김 동 하 교수 지도

석사학위 청구논문

TCIWAE: 조건부 IWAE를 활용한 재현 데이터 생성 방법론에 대한 연구

-표 형태 데이터의 재현을 중심으로-

2023

성신여자대학교 일반대학원

통 계 학 과

김 지 우

TCIWAE: 조건부 IWAE를 활용한 재현 데이터 생성 방법론에 대한 연구

-표 형태 데이터의 재현을 중심으로-

김 동 하 교수 지도

이 논문을 석사학위논문으로 제출함

2022년 11월

성신여자대학교 일반대학원
통 계 학 과
김 지 우

인 준 서

김지우의 석사학위 논문으로 인준함

2022년 11월

심사위원장 이 성 건 인

심 사 위 원 박 만 식 인

심 사 위 원 박 성 오 인

성신여자대학교 일반대학원

논 문 개 요

4차 산업혁명 시대에는 정보통신기술의 발전과 클라우드 컴퓨팅 환경의 도입으로 수많은 데이터가 빅데이터의 이름으로 수집 및 저장된다. 이러한 빅데이터는 다양한 가치 창출을 위해서 사회 전반에 공유된다. 그러나, 데이터 공유가 활발히 일어날수록 그로 인해 개인의 민감한 정보가 노출되는 프라이버시 침해에 관한 우려도 커지고 있는 실정이다. 이에 프라이버시 노출의 가능성이 있는 원본 데이터 대신 재현 데이터(synthetic data)를 생성하여 배포하는 것이 데이터 공유와 프라이버시 침해 간의 상충관계를 완화하는 대안으로 등장하였다. 재현 데이터는 원본 데이터와 통계적으로 유사한 특성을 가지면서도 임의로 생성된 데이터이기 때문에 기존의 개인정보 비식별 조치보다 프라이버시 침해로부터 안전하다. 따라서, 재현 데이터 생성방법은 프라이버시를 보장하면서 데이터를 적극적으로 공유할 수 있다는 장점이 있다.

지금까지의 연구를 살펴보면, 심층 생성 모형을 활용한 재현 데이터 생성은 대부분 GAN(Generative Adversarial Networks) 방법론에 기반을 두고 있다. GAN은 적대적 학습법(adversarial learning)을 사용하여 내재적으로 불안정하다는 구조적 한계에도 불구하고 재현 데이터 생성 방법론에 활발히 이용되고 있다. 특히, 표 데이터(table type data)의 성공적인 재현을 위해 고안된 대표적인 방법론인 Table-GAN, CTGAN(Conditional Tabular GAN)은 모두 표 데이터의 고유한 특성을 고려하여 GAN의 변형을 시도하였다. CTGAN 논문에서는 GAN과 함께 대표적인 심층 생성 모형 중 하나인 VAE(Variational AutoEncoder)를 이용한 표 데이터 재현 방법론인 TVAE(Tabular VAE)도 소개되었다. VAE는 우도 함수(likelihood function) 기반의 안정적인 학습이 가

능하다는 장점에도 표 데이터 재현 방법론에 많이 이용되지 않았다. 특히, CTGAN 논문에서는 CTGAN과 TVAE의 재현 성능을 비교한 실험의 결과를 제공하는데 TVAE는 CTGAN에 비해 복잡하지 않은 신경망 구조를 가짐에도 더 높은 재현 성능을 보여주었다. 이에 본 논문에서는 TVAE를 발전시켜 우도 함수 기반의 표 데이터 재현 방법론을 제안하고자 한다.

본 논문에서 제안하는 방법론은 TVAE를 3가지 측면에서 발전시켰다. 첫째, 제안 방법은 TVAE에서 사용하는 VAE의 목적함수보다 원본 데이터 분포의 로그 우도 함수(log likelihood)에 대한 더 정밀한 하한 값을 제공하는 IWAE(Importance Weighted AutoEncoders)의 목적함수를 사용하여 학습을 진행하였다는 것이다. 둘째, 제안 방법은 TVAE와 달리 범주형 변수의 범주 불균형 문제를 다루고자 조건부 분포를 고려한 CIWAE(Conditional IWAE)를 개발하여 IWAE를 발전시켰다는 것이다. 마지막으로, CIWAE에 TVAE가 사용하는 전처리 기법 외에 표 데이터 재현 방법론에서 사용하는 다른 전 처리 기법을 추가하여 TCIWAE(Tabular CIWAE)를 고안하였다.

제안 방법인 TCIWAE와 기존의 표 데이터 재현 방법론인 CTGAN, TVAE를 비교하기 위해서 2가지 표 데이터를 가지고 재현 데이터 생성 실험을 진행하였다.

결과적으로, 제안 방법론인 TCIWAE가 생성한 재현 표 데이터는 비교 방법론인 CTGAN, TVAE로부터 얻은 그것과 프라이버시 노출 위험 정도는 비슷하면서 여러 가지 유용성 측면에서 더 유사하였음을 확인할 수 있었다.

목 차

논문개요

Ⅰ. 서 론 ····· 1

1. 기호와 정의 ····· 4

Ⅱ. 관련 연구 ····· 5

1. 기존의 심층 생성 방법론 ····· 5

1) GAN 기반의 방법론 ····· 5

2) VAE 기반의 방법론 ····· 7

2. 표 데이터에 특화된 재현 방법론 ····· 11

1) Table-GAN ····· 11

2) CTGAN ····· 14

Ⅲ. 제안 방법 ····· 16

1. CIWAE 모델 구축 ····· 16

2. CIWAE 모델 학습 ····· 20

3. 표 데이터에 특화된 TCIWAE의 전 처리 ····· 23

Ⅳ. 실험 ····· 24

1. 실험 데이터 및 실험 과정 ····· 24

2. 실험 평가지표 ····· 26

3. 실험 결과 ····· 29

1) Adult 데이터 ····· 29

① 유용성 평가지표 : 주변확률분포 시각화 ····· 29
② 유용성 평가지표 : 분포적 유사성 ····· 43
③ 유용성 평가지표 : 머신러닝 성능의 유사성 ····· 44
④ 프라이버시 평가지표 : DCR ····· 45
⑤ 프라이버시 평가지표 : NNDR ····· 47
2) Covertype 데이터 ····· 50
① 유용성 평가지표 : 주변확률분포 시각화 ····· 50
② 유용성 평가지표 : 분포적 유사성 ····· 64
③ 유용성 평가지표 : 머신러닝 성능의 유사성 ····· 65
④ 프라이버시 평가지표 : DCR ····· 66
⑤ 프라이버시 평가지표 : NNDR ····· 68

Ⅴ. 결론 ····· **71**

참고문헌

ABSTRACT(영문초록)

Ⅰ. 서 론

데이터의 활발한 개방과 공유는 4차 산업혁명 시대의 동력이다. 이에 2012년 대한민국 정부는 정부 3.0을 도입하여 정부가 보유하고 있는 정보를 적극적으로 개방, 공유하여 국민 맞춤형 서비스를 제공하고자 하였다. 현재까지 정부는 '공공데이터포털'[1]을 개설하여 정보를 공개하고 있으며 이러한 공개 데이터를 활용한 다양한 서비스가 사회 전반에 도입되고 있다. 그러나, 데이터 개방과 공유에는 단점 역시 존재한다. 데이터를 원본 그대로 개방하고 공유하는 것은 심각한 프라이버시 침해 문제를 초래할 수 있다. 데이터 공유와 프라이버시 침해 간의 상충관계를 해소하고자 지금까지 다양한 개인정보 비식별 조치가 개발되었다. 예를 들어, $k-$익명화($k-$anonymity)는 주어진 데이터 집합에서 같은 값을 가지는 개체가 적어도 k개 이상이 되도록 함으로써 특정 개체를 식별하지 못하게 하는 방법이다. $k-$익명화와 같은 기존의 비식별화 조치는 다른 데이터에 의한 연결 공격(linkage attack)에 매우 취약하며 낮은 수준의 프라이버시를 보장한다. 이에 원본 데이터와(original data) 유사한 통계적 특징을 지니면서 실제 데이터가 아닌 가상의 재현 데이터(synthetic data)를 만드는 것에 대한 관심이 증가하고 있다.

고차원의 이미지 데이터(image type data) 생성에서 큰 성과를 보인 딥러닝을 이용한 심층 생성 모형(deep generative model)이 최근 재현 데이터 생성에도 활용되고 있다. 대표적으로, GAN(Generative Adversarial Networks, [1])과 VAE(Variational AutoEncoder, [2])는 모두 심층 인공 신경망을 이용한 비선형 인자 모형에 그 뿌리를 두고 있다. GAN은 적대적 학습법(adversarial learning)을 통해 학습하는 반면 VAE는 변분 추론(variational inference)을 통

1) 공공데이터포털: https://www.data.go.kr/

해 원본 데이터 분포의 로그 우도 함수(loglikelihood function)의 하한(Evidence Lower Bound; ELBO)을 구하여 그 하한을 최대화하는 방식으로 학습을 진행한다. GAN과 VAE 모두 학습을 마친 분포로부터 샘플링하여 쉽게 원본과 유사한 모의 데이터를 얻는다는 장점을 가진다. 적대적 학습법을 기반으로 한 GAN은 내재적으로 학습이 불안정하지만 로그 우도 함수를 최적화하는 VAE는 GAN에 비해 학습이 안정적이다.

앞서 소개한 대표적인 심층 생성 모형은 행(rows)과 열(columns)로 구성되며 열 간의 위치적 특성이 없는 표 데이터(table type data) 재현에 곧바로 사용될 수 없다. 그 이유로는 이미지 데이터와 표 데이터 간의 크게 3가지 두드러지는 차이가 있기 때문이다. 첫째, 이미지 데이터는 모두 수치형 값만을 가지나 표 데이터는 수치형과 범주형 값의 혼합으로 이루어진다. 둘째, 이미지 데이터는 변수의 위치적인 특성을 고려하나 표 데이터는 변수의 위치적 정보는 중요하지 않다. 셋째, 이미지 데이터는 각각의 값이 가지는 의미보다 그것의 결합이 가지는 의미가 더 중요하나 표 데이터는 개별 값의 정보 역시 중요하다는 것이다. 이러한 표 데이터의 특징을 고려한 표 데이터 재현 방법론이 여러 연구를 통해 소개되고 있으나 대부분은 심층 생성 모형 중 GAN을 활용하고 있다.

이에 본 논문은 GAN에 비해 안정적인 학습이 가능한 VAE를 활용하여 새로운 표 데이터 재현 방법론을 개발하고자 하였다. VAE 기반의 표 데이터 재현 방법으로는 CTGAN(Conditional Tabular GAN, [3])논문에서 소개하는 TVAE(Tabular VAE)가 있다. 해당 논문에서 TVAE의 재현 성능이 CTGAN보다 더 우수하다는 것이 입증되었다. 이 점을 고려하여 본 논문은 TVAE보다 더 발전된 형태의 우도 함수 기반 표 데이터 재현 모형을 만드는 것을 목표로 하였다. 이를 위해, TVAE가 사용하는 VAE 목적함수인

ELBO 대신 더 정밀한 하한을 제공하는 것으로 알려진 IWAE(Importance Weighted AutoEncoders, [4])의 목적함수를 적용하여 TVAE보다 높은 재현 성능 꾀했다. 또한, 범주형 변수의 불균형한 범주 문제를 다루지 않는 TVAE 대신 CTGAN과 같이 조건부 분포를 모델링하여 범주형 변수의 성공적인 재현을 추구하였다. 이로써, 새로운 형태의 재현 생성 모형과 학습 방법론인 CIWAE(Conditional IWAE)를 제안하고 제안한 CIWAE에 기존의 표 데이터를 전 처리 방법을 도입하여 최종적으로 TCIWAE(Tabular CIWAE)를 고안하였다. 표 데이터 재현에 자주 사용되는 2가지 데이터를 가지고 제안 방법인 TCIWAE과 비교 방법론인 CTGAN, TVAE로부터 재현 데이터를 생성하는 실험을 하였다. 후에 각 재현 데이터의 유용성과 프라이버시 노출 정도 비교, 평가하였다.

본 논문의 구성은 다음과 같다. 2장에서는 대표적인 심층 생성 모형의 방법론과 표 데이터 재현 방법론에 대해 자세히 설명하고, 3장에서는 본 연구에서 제안하는 새로운 모형과 학습방법인 CIWAE를 소개하고 기존의 표 데이터 전 처리 기법을 추가한 TCIWAE를 다룬다. 4장에서는 2가지 표 데이터를 제안 방법인 TCIWAE와 비교 방법인 CTGAN, TVAE를 활용하여 재현하고 유용성 측면과 프라이버시 노출 측면에서 비교 분석한 결과를 보여준다. 마지막으로 5장에서는 본 연구의 결론 및 추후 연구 방향에 대해 논의한다.

1. 기호와 정의

본 논문에서 사용되는 기호와 정의를 서술한 후 논의를 계속 이어나가고자 한다. 재현하고자 하는 p차원 랜덤 벡터 X는 수치형 변수 p_1개, 범주형 변수 p_2개로 구성되며 $X = (R_1, \cdots R_{p_1}, C_1, \cdots, C_{p_2})^T \in \mathbb{R}^p$ 라고 하고 임의의 k에 대해서 $[k] = \{1, \cdots, k\}$라고 정의하자. 이때, 수치형 변수의 값은 $R_{j_1} \in \mathbb{R}, j_1 \in [p_1]$와 같고, 범주형 변수의 범주 값은 $C_{j_2} \in [k_{j_2}]$, $j_2 \in [p_2]$와 같다. X의 분포를 P_X라고 할 때, P_X로부터 추출한 n개의 훈련 자료는 $T^{tr} = \{x_1, \cdots, x_n\}$라고 하자. 여기서 T^{tr}의 i번째 데이터는 $x_i = (r_{i,1}, \cdots, r_{i,p_1}, c_{i,1}, \cdots, c_{i,p_2})^T$ 와 같이 나타낸다. d차원의 잠재 벡터 Z는 $Z = (Z_1, \cdots, Z_d)^T \in \mathbb{R}^d$라고 하자. 더불어, d차원의 영벡터 0_d는 $0_d = (0, \cdots, 0)^T \in \mathbb{R}^d$ 로 정의한다. 주어진 벡터 x, y에 대한 결합은 기호 $\oplus$를 통해 표현한다. 예를 들어, $x = (1, 2)^T$이고 $y = (3, 4)^T$라고 한다면 $x \oplus y = (1,2,3,4)^T$와 같다. 또한, i번째 원소만 1의 값을 가지고 나머지 원소는 0의 값을 가지는 원-핫 인코딩(one-hot encoding)한 벡터는 $e_i = (0, \cdots, 1, \cdots, 0)^T$라고 하자. p차원 벡터 x에서 i번째 원소를 뺀 벡터는 $x_{-i} = (x_1, \cdots, x_{i-1}, x_{i+1}, \cdots, x_p)^T$로 표기한다.

Ⅱ.관련 연구

1. 기존의 심층 생성 방법론

본 장은 널리 사용되고 있는 대표적인 심층 생성 방법론을 소개하고자 한다. 심층 생성 방법론은 크게 두 가지 갈래로 구분된다. 하나는 GAN(Generative Adversarial Networks) 기반의 방법론이고 다른 하나는 VAE(Variational AutoEncoder) 기반의 방법론이다. 먼저 GAN의 모형과 학습 방식을 기술하고 GAN의 학습 불안정성을 개선하고자 GAN을 개량한 WGAN(Wasserstein GAN)을 설명하겠다. 다음으로 VAE의 모형과 학습 방식을 서술하고 원본 데이터 분포의 로그 우도 함수의 하한 값(Evidence Lower Bound; ELBO)을 제공하는 VAE의 목적함수보다 더 정밀한 하한을 제공함으로써 VAE를 개선한 IWAE(Importance Weighted AutoEncoders)를 다루겠다.

1) GAN 기반의 방법론

GAN [1]은 생성자 모형 $g(z;\theta)$와 구분자 모형 $d(x;\eta)$로 구성된다. 이때, 데이터 생성은 식 (2.1)과 같이 이루어진다.

$$\begin{aligned} Z &\sim N(0_{d,} I_d), \\ X|Z=z &= g(z;\theta), \end{aligned} \tag{2.1}$$

여기서 생성자 모형 $g(z;\theta)$의 입력 값인 잠재 벡터 Z는 d차원의 평균이 영벡터이고 분산이 항등 행렬인 다변량 표준정규분포를 따른다. 딥러닝 기반의 생성자 모형 $g(z;\theta)$는 입력 받은 잠재 벡터 z를 가지고 원본 데이터를 생성하는 함수이다. 반면, 구분자 모형인 $d(x;\eta)$의 입력 값인 p차원의 관측 데이터 x는 원본 데이터 혹은 생성자 모형 $g(z;\theta)$에서 생성된 데이터이다. 딥러닝 기반의 구분자 모형 $d(x;\eta)$의 역할은 입력받은 관측 데이터 x가 원본 데이터인지 아닌지를 구분하는 것이다. GAN 방법론에서 생성자 모형 $g(z;\theta)$는 원본 데이터와 최대한 유사한 데이터를 생성하여 구분자 모형 $d(x;\eta)$을 헷갈리게 하는 것을 목표로 하며 구분자 모형 $d(x;\eta)$는 원본 데이터와 생성 데이터를 잘 구분하는 것을 목표로 한다. 두 모형은 서로 상충되는 목적을 가지고 적대적 학습법(adversarial learning)을 진행하기에 GAN의 목적함수는 L_{GAN}이라 하고 식 (2.2)과 같이 표현한다.

$$\min_{\theta}\max_{\eta} L_{GAN} = \mathbb{E}_{X\sim P_X}[\log(d(X;\eta))] + \mathbb{E}_{Z\sim N(0_d, I_d)}[\log(1-d(g(Z;\theta);\eta))], \quad (2.2)$$

여기서 생성자 모형 $g(z;\theta)$는 L_{GAN}을 최소화하는 방향으로 구분자 모형 $d(x;\eta)$는 L_{GAN}을 최대화하는 방향으로 학습을 진행한다. 서로 다른 목적을 가진 모형을 동시에 최적화시키는 것은 어렵기에 GAN은 내재적으로 학습의 불안정성을 지닌다.

GAN의 학습의 불안정성을 해결하기 위해 Wasserstein-1 거리 개념을 도입하여 Wasserstein GAN(WGAN; [5])이 제안되었다. WGAN의 목적함수는 식 (2.3)과 같고 이론적 및 실험적으로 더 안정적인 학습 방법임이 알려져 있다.

$$\min_{\theta}\max_{\eta} L_{GAN} = \mathbb{E}_{X\sim P_X}[(d(X;\eta))] + \mathbb{E}_{Z\sim N(0_d,I_d)}[(1-d(g(Z;\theta);\eta))],$$
$$subject\ to\ ||d(x;\eta)||_L \leq 1, \tag{2.3}$$

이때, $||d(x;\eta)||_L$는 구분자 모형 $d(x;\eta)$의 Lipschitz 상수를 표현하는 것이며 1-Lipschitz를 제약식으로 하는 최적화를 진행한다.

앞서 식 (2.2)와 식 (2.3)을 통해 소개한 GAN과 WGAN의 목적함수를 최적화하는 학습은 경사 하강법(gradient descent)을 통해 이루어진다.

2) VAE 기반의 방법론

VAE [2]는 디코더 모형 $p(x|z;\theta)$을 통해 데이터를 생성하고자 한다. 이때, 데이터 생성은 식 (2.4)과 같이 이루어진다.

$$\begin{aligned} Z &\sim N(0_d, I_d), \\ X|Z=z &\sim p(X|z;\theta), \end{aligned} \tag{2.4}$$

GAN과 마찬가지로 d차원의 잠재 벡터 Z의 분포 $p(Z)$는 평균이 영벡터이고 분산이 항등 행렬인 다변량 표준정규분포로 가정한다. 심층 인공 신경망 기반의 디코더 모형 $p(x|z;\theta)$로부터 생성된 데이터를 원본 데이터와 유사하게 만들기 위해서 x의 로그 우도 함수 $\log p(x;\theta)$를 최대화하는 학습을 진행해야 한다. 이때, x의 로그 우도 함수 $\log p(x;\theta)$는 식 (2.5)과 같이 표현한다.

$$\log p(x;\theta) = \log \int p(z,x;\theta)dz, \tag{2.5}$$

여기서 $p(z,x;\theta)$의 적분 값은 닫힌 형태가 없을 뿐만 아니라 직접적인 근사 값을 얻는 것조차 어렵다. 따라서 VAE는 $p(z|x;\theta)$를 다루기 쉬운 변분 분포 $q(z|x;\phi)$로 근사한다. VAE 방법론이 사용하는 변분 분포 $q(z|x;\phi)$는 평균 벡터가 $\mu(x;\phi)$이고 분산이 $diag(\sigma^2(x;\phi))$인 다변량 정규분포로 가정하고 이를 VAE의 인코더 모형이라고 한다.

심층 인공 신경망 기반의 인코더 모형과 변분 추론(variational inference) 방법을 이용하여 $\log p(x;\theta)$의 하한을 계산하면 식 (2.6)과 같다.

$$\begin{aligned}\log p(x;\theta) &= \log \int \frac{p(z,x;\theta)}{q(z|x;\phi)} q(z|x;\phi)dz \\ &\geq \int \log\left[\frac{p(z,x;\theta)}{q(z|x;\phi)}\right] q(z|x;\phi)dz \\ &=: L_{VAE}(\theta,\phi;x),\end{aligned} \tag{2.6}$$

즉, VAE의 목적함수 $L_{VAE}(\theta,\phi;x)$는 로그 우도 함수 $\log p(x;\theta)$의 하한 값이기 때문에 ELBO(Evidence Lower Bound)라고 불린다.

$$\max_{\theta,\phi} \mathbb{E}_{X\sim P_X}\left[L_{VAE}(\theta,\phi;X)\right], \tag{2.7}$$

여기서 식 (2.7)을 최대화하여 얻는 $\hat{\theta}$과 $\hat{\phi}$은 완벽한 최대 우도 추정값(Maximum Likelihood Estimate; MLE)는 아니더라도 MLE와 비슷한 추정값을 줄 것으로 기대된다.

VAE의 목적함수 $L_{VAE}(\theta,\phi;x)$를 몬테카를로 적분법을 통해 근사한 값을

$\hat{L}_{VAE}(\theta,\phi;x)$라 하고 식 (2.8)과 같이 표현한다.

$$
\begin{aligned}
&z_l \sim q(z|x\phi), \quad l \in [L], \\
&\hat{L}_{VAE}(\theta,\phi;x) := \frac{1}{L}\sum_{l=1}^{L}\log\left(\frac{p(z_l,x;\theta)}{q(z_l|x;\phi)}\right),
\end{aligned}
\tag{2.8}
$$

즉, 변분 분포 $q(z|x;\phi)$로부터 추출한 L개의 샘플을 대입한 뒤 평균을 취하여 $L_{VAE}(\theta,\phi;x)$를 식 (2.8)로 근사한다.

$$
\max_{\theta,\phi}\mathbb{E}_{X\sim P_X}\left[\hat{L}_{VAE}(\theta,\phi;X)\right]. \tag{2.9}
$$

최종적으로 VAE는 식 (2.9)에 대해 경사 상승법(gradient ascent)을 이용하여 $\hat{\theta}$와 $\hat{\phi}$를 얻는다.

VAE는 우도 함수 기반의 학습이 이루어지기 때문에 적대적 학습법을 사용하는 GAN과 달리 안정적이다. 그러나 로그 우도 함수인 $\log p(x;\theta)$를 직접 최대화하지 않고 그것의 하한을 최대화하는 방식으로 학습이 이루어지기 때문에 더 정밀한 하한을 찾고자 하는 후속 연구들이 진행되었다.

