분류 모델을 통한 당뇨병 발병 예측

[24-1 데이터마이닝]

Term Project

2021150456 이예지

목차

[데이터마이닝][Term Project]

1. 서론

1-1. 연구의 필요성

질병관리청에서 수행한 2022 국민건강영양조사 결과에 따르면, 만 19세 이상 성인의 약 9% 정도가 당뇨병을 앓고 있다. 연령이 높아질수록 당뇨병 유병률이 높아지는 것을 확인할 수 있다. 남성은 만 50세 이상에서 당뇨병 유병률이 20%를 넘었으며, 여성은 만 50세 이상에서 당뇨병 유병률이 10%를 넘었다. 당뇨병은 한번 발병하면 평생 지속되며, 다양한 합병증을 유발할 수 있어 위험하다.

따라서 본 프로젝트는 다양한 신체적, 환경적 요인을 통해 당뇨병 환자인지, 당뇨병 위험군인지, 당뇨병을 가지고 있지 않은지 예측하는 모델을 구축하여 당뇨병 환자를 더 빠르고 정확하게 분류함을 목적으로 한다. 또한, 당뇨병 환자로 예측하는 요인을 탐색함으로써 당뇨병을 예방할 수 있는 방법에 대해서도 추가로 알아볼 것이다.

1-2. 연구 방법

가장 먼저 bar plot, 히스토그램, 이변량 그래프 등의 시각화를 통해 데이터셋의 분포와 특징을 확인할 것이다. 이후 전체 데이터셋과 Resampling 기법을 사용한 데이터셋에 대해 모델을 학습해 가장 성능이 뛰어난 모델을 선정하려 한다.

사용할 모델은 Multinomial Logistic Regression, Classification and Regression Tree, Naïve Bayes Classifier, Artificial Neural Network, CART Bagging, Random Forest, ANN Bagging, Adaptive Boosting, Gradient Boosting Machine이며, 데이터셋에 불균형이 존재하므로 F1-Measure를 주요 평가지표로 활용할 것이다.

2. 데이터셋 탐색

데이터셋 다운로드 링크는 다음과 같다.

Diabetes Health Indicators Dataset (kaggle.com)

2-1. 데이터셋 설명

CDC(Centers for Disease Control and Prevetion)의 2015 BRFSS Survey Data에서 파생된 데이터로, 총 253,680개의 데이터가 있으며, class imbalance가 존재한다. 결측치는 존재하지 않는다. 총 21개의 설명변수와 1개의 종속변수가 있다.

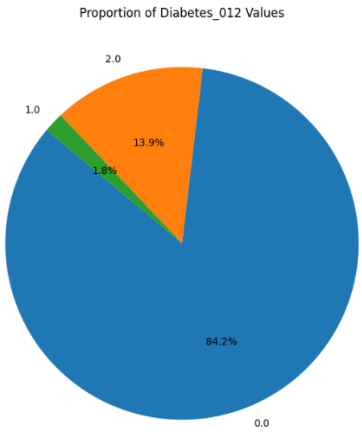
[표 1] 데이터셋 변수 설명

변수명	변수 설명
Diabetes_012	당뇨병 진단 여부.

(종속변수)	0=당뇨 없음, 1=당뇨 위험 단계, 2=당뇨
HighBP (설명변수)	고혈압 진단 여부. 0=없음, 1=있음.
HighChol (설명변수)	고콜레스테롤 진단 여부. 0=없음, 1=있음.
CholCheck (설명변수)	지난 5년 내에 콜레스테롤 검사를 받았는지 여부. 0=없음, 1=있음.
BMI (설명변수)	체질량 지수.
Smoker (설명변수)	평생 동안 100개비 이상의 담배를 피운 적이 있는지 여부. 0=없음, 1=있음.
Stroke (설명변수)	뇌졸중 진단 여부. 0=없음, 1=있음.
HeartDiseaseorAttack (설명변수)	관상 동맥 심장병 또는 심근경색 진단 여부. 0=없음, 1=있음.
PhysActivity (설명변수)	지난 30일 동안 정기적인 직업 외에 신체 활동이나 운동을 한 적이 있는 성인. 0=없음, 1=있음.
Fruits (설명변수)	하루에 한번 이상 과일을 섭취하는지 여부. 0=없음, 1=있음.
Veggies (설명변수)	하루에 한번 이상 채소를 섭취하는지 여부. 0=없음, 1=있음.
HvyAlcoholConsump (설명변수)	과음 여부. 성인 남성은 주당 14잔 이상, 성인 여성은 주당 7잔 이상의 음주. 0=없음, 1=있음.
AnyHealthcare (설명변수)	종류에 관계없이 건강 관리 보장이 있는지 여부. 0=없음, 1=있음.
NoDocbcCost (설명변수)	지난 12개월 동안 비용 때문에 병원에 가야함에도 불구하고 병원을 찾아가지 못한 적이 있는지 여부. 0=없음, 1=있음.
GenHlth (설명변수)	전반적인 건강 상태에 대한 자기 평가. 1=훌륭함, 2=매우 좋음, 3=좋음, 4=적당함, 5=좋지 않음.
MentHlth (설명변수)	지난 30일 동안 정신 건강이 좋지 않았던 날의 수. 스트레스, 우울증, 감정 문제 포함.
PhysHlth (설명변수)	지난 30일 동안 신체 건강이 좋지 않았던 날의 수. 신체 질병 및 부상 포함.
DiffWalk (설명변수)	걷거나 계단을 오르는데 심각한 어려움을 겪는지 여부. 0=없음, 1=있음.
Sex (설명변수)	응답자의 성별. 0=여성, 1=남성.
Age (설명변수)	14단계 연령 카테고리. 1=18~24세, 2=25~29세, 3=30~34세, 4=35~39세, 5=40~44세, 6=45~49세, 7=50~54세, 8=55~59세, 9=60~64세, 10=65~69세, 11=70~74세, 12=75~79세, 13=80세 이상

Education (설명변수)	**응답자가 완료한 최고 학년 또는 학교 연도. 1=학교를 간 적이 없거나 유치원만 다님, 2=1~8학년, 3=9~11학년, 4=고등학교 졸업, 5=대학 1~3학년, 6=대학졸업, 9=거부**
Income (설명변수)	**연간 가구 소득. 응답자가 특정 소득 수준에서 거부할 경우 "거부"로 기록. 1=$10,000 미만, 2=$15,000 미만, 3=$20,000 미만, 4=$25,000 미만, 5=$35,000 미만, 6=$50,000 미만, 7=$75,000 미만, 8=$75,000 이상**

다음의 그래프에서 확인할 수 있듯이, Diabetes_012(종속변수)에 클래스 불균형이 존재한다.



[그림 1] 당뇨병 여부 비율

현재 대부분의 데이터가 0 값을 가지며 (84.2%), 소수의 샘플만이 1 (1.8%) 또는 2 (13.9%) 값을 가지고 있다. 이를 통해 대부분 당뇨병이 없는 상태임을 알 수 있으며, 소수만이 경증 또는 중증의 당뇨병 상태임을 알 수 있다. 따라서 성능 지표로 Accuracy를 사용할 경우, 비율이 적은 1이나 2 클래스를 잘 맞추지 못하더라도, 즉 모든 클래스를 0으로 분류하더라도 높은 성능을 보일 수 있다. 본 프로젝트에서는 0번 클래스의 비율인 84.2%를 목표 성능으로 설정하고, 데이터셋 불균형을 고려할 수 있는 지표와 샘플링 방법을 통해 모든 클래스를 0으로 분류해서 84.2%의 성능을 보이는 모델이 아니라, 모든 클래스를 골고루 맞혀서 최소 84.2%의 성능을 보이는 모델을 만들고자 한다.

그렇기에 이러한 클래스 불균형을 고려할 수 있는 F1-Measure를 성능 지표로 사용할 것이다. 현 데이터셋은 보다 중요한 범주가 존재하므로, 균형정확도(Balance Correction Rate)보다는 F1-Measure를 주요 성능 지표로 활용할 것이다.

또한, 클래스 불균형이 존재하므로, 이후 Stratified 방식으로 추출된 데이터셋과 Resampling 기법을 적용한 데이터셋을 모델 학습에 사용할 것이다.

2-2. 데이터셋 시각화

현재 데이터셋을 확인해보면, 범주형 변수들이 숫자로 인코딩된 것을 확인할 수 있다. 따라서 1-

of-C coding은 진행하지 않는다. 이진 변환된 변수들은 범주형 변수로, 그 외의 BMI, GenHlth, MentHlth, PhysHlth, Age, Education, Income은 수치형 변수로 취급하였다.

