

머신러닝

회귀분석을 이용한 당뇨병 진단 프로젝트

TEAM B팀

팀장 박주호 | 김정현 | 임현주



목차

01. 프로젝트 개요

02. 프로젝트 팀 구성 및 역할

03. 프로젝트 수행절차 및 방법

04. 프로젝트 수행 경과

05. 자체 평가 의견

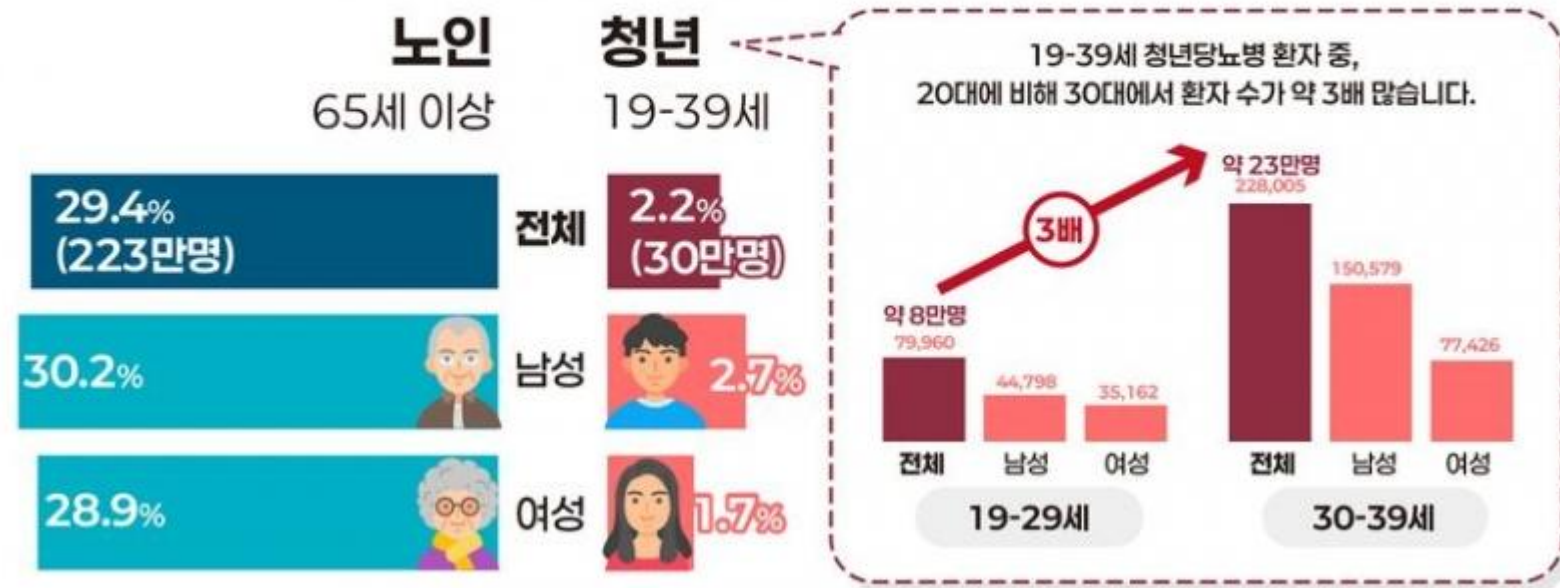
06. 마무리

01-1. 프로젝트 개요

30~40대 당뇨병 인지·치료·조절 낮아 관리 필요

당뇨병 유병자 인지·치료는 지속 개선...조절 수준은 정체

노인·청년당뇨병 유병률 비교



01

프로젝트 선정 이유

- 20대~40대 젊은 층에서 당뇨병 환자 증가
- 당뇨병 주요 영향 요인 분석 및 위험도 예측
- 조기 인지 및 치료 지원으로 예방과 건강 증진 목표

첨부1) <https://www.yakup.com/news/index.html?mode=view&cat=11&nid=288779>

첨부2) 대한당뇨병학회

01-2. 프로젝트 개요

01

프로젝트 내용

- 머신러닝 기반 당뇨병 예측 및 시각화
- 예측 유형: 이진 분류 (0: 비당뇨, 1: 당뇨)

02

활용 장비 및 재료

- Python
- Pandas, Scikit-learn, XGBoost, RandomForest, Streamlit 등

03

프로젝트 구조

- 의료현장 지원을 위한 조기 예측 도구 구축 목적
- 의료 데이터를 기반으로 당뇨병 유무 예측

02. 프로젝트 팀 구성 및 역할

박주호

팀장

- 데이터 수집
- 전처리 및 모델 학습
- PPT 작성

김정현

팀원

- 데이터 수집
- 전처리 및 모델 학습
- PPT 작성

임현주

팀원

- 데이터 수집
- 전처리 및 모델 학습
- PPT 작성

03. 프로젝트 수행 절차 및 방법

프로젝트 사전 기획 및 수행 과정

	4/2 ~ 5/2	5/2 ~ 5/9	비고
사전 기획	<div></div>		주제 선정
데이터 수집	<div></div>		데이터 선정
데이터 전처리	<div></div>		데이터 정제 및 정규화
모델링		<div></div>	모형 구현
서비스 구축		<div></div>	최적화 오류 수정
발표 자료 작성		<div></div>	PPT 작성

04. 프로젝트 수행 경과

데이터 수집

- Kaggle 데이터셋 수집

데이터 전처리

- 결측치/이상치 처리
- 범주형 변환, 파생 변수 생성
- 특성 스케일링

탐색적 데이터 분석 (EDA)

- 당뇨병 유무에 따른 변수 차이 분석
- 상관관계 히트맵

모델 정확도 향상

- 데이터 수준
- 모델링 수준
- 피처 엔지니어링 수준

모델 평가 및 비교

- 결과 비교표

Streamlit 활용

- 웹 배포

04. 프로젝트 수행 경과

데이터 수집

- 출처: Kaggle (Diabetes Dataset 기반)
- 샘플 수: 10,000개 (원본 100,000개 중 일부 발취)
- 분석 목적: 당뇨병 예측을 위한 주요 건강 지표 분석

```
import pandas as pd  
data = pd.read_csv("diabest_cut.csv")
```