VAE의 목적함수가 제공하는 하한보다 더 정밀한 하한을 계산하여 학습을 개선하고자 IWAE [4] 방법이 제안되었다. IWAE는 1개의 인코더 모형을 사용한 VAE와 달리 K개의 인코더 모형을 사용하여 기댓값을 구하고자 하였다. K개의 인코더 모형을 사용한 IWAE의 목적함수 $L_{IWAE}(\theta,\phi;x)$는 식 (2.10)과 같이 표기한다.

$$\begin{aligned}\log p(x;\theta) &\geq \mathbb{E}_{Z_1,\ \cdots,\ Z_k \sim q(Z|x;\phi)}\left[\log\left(\frac{1}{K}\sum_{k=1}^{K}\frac{p(Z_k,x;\theta)}{q(Z_k|x;\phi)}\right)\right] \\ &=: L_{IWAE}(\theta,\phi;x),\end{aligned} \tag{2.10}$$

즉, IWAE의 목적함수인 $L_{IWAE}(\theta,\phi;x)$는 K에 대한 증가함수이며 K가 무한대로 발산하면 $L_{IWAE}(\theta,\phi;x)$가 로그 우도 함수 $\log p(x;\theta)$로 수렴한다.

$$\max_{\theta,\phi}\mathbb{E}_{X\sim P_X}\left[L_{IWAE}(\theta,\phi;X)\right], \tag{2.11}$$

이때, 식 (2.11)과 같이 IWAE의 목적함수 $L_{IWAE}(\theta,\phi;x)$를 θ와 ϕ에 대해서 최대화한다. 또한, 몬테카를로 적분법을 사용하여 얻은 IWAE의 목적함수의 근사 값은 $\hat{L}_{IWAE}(\theta,\phi;x)$라 하고 식 (2.12)와 같이 표기한다.

$$\hat{L}_{IWAE}(\theta,\phi;x) := \log\left(\frac{1}{K}\sum_{k=1}^{K}\frac{p(z_k,x;\theta)}{q(z_k|x;\phi)}\right), \tag{2.12}$$

여기서 각 K개의 인코더 모형에서 1개의 샘플을 추출한 뒤 그 값을 대입하여 근사 값을 구한다.

$$\max_{\theta,\phi}\mathbb{E}_{X\sim P_X}\left[\hat{L}_{IWAE}(\theta,\phi;X)\right]. \tag{2.13}$$

최종적으로 IWAE는 경사 상승법을 이용하여 식 (2.13)에 대한 최적의 $\hat{\theta}$와 $\hat{\phi}$를 얻는다.

2. 표 데이터에 특화된 재현 방법론

본 장에서는 앞서 소개한 심층 생성 모형 중 GAN을 개량하여 표 데이터 재현을 시도한 방법인 Table-GAN, CTGAN을 순서대로 소개하고자 한다. 지금까지 표 데이터 재현에 사용된 모델은 GAN 기반의 방법론이 주를 이루고 있는데, 기존의 GAN은 이미지 데이터 생성에 특화된 모형이다. 따라서 이미지 데이터와 여러 가지 측면에서 차이를 보이는 표 데이터를 재현하기 위해서는 몇 가지 추가적인 작업이 필수적이다. 따라서 성공적인 표 데이터 재현을 위해서 각 방법론이 기존의 GAN에 추가한 데이터 전 처리 방법, 모델, 모델 학습에 대해 기술하겠다.

1) Table-GAN

Table-GAN [6]의 전 처리 과정을 소개하자면 수치형 변수는 표준화 작업 진행하여 측정단위의 영향을 받지 않게 하고 범주형 변수는 각 범주 수만큼 차원을 늘려서 해당 범주에 속하면 1의 값을 속하지 않으면 0의 값을 출력하는 원-핫 인코딩 처리를 수행한다. Table-GAN의 특징 중 하나는 이미지 생성에 특화된 심층 인공 신경망 구조로 이루어진 DCGAN(Deep Convolutional GAN, [7])을 사용하였다는 것이다. DCGAN에서 생성자 모형 $g(z;\theta)$는 디컨볼루션 인공 신경망(DeConvolutional Neural Network; DCNN, [8])으로 구성되고, 구분자 모형 $d(x;\eta)$는 컨볼루션 인공 신경망(Convolutional Neural Network; CNN, [9])의 형태를 가진다. 따라서 표 데이터의 구조를 이미지 데이터의 형태와 같은 정방행렬(square matrix)로 바꾸어주는 전 처리 작업이 필요하다. 예를 들어, 24차원의 표 데이터를 5×5

정방행렬 형태로 변환하려면 변수 순서를 고려하지 않고 값을 넣은 후 5×5의 정방행렬의 마지막 빈 원소에는 0을 넣어주는 제로 패딩(zero padding)을 해준다.

Table-GAN의 모형은 일반적인 GAN 모형에 해당하는 생성자 모형 $g(z;\theta)$와 구분자 모형 $d(x;\eta)$에 분류기 모형 $cl(x;\delta)$을 추가한다. 분류기 모형 $cl(x;\delta)$은 변수 하나 $X_j, j \in [p]$를 종속 변수로 그 외의 나머지 모든 변수 X_{-j}를 독립변수로 설정하여 종속 변수를 잘 분류하는 것을 목표로 한다. Table-GAN은 분류기 모형 $cl(x;\delta)$을 추가함으로써 표 데이터의 의미적 일치성(semantic integrity)을 높이고자 한다. 여기서 의미적 일치성이란 불가능한 사건 조합이 없음을 뜻한다. 불가능한 사건 조합으로는 혈당 변수의 값이 '$50mg/dL$'인 개체의 당뇨 유무 변수의 값이 '당뇨'인 경우이다. 이러한 조합을 생성하지 않는 재현 데이터를 만드는 것은 매우 중요하다.

Table-GAN은 3가지 목적함수를 최적화하는 학습을 진행한다. 첫 번째 목적함수는 식 (2.2)의 L_{GAN}이다. 두 번째 목적함수는 정보 손실(information loss) 함수라고 하며 재현 데이터와 원본 데이터의 통계적 유사성을 높이기 위해 고안되었다. 이때, 재현 데이터와 원본 데이터의 통계적 유사성을 측정하기 위해서 구분자 모형 $d(x;\eta)$의 최상층 은닉층 벡터 $f(x;\eta)$를 추출하여 재현 데이터와 원본 데이터의 평균과 표준편차의 유클리디안 놈을 구하고 식 (2.14)과 같이 표기한다.

$$
\begin{aligned}
L_{mean} &= \| \mathbb{E}_{X \sim P_X}[f(X;\eta)] - \mathbb{E}_{Z \sim N(0_d, I_d)}[f(g(Z;\theta);\eta)] \|_2, \\
L_{sd} &= \| \mathbb{SD}_{X \sim P_X}[f(X;\eta)] - \mathbb{SD}_{Z \sim N(0_d, I_d)}[f(g(Z;\theta);\eta)] \|_2.
\end{aligned}
\tag{2.14}
$$

최종적으로 정보 손실 함수는 식 (2.14)를 힌지 손실(hinge loss) 함수 형식으로 구성함으로써 식 (2.15)과 같이 표현된다.

$$L_{info} = \max(0, L_{mean} - \xi_{mean}) + \max(0, L_{sd} - \xi_{sd}), \tag{2.15}$$

이때, $\xi_{mean}, \xi_{sd} > 0$는 조율모수(hyperparameter)로 생성 데이터와 원본 데이터의 통계적 유사성 정도를 조절한다.

세 번째 목적함수는 분류 손실(classification loss) 함수이다. 분류기 자체의 성능을 높이는 원본 데이터를 이용한 분류기 목적함수와 재현 데이터의 의미적 일치성을 높이는 생성 데이터를 이용한 분류기 목적함수를 각각 식 (2.16)과 같이 정의한다.

$$\begin{aligned} L_{Class}^{cl} &= \mathbb{E}_{X \sim P_X}[|X_j - cl(X_{-j};\delta)|], \\ L_{Class}^{g} &= \mathbb{E}_{Z \sim N(0_d, I_d)}[|g_j(Z;\theta) - cl(g_{-j}(Z;\theta);\delta)|]. \end{aligned} \tag{2.16}$$

생성자 모형의 θ, 구분자 모형의 η, 분류기 모형의 δ는 식 (2.17)과 같이 최적화된다.

$$\begin{aligned} &\min_{\theta} \; L_{GAN} + L_{info} + L_{Class}^{g} \\ &\max_{\eta} L_{GAN} \\ &\min_{\delta} L_{Class}^{cl} \end{aligned} \tag{2.17}$$

2) CTGAN (Conditional Tabular GAN)

CTGAN [3]은 Table-GAN과 같이 범주형 변수는 원-핫 인코딩을 해주고, 수치형 변수는 특정 모드 기반 정규화(Mode-Specific Normalization; MSN)라는 전 처리를 수행한다. MSN 기법에서는 수치형 변수가 여러 개의 봉우리를 가지는 가우시안 혼합 모델(Gaussian mixture model)을 따른다고 가정한다. 이에 변분 가우시안 혼합 모델(Variational Gaussian Mixture model; VGM, [10])을 이용하여 수치형 변수의 봉우리의 개수를 추정하고 가우시안 혼합 모델로 적합시킨다. 예를 들어, p_1개의 수치형 변수 중 임의의 하나의 변수를 R_{j_1}, $j_1 \in [p_1]$라고 하자. VGM을 통해 R_{j_1}는 m_{j_1}개의 봉우리를 가지는 가우시안 혼합 모델로 적합 되고 적합된 모형은 식 (2.18)과 같다.

$$p(R_{j_1} = r) = \sum_{t=1}^{m_{j_1}} \phi_t N(r; \mu_t, \sigma_t), \tag{2.18}$$

여기서 r는 수치형 변수 R_{j_1}로부터 얻은 하나의 값, ϕ_t는 t번째 주변 가우시안 분포의 가중치, μ_t는 t번째 주변 가우시안 분포의 평균, σ_t는 t번째 주변 가우시안 분포의 표준편차이다.

R_{j_1} 변수의 i번째 관측 값 r_{i,j_1}에 대한 m_{j_1}개 각각의 주변 가우시안 분포에서의 확률 밀도 값을 ρ_t라 하고 가장 높은 확률 밀도 값을 가질 때의 가우시안 분포의 인덱스를 t^*로 둘 때 식 (2.19)와 같이 표기한다.

$$\rho_t = \phi_t N(r_{i,j_1}; \mu_t, \sigma_t), \quad i \in [n],\ t \in [m_{j_1}], \tag{2.19}$$

$$t^* = \arg\max_t \rho_t.$$

식 (2.19)를 통해 얻은 가우시안 분포의 인덱스 t^*에 해당하는 평균 μ_{t^*}와 표준편차 σ_{t^*}를 이용하여 식 (2.20)와 같이 전 처리된 값을 얻는다.

$$\alpha_{i,j_1} = \frac{r_{i,j_1} - \mu_{t^*}}{4\sigma_{t^*}}, \quad \beta_{i,j_1} = e_{t^*} \in \mathbb{R}^{m_{j_1}}, \tag{2.20}$$

이때, α_{i,j_1}는 관측 값 r_{i,j_1}이 가장 있을법한 주변 가우시안 분포(t^*번째 주변 가우시안 분포)의 정보를 바탕으로 표준화를 진행하여 얻은 값으로, -1~1 사이의 값을 가진다. β_{i,j_1}은 t^*번째만 1의 값을 가지고 나머지는 0의 값을 가지는 이진 벡터로, t^*번째 가우시안 분포의 정보를 가지고 표준화하였음을 알려준다. 최종적으로 범주형 변수를 전 처리한 원-핫 인코딩 벡터와 수치형 변수를 MSN 처리하여 얻은 α, β 벡터를 연결하여 모형에 입력할 데이터를 구성한다.

CTGAN 모형은 범주형 변수의 범주 불균형 문제를 다루고자 모든 범주형 변수 중 하나의 변수를 매 학습마다 뽑아 해당 범주형 변수 값은 경험 분포에 의해 먼저 생성한다. 후에 경험 분포로 생성한 특정 범주를 조건으로 하고 나머지 모든 변수를 조건부 분포를 통해 생성하는 방식을 취한다. 이때 조건부 분포를 모형화하기 위해서 Conditional GAN [11]을 사용하고 이를 학습하기 위해서 WGAN의 목적함수인 식 (2.3)을 이용한다.

Ⅲ. 제안 방법

본 장은 크게 제안하는 CIWAE(Conditional IWAE) 모델 구축, CIWAE 모델 학습, 표 데이터에 특화된 TCIWAE 모델의 데이터 전 처리 순으로 구성된다. 먼저, 제안 모델인 CIWAE의 구조를 설명하고, 범주형 자료의 범주 불균형 문제를 해결하고자 조건부 벡터를 어떻게 모델링하였는지 소개하겠다. 그 다음으로 제안 모델이 조건부 모형을 학습하고자 IWAE의 목적함수를 어떻게 사용하였는지를 설명하겠다. 마지막으로 표 데이터 재현 방법인 TCIWAE에서 사용한 표 데이터의 특징을 고려한 전 처리 방법을 소개하겠다. 특히, CTGAN에서 제안한 수치형 자료 전 처리 기법인 특정 모드 기반 정규화(MSN)를 수행하기 어려운 경우를 제시하고 효과적인 MSN을 수행을 위해 CTAB-GAN [12]에서 사용하는 로그 변환에 대해 설명하겠다.

1. CIWAE 모델 구축

본 논문에서 제안하는 모델은 TVAE에 CTGAN의 조건부 벡터 아이디어를 추가하였는데, 그 이유로는 TVAE가 우도 함수 기반의 목적함수를 사용하여 안정적인 학습이 가능하다는 장점을 가지나 범주형 변수의 범주 불균형 문제를 다루고 있지 않는다는 한계가 있어 이를 보완하고자 하였기 때문이다.

조건부 벡터를 만들기 위해서는 먼저 특정 범주형 변수 하나를 선택해야 한다. 예를 들어, p_2개의 범주형 변수 중 임의의 하나의 변수를 C_{J_2}라고 할 때 특정 범주형 변수 선택은 식 (3.1)과 같다.

$$J_2 \sim Multi(\frac{1}{p_2}, \cdots, \frac{1}{p_2}). \tag{3.1}$$

그 다음으로 범주형 변수의 변수 인덱스가 $J_2 = j_2$로 주어졌을 때 선택된 범주형 변수 C_{j_2}를 모델링한다. 범주형 변수 C_{j_2}가 k_{j_2}개의 범주를 가질 수 있다고 하자. 본 논문에서는 다항분포를 이용해 C_{j_2}의 분포를 규정하고 이를 식 (3.2)과 같이 표현한다.

$$C_{j_2} \sim Multi(p_{j_2,1}, \cdots, p_{j_2,k_{j_2}}). \tag{3.2}$$

최종적으로, 범주형 변수 C_{j_2}의 범주가 c로 주어졌을 때 해당 범주를 조건으로 하는 조건부 벡터를 $c^{cond}_{j_2,c}$라 하고 이를 식 (3.3)과 같이 표현한다.

$$c^{cond}_{j_2,c} = 0_{k_1} \oplus 0_{k_2} \oplus \cdots \oplus e_c \oplus \cdots \oplus 0_{k_{p_2}}, \tag{3.3}$$

즉, 범주형 변수 C_{j_2}의 범주 c에서만 1의 값을 가지고 나머지 모든 범주형 변수의 범주에서는 0의 값을 가지는 벡터가 된다.

기존의 GAN, VAE 기반의 방법론들이 가정한 것과 마찬가지로 d차원 잠재 벡터 $Z \in R^d$의 분포는 평균이 영벡터, 분산이 항등행렬인 다변량 정규분포를 가정한다. 이를 식으로 나타내면 식 (3.4)과 같다.

$$Z \sim N(0_d, I_d). \tag{3.4}$$

디코더 모형에 사용할 $Z = z$ 와 앞에서 언급한 $c_{j_2,c}^{cond}$을 연결하여 식 (3.5)와 같이 표기한다.

$$(z_1, \cdots, z_d)^T \oplus 0_{k_1} \oplus 0_{k_2} \oplus \cdots \oplus e_c \oplus \cdots \oplus 0_{k_{p_2}}. \tag{3.5}$$

식 (3.5)의 벡터를 조건부 입력 값으로 하는 디코더 모형은 식 (3.6)과 같이 표현한다.

$$R, C_{-j_2} | Z = z, C_{j_2} = c \sim p(R, C_{-j_2} | z \oplus c_{j_2,c}^{cond}; \theta), \tag{3.6}$$

이때, 수치형 변수 R과 범주형 변수 C가 조건부 독립이라고 가정하면 식 (3.7)처럼 표현할 수 있다.

$$\begin{aligned} & p(R, C_{-j_2} | z \oplus c_{j_2,c}^{cond}; \theta) \\ & = \prod_{\tilde{j}_1 = 1}^{p_1} p_{\tilde{j}_1}(R_{\tilde{j}_1} | z \oplus c_{j_2,c}^{cond}; \theta) \prod_{\tilde{j}_2 \in [p_2] \setminus j_2} p_{\tilde{j}_2}(C_{\tilde{j}_2} | z \oplus c_{j_2,c}^{cond}; \theta), \end{aligned} \tag{3.7}$$

여기서 수치형 변수는 정규분포로 가정하고, 범주형 변수는 소프트맥스(softmax) 함수를 활용한 다항분포로 가정한다.

최종적으로 본 논문에서 제안하는 모델의 데이터 생성 모형의 구축 과정은 아래와 같이 정리할 수 있다.

<데이터 생성 모형 구축 과정>

- 범주형 변수 선택: $J_2 \sim Multi(\frac{1}{p_2}, \cdots, \frac{1}{p_2})$
- 선택한 변수의 특정 범주 선택: $C_j | J_2 = j_2 \sim Multi(p_{j_2,1}, \cdots, p_{j_2,k_{j_2}})$
- 조건부 벡터: $c^{cond}_{j_2,c} = 0_{k_1} \oplus 0_{k_2} \oplus \cdots \oplus e_c \oplus \cdots \oplus 0_{k_{p_2}}$
- 잠재 벡터 추출: $Z \sim N(0_d, I_d)$
- 디코더 모형 구축: $R, C_{-j_2} | Z = z, C_{j_2} = c \sim P(R, C_{-j_2} | z \oplus c^{cond}_{j_2,c}; \theta)$

원본 데이터가 주어졌을 때 잠재 벡터를 만들어내는 인코더 모형 $q(Z|R=r, C=c; \phi)$는 VAE, IWAE와 동일하게 $N(\mu(r,c;\phi), diag(\sigma^2(r,c;\phi)))$의 확률밀도함수로 이루어진다.

2. CIWAE 모델 학습

주어진 $(R,C)=(r,c)$에 대해서 로그 우도 함수를 식 (3.8)과 같이 분해 할 수 있다.

$$\begin{aligned}
&\log p(R=r,\, C=c;\theta,p)\\
&=\log\sum_{j_2}\{p(R=r,\, C=c,\, J_2=j_2;\theta,p)\}\\
&=\log\sum_{j_2}\{p(R=r,\, C=c|\, J_2=j_2;\theta,p)\times p(J_2=j_2)\}\\
&=\log\sum_{j_2}\{p(R=r,\, C_{j_2}=c_{j_2},\, C_{-j_2}=c_{-j_2}|\, J_2=j_2;\theta,p)\times p(J_2=j_2)\}\\
&=\log\sum_{j_2}\begin{Bmatrix}p(R=r,\, C_{-j_2}=c_{-j_2}|\, J_2=j_2,\, C_{j_2}=c_{j_2};\theta,p)\times\\ p(C_{j_2}=c_{j_2}|J_2=j_2;p)\times p(J_2=j_2)\end{Bmatrix},
\end{aligned} \tag{3.8}$$

여기서 p는 각 범주형 변수마다 범주가 뽑힐 다항분포의 확률 값 전체를 뜻한다. 식 (3.8)의 실제 계산은 어렵기 때문에 몬테카를로 적분법을 활용하여 식 (3.1)로부터 j_2를 샘플링한다. 근사한 로그 우도 함수는 식 (3.9)과 같다.

$$\begin{aligned}
&\log p(R=r,\, C=c;\theta,p)\\
&\approx \log p(R=r, C_{-j_2}=c_{-j_2}|C_{j_2}=c_{j_2};\theta)p(C_{j_2}=c_{j_2};p)\\
&\approx \log p(R=r, C_{-j_2}=c_{-j_2}|C_{j_2}=c_{j_2};\theta)+\log p(C_{j_2}=c_{j_2};p).
\end{aligned} \tag{3.9}$$

따라서 로그 우도 함수를 최대화한다는 것은 식 (3.9)의 근사한 로그 우도

함수의 구성 요소인 $\log p(R=r, C_{-j_2}=c_{-j_2}|C_{j_2}=c_{j_2};\theta)$와 $\log p(C_{j_2}=c_{j_2};p)$를 각각 최대화한다는 것을 의미한다.

먼저 $\log p(C_{j_2}=c_{j_2};p)$을 최대화하는 p를 추정하고자 한다. 식 (3.2)에서 C_{j_2}를 다항분포로 모형화하였는데, 이때 범주별 비율 값인 $p_{j_2,1}, \cdots, p_{j_2,k_{j_2}}$는 최대우도 추정치인 훈련 데이터에서의 범주 비율 값으로 구한다. 예를 들어, $p_{j_2,c}$의 표본 비율이 $\hat{p}_{j_2,c}$라고 할 때 $\hat{p}_{j_2,c}$는 식 (3.10)와 같이 구한다.

$$\hat{p}_{j_2,c} = \frac{1}{n}\sum_{i=1}^{n} I(c_{ij_2}=c). \tag{3.10}$$

다음으로 $\log p(R=r, C_{-j_2}=c_{-j_2}|C_{j_2}=c_{j_2};\theta)$를 최대화하는 θ를 추정하고자 한다. 이때, IWAE [4]의 학습 방법을 도입해 식 (3.11)과 같이 표기한다.

$$\begin{aligned}
&\log p(R=r, C_{-j_2}=c_{-j_2}|C_{j_2}=c_{j_2};\theta) \\
&\geq \mathbb{E}_{Z_1,\cdots,Z_K \sim q(Z|r,c;\phi)}\left[\log\left(\frac{1}{K}\sum_{k=1}^{K}\frac{p(Z_k, r, c_{-j_2}|c_{j_2};\theta)}{q(Z_k|r,c;\phi)}\right)\right] \\
&=: L_{CIWAE}(\theta,\phi;r,c_{-j_2}).
\end{aligned} \tag{3.11}$$

식 (3.11)에 몬테카를로 적분법을 적용하여 CIWAE의 목적함수의 근사 값 $\hat{L}_{CIWAE}(\theta,\phi;r,c_{-j_2})$를 식 (3.12)와 같이 표현한다.

$$\hat{L}_{CIWAE}(\theta, \phi; r, c_{-j_2}) := \log\left(\frac{1}{K}\frac{p(z_k, r, c_{-j_2}|c_{j_2};\theta)}{q(z_k|r, c;\phi)}\right). \tag{3.12}$$

$$\max_{\theta,\phi} \mathbb{E}_{R, C\sim P_X}\left[\hat{L}_{CIWAE}(\theta, \phi; R, C_{-j_2})\right]. \tag{3.13}$$

최종적으로 식 (3.13)을 최대화하는 방향으로 경사 상승법을 이용해 최적의 $\hat{\theta}$와 $\hat{\phi}$를 구한다.