수치형 변수의 주요 통계량은 다음과 같다.

	BMI	GenHlth	MentHlth	PhysHlth	Age	Education	Income
count	253680.000000	253680.000000	253680.000000	253680.000000	253680.000000	253680.000000	253680.000000
mean	28.382364	2.511392	3.184772	4.242081	8.032119	5.050434	6.053875
std	6.608694	1.068477	7.412847	8.717951	3.054220	0.985774	2.071148
min	12.000000	1.000000	0.000000	0.000000	1.000000	1.000000	1.000000
25%	24.000000	2.000000	0.000000	0.000000	6.000000	4.000000	5.000000
50%	27.000000	2.000000	0.000000	0.000000	8.000000	5.000000	7.000000
75%	31.000000	3.000000	2.000000	3.000000	10.000000	6.000000	8.000000
max	98.000000	5.000000	30.000000	30.000000	13.000000	6.000000	8.000000

[그림 2] 수치형 변수 주요 통계량

BMI 값은 평균이 약 28.38이며, 중앙값이 27인 것을 통해 오른쪽으로 꼬리가 긴 분포를 지닌 것을 알 수 있다. 이는 아래에 그려진 히스토그램을 통해서도 확인할 수 있다. Q1, Q3 값을 통해 max가이상치임도 알 수 있다. BMI가 98인 경우는 현실에 존재하기 어려울 것으로 보이나, 존재할 가능성을 배제할 수 없으므로 그대로 사용하기로 한다.

GenHlth를 통해 전반적으로 좋은 건강 상태를 지니고 있는 것을 알 수 있다.

MentHlth를 통해 지난 30일 동안 정신 건강이 좋지 않았던 날의 평균은 약 3일 정도지만, 절반 이상의 사람이 항상 정신 건강이 좋았다고 답한 것으로 보아, 이상치의 영향이 컸을 것으로 추정된다.

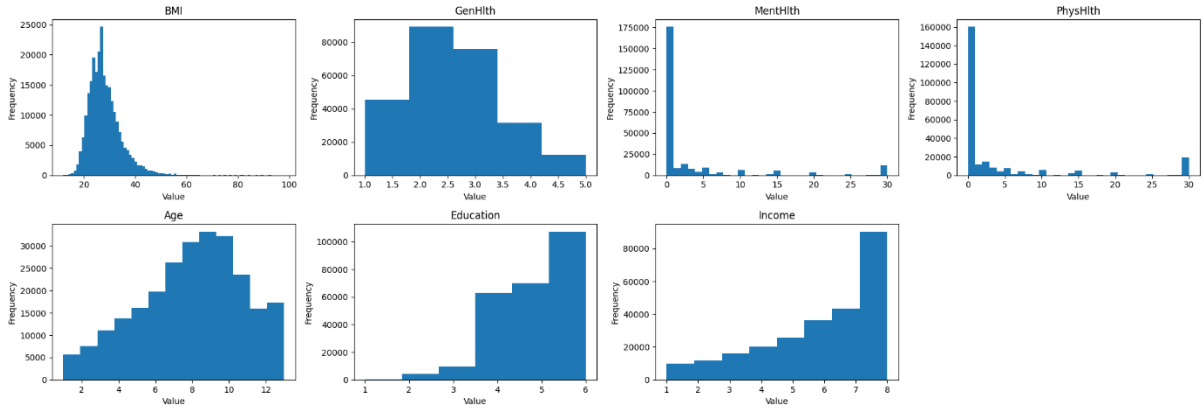
PhysHlth를 통해 지난 30일 동안 신체 건강이 좋지 않았던 날의 평균은 약 4일 정도이며, MentHlth와 마찬가지로 절반 이상의 사람이 항상 신체 건강이 좋았다고 답한 것으로 보아, 이상치의 영향이 컸을 것으로 보인다.

Age를 통해 평균적으로 나이가 8번 집단에 해당하며, Q1이 6인 것으로 보아 전반적으로 연령이 40대 이상으로 나타났다. 당뇨병과 관련된 데이터셋이다 보니, 중장년층 이상을 대상으로 조사한 것으로 보인다.

Education을 통해 전반적으로 고등학교 이상의 학력을 지닌 것을 알 수 있다. 이는 당뇨병과 직접적인 연관이 있기 보다는, 환경적인 요인과 관련이 있을 것으로 보인다.

Income은 평균이 6번 집단 정도이며, 중위값은 7번 집단에 해당하는 것을 보아, 왼쪽으로 꼬리가 긴 분포를 지닌 것을 알 수 있다. Income 역시 당뇨병과의 직접적인 연관보다는 환경적인 요인과 관련이 있을 것으로 보인다.

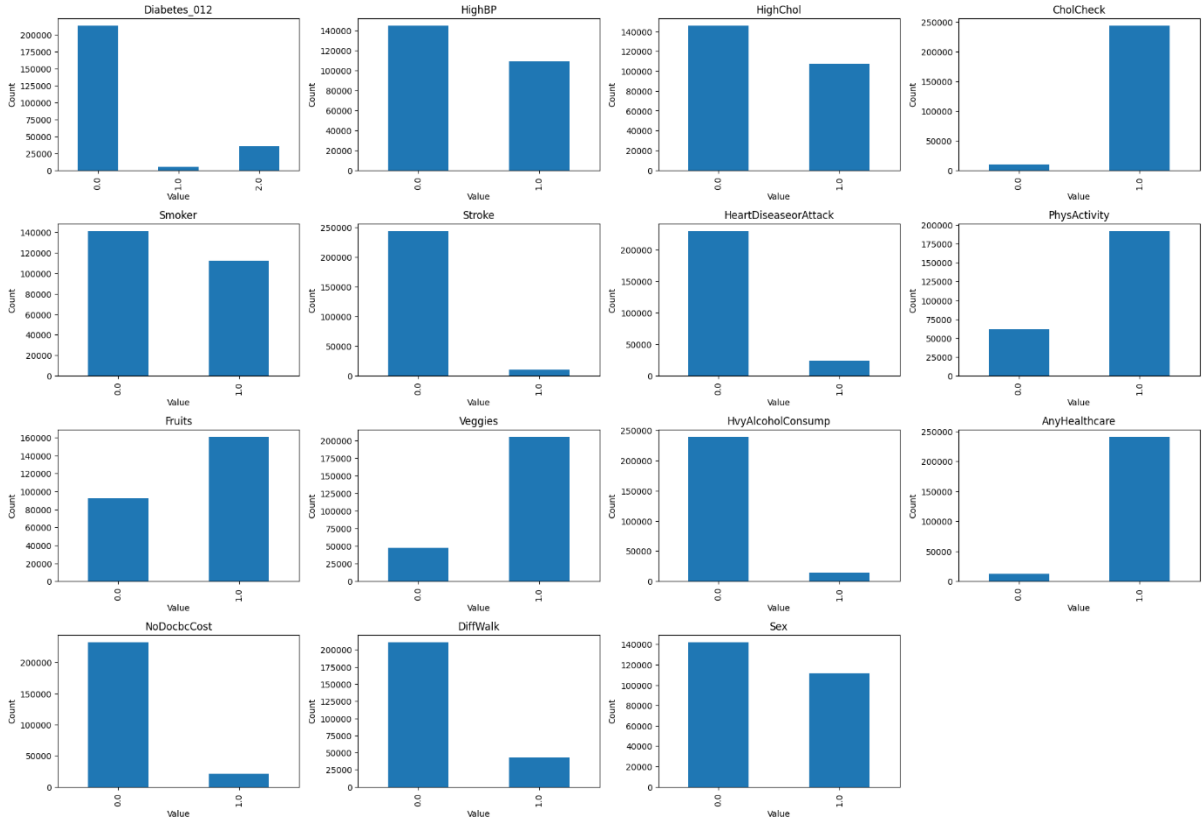
다음은 수치형 변수들을 시각화한 히스토그램이다.



[그림 3] 수치형 변수의 히스토그램 시각화

위에서 통계량을 통해 알아낸 사실을 다시 한번 눈으로 확인할 수 있으며, BMI가 정규분포의 형태를 띠고 있는 것 역시 확인할 수 있다.

다음은 범주형 변수를 막대 그래프로 시각화한 결과이다.

