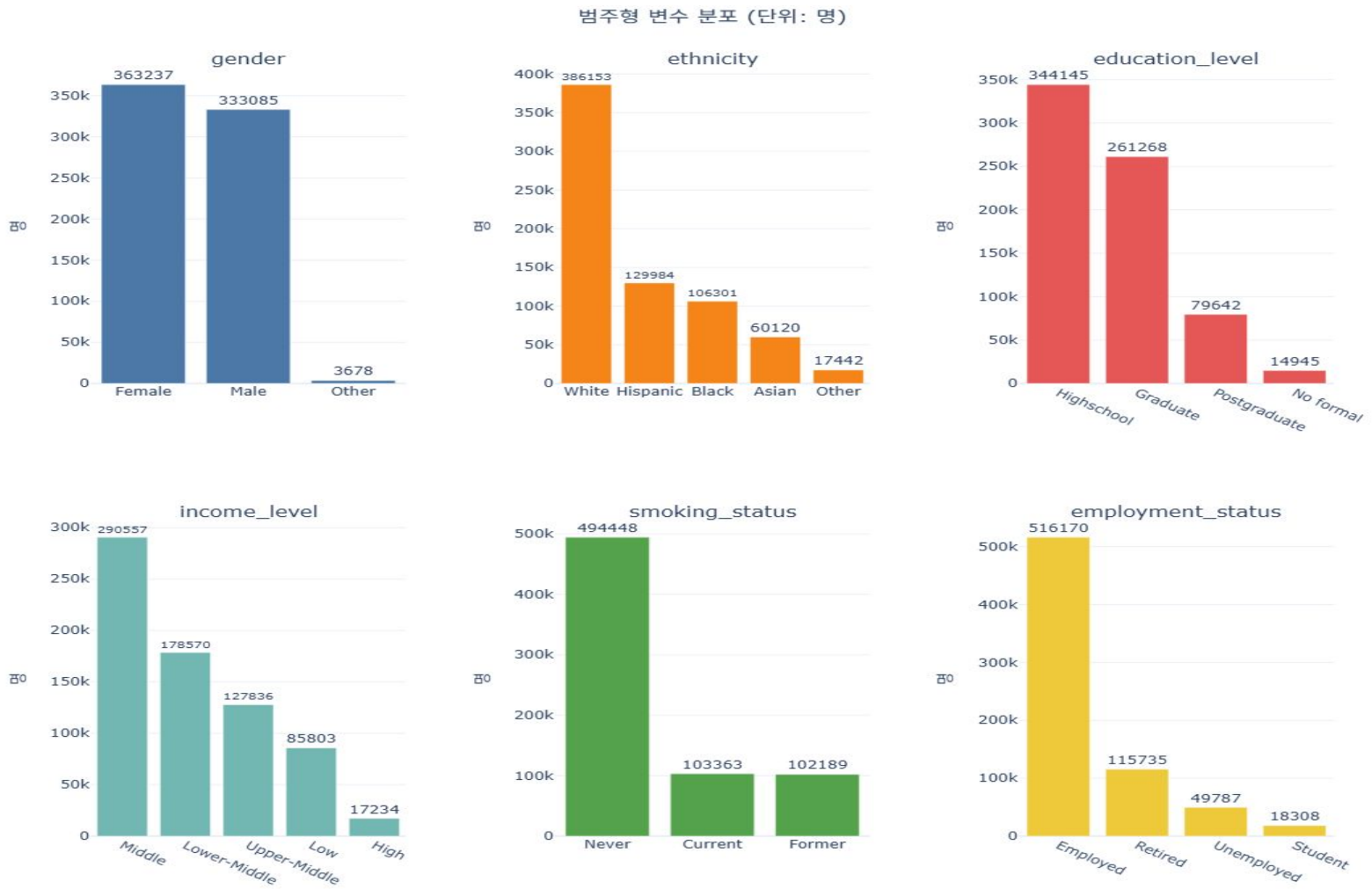


	0
family_history_diabetes	0.149401
hypertension_history	0.181990
cardiovascular_history	0.030324