04. 프로젝트 수행 경과

데이터 수집

변수명	설명
age	환자의 나이 (세)
bmi	체질량지수 (BMI = kg/m ²), 비만 여부 판단 지표
HbA1c_level	당화혈색소 수치 (%), 3개월 평균 혈당을 반영하는 지표
blood_glucose_level	혈당 수치 (mg/dL), 당뇨 진단 및 관리에 사용됨
hypertension	고혈압 여부 (0: 없음, 1: 있음)
heart_disease	심장병 병력 여부 (0: 없음, 1: 있음)
smoking_history	흡연 이력 (never, former, current 등 → 수치형으로 인코딩됨)
gender	성별 (0: 여성, 1: 남성 등으로 인코딩됨)
diabetes	당뇨병 여부 (0: 비당뇨, 1: 당뇨)

04. 프로젝트 수행 경과

데이터 전처리

결측값 처리

- bmi, age 변수에 평균값으로 대체
- 결측치로 인한 학습 오류 방지

```
data['bmi'] = data['bmi'].fillna(data['bmi'].mean())  
data['age'] = data['age'].fillna(data['age'].mean())
```

범주형 변수 인코딩

- gender, smoking_history → Label Encoding
- 문자열을 수치형으로 변환하여 모델 학습 가능하게 함

```
le = LabelEncoder()  
data['gender'] = le.fit_transform(data['gender'])  
data['smoking_history'] = le.fit_transform(data['smoking_history'])
```

04. 프로젝트 수행 경과

데이터 전처리

파생 변수 생성

- diabetes_risk: BMI ≥ 30 또는 HbA1c ≥ 6.5 → 당뇨 고위험
- high_glucose_risk: 혈당 ≥ 180 → 고혈당 위험

```
data['diabetes_risk'] =  
((data['bmi'] >= 30) | (data['HbA1c_level'] >= 6.5)).astype(int)  
data['high_glucose_risk'] =  
(data['blood_glucose_level'] >= 180).astype(int)
```

04. 프로젝트 수행 경과

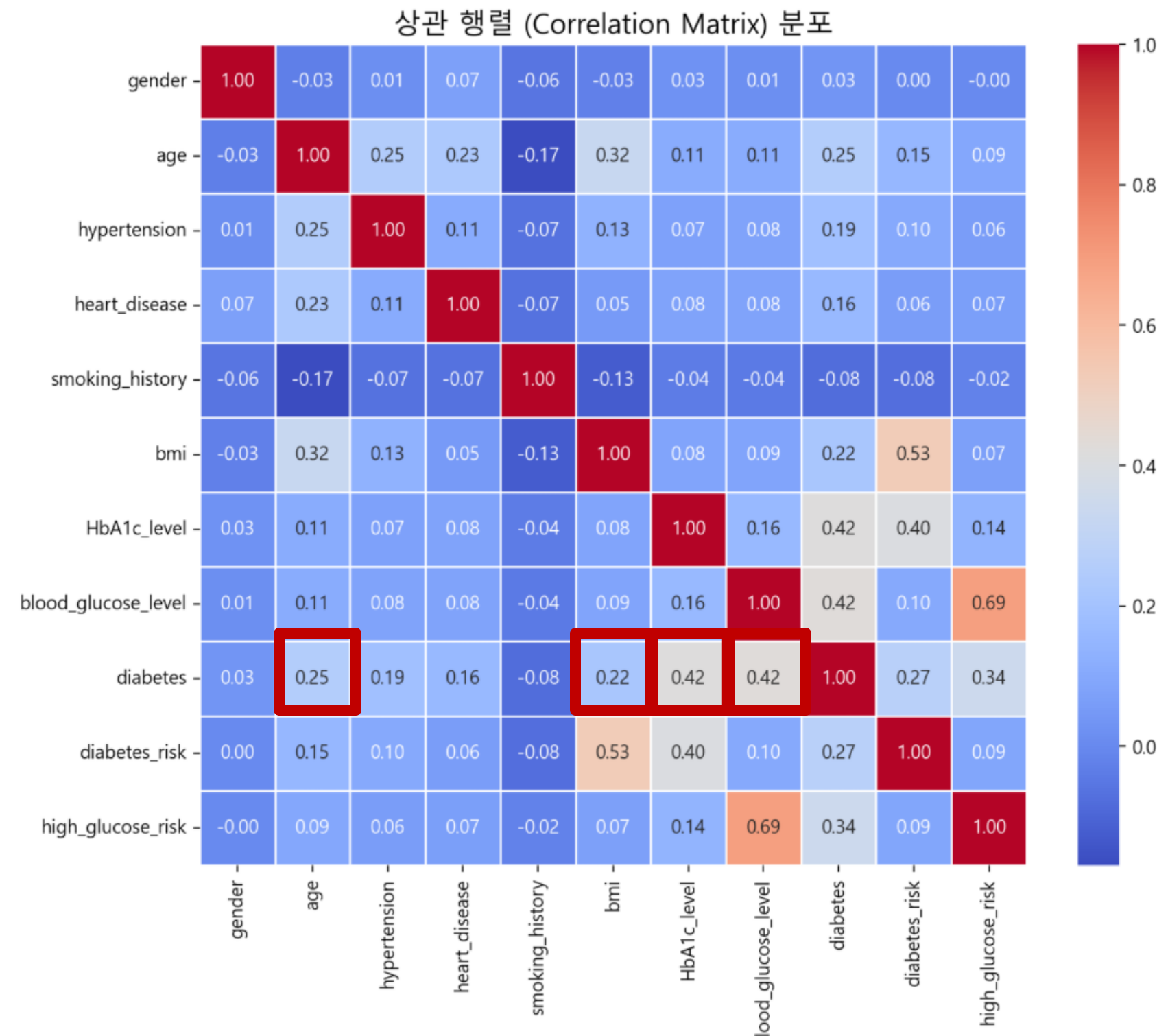
탐색적 데이터
분석 (EDA)

상관 관계 분석 (히트맵)

```
plt.figure(figsize=(12, 10))  
corr_matrix = data.corr()  
sns.heatmap(corr_matrix, annot=True, cmap="coolwarm", fmt=".2f",  
linewidths=0.5)  
plt.title("변수 간 상관관계 히트맵")  
plt.show()
```

04. 프로젝트 수행 경과

탐색적 데이터
분석 (EDA)



상관 관계 분석 (당뇨병 여부)

- BMI(0.22)는 약한 상관 관계를 보이지만, 비만일 수록 당뇨병 발생과 관련이 깊음
- HbA1c_level(0.42)는 높은 상관 관계
- 혈당 수치(0.42)와 높은 상관관계
- 연령(0.25)은 다른 변수들과 약한 상관 관계를 보이지만, 나이가 들수록 당뇨 위험이 증가하는 경향으로 보임