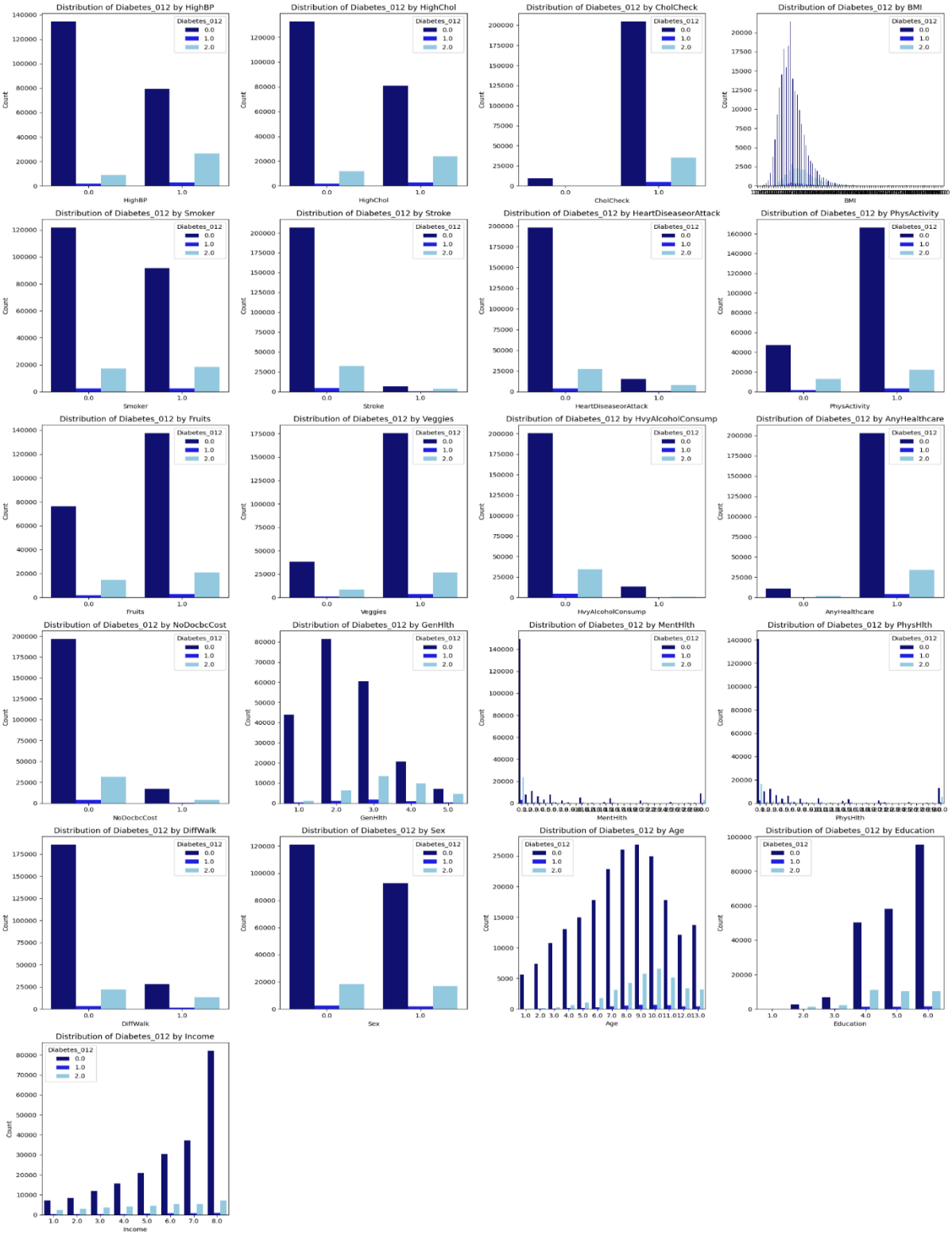


[그림 4] 범주형 변수의 막대 그래프 시각화

15개의 변수 중 약 10개 정도의 변수가 극심한 불균형을 보이고 있는 것을 확인할 수 있다. 고혈압을 진단받은 적 있는 사람은 그렇지 않은 사람과 비슷한 비율을 차지하고 있으며, 고콜레스테롤 역시 마찬가지이다. 거의 모두가 지난 5년 이내에 콜레스테롤 검사를 받았으며, 흡연자와 비흡연자는 비슷한 비율로 존재한다. 뇌졸중, 심장병 등은 진단받은 적 없는 사람이 훨씬 많았으며, 지난 30일 동안신체 활동이나 운동을 한 사람이 그렇지 않은 사람에 비해 약 3배 정도 많았다.

과일과 채소는 하루 한번 이상 섭취하는 사람이 많았으며, 과음은 하지 않은 사람이 많았다. 건강관리 보장은 있는 사람이 많았으며, 비용 때문에 병원을 찾아가지 못한 경우는 많지 않았다. 걷거나 계단을 오르기 불편하지 않은 경우가 많았고, 성별은 남녀가 비슷한 비율을 차지했다.

당뇨병 발병 여부와 각 설명변수가 어떤 관계를 가지고 있는지 확인하기 위해 당뇨병 발병 여부와 각 설명변수에 대한 이변량 그래프를 그려보았다.



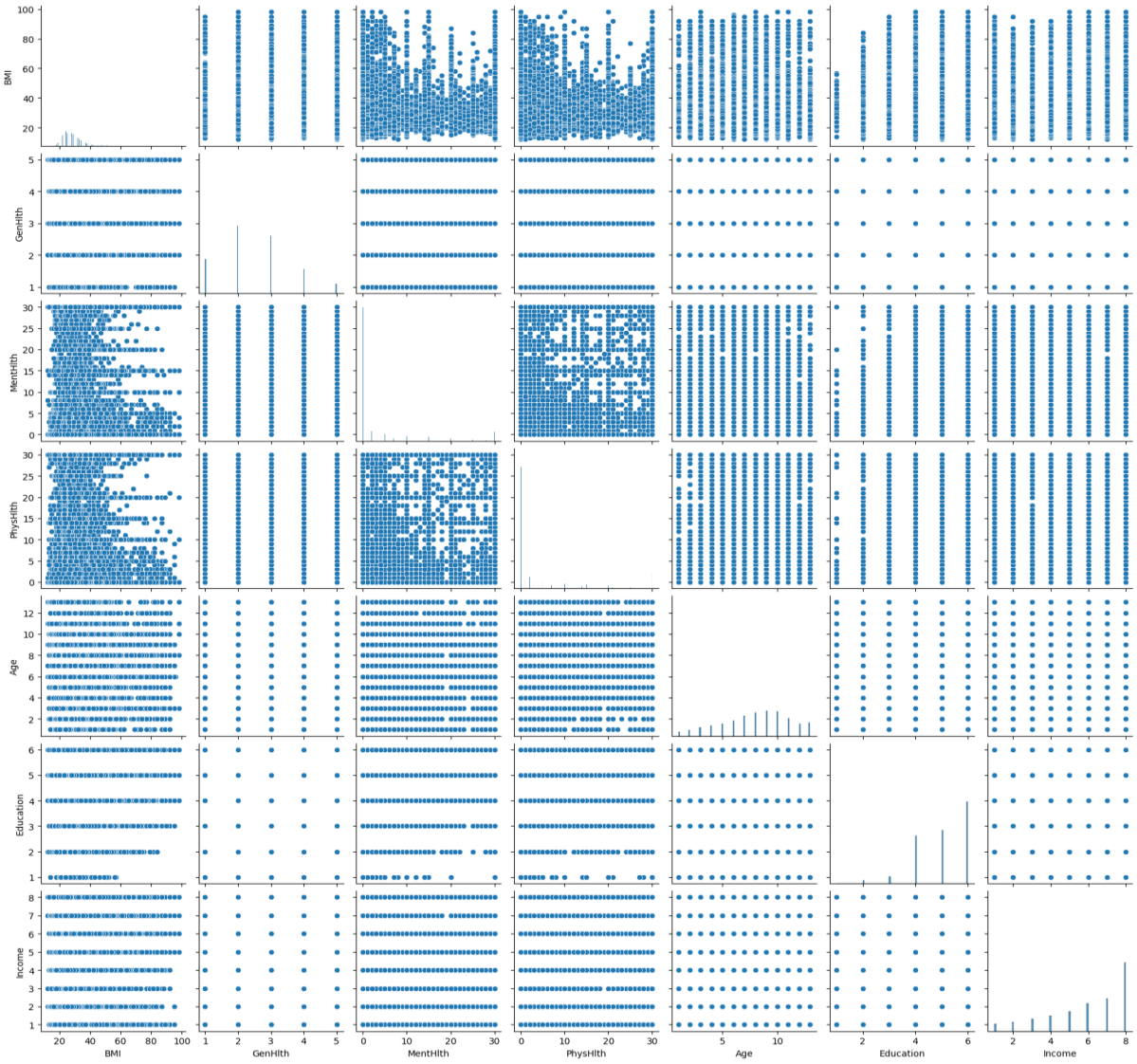
[그림 5] 당뇨병 유병 여부와 설명변수의 관계 시각화

고혈압이 있는 사람들 중 당뇨병 환자의 비율이 높고, 고콜레스테롤일 경우에도 당뇨병 환자의 비율이 높은 것을 알 수 있다. 당뇨병 환자들은 모두 5년 이내에 콜레스테롤 검사를 받은 것으로 보이는데, 이는 당뇨병 환자들이 정기적으로 검진을 받기 때문이라고 해석하는 것이 올바른 해석