cardiovascular_history
불균형이 크므로 분석 간
효과크기(Cramér's V) 반영하는
등 고려 필요

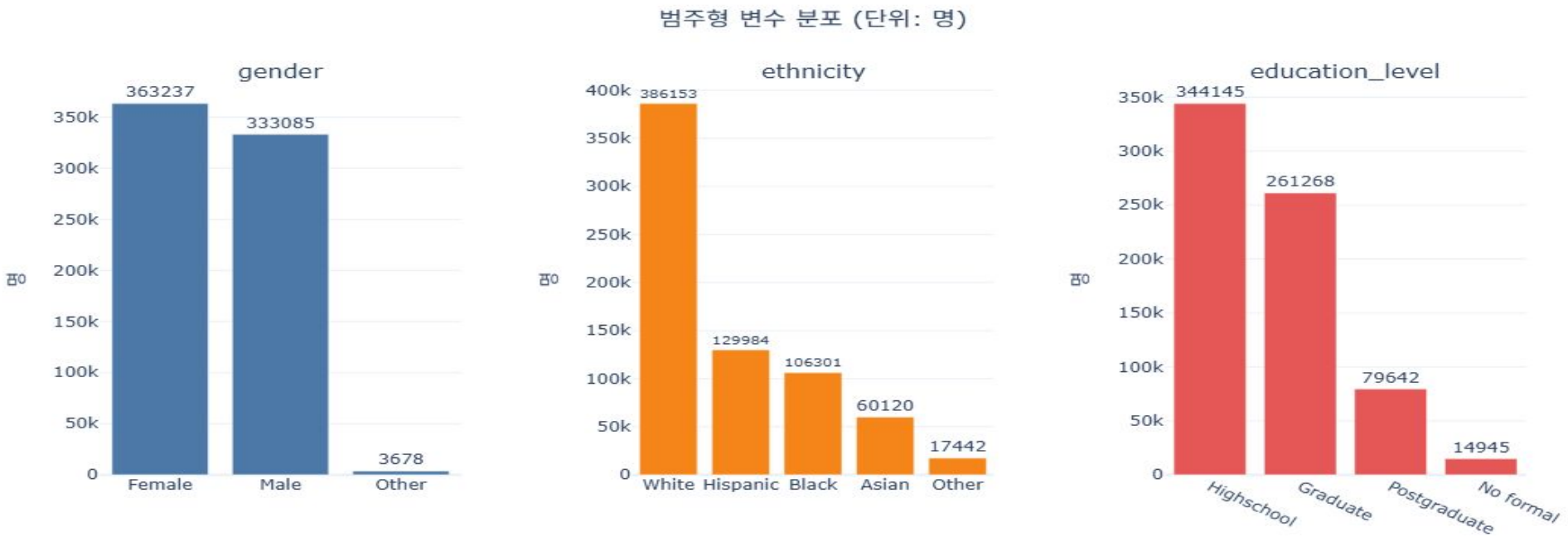
EDA(탐색적 데이터 분석) - 카테고리 변수

대상: Kaggle - Diabetes Prediction Challenge 의 train, test data set



EDA(탐색적 데이터 분석) - 카테고리 변수

대상: Kaggle - Diabetes Prediction Challenge 의 train, test data set



Gender(성별)

- 성비는 분석에 영향을 줄 만큼 크지 않음
- other은 별도 처리 필요하지 않음

Ethnicity(인종)

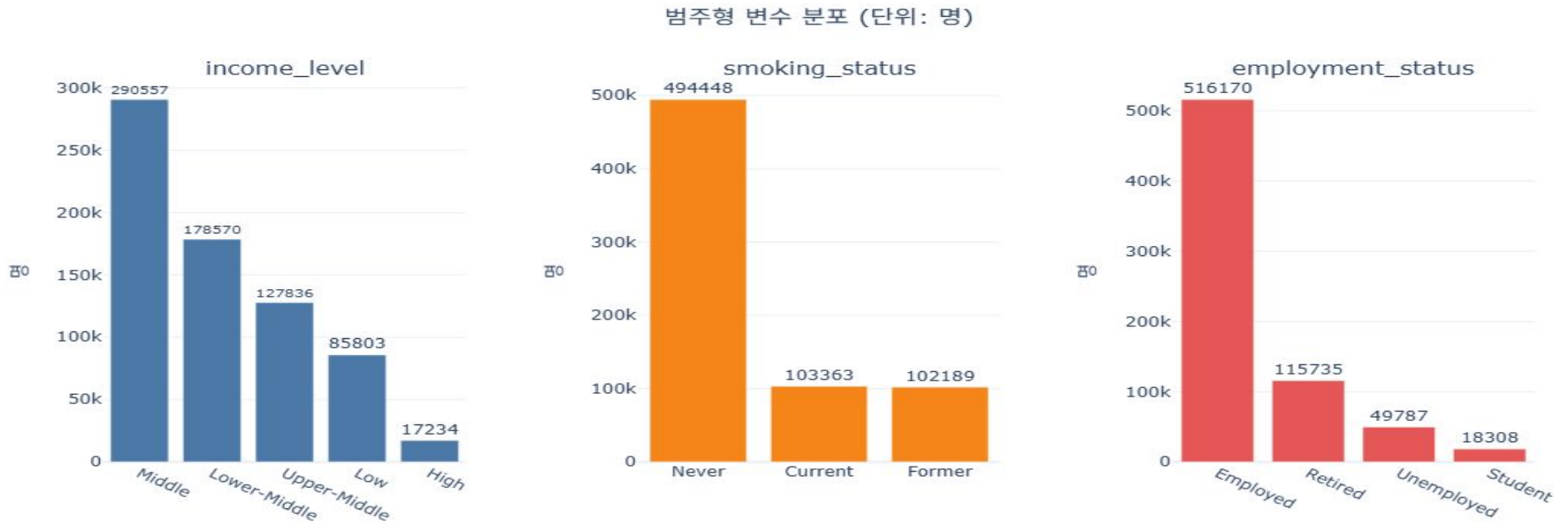
- **백인(White)** 그룹이 압도적으로 가장 높은 비중을 차지
- 그 뒤로 히스패닉(Hispanic), 흑인(Black), 아시아인(Asian) 순으로 **class 간 불균형**

Education_level(학력)

- **고졸(Highschool)** 학력자가 가장 많으며, 그 뒤 학사(Graduate) 학위
- 석/박사 이상고학력(Postgraduate), 정규 교육 부재(No formal) 비율은 상대적으로 낮음.