3. 표 데이터에 특화된 TCIWAE를 위한 전 처리

제안하는 CIWAE를 사용해 표 데이터를 성공적으로 재현하기 위해서는 표 데이터의 변수 유형에 따른 전 처리 진행이 필수적이다. 먼저, 범주형 변수는 원-핫 인코딩을 진행한다. 다음으로 수치형 변수는 CTGAN에서 제안하는 특정 모드 기반 정규화(MSN)를 수행한다. 이때, MSN은 수치형 변수의 분포가 가우시안 혼합 분포를 따른다고 가정하여 변분 가우시안 혼합 모델(VGM)로 모델링한다. 그러나 수치형 변수의 분포가 왼쪽 혹은 오른쪽으로 꼬리가 길 때 VGM이 잘 동작하지 않는다. 왼쪽 혹은 오른쪽으로 꼬리가 긴 수치형 변수는 표 데이터에서 흔히 등장하기에 이러한 문제를 해결하고자 CTAB-GAN [12]에서는 꼬리가 긴 수치형 변수에 먼저 로그 변환을 취하고 난 뒤 MSN을 시도한다. 로그 변환을 취함으로써 대부분의 데이터가 집중 되어 있는 부분과 데이터가 거의 없는 꼬리 부분 사이의 거리가 줄어 들어 VGM이 더 잘 작동하게 된다. 예를 들어, 꼬리가 긴 분포를 따르는 수치형 변수 R_j의 하한 값을 l_j라고 하고 0보다 큰 작은 임의의 값을 ϵ이라고 할 때 로그 변환한 값 $\tilde{r}_{i,j}$는 식 (3.14)과 같이 표현한다.

$$\tilde{r}_{i,j} = \begin{cases} \log(r_{i,j}) & \text{if } l_j > 0 \\ \log(r_{i,j} + \epsilon) & \text{if } l_j = 0, \epsilon > 0 \\ \log(r_{i,j} - l_j + \epsilon) & \text{if } l_j < 0, \epsilon > 0 \end{cases} \tag{3.14}$$

Ⅳ. 실험

본 장은 실험 데이터 및 실험 과정을 설명하고, 재현 데이터를 평가하기 위한 유용성 평가지표 및 프라이버시 노출 평가지표를 소개한 뒤, 실험 결과를 제시하는 순서로 구성된다. 먼저, 실험에서 사용한 2가지 데이터의 출처, 변수 유형, 변수의 특징을 밝히고 재현 데이터 생성을 위한 실험 과정을 설명하겠다. 그 다음으로는 유용성 평가지표인 원본 데이터와 재현 데이터의 확률분포 유사성, 머신러닝 성능의 유사성을 구하는 방법과 프라이버시 노출 평가지표인 원본 데이터와 재현 데이터의 거리 차이를 계산하는 방법을 소개하겠다. 마지막으로 실험 결과를 통해 제안 방법의 우수성을 다른 방법론과 비교하여 확인하겠다.

1. 실험 데이터 및 실험 과정

본 연구에서 사용한 실험 데이터는 Adult, Covertype로 총 2가지 이며 UCI machine learning repository를 통해 얻었다. 실험에 사용한 데이터는 표 자료 재현과 관련된 기존의 논문에서 자주 사용하는 데이터이다.

각 데이터에 대해 소개하자면, 첫 번째로 Adult는 인구통계학 정보를 기반으로 개인의 연 소득이 $50,000를 초과하는지 아닌지를 이진 분류(binary classification)하기 위한 데이터이다. 관측치는 48,842개이며 15개의 변수 중 9개의 변수가 범주형에 해당한다. 수치형 변수 중에서 꼬리가 긴 분포 형태를 가지는 변수는 총 3개이다.

두 번째로 Covertype은 지리, 지질학 정보를 바탕으로 해당 지역의 숲의 유

형을 판단하는 다중분류(multiclass classification)를 위한 데이터이다. 관측치는 52,292개이며 55개의 변수 중 45개의 변수가 범주형에 해당한다. 이때, 45개의 범주형 변수 중 숲의 유형을 나타내는 변수('Cover_Type')를 제외한 모든 변수는 이진 변수로 구성되었다. 수치형 변수 중에서 꼬리가 긴 분포 형태를 보이는 변수는 없었다.

위의 실험 데이터에 관한 설명을 표로 정리하면 [표 1]과 같다.

[표 1] 실험 데이터 설명

데이터	종속 변수	수치형 변수 (꼬리가 긴 변수)	범주형 변수
Adult	'income'	6(3)	9
Covertype	'Cover_Type'	10(0)	45

본 연구의 실험 과정으로는 위의 2가지 실험 데이터 중 80%를 훈련 데이터로 20%를 시험 데이터로 나누어 훈련 데이터만을 사용하여 그것과 동일한 크기의 재현 데이터를 생성하였다. 이때, Adam 옵티마이저(optimizer)를 이용해 제안 방법인 TCIWAE와 비교 방법론인 CTGAN, TVAE를 학습하였고 시험 데이터는 후에 유용성 평가지표인 머신러닝 성능을 측정할 때 사용하였다. 특히, 제안 방법인 TCIWAE는 인코더의 개수를($K=1, 10, 20, 30, 40, 50$) 다양하게 고려하여 실험하였다. 공정한 비교와 평가를 위해서 4가지 조율 모수를 모든 데이터, 모든 재현 데이터 생성 모델마다 동일하게 설정하였다. 먼저, 학습 반복 수(epochs)는 300으로, 미니 배치 크기(mini-batch size)는 500으로, 학습률(learning rate)은 0.001로 정하였다. 또한, 수치형 변수 전처리 기법인 MSN에서 추정하는 봉우리의 최대개수를 10개로 하였다.

2. 실험 평가지표

다양한 재현 데이터 생성 모델로부터 얻은 재현 데이터를 평가하고자 기존의 재현 데이터 평가지표로 널리 사용되는 유용성 평가지표와 프라이버시 노출 평가지표를 사용하였다. 유용성 평가지표는 재현 데이터가 원본 데이터의 의미 있는 정보를 잘 재현했는지를 판단한다. 반면, 프라이버시 노출 평가지표는 재현 데이터가 원본 데이터의 정보를 얼마나 잘 드러내는가를 확인한다. 다시 말해, 재현 데이터의 유용성이 높을수록 재현 데이터가 원본 데이터와 비슷하게 되어 프라이버시 노출 위험이 커지게 된다. 이에 유용성 평가지표와 프라이버시 노출 평가지표를 동시에 고려하여 재현 데이터를 평가해야 한다.

본 연구에서는 재현 데이터의 유용성을 3가지 측면에서 확인한다. 첫 번째로 재현 데이터의 주변확률분포가 원본 데이터의 그것과 유사한지를 판단하였다. 이를 위해 재현 데이터의 주변확률분포와 원본 데이터의 주변확률분포를 시각화해보거나 두 데이터 간의 분포 거리를 계산한다. 이때, 범주형 변수 간의 분포 거리는 잰슨-섀넌 발산(Jensen-Shannon divergence)을 통해 측정한다. 재현 데이터의 범주형 변수와 원본 데이터의 범주형 변수의 분포가 이질적일수록 1에 가까운 값을 두 분포가 비슷할수록 0에 가까운 값을 가진다. 반면, 수치형 변수 간의 분포 거리를 계산할 때 잰슨-섀넌 발산을 사용하게 되면 재현 데이터의 수치형 변수의 구간과 원본 데이터의 그것이 겹치지 않을 경우 계산한 값이 안정적이지 않게 된다. 이에 더 안정적인 두 수치형 변수 간의 분포 거리를 측정하고자 Wasserstein 거리를 이용하고 수치형 변수의 측정단위에 따라 Wasserstein 거리가 크게 변하므로 먼저 표준화를 한 후 분포 거리를 계산한다.

두 번째로 재현 데이터와 원본 데이터의 두 변수 간 상관성이 높은지를 판

단한다. 이때, 각 데이터마다 수치형 변수 간의 상관성은 피어슨 상관계수(Pearson's correlation coefficient)로 계산하고 수치형 변수와 범주형 변수 간의 상관성은 상관비 계수(correlation ratio)로 측정하며 범주형 변수 간의 상관성은 Theil's U 계수를 통해 구한다. 원본 데이터의 모든 변수 조합의 상관성과 재현 데이터의 그것의 차이를 계산하여 차이가 작을수록 원본 데이터의 변수 간 상관성 정보를 잘 담고 있는 재현 데이터라고 평가한다.

세 번째로 재현 데이터의 머신러닝 성능과 원본 데이터의 그것이 비슷한지를 판단한다. 재현 데이터를 다양한 머신러닝 모델(로지스틱회귀, 의사결정나무, 랜덤포레스트, 다층 퍼셉트론)의 학습 데이터로 사용했을 때 시험 데이터에서의 머신러닝의 성능과 원본 데이터를 학습 데이터로 사용한 경우의 그것과 비교하여 차이가 작을수록 재현 데이터의 유용성이 높다고 평가한다.

본 연구에서는 재현 데이터의 프라이버시 노출 위험을 2가지 측면에서 확인한다. 첫 번째로 표 데이터에서 한 행의 정보를 레코드라고 할 때, DCR(Distance to the Closest Record)은 재현 레코드와 가장 가까운 원본 레코드 사이의 거리를 통해 프라이버시 노출 정도를 판단한다. DCR를 계산하기 위해 먼저 레코드의 각 변수마다 측정단위가 다르기 때문에 변수별 표준화 작업을 진행한다. 표준화된 재현 레코드에 유클리디안 거리(Euclidean distance)가 가장 가까운 표준화된 원본 레코드를 찾고 그 거리를 DCR이라고 한다. DCR이 가까울수록 원본 레코드의 노출 위험이 커진다고 평가할 수 있을 때, 각 재현 데이터 생성 모델별로 DCR를 시각화하여 각 재현 데이터의 프라이버시 노출 위험을 순위 매긴다.

두 번째로 NNDR(Nearest Neighbor Distance Ratio)은 재현 레코드와 가장 가까운 원본 레코드와 두 번째로 가까운 원본 레코드 간의 거리의 비율을 통해 프라이버시 노출 위험을 판단한다. DCR을 계산하는 방법과 동일하게 구할

수 있으며 NNDR은 비율 값이므로 0~1 사이의 값을 가지며 0의 값에 가까울수록 원본 레코드가 재현 레코드와 거리가 가까워 노출 위험이 커진다. DCR과 마찬가지로 각 재현 자료 생성 방법론별 NNDR을 시각화하여 재현 데이터별 프라이버시 노출 위험 순위를 확인한다.

위의 실험 평가지표에 관한 설명을 표로 정리하면 [표 2]과 같다.

[표 2] 실험 평가지표 설명

유용성 측면	프라이버시 노출 위험 측면
1. 주변확률분포의 유사성 2. 두 변수별 상관성의 유사성 3. 머신러닝 모델 성능의 유사성	1. DCR 2. NNDR

3. 실험 결과

Adult, Covertype 데이터 순으로 제안 방법인 TCIWAE와 비교 방법인 CTGAN, TVAE로 만든 재현 데이터의 성능을 앞서 소개한 유용성 평가지표와 프라이버시 노출 위험 평가지표를 가지고 평가한 결과를 살펴보겠다.

1) Adult 데이터

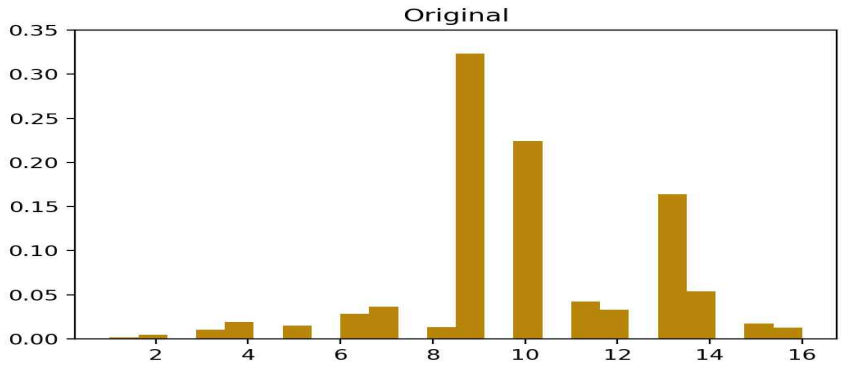
① 유용성 평가지표 : 주변확률분포 시각화

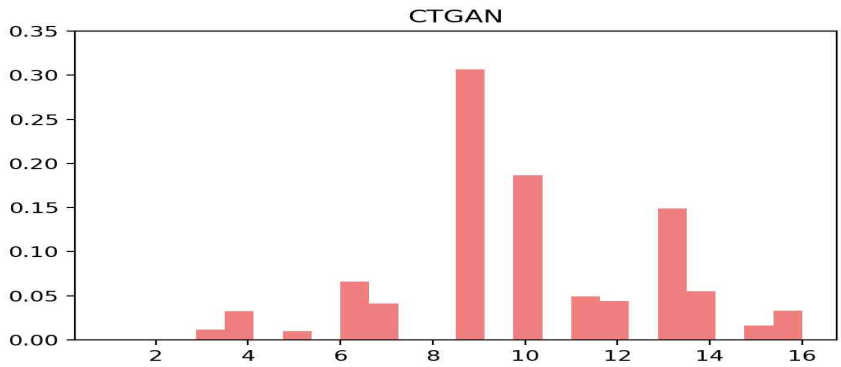
Adult 데이터의 수치형 변수와 범주형 변수의 분포를 시각화하여 각 방법론별 재현 데이터의 주변확률분포가 원본 데이터의 그것을 잘 따르고 있는지 확인하였다. 이때, 제안 방법인 TCIWAE를 통해 재현 데이터를 생성하는 과정에서 Adult 데이터의 수치형 변수 중 꼬리가 긴 변수인 'fnlwgt', 'capital-gain', 'capital-loss'는 CTAB-GAN [12]에서 제안한 로그 변환 후 MNS를 수행하여 데이터 재현을 시도하였다. 변수의 개수가 많은 관계로 [표 3]과 같이 수치형 변수 2개(로그변환 시도 유무), 범주형 변수 2개(불균형한 범주 가진 변수)를 선택하였고 이를 시각화한 결과를 [그림 1]-[그림 4]에서 확인할 수 있다.

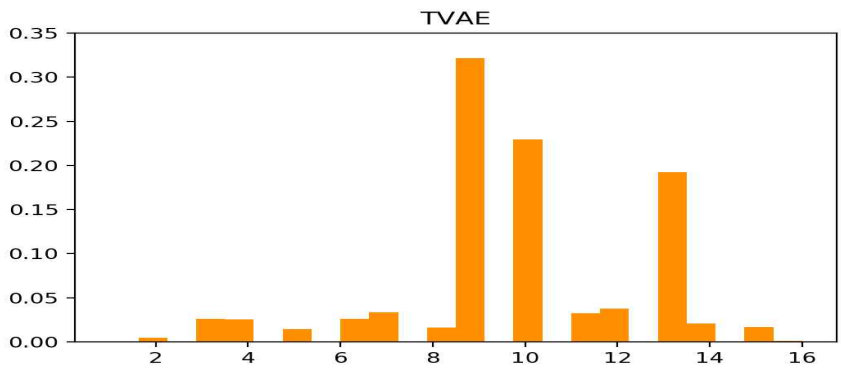
[표 3] Adult 데이터 분포 시각화 변수

수치형 변수	범주형 변수
'educational-num'	'workclass'
'fnlwgt'(로그변환)	'race'

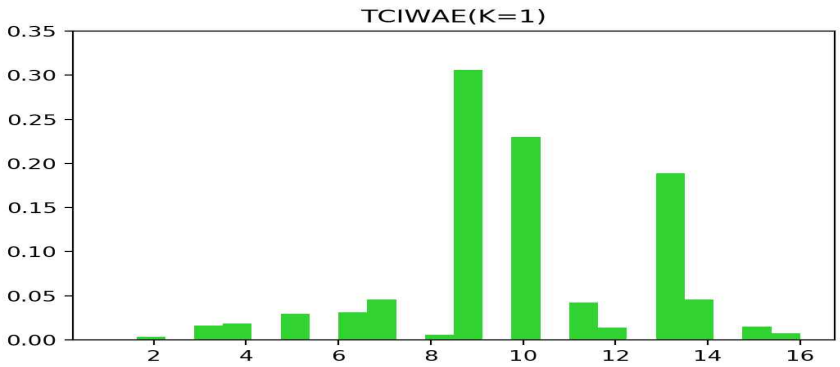
1. 수치형 변수 : ‘educational-num’

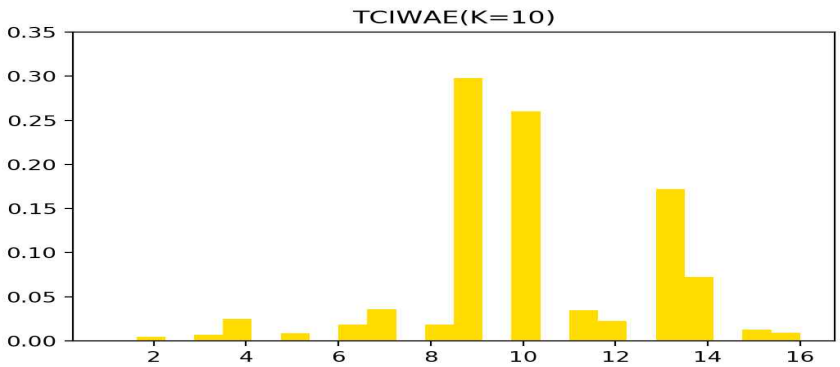


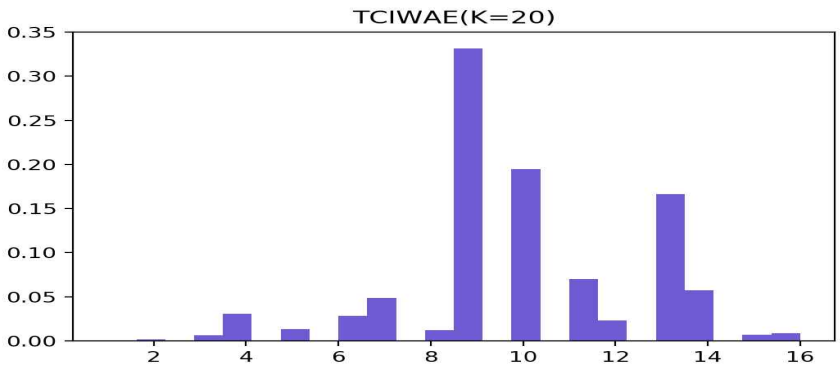




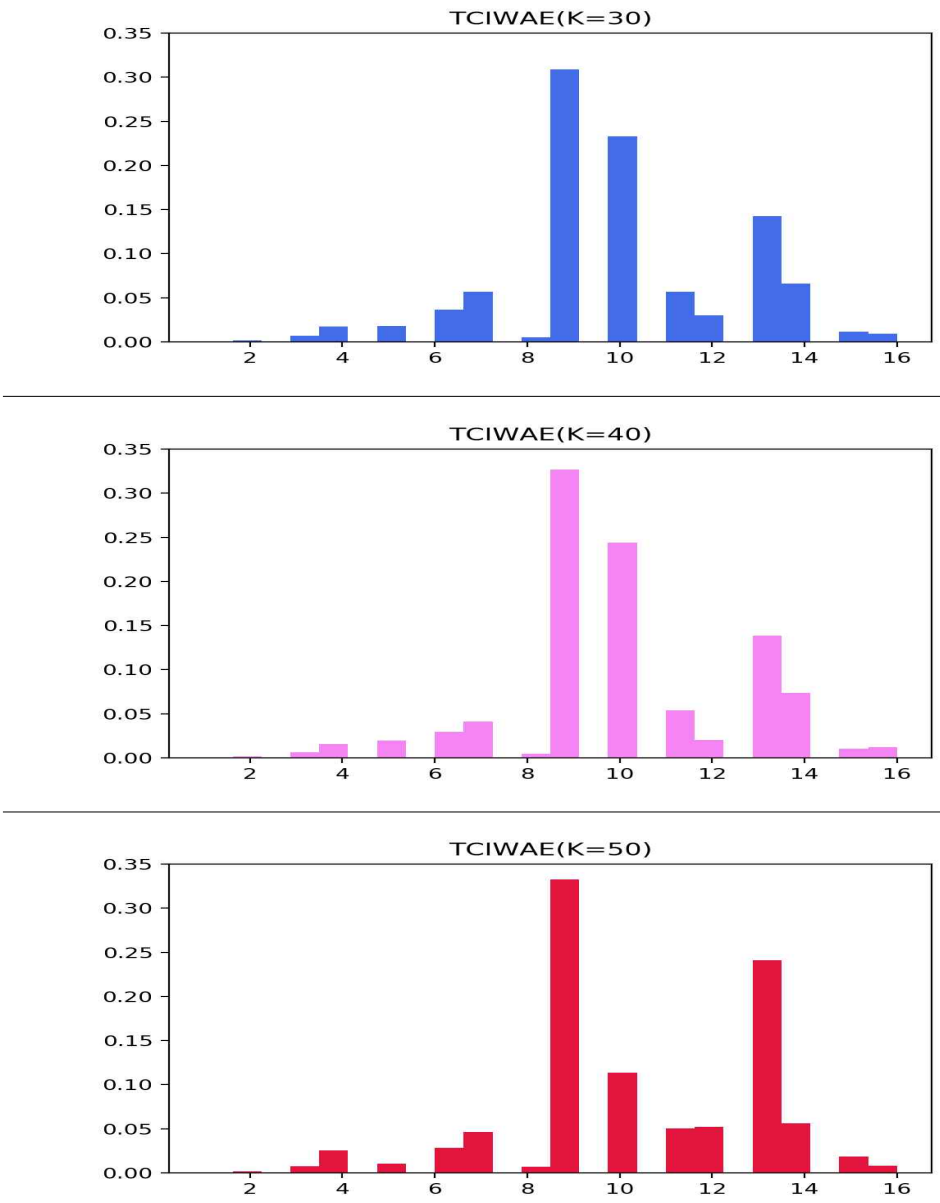
1. 수치형 변수 : ‘educational-num’





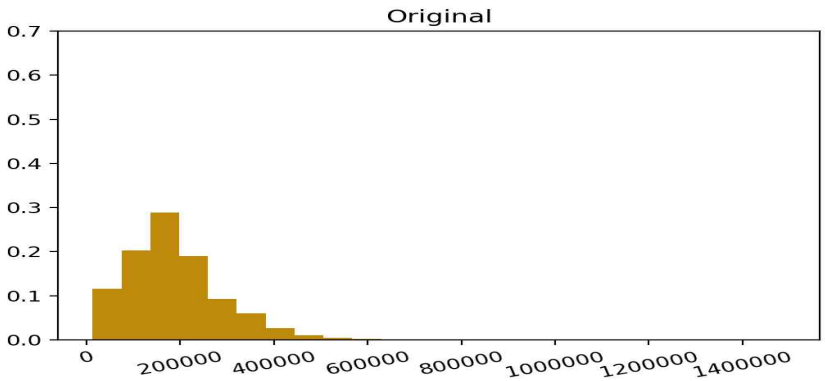


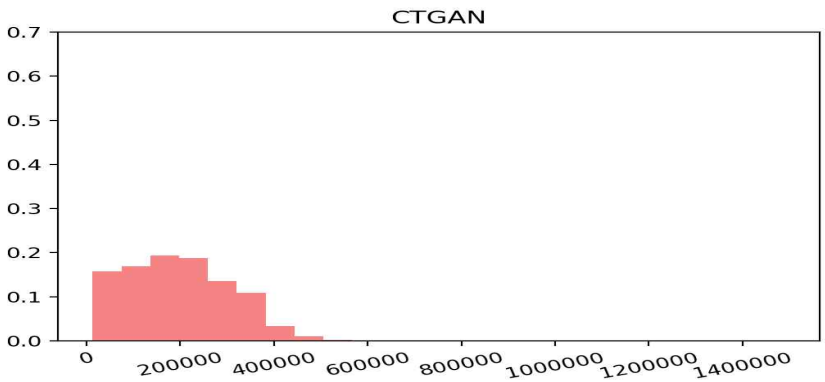
1. 수치형 변수 : ‘educational-num’