인 것으로 보인다. BMI가 높은 사람들 중 당뇨병 환자가 좀 더 많은 경향을 보이며, 흡연자 중 당뇨병 환자의 비율이 다소 높은 것을 볼 수 있다. 뇌졸중, 심장질환 등을 앓은 적 있는 사람들 중 당뇨병 환자의 비율이 높았으며, 신체활동을 하지 않는 사람들, 과일과 채소를 적게 섭취하는 사람들, 과음을 하는 사람들, 건강 관리 보장이 없는 사람들, 경제적 이유로 진료를 못받은 사람들일 수록 당뇨병 환자가 많았다. 전반적인 건강상태가 나쁠수록, 정신건강상태가 나쁠수록 , 나이가 많아질수록, 교육수준이 낮을수록, 소득수준이 낮을수록 당뇨병 환자가 많았고, 걷기 어려움이 있는 사람들, 여성보다는 남성이 당뇨병 환자가 많았다.

고혈압, 고콜레스테롤 등의 성인병 질환은 나이가 많을수록 환자가 많아진다. 이전에 다른 질병이 있었다면 건강 관리에 소홀했을 가능성이 있고, 이는 건강 관리, 건강 상태 등의 변수와 관련이 있으며, 결국 당뇨병 환자와도 연관성이 있을 것이라고 추측할 수 있다. 소득수준이 낮을수록 건강에 신경을 쓰기 어려웠을 가능성이 있으며, 그로 인해 여러 질병에 걸리게 됐을 가능성이 있다. 이처럼 설명변수들 간에도 관련성이 있으며, 단순히 분포에서 보이는 것보다 그 안에 내재되어 있는 관계를 찾아낼 필요가 있다.

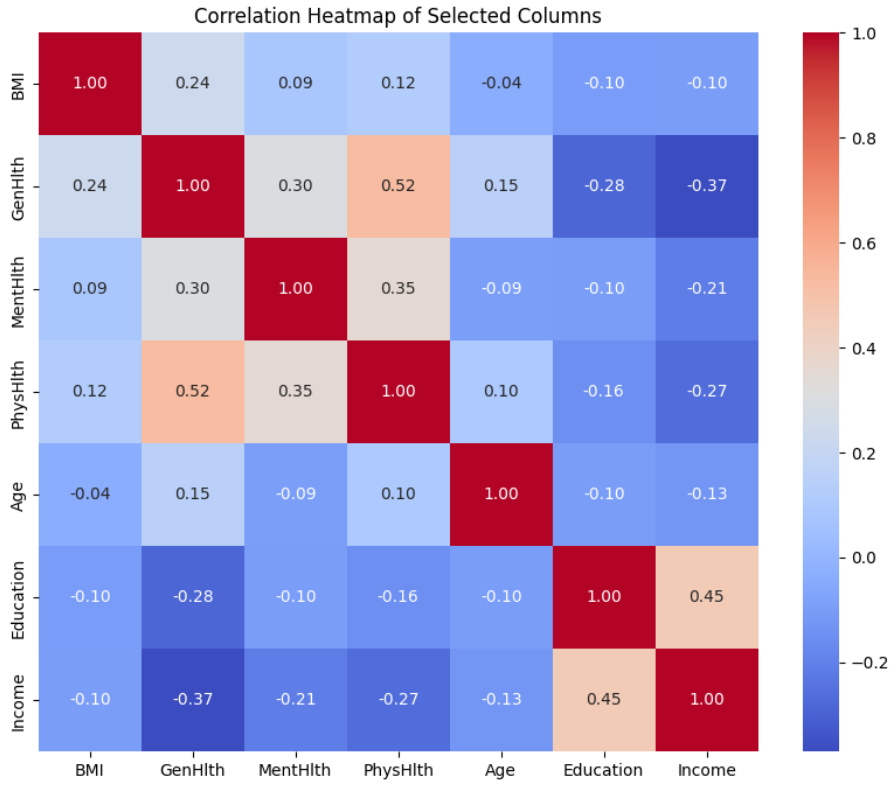
수치형 변수에 대해 산점도를 그려본 결과는 다음과 같다.



[그림 6] 수치형 변수의 산점도

값들이 정수로 이루어져 있어, 점들이 수직 혹은 수평으로 분포해있는 것을 확인할 수 있다. 따라서 산점도를 통해 상관관계를 알아보기에는 다소 무리가 있을 것으로 보인다.

상관관계를 히트맵으로 시각화한 결과는 다음과 같다.



[그림 7] 수치형 변수의 상관관계 히트맵

가장 강한 상관관계는 0.52로, 한 변수를 제거해야 할 만큼 높은 상관관계는 나타나지 않았다. 따라서 특정 변수를 제거하는 대신, 모든 변수를 사용할 것이다.

2-3. 데이터셋 생성

먼저 수치형 변수들 간의 범위가 다르므로 Scaling을 진행할 것이다. BMI 등의 설명변수에서 이상치가 존재하는 것을 확인할 수 있다. 그러나 이는 충분히 발생할 수 있는 이상치라고 판단했기에 이상치에 강건한 scaling 방식인 RobustScaler를 사용하였다.

먼저 수치형 변수들에 대해서 scaling을 진행할 것이다. 앞서 BMI 등의 설명변수에서 이상치가 존재하는 것을 확인하였다. 그러나 이는 충분히 발생할 수 있는 이상치라고 판단했기에 이상치에 강건한 scaling 방식인 RobustScaler를 사용하였다.

본 프로젝트에서는 세가지의 데이터셋을 만들어 비교할 예정이다. 종속변수의 비율에 차이가 있으므로 훈련 데이터셋 분할 시 종속변수의 비율에 맞추어 분할한 Stratified dataset, 무작위로 가장 적은 클래스의 데이터를 가장 많은 클래스만큼 복제하는 Oversampling dataset, 다수 클래스를 샘플링하고 기존 소수 샘플을 보간하여 새로운 소수 인스턴스를 합성해내는 SMOTE dataset을 만들 것이며, 데이터셋의 크기가 커 학습에 시간이 너무 오래 걸리므로 학습/검증/테스트 데이터셋은 8,000/2,000/2,000개로 설정하였다. 반복실험은 5회씩 진행하였으며, 데이터셋 성능 비교를 위해 학습 데이터로 학습 후, 검증 데이터셋으로 성능을 측정하였다. 5회씩 반복실험을 진행했기에

모든 Confusion Matrix를 본문에 나타내기는 어려움이 있어, 첨부하는 .ipynb 파일을 통해 실험 결과를 확인할 필요가 있다.

3. 모델 학습

3-1. Multinomial Logistic Regression

Multinomial Logistic Regression의 경우, 보다 정확한 분류를 위해 학습해야 하는 하이퍼파라미터가 특별히 존재하지 않는다고 생각하여, 따로 하이퍼파라미터 조정 과정을 거치지 않았다.

[표 2] Multinomial Logistic Regression의 성능 평가 결과

	F1-Measure 평균	F1-Measure 표준편차	F1-Measure 신뢰구간
Stratified	**0.3910**	**0.0139**	**[0.3718, 04103]**
Oversampled	**0.5100**	**0.0131**	**[0.4918, 0.5282]**
SMOTE	**0.5224**	**0.0099**	**[0.5086, 0.5362]**

F1-Measure의 신뢰구간을 확인해보았을 때, Stratified dataset보다는 샘플링된 데이터셋이 더 높은 성능을 보여주는 것을 알 수 있다. SMOTE dataset의 평균이 Oversampling dataset의 평균보다 높지만, 신뢰구간에서 겹치는 부분이 있어 어느 한쪽이 더 우수하다고 이야기하기 어렵다.

다음은 각 데이터셋의 Confusion Matrix의 해석에 관한 내용이다.