EDA(탐색적 데이터 분석) - 카테고리 변수

대상: Kaggle - Diabetes Prediction Challenge 의 train, test data set



Income_level(소득수준)

- **중간 소득(Middle)** 계층이 가장 높은 빈도.
- **고소득(High)** 층의 비율이 가장 적음.

Smoking_status(흡연여부)

- **비흡연자 (Never)**의 비중이 매우 높음
- 현재 흡연자(Current)와 과거 흡연자 (Former)의 비율은 비슷
- **중간정도 class 불균형**

Employment_status(고용형태)

- 취업 상태(Employed)가 대다수
- **class 분균형**

EDA(탐색적 데이터 분석) - 전처리 전략

대상: Kaggle - Diabetes Prediction Challenge 의 train, test data set

수치형 데이터

1. 각 요인별 단위가 다르 때문에 **Standard Scaling**
2. 이상치가 확인되므로 log변환 고려 / tree 기반 모델링 scaling 간 이진변수 제외
3. **BMI/허리-엉덩이 비율(WHR) 조합:** 비만도를 나타내는 두 지표를 곱하거나 조합하여 '복부 비만도'를 강조 가능.
   `복부 비만도 = BMI * waist_to_hip_ratio`
   이유: BMI가 높아도 근육질인 경우를 WHR이 보정해주고, BMI는 낮아도 배만 나온 '마른 비만'을 잡아낼 수 있다.
4. **연령(Age)과 BMI**가 타겟과 매우 강한 상관관계보이므로 이 두 변수의 분포가 특정 구간에 쏠려 있다면 구간화를 통해 "고령층", "고비만군" 등의 범주형 변수를 병행해서 생성하는 전략

범주형 데이터

1. **명목형 변수 (Nominal): One-Hot Encoding**
   `employment_status`에서 샘플 수가 매우 적은 'Student' 같은 범주는 인코딩 전, 'Unemployed'와 합쳐서 'Non-working'으로 그룹화하는 것이 모델의 일반화에 유리
2. **순서형 변수 (Ordinal): Ordinal Encoding**
3. **불균형 해소를 위한 범주 통합 (Binning)**

   Smoking Status: `Never`가 압도적이므로, [Never(0) vs Former/Current(1)]로 이진화(Binary)하여 '흡연 경험 유무'로 단순화하는 것이 효과적일 수 있음

   Ethnicity: `White` 이외의 소수 인종 데이터가 너무 적다면, 인종별 유전적 차이가 크지 않다는 가정하에 유사한 그룹끼리 묶어 범주의 개수를 줄여주는 것이 좋음

SECTION 03 STATISTIC ANALYSIS

통계적 분석

- 단변량 분석(T-test/Chi-square test)
- 다변량 분석(로지스틱회귀)
- 핵심 요인 도출

03

단변량 분석 결과(T-test)

통계적 유의성 (Statistical Significance)

배경 변수 (Demographic Factors)

Significant

Age의 t = 137.33, f = 18859.89, p<0.01로 나이 차이는 그냥 생긴게 아님이 확실함, 나이가 많을수록 당뇨 확률이 높음

생활습관 요인 (Lifestyle Factors)

Significant

Diet_score, Physical_activity 두 변수 모두 종속변수의 변화를 설명하는 핵심 예측 변수임을 확인, 이는 건강한 식단과 꾸준한 신체활동의 병행이 결과 지표 개선에 효과를 줌