04. 프로젝트 수행 경과

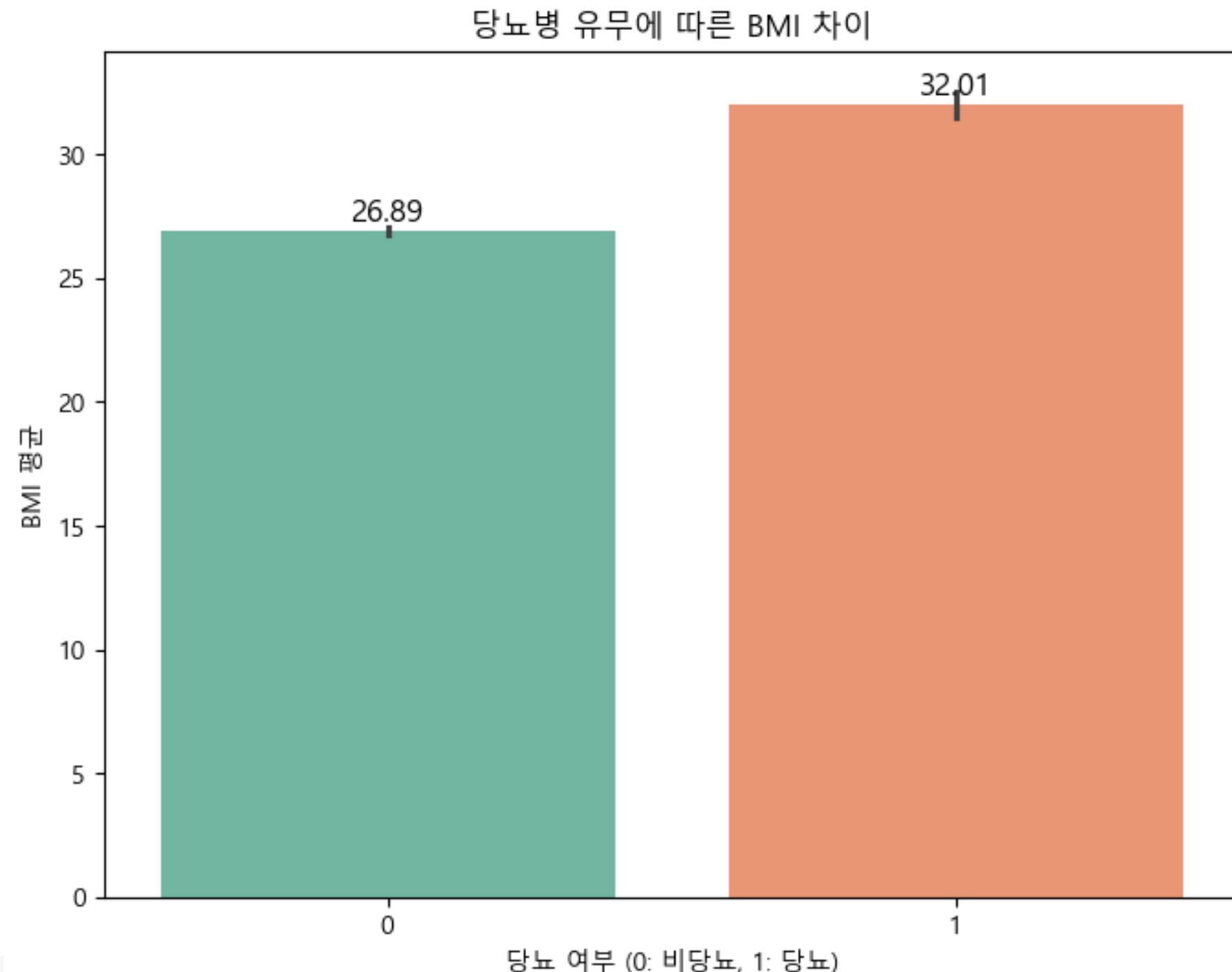
탐색적 데이터
분석 (EDA)

당뇨병 유무에 따른 BMI 분포 (막대그래프)

```
plt.figure(figsize=(8, 6))
sns.barplot(x='diabetes', y='bmi', data=data, palette="Set2")
plt.title("당뇨병 유무에 따른 BMI 차이")
plt.xlabel("당뇨 여부 (0: 비당뇨, 1: 당뇨)")
plt.ylabel("BMI 평균")
plt.show()
```

04. 프로젝트 수행 경과

탐색적 데이터
분석 (EDA)



당뇨병 유무에 따른 BMI 분석 결과

- BMI가 32(kg/m^2) 이상인 사람들의 평균이 당뇨병 유무가 1인 그룹에서 더 높게 나타남
- BMI가 32(kg/m^2) 이상인 사람일수록 당뇨병에 걸릴 확률이 높다는 의미