가장 먼저 Stratified dataset의 경우를 보면, 반복실험을 진행하는 동안 모두 1번 클래스로 예측한 경우가 존재하지 않았다. 이는 1번 클래스에 속하는 데이터가 가장 적었기 때문으로 예상되며, 2번 클래스를 틀리는 경우가 많고, 1번 클래스의 영향으로 F1-Measure가 높지 않았던 것으로 보인다.

Oversampling dataset을 보면, 각 클래스에 속하는 데이터의 비율이 균등해져서, 1번 클래스로 예측한 데이터가 훨씬 많아진 것을 확인할 수 있다. 0번 클래스를 맞히는 비율이 줄어들었으나, 1번 클래스를 맞히는 경우가 생겼고, 2번 클래스를 맞추는 비율은 비슷하기 때문에 Stratified dataset보다 F1-Measure의 값이 높아질 것임을 예측할 수 있다.

SMOTE dataset을 보면, Oversampling dataset과 비슷한 정도로 클래스를 맞히는 것을 볼 수 있다. 따라서 SMOTE dataset과 Oversampling dataset의 성능은 비슷할 것으로 예측할 수 있다.

Stratified dataset은 소수 클래스를 학습하기 어려웠기 때문에 다른 두 데이터셋에 비해 낮은 성능을 보인 것이 납득 가능하다.

3-2. Classificatigression Tree (CART)

다음은 Classification and Regression Tree를 학습하기 위해 사용한 하이퍼파라미터와 그 후보에 관한 설명이다.

[표 3] (CART) 사용한 하이퍼파라미터 및 후보값

종류	범위	설명
criterion	["gini", "entropy"]	분기 시 불순도를 측정하는 함수. “gini”는 Gini impurity를, “entropy”는 Information gain을 나타낸다.
max_depth	[10, 20, None]	Tree의 최대 깊이. None으로 설정할 경우 모든 leaf node가 pure하거나 모든 leaf node가 min_samples_split sample보다 적을 때까지 팽창한다.
min_samples_leaf	[5, 10, 20]	leaf node에 요구되는 최소 샘플 수.

Stratified dataset의 훈련 데이터셋을 기준으로 Full Tree를 만들어보았을 때, depth가 26이었으므로, 과적합을 방지하기 위해 max_depth의 후보값을 10과 20으로 선정하였다. 만약의 경우에 대비해 None 역시 추가하였다.

각 데이터셋별 최적의 하이퍼파라미터 조합은 다음과 같았다.

[표 4] (CART) 최적의 하이퍼파라미터 조합

Dataset	Stratified	Oversampled	SMOTE
criterion	"entropy"	"gini"	"gini"
max_depth	20	10	20
min_samples_leaf	5	5	20

위에서 구한 최적의 하이퍼파라미터 조합으로 각 데이터셋을 학습한 결과는 다음과 같다.

[표 5] CART의 성능 평가 결과

	F1-Measure 평균	F1-Measure 표준편차	F1-Measure 신뢰구간
Stratified	0.4003	0.0060	[0.3919, 0.4087]
Oversampled	0.5054	0.0077	[0.4947, 0.5161]
SMOTE	0.5802	0.0071	[0.5703, 0.5900]

앞서 보았던 Multinomial Logistic Regression과 동일하게 Stratified dataset의 성능이 가장 낮았으며, 이번에는 SMOTE가 Oversamling보다 확실하게 더 좋은 성능을 보여주었음을 알 수 있다.

Multinomial Logistic Regression보다 CART의 성능이 더 뛰어날 것으로 기대했지만, 수직으로만 분류 경계면을 만들 수 있는 Decision Tree의 한계 때문인지 생각보다 성능에 큰 차이가 나지 않았다.

다음은 각 데이터셋의 Confusion Matrix의 해석에 관한 내용이다.

먼저 Stratified dataset을 보면, 1번 클래스로 예측한 경우가 있으나, 정답을 맞힌 경우는 없는 것으로 확인된다. 2번 클래스 역시 정답보다 오답의 비율이 더 높아, F1-Measure 값이 높지 않을 것이라고 예상할 수 있다.

Oversampling dataset을 보면, 모든 클래스의 비율이 비슷해진 것을 확인할 수 있으며, 0번 클래스는 정답이, 1번 클래스는 오답이 조금 더 많고, 2번 클래스는 정답과 오답의 비율이 비슷한 것으로 보아 성능이 0.5 근방일 것이라 예상할 수 있다. 1번 클래스를 맞히는 경우가 많아졌기 때문에 Stratified dataset보다는 당연히 성능이 높을 것이라고도 예상 가능하다.

SMOTE dataset을 보면, 0번과 1번 클래스의 정답 비율이 더 늘어난 것을 확인할 수 있다. 따라서 SMOTE dataset의 성능이 가장 좋을 것이라고 기대할 수 있다.

3-3. Naïve Bayes Classifier

Naïve Bayes Classifier 역시 보다 정확한 분류를 위해 조정해야 하는 하이퍼파라미터가 특별히 존재하지 않는다고 생각하여, 따로 하이퍼파라미터 조정 과정을 거치지 않았다.

[표 6] Naive Bayes Classifier의 성능 평가 결과

	F1-Measure 평균	F1-Measure 표준편차	F1-Measure 신뢰구간
Stratified	**0.4226**	**0.0081**	**[0.4113, 0.4339]**
Oversampled	**0.4517**	**0.0201**	**[0.4237, 0.4797]**
SMOTE	**0.5121**	**0.0198**	**[0.4847, 0.5396]**

Stratified dataset과 Oversampling dataset의 신뢰구간에 중복되는 부분이 생겨 Oversampling dataset이 Stratified dataset보다 우수하다고 이야기하기 어렵다. 그러나, SMOTE dataset의 경우, 다른 두 데이터셋보다 우수함을 알 수 있다.

앞서 데이터셋 탐색에서 이야기했듯이, 각 설명변수는 독립이 아닐 가능성이 높다. 그러나 Naïve Bayes Classifier의 경우, 각 설명변수의 독립을 가정하므로, 성능이 높지 않을 것이라고 사전에 예상할 수 있다.

다음은 각 데이터셋의 Confusion Matrix의 해석에 관한 내용이다.

먼저 Stratified dataset을 보면, 1번 클래스로 예측하는 경우가 늘어났고, 1번 클래스를 맞힌 경우도 발생하였으나, 그 영향이 미미할 것으로 예상된다. 앞선 두 모델보다 2번 클래스로 예측한 경우가 더 많고, 더 많이 맞혔기 때문에 이전 모델들에 비해 F1-Measure 값이 상승했을 것이라 예상할 수 있다.

Oversampling dataset의 경우를 보면, 1번 클래스로 예측한 경우가 적고, 0번 클래스와 2번 클래스로 예측한 비율이 비슷함을 알 수 있다. 1번 클래스는 오답의 비율이 더 높고, 0번 클래스와 2번 클래스는 정답과 오답의 비율이 비등비등하기 때문에 성능이 0.5 이하일 것으로 추측할 수 있다.

SMOTE dataset의 경우를 보면, 1번 클래스로 예측한 경우가 증가했으며, 0번 클래스의 정답 비율이 높아진 것을 확인할 수 있다. 그러나 아직 만족할 만한 성능에 도달하지 못했기에, 다른 모델들을 추가적으로 학습해보려 한다.

3-4. Artificial Neural Network (ANN)

다음은 Artificial Neural Network를 학습하기 위해 사용한 하이퍼파라미터와 그 후보에 관한 설명이다. activation function은 'relu', learning_rate_init은 0.001로 설정한 후, Grid Search 방식을 통해

최적의 하이퍼파라미터 조합을 찾아보았다.

[표 7] (ANN) 사용한 하이퍼파라미터 및 후보값

종류	범위	설명
hidden_layer_sizes	[(50,), (100,), (50,50), (100,100)]	hidden layer 내에 있는 hidden neuron의 개수. i번째 요소는 i번째 hidden layer를 나타낸다.
max_iter	[150, 300]	최대로 반복 가능한 횟수를 의미한다. learning_rate_init과 함께 gradient의 수렴에 영향을 미친다.

hidden layer의 개수는 3개 이상 쓰더라도 성능 개선이 많이 이루어지지 않는다는 점을 고려하여 최대 layer의 수를 2개로 설정하였다. max_iter의 경우, learning_rate_init이 작기 때문에 gradient가 충분히 수렴할 수 있도록 최대 300으로 설정하였다.

각 데이터셋별 최적의 하이퍼파라미터 조합은 다음과 같았다.