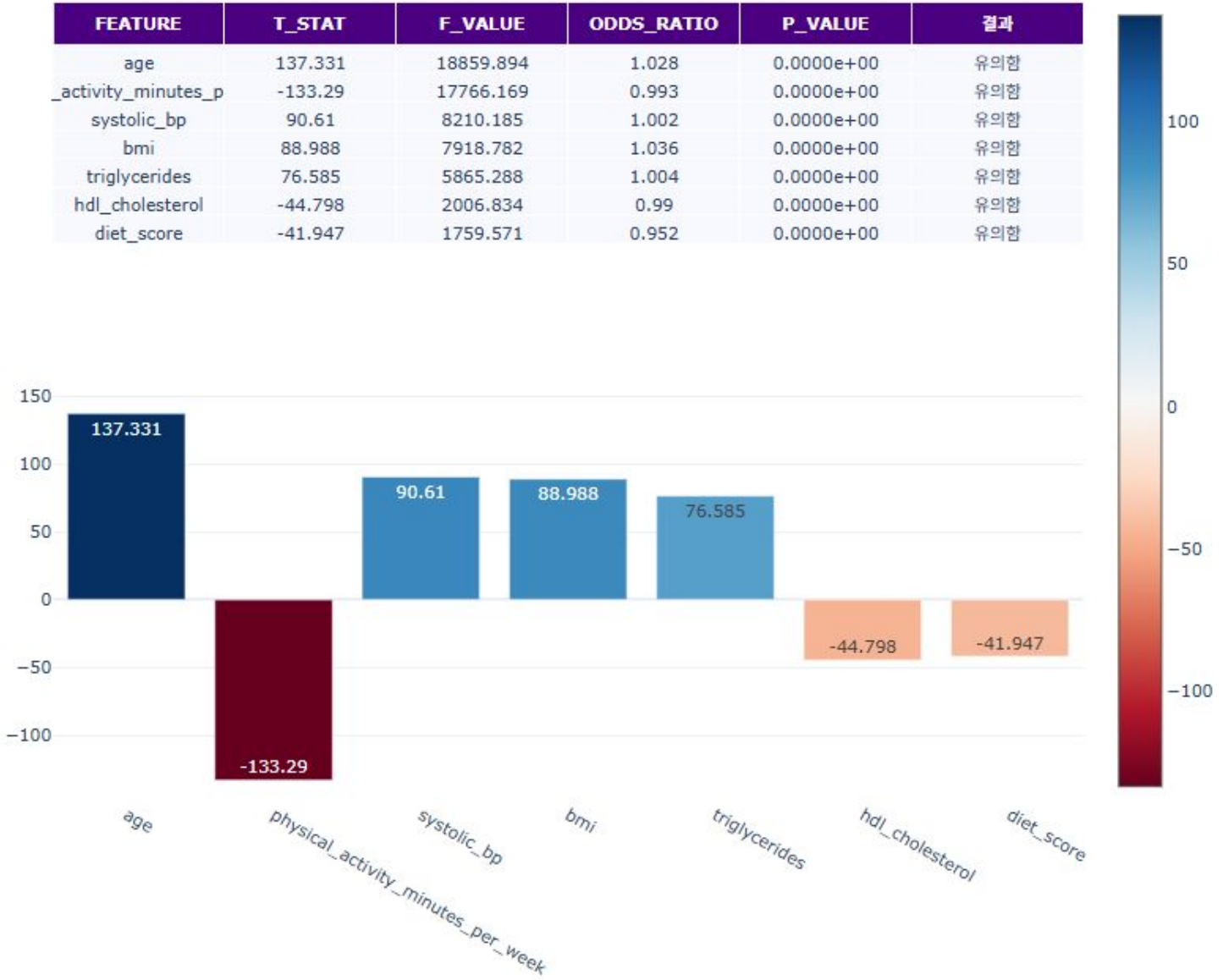
신체/대사 상태지표 (Physiological & Metabolic Lndicators)

Significant

Systolic_bp, Triglycerides, Hdl_cholesterol,Bmi은 집단 간 유의한 차이를 보여 당뇨 발생과 밀접한 관련이 있으며, 이들 지표의 복합적인 관리가 대사 건강 예측에 핵심적임을 시사

시각적 근거 (Visual Evidence)

FEATURE	T_STAT	F_VALUE	ODDS_RATIO	P_VALUE	결과
age	137.331	18859.894	1.028	0.0000e+00	유의함
_activity_minutes_p	-133.29	17766.169	0.993	0.0000e+00	유의함
systolic_bp	90.61	8210.185	1.002	0.0000e+00	유의함
bmi	88.988	7918.782	1.036	0.0000e+00	유의함
triglycerides	76.585	5865.288	1.004	0.0000e+00	유의함
hdl_cholesterol	-44.798	2006.834	0.99	0.0000e+00	유의함
diet_score	-41.947	1759.571	0.952	0.0000e+00	유의함



단변량 분석 결과(Chi-square)

시각적 근거 (Visual Evidence)