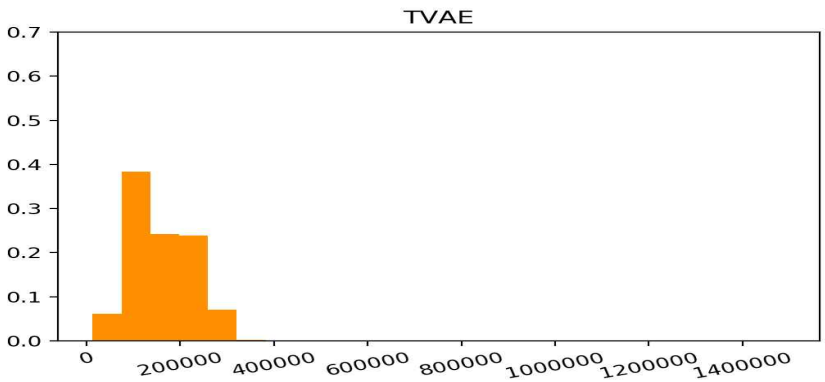


[그림 1] 각 재현 데이터별 ‘educational-num’ 변수의 분포 시각화

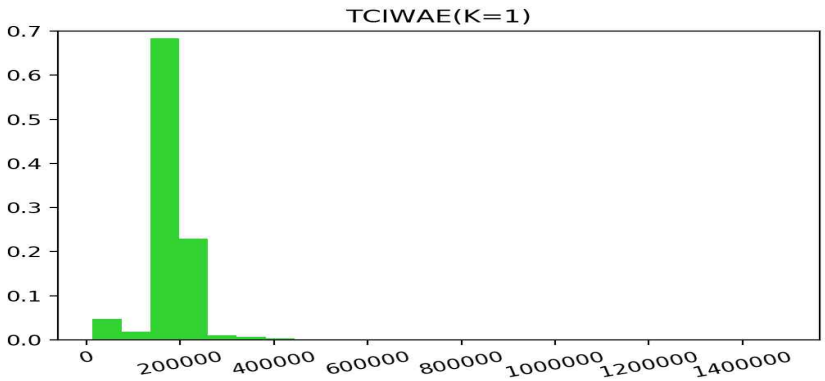
2. 수치형 변수 : ‘fnlwgt’(로그변환)

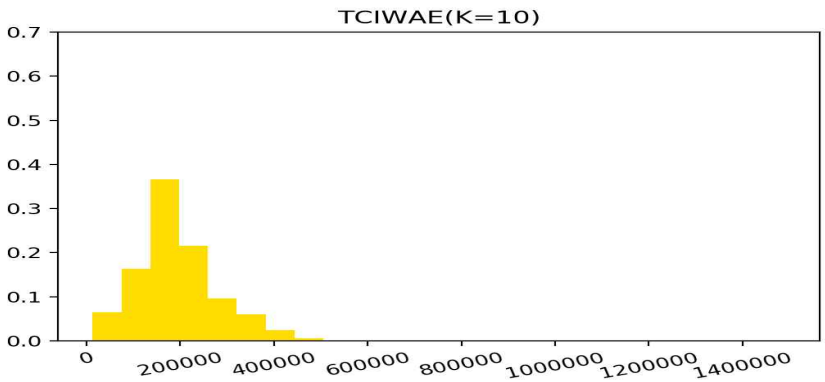


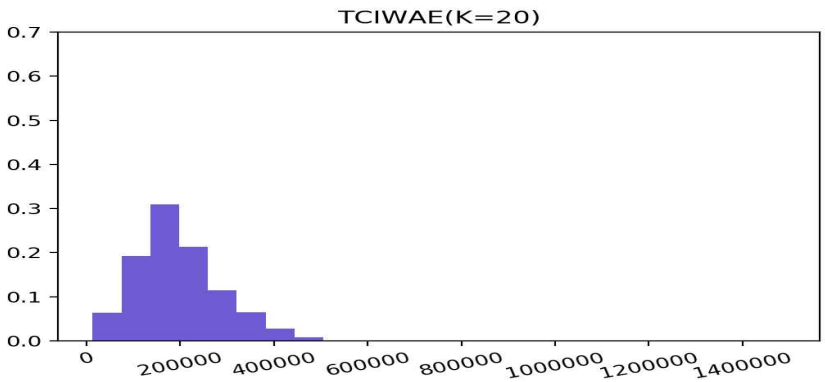




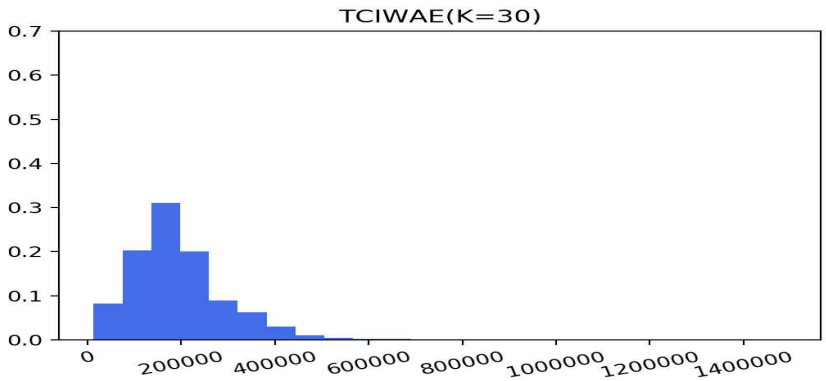
2. 수치형 변수 : ‘fnlwgt’(로그변환)

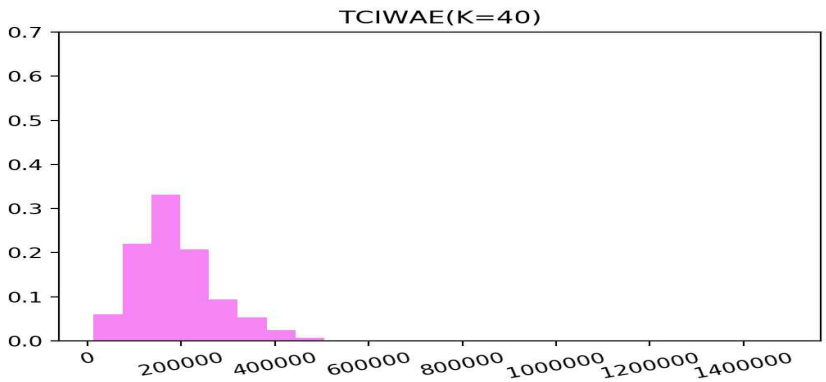


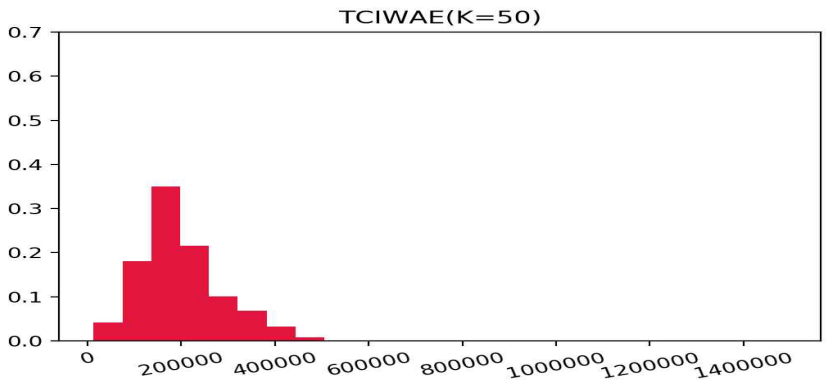




2. 수치형 변수 : ‘fnlwgt’(로그변환)

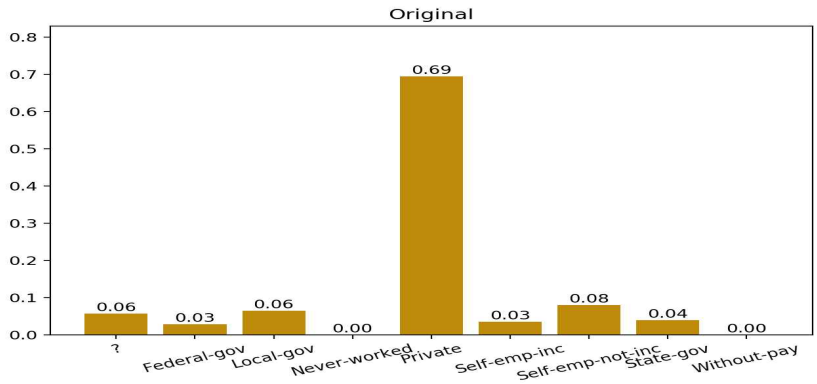


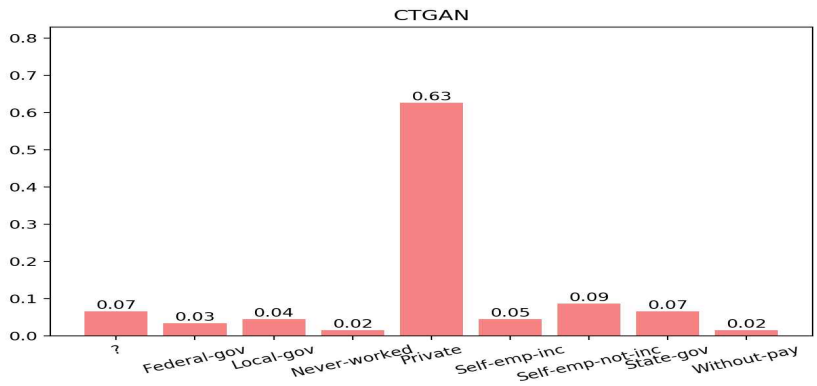


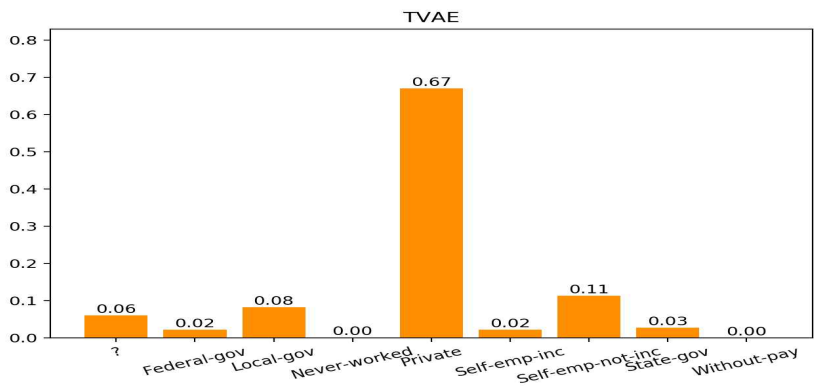


[그림 2] 각 재현 데이터별 ‘fnlwgt’ 변수의 분포 시각화

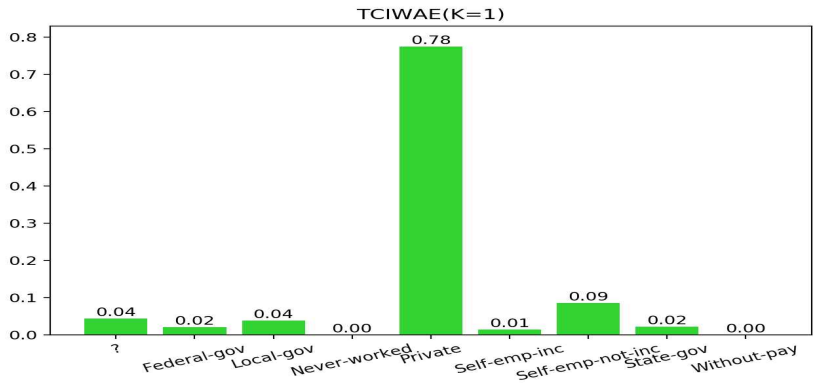
3. 범주형 변수 : ‘workclass’

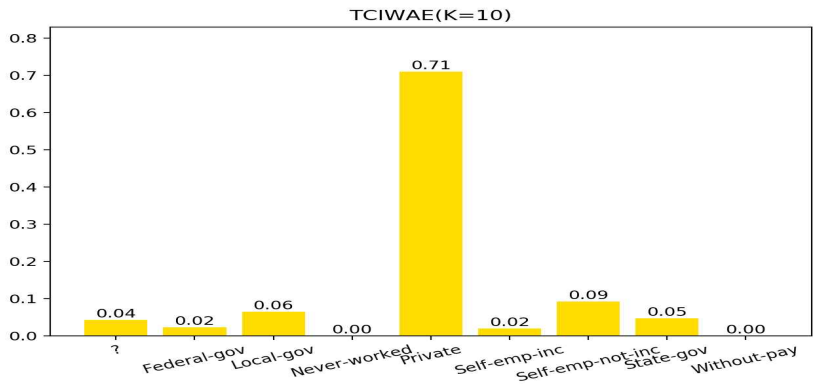


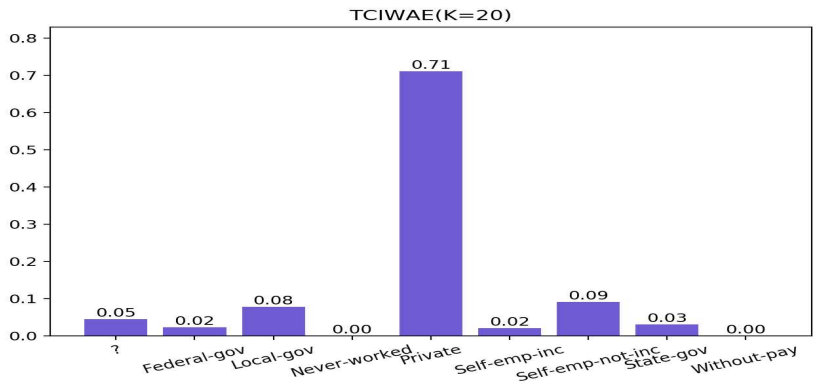




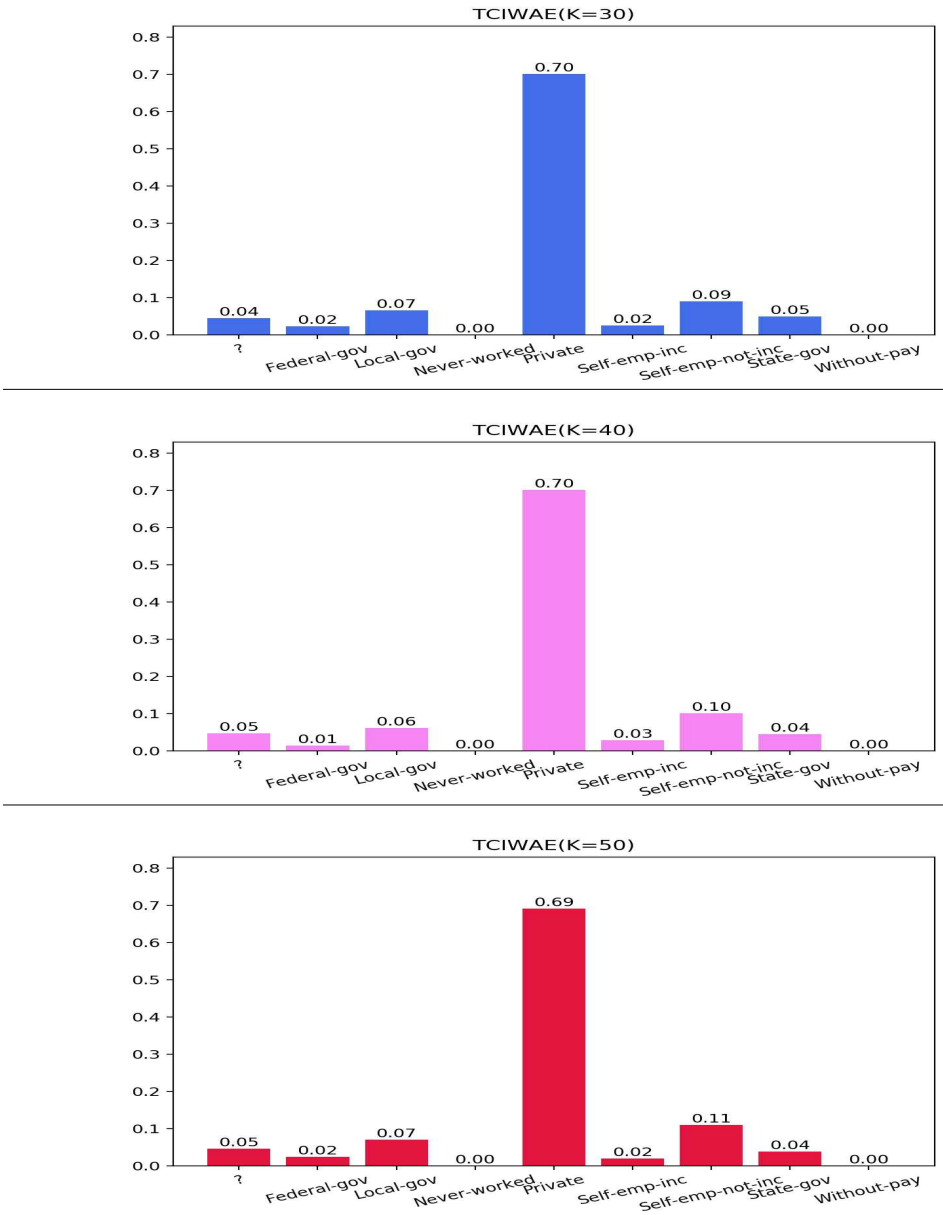
3. 범주형 변수 : ‘workclass’





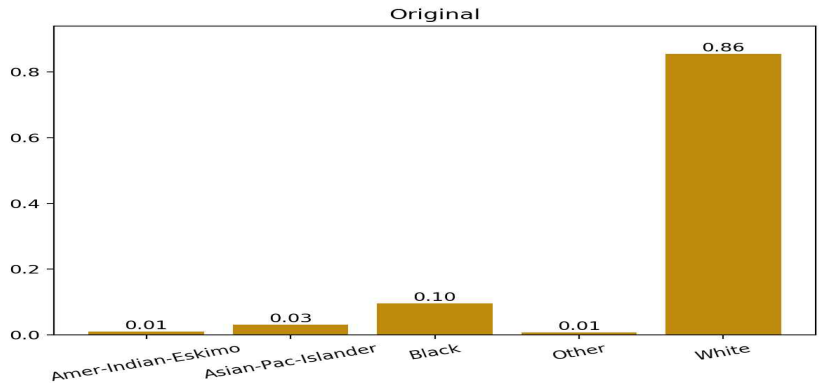


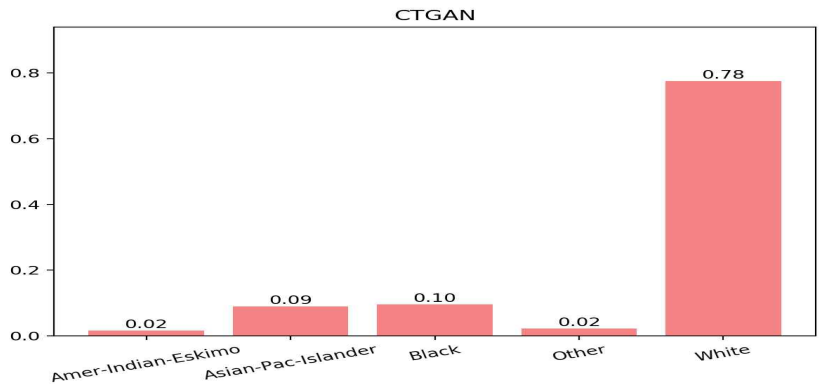
3. 범주형 변수 : ‘workclass’

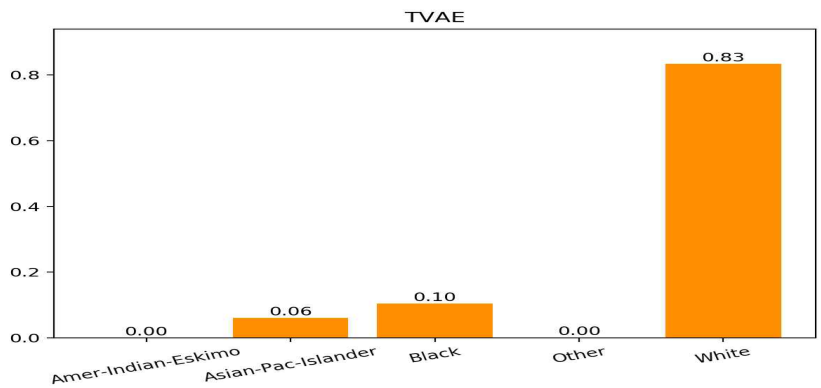


[그림 3] 각 재현 데이터별 ‘workclass’ 변수의 분포 시각화

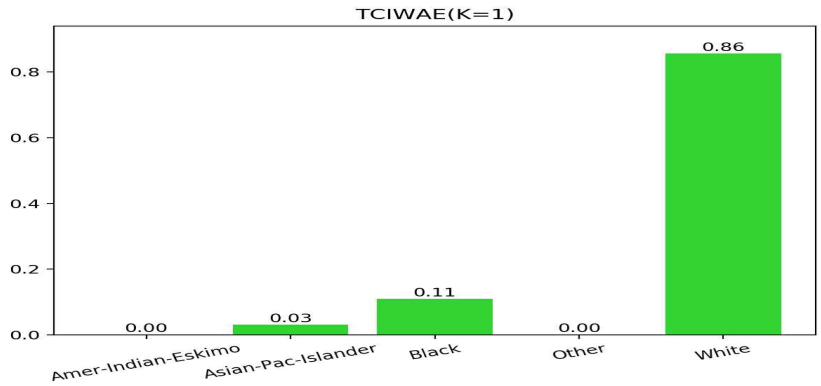
4. 범주형 변수 : ‘race’

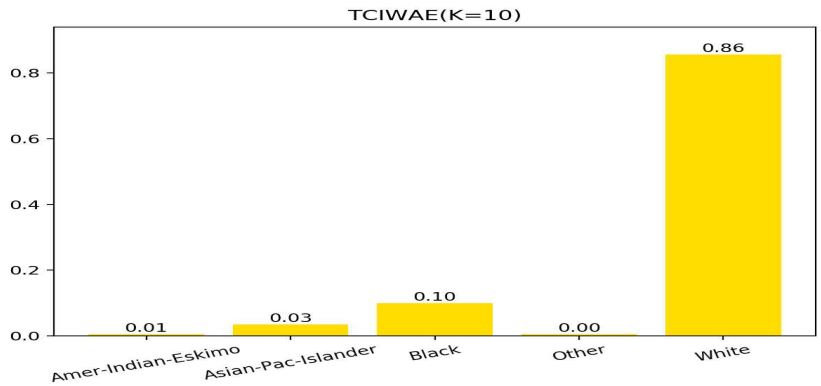

Original
0.8
0.6
0.4
0.2
0.0
0.01
0.03
0.10
0.01
0.86
Amer-Indian-Eskimo
Asian-Pac-Islander
Black
Other
White

