발간등록번호
11-1790377-100001-01

2024. 12

데이터의 안전한 활용을 위한

합성데이터 생성·활용 안내서

개인정보보호위원회

안내사항

발간 목적

본 안내서는 「개인정보보호법」을 준수하며 합성데이터를 생성, 활용할 수 있도록 참고 방법 및 절차 등을 안내할 목적으로 마련되었으며, 합성데이터를 업무 등에 적용하는 담당자 등이 활용할 수 있습니다.

제·개정 이력

개인정보보호 관련 법·제도 및 환경을 반영하여 다음과 같이 제정하였습니다.

일자	주요 내용
'24. 12. 발간	합성데이터 생성 및 활용 방법, 절차, 유의사항 등 안내

재검토 기한

안내서의 최신성을 유지하기 위해 발간일(2024년 12월)을 기준으로 매 3년이 되는 시점(매 3년째의 12.31.까지를 말함)마다 보완 및 개선 등의 조치를 취할 예정입니다.

저작권 표시

본 안내서 내용의 무단전재를 금하며, 가공·인용할 때는 출처를 밝혀 주시기 바랍니다.

* 출처 : 개인정보보호위원회, 「합성데이터 생성·활용 안내서」 2024.12.

문의처

안내서 내용 관련 문의는 소관 법령별로 다음의 연락처로 주시기 바랍니다.

- 개인정보 보호법 : 개인정보보호정책과(☎02-2100-3057, 3047)
  개인정보 법령해석 지원센터(☎02-2100-3043)
- 합성데이터 생성·활용 안내서 : 신기술개인정보과(☎02-2100-3068)

관계 법령

「개인정보 보호법」 제15조, 제17조, 제18조, 제26조, 제28조의2, 제58조의2 등

※ 법령 최신 자료는 국가법령정보센터(www.law.go.kr), 개인정보 보호 안내서 최신 자료는 개인정보보호위원회 누리집*, 개인정보 포털**을 참고

* 개인정보보호위원회 누리집(www.pipc.go.kr) : 법령 > 법령정보 > 안내서

** 개인정보 포털(www.privacy.go.kr) : 자료 > 자료보기 > 안내서

목 차

Ⅳ

제4장
활용 안내사항

부록

제1장

안내서 개요

Ⅰ 안내서 개요

1. 목적

■ **인공지능(AI) 등 신기술 발전으로 학습에 필요한 데이터 수요가 급증하고 있으나, 현실에서는 데이터 활용이 쉽지 않은 상황**

- AI가 현실에 부합하는 서비스를 제공하기 위해서는 개인정보를 포함한 데이터가 필요하나, 법적 한계 등으로 직접적 활용에 한계
- 공개 데이터의 생성 속도는 한정적*으로 학습데이터 부족이 우려

* 인간이 생성한 공개 텍스트 데이터의 유효 재고량은 약 300조 토큰 규모로 '26년~'32년 사이 AI 언어모델에 완전히 활용될 것을 예측('24. 美 Apoch AI)

■ **합성데이터가 이러한 문제를 해결하는 대안으로 각광받고 있으나, 합성데이터의 성격, 활용조건 등이 불분명하여 민간의 불확실성 증대**

- 의료·제조 분야 등에서 이미 합성데이터를 연구·개발에 활용*하고 있지만, 관련된 익명성 판단 기준은 부족한 상황

* (예시1) 다양한 도로상황, 날씨, 위험 등을 생성하여 자율주행 알고리즘에 학습
(예시2) 희귀 질환의 합성데이터를 생성하여 의료기기 진단 정확도 상승
※ 현재('24.11.) 합성데이터의 개인정보 안전성 검증에 대해서는 논의가 진행 중으로, 표준화된 방법론은 없는 것으로 보임

■ **본 안내서는 선행 연구(합성데이터 생성 참조모델, '24.5)를 기반으로 「개인정보보호법」을 준수하면서 합성데이터를 생성·활용할 수 있도록 관련 내용을 안내하고자 함**

- 본 안내서는 합성데이터 생성 절차를 규율하거나 방법론 등을 표준화하고자 하는 의도가 없으며, 합성데이터를 생성하고자 하는 자는 누구든지 참고할 수 있는 권고 및 안내의 용도로 마련되었음
- 합성데이터 생성 및 검증 기술은 현재도 지속 발전하고 있음에 따라 본 안내서도 개정수요에 맞춰 업데이트 될 계획임

[참고 : 합성데이터 관련 논의 경과]

• 「합성데이터 생성 참조모델」 5종 모델 발표('24. 5.)

- 보건의료, 유통, 공공안전, 금융 등 총 5종*의 합성데이터 셋을 공개**하고, 각 합성데이터를 생성한 절차와 데이터에 대한 설명을 포함함

* 구강 이미지, 안전모 착용 이미지, 혈당 측정정보, 통신사 멤버십 사용내역, 기업주주 대표자 정보

** 가명정보지원플랫폼(dataprivacy.go.kr)에서 다운로드 가능하므로, 실 데이터와 연계하여 본 안내서 참고 가능

• 안내서 마련을 위한 연구반 운영(6회, '24. 6.~10.)

- 산업계·학계·법조계 전문가로 구성된 연구반 진행, 법·기술적 주요 쟁점사항 논의 및 안내서(안)에 대한 의견 수렴

회차	주제	회차	주제
1차	합성데이터가 익명 데이터로 인정받기 위한 조건 논의	4차	안내서 초안 1차 서면 검토
2차	안내서에 담을 생성방법론 확정	5차	안내서(안)에 대한 공유 및 자유토론
3차	안내서에서 제시할 안전성, 유용성 측도 토론	6차	안내서 최종 검토

2. 적용대상

■ 본 안내서는 원본데이터에 개인정보가 포함된 경우를 대상으로 함

- 원본데이터에 개인정보가 없는 합성데이터는 외부 공격 등에 의해 노출될 개인정보가 없으므로 본 안내서의 적용 대상에서 제외

■ 완전 합성데이터를 기준으로 서술함에 따라 부분 합성데이터를 생성할 때는 합성데이터로 대체하고자 하는 영역에 본 안내서 적용이 가능

구분	설명
완전 합성데이터	- 생성하고자 하는 합성데이터에 원본데이터가 전혀 없이 모두 가상으로 생성된 데이터
부분 합성데이터	- 원본데이터 중 일부 데이터셋 또는 일부 속성변수를 선택하여 합성데이터로 대체한 데이터 - 다른 속성은 그대로 두고, 민감성이 높거나 공개가 어려운 데이터만 합성데이터로 대체하는 방식 등으로 활용 ※ 생성된 합성데이터에 원본데이터 레코드가 존재하므로 유용성이 높아질 수 있지만, 합성데이터 레코드와 원본데이터 레코드 간 연결 가능성이 커져 안전성이 낮아질 수 있음

3. 용어 정리

구분	용어설명
합성데이터	컴퓨터 시뮬레이션 또는 알고리즘에 의해 생성된 정보로, 원본데이터의 구조적 및 통계적 속성을 재현한 데이터 - 숫자나 텍스트, 이미지, 비디오, 표 등 다양한 형식 데이터일 수 있음
원본데이터 (합성대상)	합성데이터를 생성하기 위해 사용되는 데이터
합성	컴퓨터 시뮬레이션 또는 알고리즘 등을 통해 원본데이터의 통계적 속성 등을 재현한 합성데이터를 만들어 내는 것
레코드	데이터베이스 테이블(DB table)에서 가로 방향의 한 줄로 나타내는, 줄, 행(行, row) 또는 튜플(tuple)
합성데이터 생성자	실제로 합성데이터를 생성하는 공공기관, 법인, 단체 또는 개인 등
합성데이터 이용자 (활용자)	합성데이터를 활용하는 공공기관, 법인, 단체, 개인, 또는 불특정 다수 등 제 3자
유용성 검증	합성데이터와 원본데이터의 통계적 분포가 얼마나 유사한지, 동일한 목표를 달성할 수 있는지 등을 검증하는 단계 (행위)
안전성 검증	생성된 합성데이터를 통해 원본데이터 내 개인이 식별될 가능성이 있는지 등을 검증하는 단계 (행위)
개인정보 (개인정보보호법 제2조)	살아있는 개인에 관한 정보로서 다음의 정보를 포함함 - 성명, 주민등록번호 및 영상 등을 통하여 개인을 알아볼 수 있는 정보 - 해당 정보만으로는 특정 개인을 알아볼 수 없더라도 다른 정보와 쉽게 결합하여 알아볼 수 있는 정보
가명정보 (개인정보보호법 제2조)	개인정보를 가명처리 함으로써, 원래의 상태로 복원하기 위한 추가정보의 사용·결합 없이는 특정 개인을 알아볼 수 없는 정보 ※ 가명정보도 개인정보의 범주에 포함
익명정보 (개인정보보호법 제58조의2)	시간·비용·기술 등을 합리적으로 고려할 때 다른 정보를 사용하여도 더 이상 개인을 알아볼 수 없는 정보

제2장

합성데이터 소개

Ⅱ 합성데이터 소개

1. 합성데이터 정의

- 합성데이터(synthetic data)는 원본데이터의 형식과 구조 및 분포 특성을 학습하여 생성된 모의(simulated) 데이터임
- 합성데이터(synthetic data)는 개인의 프라이버시를 보호하는 동시에 산업적으로 효용성이 높은 데이터를 활용할 수 있는 방법임

■ 합성데이터는 특정 목적을 위해 원본데이터의 형식과 구조 및 통계적 분포 특성과 패턴을 학습하여 생성한 모의(simulated) 또는 가상(artificial) 데이터임

※ Synthetic data는 재현데이터로 번역되기도 하는데, 본 안내서에서는 합성과 재현을 구분하지 않고 동일한 의미로 서술

- 합성데이터는 가상 데이터이기 때문에, 잘 생성된 합성데이터는 원본데이터의 개인 식별정보나 민감정보를 외부에 직접적으로 노출하지 않아 개인정보 이슈를 해결하는 하나의 방법이 될 수 있음
- 합성데이터는 데이터 부족 문제나 데이터를 수집·이용하기 어려운 상황 등에서 합리적인 대안이 될 수 있음

▶ **데이터 증강** : 기존 데이터를 바탕으로 데이터가 부족한 범주 데이터를 추가로 생성하여 더욱 정확한 모델 학습을 하는 데 유용할 수 있음

▶ **데이터 다양성 증가** : 다양한 시나리오와 조건을 반영한 데이터를 생성함으로써 모델의 일반화 능력을 향상시킬 수 있음

▶ **프라이버시(privacy) 보호** : 실제 개인정보나 민감정보를 포함하지 않기 때문에, 개인정보보호 규정을 준수하면서 데이터를 생성하고 사용할 수 있음

▶ **데이터 접근성** : 데이터 구입 비용이 크거나 데이터의 사용기간이 한정된 경우, 개인정보의 목적 외 이용·제공 제한에 따라 데이터 공유가 어려운 상황 등에서 합성데이터를 사용하여 데이터 활용이 가능함

2. 합성데이터 유형 및 사례

1 (원본데이터 형태) 정형 합성데이터와 비정형 합성데이터로 구분

▶ 아래 사례별 세부내용은 부록5, 6 및 「합성데이터 생성 참조모델('24.5.)」 참고
▶ 참조모델 합성데이터(5종)* 다운로드 : 가명정보지원플랫폼 (dataprivacy.go.kr)
* 구강 이미지, 안전모 착용 이미지, 혈당 측정정보, 통신사 멤버십 사용내역, 기업주주 대표자 정보

• 정형 합성데이터 : 원본데이터가 행과 컬럼으로 이루어진, 테이블 형태(CSV 파일 등) 데이터로부터 생성된 합성데이터

※ 수치형, 문자형, 범주형, 날짜형 등 다양한 형태의 컬럼(항목)으로 구성

- 정형 합성데이터 사례 ❶

※ 멤버십 앱 사용고객 및 제휴사 선호 분석을 위한 합성데이터 생성
▶ (활용 목적) 통신사 맴버십 고객의 쿠폰 사용/미사용 특성을 분석하여 쿠폰 상품개발 등 기획
▶ (원본데이터) 통신사가 보유한 '22. 6. ~ '23. 4. 데이터 중 월 2만건 표본 데이터
▶ (합성데이터 생성) 고객연령, 고객성별, 주소, 쿠폰유형, 발행요일, 발행월, 발행시각 등 102,503건 생성

- 정형 합성데이터 사례 ❷

※ 헬스 케어기기의 오차 정밀 보정을 위한 합성데이터 생성
▶ (활용 목적) 혈당기기 정밀 보정을 위해 자사와 타사의 혈당 기기 측정데이터 비교 분석
▶ (원본데이터) 고객의 혈당 측정 결과를 관리하는 서비스 업체가 보유한 고객 단위 혈당 측정 데이터
▶ (합성데이터 생성) 의료정보측정시간, 나이, 혈당, 식사여부, 식사량, 측정기기 등 723건 생성

합성데이터 생성 결과					
측정 시간	측정자 나이	측정자 혈당값	측정 전 식사 시간	직전 섭취 칼로리	측정기기
오전 10시	57세	137	식사 후 8시간	0 kcal	A사 측정기기
오후 5시	52세	126	식사 후 3시간	500 kcal	B사 측정기기
오후 11시	53세	120	식사 후 5시간	680 kcal	A사 측정기기
오전 3시	61세	126	식사 후 6시간	388 kcal	A사 측정기기
오후 10시	53세	137	식사 후 4시간	90 kcal	C사 측정기기

• 비정형 합성데이터 : 원본데이터가 정형 데이터가 아닌 데이터로부터 생성된 합성데이터

※ 비정형 데이터는 텍스트, 이미지, 영상, 음성 등이 있음

- 비정형 합성데이터 사례 ❶

※ 구강 내 질환 진단 및 예방을 위한 AI 학습용 합성데이터 생성

▶ (활용 목적) 희소한 충치 데이터를 합성데이터로 생성하여 AI 충치 진단 솔루션 개발

▶ (원본데이터) IRB 공동연구와 위탁계약, 제공협약을 통한 총 500명의 상악/하악 구강 데이터

▶ (합성데이터 생성) 512x512 해상도의 상악치/하악지 구강 이미지 데이터 1,000장 생성

- 비정형 합성데이터 사례 ❷

※ 안전사고 및 재난 감지 엔진 구축을 위한 AI 학습용 합성데이터 생성

▶ (활용 목적) 안전보호구 착용상태를 자동 감지하는 기술 개발을 위해 안전보호구 착용 이미지 데이터 필요

▶ (원본데이터) 개인정보 수집·이용에 동의한 대상자들에 한해 작업 현장에서 안전보호구를 착용하고 전신/상반신을 촬영한 432장의 이미지

▶ (합성데이터 생성) 512x512 해상도의 안전보호구 가상 이미지(가상 얼굴 포함) 5,500장 생성

2 (처리 목적) 공개용, 특정 기관 내부에서 분석 및 AI 학습용, 교육용, 기술 검증용 등으로 구분[1)]

- 공개용 합성데이터 : 불특정 다수가 사용할 수 있도록 공개용으로 만들어진 데이터

※ 공개용 사례 (통계청)
▶ (활용목적) 정보공개로 인한 개인정보 노출 최소화 등을 위한 합성데이터 활용방안 연구
▶ (합성데이터 내용) 기업통계등록부에 포함된 숙박 및 음식점 사업체로 합성데이터 생성
▶ (공개 서비스 기간) '23. 6. 28(수) ~ 9.30(토) 기간에 한시적으로 다운로드 허용

- 통계 분석 및 학습용 합성데이터 : 특정 기관 내부 활용을 위해 만들어진 데이터

※ 내부 분석용 사례 (A사)
▶ (활용목적) '18년 이후 급증한 제주도 전출 인구에 따른 인구구조 변화 및 영향 등 분석
▶ (합성데이터 내용) 개인 전출입 지역, 연령, 소득에 대해 원본데이터 분포와 유사한 약 50만 건의 합성데이터를 생성

- 교육용 합성데이터 : 다양한 분석 교육을 위해 만들어진 데이터

※ 교육용 사례 (B사)
▶ (활용목적) 개인신용정보 DB를 사용하여 시계열적 분석을 할 수 있도록 교육용 합성데이터 생성
▶ (합성데이터 내용) 원본데이터의 차주정보, 대출정보, 연체정보 등 26개 항목의 특성을 따르는 통계 모형을 도출해 180만명 가상 차주에 대한 25개월분의 합성데이터를 생성

- 기술 검증용 합성데이터 : IT 시스템 위탁 개발시, 개인정보가 포함된 테스트 데이터를 직접 제공하지 않고 합성데이터를 제공하여 검증 가능

기술 검증용 합성데이터 예시

1) 출처 : 유엔 유럽경제위원회(UNECE), 2022

3. 합성데이터 동향

합성데이터 생성 기술 동향

- (초창기) 분포추정과 결측 대체 방법을 활용한 기법이나 의사결정나무 기반으로 한 생성 방법이 주로 개발되었으며, 이때 개발된 방법 중에서 synthpop 등 현재까지도 활발하게 사용되는 방법들이 있음
- (2010년대 중반 이후) 비정형데이터 중심으로, 딥러닝 기반의 생성형 모형을 이용한 합성데이터 기법들이 소개되기 시작하였음
- (최근) 합성데이터에 차등 프라이버시를 사용하는 방식이 재식별 위험을 줄이는 메커니즘으로 논의되고 있으며, 합성데이터의 유용성을 유지하며 차등 프라이버시를 적용하려는 연구가 진행 중임

해외 합성데이터 관련 제도 : 가이드라인 등

- 합성데이터의 활용성이 증대됨에 따라 유엔 유럽 경제위원회(UNECE), 싱가포르 개인정보보호위원회(PDPC) 등에서 합성데이터를 생성하거나 활용하는 데 도움을 주는 가이드라인을 발표함

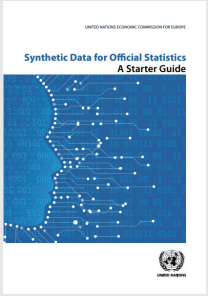
기관명	가이드라인 제목
UNECE(유엔)	• Synthetic Data for Official Statistics – A Starter Guide (2022)
PDPC(싱가포르)	• Proposed Guide on Synthetic Data Generation (2024)
UNU(유엔)	• Recommendations on the Use of Synthetic Data to Train AI Models (2024)

유엔 유럽 경제위원회(UNECE), 2022

<소개>

본 가이드라인은 NSO(National Statistical Offices)에서 근무하는 사람들을 위해 작성된 것으로, 데이터에 접근하는 방법으로 합성데이터를 사용하려는 사용자를 위한 것

NSOs에서 성공적으로 적용한 합성데이터 사례를 강조하고, 합성데이터를 생성하기 위한 다양한 방법을 소개



- **합성데이터 정의** : 합성데이터는 분석적(analytical) 가치를 지니고 있으며 높은 수준의 노출(disclosure) 통제를 유지하는 무작위적으로(stochastically) 생성된 데이터

> Synthetic data is defined as being stochastically generated data that has analytical value, and which maintains high levels of disclosure control.

- **합성데이터 생성방법** : 조건부 분포를 활용한 완전 대체법(FCS), 시뮬레이션 데이터, 딥러닝 방법 소개. 데이터 유형에 따라 적절한 방법론 및 생성방식이 고려되어야 함

- **합성데이터 공개 고려 사항** : 합성데이터 효용성과 공개위험 간의 상충관계를 설명하고, 공개 시 신원노출(identity disclousure), 속성노출(attribute disclosure)의 두가지 위험성이 있음을 언급하며 해당 정보를 삭제할 것을 안내. 또한, 개인정보 보호 기법으로 차등 프라이버시(DP), k-익명성, l-다양성, t-근접성, 동료평가 등을 소개

- **합성데이터의 유용성 측정** : 신뢰구간 중첩, 일반적인 단일값 형태로 제공되는 유용성 측도 등 방법론을 제시하고, 유용성을 평가하는 방법을 사례로 안내

싱가포르 개인정보보호위원회(PDPC), 2024

<소개>

해당 가이드라인에서는 합성데이터를 개인정보보호 강화 기술(PET) 중 하나로 소개하며, 합성데이터를 사용하여 정형 데이터(정형 합성데이터)를 생성하는 것을 중점으로 다룸

합성데이터는 기본적으로 허구의 데이터이나 재식별 위험이 존재함을 전제하고 있으며, 이러한 위험을 최소화하기 위해 채택할 수 있는 모범 사례, 위험 평가 및 고려 사항 그리고 거버넌스 통제(governance controls), 계약 프로세스(contractual process), 잔여 위험을 완화하기 위한 기술적 조치 등을 포함

▶ **합성데이터 정의** : 합성데이터는 특정 목적을 위해 구축된 수학적 모델(인공지능(AI)/머신러닝(ML) 모델 포함) 또는 알고리즘을 사용하여 생성된 인공데이터

> Synthetic data is commonly referred to as artificial data that has been generated using a purpose-built mathematical model (including artificial intelligence (AI)/ machine learning (ML) models) or algorithm.