[표 8] (ANN) 최적의 하이퍼파라미터 조합

Dataset	Stratified	Oversampled	SMOTE
hidden_layer_sizes	(50, 50)	(100, 100)	(100, 100)
max_iter	300	300	150

다음은 각 데이터셋에 대해 성능을 평가한 결과이다.

[표 9] ANN의 성능 평가 결과

	F1-Measure 평균	F1-Measure 표준편차	F1-Measure 신뢰구간
Stratified	0.4223	0.0101	[0.4083, 0.4364]
Oversampled	0.5505	0.0085	[0.5388, 0.5623]
SMOTE	0.5885	0.0129	[0.5706, 0.6063]

다른 모델에 비해 Stratified dataset의 평균이 조금 더 높고, Oversampling data와 SMOTE dataset의 평균 차이가 작은 것이 눈에 띈다. ANN에서도 SMOTE가 가장 좋은 성능을 보여준 데이터셋이다.

다음은 각 데이터셋의 Confusion Matrix의 해석에 관한 내용이다.

Stratified dataset을 보면, 1번 클래스로 예측하고, 맞은 경우가 존재하는 것이 눈에 띈다. 다만 그 정답률이 높지는 않아 큰 영향을 미치지는 못했지만, 다른 모델들보다 좋은 성능을 보여줄 수 있었던 이유로 보인다.

Oversampling dataset을 보면, 대체적으로 0번 클래스보다 1번 클래스를 맞힌 경우가 더 많은 것이 눈에 띈다. 이 역시 다른 모델들에 비해 높은 F1-Measure 평균이 나오게 된 이유로 보인다.

SMOTE dataset을 보면 Oversampling dataset에 비해 1번 클래스를 맞히는 경우가 살짝 줄어든 것을 볼 수 있다. 그러나 전체적인 정답 수가 증가해 Oversampling dataset보다 높은 성능을 보여줄 수 있었던 것으로 보인다.

3-5. CART Bagging

3-2. CART에서 선택된 최적의 하이퍼파라미터 설정을 유지하며 base learner의 수만 30, 50, 100으로 증가시켜 보았다. Stratified dataset과 SMOTE dataset은 100, Oversampling dataset은 50이 최적인 것으로 드러났다. 각각의 데이터셋에 대해 최적의 하이퍼파라미터로 세팅을 한 후 얻은 성능 평가 결과이다.

[표 10] CART Bagging의 성능 평가 결과

	F1-Measure 평균	F1-Measure 표준편차	F1-Measure 신뢰구간
Stratified	**0.4007**	**0.0093**	**[0.3878, 0.4135]**
Oversampled	**0.5561**	**0.0106**	**[0.5414, 0.5708]**
SMOTE	**0.6456**	**0.0048**	**[0.6389, 0.6523]**

이전 결과들과 동일하게 Stratified dataset < Oversampling dataset < SMOTE dataset 순으로 성능이 좋은 것을 확인할 수 있다. 이전에 비해 각 데이터셋 간의 성능 차이가 더 커진 것을 볼 수 있다. 이전 단일 CART에 비해 성능의 최고점이 올라간 것을 확인할 수 있다. 이는 여러 base learner가 학습한 뒤, 그 결과들을 가지고 다시 예측을 하는 Bagging의 특성으로 인한 것으로 예상된다.

다음은 각 데이터셋의 Confusion Matrix의 해석에 관한 내용이다.

Stratified dataset의 경우, 다시 1번 클래스로 예측하는 경우가 사라진 것을 확인할 수 있다. 이로 인해 성능이 이전 Multinomial Logistic Regression이나 CART 모델과 비슷할 것이라 예측할 수 있다.

Oversampling dataset의 경우, 0번, 1번, 2번 클래스를 모두 골고루 맞히고 있는 것을 확인할 수 있다. 특히 0번 클래스를 잘 맞히는 것을 확인할 수 있다.

SMOTE dataset의 경우, 0번 클래스의 정답 비율이 상당히 높으며, 1번과 2번 클래스 역시 준수하게 맞히는 것을 볼 수 있다.

3-6. Random Forest

3-2. CART에서 선택된 최적의 하이퍼파라미터 설정을 유지하며 base learner의 수만 30, 50, 100으로 증가시켜 보았다. Stratified dataset은 30, 나머지 두 데이터셋은 100개의 base learner를 사용하는 것이 가장 성능이 좋았다. Base learner의 수가 많아진다고 해서 무조건 성능이 높아지는 것은 아니라는 사실을 알 수 있다.

[표 11] Random Forest의 성능 평가 결과

	F1-Measure 평균	F1-Measure 표준편차	F1-Measure 신뢰구간
Stratified	**0.3716**	**0.0104**	**[0.3571, 0.3860]**
Oversampled	**0.5537**	**0.0109**	**[0.5385, 0.5689]**
SMOTE	**0.6465**	**0.0100**	**[0.6326, 0.6605]**

CART Bagging과 성능 차이가 크지 않은 것을 확인할 수 있다. 이는 두 모델이 같은 base learner를 사용했기 때문에 크게 차이가 벌어지지 못한 것으로 보인다. 다만, Random Forest는 몇몇 설명변수를 일부러 제외하기 때문에, CART Bagging보다 조금 더 유연한 예측이 가능할 수 있을 것으로 보인다.

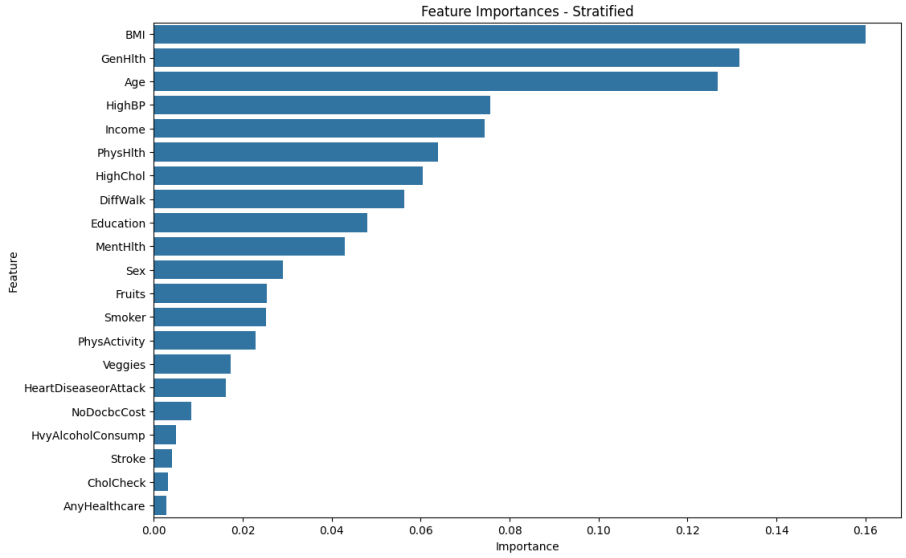
다음은 각 데이터셋의 Confusion Matrix의 해석에 관한 내용이다.

먼저 Stratified dataset을 보면, 1번 클래스로 예측한 경우는 존재하지 않는다. 이전에 보았던 다른 모델들의 Stratified dataset과 비슷한 Confusion Matrix를 지니고 있어, 성능 역시 비슷할 것으로 보인다.

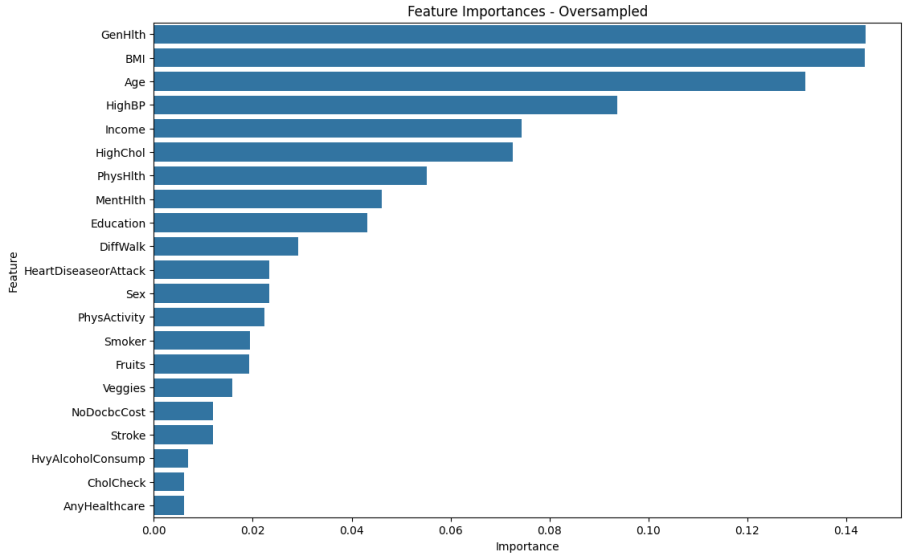
Oversampling dataset을 보면, 각 클래스에 대해 정답과 오답 비율이 비슷하거나, 정답 비율이 더 높은 것을 알 수 있다. 따라서 성능이 0.5를 넘길 것으로 보인다.