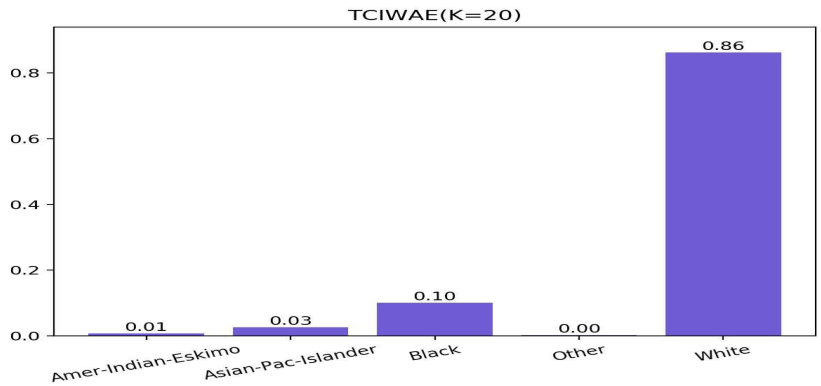

CTGAN
0.8
0.6
0.4
0.2
0.0
0.02
0.09
0.10
0.02
0.78
Amer-Indian-Eskimo
Asian-Pac-Islander
Black
Other
White


TVAE
0.8
0.6
0.4
0.2
0.0
0.00
0.06
0.10
0.00
0.83
Amer-Indian-Eskimo
Asian-Pac-Islander
Black
Other
White

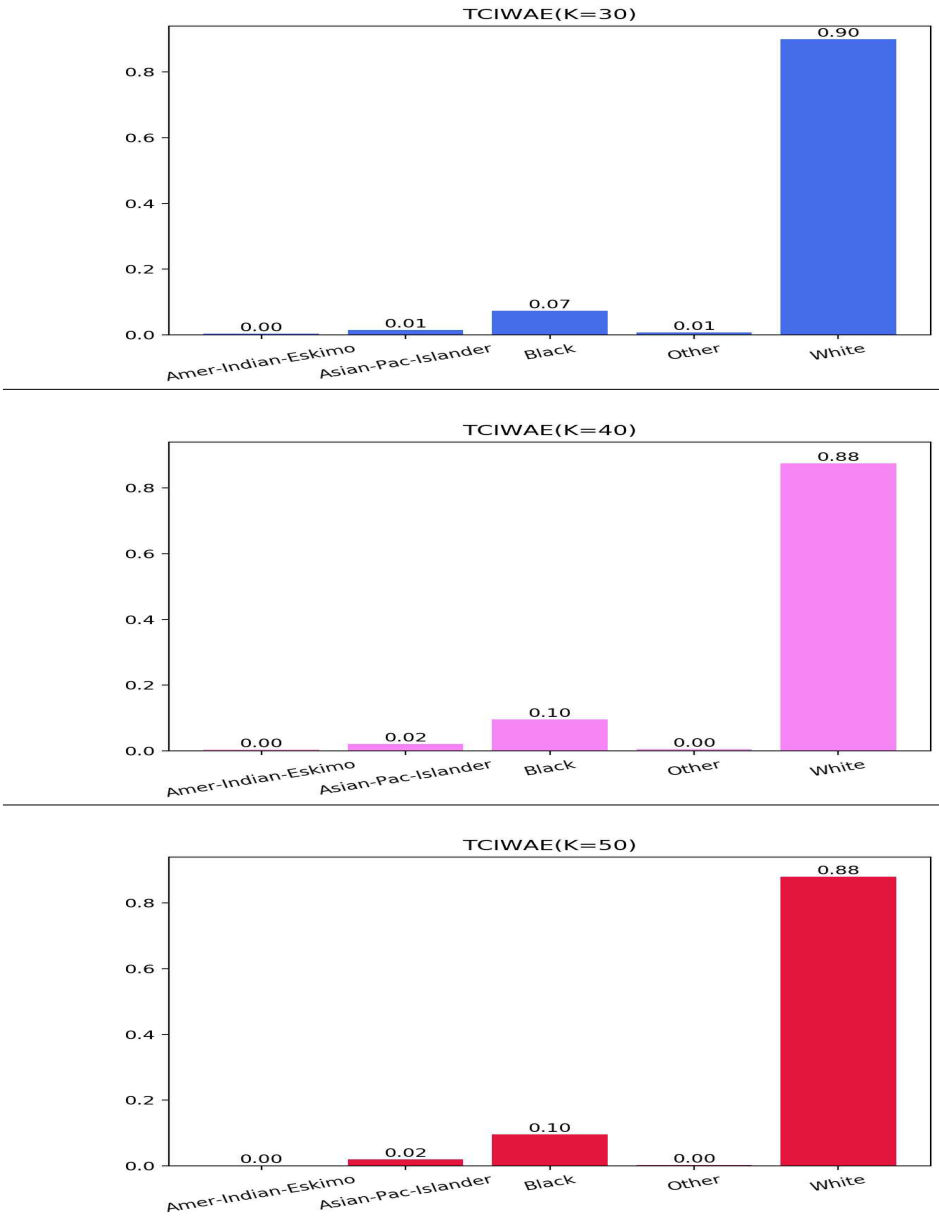
4. 범주형 변수 : ‘race’


TCIWAE(K=1)
0.0
0.2
0.4
0.6
0.8
0.00
0.03
0.11
0.00
0.86
Amer-Indian-Eskimo
Asian-Pac-Islander
Black
Other
White


TCIWAE(K=10)
0.0
0.2
0.4
0.6
0.8
0.01
0.03
0.10
0.00
0.86
Amer-Indian-Eskimo
Asian-Pac-Islander
Black
Other
White


TCIWAE(K=20)
0.0
0.2
0.4
0.6
0.8
0.01
0.03
0.10
0.00
0.86
Amer-Indian-Eskimo
Asian-Pac-Islander
Black
Other
White

4. 범주형 변수 : ‘race’



[그림 4] 각 재현 데이터별 ‘race’ 변수의 분포 시각화

[그림 1]-[그림 4]는 [표 3]에서 언급한 변수의 주변확률분포를 시각화한 결과이다. 원본 데이터의 분포는 'Original'이라고 표기하고, 재현 데이터의 분포는 사용한 방법론의 이름을 적어 나타내었다.

[그림 1]을 살펴보았을 때, 제안 방법인 TCIWAE를 통해 재현한 경우 대체로 원본 분포의 경향성을 따랐으며 아주 낮은 빈도의 값의 생성도 성공적으로 해내었다. 또한, $K=1$일 때보다 K가 1보다 클 때 더 높은 재현 결과를 보여주었다. 반면, 비교 방법인 CTGAN과 TVAE를 통해 재현한 경우는 TCIWAE에 비해 낮은 빈도를 가지는 값의 재현이 잘 되지 않았으며 특정 값의 빈도가 원본 분포의 그것보다 너무 높거나 낮은 경우가 꽤 있었다.

[그림 2]를 살펴보았을 때, 제안 방법인 TCIWAE는 원본 변수의 경향성을 잘 파악하였고 특히 $K=30$인 경우 오른쪽 꼬리 값도 비교 방법론보다 더 잘 재현해냈음을 확인할 수 있다. 반면, 비교 방법인 CTGAN을 통해 재현한 경우는 원본 변수의 최빈 값의 빈도를 잘 재현하지 못하였다. 또한, TVAE를 통해 재현한 경우도 최빈 값의 빈도를 너무 많이 생성하고 오른쪽 꼬리 값도 잘 재현해내지 못했다.

[그림 3]을 살펴보았을 때, 제안 방법인 TCIWAE와 TVAE로 재현한 경우 대체로 원본 변수의 범주 경향성을 반영한 재현을 하였으나 제안 방법이 범주별 정교한 재현을 더욱 잘 해낸 것을 확인할 수 있다. 반면, CTGAN을 통해 재현한 경우는 원본 변수의 최빈 범주의 빈도를 약간 낮게 소수의 범주를 약간 높게 재현하였다.

[그림 4]를 살펴보았을 때, 제안 방법인 TCIWAE으로 재현한 경우 'White' 범주를 원본 빈도만큼 재현하였고 대체로 소수의 범주도 원본의 그것과 비슷하게 재현하였음을 확인할 수 있었다. 반면, CTGAN과 TVAE를 통해 재현한 경우는 'White' 범주는 원본보다 약간 낮게 그 외의 소수의 범주는 약간 높게 재현하였다.

② 유용성 평가지표 : 분포적 유사성

[표 4] Adult 데이터의 분포적 유사성

	평균 W거리	평균 J-S거리	상관성 차이
CTGAN	0.015	0.103	0.913
TVAE	0.018	0.087	0.965
TCIWAE($K=1$)	0.019	0.095	0.988
TCIWAE($K=10$)	0.016	**0.062**	**0.503**
TCIWAE($K=20$)	0.011	0.076	0.605
TCIWAE($K=30$)	0.011	0.067	0.641
TCIWAE($K=40$)	**0.009**	0.070	0.631
TCIWAE($K=50$)	0.017	0.069	0.518

Adult 데이터의 모든 변수마다 재현 변수와 원본 변수 간의 분포 거리를 계산하였다. 이때, 수치형 변수의 분포 거리는 W거리 즉, Wasserstein 거리로 구하고 범주형 변수의 분포 거리는 J-S거리 즉, 잰슨-섀넌 거리로 측정하였다. [표 4]는 모든 수치형 변수 간의 거리 평균, 모든 범주형 변수 간의 거리 평균, 원본 데이터와 재현 데이터의 두 변수별 상관성 차이를 보여준다.

[표 4]의 각 방법론별 평균 W거리의 결과를 통해 제안 방법인 TCIWAE ($K=40$)가 평균 거리가 가장 짧아 원본 데이터의 수치형 변수를 잘 재현하였다고 판단할 수 있었다. 더불어, 평균 J-S거리 결과를 통해 TCIWAE($K=10$)가 평균 거리가 가장 가까워 원본 데이터의 범주형 변수를 잘 재현하였다고 판단할 수 있었다. 또한, 원본 데이터와 재현 데이터 간의 상관성 차이가 가장 작은 TCIWAE($K=10$)가 원본 변수 간의 상관성을 잘 재현해냈다고 결론 지을 수 있었다.

③ 유용성 평가지표 : 머신러닝 성능의 유사성

[표 5] Adult 데이터의 머신러닝 성능의 유사성

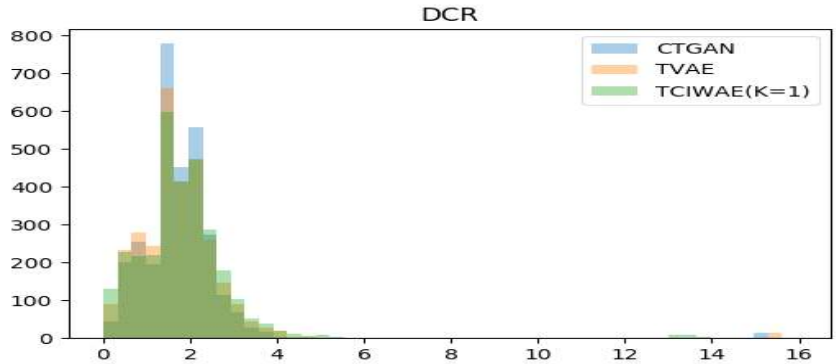
	ACC	AUC	F1-Score
CTGAN	3.022	0.043	0.060
TVAE	4.179	0.057	0.099
TCIWAE($K=1$)	4.077	0.079	0.107
TCIWAE($K=10$)	2.907	0.032	**0.033**
TCIWAE($K=20$)	2.464	**0.027**	0.040
TCIWAE($K=30$)	2.613	**0.027**	0.047
TCIWAE($K=40$)	2.672	0.028	**0.033**
TCIWAE($K=50$)	**2.447**	0.032	0.043

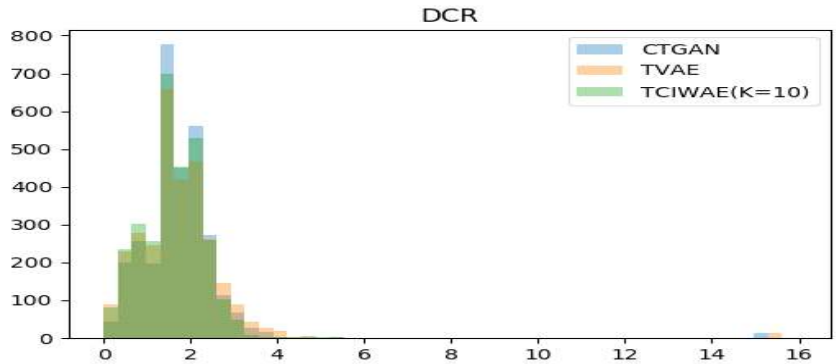
종속 변수인 'income'을 분류하는 원본 데이터로 훈련한 4가지 머신러닝 모델과 재현 데이터로 훈련한 4가지 머신러닝 모델 간의 정확도(ACC), AUC, F1-Score 차이를 평균 낸 값은 [표 5]와 같다. 서로 다른 데이터로 훈련한 두 머신러닝 모델 간의 성능 차이가 작을수록 원본 데이터의 정보를 잘 담고 있는 재현 데이터라고 평가하고자 한다.

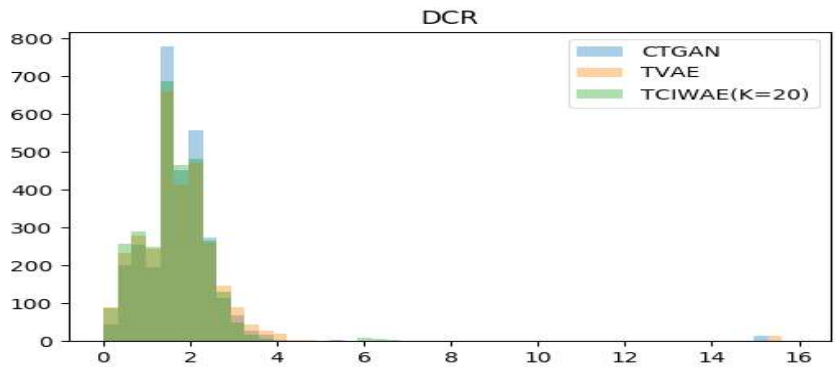
[표 5]의 각 방법론별 머신러닝 성능 결과를 통해 제안 방법인 TCIWAE에서 K가 1보다 클 때 모든 지표에서 비교 방법인 CTGAN, TVAE보다 높은 성능을 보였음을 확인할 수 있었다. 특히, 종속 변수인 'income'은 불균형한 이진 범주로 구성되어 있기 때문에 머신러닝 모델의 성능을 더욱 잘 평가하기 위해서는 F1-Score를 확인하는 것이 좋다. F1-Score 차이는 제안 방법인 TCIWAE($K=10, 30$)인 경우 가장 작아 제안 방법으로 재현한 데이터가 원본 데이터의 유의미한 정보를 잘 재현하였다고 평가할 수 있었다.

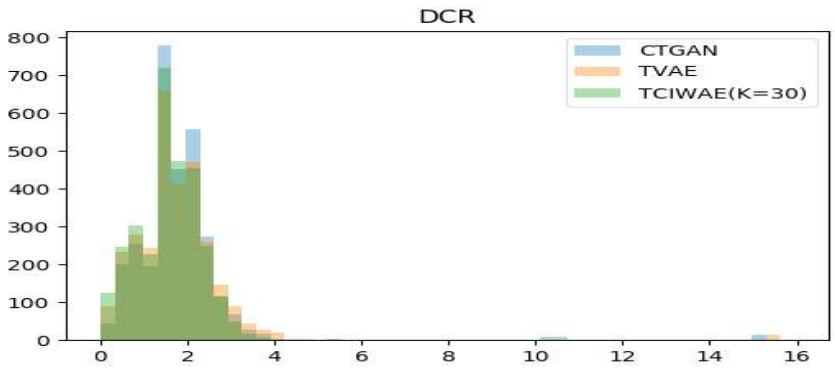
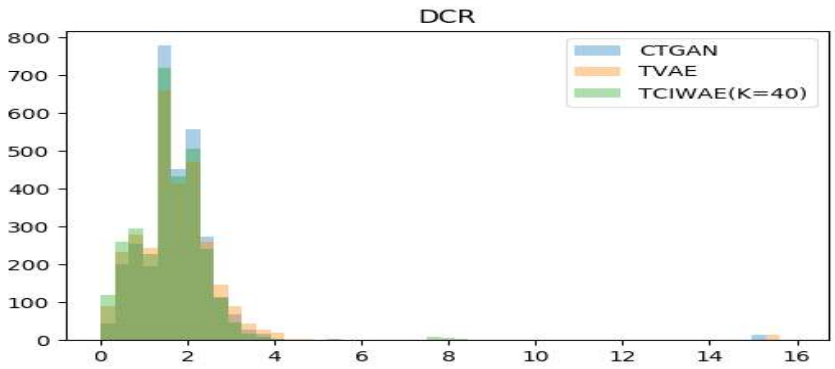
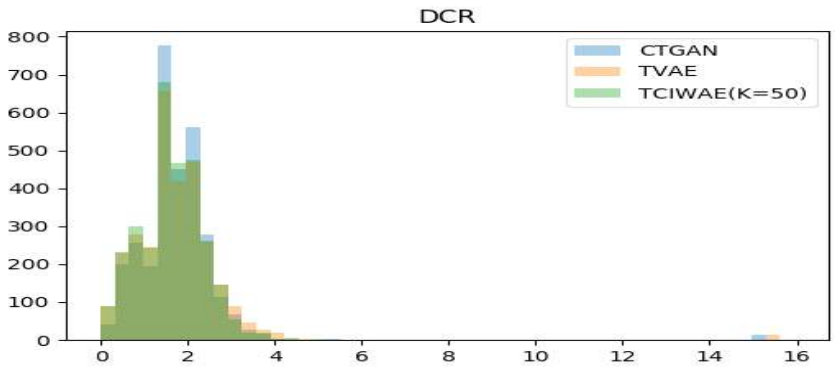
④ 프라이버시 평가지표 : DCR

DCR(Distance to the Closest Record))





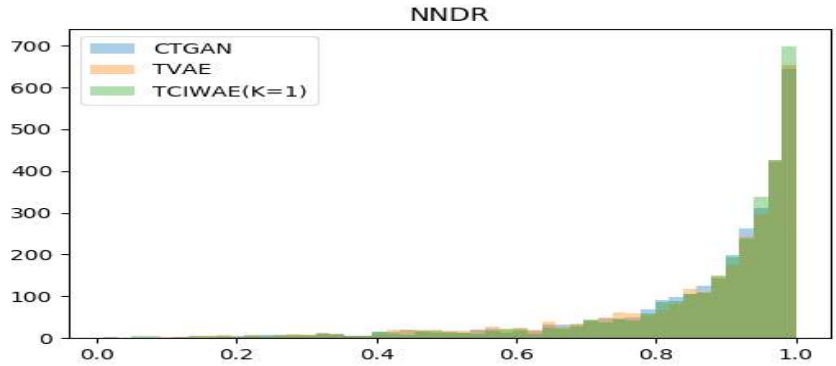


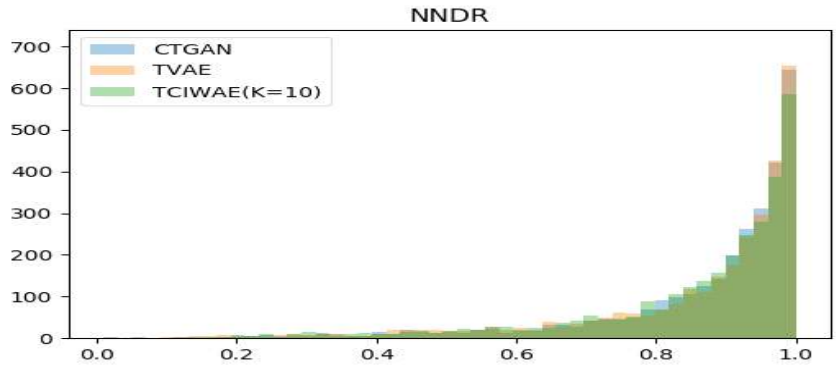
DCR(Distance to the Closest Record))




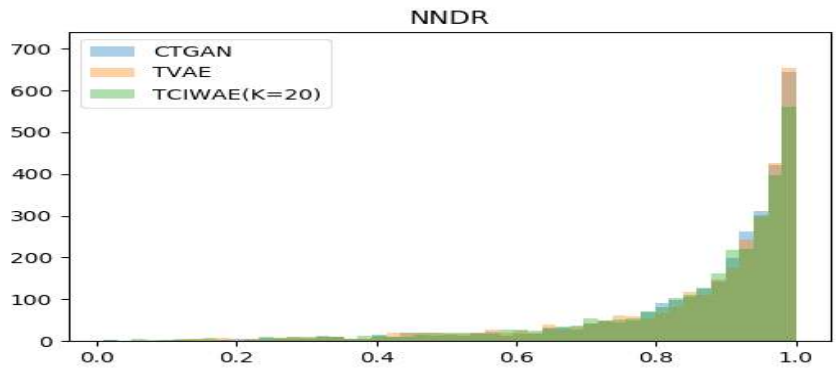
[그림 5] 각 재현 데이터별 DCR의 히스토그램 시각화

⑤ 프라이버시 평가지표 : NNDR

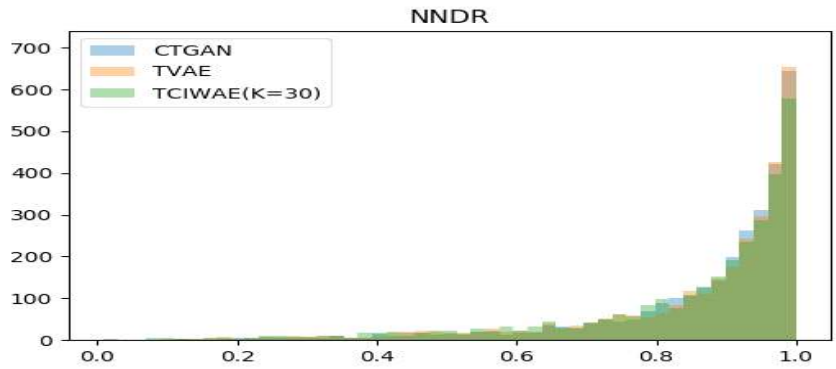
NNDR(Nearest Neighbor Distance Ratio)

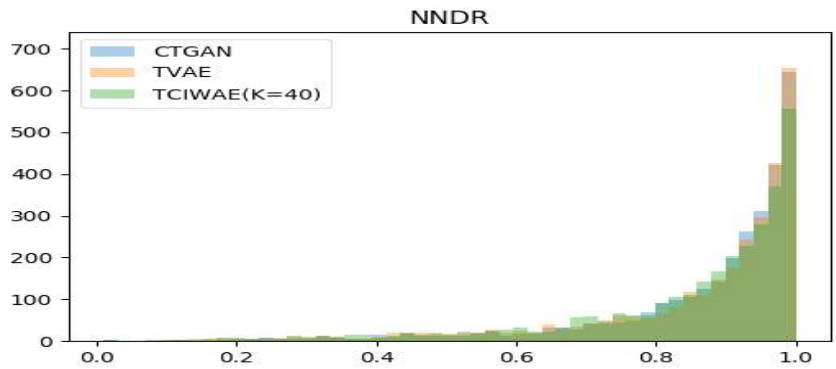


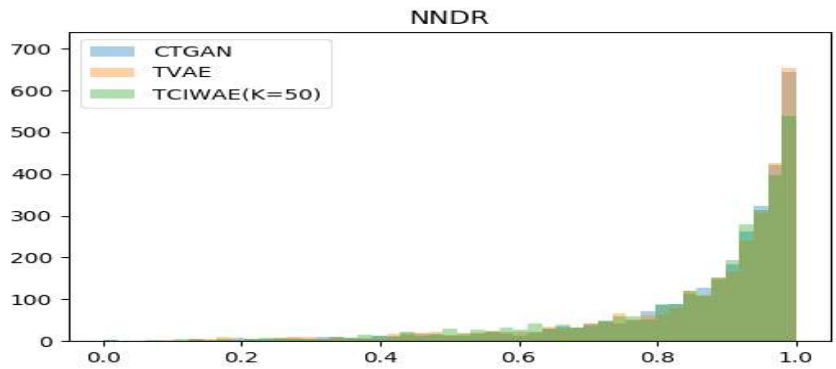




NNDR(Nearest Neighbor Distance Ratio)







[그림 6] 각 재현 데이터별 NNDR의 히스토그램 시각화

[그림 5]-[그림 6]는 제안 방법인 TCIWAE와 비교 방법인 CTGAN, TVAE로 만든 재현 데이터의 프라이버시 노출 위험 정도를 평가하기 위한 지표인 DCR과 NNDR를 계산한 값을 히스토그램으로 그린 결과이다. 시각화 그림의 x축은 DCR 혹은 NNDR의 값을 y축은 DCR 혹은 NNDR의 값을 갖는 빈도를 나타낸다. DCR과 NNDR 모두 작을수록 프라이버시 노출 위험이 큰 것으로 판단한다.