▶ **합성데이터 생성 시 권고사항** : 합성데이터를 생성하기 위해 데이터 파악, 데이터 준비, 합성데이터 생성, 재식별 위험 평가, 잔여 위험 관리의 5단계 접근방식을 권고

▶ **합성데이터 생성 방법 예시** : 베이지안 네트워크, 조건부-코플라, 조건부 확률 분포 기반 데이터 합성, 순차 트리 기반 합성(SEQ), 생성적 적대 네트워크(GAN) 등 간략 소개

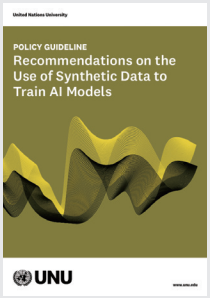
▶ **재식별 위험** : 구별(Single out) 공격, 연결(Linkability) 공격, 추론(Inference) 공격에 대해 설명

유엔 유럽 경제위원회(UNECE), 2022

<소개>

해당 가이드라인은 AI 모델 훈련 시, 합성데이터를 책임감 있게 사용하기 위한 권장 사항을 제시하며, 합성데이터가 사용되는 모든 시스템에 적용할 수 있는 정책 가이드라인(policy guideline)을 소개

합성데이터의 이점, 합성데이터 사용 시 주요 위험, 권장하는 기술적·정책적 조치 등을 다룸



▶ **합성데이터 정의** : 합성데이터는 컴퓨터 시뮬레이션이나 알고리즘에 의해 실제 데이터의 일부 구조적(structural) 및 통계적(statistical) 속성을 재현(reproduce)하여 생성된 정보

> Synthetic Data are information created by computer simulations or algorithms that reproduce some structural and statistical properties of real-world data.

▶ **윤리적 측면** : 인공지능 개발 시 데이터 편향성 등이 세계적인 영향을 미칠 수 있다고 보고, 인권·정의·평등 등 다양한 가치를 염두에 두어야 하며, 글로벌 격차를 해소하는 방향으로 인공지능 모델을 학습시킬 때 합성데이터가 사용될 수 있다고 설명. 그러나 합성데이터 역시 낮은 데이터 품질, 오용, 편향성 전파 등의 위험이 있다고 언급

▶ **권장사항** : 권장하는 기술적 조치로는 편향 완화, 합성데이터 생성 시 다양한 방법 사용, 모델 검증 및 평가 등. 권장하는 정책적 조치로는 전 세계 품질 기준 및 보안 조치 수립, 지역적 품질 기준 및 보안 조치 시행, 합성데이터를 고려한 윤리적 가이드라인 작성 등을 권장함

4. 합성데이터 생성·활용 시 고려사항

1 (법적 성격) 합성데이터는 개인정보 보호 강화기술(PET, Privacy Enhancing Technology)의 하나로, 개인정보 침해 위험 없이 안전하게 데이터를 활용하는 방법임

- 현행법 체계 상 완전 합성데이터의 경우 설정된 안전성 수준에 따라 익명정보 또는 가명정보로 판단될 수 있음

  ※ 부분 합성데이터는 개인정보가 포함돼 있을 가능성 때문에 단순히 익명 또는 가명정보로 분류하는 것이 적절치 않음

- 합성데이터는 유형에 맞는 생성방법론, 정량·정성 검증절차 등을 거쳐 안전성이 인정되면, 익명정보로 활용될 수 있음

  ※ 일반대중 공개 목적으로 합성데이터 생성 시 높은 수준의 안전기준을 설정한 후 본 안내서에 소개된 절차를 충실히 이행하면 익명정보로 활용 가능

  ※ 활용환경의 안전성이 높거나 권한있는 부서 내부활용 등을 위해 당초 안전성 수준을 낮게 설정한 경우, 또는 합성결과 최종적으로 안전성이 낮은 경우에는 가명정보로 활용·관리 하는 것이 바람직함

2 (주요 권장사항) 합성데이터를 생성·활용하고자 할 때는 다음과 같은 기본 원칙을 고려해야 함

- 안전기준 설정 : 합성데이터의 유용성과 안전성 간 상충관계를 고려하여, 활용목적 등에 따른 안전기준을 먼저 설정해야 함

  ※ 이는 제3장의 임계값 산정과도 연결

▶ (사례 1) 유용성에 중점을 두는 상황
: 안전한 내부 폐쇄환경에서의 활용, 침해 위험성이 낮은 환경에서 대규모 언어모델의 토큰(벡터) 단위 AI 학습에 활용

▶ (사례 2) 안전성에 중점을 두는 상황
: IT시스템 위탁 개발 시 활용, 합성데이터의 외부 공개

- 원본데이터 전처리 : 원본데이터에 대한 분석을 바탕으로 합성에 불필요한 영역 삭제 및 데이터 정제 등을 수행하는 것이 권장됨
  - 이는 합성데이터 결과물의 유용성을 높일 수 있음. 고유식별정보 등 식별자에 대한 삭제·암호화 등 비식별처리를 통해 원본데이터의 안전성을 높이면 합성데이터의 안전성 역시 높아짐

    ※ 식별자가 합성에 필요한 항목이라면 원본데이터 비식별처리의 생략도 가능
- 안전성 검증 : 생성된 합성데이터에 개인 식별가능정보가 존재하는 상황*에 대비하여 정량·정성적 차원의 안전성 검증**을 거쳐야 함
  - 생성된 합성데이터를 통해 원본데이터 내 개인이 식별될 가능성이 있는지 검증하는 것으로, 익명정보로 인정받기 위한 필수 절차임

    * 합성데이터를 통해 특정 개인이 연결·식별되는 신원노출(identity disclosure), 개인의 민감 속성에 대한 추론이 이루어지는 속성노출(attribute disclosure) 등[2)]

    ** 안전성 검증 방법론과 지표들은 합성데이터 생성자가 직접 정해 사용할 수 있고, 본 안내서에서는 선행연구(합성데이터 생성 참조 모델, '24.5.)에서 사용된 방법론을 안내(구별위험도, 연결위험도, 추론위험도)
- 안전한 관리 : 합성데이터를 일반대중에 공개하는 경우 등에 있어서는 재식별 등 잔여 위험* 가능성에 대비하여, 관리계획을 마련·이행할 필요가 있음

  * 멤버십 추론 공격 발생, 전문가의 악의적 이용, 기술발전에 따른 식별 가능성 증가 등(자세한 내용은 3장 5절 참고)

3 (개인정보보호법 관련성) 합성데이터 생성·활용의 적법근거는 「개인정보보호법」 제15조, 제17조, 제18조, 제26조, 제58조의2 등임

- 개인정보 수집·이용 및 목적외 이용 : 익명정보로 합성데이터를 생성·활용할 시 「개인정보보호법」 적용이 배제(제58조의2)됨
  - 합성데이터 전체가 익명정보인 경우가 아니라면 수집·이용 및 목적 외 이용 등 적법요건 확인이 필요함(개인정보보호법 제15조, 제18조)

2) 유엔 유럽경제위원회(UNECE), 2022

<합성데이터의 적법요건 확인 필요성>

합성 정도	구분		내 용
완전 합성	목적내 활용		「개인정보보호법」에 따라 합성데이터 생성·활용 가능
	목적외 활용	익명정보	
		가명정보	가명정보의 처리 등(개인정보보호법 제28조의2)에 따른 적법 요건 확인 : 통계작성·과학적 연구·공익적 기록보존 등 목적 해당 여부
부분 합성	개인식별이 가능한 부분 전체를 합성데이터로 대체		완전 합성과 동일하게 수집목적, 합성데이터 성격(익명정보 여부)에 따라 동의 필요성 등 적법요건 확인
	그 외		합성데이터 결과물의 개인정보 포함여부를 먼저 확인하고, 개인정보가 포함돼 있으면서 수집 목적 외로 활용된다면 별도 동의 등 적법요건을 갖추어야 함

• 원본데이터 제공 : 원본데이터 보유자가 아닌 제3의 외부 기관이 합성데이터를 생성하고자 하는 경우 아래 방법 등을 고려할 수 있음

< 예시 >

1) 개인정보 처리 위·수탁 계약을 체결하고 합성데이터 생성(개인정보보호법 제26조)
2) 개인정보 제3자 제공 동의를 별도로 받은 후, 원본데이터를 합성데이터 생성 기관에 제공하여 합성데이터 생성(개인정보보호법 제17조)
3) 원본데이터를 합성데이터 생성 목적으로 가명처리(개인정보보호법 제28조의2)한 뒤, 합성데이터 생성 기관에 제공하여 합성데이터 생성

원본데이터 보유자(A)	⇒	합성데이터 생성자(B)
(A = 합성데이터 활용자)	[1]위·수탁계약	[1]원본데이터로 합성데이터 생성
-	[2]제3자제공 동의 후 제공	[2]원본데이터로 합성데이터 생성
-	[3]가명처리 후 반출 (별도 동의 X)	[3]가명정보로 합성데이터 생성

※ 합성데이터 생성·활용에 대한 개인정보 수집·이용 동의를 별도로 받은 경우는 위의 논의사항을 고려하지 않아도 됨(개인정보보호법 제15조)

- [1]위·수탁 계약을 체결할 때*와 [2]제3자 제공 동의를 별도로 받는 경우** 원본데이터 보유자(A), 합성데이터 생성자(B) 간 법률관계, 책임이 상이

* 1) 위·수탁 계약 : 문서로 위탁하여야 하고, 업무내용 및 수탁자(B)를 공개해야 하며, 위탁자(A)는 관리감독 책임 등을 짐(개인정보보호법 제26조 등)

** 2) 제3자 제공 동의 : 원본데이터 보유자(A)는 정보주체에게 개인정보를 제공받는 자(B) 및 이용목적 등을 알리고 동의를 받아야 함(개인정보보호법 제17조 등)

4 (합성데이터의 한계) 합성데이터에서 개인정보가 재식별될 가능성 및 합성데이터 자체가 허위정보 또는 편향된 정보로서 사회적 부작용을 유발할 가능성이 있음

- 특히 안전기준 설정 시 유용성이 높고 안전성이 낮은 경우 개인정보 재식별 등이 가능하므로 활용환경의 보호조치, 활용 이력 모니터링 등 관리 수준을 높여야 함
- 생성된 합성데이터의 품질이 낮거나, 합성데이터가 허위정보로서 AI 학습에 활용되어 허위정보를 지속 확대·생성하는 문제 등[3]을 방지하기 위해 안전성 검증 뿐만 아니라 유용성 검증도 함께 수행해야 함

※ 합성데이터에는 원본데이터가 가지고 있는 통계적 특성, 경향 등은 재현될 수 있지만, 원본데이터의 컬럼 합계, 평균 등의 값이 합성데이터와 정확히 일치하는 것은 아님

3) 유엔 대학교(UNU), 2024

제3장

합성데이터 생성 및 활용

Ⅲ 합성데이터 생성 및 활용

<개 요>

■ **합성데이터 생성 절차는 ①사전준비, ②합성데이터 생성, ③안전성 및 유용성 검증, ④심의위원회 평가, ⑤활용 및 안전한 관리 5단계로 이루어짐**

※ 4, 5단계는 공개가능한 익명수준의 합성데이터를 생성하고자 할 때 참고가능한 절차이며, 활용목적·환경 등에 따라 절차의 생략이나 통합도 가능

1단계		2단계		3단계		4단계		5단계
사전준비	↔	합성데이터 생성	↔	안전성 검증 및 유용성 검증	↔	심의위원회 평가	↔	활용 및 안전한 관리

- 각 절차는 기본적으로 순차적이며, 필요한 경우 각각의 단계에서 이전 단계로 회귀하여 각 단계 목적이 달성될 때까지 반복 수행함
- 3단계 안전성 검증과 유용성 검증의 순서는 중요하지 않음
  - 안전성·유용성을 동시에 충족하는 합성데이터 생성을 목표로 함
  - 활용목적·범위 등을 고려하여 안전성·유용성 수준을 상황에 따라 달리 설정할 수 있으나, 이 경우에도 개인정보가 안전할 수 있도록 적절한 관리가 필요함

■ **합성데이터 생성 절차는 다음과 같은 세부 절차를 가짐**

사전준비 ▶	합성데이터 생성 ▶	안전성 및 유용성 검증 ▶	심의위원회 평가 ▶	활용 및 안전한 관리
①합성데이터 활용목적 및 활용범위 설정	①원본데이터 탐색적 분석	①측정지표 결정	내외부 전문가 평가	①활용
②합성데이터 생성·활용 주체 설정	②원본데이터 전처리	②지표별 임계값 산출		②안전한 관리
③원본데이터 이해 및 생성계획 마련	③합성데이터 생성	③안전성 및 유용성 측정		
④원본데이터 확보	④합성데이터 후처리	④검토 및 후처리		

1. [1단계] 사전준비

• 합성데이터 활용목적과 범위, 생성·활용 주체를 설정하고, 원본데이터의 확보 등 합성데이터 생성을 위한 사전준비를 수행

[1단계] - 1 합성데이터 활용목적 및 활용범위 설정

※ 활용 목적·범위를 사전에 정의함에 따라, 추후 절차에서 원본데이터의 어떤 특성을 얼마나 보존할지, 어떤 생성기법을 사용할지, 유용성·안전성을 어느 수준으로 검증할지 등 전반적인 방향성을 결정하는데 참고될 수 있음

• (활용목적) 합성데이터 생성 전, 합성데이터가 생성·활용되는 목적을 사전에 구체적으로 설정

* 활용 목적 예시[4]

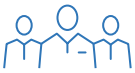
- AI 모델을 위한 학습용 데이터셋(합성데이터) 생성

▶ AI 모델 개발을 위해 대량의 데이터 셋을 확보해야 하나, 비용, 법적 문제, 개인정보 및 소유권 문제 등으로 인해 합성데이터를 생성하여 대신 활용하고자 하는 경우

▶ 보유한 데이터셋이 대용량 AI 모델 구축에 충분하지 않아, 합성데이터를 생성하여 더 많은 데이터 셋을 확보하고자 하는 경우

▶ 재난 상황, 희소질환 등 확보하기 어려운 데이터 유형의 합성데이터를 생성하여, AI 모델 훈련 시 발생하는 데이터 불균형 문제를 해결하려는 경우

- 데이터 분석 및 협업을 위한 합성데이터 생성



▶ 보건의료, 신용정보 등 개인정보를 다루는 업종의 데이터 분석 업무에서 외부와 데이터를 공유해야 하는 경우

▶ 민감한 정보를 노출하지 않으면서, 비교적 자유롭게 데이터를 제공하여 외부 업무 관계자들과 협업하고자 하는 경우

- 소프트웨어 테스트를 위한 합성데이터 생성

▶ 소프트웨어 개발 시 활용하는 데이터가 개인정보일 때, 장소에 구애받지 않고 소프트웨어를 테스트하고자 하는 경우

4) 싱가포르 개인정보보호위원회(PDPC), 2024

• (활용범위) 합성데이터 활용 목적을 참고하여 생성할 합성데이터의 활용 환경, 활용 형태 등이 특정되는 경우 명시

활용 환경	설명	필요 안전수준
ⓐ 완전 개방 (open)	외부에서 별다른 통제 없이 자유롭게 접근할 수 있는 환경	최고 ▲ ▼ 보통
ⓑ 부분 개방 (semi-open)	폐쇄 환경은 아니지만, 취급자 또는 제공자의 통제가 가능한 환경	
ⓒ 폐쇄적 환경에서의 활용 (closed)	외부와의 연결이 없거나, 있더라도 데이터의 취급 시 통제가 강제되는 환경	

활용 환경	설명	필요 안전수준
① 외부 공개	생성한 합성데이터를 외부에 공개하여 별다른 통제 없이 자유롭게 활용하는 형태	최고 ▲ ▼ 보통
② 불특정 제3자 제공	현재는 불특정이나 향후에는 제공받는 자가 특정될 수 있는 경우	
③ 제3자 제공	합성데이터를 생성하여 특정 제3자에게 제공	
④ 타 부서 이용 (제공)	처리권한을 가진 개인정보취급자가 보유한 개인정보로 합성데이터를 생성하여, 다른 부서에 제공하는 경우	
⑤ 내부 이용	처리권한을 가진 개인정보취급자가 보유한 개인정보로 합성데이터를 생성하여, 내부에서 직접 활용하는 경우	

< 활용목적 및 활용범위 설정 예시 >

▶ **(상황)**
- A 제휴사는 B 통신사에서 고객 단위 멤버십 사용내역 데이터를 받아 고객의 쿠폰 선호도를 추정하는 모델을 개발하고자 함
- B 통신사는 개인정보가 포함된 데이터를 A 제휴사에게 전달할 수 없음
- B 통신사는 합성데이터를 생성, 안전성 평가 후 A 제휴사에게 전달하고자 함

▶ **(활용목적)** 통신사 멤버십앱 사용 데이터를 이용한 쿠폰 추천 AI 모델 개발
(활용범위) 제3자 제공 및 폐쇄적 환경에서 활용

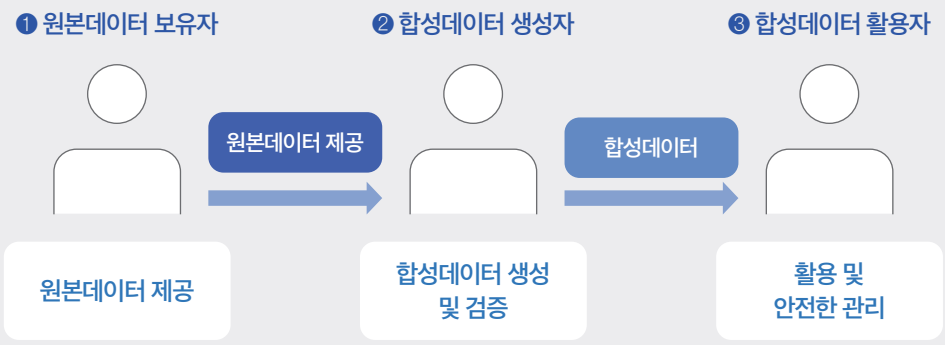
<예시> AI 모델을 위한 학습용 데이터셋 생성 + 특정 제3자 제공

[1단계] - ❷ 합성데이터 생성·활용 주체 설정

▶ 합성데이터 생성·활용 주체와 관련된 고려사항은 2장 4절 참고

- 원본데이터 확보, 합성데이터 생성과 활용에 대한 체계를 정의하고 검토함에 있어 각각의 주체를 명확히 해야 함
  - 접근제어, 인증, 데이터 파기 등 관련 정책 및 절차 마련을 위함

[합성데이터 생성 주체 개요]



※ ❶, ❷, ❸은 각각 다수가 될 수 있으며, 동일한 주체가 수행할 수도 있고 서로 다른 주체가 수행할 수도 있음
※ ❸의 경우 특정되지 않는 경우가 존재할 수 있음 (외부 공개 등)
※ 원본데이터 보유자는 본인이 아닌 제3의 주체가 합성데이터를 활용하는 경우(❶≠❸), 합성데이터 사용기간 설정, 일회적 사용 후 파기 등 강화된 '활용 및 안전한 관리' 전제 하에 원본데이터를 제공할 수 있음

❶ 원본데이터 제공자 선정

- 합성데이터의 원본데이터를 제공하는 기관(또는 개인)을 명시

▶ 원본데이터의 제공자와 원본데이터의 보유자가 다른 경우 원본데이터 보유자도 명시
(원본데이터 활용을 위한 적법성 확인 및 신뢰성 검증 등을 위해 별도 표시 필요)

❷ 합성데이터 생성자 선정

- 합성데이터를 생성하는 주체가 되는 기관(또는 개인)을 명시

▶ 생성자는 제공자로부터, 개인정보 처리 위·수탁계약, 제3자 제공 동의 등을 통해 원본데이터를 받아 합성데이터를 생성할 수 있음(개인정보보호법 제17조, 제26조 등 참고)

❸ 합성데이터 활용자 선정

- 합성데이터에 대한 이용 기관(또는 개인)이 명확한 경우, 그 대상을 명시

▶ 일반 대중 공개 등의 경우 활용자를 사전에 특정하기 어려울 수 있음

• 활용 목적, 활용 범위, 생성·활용주체 등을 정의하여 합성데이터 활용 계획서 작성 가능

합성데이터 활용 계획서 작성 예시

활용 목적	A 제휴사는 B 통신사가 보유한 멤버십 사용내역 데이터를 이용하여 어떤 고객이 어떤 쿠폰을 받고 사용하는지를 추정하는 AI 모델을 개발하고자 함 데이터는 멤버십 사용 내용, 나이와 성별, 시군구 단위 주소 등 고객의 개인정보이며, 이를 합성데이터로 생성, 익명정보로 처리하여 AI모델 개발을 위한 학습용 데이터로 활용하고자 함 B 통신사는 합성데이터 생성 업무를 C 사에 위탁하고, C 사가 합성데이터를 생성, 익명정보로 처리하면, 합성데이터는 A 제휴사에 전달되어 내부에서 활용될 예정		
생성/활용 주체	원본데이터 보유자	합성데이터 생성자	합성데이터 활용자
	B 통신사	B 통신사 (위탁) C사	A 제휴사
활용 환경	☐ 완전개방 환경 *	☐ 부분개방 환경 **	☑ 폐쇄적 환경
	* 외부에서 별다른 통제 없이 자유롭게 접근 가능한 환경 ** 폐쇄 환경은 아니지만, 취급자 또는 제공자의 통제가 가능한 환경		
반복 제공 여부	☑ 1회 제공	☐ 반복 제공 예정 (　회 예정)	
제공 방법	☐ 온라인	☑ 오프라인	

※ 또한, '합성데이터 생성 기초자료 명세서'를 작성하고 추후 평가 등에 활용 가능(아래 서식의 세부내용은 후술하는 각 단계의 설명 참고)

합성데이터 생성 기초자료 명세서 (예시)	
합성데이터명	B 통신사의 회원별 멤버십 사용내역 합성데이터
생성 기간	2024년 1월 1일 ~ 2024년 1월 30일