04. 프로젝트 수행 경과

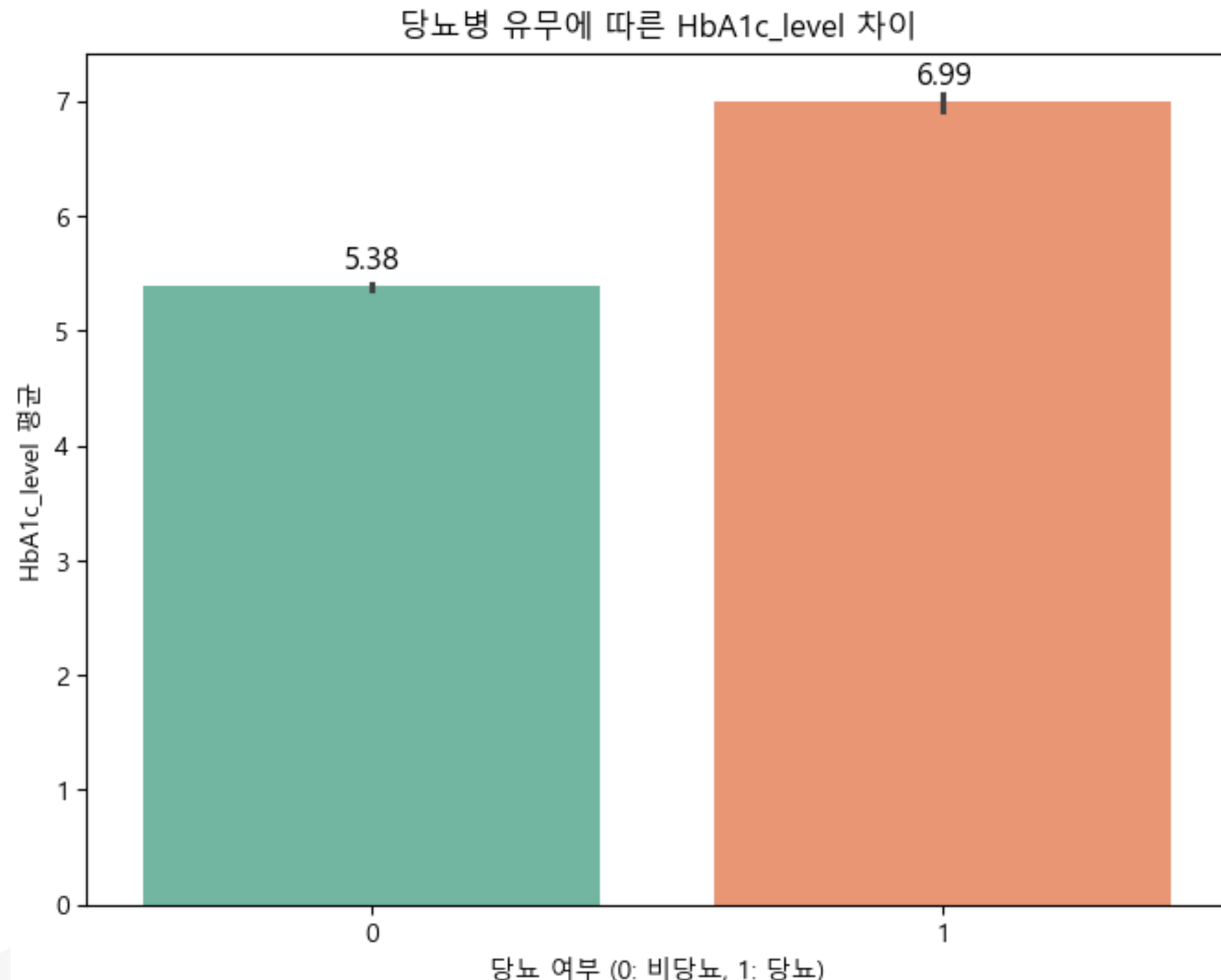
탐색적 데이터
분석 (EDA)

당뇨병 유무에 따른 HbA1c_level 분포 (막대그래프)

```
plt.figure(figsize=(8, 6))
sns.barplot(x='diabetes', y='HbA1c_level', data=data, palette="Set2")
plt.title("당뇨병 유무에 따른 HbA1c_level 차이")
plt.xlabel("당뇨 여부 (0: 비당뇨, 1: 당뇨)")
plt.ylabel("HbA1c_level 평균")
plt.show()
```


04. 프로젝트 수행 경과

탐색적 데이터
분석 (EDA)



당뇨병 유무에 따른 HbA1c 분석 결과

- HbA1c가 6.99(%) 이상인 사람들의 평균이 당뇨병 유무가 1인 그룹에서 더 높게 나타남
- HbA1c가 6.99(%) 이상인 사람일수록 당뇨병에 걸릴 확률이 높다는 의미

04. 프로젝트 수행 경과

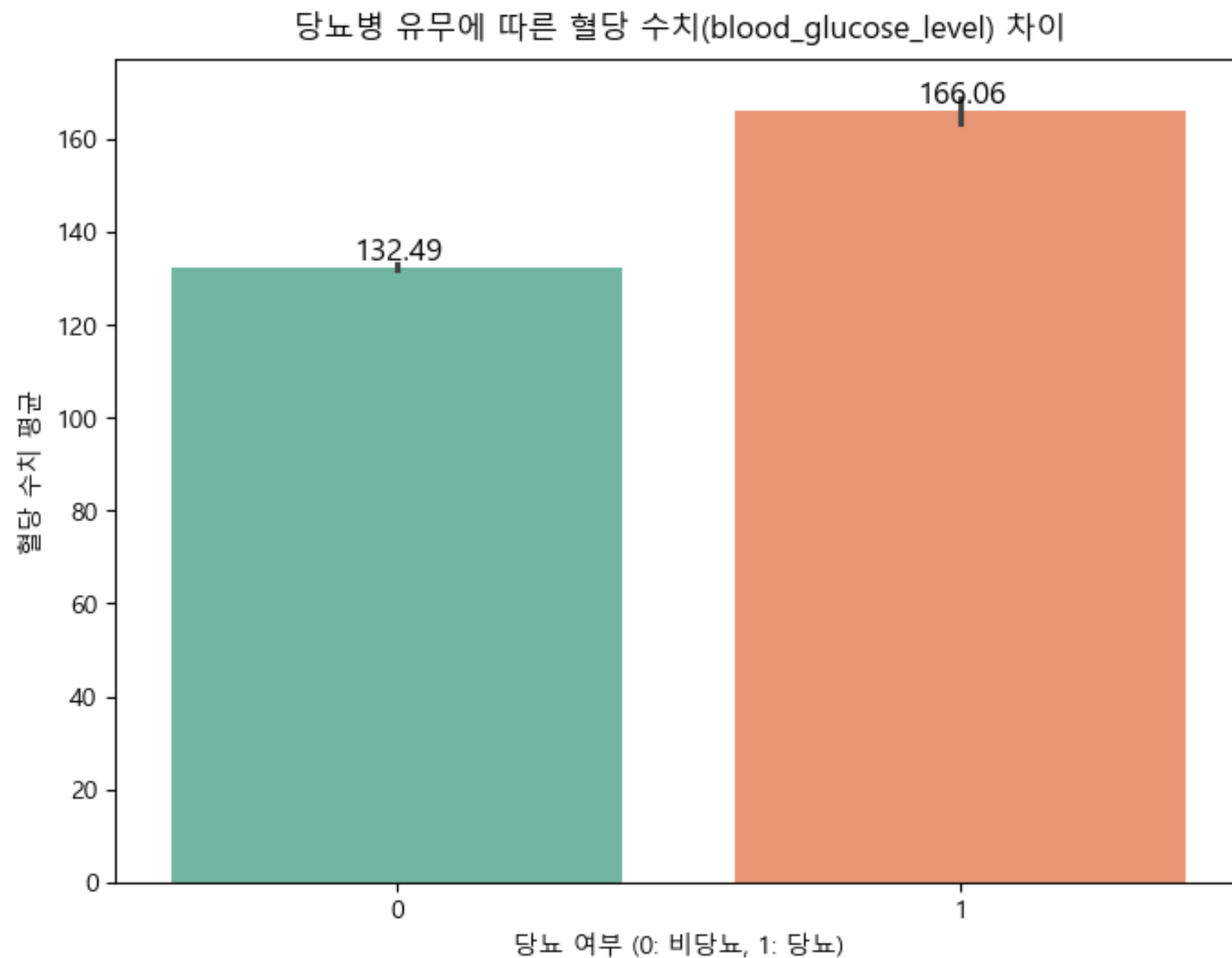
탐색적 데이터
분석 (EDA)