SMOTE dataset을 보면, Oversampling dataset에 비해 각 클래스에 대한 정답 비율이 높아진 것을 확인할 수 있다. 이는 SMOTE dataset의 데이터 증강 방식이 더욱 효과적이었다는 것을 의미한다.

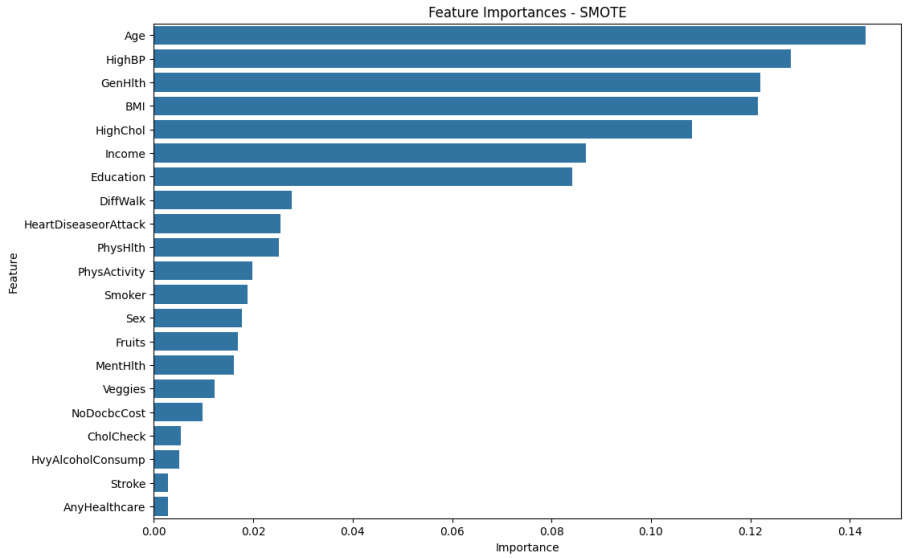
추가적으로, 각 데이터셋에 대해 최적의 하이퍼파라미터를 찾는 과정에서 얻어진 best model의 변수 중요도는 다음과 같다



[그림 8] (Random Forest) Stratified Dataset의 변수 중요도



[그림 9] (Random Forest) Oversampling Dataset의 변수 중요도



[그림 10] (Random Forest) SMOTE Dataset의 변수 중요도

데이터셋별로 중요도 순위는 다르지만, 대체적으로 순위권에 있는 변수들은 동일한 것으로 보인다. BMI, Age, GenHlth, HighBP 등이 그 예시이며, AnyHealthcare, CholCheck 등의 변수는 공통적으로 중요도가 낮은 것으로 나타났다. 상위권에 있는 변수만을 이용하여 모델을 학습하면 데이터셋의 복잡도가 줄어 일반화 능력이 높아지는 결과가 나타날 수 있을 것으로 보인다.

3-7. Adaptive Boosting (AdaBoost)

3-2. CART에서 선택된 최적의 하이퍼파라미터 설정을 유지하며 base learner의 수만 30, 50, 100으로 증가시켜 보았다. Stratified dataset은 30, 나머지 두 데이터셋은 100개의 base learner를 사용하는 것이 가장 성능이 좋았다.

[표 12] AdaBoost의 성능 평가 결과

	F1-Measure 평균	F1-Measure 표준편차	F1-Measure 신뢰구간
Stratified	**0.4094**	**0.0069**	**[0.3998, 0.4190]**
Oversampled	**0.5130**	**0.0098**	**[0.4993, 0.5266]**
SMOTE	**0.6721**	**0.0086**	**[0.6602, 0.6841]**

다른 모델들과 동일하게 Stratified dataset < Oversampling dataset < SMOTE dataset 순으로 높은 성능을 보였다. 모두 매우 작은 표준편차를 지닌 것도 눈에 띈다.

다음은 각 데이터셋의 Confusion Matrix의 해석에 관한 내용이다.

Stratified dataset의 경우 다른 모델들과 마찬가지로 1번 클래스로 예측한 경우가 거의 없다. 0번 클래스는 정답 비율이 높고, 2번 클래스는 0.5 정도에서 변동이 있다. 특정 클래스에 대한 정답 비율이 매우 낮기 때문에 F1-Measure 값이 높게 나올 수 없었다.

Oversampling dataset의 경우, 특출나게 특정 클래스의 정답 비율이 높다던가, 오답 비율이 높은 현상은 발생하지 않았다. 따라서 F1-Measure의 값이 0.5 근방에 있을 것으로 예상된다.

SMOTE dataset의 경우, 0번 클래스의 정답률이 높은 것을 확인할 수 있다.

3-8. Gradient Boosting Machine (GBM)

3-2. CART에서 선택된 최적의 하이퍼파라미터 설정을 유지하며 base learner의 수만 30, 50, 100으로 증가시켜 보았다. Stratified dataset과 Oversampling dataset은 50, SMOTE dataset은 100개의 base learner를 사용하는 것이 가장 성능이 좋았다.

[표 13] GBM의 성능 평가 결과

	F1-Measure 평균	F1-Measure 표준편차	F1-Measure 신뢰구간
Stratified	**0.4059**	**0.0073**	**[0.3958, 0.4159]**
Oversampled	**0.5878**	**0.0121**	**[0.5710, 0.6046]**
SMOTE	**0.7528**	**0.0044**	**[0.7466, 0.7590]**

SMOTE dataset의 성능이 눈에 띄게 개선된 것을 확인할 수 있다. 본 데이터셋에 대해 Gradient Boosting Machine이 가장 효과적이며, SMOTE dataset을 통해 데이터를 증강했을 때 제 성능을 온전히 발휘할 수 있었던 것으로 보인다. 데이터셋들 간 F1-Measure의 평균 차이도 눈에 띈다.

다음은 각 데이터셋의 Confusion Matrix의 해석에 관한 내용이다.

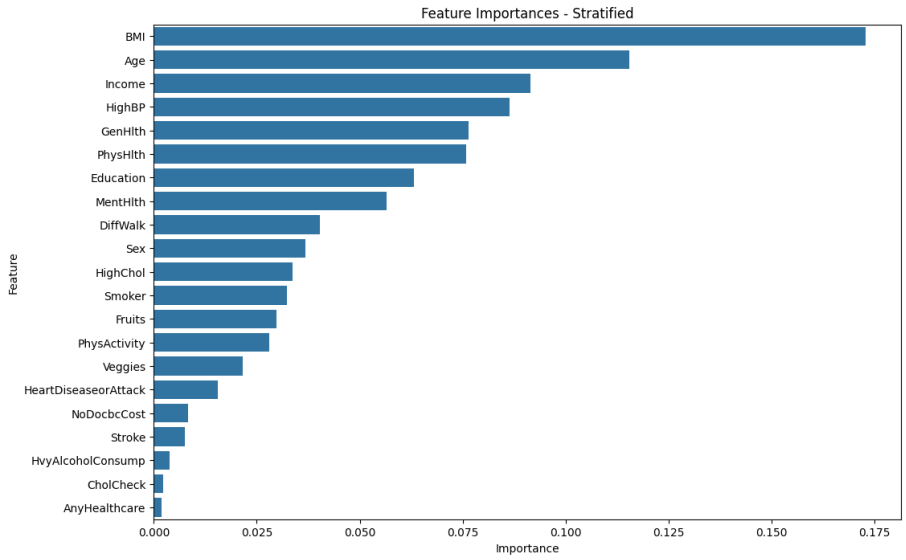
Stratified dataset의 경우 1번 클래스로 예측한 경우는 존재하나, 전부 정답을 맞히지 못했다. 2번 클래스도 오답 비율이 높아, F1-Measure 값이 높지 않을 것임을 예상할 수 있다.

Oversampling dataset의 경우, 각 클래스별 정답 개수가 비슷해진 것이 눈에 띈다. 각 클래스의 정답 비율이 50%를 넘어가는 것으로 보아, F1-Measure 역시 0.5를 넘을 것으로 예상할 수 있다.