원본데이터 특성			
원본데이터명	B 통신사의 회원별 멤버십 사용내역 데이터		
원본데이터 유형	☑ 개인정보	☐ 가명정보	☐ 기타 ( )
데이터 업종	유통 분야		
데이터 속성	• B 통신사의 멤버십 사용 내역 (A 제휴사 연계 활용 정보) • 해당 데이터에는 B 통신사를 이용하는 고객 개인을 식별할 수 있는 고객 고유 식별자, 전화번호 등이 존재하며, 그 외 고객의 성별, 나이, 주소 등의 정보가 담긴 개인정보가 포함되어 있음 • '22.6.~'23.4. 중 매월 2만건씩 약 20만건 추출		
데이터 형식	☑ 정형데이터	☐ 비정형 데이터(이미지)	☐ 비정형 데이터(기타)
	* 해당 데이터는 CSV 형식의 정형데이터		

합성데이터 생성계획		
번호	구분	검토사항
1	환경 및 데이터 준비	• B 통신사(원본데이터 제공자) 개인정보 시스템에서 개인정보취급자가 직접 원본데이터 추출 후 안전한 장소로 이동 • 데이터 분석 및 합성데이터 생성 환경은 합성데이터 생성자가 요구하는 환경으로 개인정보처리자가 직접 준비
2	데이터 이해	• 식별자 : 고객번호(삭제), 상세 주소(읍면동 단위 일반화 처리), 전화번호(삭제) 준식별자 : 성별, 읍면동 단위 거주지 민감정보 : 없음 • 원본데이터 전처리를 통해 데이터의 위험성을 낮췄고, 폐쇄적 환경의 분석 모형 개발에 사용되므로 유용성 확보에 중점을 두고 합성데이터 생성 필요
3	합성데이터 생성	• B 통신사(원본데이터 제공자) 내 보호조치가 갖추어진 안전한 장소에서 합성데이터 생성자(위탁처리-C사)가 권한을 부여받아 개인정보처리자 관리감독 하에 합성데이터 생성 • 원본데이터의 크기를 고려하여 메모리 64GB 이상, GPU 4080Ti 12G 이상 고성능 컴퓨터 필요 • 원본데이터 특성을 고려하여 생성 기법은 가우시안 혼합 모델(GMM), synthpop-CART, CTGAN을 사용하고, 가장 성능이 우수한 기법을 최종 선별하여 합성데이터 생성

[1단계] - 3 원본데이터 이해 및 생성계획 마련

- 원본데이터를 이해하고, 원본데이터를 통해 어떻게 합성데이터를 생성할 것인지 계획을 작성하고 검토

- (원본데이터 이해) 원본데이터를 확보하기 이전에, 검토할 수 있는 원본데이터의 유형 및 속성 등 특성을 파악

  - 원본데이터 사용 시 고려해야 할 데이터 업종, 속성, 형식을 검토

검토 항목	검토 내용	예시
원본데이터 업종	특별법 적용 등을 검토하기 위해 업종을 명확하게 파악	보건의료, 신용정보
원본데이터 속성	원본데이터의 개인정보 관련 위험성에 대해 사전 검토	식별가능성, 개인 식별 시 파급력 등
원본데이터 형식	합성데이터를 생성하는 데 직접 사용되는 데이터의 형식으로, 합성데이터 생성방법을 선택하거나, 안전성/유용성 측정 시 지표 선택에 참고될 수 있으므로 명시 * 합성데이터 생성 전, 전처리 등으로 형식이 변경되는 경우 변경 후의 형식만을 고려	정형데이터, 비정형데이터(이미지, 텍스트), 기타

- (생성계획 마련) 원본데이터 이해를 바탕으로 합성데이터 처리방법, 처리자 등을 검토하여 종합적 생성계획 마련

  - 원본데이터 확보 전일지라도 식별자 등에 대한 처리계획, 원본데이터 특성에 적합한 생성방법 상정 등 대략적인 처리방법을 계획하여 정리

※ 현 단계에서 특정되지 않는 내용은 다음 절차에서 수정·보완될 수 있으며, 특히, 심의위원회 평가 서식(양식)에서 보완되어 검토되어야 함

검토항목	검토 내용	예 시
처리방법	식별자 등 처리방식, 컴퓨터 환경, 합성데이터 생성방법 등 검토	• 식별자에 대해선 삭제 • 합성데이터를 생성하는 컴퓨터 환경(필요 소프트웨어, 하드웨어 요구사항 등) 구비 • 원본 보유기관 사옥에서 합성데이터 생성 • 정형데이터 특성을 고려하여 가우시안 혼합모델 사용 예정
처리자	데이터 분석, 전처리, 합성데이터 생성 등을 수행하는 실질적 주체를 명확히 검토(개인정보취급자, 위·수탁처리, 기타 등)	• B통신사(원본데이터 제공자)와의 위·수탁 계약을 통해 C사가 합성데이터를 생성 • 원본데이터는 B통신사의 개인정보처리자(취급자)가 자체 시스템에서 직접 추출하여 준비

[1단계] - 4 원본데이터 확보

- 원본데이터 확보 시 필요한 동의서·계약서 등 적법절차를 확인하고, 필요 서류를 작성하여 원본데이터 확보
  - 생성할 합성데이터가 익명정보가 아닐 경우, 합성데이터 활용 목적으로 개인정보 수집 동의를 받아야 할 수도 있음(안내서 2장 4절 참고)

※ 합성데이터를 가명정보로 활용·관리하는 경우, 「가명정보 처리 가이드라인」 참고

- 원본데이터가 특정 법률 조항을 준수해야 하는 경우, 해당 법률에 따라 적정성 검토

※ (예시) 보건의료데이터를 활용하여 합성데이터를 생성하는 경우, 데이터의 생명윤리법 적용 여부와 IRB(Institutional Review Board, 기관생명윤리위원회) 심의 대상여부를 확인하고, 심의 대상이면 심의 필요

- 합성데이터 생성 업무를 위탁하는 경우, 개인정보 처리 위·수탁계약서를 작성하고 업무내용 및 수탁자를 공개(개인정보보호법 제26조)

※ 개인정보 처리 위·수탁(합성데이터 관련) 계약서 작성 예시[5]

계약 주체	• 위탁자는 자신이 보유하고 있는 원본데이터를 수탁자에게 제공하고 합성데이터 생성 업무를 위탁 • 수탁자는 합성데이터 생성자로서 위탁자의 관리 하, 실제 합성데이터 생성 업무를 수행
위탁 업무의 목적 및 범위와 위탁 기간	• 위탁 업무의 목적 및 범위는 합성데이터 위탁 생성임 • 수탁자는 원본데이터를 위탁받은 범위 이외로 처리하지 않아야 함 • 합성데이터를 생성하는 업무의 처리 기간 명시
재위탁 제한 등	• 사전 승낙을 얻은 경우를 제외하고, 계약상의 권리와 의무의 전부 또는 일부를 제3자에게 양도하거나 재위탁할 수 없음 명시
위탁자의 관리감독 및 교육	• 위탁자의 자료제출 요구, 현장방문 등에 수탁자는 성실히 임해야 하고, 위탁자는 관리·감독 외 필요시 수탁자 대상 교육도 진행
안전성 확보조치 관련 사항	• 접근통제 및 접근 권한의 제한 조치 등 관련 내용 제시(「개인정보의 안전성 확보조치 기준 안내서」 참고)
손해배상 등 책임에 관한 사항	• 수탁자가 불법적으로 개인정보를 유출하는 경우, 위탁자가 손해배상 책임을 지고, 수탁자도 과징금·과태료 등의 제재대상에 포함됨

5) 「개인정보 처리 통합 안내서('25. 개인정보위)」 참고

2. [2단계] 합성데이터 생성

• 실제 원본데이터를 분석하여 전처리를 수행하고, 알고리즘 등을 적용하여 합성데이터 생성 후 후처리를 진행

[2단계] - 1-1 원본데이터 탐색적 분석 (정형)

• 실제 원본데이터를 확보하여 데이터 컬럼, 레코드에 대한 분석을 통해 원본데이터의 세부 속성 파악

- 식별자, 준식별자, 민감정보 여부 구체적 검토
- 컬럼별 분포 특성 파악(분포 모양, 극단값, 범주형 변수 빈도, 컬럼간 관계 등)
- 각 컬럼 단위로 수치형·범주형·날짜 등 성격을 이해하고, 불필요한 컬럼이 있는지 파악해야 함

식별 위험성 파악	• 개인이 식별될 위험이 있는 식별자, 준식별자 등이 있는지 파악함 • 노출 시 위험성이 있는 민감정보가 있는지 파악함 • 식별자가 필요한 경우는 식별자를 일련번호로 대체하고, 식별자 내에 필요한 정보가 있을 시 해당 정보만 추출함(주민등록번호의 출생년도 등) • 분석에 필요 없는 식별자, 준식별자, 민감정보가 있는지 파악함
극단값, 범주형 컬럼 빈도 분석	• 수치형 컬럼에 극단값이 존재할 경우, 분석에 영향을 주지 않는 범위에서 일반화 혹은 삭제 고려 • 범주형 컬럼의 저빈도 범주에 대해 분석에 영향을 주지 않는 범위에서 범주 통합 등 일반화 수행 고려 ※ 원본데이터의 극단값을 반드시 제거해야 하는 것은 아님
원본데이터 세부속성 파악	• 컬럼 특성, 컬럼 간 관계 등 데이터의 항목에 대해 파악함 • '사전준비' 단계에서 파악한 데이터 특성이나 규칙 등이 실제 원본데이터와 다르거나, 원본데이터의 규칙을 파악할 수 없는 경우 원본데이터 보유자 등과 충분한 소통을 통해 명확히 이해함 • '고속도로 입차 시간은 출차 시간보다 빨라야 함' 등의 논리적 제약사항을 원본데이터에서 점검하고 만족하지 않을 시, 원인을 파악하여 조치함

[원본데이터 탐색적 분석 예시]

- 고객 번호 컬럼은 식별자이고 행정동 단위 주소지, 직업 등은 준식별자, 병명은 민감정보임
- 성별 컬럼이 1, 2로 입력되었다면 이는 범주형으로 정의해야 함
- 나이 컬럼은 정수형이고 0세에서 120세 이내임
- 접수 시각, 처리 시각 컬럼의 경우 처리 시각은 접수 시각보다 빠를 수 없음

• 분석 결과에 따라 다음 '원본데이터 전처리' 단계 또는 '합성데이터 후처리' 단계에서 데이터의 처리를 수행해야 함

※ 식별위험 관련 항목별로 처리계획을 작성할 수 있음

개인정보 처리계획 예시 (정형데이터)						
연번	항목명	구분		처리방법		세부방법 및 처리수준
		정보영역	설명			
1	B통신사 멤버십 데이터	회원id	내부 고객관리 id	전처리 시	삭제	삭제
		발행일자	일자, 시각	전처리 시	일반화	월, 시각, 요일만 유지
		고객연령	1세 단위 나이	전처리 시	유지	그대로 사용
		...				
2	...	-				

[2단계] - 2-2 원본데이터 탐색적 분석 (비정형-이미지)

• 실제 원본데이터를 확보하여 식별가능성이 높은 영역, 메타데이터(데이터를 설명하는 정보), 이미지 자체의 특성 등 세부 속성을 파악

- 메타데이터에 포함된 정형 식별자나 이미지 자체에 포함된 식별 가능한 영역(특정 장소명 등) 등 처리가 필요한 사항 파악

합성데이터 활용 필요/불필요 영역 파악	• 각 이미지 영역이 합성데이터 생성에 필요한지 파악 • 활용에 필요한 영역은 보존, 그렇지 않은 영역은 삭제 및 비식별처리될 수 있도록 구분
이미지에 메타데이터 포함 여부	• 메타데이터의 삭제 또는 비식별 처리를 고려 • 원본데이터에 메타데이터가 포함되지 않은 경우 고려대상 아님
식별 가능성이 높은 영역, 특이치 포함 여부 파악	• 사진에 포함된 사람의 이름표, X-ray 사진에 의사가 직접 표시한 환자 번호 등 식별 가능성이 큰 영역은, 삭제하거나 마스킹 등 비식별 처리를 고려 • 일부 이미지가 개인이 특정될 정도로 희귀한 내용을 포함하는 경우, 해당 영역을 삭제하거나 비식별 처리 고려(예시: 단 1명만이 가지고 있는 문신이 촬영된 신체 사진)
이미지 데이터 자체 특성 파악	• 이미지 자체의 성질이 특수한 경우, 이러한 성질만으로 재식별 가능성이 존재하는지, 합성데이터 생성에 어떠한 영향을 끼치는지 검토(사람의 얼굴, 흉터, 신체 외형 등) • 재식별 가능성과 노출 시 민감도를 같이 고려

• 분석 결과에 따라 다음 '원본데이터 전처리' 단계 또는 '합성데이터 후처리' 단계에서 데이터의 처리를 수행해야 함

※ 식별위험 관련 항목별로 처리계획을 작성할 수 있음

개인정보 처리계획 예시 (비정형 이미지 데이터)						
연번	항목명	구분		처리방법		세부방법 및 처리수준
		정보 영역	설명			
1	구강 촬영 이미지	구개	입천장	전처리 시	유지	별도 처리하지 않음
		코		전처리 시	마스킹	연구에 필요 없으므로, 마스킹 처리하여 안전하게 활용

[2단계] - 2-1 원본데이터 전처리 (정형)

• '원본데이터 탐색적 분석' 단계의 검토 결과를 바탕으로, 합성데이터 생성 전 식별자 처리, 컬럼 정리 등 데이터를 적절히 정제

- 불필요한 식별자는 삭제하고, 식별자가 필요한 경우 식별자를 다른 일련번호 등으로 변환하여 식별성을 최대한 제거
- 특이정보*는 식별성을 증가시킬 수 있지만, 분석에 중요한 데이터일 가능성이 동시에 존재하므로, 분석에 영향을 주지 않는 범위에서 레코드 삭제·로그 변환 등의 처리가 가능

*** 특이정보란?**

전체 데이터에 식별가능성을 가지는 고유(희소)값, 편중된 분포를 가지는 단일·다중 이용항목

(예시) 희귀 성씨 등 특이한 값, 국내 최고령, 고액 체납 금액 등 극단 값, 특정 데이터 분석 집단에서 희소한 값

- 불필요한 컬럼은 삭제하고, 너무 낮은 빈도를 가지는 범주는 유사한 범주와 합치는 일반화 처리를 통해 식별 가능성을 완화

※ 불필요한 컬럼 삭제, 적절한 범주 일반화는 안전성뿐 아니라 유용성 증대에도 도움이 되므로 분석 담당자 혹은 AI 모형 개발자와 상의하여 처리하는 것이 바람직함

- 처리 환경 제약으로 인해 데이터의 형태나 규모를 축소할 수도 있음

[예시]

< 식별자 삭제 : 개인 건강 정보 데이터 >

• '주민등록번호'는 유일한 식별자에 해당하므로 삭제함

● 원본데이터

주민등록번호	나이	성별	키	몸무게	...
240101-1111111	58세	남	176	89	...
240101-2222222	48세	여	155	46	...
240101-3333333	34세	남	185	60	...
240101-4444444	62세	여	161	70	...

삭제

[예시]

< 식별성 최소화 : 전자상거래 주문 데이터 >

- '주문자 ID'와 같이 여러 개 존재하는 중복 식별자는 암호화나 일련번호 등 재식별이 불가능한 형식으로 변환 후 활용할 수 있음

● 원본데이터

주문자 ID	회원 등급	구매 금액	결제 상태	구매 날짜
CUST0002	Silver	650,634	완료	2024-06-15
CUST0003	Bronze	404,572	대기 중	2024-06-21
CUST0004	Gold	624,682	완료	2024-07-19
CUST0003	Bronze	334,489	완료	2024-01-28

● 원본데이터

주문자 ID	회원 등급	구매 금액	결제 상태	구매 날짜
9b3e0f8a7c	Silver	650,634	완료	2024-06-15
e6f74c8f3c	Bronze	404,572	대기 중	2024-06-21
5a8b9c0d1e	Gold	624,682	완료	2024-07-19
e6f74c8f3c	Bronze	334,489	완료	2024-01-28

변환

< 특이정보 : 연간 소득 데이터 >

- 연간 소득 데이터에서 연간 소득이 관측된 범위에서 많이 벗어나 다른 정보와 뚜렷하게 구별되는 행은 특이정보(극단 값)에 해당함
- 희소한 직업(국회의원)의 경우, 다른 정보(나이, 거주지역 등)와 결합하여 쉽게 개인을 식별할 수 있는 특이정보에 해당함

● 원본데이터

나이	성별	직업	거주 지역	연간 소득	...
47	여	교사	부산	58,430,000	...
59	남	국회의원	서울	92,842,000	...
37	남	자영업	인천	45,294,000	...
29	여	개발자	경기	55,932,000	...
55	남	자영업	서울	4,332,463,000	...
48	여	공무원	대전	63,225,000	...
28	남	공무원	대구	39,842,000	...

[2단계] - ❷-2 원본데이터 전처리 (비정형-이미지)

- '원본데이터 탐색적 분석' 단계의 결과를 바탕으로, 합성데이터 생성 전 합성데이터의 품질과 안전성을 높이기 위해 원본데이터를 수정
  - 메타데이터가 합성데이터 생성 목적과 관련이 없는 경우 삭제, 필요한 경우 비식별 처리하거나 별도로 메타데이터 자체를 합성 하고 안전성 검증 필요
  - 식별가능성이 높거나, 합성데이터에 필요하지 않은 이미지 영역은 삭제, 비식별처리를 진행
  - 이미지 크기 정규화 등 합성데이터 생성에 적합하도록 형태 등 변경

[예시]

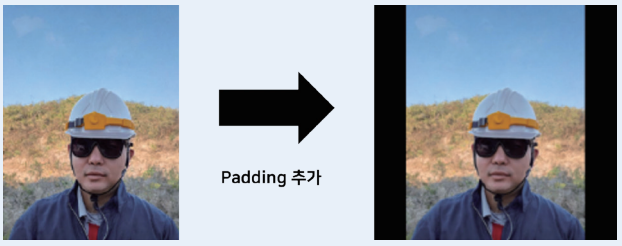
< 식별가능성 축소 >

- 이미지 데이터의 메타데이터 K-익명성 적용
- 구강 이미지에서 분석에 불필요한 영역을 블러링 처리

※ 출처 : 개인정보보호위원회, 가명정보 처리 가이드라인(2024.2.4 개정)

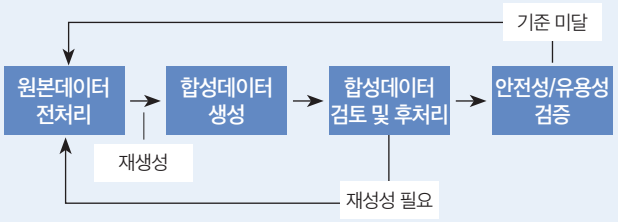
< 합성데이터 생성에 적합하도록 형태 등 변경 >

- 분석에 필요한 데이터 형태로 원본데이터 형태나 값을 변환
- 생성모형에서 요구하는 이미지 크기(예 : 1024, 1024)로 일괄 변경
- 촬영된 이미지들의 비율이 서로 다른 경우, padding을 추가하는 등 이미지 사이즈 정규화 처리



※ 합성데이터 재생성 시 전처리

합성데이터 검토 결과 또는 안전성·유용성 검증 결과가 만족스럽지 않아 합성데이터를 다시 생성해야 하는 경우, 전처리를 다시 수행할 수도 있음



(예시) 안전성 검증 결과에 따라 특정 레코드를 삭제하거나 특정 속성에 대한 재범주화 처리

[2단계] - 3 합성데이터 생성

• 합성데이터 생성방법론을 확정하고, 해당 알고리즘을 적용하여 합성데이터 생성

- 합성데이터 활용 목적부터 합성데이터 생성 방법별 장단점, 매개변수(parameter)까지 전반적으로 고려하여 사용할 생성방법 결정

고려사항	내용
합성데이터 활용 목적·범위	• 합성데이터 활용 목적, 활용 환경·형태에 따라 안전성·유용성 수준에 대한 의사결정을 해야 함
원본데이터 유형	• 원본데이터 유형(정형, 비정형 등)에 따라 적절한 생성방법을 결정해야 함 * (예시) 정형 데이터는 CTGAN, 비정형 데이터(이미지)는 StyleGan 등
생성방법의 장·단점, 개인식별 위험성* * 원본데이터와 완전히 동일한 데이터가 생성될 가능성 등	• 데이터의 특성, 매개변수 설정에 따라 생성방법의 성능이 달라지므로 다각도의 검토가 필요함 ※ 생성방법, 매개변수 설정에 의해 지나치게 높은 재현율을 보이는 등 개인식별 위험성이 존재하거나, 학습이 충분하지 않아 유용성이 너무 떨어지는 경우도 있음 ※ 정형 합성데이터의 경우, 원본데이터의 특정 레코드와 동일한 레코드가 생성될 가능성이 존재함. 특히, 범주형 컬럼 비중이 높은 정형데이터의 경우 더욱 가능성이 높음
논리적 제약	• 원본데이터 컬럼 간의 논리적 제약이 합성데이터에 적용되도록 생성 모델에 반영할 필요가 있음 *원본데이터 컬럼 간의 논리적 제약이 생성모형에 반영되지 않은 경우는 다음 '합성데이터 후처리' 단계에서 바로잡음
매개변수 설정	• 생성모형은 이미 개발된 소프트웨어를 사용할 수 있고 이 경우, 모형의 매개변수 설정에 따라 유용성, 안전성에 영향을 주기 때문에 격자 탐색 (grid search)* 등을 이용해 최적의 매개변수를 찾는 것이 중요함 * 머신러닝에서 매개변수의 탐색을 위한 방법 중 하나로, 주어진 매개변수의 모든 조합을 시도하여 최적의 매개변수를 탐색하는 알고리즘
기타 주의사항	• 공격자가 합성데이터를 생성한 알고리즘, 코드 등을 확보하였을 때 이를 이용하여 합성데이터에서 역으로 원본데이터를 식별할 수 없도록 비가역적 알고리즘을 사용해야 함  • 원본데이터를 레코드 단위로 변형하여 합성데이터의 레코드를 생성하는 것은 안전성을 보장하는 합성데이터 생성 방법이 아님 * (예시) 원본 레코드에 임의 난수를 추가하여 변형함