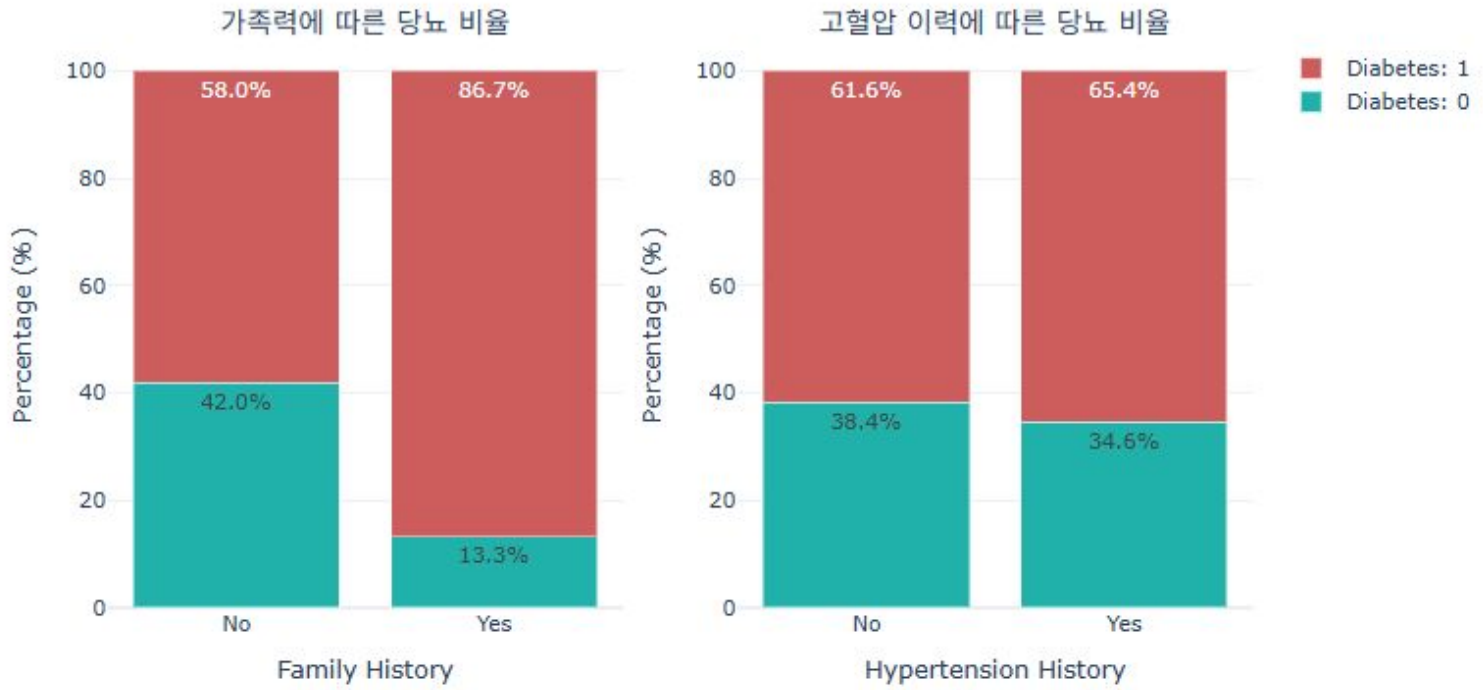
Feature (독립변수)	Chi2	P-Value	Odds Ratio	결과분석
family_history_diabetes	31182.4325	0.0000e+00	4.7255	유의한 상관관계 있음
hypertension_history	628.9778	8.3395e-139	1.1766	유의한 상관관계 있음

통계적 유의성 (Statistical Significance)

유전/병력 요인 (Genetic & Medical History)

Significant

Family_history, Hypertension History 두 변수 사이에 통계적으로 매우 유의미한 연관성이 확인되었는데, 가족 내 고혈압 내력이 있는 집단에서 발생빈도 유의하게 높으며 강력한 위험인자임을 나타냄



가족력, 고혈압 병력 2개의 범주형 요인을 카이제곱 검정을 통해 빈도 차이를 분석

로지스틱 회귀 분석 결과

Target Variable
Logistic Regression

주요 통계적 유의성 (Statistical Significance)

핵심 위험 인자 (Critical Risk Factors)
- 가장 치명적 유전인자 : 가족력(+1.92)은 전체 변수 중 압도적인 위험도, 연령(+0.25)보다 당뇨 위험을 약 7.7배 더 강력하게 높이는 핵심 인자

- 신체대사지표 : 중성지방(+1.24), BMI(+0.77) 수치가 높을수록 위험이 급증하며, 특히 BMI는 연령보다 위험도를 약 3배 상승시킴

- 기저 질환의 영향 : 고혈압 병력은(+0.65) 단순한 수축기혈압 수치(+0.16)보다 당뇨 위험을 약 4배 더 높이는 유의미한 과거력 인자로 확인