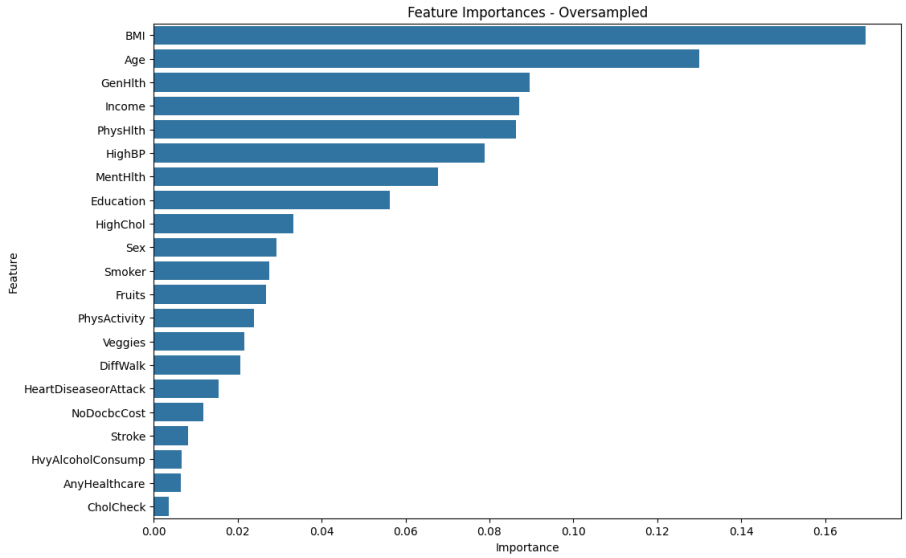
SMOTE dataset의 경우, 모든 모델과 비교하여도 가장 높은 수준의 1번 클래스 정답률을 보여주

며, 0번 클래스 역시 매우 높은 정답률을 보여준다. 2번 클래스를 1번 클래스로 잘못 예측하는 경우와 2번 클래스를 1번 클래스로 잘못 예측하는 경우는 여전히 꽤 많지만, 이는 훈련 데이터셋의 크기를 늘림으로써 해결할 수 있을 것으로 보인다.

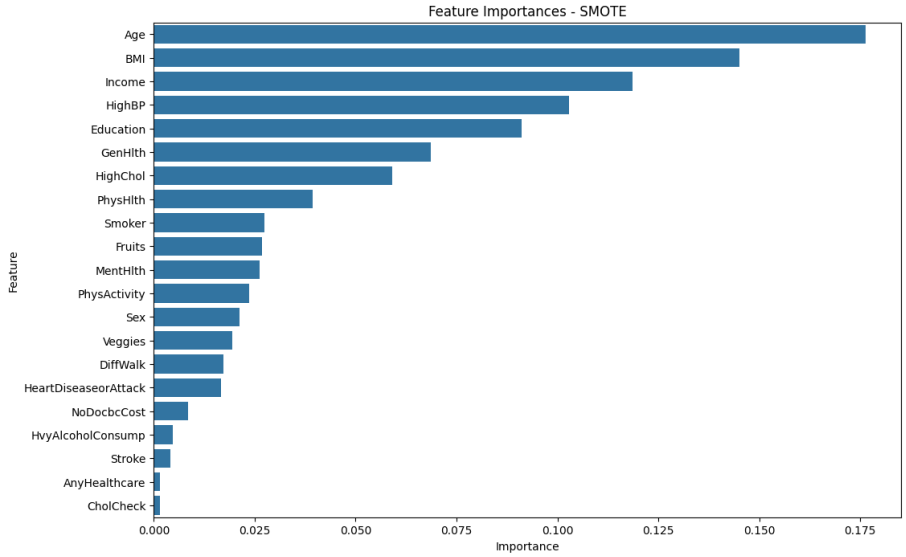
추가적으로, 각 데이터셋에 대해 최적의 하이퍼파라미터를 찾는 과정에서 얻어진 best model의 변수 중요도는 다음과 같다.



[그림 11] (GBM) Stratified Dataset의 변수 중요도



[그림 12] (GBM) Oversampling Dataset의 변수 중요도



[그림 13] (GBM) SMOTE Dataset의 변수 중요도

BMI, Age, Income 등의 변수가 모든 데이터셋에서 상위권에 위치해있으며, 이는 Random Forest와도 동일한 결과이다. 모든 데이터셋에서 CholCheck, AnyHealthcare 등의 변수가 하위권에 위치한 것 역시 Random Forest와 동일하다. 따라서 다른 모델들에서도 위와 같은 변수들이 중요했을 것으로 추측된다.

4. 최종 모델 선정

거의 모든 경우에 SMOTE dataset이 가장 우수한 성능을 보여준 것을 알 수 있다. 이는 소수 클래스를 증강시켜 더 학습이 잘 될 수 있도록 했기 때문이며, Oversampling dataset에 비해 더 높은 성능을 보인 이유는 단순히 데이터셋을 복제하는 것이 아니라, 보간을 통해 데이터를 합성했기 때문으로 보인다.

다음은 8가지 모델 중 가장 높은 성능을 보인 GBM – SMOTE 모델을 가지고 SMOTE의 테스트 데이터셋을 학습한 결과이다.

F1-Measure는 0.7222의 값을 가지므로, 검증 데이터셋에 대한 성능과 테스트 데이터셋에 대한 성능이 크게 차이나지 않아, 모델이 과적합되지 않았음을 알 수 있다.

테스트 데이터셋에 대한 Confusion Matrix는 다음과 같다.

[표 14] 테스트 데이터셋에 대한 GBM - SMOTE의 Confusion Matrix

Confusion Matrix	0	1	2
0	596	1	70
1	29	468	170
2	101	178	387

0번 클래스와 1번 클래스에 대한 정답 비율이 매우 높은 것을 확인할 수 있다. 2번 클래스를 더욱 잘 맞힐 수 있도록 모델을 학습시키면 정확도가 개선될 수 있을 것으로 보인다.

5. 결론

거의 모든 경우에 SMOTE dataset이 가장 우수한 성능을 보여준 것을 알 수 있다. 이는 소수 클래스를 증강시켜 더 학습이 잘 될 수 있도록 했기 때문이며, Oversampling dataset에 비해 더 높은 성능을 보인 이유는 단순히 데이터셋을 복제하는 것이 아니라, 보간을 통해 데이터를 합성했기 때문으로 보인다.

Confusion Matrix를 보았을 때, 대부분 Stratified dataset에서는 1번 클래스를 예측한 경우가 거의 없다시피 했고, 0번 클래스를 맞히는 비율이 매우 높았다. SMOTE dataset에서도 다른 클래스에 비해 0번 클래스를 더 잘 맞히는 경향이 나타났다. 이는 1번과 2번 클래스를 증강했음에도 불구하고, 정확히 구별해내기엔 훈련 데이터셋의 크기가 작았기 때문일 수 있다.

컴퓨팅 파워의 부족으로 학습 데이터를 8천개로 줄였지만, 그럼에도 불구하고 테스트 데이터셋에 대해 72% 정도의 성능이 나왔기에 최종모델을 충분히 예측 모델로 사용할 만하다고 생각한다. 다만, 실제로 사용할 때에는 나머지 데이터도 모두 학습에 사용하여 조금 더 성능을 끌어올리는 것이 좋아 보인다. 하이퍼파라미터 조합 역시 제한적으로 실험을 진행하였기 때문에, 더욱 다양한 하이퍼파라미터와 더 많은 후보로 실험한다면 더 좋은 하이퍼파라미터 조합을 찾을 수 있을 것으로 보인다.

또한, 너무 오랜 시간이 소요되어 실험하지 못했던 ANN Bagging 등 추가적으로 다른 모델과 비교해봄으로써 더 나은 모델을 찾을 수 있을 것으로 보인다.

변수 중요도를 통해 BMI, Age, Income 등의 변수가 당뇨병 여부를 예측하는데 중요하고, CholCheck, AnyHealthcare 등의 변수는 중요하지 않음을 알 수 있었다. Age는 건강 상태와, Income은 생활환경 등과 연관이 있어 많은 변수들을 아우르는 상위 개념의 변수이므로, 그 중요성이 높게 측정될 수밖에 없었을 것으로 보인다. 따라서 결과를 해석할 때에는 단순히 나이가 많을수록, 소득이 적을수록 당뇨에 걸릴 확률이 높다고 해석하기 보다는, 나이가 들어감에 따라 건강에 조금 더 신경을 쓰고, 좋은 식습관, 생활습관 등을 유지할 수 있도록 노력할 필요가 있다고 해석하는 것이 더 적절해보인다.