• 생성 방법은 크게 통계 모형 기반과 AI 모형 기반으로 나눌 수 있음

※ 동일한 원본데이터일지라도 생성 방법론에 따라 합성데이터의 안전성·유용성이 달라질 수 있으므로, 최적의 생성 방법론을 찾는 것이 중요함

- **통계 모형 기반 생성 방법** : 원본데이터의 통계적 분포 특성을 추정하여 원본과 유사한 분포 특성을 가진 가상의 데이터를 생성하는 방법

<장점> 생성 시간이 짧고 상대적으로 유용성이 높은 데이터를 생성할 수 있음

<단점> 시계열 데이터나 네트워크 데이터, 비정형데이터와 같이 복잡한 데이터 구조일 경우, 통계 기반 생성 방법이 이러한 복잡성을 충분히 반영하기 어려울 수 있음

<주요 방법> Synthpop : 조건부 확률분포를 이용해 순차적으로 데이터를 생성하는 방법으로 생성 속도가 빠르지만 원본과 같은 레코드가 생성될 가능성이 높음

가우시안 혼합 모델(Gaussian Mixture Model) : 원본데이터를 여러 정규분포 혼합 모형으로 가정하여 합성데이터를 생성하는 방법으로, 정규분포 가정이 적합하지 않은 경우 합성데이터 품질이 떨어지는 단점 존재

베이지안 네트워크(Bayesian Network) : 변수 간의 조건부 의존관계를 그래프로 표현하며 조건부 확률분포를 통해 데이터를 생성하는 방법으로, 변수가 많을 경우 기하급수적으로 계산비용 증가 가능

<예시> • 대형할인점에서 고객의 월별 구매 데이터를 분석하여 마케팅 전략을 최적화하기 위해 Synthpop 알고리즘을 사용하여 합성데이터 생성

- **AI 모형 기반 생성 방법** : 딥러닝 생성형 모델을 이용하여 원본데이터의 분포 특성을 학습시키고 원본데이터의 분포 특성에 근사하는 데이터를 생성하는 방법

<장점> 정형 데이터뿐만 아니라 이미지, 텍스트, 오디오, 동영상 등 다양한 데이터 유형의 복잡한 패턴과 상호작용 학습이 가능함

<단점> 딥러닝 모델의 학습 시간과 비용이 많이 소요됨. 또한 모형의 매개변수(parameter) 조정이 복잡하므로 최적의 성능을 위해 많은 실험과 조정이 필요함. 과적합이 일어날 경우, 원본데이터를 거의 그대로 생성하는 문제점도 존재함

<주요 방법> 변분 오토인코더(VAE) : 인코더와 디코더로 구성된 신경망 구조를 활용하여 학습하는 모형으로, 상대적으로 계산량이 적은 대신 복잡한 패턴을 학습하는데 한계

생성적 적대 신경망(GANs) : 생성자와 판별자라는 두 신경망이 성능을 최적화해 데이터를 생성하는 방법으로, 이미지 데이터 합성 성능이 뛰어나지만 비용·시간이 크게 소요

확산모델(Diffusion Models) : 점진적 잡음(noise)을 추가하여 확률분포에 가깝도록 확산시킨 후 역으로 확률적 모형화를 통해 잡음을 제거해 나가면서 데이터를 생성하는 방법으로, 고차원의 복잡한 데이터도 효과적으로 합성할 수 있으나 학습과정이 복잡하고 시간·비용이 상당 소요

<예시> 생성적 적대 신경망(GANs)을 사용하여 MRI, CT 등 의료 합성 이미지 데이터를 생성, 다양한 질병 상태 진단 기술 개발

[잘못된 합성데이터 생성 예시]

<정형 데이터>

- 다음과 같이 건강 정보 데이터에서 원본데이터의 레코드를 참조하여 변형 생성하거나, 잡음(noise)을 추가하여 생성한 경우는 올바른 합성데이터 생성 방법이라고 할 수 없음

● 원본데이터

나이	BMI	수축기_혈압	이완기_혈압	혈당	...
58	21.103	110	87	133	...
71	19.504	122	75	72	...
48	32.807	101	74	170	...
34	28.401	111	89	120	...

● 합성데이터

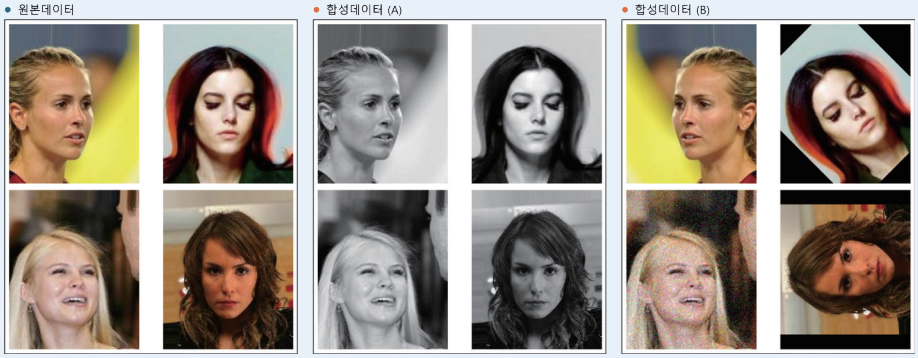
나이	BMI	수축기_혈압	이완기_혈압	혈당	...
59	21.101	113	83	132	...
74	19.503	124	77	73	...
46	32.802	103	72	172	...
35	28.4.07	112	90	124	...

<비정형 데이터>

- 다음과 같이 인물 사진 데이터에서 합성데이터(A)나 합성데이터(B)처럼 원본데이터를 이미지 단위로 변형하는 방법은 적절한 생성 방법이 아님
- 합성데이터(A)는 원본데이터를 흑백으로 처리하였고, 합성데이터(B)는 뒤집기, 회전, 잡음(noise) 추가를 통해 이미지 한 장 단위(sample)로 변형한 예시임. 이처럼 잡음 추가(noise addition)*, 크기조정, 자르기, 색상 조정 등은 합성데이터 생성이라고 할 수 없음

  * 잡음의 크기가 큰 경우 안전성이 보장될 수도 있으나, 안전성이 보장되는 수준에서는 유용성이 보장되지 않을 수 있음

● 원본데이터 ● 합성데이터 (A) ● 합성데이터 (B)



▶ 생성 방법별 세부 내용은 부록1 '합성데이터 생성방법론'을 참고

※ 합성데이터 생성 목적 및 향후 기술 발전 등에 따라 다양한 생성 방법을 자유롭게 개선 또는 변경할 수 있고, 새로운 기법을 활용할 수도 있음

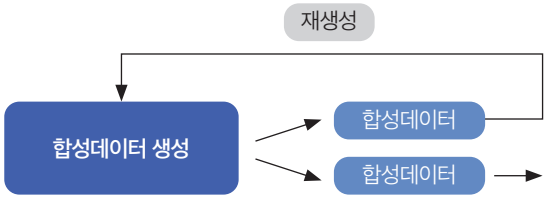
[2단계] - 4 합성데이터 후처리

• 생성된 합성데이터의 형식, 규칙, 논리적 제약 등을 원본데이터와 비교하고 추가 생성, 재생성, 일부 데이터 삭제 등의 후처리가 필요한지 검토하여 처리

※ 후처리를 수행하는 경우 데이터의 유용성을 크게 훼손하지 않아야 함

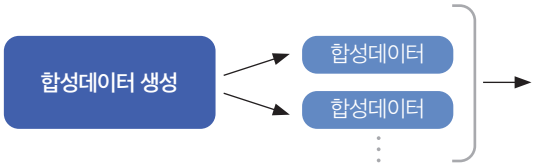
- 합성데이터 활용 시에 중요하게 다뤄질 속성 등이 온전하게 생성되었는지 검토

▶ 일부 데이터만 중요 속성에 문제가 있는 경우 삭제 등의 후처리가 가능



- 생성된 합성데이터가 원래 목표하였던 수량 또는 규모를 충족하는지 검토

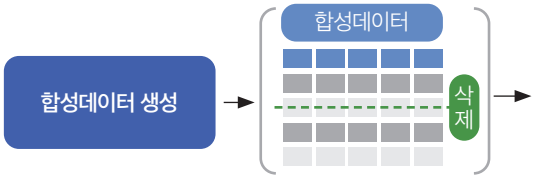
▶ 합성데이터의 생성 목적에 구체적인 수량이 명시되고, 그보다 적게 생성된 경우 필요한 만큼 합성데이터를 추가로 생성할 수 있음



- 생성된 합성데이터의 형식/정의/구조 또는 논리적 제약 등을 검토

▶ '원본데이터 탐색적 분석' 과정의 결과를 바탕으로 검토

▶ 일부 데이터에 문제가 있는 경우 삭제할 수 있으며, 그 비율이 높으면 재생성도 고려할 수 있음



(정형 예시1) 키는 음수가 될 수 없음
(정형 예시2) 고속도로 입차 시간은 반드시 출차 시간보다 빨라야 함
(정형 예시3) 원본데이터의 컬럼 별 범주, 소수점 자릿수 등이 일치하는지를 검사함

(비정형 예시1) 사람 얼굴의 이목구비가 의도한 위치에 있는지 검사함
(비정형 예시2) 성인의 치아 개수는 총 32개이므로 그 이상의 치아가 나타난 이미지는 배제

<택배 배송 데이터 예시>

- 다음과 같이 택배 배송 데이터에서 '배송지'가 시/구/동 등으로 분할되어 있을 때 시/구/동 간 매칭이 틀리거나, '배송 도착' 날짜가 '배송 출발' 날짜보다 이른 경우 등 합성데이터를 생성하는 과정에서 논리적으로 맞지 않는 데이터가 생성될 가능성이 있음
- 이러한 제약조건이 존재하는 원본데이터의 경우 생성 알고리즘에서 제약조건이 지켜지도록 반영하거나, 생성 결과에서 후처리하여 반영함

• 원본데이터

배송지_시	배송지_구	배송지_동	배송 출발	배송 도착	...
서울특별시	영등포구	여의도동	2024-01-17	2024-01-29	...
인천광역시	미추홀구	용현동	2024-02-03	2024-02-06	...
대전광역시	서구	둔산동	2024-03-01	2024-03-07	...
부산광역시	연제구	연산동	2024-03-22	2024-01-29	...

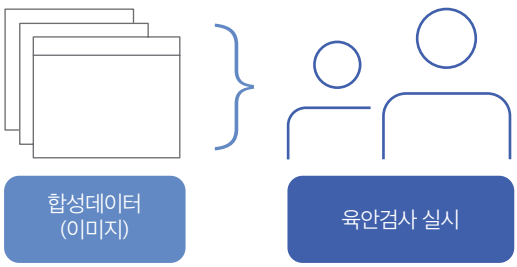
• 합성데이터

배송지_시	배송지_구	배송지_동	배송 출발	배송 도착	...
서울특별시	영등포구	여의도동	2024-02-03	2024-02-06	...
인천광역시	미추홀구	용현동	2024-01-17	2024-01-29	...
대전광역시	서구	둔산동	2024-03-01	2024-03-07	...
인천광역시	영등포구	연산동	2024-03-22	2024-01-29	...

삭제

- 비정형의 경우 합성데이터 규모에 따라 위 항목들 외 육안 검사도 시행하는 것을 권장

▶ 합성데이터(이미지)에 원본데이터와 매우 유사하여 개인식별이 가능한 이미지가 존재하는지 검사



[참고 : 합성데이터 생성 체크리스트]

< 정형 합성데이터 생성 체크리스트 >

단 계	검토 항목
1 원본데이터 탐색적 분석	① 원본데이터는 식별 정보를 포함하고 있는가?
	※ 그 자체로 식별될 위험이 있거나 다른 항목과 쉽게 결합하여 식별될 가능성이 있는 항목(주민등록번호, 전화번호, 상세주소 등)
	② 원본데이터는 서로 다른 데이터를 조합하여 식별가능성이 높아질 수 있는 정보(준식별자 등)를 포함하고 있는가?
	③ 원본데이터는 민감한 정보, 공개시 사회적 파장이 큰 정보를 포함하고 있는가?
	④ 원본데이터에 식별성이 큰 극단값이나 특이정보가 존재하는가?
	※ 극단값은 그 정보만으로 개인을 식별할 수 있는 정보는 아니더라도 고유(희소)한 특성 때문에 개인을 알아볼 가능성이 높은 항목을 말함(희귀 성씨 등 특이한 값, 국내 최고령, 고액 체납 금액 등 극단 값, 특정 데이터 분석 집단에서 희소한 값)
	⑤ 합성데이터를 활용하기 위해 보존 및 재현이 필요한 컬럼과 속성을 사전에 파악하였는가?
	⑥ 원본데이터가 가지고 있는 분포를 합성데이터에 온전히 재현하기 위하여, 각 컬럼 또는 정보 속성 간의 관계를 파악하였는가?
	⑦ 파악한 개인정보 속성, 논리적 제약, 컬럼 및 속성 간의 관계 등을 문서화 하였는가?
2 원본데이터 전처리	① 분석에 필요하지 않은 컬럼 속성 등을 전처리 단계에서 삭제하였는가?
	② 식별자가 분석에 필요한 경우, 식별자를 일련번호 등으로 변환하여 식별성을 최대한 제거하였는가?
	※ 식별자는 암호화나 일련변호 등 재식별이 불가능한 형식으로 변환 후 합성데이터를 생성하여 활용할 수 있음
	③ 특이정보 또는 극단값의 경우, 분석 목적 달성이 가능한 범위에서 삭제나 일반화 등의 적절한 처리를 하였는가?
	※ 사분위수 범위(IQR), Z-score 등을 활용하여 극단값을 찾을 수 있음

단 계	검토 항목
3 합성데이터 생성	① 합성데이터를 생성하는 목적 및 활용 환경 등을 고려하여 생성 알고리즘을 선택하였는가?
	② 합성데이터 생성 시 원본데이터의 유형/형식을 고려한 생성 알고리즘을 선택하였는가?
	※ 원본데이터의 유형이 정형데이터인 경우, 비정형 합성데이터 생성 알고리즘을 사용할 수 없음 ※ 원본데이터가 시계열 자료인 경우, synthpop, GMM, CTGAN 등의 알고리즘을 직접 사용할 수 없음
	③ 생성하는 알고리즘의 특징, 장/단점을 명확하게 파악하여, 적합한 생성방법을 선택하였는가?
	④ 합성데이터 생성에 사용되는 매개변수(parameter)는 생성 단계에서 과적합, 과소적합 등 안전성과 유용성에 안 좋은 영향을 유발할 위험성이 있는가?
	※ 학습 시 과적합 하도록 매개변수를 설정하면 유용성은 좋지만, 안전성에 문제가 발생하므로 생성 단계에서 매개변수 설정이 중요함
	⑤ 합성데이터 생성에 사용되는 알고리즘이 원본데이터에 레코드 단위로 잡음을 추가하지 않고, 비가역적인 알고리즘인가?
	※ 레코드 단위 변조 방식을 사용할 수 없음 ※ 합성데이터에서 원본데이터 역추적이 불가능한 비가역적 알고리즘을 선택하여야 함
4 합성데이터 후처리	① 합성데이터 활용 시 필요한 컬럼, 속성 등이 누락없이 온전히 보존 및 재현되었는가?
	② 생성된 합성데이터는 원본데이터의 형식/구조/규칙과 동일한가?
	③ 생성된 합성데이터에 원본데이터가 가지고 있는 논리적 제약이 온전히 반영되었는가?
	※ 혈압은 음수값이 있을 수 없고 '인천광역시-영등포구'는 실제 데이터 상 있을 수 없음. 또한, 고속도로 출차 시간은 반드시 입차 시간보다 커야 함
	④ 일정 수준의 합성데이터 구축 수량이 필요한 경우, 생성된 합성데이터는 그 수량을 만족하는가?

< 비정형 합성데이터 생성 체크리스트 >

단 계	검토 항목
1 원본데이터 탐색적 분석	① 이미지를 영역별로 분류하여 영역별 속성과 특징을 검토하였는가?
	② 합성데이터 생성 목적을 고려하여, 합성데이터 활용에 필요한 영역과 그렇지 않은 영역을 구분하였는가?
	③ 원본데이터에 메타데이터가 포함되어 있고, 식별 가능성 높은 정보를 포함하고 있는가?
	④ 원본데이터에 개인 식별 가능성이 높은 영역이 존재하는가?
	※ 원본데이터 이미지의 일부분이 개인을 추정 가능할 정도로 식별가능성이 높은 영역을 의미함(X-ray 사진에 의사가 직접 표시한 환자번호, 차량 주행 데이터에서 선명하게 촬영된 영상 속 행인의 얼굴 이미지 등)
	⑤ 이미지 자체 특성에 따른 재식별 가능성이 존재하는 요소를 사전에 파악하였는가 (예시: 사람의 얼굴 이미지 등)?
	⑥ 원본데이터에 식별성이 큰 극단값이나 특이정보가 존재하는가?
	※ 그 정보만으로 개인을 식별할 수 있는 정보는 아니더라도 고유(희소)한 특성 때문에 개인을 알아볼 가능성이 높은 이미지를 의미함(촬영된 신체의 문신, 촬영된 X-ray 이미지의 골격이 희소하여 개인을 알아볼 수 있는 경우)
	⑦ 파악한 각 이미지 영역의 속성, 특징, 논리적 제약 등을 문서화 하였는가?
2 원본데이터 전처리	① 원본데이터가 메타데이터를 포함하고, 이를 사용하지 않는 경우 삭제하였는가?
	※ 데이터에서 환자를 특정할 수 있는 식별정보(진료번호 등)를 삭제하거나, DICOM 등 데이터 포멧을 변경하여 메타데이터 제거
	② 원본데이터에 포함된 메타데이터를 사용하는 경우 비식별처리 등 필요한 처리를 거치고, 적정한 안전성 조치를 취하였는가?
	③ 합성데이터 활용에 필요하지 않은 영역에 대해서 삭제 또는 비식별처리를 진행하였는가?
	※ 구강 촬영 의료이미지 데이터를 사용하여 충치 진단솔루션을 개발하려고 하는 경우, 배경, 코, 촬영 기구 등은 합성데이터 활용에 필요하지 않으므로 삭제
	④ 식별가능성이 높은 영역에 대하여 삭제 또는 비식별처리를 진행하였는가?
	※ 의료이미지에 환자를 특정할 수 있는 식별정보(진료번호 등)가 존재하는 경우 비식별처리 필요
	⑤ 식별성이 큰 특이값 이미지를 삭제하거나, 해당 이미지의 특정될 수 있는 부분을 비식별처리 하였는가?
	※ 원본데이터에 촬영된 신체의 문신이 희소하여 개인을 알아볼 위험성이 존재하는 경우, 해당 이미지의 부분만 마스킹 처리

단 계	검토 항목
3 합성데이터 생성	① 합성데이터를 생성하는 목적 및 활용 환경 등을 고려하여 생성 알고리즘을 선택하였는가?
	② 합성데이터 생성 시 원본데이터의 유형/형식을 고려한 생성 알고리즘을 선택하였는가?
	※ 원본데이터의 유형이 이미지인 경우, 정형 합성데이터 생성 알고리즘을 사용할 수 없음
	③ 생성하는 알고리즘의 특징, 장/단점을 명확하게 파악하여, 적합한 생성방법을 선택하였는가?
	④ 합성데이터 생성에 사용되는 매개변수(parameter)는 생성 단계에서 과적합, 과소적합 등 안전성과 유용성에 안 좋은 영향을 유발할 위험성이 있는가?
	※ 학습 시 과적합 하도록 매개변수를 설정하면 유용성은 좋지만, 안전성에 문제가 발생하므로 생성 단계에서 매개변수 설정이 중요함
	⑤ 원본데이터를 이미지 단위로 변형하거나, 채도 조절, 회전 하는 방법만으로 합성데이터를 생성하지 않고, 비가역적 알고리즘을 사용하여 합성데이터를 생성하였는가?
	※ 이미지의 뒤집기, 회전, 잡음(noise) 추가 등 이미지 변형은 합성데이터 생성이라고 할 수 없음 ※ 합성데이터에서 원본데이터 역추적이 불가능한 비가역적 알고리즘을 선택하여야 함
4 힙성데이터 후처리	① 합성데이터 활용 시 필요한 이미지 영역들이 온전히 보존 및 재현되었는가?
	② 생성된 합성데이터 이미지들이 원본과 비교하였을 때 의미상으로 동일한가?
	※ 논리적으로 존재할 수 없는 이미지 생성 여부를 검증함
	③ 일정 수준의 합성데이터 구축 수량이 필요한 경우, 생성된 합성데이터는 그 수량을 만족하는가?