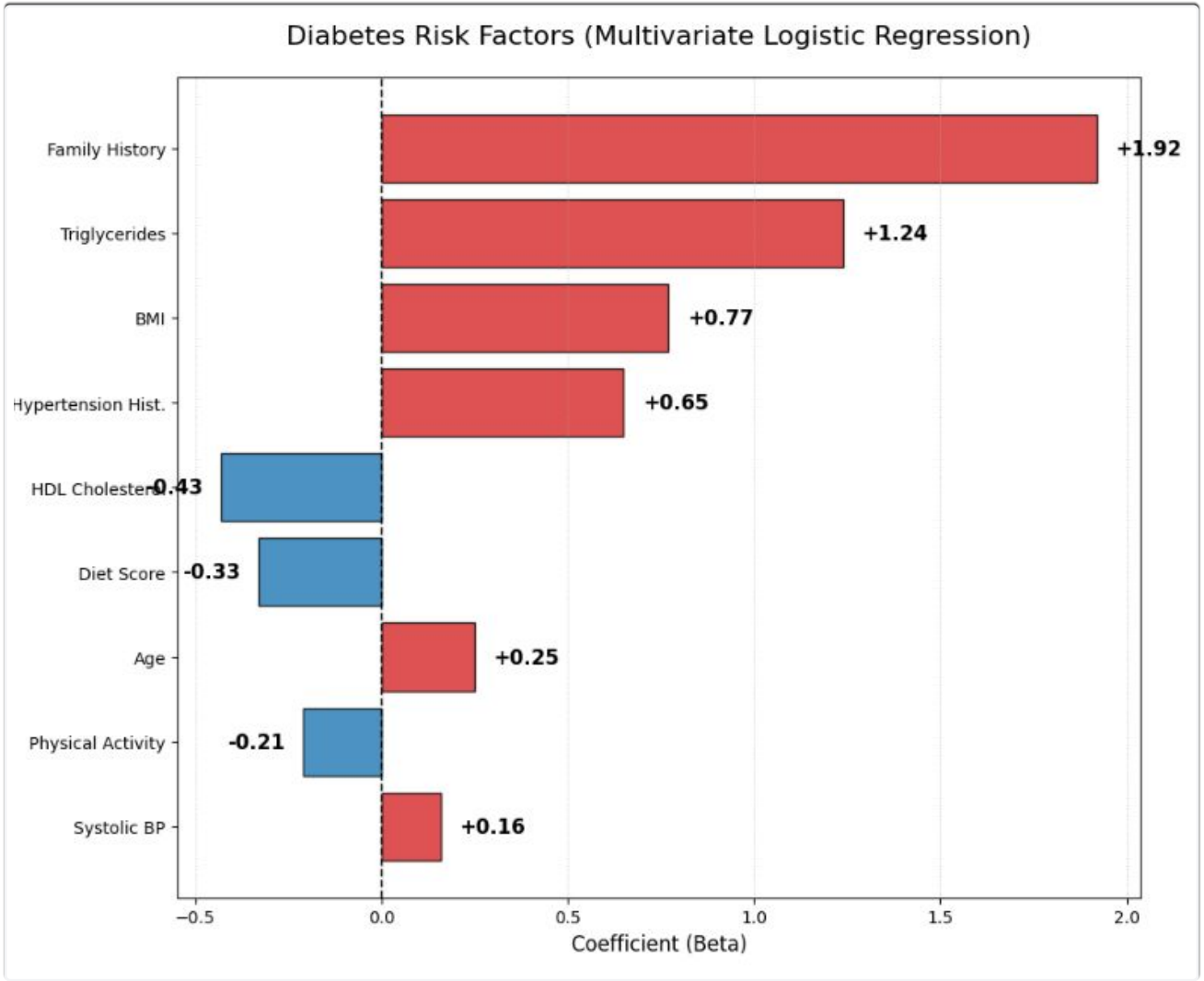
보호 요인 (Protective Factors)
-최고의 방어 인자: HDL 콜레스테롤(-0.43)은 가장 강력한 보호 효과를 나타내며, 신체활동량(-0.21)보다 약 2배 더 높은 위험 감소 기여도를 보임

-생활습관의 예방력 : 식단관리(-0.33)는 활동량(-0.21)보다 약 1.5배 더 효과적으로 당뇨 위험을 낮추는 보호 요인으로 분석됨

해석 (Interpretation)
-관리의 효율성 : 나이(연령+0.25)보다 비만(BMI+0.77)이 위험도를 약 3.1배 더 높이므로 노화보다 체중 관리가 예방에 훨씬 효과적

-통제가능한 변수의 중요성 : 가족력과 연령은 바꿀 수 있지만 중성지방과 BMI를 낮추고 HDL을 높이는 대사적 노력이 당뇨 발병을 막는 열쇠

요인별 상대적 영향력 (Standardized Coefficients)



핵심 요인 도출

배경 변수 (Demographic Factors)
핵심변수 : Age

Cohen's d: -0.337

연령이 증가함에 따라 췌장의 베타 세포 기능이 저하되고 인슐린 저항성이 증가하는 것은 당뇨병 발병의 전형적인 경로

연령은 비만, 신체 활동 부족 등 다른 가변적 위험 요인들에 노출된 기간을 대변하는 대리 변수

생활습관 요인 (Lifestyle Factors)
핵심변수 : Diet_score, Physical_activity

Cohen's d: 0.104 / 0.356

Physical_activity: 근육의 인슐린 감수성을 직접적으로 개선하여 혈당 조절 능력을 높이는 생리학적 경로를 가짐

Diet_score: 장기적인 대사 상태를 결정하는 기초 변수. 인슐린 저항성에 영향을 주는 변수

신체/대사 상태지표 (Physiological & Metabolic indicators)
핵심변수 : Bmi, Systolic_BP, Triglycerides, HDL_Colesterol

당뇨병의 직접적인 생체 신호로써, 상호 상관성이 높으므로 모델링 전 표준화 필수적이며 특히 Bmi, Triglycerides는 당뇨 진단의 핵심 수치적 근거가 됨

유전/병력 요인 (Genetic & Medical History)
핵심변수 : Family_history, Hypertension_history

카이제곱 분석으로 유의성이 검증된 범주형 변수로써 선천적 취약성을 대변하는 강력한 이진특징으로 작용함

분류된 4대 요인을 머신러닝(XGBoost, LightGBM)에 투입할 수 있도록 표준화 및 전처리 코드 제공

SECTION 04 MACHINE LEARNING ANALYSIS

머신러닝 분석

- 데이터 전처리
- 모델링 및 하이퍼파라미터 설정
- 모델 성능 평가

04

모델링 전략 및 실험 설계

예측 모델 및 전처리 (Models & Preprocessing)

CATEGORY	DETAILS & LIBRARIES
Baseline Models 기본 성능 확인	Linear Logistic Regression Tree Decision Tree Classifier
Ensemble Models 고성능 부스팅 알고리즘	Boosting XGBoost (Extreme Gradient Boosting) Boosting LightGBM (Light Gradient Boosting)
Preprocessing Column Transformer	수치형: StandardScaler (표준화) 범주형: OneHotEncoder / OrdinalEncoder

검증 전략 (Validation)

Stratified K-Fold CV

타겟 클래스 비율을 유지하며 5-Fold 교차 검증 수행 (n_splits=5). 데이터 불균형을 고려한 안정적인 평가.