당뇨병 유무에 따른 blood_glucose_level 분포 (막대그래프)

```
plt.figure(figsize=(8, 6))
sns.barplot(x='diabetes', y='blood_glucose_level', data=data, palette="Set2")
plt.title("당뇨병 유무에 따른 혈당 수치(blood_glucose_level) 차이")
plt.xlabel("당뇨 여부 (0: 비당뇨, 1: 당뇨)")
plt.ylabel("혈당 수치 평균")
plt.show()
```

04. 프로젝트 수행 경과

탐색적 데이터
분석 (EDA)



당뇨병 유무에 따른 blood_glucose_level 분석 결과

- 혈당 수치가 166.06(mg/dL) 이상인 사람들의 평균이 당뇨병 유무가 1인 그룹에서 더 높게 나타남
- 혈당 수치가 166(mg/dL) 이상인 사람일수록 당뇨병에 걸릴 확률이 높다는 의미

04. 프로젝트 수행 경과

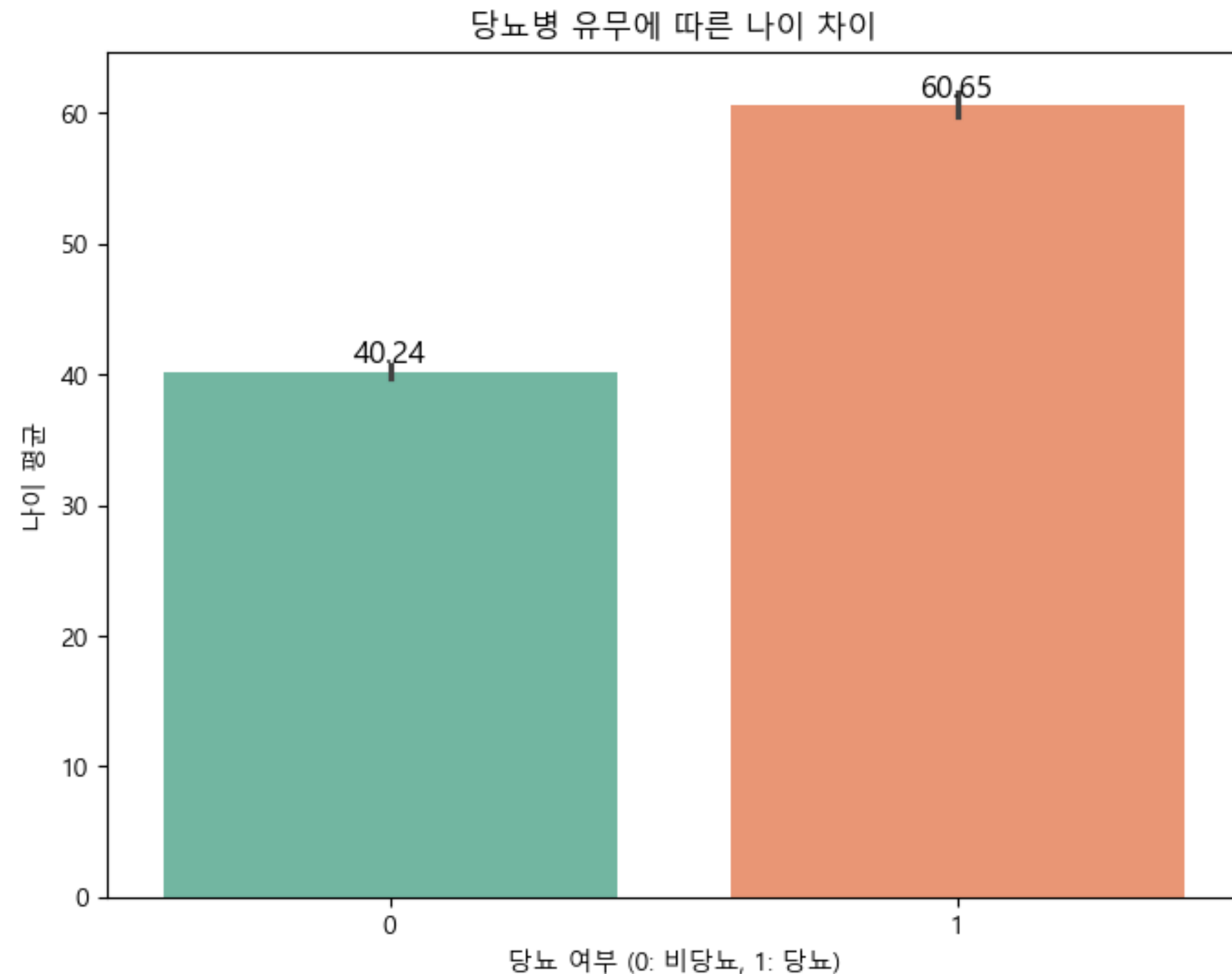
탐색적 데이터
분석 (EDA)

당뇨병 유무에 따른 age 분포 (막대그래프)

```
plt.figure(figsize=(8, 6))
sns.barplot(x='diabetes', y='age', data=data, palette="Set2")
plt.title("당뇨병 유무에 따른 나이 차이")
plt.xlabel("당뇨 여부 (0: 비당뇨, 1: 당뇨)")
plt.ylabel("나이 평균")
plt.show()
```

04. 프로젝트 수행 경과

탐색적 데이터
분석 (EDA)



당뇨병 유무에 따른 age 분석 결과

- 연령대가 60세 이상인 사람들의 평균이 당뇨병 유무가 1인 그룹에서 더 높게 나타남
- 연령대가 60세 이상인 사람일수록 당뇨병에 걸릴 확률이 높다는 의미

04. 프로젝트 수행 경과

모델 정확도 향상

모델 정확도 개선 전 데이터 준비

- 특성과 타겟 분리

```
X = data.drop(columns=['diabetes'])  
y = data['diabetes']
```

- 데이터 분할

```
X_train, X_test, y_train, y_test = train_test_split(X, y, test_size=0.2,  
random_state=42)
```

04. 프로젝트 수행 경과

모델 정확도 향상

데이터 수준

- 스케일링

```
scaler = StandardScaler()  
X_train_scaled = scaler.fit_transform(X_train)  
X_test_scaled = scaler.transform(X_test)
```