3. [3단계] 안전성 및 유용성 검증

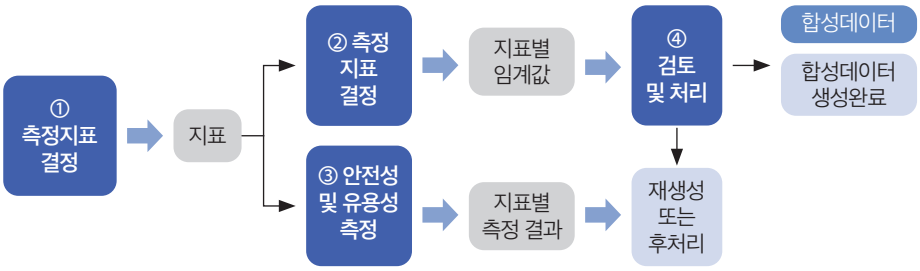
(안전성 검증) : 생성된 합성데이터를 통해 원본데이터 내 개인이 식별될 가능성이 있는지 검증하는 단계로 익명성 인정을 위해 필수절차임

▶ 생성과정과 방법에 따라 합성데이터로부터 개인이 식별될 위험이 존재할 수 있으므로 검증 필요

(유용성 검증) : 합성데이터와 원본데이터의 분포 특성이 얼마나 유사한지, 같은 목표를 달성할 수 있는지 등을 검증하는 단계

▶ 합성데이터는 실제 데이터 대신 활용되므로 합성데이터가 원본데이터를 활용하는 것과 유사한 결과를 가져올 수 있는지 검증 필요

• 안전성, 유용성 검증은 병렬적이면서, 동일한 내부 단계로 수행됨 (①~④)



[3단계] - 1 측정지표 결정

• 안전성 또는 유용성을 측정하기 위한 정량·정성적 지표를 선택

- 합성데이터의 안전성, 유용성을 측정하는 방법은 정형·비정형 등 데이터의 유형, 특징마다 다르게 적용해야 함

가. 안전성 검증

ⅰ) 정형 합성데이터 안전성 검증지표

• 합성데이터에 원본데이터의 개인정보가 존재할 가능성이 있음

- 합성데이터의 레코드가 원본데이터의 레코드와 우연히 일치하거나, 원본데이터의 특성을 학습하는 과정에서 과적합(over-fitting)이 발생하여 원본데이터와 매우 유사하게 생성될 가능성이 존재함

• 합성데이터의 안전성은 원본데이터 자체의 안전성에 영향을 받음

- 원본데이터 내에 유일한 레코드가 많을수록, 준식별자가 많을수록 안전성이 낮기 때문에 안전성 확보를 위한 적절한 처리를 해야 함

- 원본데이터 자체의 안전성과 생성된 합성데이터의 안전성을 비교하여 식별위험성이 얼마나 감소했는지 확인 필요

• 본 안내서에서는 개인정보의 익명화에 사용하는 평가 기준[6]을 활용하여 다음과 같은 세 가지 안전성 지표를 제시함

- 구별 위험도 : 합성데이터 내에 원본데이터와 같은 레코드가 존재할 위험성

※ 합성데이터 내 모든 레코드에 대해 원본데이터 레코드와 비교한 값(같은 값이 있으면1, 없으면0)의 평균을 계산

- 연결 위험도 : 공격자가 원본데이터의 준식별자를 알고 있을 때, 합성데이터를 통해 민감정보를 유추해낼 위험성

※ 연결위험도는 CAP(Correct Attribution Probability)으로 측정 가능

- 추론 위험도 : 합성데이터 내 원본데이터와 같은 레코드는 없지만, 매우 유사하여 특정 개인의 정보를 추론해 낼 위험성

※ 합성데이터와 원본데이터의 두 레코드가 유사하다는 것은 레코드 사이 수학적 거리가 가깝다는 것으로 나타낼 수 있음

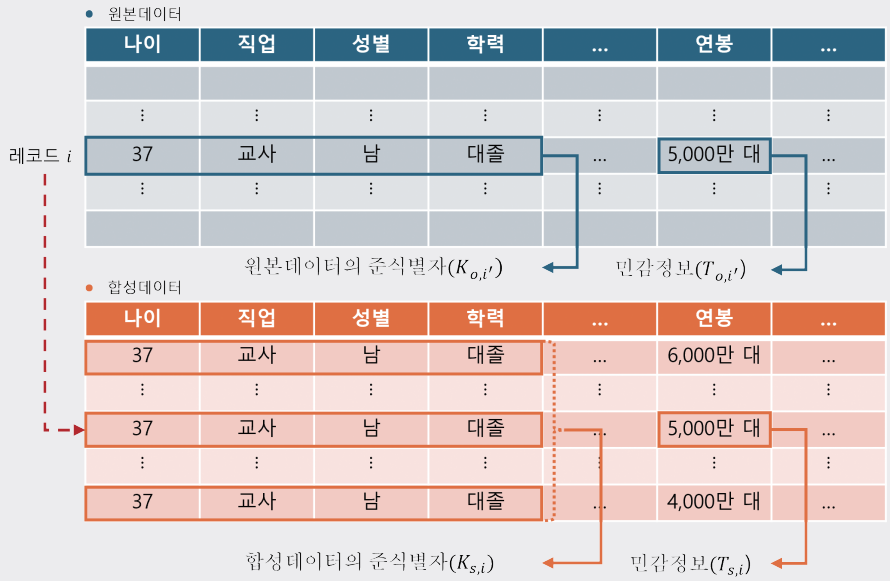
구분	구별 위험도	연결 위험도	추론 위험도
지표	Single out risk Index	Attrribute disclosure risk Index	Inferential disclosure risk Index
세부 설명	• 합성데이터의 특정 레코드(S)를 원본데이터(R) 내에서 찾을 수 있는 비율	• 원본데이터와 합성데이터의 같은 준 식별자(K) 정보를 통해 원본데이터의 민감정보(T)를 알게 될 확률	• 합성데이터의 레코드와 동일한 원본 내 레코드는 없지만, 매우 유사하여 추론을 통해 특정 개인의 민감정보가 유출될 가능성
수식	$\frac{1}{n_s}\sum_{i=1}^{n_s} I(S_i = R_j),$ $\frac{1}{n_s}\sum_{i=1}^{n_s} \frac{I(S_i = R_j)}{f(R_j)}$	$CAP_{s,i} = \frac{\sum_{j=1}^{n_s} I((T_{s,j} = T_{o,i}) \cap (K_{s,j} = K_{o,i}))}{\sum_{j=1}^{n_s} I(K_{s,j} = K_{o,i})}$	$A = \frac{1}{n_s}\sum_{i=1}^{n_s} I(d_S < d_O),$ • d_s : 합성 레코드와 원본에서 합성 레코드와 가장 가까운 레코드의 거리 • d_o: 해당 원본 레코드와 가장 가까운 원본 레코드의 거리
해석	0에 가까울수록 안전성이 높음	0에 가까울수록 안전성이 높음	• A=0 : 추론 위험이 없으나 유용성이 매우 떨어짐 • A=1 : 원본 레코드와 매우 유사하여 추론 위험성이 높음 • A=0.5 : 유용성, 안전성 동시 만족

※ 위의 안전성 지표는 참고를 위한 예시이며, 합성데이터 안전성을 증명할 수 있는 다른 지표와 기준들도 충분히 사용할 수 있음

6) ARTICLE 29 DATA PROTECTION WORKING PARTY, Opinion 05/2014 on Anonymisation Techniques

<연결위험도 측정 예시>

• 아래는 인구 소득 데이터에서 연결 위험도(CAP)를 산출하는 예시임



• 원본데이터의 i번째 레코드는 준식별자가 (나이 : 37세, 직업 : 교사, 성별 : 남, 학력 : 대졸)이고, 민감정보는 (연봉 : 5,000만 대)임
• 해당 원본데이터의 i번째 레코드에 대해, 합성데이터에서 준식별자가 같은 레코드의 개수를 확인함
• 합성데이터에서 원본데이터의 i번째 레코드와 준식별자가 같은 레코드 중 준식별자와 민감정보가 모두 같은 레코드의 개수를 확인함

※ CAP 산출 예시

$$CAP_{s,i}(\text{레코드 } i\text{의 } CAP) = \frac{\sum_{j=1}^{n_s} I((T_{s,j} = T_{o,i}) \cap (K_{s,j} = K_{o,i}))}{\sum_{j=1}^{n_s} I(K_{s,j} = K_{o,i})}$$

$$= \frac{\text{원본데이터의 레코드 } i\text{와 준식별자, 민감정보가 모두 같은 합성데이터의 레코드 개수}}{\text{원본데이터의 레코드 } i\text{와 준식별자가 동일한 합성데이터의 레코드 개수}}$$

$$= \frac{1}{3} = 0.333$$

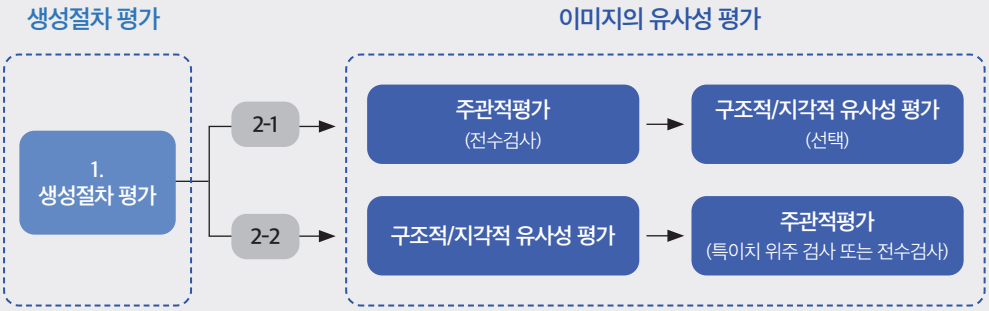
▶ 세부내용은 부록2 '합성데이터 안전성 검증 기준' 참고

ii) 비정형 합성데이터 안전성 검증지표

- 비정형 합성데이터는 이미지가 정확하게 같아야 식별되는 것이 아니고, 이미지의 유사성이 높으면 식별성이 높아지기 때문에 정형 합성데이터에 비해 식별 여부의 판단이 어려움
- 비정형 합성데이터는 생성된 합성데이터의 안전성을 정량적으로만 판단하는 것에는 한계가 있음
- 따라서 합성데이터의 안전성을 검증하기 위해서는 데이터 자체의 위험성 뿐만 아니라 생성과정 전반에 대한 절차적 타당성과 생성된 합성데이터 자체의 안전성을 모두 평가하여야 함

① 생성과정 평가 : 데이터 전처리 여부, 알고리즘의 적정성 등 절차적 타당성 평가
▶ 원본데이터에 개인식별 가능한 요소가 포함된 경우, 전처리 시 해당 요소가 활용 목적 범위에 포함되지 않는다면 비식별처리가 필요
▶ 비가역적 알고리즘을 사용하여 완전재현 방식으로 합성데이터를 생성할 필요

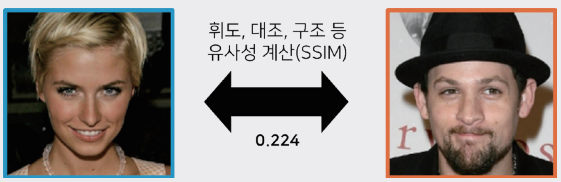
② 이미지 유사성 평가 : 절차적 타당성을 준수하였음에도 원본과 거의 동일한 합성데이터가 생성된 경우를 검증
▶ 구조적·지각적 유사성에 대한 정량적 평가, 육안을 통한 정성적 평가 수행



<구조적 유사성 평가 예시>

▶ 원본데이터와 합성데이터의 모든 이미지 간 유사성을 비교하여 구조적으로 과도하게 비슷하게 생성된 이미지가 있는지를 측정
▶ 이미지의 구도, 휘도, 대비를 하나의 품질 점수로 통합한 지표인 SSIM(Structural Similarity Index Measure) 지표 사용
▶ 두 이미지를 비교하는 지표로, 서로 다른 유사성을 가진 이미지일수록 0에 가까움

$$SSIM(x,y) = \frac{(2\mu_x\mu_y + c_1)(2\sigma_{xy} + c_2)}{(\mu_x^2 + \mu_y^2 + c_1)(\sigma_x^2 + \sigma_y^2 + c_2)}$$



<지각적 유사성 평가 예시>

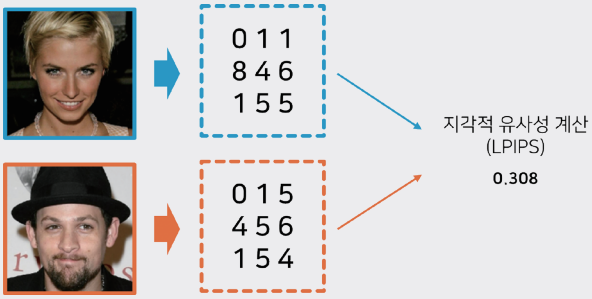
▶ 사람의 시각인지능력 측면에서 원본데이터와 합성데이터 간 시각적 유사성을 측정

▶ LPIPS(Learned Perceptual Image Patch Similarity) 지표 사용

▶ 두 이미지를 비교하는 지표로, 서로 다른 유사성을 가진 이미지일수록 1에 가까움

* 다른 지표와의 통일 및 점수산출 간편성을 위해 1에서 뺀 값을 사용하여, 서로 다른 유사성을 가진 이미지일수록 0에 가까워지도록 사용 가능 (지각적 유사도= 1-LPIPS)

$$LPIPS = \sum_l \frac{1}{H_l W_l} \sum_{h,w} \| w^l \odot (\hat{y}^l_{hw} - \hat{y}^l_{0hw}) \|_2^2$$



▶ 세부내용은 부록2 '합성데이터 안전성 검증 기준' 참고

- 합성데이터 안전성 검증에 대한 방법론들과 지표들은 여러 방법 중 일부를 제시한 것으로 표준을 규정하는 것이 아니며, 향후 더 폭넓은 연구가 이루어져야 함
- 안전성 검증지표와 방법은 합성데이터 생성자가 직접 정해 사용할 수 있지만, 객관적으로 타당성이 인정되는 방법론이어야 함

나. 유용성 검증

• 합성데이터의 유용성은 아래 측도 등으로 평가할 수 있음

- 합성데이터의 통계적 분포가 원본데이터의 분포와 유사해야 하고, 원본데이터의 분석 결과와 합성데이터의 결과가 유사해야 함

ⅰ) 정형 합성데이터 유용성 측정 방법(예시)

측도	설명	측도 예시 및 수식
일차원 분포 유사성	• 원본데이터와 합성데이터 컬럼별로 분포 유사성 측정	- 얀센 샤논 발산 (Jenson- Shannon divergence) $JSD(P \parallel Q) = \frac{1}{2} D_{KL}(P \parallel \frac{(P+Q)}{2}) + \frac{1}{2} D_{KD}(Q \parallel \frac{(P+Q)}{2})$ - 콜모고로프-스미르노프 검정 (Kolmogorov Smirnov Test) $D_{n,m} = \max\lvert F_{1,n}(x) - F_{2,m}(x)\rvert$ - 카이제곱 검정 (Chi-square Test) $\chi^2 = \sum_{i=1}^{c} \frac{(O_i - E_i)^2}{E_i} \sim \chi^2(c-1)$
2차원 관계 유사성	• 각 컬럼 간 존재하는 상관관계의 유사성 검증 • 원본과 합성데이터의 컬럼 간 상관계수 행렬 차이 값 표준편차를 계산 ※ 수치형·범주형 간 상관관계 산출식은 없음 ※ 전체 데이터의 2차원 관계 유사성을 계산하기 위해서 분산분석의 결정계수 활용 (부록3 '합성데이터 유용성 검증 방법' 참고)	수치형 : 피어슨 상관계수 $r_{XY} = \frac{\sum_{i}^{n}(X_i - \overline{X})(Y_i - \overline{Y})}{\sqrt{\sum_{i}^{n}(X_i - \overline{X})^2}\sqrt{\sum_{i}^{n}(Y_i - \overline{Y})^2}}$ 범주형 : 크래머 V 상관계수 (Cramer's V) $V = \sqrt{\frac{\chi^2}{n(k-1)}}$ 범주형, 수치형 : 분산분석(ANOVA)의 결정계수(SSR/SST) SST(Sum of Square Total) = SSE(Sum of Square Error) + SSB(Sum of Square Between)
구별 불가능성 (성향점수)	• 원본데이터와 합성데이터를 섞은 데이터셋에 대해 원본인지 합성데이터인지 구분하는 분류 모형을 만들고 분류 성능으로 평가	- 성향점수평균제곱오차(pMSE) $pMSE = \frac{1}{n_o + n_s} \sum_{i=1}^{n_o+n_s} (\hat{p_i} - \frac{n_s}{n_o + n_s})^2$ - PO50 - SPECKS
모형 성능 유사성	• 활용 목적에 따라 합성데이터가 원본데이터와 동일한 성능을 내는지를 검증 • 회귀계수 신뢰구간 비교 등	- 분류 모형 : 정확도(Accuracy), 재현율(Recall), 정밀도(Precision), F1 점수(score) - 회귀모형 : 평균제곱오차(MSE) $CI = \frac{1}{2}\left[\frac{U_i - L_i}{U_{ori} - L_{ori}} + \frac{U_i - L_i}{U_{syn} - L_{syn}}\right]$

ii) 비정형 합성데이터 유용성 측정 방법 (예시)

측도	설명	측도 예시 및 수식
모델 성능	• 활용 목적에 따라 합성데이터가 원본	정확도(Accuracy), 재현율(Recall), 정밀도(Precision), F1 score 등
Visual Turing Test	• 사람의 육안 검사 등을 통해, 원본과 합성데이터를 맞추는 실험	정확도(Accuracy), 재현율(Recall), 정밀도(Precision), F1 score 등 $VTT\text{결과} = \frac{\text{응답자가 찾아낸 합성데이터의 개수}}{\text{전체 합성데이터의 개수}}$
이미지 품질	• 원본데이터와 합성데이터 간 품질을 정량적으로 측정하여 비교 (통계적으로 얼마나 유사한지 등)	Inception score, FID score 등 $FID(x,g) = \parallel \mu_x - \mu_g \parallel_2^2 + Tr(\Sigma_x + \Sigma_g - 2(\Sigma_x\Sigma_g)^{\frac{1}{2}})$

▶ 세부내용은 부록3 '합성데이터 유용성 검증 방법' 참고

[3단계] - ❷ 지표별 임계값 산출

• 안전성을 측정하기 위한 지표의 기준값(임계값) 수준을 산출

- 안전성 검증지표를 통해 측정된 값이 안전한 것인지 평가할 수 있는 기준이 필요함에 따라, 판단기준이 되는 임계값을 산출하여 임계값 만족여부를 확인
- 지표별로 절대적 임계값이 존재하는 경우, 해당 임계값으로 설정

  * (예시) '연결 위험도'의 임계값은 절대값 지정이 가능

- 절대적 임계값이 존재하지 않는 경우, 원본데이터의 종류와 활용 목적에 따라 합리적인 임계값 수준을 정의하고 그 근거를 명시 및 보관

  * (예시) '구별 위험도', '추론 위험도'의 경우 적절함 임계값 설정 가능
  ※ 합성데이터의 활용목적과 활용범위에 따라 필요 안전수준을 구분하여 높은 안전성이 요구되는 경우에는 엄격한 임계값 기준을, 보통 수준인 경우에는 상대적으로 완화된 임계값 기준을 적용할 수 있음

▶ 부록4 '정형 데이터 검증 지표 임계값 산정 방법' 참고

[3단계] - 3 안전성 및 유용성 측정

- 검증지표, 임계값 산정 등을 바탕으로 합성데이터로부터 원본데이터를 식별할 수 있는지를 다각도로 검토함(안전성)
- 합성데이터의 통계적 분포가 원본데이터와 얼마나 유사한지 등을 검토함(유용성)

※ 비정형 데이터는 정형 데이터의 안전성·유용성 검증지표로는 측정할 수 없음
※ 안전성·유용성 측정 시 활용되는 원본데이터는 전처리 작업이 완료된 원본데이터임

▶ (안전성 참고) 부록2 '합성데이터 안전성 검증 기준' 참고
▶ (유용성 참고) 부록3 '합성데이터 유용성 검증 방법' 참고

[3단계] - 4 검토 및 후처리

- 지표별로 정의된 임계값, 측정결과, 데이터 탐색적 분석에서 파악된 제약사항 등을 종합적으로 검토하여 합성데이터의 생성 결과를 검토하고, 필요 시 일부 레코드 삭제 등 후처리 진행

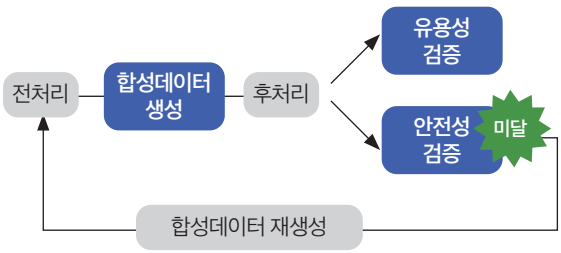
※ 임계값을 충족하지 못하는 합성데이터는 제외하고, 임계값을 충족하는 합성데이터가 목표규모보다 많은 경우 유용성이 높은 순으로 채택하는 방법 등 사용