[그림 5]을 살펴보았을 때, 제안 방법인 TCIWAE를 통해 재현한 경우는 모든 K에 대해서 비교 방법인 CTGAN, TVAE와 비슷한 정도의 프라이버시 노출 위험 수준을 보인다고 판단할 수 있었다. [그림 5]에서 DCR 값은 대체로 $0-6$ 사이에 존재하나 비교 방법인 CTGAN과 TVAE에서 일부 값은 14이상의 큰 값을 가지기도 하였다. DCR이 작을수록 프라이버시 노출 위험이 커지나 그렇다고 DCR이 클수록 좋은 것은 아니다. DCR이 클수록 원본 데이터와 이질적인 데이터라는 뜻이며 이는 유용성이 떨어진다는 것을 의미이다. 이를 통해 CTGAN과 TVAE의 유용성이 제안 방법인 TCIWAE보다 낮다는 것도 간접적으로 확인할 수 있었다.

[그림 6]를 살펴보았을 때, 제안 방법인 TCIWAE의 모든 K에서의 NNDR이 비교 방법인 CTGAN, TVAE의 NNDR보다 약간 왼쪽으로 꼬리가 긴 것으로 보이며 이는 프라이버시 노출 위험이 약간 높음을 의미하지만 미미한 수준이라고 평가할 수 있었다.

Adult 데이터를 다양한 표 데이터 재현 방법론을 이용하여 재현한 결과, 제안 방법인 TCIWAE가 유용성 측면에서 비교 방법론인 CTGAN, TVAE보다 모든 지표에서 더 좋은 결과를 보였고 높아진 유용성에 비해 프라이버시 노출 측면에서는 비교 방법보다 약간 높은 수준의 위험도 밖에 보이지 않았다고 판단할 수 있겠다.

2) Covertype 데이터

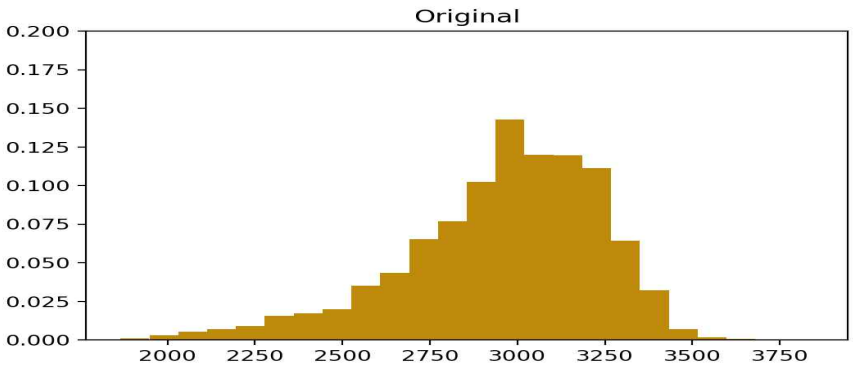
① 유용성 평가지표 : 주변확률분포 시각화

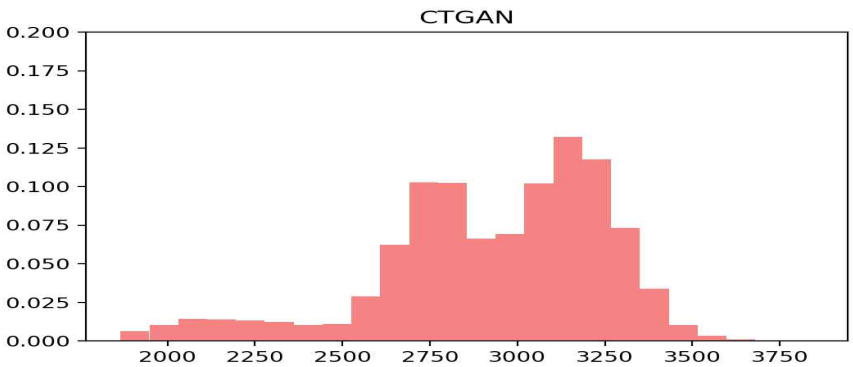
먼저 Covertype 데이터도 변수의 개수가 많아 Adult 데이터와 동일하게 수치형 변수와 범주형 변수 각각 2개씩 선택하여 각 변수의 원본 데이터와 재현 데이터의 주변확률분포를 시각화하였다. 이때, 제안 방법인 TCIWAE를 통해 재현 데이터를 생성하는 과정에서 Covertype 데이터의 수치형 변수는 꼬리가 긴 변수가 없어서 로그 변환 후 MNS를 수행하는 작업을 실시하지 않았다. 사용한 변수는 [표 6]을 통해 알 수 있으며 변수를 시각화한 결과는 [그림 7]-[그림 10]에서 확인할 수 있다.

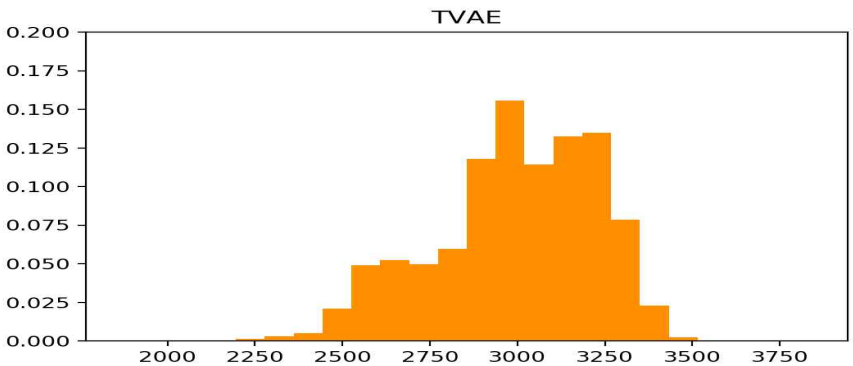
[표 6] Covertype 데이터 분포 시각화 변수

수치형 변수	범주형 변수
‘elevation’	‘wild_area_3’
‘vertical_distance_to_hydrology’	‘Cover_Type’

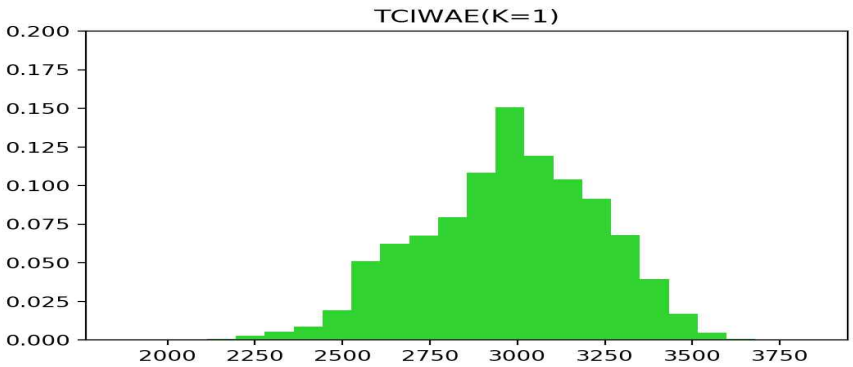
1. 수치형 변수 : ‘elevation’

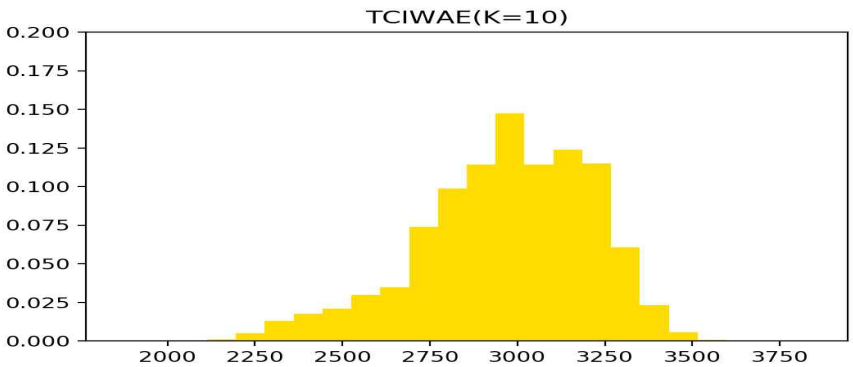


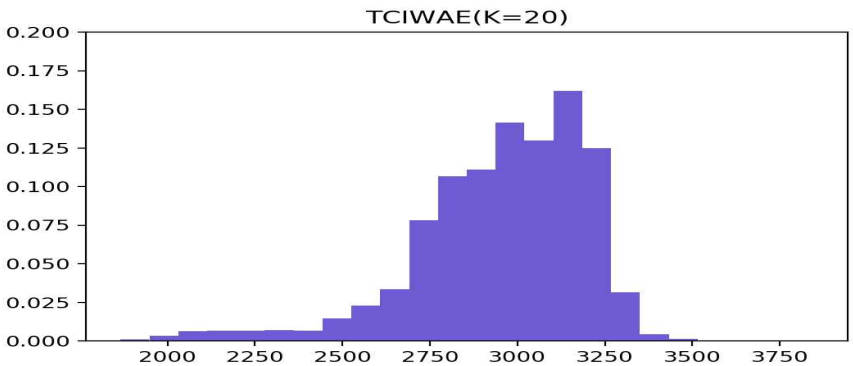




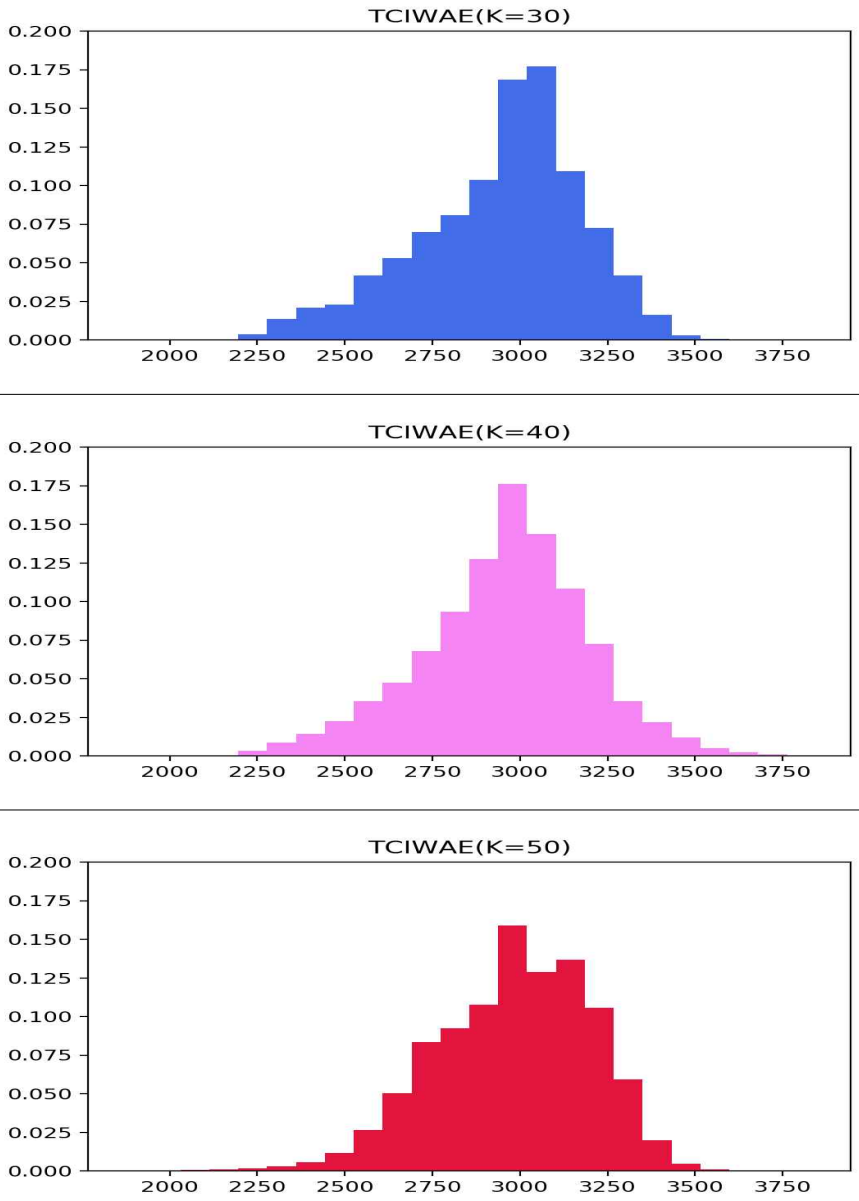
1. 수치형 변수 : ‘elevation’





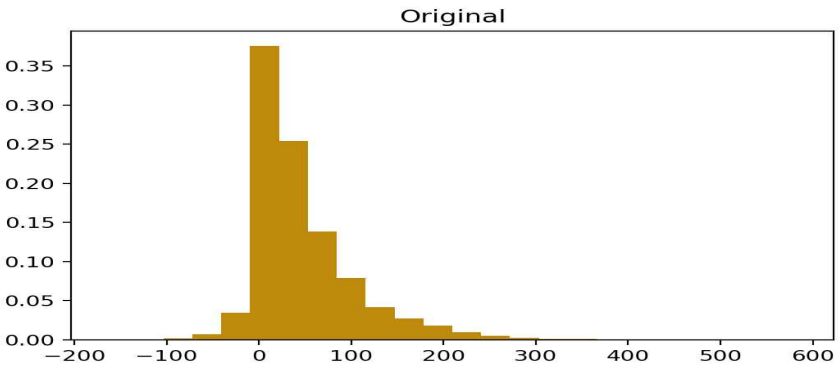


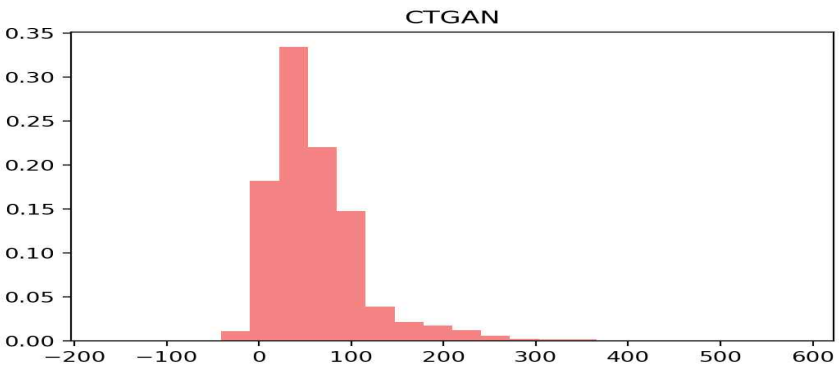
1. 수치형 변수 : ‘elevation’

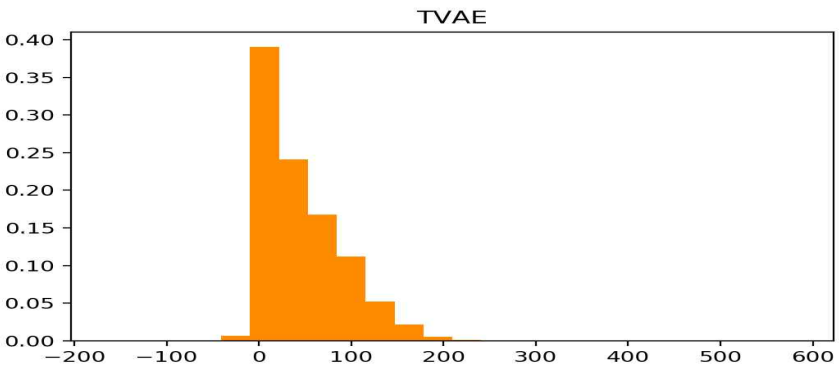


[그림 7] 각 재현 데이터별 ‘elevation’ 변수의 분포 시각화

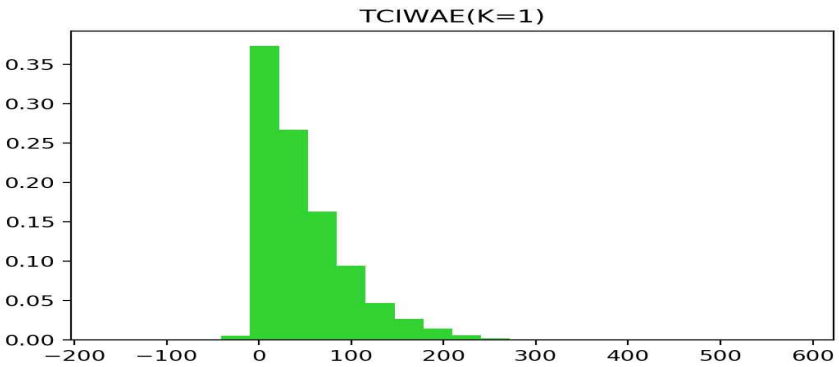
2. 수치형 변수 : ‘vertical_distance_to_hydrology’

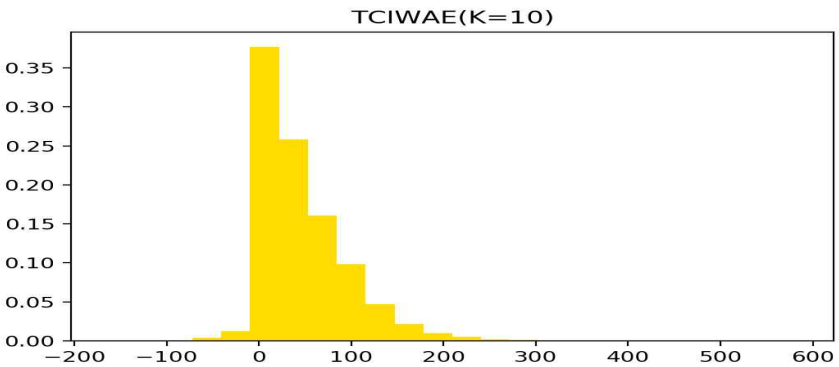


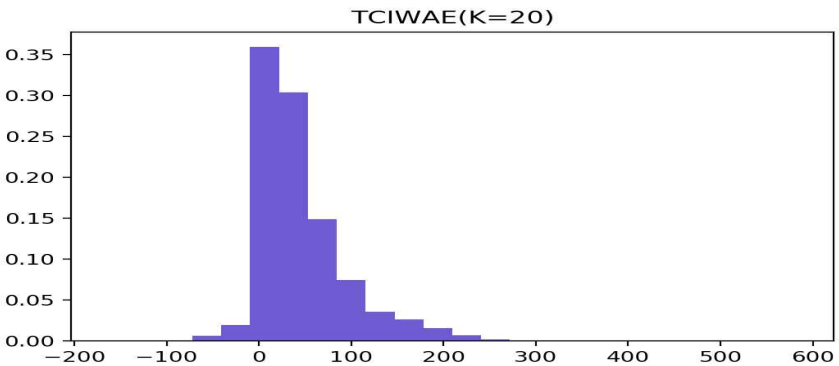




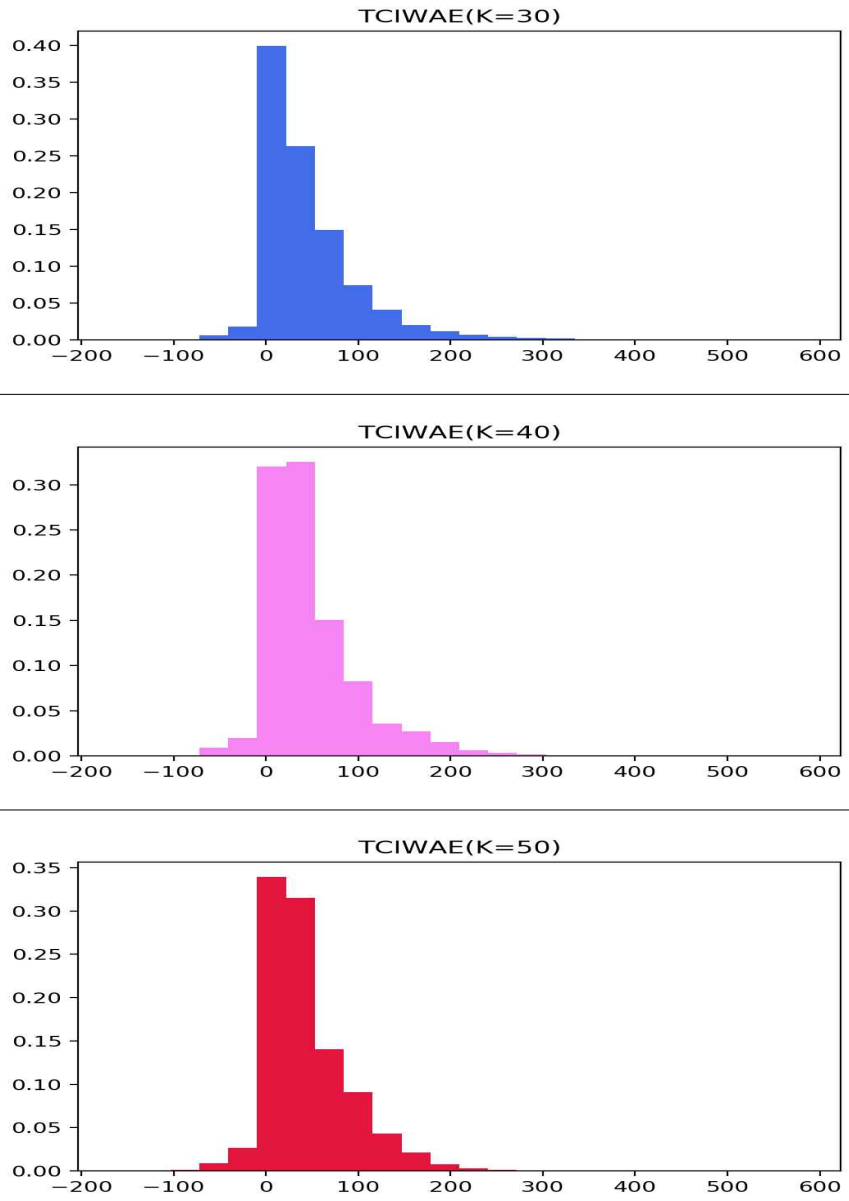
2. 수치형 변수 : ‘vertical_distance_to_hydrology’