- 클래스 불균형 해결

```
smote = SMOTE(random_state=42)  
X_train_resampled, y_train_resampled = smote.fit_resample(X_train_scaled,  
y_train)
```

04. 프로젝트 수행 경과

모델 정확도 향상

모델링 수준

- 로지스틱 회귀 (Logistic Regression)

```
log_model = LogisticRegression(max_iter=1000, random_state=42)
```

- 랜덤 포레스트 (Random Forest)

```
rf_model = RandomForestClassifier(random_state=42)
```

- XGBoost (eXtreme Gradient Boosting)

```
xgb_model = XGBClassifier(random_state=42)
```


04. 프로젝트 수행 경과

모델 정확도 향상

모델링 수준

- 교차 검증

```
skf = StratifiedKFold(n_splits=5, shuffle=True, random_state=42)
for model, name in zip([log_model, rf_model, xgb_model], ['Logistic', 'Random Forest',
'XGBoost']):
    scores = []
    for train_idx, val_idx in skf.split(X_train_resampled, y_train_resampled):
        model.fit(X_train_resampled[train_idx], y_train_resampled[train_idx])
        pred = model.predict(X_train_resampled[val_idx])
        acc = accuracy_score(y_train_resampled[val_idx], pred)
        scores.append(acc)
```

04. 프로젝트 수행 경과

모델 정확도 향상

모델링 수준

- 하이퍼파라미터 튜닝(GridSearchCV)

```
param_grid = {'n_estimators': [100, 200], 'max_depth': [3, 5, 7]}  
grid = GridSearchCV(rf_model, param_grid, cv=3, scoring='accuracy')  
grid.fit(X_train_resampled, y_train_resampled)
```

04. 프로젝트 수행 경과

모델 정확도 향상

모델링 수준

- 앙상블 모델 활용 (Voting)

```
voting = VotingClassifier(estimators=[
    ('lr', log_model),
    ('rf', grid.best_estimator_),
    ('xgb', xgb_model)], voting='soft')
voting.fit(X_train_resampled,
y_train_resampled)
y_pred = voting.predict(X_test_scaled
```

- 앙상블 모델 활용 (Stacking)

```
stacking = StackingClassifier(
    estimators=[('lr', log_model),
    ('rf', grid.best_estimator_)],
    final_estimator=XGBClassifier
    (random_state=42), cv=3)
stacking.fit(X_train_resampled,
y_train_resampled)
stack_pred = stacking.predict(X_test_scaled)
```

04. 프로젝트 수행 경과

모델 정확도 향상

피처 엔지니어링 수준

- 상관성 낮은 변수 제거

```
corr_matrix = data.corr()
low_corr_features = corr_matrix['diabetes'][abs(corr_matrix['diabetes'])
< 0.05].index.tolist()
print("제거된 변수:", low_corr_features)
data_reduced = data.drop(columns=low_corr_features)
```

- 결과

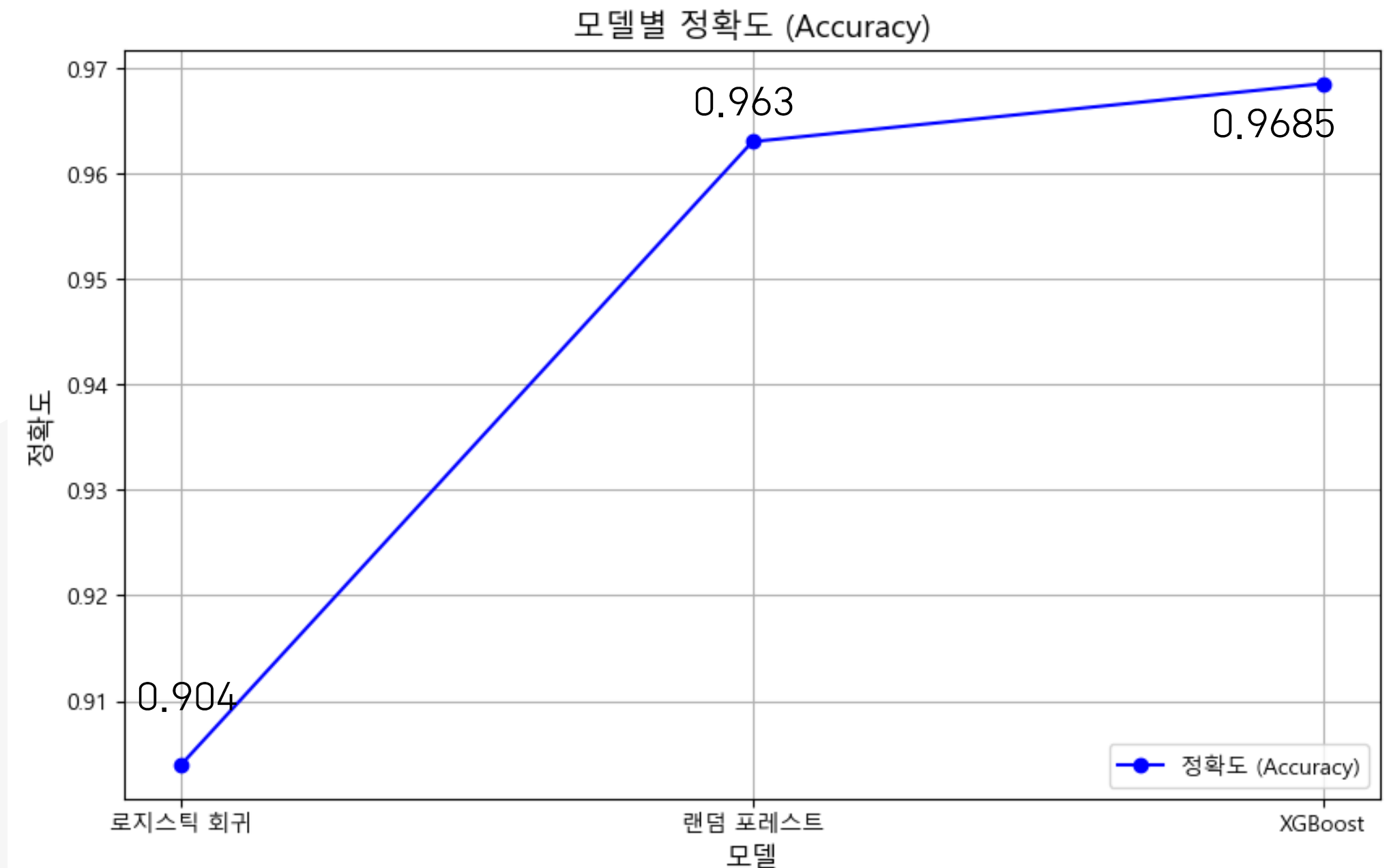
제거된 변수: ['gender']