- 측정 내용은 문서화하여 심의위원회 평가 시 활용. 특히, 안전성의 경우 검토 결과서 작성이 중요

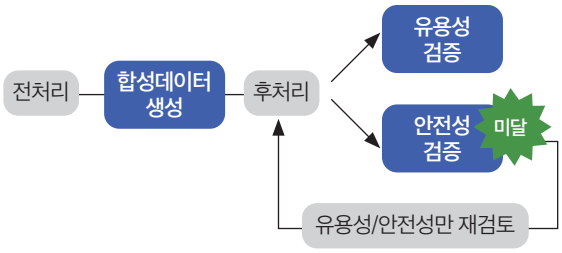
▶ 부록5, 6 내 '합성데이터 안전성 자체 검토 결과서' 참고

- 검토결과 기준을 만족하지 못하는 경우, 재처리 필요 여부를 결정하여 이전 단계로 회귀하거나, 후처리를 수행

▶ 합성데이터를 다시 생성하는 경우, 합성데이터 생성 전 전처리 단계로 돌아가 방법론 등을 변경 후, 재생성



▶ 합성데이터를 다시 생성하지 않고 일부 레코드의 삭제 등 후처리 후 안전성 및 유용성만을 재검증하는 것도 가능함



4. [4단계] 심의위원회 평가

• 합성데이터를 익명정보로 평가 받기 위해 심의위원회를 구성하여 충분히 안전하게 생성되었는지 최종 검토

• 합성데이터의 대외공개 등을 위해 보다 객관적·전문적인 검토가 필요하다고 판단될 시, 내외부 전문가 심의를 진행

- (기초자료 작성 및 필요서류 확인) 합성데이터 심의에 필요한 정보들이 포함된 기초자료를 작성
  ※ [1]~[3]단계에서 작성하였던 필요서류와 법률사항에 대한 점검 수행, 자료 보완
- (심의위원회 구성) 합성데이터 전문가, 법률 전문가, 개인정보 전문가 등이 포함된 합성데이터 심의위원회를 구성
  ※ 심의 위원은 최소 3명 이상(과반수는 합성데이터 생성·활용에 관여하지 않은 외부 인원)으로 구성할 것을 권고
- (심의 진행) 1~3단계의 합성데이터 생성 절차와 합성데이터의 생성 결과를 검토하여 합성데이터가 활용 목적, 활용 범위를 고려했을 때 적합하게 생성되었는지 평가

[심의 단계별 검토 및 처리 표]

구분	검토 내용	부적정 시 고려사항
① 활용 목적, 활용 범위 설정 등	• 사전준비 단계에서 필요서류가 적합하게 작성되었는지 검토 • 활용 목적 및 범위를 구체적으로 설정하였는지 확인 • 원본데이터에서 합성데이터 생성·활용에 불필요한 부분은 제외하였는지 확인	자료 보완작성, 목적 명확화 및 재설정 등
	검토자료 : 합성데이터 활용 계획서, 개인정보 처리 위·수탁(합성데이터 관련) 계약서 ※ 필요 시 제3자 제공에 대한 계약서, IRB 심의결과통지서, 원본데이터 이용 협약 등	
② 식별 위험성	• 원본데이터의 식별 위험성 판단 항목을 누락없이 검토하였는지 확인	식별 위험성 재검토
	검토자료 : 원본데이터 명세서, 항목별 개인정보 처리계획	
③ 합성데이터 생성	• 원본데이터 특성 검토 결과 및 합성데이터 생성계획을 반영하여 합성데이터를 생성하였는지 검토 • 합성데이터 생성 모형이 적합한지 검토	합성데이터 재생성, 생성계획 보완
	검토자료: 합성데이터 생성 기초자료 명세서 ※ 그 외 합성데이터 생성 시 사용한 방법을 설명할 수 있는 자료	
④ 안전성 검증	• 생성된 합성데이터에 대해 안전성 검증을 수행하였고, 지표와 임계값 등을 설정였으며 결과는 기준에 충족하는지 검토 • 안전성을 만족시키기 위해, 필요한 경우 합성데이터 재생성 및 후처리 등을 올바르게 수행하였는지 검토	합성데이터 재생성 및 후처리 등
	검토자료 : 합성데이터 명세서, 합성데이터 안전성 자체 검토 결과서* * 안전성 검증 단계에서 측정된 정량·정성적 결과에 대한 자료(자유서식) ※ 필요 시 안전성 측정 시 사용한 지표 또는 생성모델의 신뢰성을 설명할 수 있는 근거 자료	
⑤ 활용환경의 보호조치	• 합성데이터를 활용하기 위해 활용환경의 보호조치가 필요한 경우, 기술적/관리적 보호조치를 설정하였는지 검토 • 합성데이터의 재식별 위험 등에 대한 관리계획이 있는지 검토	합성데이터 재생성 또는 활용환경 변경
	검토자료 : 자체 관리계획서	

- (보완 및 재평가) 심의결과가 부정적인 경우, 심의위원회의 의견을 반영하여 보완 또는 재평가받아야 함

5. [5단계] 활용 및 안전한 관리

• 구축한 합성데이터를 활용 목적·범위에 맞게 활용하고 후속조치로서 관리하는 단계

[5단계] - 1 활용

• 합성데이터의 안전성 검증, 심의위원회 평가 등이 완료된 경우 해당 데이터를 이용 또는 제공할 수 있음

- 단, 계약 등으로 인해 활용에 제약이 있는 합성데이터의 경우 허가된 목적·범위·대상 내에서만 활용되도록 관리체계* 마련 필요

* 안전조치 시행, 사용기간 확인, 파기 이전까지 주기적 위험성 검토 등

[5단계] - 2 안전한 관리

• 설정한 합성데이터 활용 목적 및 범위에 맞게 활용되도록 관리

가. 익명정보 성격의 합성데이터

• 기본적으로 익명정보로 인정받은 합성데이터는 자유롭게 활용이 가능하고 별도의 관리는 필요하지 않음

- 다만, 추후 식별자·준식별자 관련 정보로 인한 재식별 가능성 등 잔여 위험*을 판단하고, 위험 예상 시 모니터링, 합성데이터 사용기간 설정 등을 권고

*** [잔여 위험 예시 및 대응방향][7]**

항목	고려사항 및 대응방향
① 합성데이터에서 발견될 수 있는 새로운 식별 위험 대비	• 합성데이터 분석 또는 다른 데이터셋과 함께 분석함으로써 원본데이터에 대한 추론 식별 등이 얻어질 수 있음 • 합성데이터 생성자는 이러한 식별위험이 민감하거나 오해를 불러일으킬 수 있는지 평가해야 함
② 멤버십 추론 공격으로 인한 잠재적 영향 대비	• 멤버십추론공격이란, 공격자가 합성데이터의 정보를 사용하여 특정 대상 그룹이 원본데이터에 포함되어 있었다는 것을 확인하는 것 • 멤버십 추론 공격이 발생하였을 때의 잠재적 영향도를 고려해야 함
③ 악의적인 이용 대비	• 합성데이터 이용자가 합성데이터를 유출하는 위험이 발생할 수 있음 • 특수한 기술이나 지식을 가진 합성데이터 이용자가 공개된 지식이나 전문 지식을 조합하여 합성데이터에서 개인을 재식별할 수 있는 가능성을 고려해야 함
④ 변화하는 환경에 대한 모니터링	• 합성데이터의 재식별 위험 가능성은 시간이 지남에 따라 증가 • 컴퓨팅 파워의 증가와 데이터 연결 기술의 발전으로 인한 재식별 가능성을 고려하고 모니터링 해야 함
⑤ 합성데이터 생성모델 및 매개변수 유출 대비	• 합성데이터를 생성하는 데 사용한 생성모델과 매개변수(parameter)는 공격자가 모델 반전 공격(model inversion attack)을 수행하는데 유용한 정보일 수 있음

나. 가명정보 성격의 합성데이터

• 「개인정보보호법」상 안전조치, 재식별 가능성 모니터링, 기록관리, 파기 등 적절한 관리*를 수행

* 세부 내용은 「가명정보 처리 가이드라인」 참고

▶ 부록7 내 '양식 1-6) 합성데이터 관리대장' 참고

7) 싱가포르 개인정보보호위원회(PDPC), 2024

제4장

활용 안내사항

Ⅳ 활용 안내사항

1. 본 안내서와 「합성데이터 생성모델('24.5, 개인정보위)」, 실제 데이터*를 연계하여 업무·교육 등에 활용 가능

* 가명정보지원플랫폼(dataprivacy.go.kr)

2. 본 안내서를 참고함에 있어 다음과 같은 사항에 유의하여야 함

① 참고용 절차·방법론의 제시

- 안내서에서 소개한 합성데이터 생성 절차 및 생성방법, 안전성·유용성 검증지표 등은 하나의 예시를 제시한 것임.

※ 시계열 자료와 같이 레코드 간 상관관계가 존재할 경우 본 안내서에서 제시하는 방법론을 그대로 적용할 수 없고 수정 적용해야 함

- 절차를 의무화하거나 표준화하고자 하는 의도가 없으며, 생성·활용자의 판단에 따라 적절한 절차를 거쳐 다양한 방법론, 지표들을 자유롭게 사용할 수 있음.
- 추후 주요 방법론이 형성되거나 합성데이터 생성 절차 등에 대한 사회적 합의가 이루어진다면 관련 내용을 담아 안내서도 개정될 예정임.

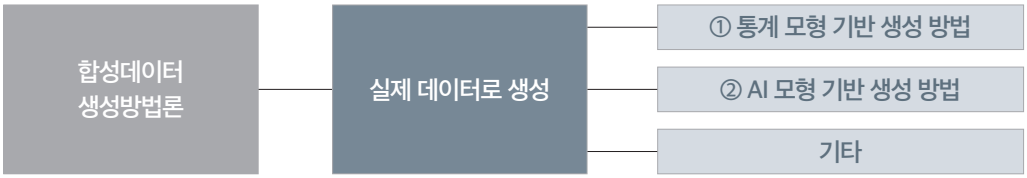
② 비정형 데이터 관련 추가 연구 필요

- 본 안내서는 이미지 형식의 비정형 합성데이터에 대해서도 다루고 있으나, 비정형데이터의 안전성·유용성 검증 등에 대해서는 여전히 추가 연구가 필요한 상황임.
- 향후 기술발전 및 연구결과에 따라 비정형 합성데이터에 대해서는 추가 안내가 필요할 것으로 보여짐.
- 또한, 이미지가 아닌 영상, 음성 및 멀티모달 데이터 등 다양한 비정형 합성데이터에 대한 안내도 추후 과제로 남아있음.

부록

Ⅴ 활용 안내사항

부록1. 합성데이터 생성 방법론



❶ 통계 모형 기반 생성 방법

▶ 원본데이터의 통계적 분포 특성을 추정하여 원본과 유사한 분포 특성을 가진 가상의 데이터를 생성하는 방법

<장점> 통계 모형 기반 생성 방법은 생성 시간이 짧고 상대적으로 유용성이 높은 데이터를 생성할 수 있음

<단점> 시계열 데이터나 네트워크 데이터, 비정형데이터와 같이 복잡한 데이터 구조일 경우, 통계 모형 기반 생성 방법이 이러한 복잡성을 충분히 반영하기 어려울 수 있음

<예시>
- 고객의 거래 금액, 거주지, 신용도 등을 통해 카드 연체 여부를 예측하는 모형 개발을 위해 Synthpop 패키지로 합성데이터 생성
- 대형할인점에서 고객의 월별 구매 데이터를 분석하여 마케팅 전략을 최적화하기 위해 Synthpop 알고리즘을 사용하여 합성데이터 생성

[통계 모형 기반 생성 방법 - 주요 생성 방법]

생성 방법	설명	
Synthpop	• 조건부 확률분포를 이용해 순차적으로 데이터 생성 • 의사결정나무, 회귀모형 등의 알고리즘을 사용한 방법 (CART, Linear regression, Logistic regression, Random sample)	
	장점	생성 속도가 빠르고 CART 기법의 경우 타 기법 대비 유용성이 높은 편
	단점	CART 모형의 경우, 원본데이터를 트리에 넣고 임의 추출하는 방식으로 원본과 같은 레코드가 생성될 가능성이 타 기법에 비해 높음
가우시안 혼합 모델 (Gaussian Mixture Model)	• 원본데이터를 여러 정규분포 혼합 모형으로 가정하여 합성데이터를 생성하는 방법	
	장점	범주형 컬럼이 많지 않고 연속형 컬럼이 많은 경우에 효과적임
	단점	정규분포 가정이 적합하지 않은 경우, 합성데이터 품질이 떨어질 수 있음
베이지안 네트워크 (Bayesian Network)	• 변수 간의 조건부 의존관계를 그래프로 표현하며 조건부 확률분포를 통해 데이터 생성	
	장점	변수 간의 조건부 의존관계를 잘 생성함
	단점	변수가 많을 경우, 가능한 네트워크 구조의 경우의 수가 기하급수적으로 증가하고 계산 비용 증가

❷ AI 모형 기반 생성 방법

▶ 딥러닝 생성형 모델을 이용하여 원본데이터의 분포 특성을 학습시키고 원본데이터의 분포 특성에 근사하는 데이터를 생성하는 방법

<장점> 정형 데이터뿐만 아니라 이미지, 텍스트, 오디오, 동영상 등 다양한 데이터 유형의 복잡한 패턴과 상호작용 학습이 가능함

<단점> 딥러닝 모델의 학습 시간과 비용이 많이 소요됨. 또한 모형의 매개변수(parameter) 조정이 복잡하므로 최적의 성능을 위해 많은 실험과 조정이 필요함. 과적합이 일어날 경우, 원본데이터를 거의 그대로 생성하는 문제점도 존재함

<예시> 딥러닝 생성모델인 GANs를 사용하여 MRI, CT 등 의료 합성 이미지 데이터를 생성하여 다양한 질병 상태 진단 기술 개발

[AI 모형 기반 생성 방법 - 주요 생성 방법]

생성 방법	설명	
변분 오토인코더 (Variational Autoencoders)	• 인코더(encoder)와 디코더(decoder)로 구성된 신경망 구조를 활용하여 원본데이터 모수의 잠재 공간(latent space) 표현을 학습하는 모형 • 인코더(encoder)는 원본데이터를 정규분포를 따르는 확률 분포의 매개변수(평균과 분산) 잠재 공간으로 매핑. 잠재 공간에서 디코더를 이용해 원본데이터와 유사한 합성데이터를 생성함	
	장점	GAN에서 비해 계산량이 적음
	단점	GAN에 비해 복잡한 패턴을 학습하는 데 한계가 있음
생성적 적대 신경망 (GANs)	• 생성자(generator)와 판별자(discriminator)라는 두 신경망을 사용 • 생성자와 판별자로 성능을 최적화해 데이터를 생성 • GAN 단독 모형으로 유용성 확보가 어려워 데이터에 따라 CTGAN, Style GAN 등 GAN 변형 모형을 사용함	
	장점	이미지와 같은 비정형데이터 합성 성능이 뛰어남
	단점	유용성을 확보하기 위해 미세조정(fine-tuning)이 필요하고 한번 생성에 비용과 시간이 크게 소요
확산모델 (Diffusion Models)	• 원본데이터에 점진적인 여러 단계의 잡음(noise)을 추가하여 확률적 분포에 가깝도록 확산시킨 후 역으로 확률적 모형화(modelling)를 통해 잡음(noise)를 제거해 나가면서 데이터를 생성	
	장점	고차원 데이터의 복잡한 분포를 효과적으로 모형화(modelling)하여 GAN보다 안정적이고 뛰어난 고품질 데이터를 생성
	단점	여러 단계의 확산과 역확산 과정으로 인해 학습과정이 복잡하고 구현과 미세조정(fine-tuning)이 어려움. 상당한 시간과 비용이 요구됨

부록2. 합성데이터 안전성 검증 기준

▶ 아래 검증 기준은 참고를 위한 예시이며, 합성데이터 안전성을 증명할 수 있는 다른 지표들도 사용할 수 있음

[정형 합성데이터 안전성 검증]

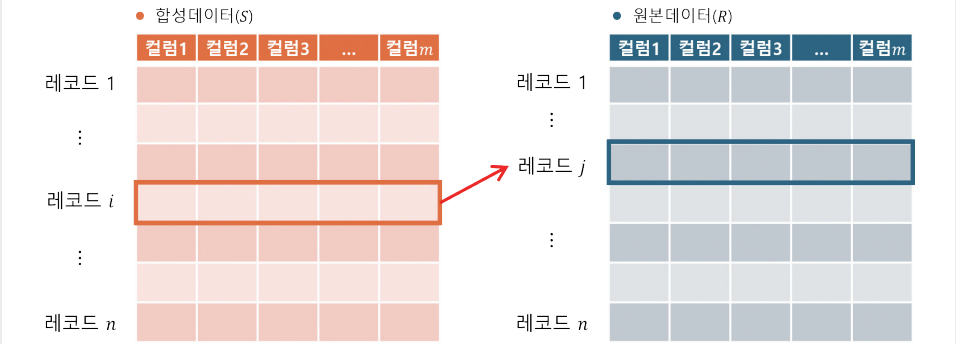
❶ 구별(Single out) 위험도

▶ 합성데이터 내 원본데이터와 같은 레코드가 존재할 위험성을 나타냄

$$\text{구별 위험도 산출식} : \frac{1}{n}\sum_{i=1}^{n} I(S_i = R_j)$$

1. 합성데이터의 i번째 레코드(S_i)를 선택함
2. 해당 레코드 와 원본데이터의 레코드 전체를 비교하여 같은 레코드(R_j)가 있으면 1, 없으면 0을 기록함
   * $I(S_i = R_j)$는 지시함수임
3. (1~2) 과정을 합성데이터의 모든 레코드 (S_1, S_2, ⋯, S_n)에 대해 n(합성데이터 레코드 수)회 반복하여 전체 결과에 대한 평균을 계산함

<예시>



- 구별 위험도가 0이면 원본데이터와 동일한 레코드가 존재하지 않는다는 것을 의미함
- 구별 위험도를 낮추기 위해 합성데이터에서 원본과 같은 레코드를 삭제할 수 있음
  ※ 단, 같은 레코드를 삭제할 시 유용성이 낮아질 수 있고, 원본데이터 내 레코드 간 중복도가 높으면 많은 레코드가 삭제될 수 있음
- 이때는 위 식을 변형해 원본데이터 내 중복 레코드 수 $f(R_j)$를 고려하여 산출할 수 있음

$$\frac{1}{n}\sum_{i=1}^{n}\frac{I(S_i = R_j)}{f(R_j)}$$

- $f(R_j)$는 원본데이터의 j번째 레코드와 원본데이터 내에 같은 레코드 수를 뜻함

❷ 연결(Linkability) 위험도

▶ 공격자가 원본데이터의 준식별자를 알고 있을 때, 합성데이터를 통해 개인의 민감정보를 유추해 낼 위험성을 나타냄

▶ 연결 위험도는 아래 CAP(Correct Attribution Probability)[8]으로 측정할 수 있음

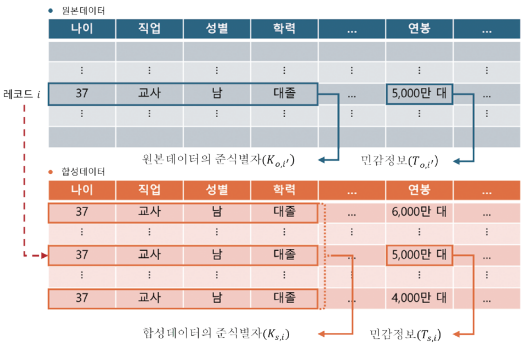
▶ CAP은 준식별자, 민감정보는 모두 범주형일 때 적용할 수 있음(수치형은 적절한 범주화후 계산해야 함)

연결 위험도 산출식 : $$CAP_{s,i} = \frac{\sum_{j=1}^{n_s} I((T_{s,j} = T_{o,i}) \cap (K_{s,j} = K_{o,i}))}{\sum_{j=1}^{n_s} I(K_{s,j} = K_{o,i})}$$

1. 원본데이터의 i번째 레코드를 선택함
2. 해당 레코드 준식별자 $K_{o,i}$와 합성데이터 레코드 준식별자를 비교하여, 같은 준식별자($K_{s,j}$)가 있으면 1, 없으면 0을 기록함
   * $I(K_{s,j} = K_{o,i})$는 지시함수임
3. 해당 레코드와 같은 준식별자를 가진 합성 레코드 중에서 민감정보 $T_{s,j}$와 $T_{o,i}$가 같은지 확인하여, 있으면 1, 없으면 0을 기록함
   * $I(K_{s,j} = K_{o,i}, T_{s,j} = T_{o,i})$는 지시함수임
4. (1~3) 과정을 원본데이터의 모든 레코드(O_1, O_2, ⋯, O_n)에 대해 n(원본데이터 레코드 수)회 반복하여 각각에 대해 원본데이터와 얼마나 일치하는지를 계산함
5. CAP은 원본 레코드 단위로 계산되고, 임계값 기준보다 높은 레코드는 해당 합성 레코드를 일부 삭제하여 CAP을 임계값 이내로 낮출 수 있음

<인구 소득 데이터 예시>

• 아래는 인구 소득 데이터에서 연결 위험도(CAP)을 산출하는 예시임



• 원본데이터의 ¿번째 레코드는 준식별자가(나이 : 37세, 직업 : 교사, 성별 : 남, 학력 : 대졸)이고, 민감정보는(연봉 : 5,000만 대)임

8) Taub, J., Elliot, M., Pampaka, M. and Smith, D. (2018). Differential Correct Attribution Probability for Synthetic Data: An Exploration. J. Domingo-Ferrer and F. Montes (Eds.): PSD 2018, LNCS 11126. pp. 122-137

<인구 소득 데이터 예시>

- 해당 원본데이터의 i번째 레코드에 대해, 합성데이터에서 준식별자가 같은 레코드의 개수를 확인함
- 합성데이터에서 원본데이터의 i번째 레코드와 준식별자가 같은 레코드 중 준식별자와 민감정보가 모두 같은 레코드의 개수를 확인함

※ CAP 산출 예시

$$CAP_{s,i}(\text{레코드 } i\text{의 } CAP) = \frac{\sum_{j=1}^{n_s} I((T_{s,j} = T_{o,i}) \cap (K_{s,j} = K_{o,i}))}{\sum_{j=1}^{n_s} I(K_{s,j} = K_{o,i})}$$

$$= \frac{\text{원본데이터의 레코드 } i\text{와 준식별자, 민감정보가 모두 같은 합성데이터의 레코드 개수}}{\text{원본데이터의 레코드 } i\text{와 준식별자가 동일한 합성데이터의 레코드 개수}}$$

$$= \frac{1}{3} = 0.333$$

- 해당 지표는 준식별자(K)와 민감정보(S)가 모두 범주형 변수일 경우 사용 가능
  ※ 준식별자나 민감정보가 수치형 변수일 경우, 지표를 산출하는 과정에서 범주형 변수로 변환해 계산할 수 있음
- 준식별자(K)는 여러 개의 변수를 가질 수 있고, 민감정보(S)가 여러 개일 경우 각각의 민감정보에 대한 CAP을 별도로 구해야 함
- CAP은 원본데이터 레코드별로 값이 측정되고 이 값이 작을수록 안전성이 높음
- 적정한 CAP 임계값을 설정하고 안전성 확보를 위해 합성데이터 내 임계값을 증가시키는 레코드를 삭제하여 연결 위험을 감소시킬 수 있음
- CAP의 임계값은 생성자가 절대적인 기준으로 설정함
- 임계값은 0.7 이하의 값으로 설정하는 것이 바람직하고, 너무 작으면 만족하지 않는 합성데이터의 레코드가 많이 삭제되어 유용성이 감소하므로 적절하게 설정해야 함

❸ 추론(Inference) 위험도[9][10]

▶ 합성데이터 내 원본데이터와 같은 레코드는 없지만 매우 유사하여 특정 개인의 정보를 추론해 낼 위험성을 나타냄

▶ 합성데이터와 원본데이터의 두 레코드가 유사하다는 것은 레코드 사이 수학적 거리가 가깝다는 것으로 나타낼 수 있음

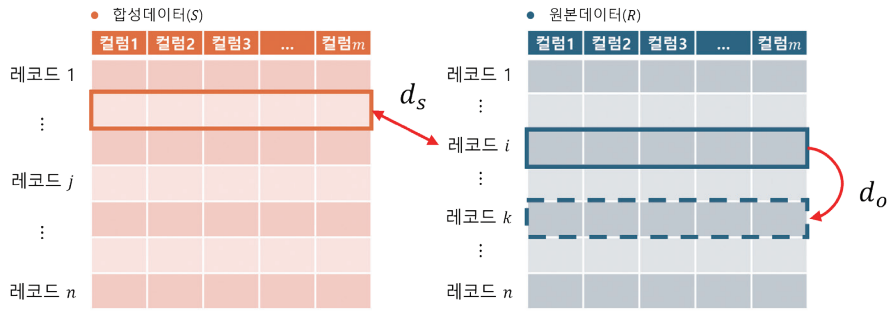
추론 위험도 산출식 : $p_A = \frac{1}{n}\sum_{i=1}^{n_s} I(d_s < d_o)$

9) Donghoon Jeong, Joseph H. T. Kim, and Jongho Im. (2023). A New Global Measure to Simultaneously Evaluate Data Utility and Privacy Risk

10) Alaa, A., Van Breugel, B., Saveliev, E. S., and van der Schaar, M. How faithful is your synthetic data? samplelevel metrics for evaluating and auditing generative models. In International Conference on Machine Learning(ICML). PMLR, 2022.