PRIMARY METRIC

ROC-AUC Score

보조 지표: 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)

Validation Metric
ROC-AUC Score

MODEL NAME	ROC-AUC ⭐	ACCURACY	PRECISION	RECALL	F1 SCORE
Logistic Regression Baseline Linear	0.6932	0.6258	0.7541	0.5929	0.6639
Decision Tree Basic Tree	0.6939	0.6302	0.7522	0.6064	0.6715
XGBoost Gradient Boosting	0.7212	**0.6809**	0.7029	0.8450	**0.7675**
LightGBM **Best Model**	**0.7246**	0.6538	**0.7726**	**0.6301**	**0.69414**

최종 선정 모델: LIGHTGBM

4개 모델 중 ROC-AUC (0.7246) 점수가 가장 높으며, PRECISION(0.7725)에서도 우수한 성능을 보여 최종 모델로 선정하였습니다. XGBoost와 성능 차이는 미미했으나, 학습 속도와 일반화 성능 측면을 고려하였습니다.

성능 분석 인사이트

선선형 모델인 Logistic Regression과 Decision Tree는 낮은 Recall과 ROC-AUC를 보였습니다.
트리 기반 앙상블 모델인 XGBoost와 LightGBM이 비선형 패턴을 더 잘 학습해 Accuracy와 F1 Score가 크게 향상되었으며, XGBoost는 Recall, LightGBM은 Precision과 ROC-AUC에서 우수한 결과를 나타냈습니다.

최우수 모델 선정: LightGBM

Validation ROC_AUC
ROC-AUC: 0.7246

모델 선정 근거 및 최적화

최고 예측 성능 (Best Performance)

Rank 1

LightGBM은 검증 데이터셋에서 ROC-AUC 0.7245으로 가장 높은 성능을 기록하였으며, Logistic Regression(0.6932) 및 Decision Tree(0.6939) 대비 우수한 예측 성능을 보임

XGBoost(0.7213)와 유사한 성능을 보였으나,
계산 효율성 및 약간의 성능우수하단 점에서 LightGBM이 보다 더 유리하다고 판단

최적화 전략 (Optimization)

RandomizedSearchCV

Stratified K-Fold 교차검증 기반 하이퍼파라미터 튜닝 수행

Randomized Search 범위 확장
Train + Validation 데이터 통합 학습
모델 특성에 맞춘 Feature 및 Pipeline 조정

Objective: 일반화 성능(ROC-AUC 향상)

선정 이유 (Why LightGBM?)

LightGBM은 복잡한 데이터 분포를 효율적으로 학습할 수 있는 모델로 알려져 있으며, 교차검증 성능과 Validation 성능 간 차이가 크지 않아 비교적 안정적인 성능을 보임. 또한 이상치와 치우친 데이터에 강한 장점이 있음음

CONCLUSION
Validation ROC-AUC 기준으로 가장 높은 성능을 기록하며, 교차검증 결과와 Validation 성능이 유사하게 나타난 LightGBM을 당뇨병 예측을 위한 최종 모델로 선정

ROC - AUC 비교 (Validation Set)

SECTION 05 Results Analysis & Interpretation

분석 결과 및 해석

- 주요 요인 요약
- 모델 비교 및 최종 모델선정

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)

INTERNAL VALIDATION

0.69515

5-Fold Stratified CV Average

Public LB — Pending

Private LB — Hidden

검증 점수와 유사한 수준 예상

핵심 분석 결과 (Key Insights)

위험 및 보호 요인 식별

통계적 유의성 검증 완료

핵심 위험 인자 (Risk Factors)

가족력, 연령, 체질량지수(BMI), 수축기혈압, 중성지방,고혈압 병력

주요 보호 요인 (Protective Factors)

HDL_콜레스테롤, 신체활동, 식단점수

최적 모델 선정

LightGBM 모델 우수성 확인

선형 모델(Logistic Regression) 대비 LightGBM이 변수 간의 복잡한 비선형 관계와 상호작용을 더 효과적으로 포착하여 가장 우수한 성능을 기록함

BEST PERFORMANCE

Validation ROC-AUC **0.7246**

SECTION 06 Appendix

부록

- 참고문헌

06

참고문헌

Albers, J. D., Koster, A., Sezer, B., Meisters, R., Chan, J. A., Wesselius, A., Schram, M. T., De Galan, B. E., Lakerveld, J., & Bosma, H. (2025). Socioeconomic position and type 2 diabetes: Examining the mediating role of social cohesion—The Maastricht Study. *Social Science & Medicine*, *376*, 118046. https://doi.org/10.1016/j.socscimed.2025.118046