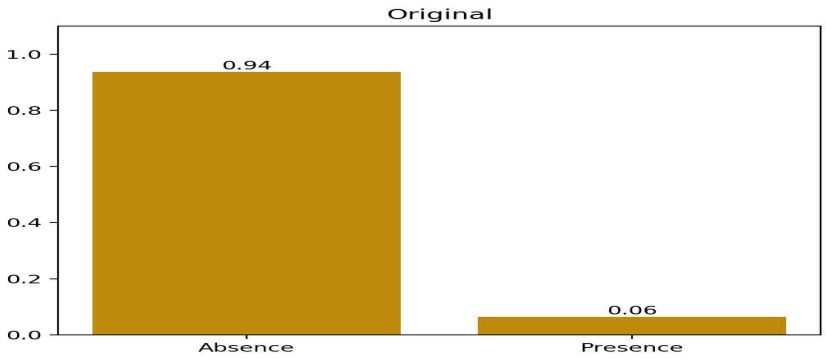


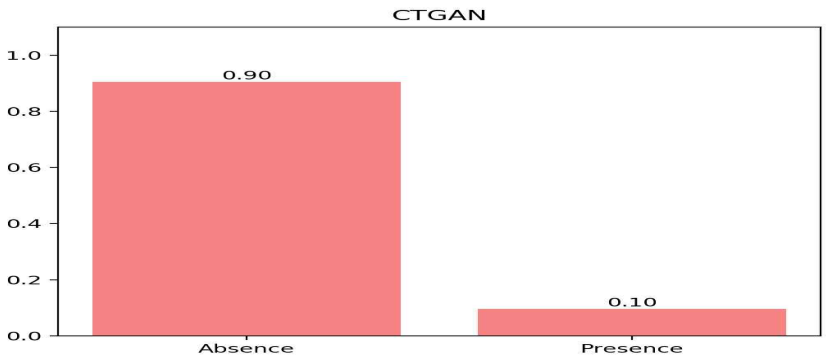
2. 수치형 변수 : ‘vertical_distance_to_hydrology’

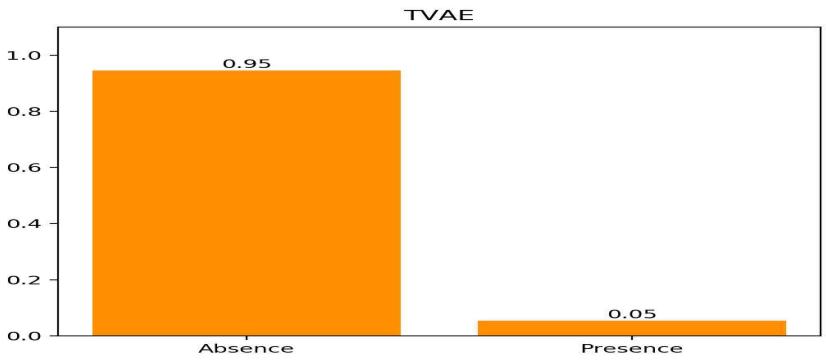


[그림 8] 각 재현 데이터별 ‘vertical_distance’ 변수의 분포 시각화

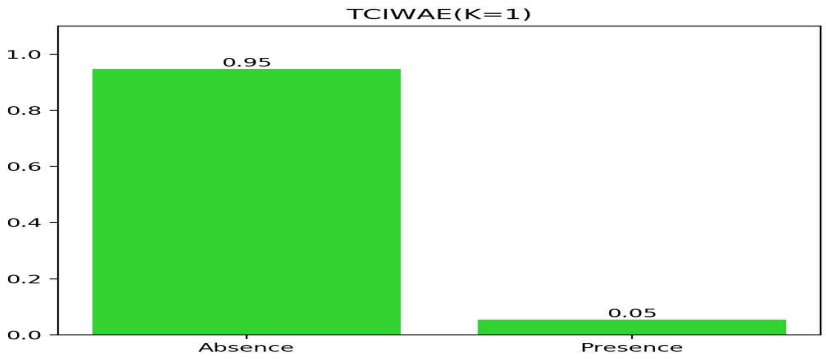
3. 범주형 변수 : ‘wild_area_3’

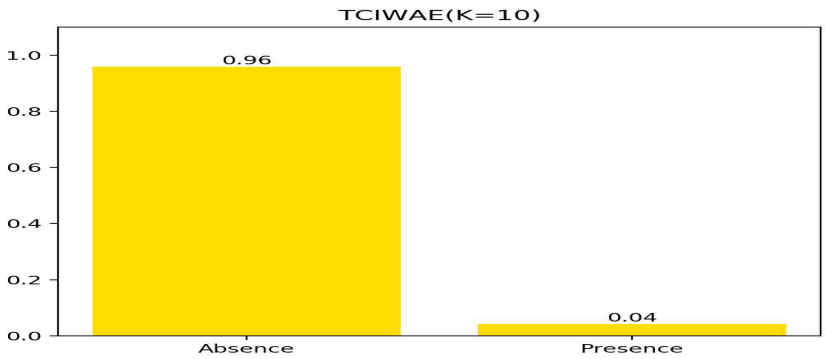


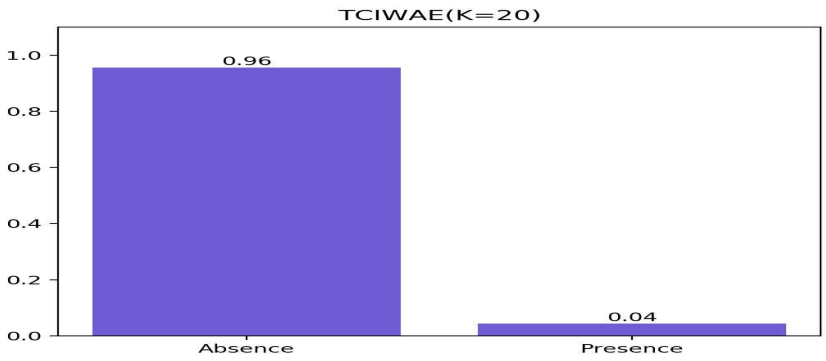




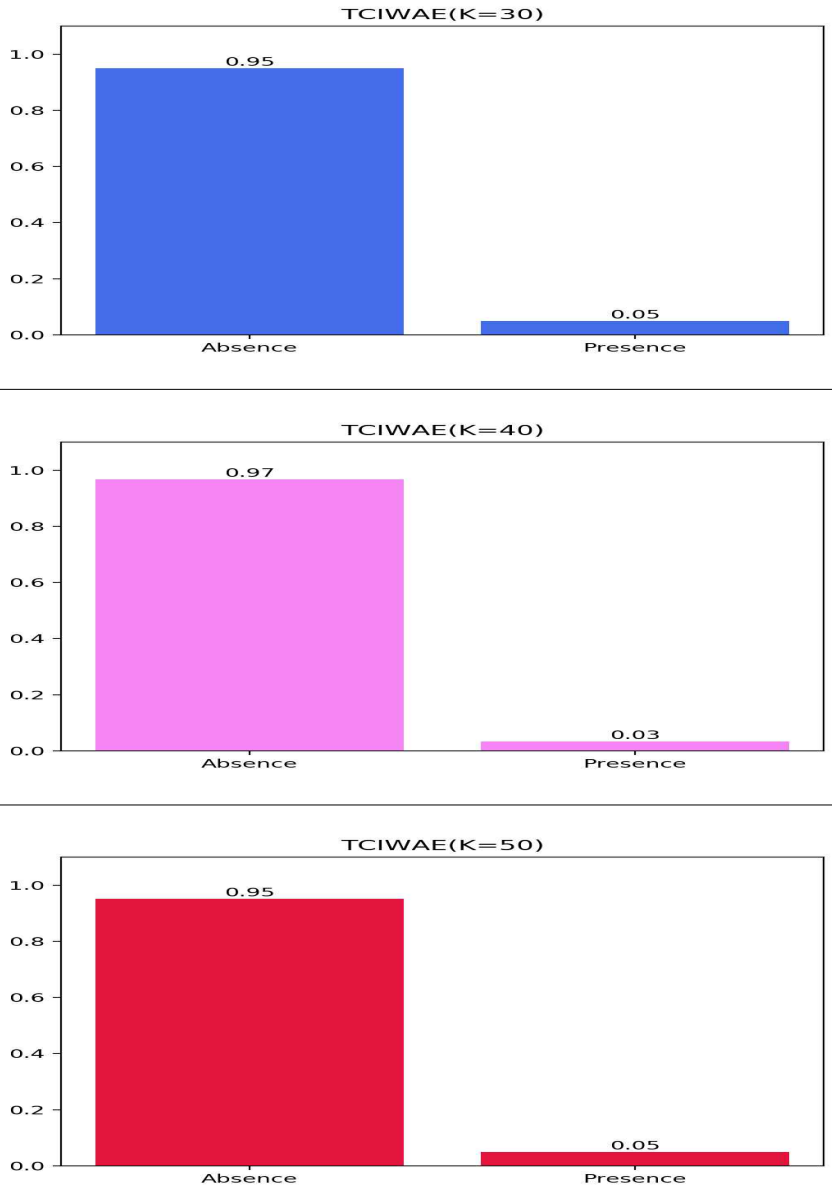
3. 범주형 변수 : ‘wild_area_3’





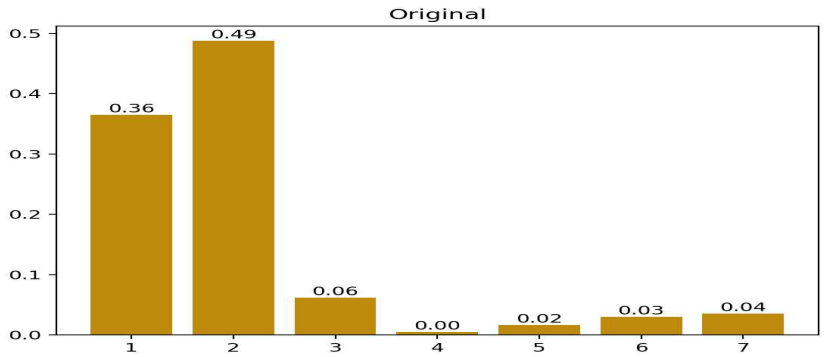


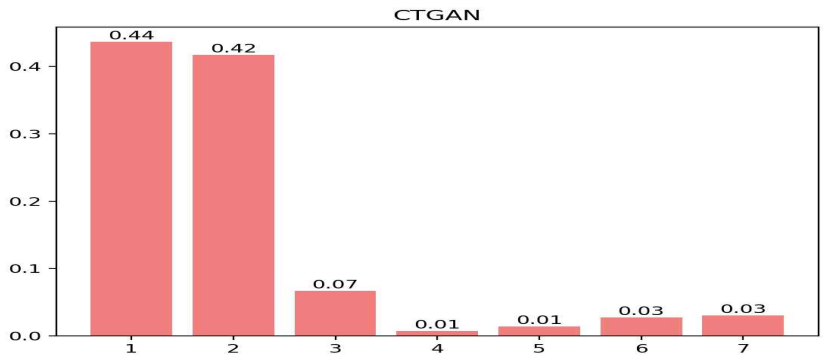
3. 범주형 변수 : ‘wild_area_3’

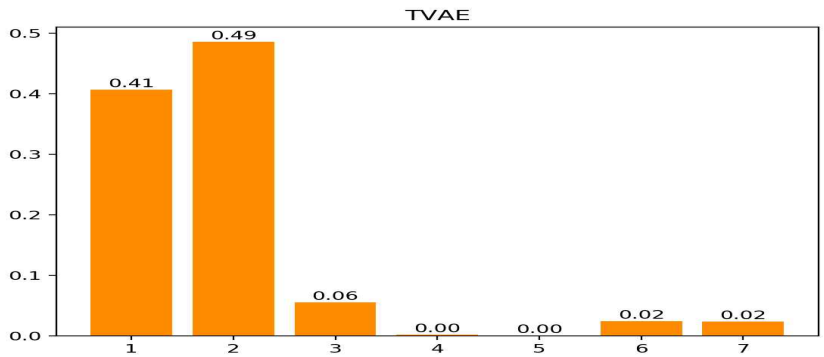


[그림 9] 각 재현 데이터별 ‘wild_area_3’ 변수의 분포 시각화

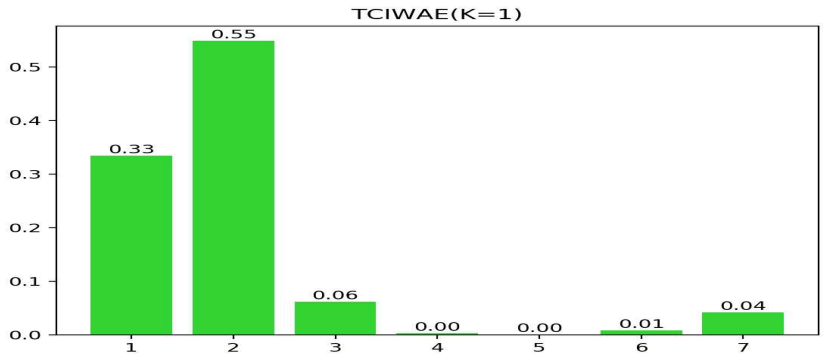
4. 범주형 변수 : 'Cover_Type'

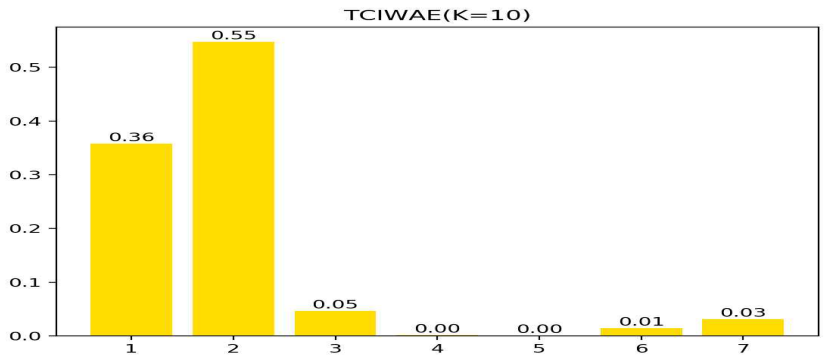


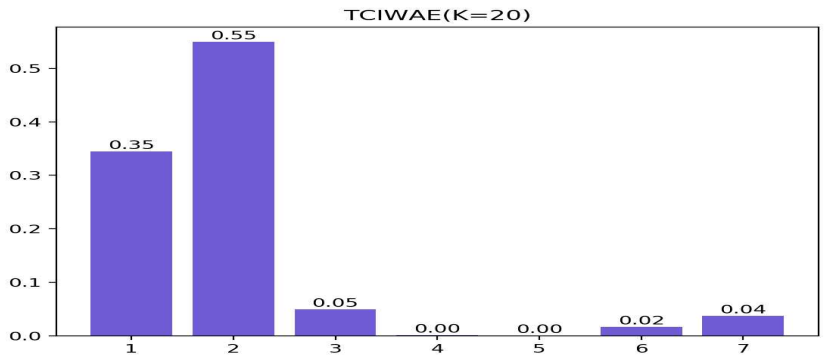




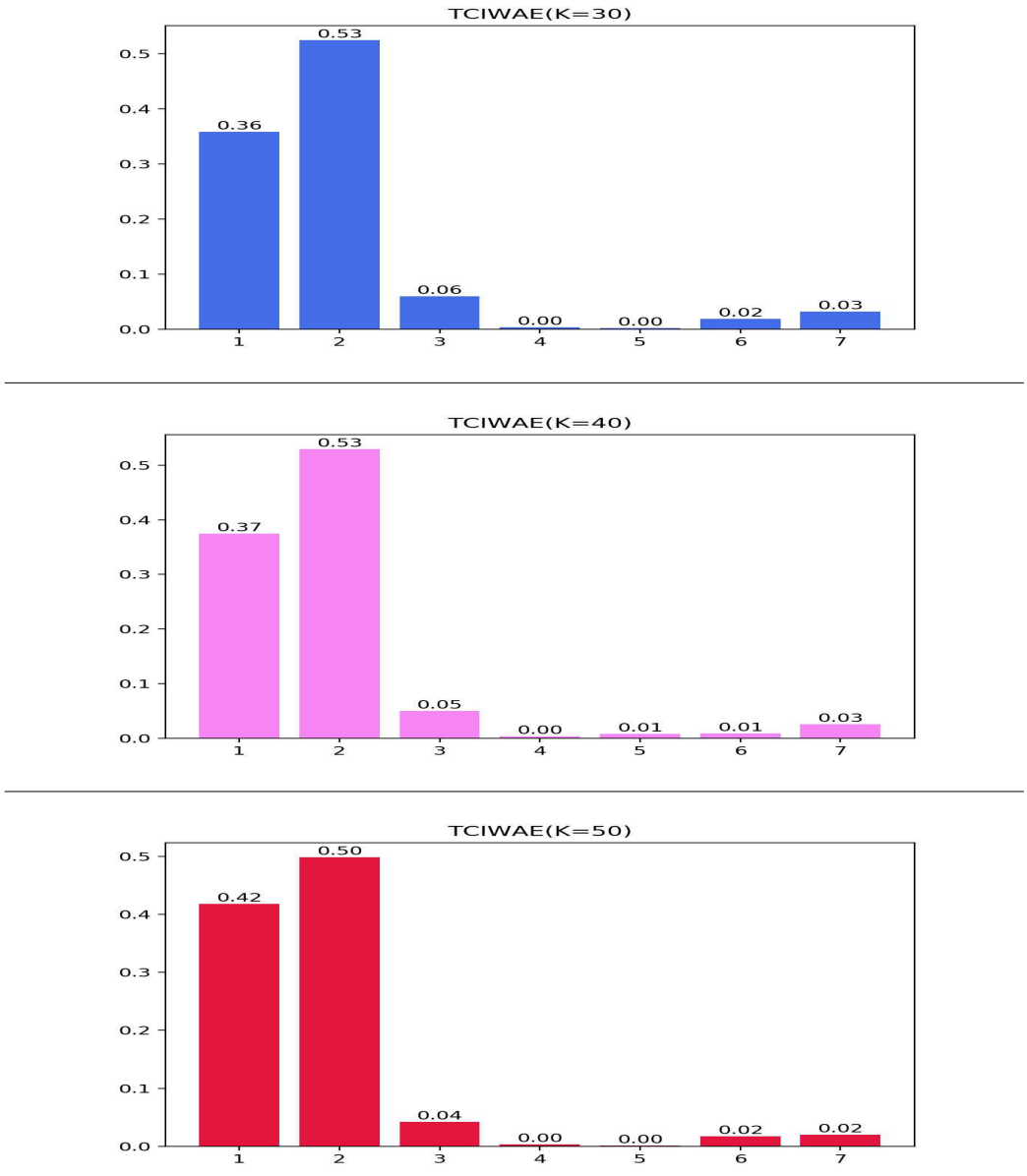
4. 범주형 변수 : ‘Cover_Type’







4. 범주형 변수 : ‘Cover_Type’



[그림 10] 각 재현 데이터별 ‘Cover_Type’ 변수의 분포 시각화

[그림 7]을 살펴보았을 때, 제안 방법인 TCIWAE를 통해 재현한 경우 원본 분포의 최빈 값과 왼쪽으로 꼬리가 긴 값의 생성을 잘 해내었다. 특히, $K=50$ 일 때 원본 변수의 분포가 가장 비슷한 분포 모습을 보였다. 반면, 비교 방법인 CTGAN을 통해 재현한 경우는 왼쪽으로 꼬리가 긴 값을 재현해내면서 가장 최빈 값은 원본 확률보다 낮게 재현하여 원본의 경향성이 왜곡되었다. 또한, TVAE를 통해 재현한 경우는 왼쪽으로 꼬리가 긴 값의 재현이 잘 되지 못하였다.

[그림 8]을 살펴보았을 때, 제안 방법인 TCIWAE를 통해 재현한 경우 K가 1보다 클 때 대체로 왼쪽과 오른쪽 꼬리의 소수의 값 재현을 잘할뿐더러 최빈 값도 원본 변수와 비슷한 확률로 생성하였다. 반면, 비교 방법인 CTGAN을 통해 재현한 경우는 최빈 값이 원본 변수와 다른 값에서 생성되었다. 또한, TVAE를 통해 재현한 경우는 최빈 값은 잘 재현하였으나 왼쪽과 오른쪽으로 꼬리가 긴 값의 재현을 잘 하지 못했다.

[그림 9]을 살펴보았을 때, 제안 방법인 TCIWAE에서 $K=10, 20$ 일 때 원본 변수의 이진 범주 각각의 빈도를 정확하게 재현하였다. 반면, CTGAN을 통해 재현한 경우는 'Presence'의 범주를 원본 범주의 빈도보다 더 많이 생성하였다. TVAE를 통해 재현한 경우는 원본 범주 빈도와 비슷하게 재현하였으나 가장 정확한 재현은 제안 방법인 TCIWAE를 통해 얻었다.

[그림 10]을 살펴보았을 때, 제안 방법인 TCIWAE을 통해 재현한 경우 K가 1보다 클 때 원본 변수의 범주별 확률에 비슷한 값을 생성하였다. 반면, 비교 방법인 CTGAN을 통해 재현한 경우는 낮은 빈도의 범주는 잘 재현하였지만 높은 빈도의 범주 값 생성이 잘 되지 않았다. TVAE를 통해 재현한 경우는 대체로 원본 분포를 따르고 있으나 가장 정교한 재현은 제안 방법인 TCIWAE를 통해 얻었다.

② 유용성 평가지표 : 분포적 유사성

[표 7] Covertype 데이터의 분포적 유사성

	평균 W거리	평균 J-S거리	상관성 차이
CTGAN	0.033	0.047	3.534
TVAE	0.022	0.020	2.254
TCIWAE($K=1$)	0.022	**0.018**	2.224
TCIWAE($K=10$)	0.019	0.025	2.284
TCIWAE($K=20$)	0.021	0.019	**2.154**
TCIWAE($K=30$)	**0.018**	0.021	2.398
TCIWAE($K=40$)	0.025	0.024	2.570
TCIWAE($K=50$)	0.023	0.023	2.259

다음으로 Covertype 데이터의 분포적 유사성은 Adult 데이터의 그것과 동일하게 모든 변수마다 재현 변수와 원본 변수 간의 분포 거리와 각 데이터별 두 변수 간 상관성을 계산하였다. 이를 [표 7]와 같이 나타내었다.

[표 7]의 각 방법론별 평균 W거리의 결과를 통해 제안 방법인 TCIWAE가 모든 K에 대해서 비교 방법론보다 낮거나 같았으며 특히, K가 30일 때 평균 거리가 가장 짧아 수치형 변수 재현 성능이 가장 높음을 확인할 수 있었다. 더불어, 평균 J-S거리 결과를 통해 TCIWAE($K=1$)가 평균 거리가 가장 가까워 범주형 변수 재현 성능이 가장 높다고 판단할 수 있었다. 또한, 상관성 차이 결과를 통해 TCIWAE($K=20$)가 원본 변수 간의 상관성을 다른 방법론에 비해 잘 재현해냈다고 결론지을 수 있었다.