1. 합성데이터의 i번째 레코드(S_i)를 선택함
2. 해당 레코드 S_i와 원본데이터 내 모든 레코드를 비교하여 거리(distance)가 가장 가까운 레코드(R_j)와의 거리를 d_s로 지정함
3. 레코드(R_j)와 원본데이터 내 모든 레코드를 비교하여 가장 거리가 가까운 레코드의 거리를 로 지정함
4. d_s와 d_o의 거리를 비교하여, d_s이 더 작으면 1, 아니면 0을 기록함
   ※ $I(d_s < d_o)$는 지시함수이고 n은 n_s에서 $d_s = d_o$인 경우의 수를 빼야 함
5. (1~4) 과정을 합성데이터의 모든 레코드 ($S_1, S_2, \cdots, S_{n_s}$)에 대해 n_s(합성데이터 레코드 수)회 반복하여 $d_s = d_o$인 경우의 수만큼 제외하고 나머지 n개의 비율을 계산함

<예시>



- 과적합된 합성데이터가 생성되면 원본데이터와 비슷한 데이터가 많아 추론 위험도가 높아짐
- 추론 위험도가 1에 가까우면 추론 위험이 높다는 의미이고, 추론 위험도가 0에 가까우면 합성데이터가 원본데이터와 유사하지 않아 추론 위험도가 작다는 것을 의미함
- 추론 위험도가 낮으면 유용성 역시 낮아짐
- 유용성과 안전성이 높은 합성데이터의 이론적 추론 위험도는 0.5임
- 거리를 계산하는 방법은 유클리드 거리, 가워(Gower)거리 등을 사용할 수 있음

[비정형 합성데이터 안전성 검증]

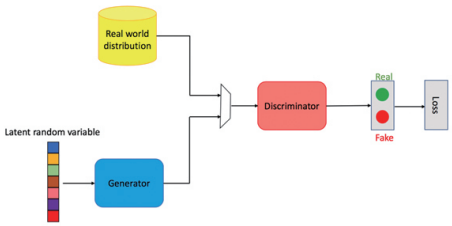
❶ 생성과정 평가

▶ 알고리즘의 적정성 또는 데이터 전처리 여부 등, 합성데이터 생성과정에서의 절차적 타당성을 평가
▶ 비정형 합성데이터의 안전성 확보를 위해서는 절차적 타당성을 검증하는 생성과정 평가가 필수

• 비가역적 알고리즘의 사용 여부
※ 합성데이터에서 원본데이터를 역추적할 수 없는 비가역적 알고리즘을 선택하여야 함

<예시>

• GAN은 생성적 적대 신경망의 약자로서 생성자(Generator)와 판별자(Discriminator)라는 2개의 신경망이 서로 적대적으로 경쟁하는 형태의 생성 알고리즘
• 생성자는 임의의 잡음(noise)로부터 이미지를 생성하며, 판별자는 해당 이미지의 진짜와 가짜 여부를 예측하는 형태로 구성
• 암시적 밀도(ImplicitDensity) 모형* 으로 분류되는 대표적인 알고리즘



출처 : Gilad Cohen and Raja Giryes, Generative Adversarial Networks(2022)

* 암시적 밀도(ImplicitDensity) 모형 : 데이터의 확률 분포를 직접 모형화하지 않고, 데이터가 만족해야 하는 조건이나 제약을 모형화 하도록 설계된 모형

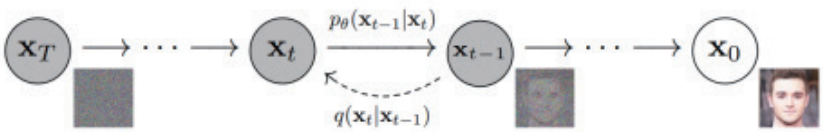
• 완전재현 방식의 합성데이터 생성
※ 비정형 합성데이터는 생성방법이 정형에 비해서 다양하며, 시간이 지남에 따라 다양한 방법들이 계속 등장하고 있으므로, 알고리즘뿐만 아니라, 어떤 방식으로 합성데이터를 생성하였는지도 평가해야 함
※ 안전한 합성데이터를 만들기 위해서는 완전재현 방식의 이미지 생성방법을 사용해야 함

① 완전재현 방식으로 이미지를 생성하는 경우

▶ 합성데이터 생성 시 이미지 모두 가상으로 생성되어야 함

<예시>

※ 이미지 생성 (image generation)
▶ 임의의 잡음(noise) 또는 텍스트 등의 입력을 기반으로 완전히 가상의 새로운 이미지를 생성하는 이미지 처리기술



출처 : Jonathan Ho 외 2인, Denoising Diffusion Probabilistic Models(2020)

② 완전재현 방식으로 이미지를 생성한 것으로 볼 수 없는 경우

▶ 이미지 일부만을 생성, 조작 등으로 합성데이터로 만드는 경우는 원본데이터 일부가 남게 됨

▶ 이미지의 단순 화질 개선, 명도, 채도 변경 등은 적절한 합성데이터 생성으로 볼 수 없으므로 해당 방법만으로 합성데이터를 생성할 수는 없음

<예시>

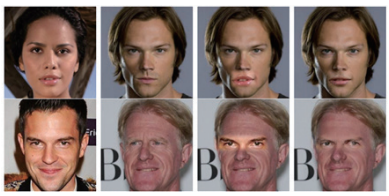
※ 이미지 인페이팅(Image inpainting)

▶ 이미지에서 누락된 영역을 재구성하여 이미지를 생성하는 이미지 처리기술

출처 : Dongsik Yoon 외 4인, DIFAI: Diverse Facial Inpainting using StyleGAN Inversion(2023)

※ 이미지 조작(Image Manipulation)

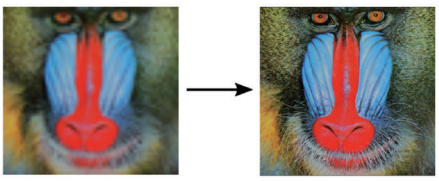
▶ 원하는 출력에 따라 이미지를 변환하는 알고리즘을 통해 이미지를 생성하는 이미지 처리기술



출처 : Andreas Lugmayr외 3인, SRFlow: Learning the Super-Resolution Space with Normalizing Flow(2020)

※ 초해상도(super resolution)

▶ 딥러닝 기반의 AI 모형 등을 통하여 저해상도, 저화질의 이미지를 고해상도, 고화질의 이미지로 변환하는 이미지 처리기술



출처 : Christian Ledig 외 10 인, Photo-Realistic Single Image Super-Resolution Using a Generative Adversarial Network(2017)

• 메타데이터의 삭제 및 비식별처리

※ 메타데이터가 합성데이터 생성 목적과 관련이 없는 경우에는 삭제 필요
※ 메타데이터가 합성데이터의 활용에 꼭 필요한 경우, 비식별 처리하거나 별도로 재현할 수 있어야 하며, 안전성 조치를 추가로 취하여야 함

<예시>

▶ 데이터에서 환자를 특정할 수 있는 식별정보(진료번호 등)를 삭제하거나, DICOM 등 데이터 포멧을 변경하여 메타데이터 제거

• 데이터 전처리 여부

※ 이미지의 영역을 구분하고, 특정 영역이 합성데이터의 활용에 필요하지 않은 경우 삭제하거나 비식별 조치 필요
※ 데이터의 특성에 따라 식별성이 높은 영역을 검토하고, 활용 목적을 고려하여 삭제하거나 비식별 조치 필요
※ 원본데이터 일부 이미지가 희소하여 개인을 특정할 수 있는 이미지를 포함하는 경우, 해당 이미지만 삭제하거나, 문제가 되는 부분을 비식별처리

<예시>

▶ 구강 촬영 의료이미지 데이터를 사용하여 충치 진단 솔루션을 개발하려고 하는 경우, 배경, 코, 촬영 기구 등은 합성데이터 활용에 필요하지 않으므로 삭제
▶ 의료이미지에 환자를 특정할 수 있는 식별정보(진료번호 등)등이 존재하는 경우 비식별 처리 필요

출처 : Yifan Zhang 외 9인, Children's dental panoramic radiographs dataset for caries segmentation and dental disease detection(Scientific Data, 2023)

▶ 원본데이터에 촬영된 신체의 문신이 희소하여 개인을 알아볼 위험성이 존재하는 경우, 해당 이미지의 부분만 마스킹 처리

❷ 이미지 유사성 평가

▶ 안전한 비정형 이미지 합성데이터를 생성하기 위한 절차적 타당성을 모두 준수하였음에도, 원본 이미지와 거의 동일한 합성데이터가 생성된 경우를 검증

<예시>

▶ 모형의 과적합 등으로, 원본이미지가 합성데이터에 거의 동일하게 보존되는 경우
▶ 우연하게 원본이미지가 합성데이터에 거의 동일하게 생성되는 경우

※ 이미지 유사성을 비교하여 원본이미지와 합성이미지간 과도하게 유사한 이미지가 생성되지는 않았는지 평가
※ 합성데이터와 원본데이터 간 1대1로 이미지의 유사성을 판단하며, 유사성이 높게 측정된 이미지에 대해서는 삭제 처리 필요

▶ 유사성 평가의 경우, 각 유사도를 비교하는 정량적 평가와 사람의 육안으로 평가하는 주관적 평가(전수검사 등)가 병행될 수 있음

※ 단, 사람의 주관적 평가(육안평가) 시 전체검사를 진행하는 경우, 구조적/지각적 유사성 등의 정량적 평가는 선택적으로 수행될 수 있음

• 구조적 유사성 검증

※ 원본-합성데이터셋의 모든 이미지간 유사성을 비교하여 구조적으로 과도하게 비슷하게 생성된 이미지가 있는지를 측정
※ 휘도, 대조, 구조 등을 기반으로 원본-합성이미지 간 형태적 유사성을 측정

<예시>

▶ 이미지의 구도, 휘도, 대비를 하나의 품질 점수로 통합한 지표인 SSIM(Structural Similarity Index Measure) 등의 지표를 사용 가능

▶ 두 이미지를 비교하는 지표로, 서로 다른 유사성을 가진 이미지일수록 0에 가까운 점수

$$SSIM(x,y) = \frac{(2\mu_x\mu_y + c_1)(2\sigma_{xy} + c_2)}{(\mu_x^2 + \mu_y^2 + c_1)(\sigma_x^2 + \sigma_y^2 + c_2)}$$

* 구조적 유사성 측정 예시

• 지각적 유사성 검증

※ 원본-합성데이터셋의 모든 이미지간 유사성을 비교하여 지각적으로 과도하게 비슷하게 생성된 이미지가 있는지를 측정
※ 사람의 시각인지능력 측면에서 원본-합성이미지 간 시각적 유사성을 측정

<예시>

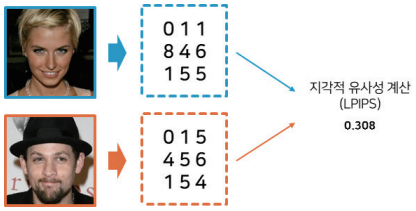
▶ LPIPS(Learned Perceptual Image Patch Similarity) 등의 지표 사용 가능

▶ 두 이미지를 비교하는 지표로, 서로 다른 유사성을 가진 이미지일수록 1에 가까운 점수

* 다른 지표와의 통일 및 점수산출 간편성을 위해 1에서 뺀 값을 사용하여, 서로 다른 유사성을 가진 이미지일수록 0에 가까워지도록 사용 가능(지각적 유사도= 1-LPIPS)

$$LPIPS = \sum_l \frac{1}{H_l W_l} \sum_{h,w} \| w^l \odot (\hat{y}^l_{hw} - \hat{y}^l_{0hw}) \|_2^2$$

* 지각적 유사성 측정 예시



• 주관적 평가

※ 정량평가의 한계를 보완하기 위해 전문가·관계자가 합성이미지의 원본 유사성·개인식별성 등을 직접 검수(정성평가)
※ 비정형데이터는 형태가 다양하고 예상치 못한 요인이 많아 정량적 지표로만 안전성을 검증하는 것은 한계가 존재하므로, 사람의 주관적 검증을 통해 최종 확인

<예시>

▶ 유사성 검증지표(구조적 유사성, 지각적 유사성)가 비교적 높게 측정된 이미지 등에 대해서만 선별 검사하고, 적절하지 않다고 판단되는 이미지는 삭제

▶ 유사성 검증지표(구조적 유사성, 지각적 유사성)를 측정하지 않거나, 높은 수준의 안전성을 보장해야 하는 경우 사람의 전수 육안검사를 진행

부록3. 합성데이터 유용성 검증방법

▶ 아래 검증 방법은 참고를 위한 예시이며, 합성데이터의 유용성을 증명할 수 있는 다른 지표들도 사용할 수 있음

[정형 합성데이터 유용성 검증]

❶ 일차원 분포 유사성

▶ 각 컬럼 간 원본데이터와 합성데이터가 유사한 분포를 따르는지 평가하여 유용성을 검증

- 얀센 샤논 발산 (Jensen-Shannon Divergence)
  - 확률 분포 간의 차이를 계산하여 원본데이터와 합성데이터의 분포 유사성을 확인함
  - 쿨백 라이블러 발산 (Kullback-Leibler Divergence)의 대칭성을 개선하여 두 분포의 평균적 차이를 계산함

$$JSD(P \parallel Q) = \frac{1}{2} D_{KL}(P \parallel \frac{(P+Q)}{2}) + \frac{1}{2} D_{KD}(Q \parallel \frac{(P+Q)}{2})$$

※ 거리가 가까울수록, 0에 가까울수록 좋은 품질의 합성데이터

- 콜모고로프-스미르노프 통계량(Kolmogorov-Smirnov Test)
  - 연속형 컬럼의 각 컬럼별 누적확률분포의 유사성을 검증

$$D_{n,m} = \max|F_{1,n}(x) - F_{2,m}(x)|$$

- 카이제곱 통계량(Chi-Square test)
  - 범주형 컬럼의 각 컬럼별 누적확률분포의 유사성을 검증

$$\chi^2 = \sum_{i=1}^{c} \frac{(O_i - E_i)^2}{E_i} \sim \chi^2(c-1)$$

❷ 2차원 관계 유사성

▶ 원본데이터 각 컬럼 간 존재하는 상관관계의 유사성을 검증

1. 원본데이터와 합성데이터의 컬럼 간 상관계수 행렬 R_{ori}, R_{syn}를 계산함

수치형 변수 간 상관관계 : 피어슨 상관 계수

$$r_{XY} = \frac{\sum_{i}^{n}(X_i - \overline{X})(Y_i - \overline{Y})}{\sqrt{\sum_{i}^{n}(X_i - \overline{X})^2}\sqrt{\sum_{i}^{n}(Y_i - \overline{Y})^2}}$$

※ +1은 완벽한 양의 선형 상관 관계, -1은 완벽한 음의 상관 관계, 0은 선형 상관 관계가 없음을 의미함

범주형 변수 간 상관관계 : 크래머 V 상관계수

$$V = \sqrt{\frac{\chi^2}{n(k-1)}}$$

※ 0에 가까울수록 두 변수 간 관계가 없음을, 1에 가까울수록 강한 상관관계가 있음을 의미함
- χ^2: 카이제곱(Chi-square) 통계량
- n : 전체 표본 수
- k : 행과 열 중 더 작은 차원

범주형 변수와 수치형 변수 간 상관관계 :
- ANOVA(Analysis Of Variance) 분석의 SST(Sum of Square Total) = SSE(Sum of Square Error) + SSB(Sum of Square Between) 관계를 이용할 수 있음
- 총 변동 중 그룹 간 변동의 비율, 즉 SSB/SST의 비율을 이용하여 연속형 변수와 범주형 변수의 상관관계로 활용함

2. 두 상관계수 행렬 R_{ori}, R_{syn}의 차이값 행렬 D를 계산함
3. 차이값 행렬 D의 표준편차 σ_D를 통하여 상관관계의 유사성을 계산함

$$D = R_{\mathrm{ori}} - R_{syn}$$

$$\sigma_D = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(d_i - \overline{d})^2}$$

※ 값이 작을수록, 0에 가까울수록 좋은 품질의 합성데이터
- d_i는 차이값 행렬 D의 각 요소임
- $\overline{d}$는 차이값 행렬 D의 평균임

- 상관계수는 원본데이터 컬럼 간 선형관계를 측정하기 때문에 비선형 관계는 측정하지 못한다는 단점이 있음
- 연속형 변수의 상관분석은 주로 피어슨 상관계수(Pearson Correlation Coefficient), 범주형 변수의 경우는 크래머 V (Cramer's V) 상관계수를 사용함

※ 연속형 변수와 범주형 변수의 상관계수 계산 방식
- ANOVA(Analysis Of Variance) 분석의 SST(Sum of Square Total) = SSE(Sum of Square Error) + SSB(Sum of Square Between) 관계를 이용할 수 있음
- 총 변동 중 그룹 간 변동의 비율, 즉 SSB/SST의 비율을 이용하여 연속형 변수와 범주형 변수의 상관관계로 활용함

❸ 구별 불가능성 (pMSE, propensity score Mean Squared Error)

▶ 원본데이터와 합성데이터를 섞은 데이터셋에 대해 원본데이터인지 합성데이터인지 구분하는 머신러닝 모형을 만들고, 이를 정확히 구분할 확률을 유사성으로 계산

1. 원본데이터와 합성데이터 셋 구성
2. 원본데이터와 합성데이터를 행간 병합하여 둘을 구분하는 종속변수를 생성함
3. 원본데이터와 합성데이터를 구분하는 분류 모형을 생성함
4. 전체 데이터의 각 레코드마다 예측값($\hat{p}$)을 계산하고, 모든 예측값을 통해 구별 불가능성(pMSE)을 계산함

$$\text{구별 불가능성}\,(pMSE) = \frac{1}{n_o + n_s}\sum_{i=1}^{n_o+n_s}(\hat{p_i} - \frac{n_s}{n_o + n_s})^2$$

※ 값이 작을수록, 0에 가까울수록 좋은 품질의 합성데이터
- $\hat{p_i}$는 i번째 레코드를 합성데이터로 예측할 확률을 의미함
- n_o는 원본데이터의 레코드 수, n_s는 합성데이터의 레코드 수를 의미함

❹ 모형 성능 유사성

▶ 활용 목적에 따라 합성데이터가 원본데이터와 유사한 성능을 내는지를 검증

▶ 성능검증 방법에 따라 정확도(Accuracy), 재현율(Recall), 정밀도(Precision), F1 score 등 다양한 지표 활용 가능