04. 프로젝트 수행 경과

모델 평가 및 비교

모델별 정확도 (Accuracy)

- XGBoost의 정확도가 가장 높은 것을 알 수있음



04. 프로젝트 수행 경과

모델 평가 및 비교

성능 요약 보고서 (classification_report)

로지스틱 회귀

	precision	recall	f1-score	support
비당뇨	0.992	0.903	0.945	1836.000
당뇨	0.457	0.915	0.610	164.000
accuracy	0.904	0.904	0.904	0.904
macro avg	0.724	0.909	0.778	2000.000
weighted avg	0.948	0.904	0.918	2000.000

랜덤 포레스트

	precision	recall	f1-score	support
비당뇨	0.979	0.981	0.980	1836.000
당뇨	0.781	0.762	0.772	164.000
accuracy	0.963	0.963	0.963	0.963
macro avg	0.880	0.872	0.876	2000.000
weighted avg	0.963	0.963	0.963	2000.000

XGBoost

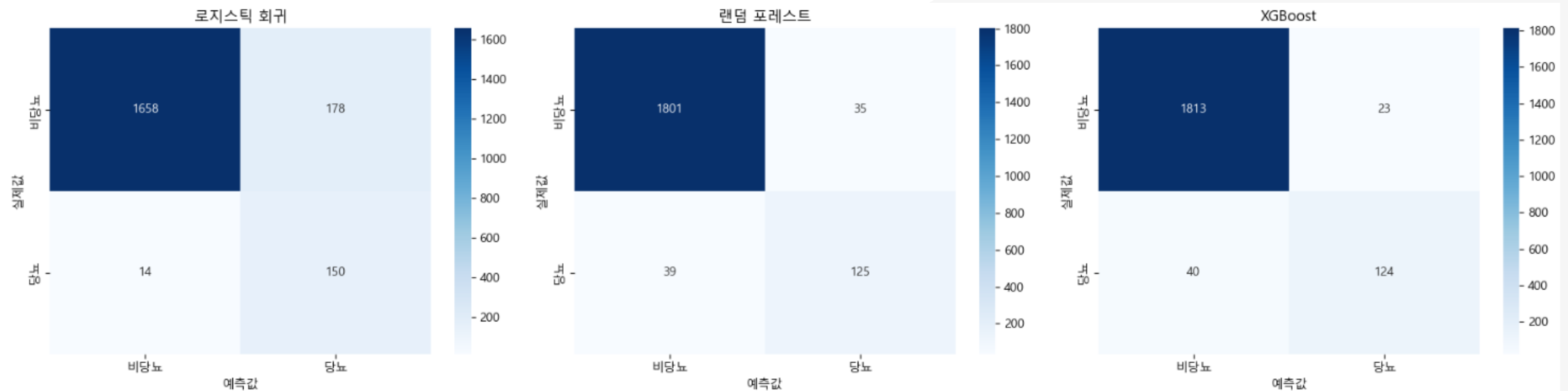
	precision	recall	f1-score	support
비당뇨	0.978	0.987	0.983	1836.000
당뇨	0.844	0.756	0.797	164.000
accuracy	0.968	0.968	0.968	0.968
macro avg	0.911	0.872	0.890	2000.000
weighted avg	0.967	0.968	0.968	2000.000

- 당뇨 예측 성능에서는 랜덤 포레스트와 XGBoost가 우수하며, 로지스틱 회귀는 당뇨 예측에서 성능이 떨어짐

04. 프로젝트 수행 경과

모델 평가 및 비교

혼동 행렬 (confusion_matrix)



- 비당뇨 예측에서는 모든 모델이 높은 정확도를 보였고, 특히 XGBoost가 당뇨 예측에서 가장 우수한 성능을 나타냄

04. 프로젝트 수행 경과

Streamlit 앱 개발

04. 프로젝트 수행 경과

sidebar: 사용자 입력


탭 1: 데이터 탐색

탭 2: 예측 결과

탭 3: 모델 성능 시각화

탭 4: 인사이트 및 요약

‘비당뇨’인 사용자 데이터

 신규 이용자 정보 입력

나이

79- +

BMI

28.73- +

HbA1c 수치

6.60- +

혈당 수치

160- +

성별

남성▼

흡연 이력

No Info▼

‘당뇨’인 사용자 데이터

 신규 이용자 정보 입력

나이

80- +

BMI

32.32- +

HbA1c 수치

8.20- +

혈당 수치

159- +

성별

여성▼

흡연 이력

No Info▼

04. 프로젝트 수행 경과

sidebar: 사용자 입력

탭 1: 데이터 탐색

탭 2: 예측 결과

탭 3: 모델 성능 시각화

탭 4: 인사이트 및 요약

당뇨 예측 시스템

데이터 탐색 예측 결과 모델 성능 인사이트 요약


데이터 탐색 (EDA)

데이터 샘플


	gender	age	hypertension	heart_disease	smoking_history	bmi	HbA1c_level	blood_glucose_
0	0	80	0	1	2	25.19	6.6	
1	0	54	0	0	2	27.32	6.6	
2	1	28	0	0	2	27.32	5.7	
3	0	36	0	0	0	23.45	5	
4	1	76	1	1	0	20.14	4.8	

EDA 설명

- 이 섹션에서는 각 변수의 분포를 살펴보고, 데이터 간 상관 관계를 분석하여 당뇨병의 위험 요인을 이해합니다.
- 데이터의 주요 변수에 대한 시각화 및 인사이트를 제공합니다.