Duan, M.-J. F., Zhu, Y., Dekker, L. H., Mierau, J. O., Corpeleijn, E., Bakker, S. J. L., & Navis, G. (2022). Effects of Education and Income on Incident Type 2 Diabetes and Cardiovascular Diseases: A Dutch Prospective Study. *Journal of General Internal Medicine*, *37*(15), 3907–3916. https://doi.org/10.1007/s11606-022-07548-8

Hill-Briggs, F., Adler, N. E., Berkowitz, S. A., Chin, M. H., Gary-Webb, T. L., Navas-Acien, A., Thornton, P. L., & Haire-Joshu, D. (2021). Social Determinants of Health and Diabetes: A Scientific Review. *Diabetes Care*, *44*(1), 258–279. https://doi.org/10.2337/dci20-0053

Ismail, L., Materwala, H., & Al Kaabi, J. (2021). Association of risk factors with type 2 diabetes: A systematic review. *Computational and Structural Biotechnology Journal*, *19*, 1759–1785. https://doi.org/10.1016/j.csbj.2021.03.003

Katiyar, D. A. (2023). *Diabetes mellitus: Risk Factors Contributing to Type 2Diabetes*.

Kim, H.-J., Hwang, J.-E., & Boo, Y.-K. (2023). Lifestyle Factors Affecting Blood Sugar Control by Workers with Type 2 Diabetes using the Korean National Health and Nutrition Examination Survey, 2016-2020. *Journal of the Korea Academia-Industrial cooperation Society*, *24*(6), 105–115. https://doi.org/10.5762/KAIS.2023.24.6.105

Kim, S.-R., Han, K., Choi, J.-Y., Ersek, J., Liu, J., Jo, S.-J., Lee, K.-S., Yim, H. W., Lee, W.-C., Park, Y. G., Lee, S.-H., & Park, Y.-M. (2015). Age- and Sex-Specific Relationships between Household Income, Education, and Diabetes Mellitus in Korean Adults: The Korea National Health and Nutrition Examination Survey, 2008-2010. *PLOS ONE*, *10*(1), e0117034. https://doi.org/10.1371/journal.pone.0117034

Kim D-.J. (2008). The association of socioeconomic status with diabetes, and cardiovascular disease. The Korean Journal of Medicine (Korean J Med), 74(4), 349-357.

Ley, S. (2023). *Chapter 13: Risk Factors for Type 2 Diabetes*.

Liu, C., He, L., Li, Y., Yang, A., Zhang, K., & Luo, B. (2023). Diabetes risk among US adults with different socioeconomic status and behavioral lifestyles: Evidence from the National Health and Nutrition Examination Survey. *Frontiers in Public Health*, *11*, 1197947. https://doi.org/10.3389/fpubh.2023.1197947

Risk Factors of Diabetes. (2023).

Sharma, U. K., Pujani, M., & . A. (2024a). Type-II-Diabetes Mellitus- Etiology, Epidemiology, Risk Factors and Diagnosis and Insight into Demography (Urban Versus Rural). *International Journal of Health Sciences and Research*, *14*(1), 283–290. https://doi.org/10.52403/ijhsr.20240136

Sharma, U. K., Pujani, M., & . A. (2024b). Type-II-Diabetes Mellitus- Etiology, Epidemiology, Risk Factors and Diagnosis and Insight into Demography (Urban Versus Rural). *International Journal of Health Sciences and Research*, *14*(1), 283–290. https://doi.org/10.52403/ijhsr.20240136

Song, Z., Yang, R., Wang, W., Huang, N., Zhuang, Z., Han, Y., Qi, L., Xu, M., Tang, Y., & Huang, T. (2021). Association of healthy lifestyle including a healthy sleep pattern with incident type 2 diabetes mellitus among individuals with hypertension. *Cardiovascular Diabetology*, *20*(1), 239. https://doi.org/10.1186/s12933-021-01434-z

The InterAct Consortium. (2012). Long-Term Risk of Incident Type 2 Diabetes and Measures of Overall and Regional Obesity: The EPIC-InterAct Case-Cohort Study. *PLoS Medicine*, *9*(6), e1001230. https://doi.org/10.1371/journal.pmed.1001230

Wang, P., Gao, X., Willett, W. C., & Giovannucci, E. L. (2024). Socioeconomic Status, Diet, and Behavioral Factors and Cardiometabolic Diseases and Mortality. *JAMA Network Open*, *7*(12), e2451837. https://doi.org/10.1001/jamanetworkopen.2024.51837

Zhou, P., Geng, X., Zhang, C., Li, T., Li, F., Cao, Y., Mao, D., Liu, Y., Wang, N., Li, K., Xiao, Z., Shang, X., Feng, C., & Zong, G. (2025). The influence of socioeconomic disparities on the association of physical behavior with incident type 2 diabetes. *BMC Public Health*, *25*(1), 3579. https://doi.org/10.1186/s12889-025-24359-8