③ 유용성 평가지표 : 머신러닝 성능의 유사성

[표 8] Covertype 데이터의 머신러닝 성능의 유사성

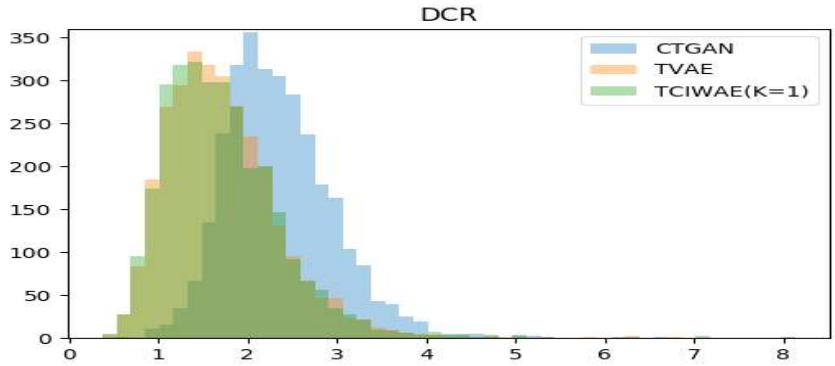
	ACC	AUC	F1-Score
CTGAN	37.714	0.317	0.401
TVAE	17.913	0.117	0.208
TCIWAE($K=1$)	13.192	0.098	0.139
TCIWAE($K=10$)	16.271	0.119	0.162
TCIWAE($K=20$)	13.692	0.094	0.143
TCIWAE($K=30$)	13.130	0.094	0.139
TCIWAE($K=40$)	**12.406**	**0.090**	**0.129**
TCIWAE($K=50$)	14.057	0.095	0.153

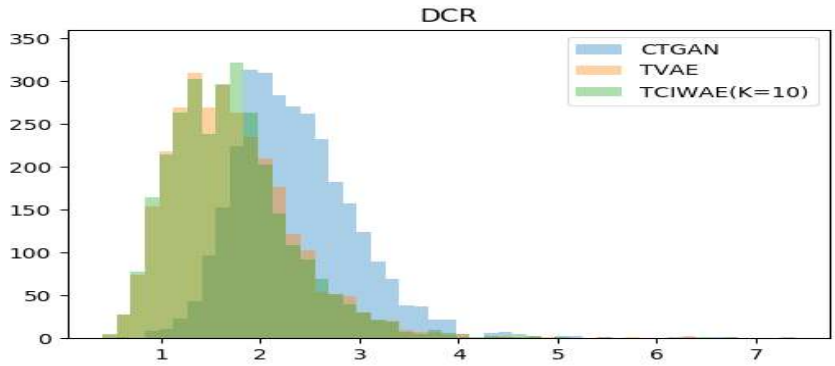
종속 변수인 'Cover_Type'을 분류하는 원본 데이터로 훈련한 4가지 머신러닝 모델과 재현 데이터로 훈련한 4가지 머신러닝 모델 간의 정확도(ACC), AUC, F1-Score 차이를 평균 낸 값은 [표 8]와 같다.

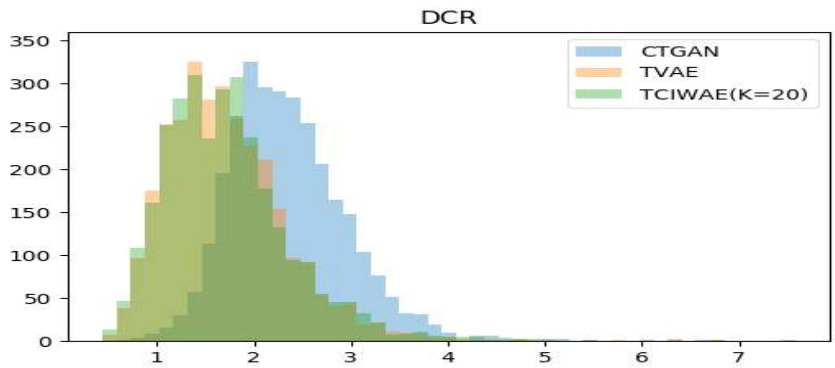
[표 8]의 각 방법론별 머신러닝 성능 결과를 통해 제안 방법인 TCIWAE가 AUC를 제외한 모든 지표에서 비교 방법인 CTGAN, TVAE보다 뛰어난 성능을 보였음을 확인할 수 있었다. 특히, 제안 방법 TCIWAE($K=40$)를 통해 재현한 데이터로 훈련한 머신러닝 모델의 성능이 ACC, AUC, F1-Score에서 원본 데이터로 훈련한 머신러닝 모델의 성능과 가장 유사하였다. 이를 통해 TCIWAE($K=40$)로 재현한 데이터가 다른 방법론에 비해 유용성이 높다고 평가할 수 있었다.

④ 프라이버시 평가지표 : DCR

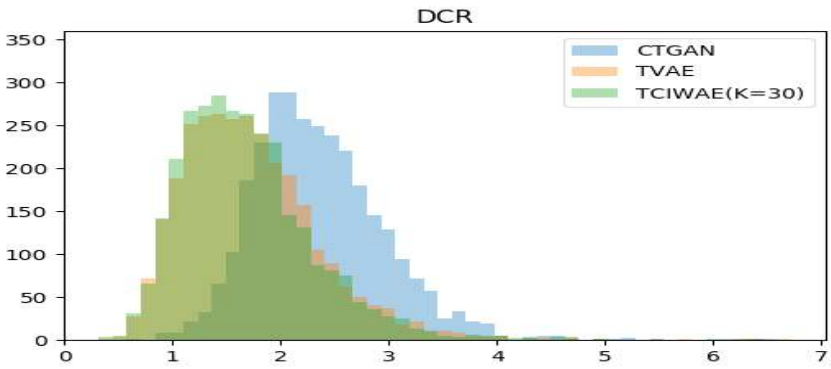
DCR(Distance to the Closest Record)

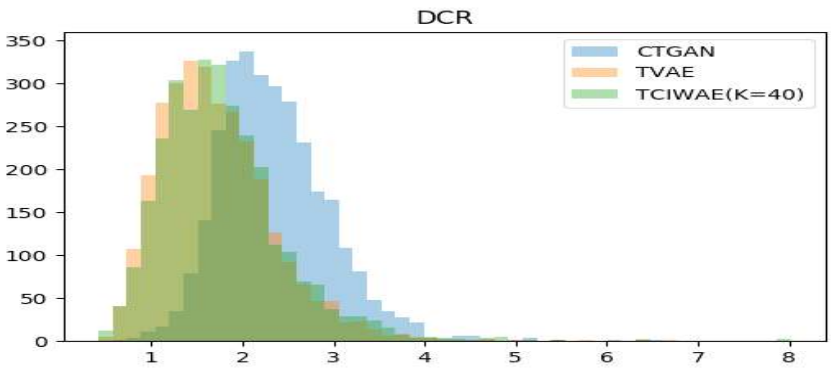


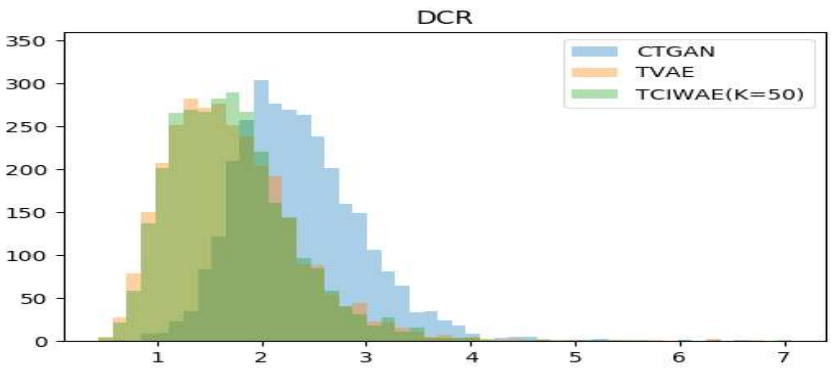




DCR(Distance to the Closest Record))



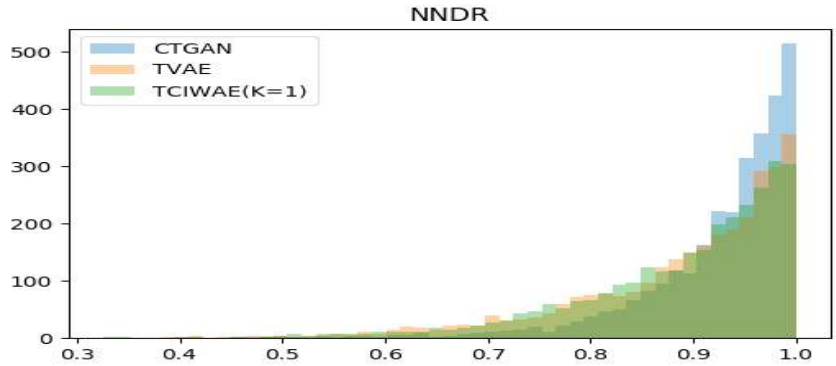


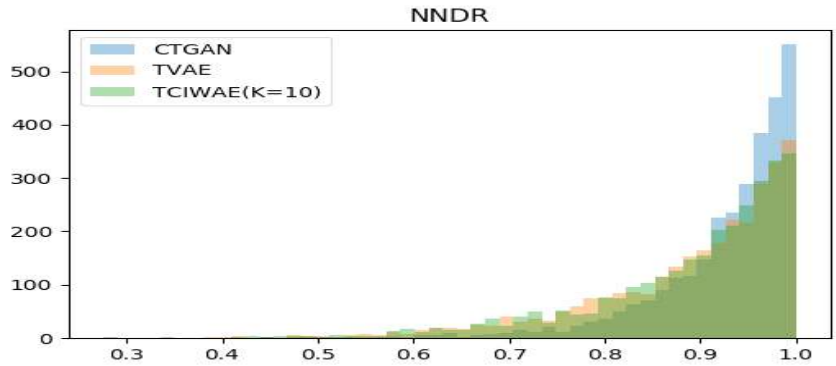


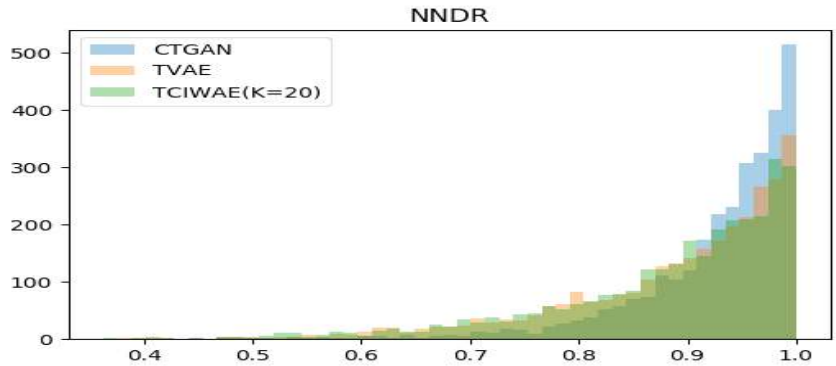
[그림 11] 각 재현 데이터별 DCR의 히스토그램 시각화

⑤ 프라이버시 평가지표 : NNDR

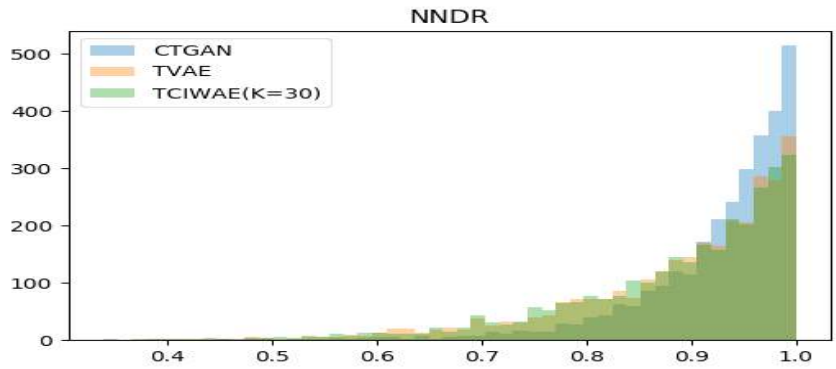
NNDR(Nearest Neighbor Distance Ratio)

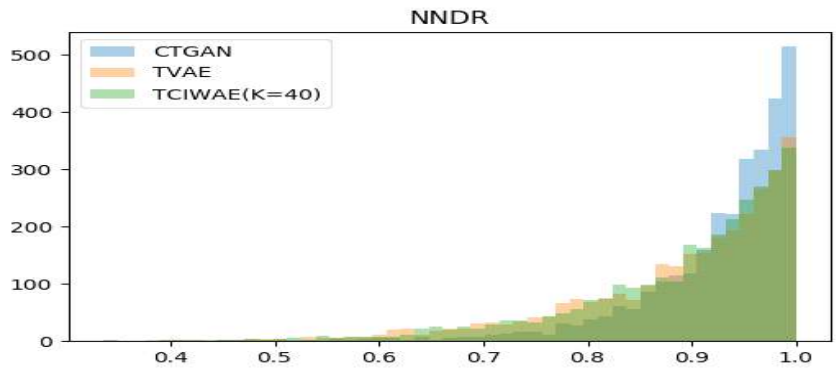


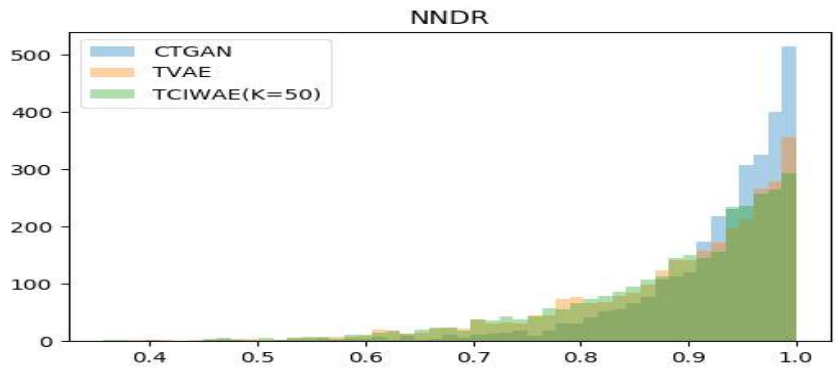




NNDR(Nearest Neighbor Distance Ratio)







[그림 12] 각 재현 데이터별 NNDR의 히스토그램 시각화

[그림 11]-[그림 12]는 제안 방법인 TCIWAE와 비교 방법인 CTGAN, TVAE로 생성한 재현 데이터의 프라이버시 노출 정도를 파악하고자 DCR과 NNDR 지표를 계산한 값을 히스토그램으로 그린 결과이다.

[그림 11]을 살펴보았을 때, 제안 방법인 TCIWAE를 통해 재현한 데이터의 DCR은 모든 K에 대해서 비교 방법인 CTGAN의 그것보다 더 낮은 값을 가질 때도 있음을 확인할 수 있었다. 이를 통해 제안 방법인 TCIWAE를 통해 재현한 데이터의 프라이버시 노출 위험 정도가 비교 방법인 CTGAN의 그것보다 높다고 판단할 수 있었다. 그러나 제안 방법인 TCIWAE의 DCR은 비교 방법인 TVAE의 그것과 비슷한 정도의 값을 가지므로 TVAE와 비슷한 프라이버시 노출 위험을 보이나 유용성 측면에서 크게 향상되었다고 평가할 수 있었다.

[그림 12]를 살펴보았을 때, [그림 11]의 DCR의 결과와 비슷하게 제안 방법인 TCIWAE의 NNDR이 모든 K에 대해서 비교 방법인 CTGAN의 그것보다는 낮은 값을 가져 높은 위험도를 보이는 반면 비교 방법인 TVAE의 그것과 비슷한 값을 가져 TVAE 수준의 프라이버시 노출 위험을 가진다고 평가할 수 있었다.

Covertype 데이터를 다양한 표 데이터 재현 방법론을 활용해 재현한 결과, 제안 방법인 TCIWAE가 모든 유용성 지표에서 비교 방법인 CTGAN, TVAE보다 높은 결과를 얻었고 프라이버시 노출 위험은 TVAE 수준임을 확인할 수 있었다. 즉, 제안 방법인 TCIWAE는 TVAE와 비슷한 노출 위험도를 가지지만 TVAE보다 유용성 측면이 향상되었다.

Ⅴ. 결론

본 논문에서는 안정적인 학습 기반의 성공적인 표 데이터 재현을 위해서 기존의 우도 함수 기반의 표 데이터 재현 방법론인 TVAE를 보완하여 새로운 방법론인 TCIWAE를 제안하였다. 제안 방법인 TCIWAE는 TVAE가 사용하는 VAE 목적함수 대신 원본 데이터 분포의 로그 우도 함수에 더 정밀한 하한 값을 제공하는 것으로 알려진 IWAE의 목적함수를 사용하여 재현 성능을 높이고자 하였다. 또한, 범주형 변수의 범주 불균형 문제를 해결하고자 조건부 분포를 모델링 하였다. 마지막으로 표 데이터에 특화된 기존의 표 데이터 전 처리 기법을 추가하였다.

본 연구에서 제안하는 TCIWAE와 기존의 표 데이터 재현 방법론인 CTGAN과 TVAE를 이용해 2가지 데이터를 생성한 결과 제안 방법인 TCIWAE가 비교 방법론과 비슷한 프라이버시 노출 위험을 보이면서도 더 뛰어난 재현 성능을 보임을 여러 평가지표를 통해 확인할 수 있었다. 실험의 주요 결과는 다음과 같다. 첫째, 유용성 측면에서 제안 방법인 TCIWAE는 CTGAN보다 안정적이며 뛰어난 재현 성능을 보였고 K가 1보다 클 때 TVAE보다 더 정교한 재현 데이터를 생성해냈다. 둘째, 프라이버시 노출 위험 측면에서 제안 방법인 TCIWAE는 적어도 TVAE와 비슷한 정도의 노출 위험을 보였다.

결과적으로 제안 방법인 TCIWAE는 K가 1보다 클 때 프라이버시 노출 위험은 비교 방법론과 비슷하면서 유용성 측면이 향상된 재현 데이터를 생성하기에 우수하다.

추후에는 비대칭 수치형 변수에 알맞은 전 처리 작업 개발과 불가능한 사건 조합을 방지하는 패널티 함수 개발을 시도하고자 한다.

참 고 문 헌

[1] Goodfellow, I., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., Courville, A., Bengio, Y. (2020). Generative Adversarial Networks. Communications of the ACM, 63(11), 139-144.

[2] Kingma, D., Welling, M. (2014). Auto-Encoding Variational Bayes. 2nd International Conference on Learning Representations.

[3] Xu, L., Skoularidou, M., Cuesta-Infante, A., Veeramachaneni, K. (2019). Modeling Tabular Data using Conditional GAN. Advances in Neural Information Processing Systems 32: Annual Conference.

[4] Burda, Y., Grosse, R., Salakhutdinov, R. (2016). Importance Weighted Autoencoders. 4th International Conference on Learning Representations.

[5] Arjovsky, M., Chintala, S., Bottou, L. (2017). Wasserstein Generative Adversarial Networks. In International Conference on Machine Learning. PMLR, 214-223.

[6] Xu, L., Veeramachaneni, K. (2018). Synthesizing Tabular Data using Generative Adversarial Networks. arXiv preprint arXiv:1811.11264.

[7] Radford, A., Metz, L., Chintala, S. (2016). Unsupervised Representation Learning with Deep Convolutional Generative Adversarial Networks. 4th International Conference on Learning Representations.

[8] Zeiler, M., Krishnan, D., Taylor, G., Fergus, R. (2010). Deconvolutional networks. 23th Conference on Computer Vision and Pattern.

[9] Krizhevsky, A., Ilya, S., Hinton, G. (2017). ImageNet Classification with Deep Convolutional Neural Networks. Communications of the ACM, 60(6), 84-90.

[10] Bishop, C. (2006). *Pattern Recognition and Machine Learning.* Springer.

[11] Mirza, M., Osindero, S. (2014). Conditional Generative Adversarial Nets. arXiv preprint arXiv:1411.1784.

[12] Zhao, Z., Kunar, A., Birke, R., Chen, L. (2021). CTAB-GAN: Effective Table Data Synthesizing. Asian Conference on Machine Learning. PMLR, 157, 97-112.

ABSTRACT

TCIWAE: A Study of Synthesizing Tabular Data Using Conditional IWAE

Jiwoo Kim
Department of Statistics
Graduate School of
Sungshin Women's University

As data sharing has received much attention, there is a growing concern about the risk of privacy exposure. To overcome the trade-off between data sharing and privacy risk, generating and distributing synthetic data, instead of the original data, is considered as a good alternative. Synthetic data refer to simulated data of the original ones, but still containing meaningful information. Well-generated synthetic data can encourage active data sharing while protecting the privacy of the original data.

Most of the existing studies of the tabular data synthesis employing deep generative models consider Generative Adversarial Networks(GAN). However GAN often gives unstable results since the goal of its objective function is to find a saddle point with a min-max learning procedure.

Thus, it is hard to apply to various real-world domains. On the other hand, there is another learning framework which is based on the log-likelihood-based learning method. It is known to be more stable and powerful than GAN, but fewer researches have been done.

In this study, we devise a new log-likelihood-based learning framework to generate synthetic tabular data successfully. We focus on the Variational AutoEncoder(VAE), one of the most widely used log-likelihood methods, and improve it in three ways. First, instead of using the Evidence Lower Bound(ELBO), we adapt the objective function of Importance Weighted AutoEncoders(IWAE) that gives a tighter lower bound of the log-likelihood. Second, to handle imbalanced categorical variables more efficiently, we apply IWAE to a conditional generative model, developing Conditional IWAE(CIWAE). Lastly, by combining an existing pre-processing method specializing in tabular data to CIWAE, we devise a new synthesizer called Tabular CIWAE(TCIWAE).

To demonstrate the superiority of our method, we conduct various experiments of tabular data synthesis using two benchmark data sets. We show that our method generates more realistic tabular data with similar leakage of privacy information compared to other baseline methods.

감사의 글

대학원 생활을 시작한 것이 엊그제 같은데 벌써 졸업의 시간이 다가오게 되었습니다. 돌아보니, 2년간의 대학원 생활을 통해 꿈을 꾸게 되었고, 좋은 사람들을 얻었으며, 많은 경험을 해볼 수 있었습니다. 제가 대학원 생활을 통해 많은 것을 누릴 수 있었던 것은 저를 사랑으로 돌봐주신 많은 분들 덕분입니다. 이 글을 통해서 감사 인사드리고자 합니다.

부족함이 많은 저를 지도 학생 삼아주셔서 연구자의 자질을 배울 수 있도록 본을 보여주시고, 항상 모든 질문에 친절하게 답해주시며, 모든 대학원 생활을 섬세하게 돌봐주신 김동하 지도교수님께 진심으로 감사를 드립니다. 한 번씩 찾아뵐 때마다 좋은 말씀을 들려주신 이종협 교수님, 여러 가지 경험을 해볼 수 있도록 좋은 기회를 주신 이성건 교수님, 대학원 생활의 전반을 살뜰하게 돌봐주셔서 부족함이 없게 해주신 박만식 교수님, 다사다난했던 여러 일들 가운데 항상 도움 주셨던 박성오 교수님, 잘할 수 있을 것이라 응원해주시고 격려해주신 정호현 교수님, 앞으로의 진로에 큰 관심과 격려를 해주시며 학업적인 부분에 많은 도움을 주신 박관영 교수님께 깊은 감사를 드립니다.

홀로 있었던 대학원 생활에 든든한 동반자 되어준 윤아, 세리언니께 진심으로 감사드리며 부족한 언니였지만 항상 격려해주고 응원해준 서연, 나경, 수지, 윤진에게도 고마움을 전합니다. 앞으로 대학원 생활을 시작하게 된 혜민, 서영이도 응원하고 축복합니다.

저의 대학원 여정에 항상 은혜를 주신 하나님께 깊은 감사와 영광을 드립니다. 더불어, 언제나 저를 믿고 응원해주신 사랑하는 우리 아빠, 엄마, 동생 지은이, 외대교회 목사님, 사모님을 비롯한 지체들 너무 사랑하고 감사드립니다.

앞으로 많은 도움을 받은 만큼 많은 사람들을 돕는 김지우가 되겠습니다.