 변수별 시각화 보기



 상관 행렬 보기



 변수별 시각화 보기



04. 프로젝트 수행 경과 **Streamlit** 활용

sidebar: 사용자 입력

탭 1: 데이터 탐색

탭 2: 예측 결과

탭 3: 모델 성능 시각화

탭 4: 인사이트 및 요약

비당뇨 사용자 입력 데이터 결과

XGBoost 모델이 비당뇨를 정확하게 예측


당뇨 예측 시스템

데이터 탐색 **예측 결과** 모델 성능 인사이트 요약

 신규 환자 예측 결과


 XGBoost 모델 결과

✓ 예측된 결과: **비당뇨**

 당뇨 확률: 33.17%


 Random Forest 모델 결과

✓ 예측된 결과: **당뇨**

 당뇨 확률: 59.26%

 Logistic Regression 모델 결과

✓ 예측된 결과: **당뇨**

 당뇨 확률: 93.22%

04. 프로젝트 수행 경과 **Streamlit** 활용

sidebar: 사용자 입력

탭 1: 데이터 탐색

탭 2: 예측 결과

탭 3: 모델 성능 시각화

탭 4: 인사이트 및 요약

당뇨 사용자 입력 데이터 결과

모든 모델이 당뇨를 정확하게 예측


당뇨 예측 시스템

데이터 탐색 **예측 결과** 모델 성능 인사이트 요약

 신규 환자 예측 결과


 XGBoost 모델 결과


✓ 예측된 결과: 당뇨

 당뇨 확률: 99.84%


 Random Forest 모델 결과

✓ 예측된 결과: 당뇨

 당뇨 확률: 94.29%

 Logistic Regression 모델 결과

✓ 예측된 결과: 당뇨

 당뇨 확률: 99.94%

04. 프로젝트 수행 경과 **Streamlit** 활용

sidebar: 사용자 입력

탭 1: 데이터 탐색

탭 2: 예측 결과

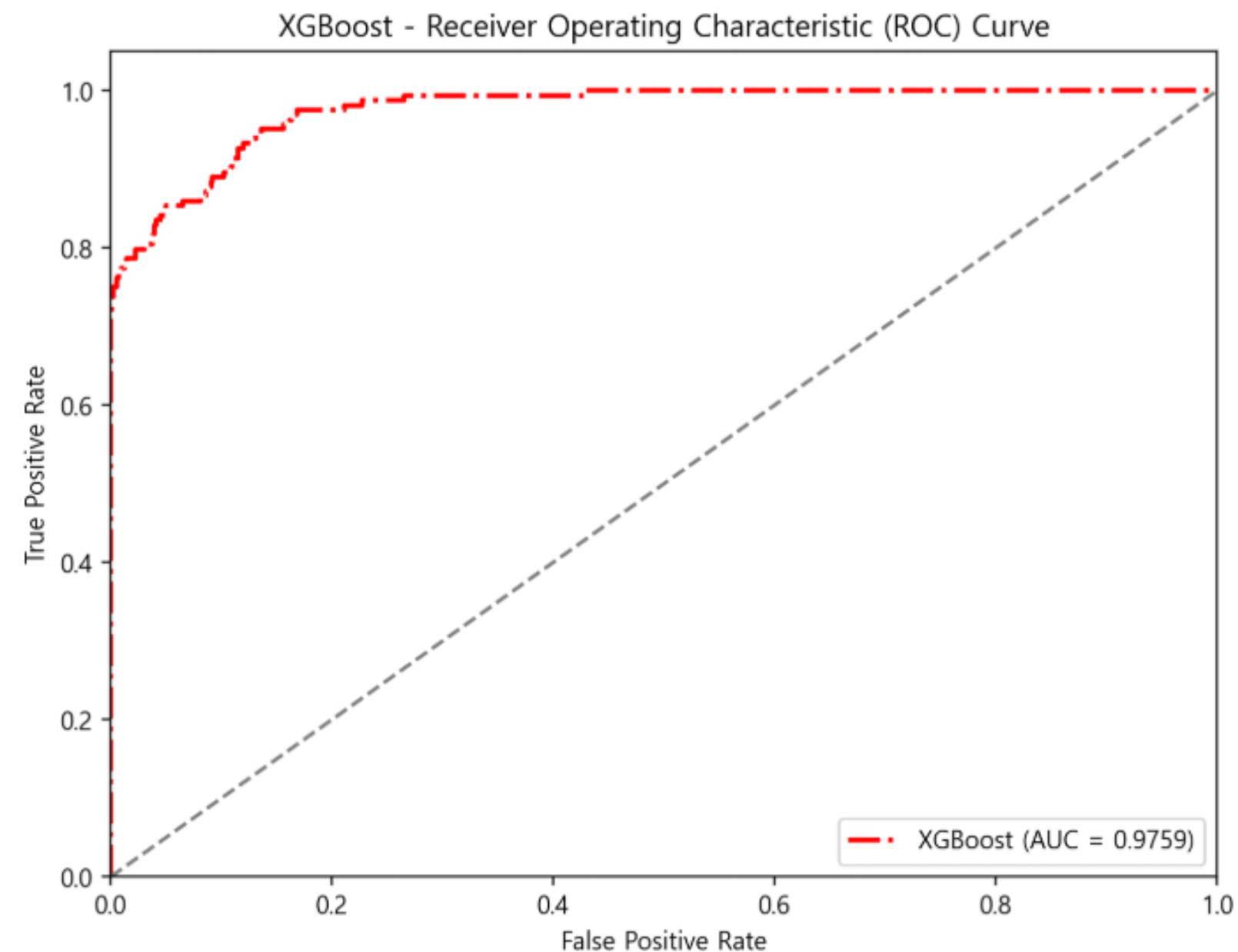
탭 3: 모델 성능 시각화

탭 4: 인사이트 및 요약

ROC 곡선 결과 해석

빨간 선(AUC)이 1에 가까울수록 우수한 모델임을 나타내는데 위 그래프의 AUC는 0.9759로 거의 완벽에 가까운 분류 능력을 보여줌

XGBoost의 ROC Curve



04. 프로젝트 수행 경과 **Streamlit** 활용

sidebar: 사용자 입력

탭 1: 데이터 탐색

탭 2: 예측 결과

탭 3: 모델 성능 시각화

탭 4: 인사이트 및 요약



당뇨 예측 시스템

데이터 탐색 예측 결과 모델 성능 **인사이트 요약**



인사이트 요약

- HbA1c 수치, BMI, 혈당 수치가 높을수록 당뇨 확률이 증가합니다.
- 흡연 이력은 간접적으로 당뇨와 관련이 있을 수 있습니다.
- XGBoost 모델이 전반적으로 가장 높은 성능을 보였습니다.

04. 프로젝트 수행 경과

Streamlit 앱 시연

05. 자체 평가 의견

당화혈색소(HbA1c) 수치 예측 시스템 구축

당화혈색소 수치를 병원에 가지 않고도 쉽게 예측할 수 있도록, 비의료 환경에서도 활용 가능한 머신러닝 기반 모델을 추가로 구축하고자 함.

서비스 다각화 계획

당뇨병 예측 시스템 구축에 그치지 않고, 당뇨병과 밀접한 관련이 있는 고혈압, 심혈관질환 등 만성질환 예측 기능까지 확장하여, 통합적인 만성질환 관리 및 예방이 가능한 시스템으로 발전시키고자 함.

Q & A

감사합니다