▶ 정형 합성데이터의 경우 선형회귀모형을 활용할 수 있음

1. 원본데이터와 합성데이터를 학습데이터로 하여 판별모델 2개를 구축
2. 모델의 성능을 정량적으로 측정할 수 있는 지표 [정확도(Accuracy), 재현율(Recall), 정밀도(Precision), F1 score 등]를 활용하여, 원본데이터로 학습한 모델과 합성데이터로 학습한 모델의 성능을 측정
3. 측정한 지표의 점수를 비교

※ 선형회귀모형을 사용할 경우
- 원본데이터와 합성데이터를 각각 학습 데이터로 사용하여 두 개의 선형 회귀 모형을 적합한 뒤, 계산된 회귀계수의 신뢰구간이 겹치는 정도를 유용성 지표로 활용할 수 있음(아래 수식 참고)

$$\text{회귀계수 신뢰구간 비교}(CI) = \frac{1}{2}\left[\frac{U_i - L_i}{U_{ori} - L_{ori}} + \frac{U_i - L_i}{U_{syn} - L_{syn}}\right]$$

※ 1에 가까울수록 좋은 품질의 합성데이터
- (L_{ori}, U_{ori})는 원본데이터 변수의 회귀계수 신뢰구간이고, (L_{syn}, U_{syn})는 합성데이터 변수의 회귀계수 신뢰구간을 의미함
- (L_i, U_i)는 두 신뢰구간이 겹치는 부분의 양 끝단을 의미함

• 원본데이터로 학습한 모델의 성능과, 합성데이터로 학습한 모델의 성능이 비슷한 경우, 합성데이터는 원본데이터의 성능과 어느정도 동일하다고 할 수 있음

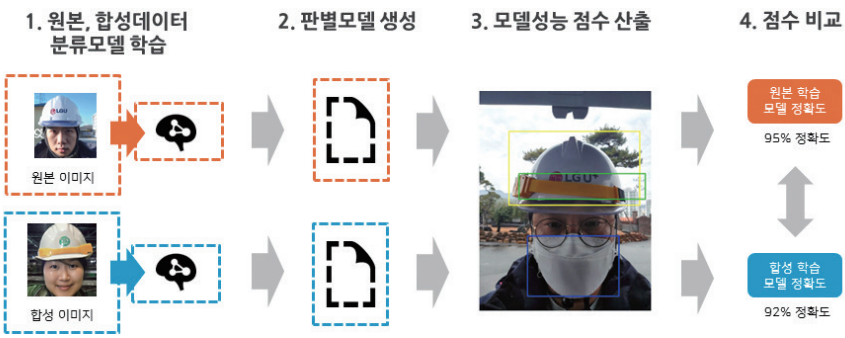
[비정형 합성데이터 유용성 검증]

❶ 모형 성능 유사성

▶ 활용 목적에 따라 합성데이터가 원본데이터와 동일한 성능을 내는지를 검증

▶ 성능검증 방법에 따라 정확도(Accuracy), 재현율(Recall), 정밀도(Precision), F1 score 등 다양한 지표 활용 가능

1. 원본데이터와 합성데이터를 학습데이터로 하여 판별모델 2개를 구축
2. 모델의 성능을 정량적으로 측정할 수 있는 지표 [정확도(Accuracy), 재현율(Recall), 정밀도(Precision), F1 score 등]를 활용하여, 원본데이터로 학습한 모델과 합성데이터로 학습한 모델의 성능을 측정
3. 측정한 지표의 점수를 비교



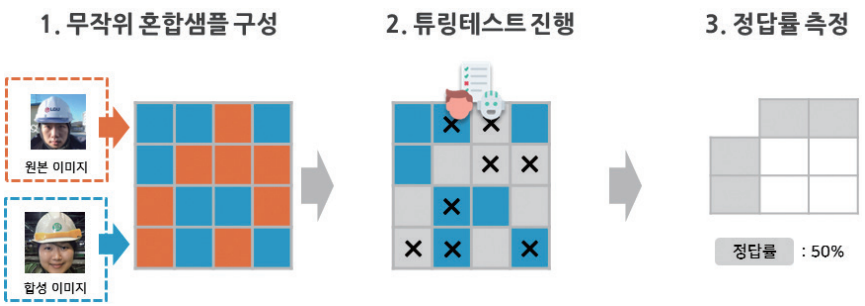
• 원본데이터로 학습한 모델의 성능과, 합성데이터로 학습한 모델의 성능이 비슷한 경우, 합성데이터는 원본데이터의 성능과 어느정도 동일하다고 할 수 있음

❷ Visual Turing Test(VTT)

▶ 사람의 육안 검사 등을 통해, 원본과 합성데이터를 맞추는 실험

▶ 전문가·관계자를 대상으로 원본이미지와 합성이미지를 동일한 수로 무작위 선별하고, 무작위 이미지 중 합성이미지를 개수만큼 구분해내는 테스트를 수행하고 정답률을 측정

1. 원본과 합성이미지를 무작위로 선별하여, 각각 동일한 수의 혼합샘플 구성
2. 튜링테스트 진행(혼합샘플에서 합성이미지를 개수만큼 선별)
3. 전체 합성데이터 중 올바르게 찾아낸 합성데이터의 갯수의 비율을 측정



• 합성이미지를 선별하는 사람은 전체 혼합샘플에서 합성이미지의 개수(혼합샘플은 원본과 합성이미지의 개수가 동일하므로, 전체 혼합샘플의 절반) 만큼 선별

$$VTT결과 = \frac{응답자가\ 찾아낸\ 합성데이터의\ 개수}{전체\ 합성데이터의\ 개수}$$

• 사람이 비교하였을 때 원본데이터와 합성데이터를 잘 구별해내지 못하는 경우, 합성데이터는 원본데이터만큼 잘 만들어진 것으로 볼 수 있음

• 가장 잘 생성된 합성데이터라고 가정 시 정답률 50%가 이상치이며, 유용성 수준에 따라 10~20%의 여유(40%~60%, 30%~70%)를 둘 수 있음

❸ 이미지 품질 검증

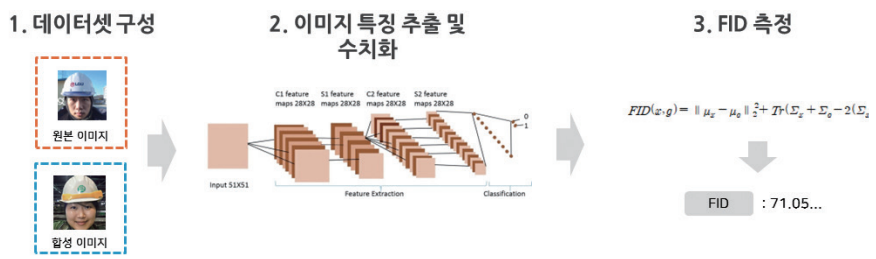
▶ AI로 생성된 이미지의 품질을 평가하는 지표를 이용하여 원본데이터와 합성데이터 간 품질을 정량적으로 측정하여 비교하는 방법

▶ 원본이미지와 합성이미지의 데이터 분포 간 거리를 비교하여 학습이미지가 원본 이미지의 분포를 얼마나 잘 담고 있는지를 검증

$$FID(x,g) = \| \mu_x - \mu_g \|_2^2 + Tr(\Sigma_x + \Sigma_g - 2(\Sigma_x\Sigma_g)^{\frac{1}{2}})$$

※ 거리가 가까울수록, 0에 가까울수록 좋은 품질의 합성데이터

1. 원본데이터와 합성데이터 셋 구성
2. 각 이미지의 특징(feature) 추출 및 수치화
3. FID 점수 측정



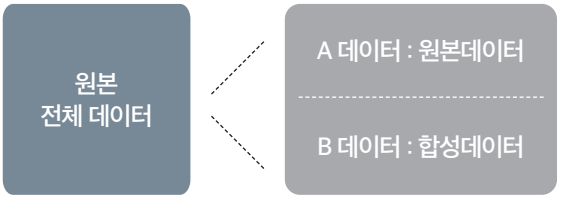
부록4. 정형 데이터 검증 지표 임계값 산정 방법

▶ 아래 임계값 산정 방법은 참고를 위한 예시이며, 합성데이터의 유용성과 안전성을 증명할 수 있는 임계값은 다른 방식으로도 사용할 수 있음

[정형 합성데이터 임계값 산정]

구별 위험도, 추론 위험도 임계값 설정 방법

[단계 1] 원본데이터를 50 : 50 크기로 무작위 분할하여 각각 데이터를 원본데이터 A, 합성데이터 B로 가정함



[단계 2] (A), (B) 두 데이터를 비교하여 구별 위험도, 추론 위험도 지표 측정

[단계 3] 구별 위험도 $\hat{p}$의 경우, 지표값을 아래의 식으로 보정함

$$p^* = 1-(1-\hat{p})^2$$

- $\hat{p}$은 단계 2에서 원본데이터 레코드 수가 $n/2$개, 즉 50% 데이터로 구한 구별 위험도이고, p^*은 원본데이터의 레코드 수를 n개로 가정하고 구한 보정 값임
- 실제 합성데이터 생성 시, 원본데이터 $n/2$개와 비교하므로 이 보정 식을 써야 정확한 임계값임
- 보정 식을 사용하지 않으려면 원본데이터 중 개를 임의로 추출하여 지표를 계산할 수 있음
- 추론 위험도는 원본 개수에 영향을 받지 않아 보정하지 않음

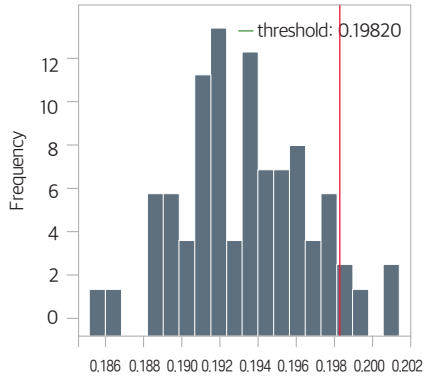
[단계 4] 위 과정을 충분히 반복한 후, 누적된 지표값의 분위 수(예: 지표값의 90%, 95%, 99% 분위수)를 임계값으로 설정

- 합성데이터의 활용목적과 활용범위에 따라 임계값 설정 수준을 설정할 수 있음
  ▶ 높은 안전성이 요구되는 경우에는 비교적 낮은 분위 수(예 : 90분위 수) 설정
  ▶ 보통 수준인 경우, 비교적 높은 분위 수(예 : 99분위 수) 설정

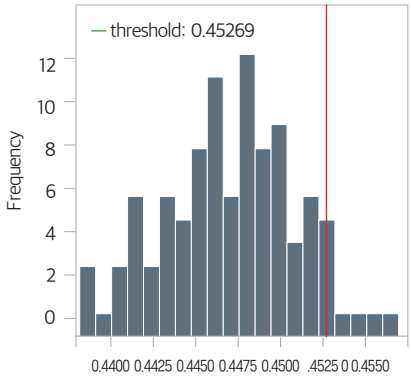
<예시>

• 원본데이터를 50:50으로 무작위 분할하여 각각 원본데이터 A와 합성데이터 B로 가정하고 이를 100회 정도 반복하여 누적된 2가지 안전성 지표 분포를 히스토그램으로 나타냄

<구별 위험도 지표값과 임계값>



<추론 위험도 지표값과 임계값>



• 100회 반복하여 측정한 각 위험도 지표값을 히스토그램으로 나타내고
• 이 지표의 95%분위수(빨간색 점선)을 임계값으로 설정함

위험도 기준	임계값
구별 위험도	0.198
추론 위험도	0.452

• 합성데이터의 각 지표값이 임계값보다 같거나 작으면 안전하다고 판단함
• 추론 위험도의 경우, 임계값보다 작거나, 0.5보다 작으면 안전하다고 판단함

연결 위험도 임계값 설정 방법

• 연결위험도의 임계값은 절대값 지정이 가능함
• 예를 들어 연결위험도 CAP 임계값을 0.7로 정하면 원본과 준식별자 조합이 같은 합성 레코드 중 민감정보까지 같은 레코드가 70% 미만인 경우, 안전한 레코드로 판단함
• 70% 이상인 경우, 합성 레코드의 일부분을 삭제하여 임계값 수준을 맞출 수 있음
• CAP 임계값이 작으면 합성 레코드 중 이를 만족하지 않는 레코드를 많이 삭제하기 때문에 유용성이 저하됨
• CAP 임계값이 크면 연결 위험이 높아 안전성이 저하되므로 적절하게 설정하는 바람직함

부록5. 정형 합성데이터 생성 예시

시나리오 : 통신사 멤버십 사용내역 합성데이터

▶ A 제휴사는 B 통신사의 고객 단위 멤버십 사용내역 데이터를 이용하여 합성데이터를 생성하고, 이를 이용해서 어떤 고객이 어떤 쿠폰을 받고 사용하는지를 추정하는 모델을 개발하고자 함

- 합성데이터 생성 및 활용 목적
  - 제휴사 선호도 AI모델 개발을 위한 통신사 멤버십앱 사용내역 학습용 데이터셋 생성
- 원본데이터 특징
  - 정형데이터(CSV) 형태의 통신사 멤버십 사용내역 로그(멤버십연계 제휴사 연계 활용 정보)
  - 회원ID, 이동통신사 명, 제휴컨텐츠 유형명, 제휴컨텐츠 발행일자·시각, 제휴사명, 고객연령, 성별, 시군구 주소 등의 정보 포함('22.6.~'23.4. 중 매월 2만건씩 약 20만건 추출)
- 데이터 활용 범위
  - 합성데이터를 생성 및 활용하는 주체는 A제휴사로 한정되어 있음
  - A제휴사 내부에서 합성데이터를 활용하며, 환경은 데이터의 취급 시 통제가 강제되는 폐쇄적 환경임

① 사전준비 단계에서 합성데이터 활용 목적, 활용 범위(활용환경·형태), 생성·활용 주체 등을 정의하여 계획 마련

합성데이터 활용 계획서

참여자	조직/부서명	A 제휴사 데이터 분석팀		
	담당자 직위	팀장	담당자 성명	홍길동
	전화번호	02-1234-5678	이메일 주소	-

활용 목적		A 제휴사는 B 통신사에서 사용자가 앱에서 쿠폰을 받고, 사용한 멤버십 사용내역 데이터를 이용하여 어떤 고객이 어떤 쿠폰을 받고 사용하는지를 추정하는 AI 모델을 개발하고자 함 데이터는 멤버십 사용 내용 데이터, 나이와 성별, 시군구 단위 등의 고객 정보 등의 개인정보이며, 이를 합성데이터로 생성, 익명정보로 처리하여 모델 개발을 위한 학습용 데이터로 활용하고자 함 B 통신사는 합성데이터 생성 업무를 C 사에 위탁하고, C 사가 합성데이터를 생성, 익명정보로 처리하면, 합성데이터는 A 제휴사에 전달되어 내부에서 활용될 예정		
활용 구분	활용형태	☐ 내부 이용	☐ 타 부서 이용(제공)	☑ 특정 제3자 제공
		☐ 불특정 3자 제공		
		☐ 데이터 완전 공개		
	생성/활용 주체	원본데이터 보유자	합성데이터 생성자	합성데이터 수요자
		B 통신사	B 통신사 (위탁) C사	A 제휴사
	이용 장소	☐ 완전개방 환경* ☐ 부분개방 환경** ☑ 폐쇄적 환경 * 인터넷이 연결된 다른 정보의 유입이 가능한 분석 환경 ** 폐쇄환경은 아니지만 취급자 또는 제공자의 통제가 가능한 내부의 분석환경을 말함		
	반복 제공 여부	☑ 1회 제공 ☐ 반복 제공 예정 (회 예정)		
	제공 방법	☐ 온라인 ☑ 오프라인		
기타	원본데이터 명칭	B 통신사 멤버십앱 사용내역 데이터		
	원본데이터 내역	B통신사가 2022년 6월부터 2023년 4월까지 고객으로부터 수집한 멤버십 사용 내역 데이터		
	합성데이터 이용 기간	없음		

첨부서류	① 합성데이터 처리 위탁계약서 ② 데이터 공급 협약서

② 사전 준비단계의 원본데이터 특성 검토와 합성데이터 생성계획 등을 담은 합성데이터 생성 기초자료 명세서 작성

합성데이터 생성 기초자료 명세서	
합성데이터명	B 통신사의 회원별 멤버십 사용내역 합성데이터
생성 기간	2024년 1월 1일 ~ 2024년 1월 30일

원본데이터 특성			
원본데이터명	B 통신사의 회원별 멤버십 사용내역 데이터		
데이터 유형	☑ 개인정보	☐ 가명정보	☐ 기타 ( )
데이터 업종	유통 분야		
데이터 속성	• B 통신사의 멤버십 사용 내역 (A 제휴사 연계 활용 정보) • 해당 데이터에는 B 통신사를 이용하는 고객 개인을 식별할 수 있는 고객 고유 식별자, 전화번호 등이 존재하며, 그 외 고객의 성별, 나이, 주소 등의 정보가 담긴 개인정보가 포함되어 있음 • '22.6.~'23.4. 중 매월 2만건씩 약 20만건 추출		
데이터 형식	☑ 정형데이터	☐ 비정형 데이터(이미지)	☐ 비정형 데이터(기타)
	* 해당 데이터는 CSV 형식의 정형데이터		

합성데이터 생성계획		
번호	구분	검토사항
1	환경 및 데이터 준비	• B 통신사(원본데이터 제공자) 개인정보 시스템에서 개인정보취급자가 직접 원본 데이터 추출 후 안전한 장소로 이동 • 데이터 분석 및 합성데이터 생성 환경은 합성데이터 생성자가 요구하는 환경으로 개인정보처리자가 직접 준비
2	데이터 이해	• 식별자 : 고객번호(삭제), 상세 주소(읍면동 단위 일반화 처리), 전화번호(삭제) 준식별자 : 성별, 읍면동 단위 거주지 민감정보 : 없음 • 원본데이터 전처리를 통해 데이터의 위험성을 낮췄고, 폐쇄적 환경의 분석 모형 개발에 사용되므로 유용성 확보에 중점을 두고 합성데이터 생성 필요
3	합성데이터 생성	• B 통신사(원본데이터 제공자) 내 보호조치가 갖추어진 안전한 장소에서 합성데이터 생성자(위탁처리-C사)가 권한을 부여받아 개인정보처리자 관리감독 하에 합성데이터 생성 • 원본데이터의 크기를 고려하여 메모리 64GB 이상, GPU 4080Ti 12G 이상 고성능 컴퓨터 필요 • 원본데이터 특성을 고려하여 생성 기법은 가우시안 혼합 모델(GMM), synthpop-CART, CTGAN을 사용하고, 가장 성능이 우수한 모델을 최종선별하여 합성데이터 생성

③ 원본데이터 확보 후 탐색적 분석을 수행하는 과정에서 명세서를 작성하여 정리하고, 항목별 처리계획도 마련

원본데이터 명세서

1) 개요

항목	내용	비고
데이터명	통신사 멤버십 사용내역 로그	-
데이터 유형	정형데이터	csv 파일
데이터 규모	약 200,000건	-
특이사항		

2) 개인정보 유형 분류표

2-1) 정보 개요

연번	정보영역	수량	비중	정보영역설명
1	고객 정보	5	16%	쿠폰을 사용한 고객의 정보
2	멤버십 사용내역정보	25	84%	발행 및 사용 쿠폰 정보
전체	-	30	100%	-

2-1) 정보 상세

연번	자료형태	정보영역	항목명	항목 설명
1	범주형	고객 정보	회원ID	고객을 식별하는데 사용되는 고유 회원ID
2	범주형	멤버십 사용내역정보	이동통신사명	회원의 이동통신사 명
3	범주형	멤버십 사용내역정보	컨텐츠유형명	영화할인쿠폰, 도서할인쿠폰 등의 쿠폰 명칭
4	범주형	멤버십 사용내역정보	쿠폰사용유형명	컨텐츠의 타입 명칭
5	수치형	멤버십 사용내역정보	발행일자	쿠폰 발행일(1~31)
6	수치형	멤버십 사용내역정보	컨텐츠발행시각	쿠폰 발행시각(0~23)
7	범주형	멤버십 사용내역정보	제휴사명	컨텐츠 제휴사 명
~	~			
28	수치형	고객 정보	고객연령	0~100세 이하의 1단위 연령
29	범주형	고객 정보	성별구분코드	(1,2) - (남/녀)
30	범주형	고객 정보	시군구주소	쿠폰을 사용한 주소(시/군/구)

항목별 개인정보 처리 계획

① 개인정보 처리계획						
연번	항목명	정보영역		처리방법		세부방법 및 처리수준
		항목명	설명			
1	B통신사 멤버쉽 데이터	회원id	내부 고객관리 id	전처리 시	삭제	삭제
		통신사명	-	전처리 시	삭제	삭제
		컨텐츠 유형명	-	전처리 시	일부 유지	빈도분석을 통해 빈도가 높은 컨텐츠 선택
		발행일자	-	전처리 시	일부 유지	분석에 불필요한 구체적 일자 정보에서 월, 시각, 요일만 추출
		제휴사명	-	전처리 시	삭제	삭제
		고객연령	1세 단위 나이	전처리 시	유지	그대로 사용
		고객성별		전처리 시	유지	그대로 사용
		~				
		시군구 주소	상세주소 (시·군·구 단위)	전처리 시	일부 유지	모형 개발에 불필요한 상세 주소는 삭제하고 동단위 정보만 유지

④ 생성된 합성데이터 결과를 검토하고, 안전성 검증 등이 적절히 수행되었는지 종합적으로 검토

합성데이터 명세서

1) 개요

항목	내용	비고
데이터명	회원별 멤버십 사용내역 합성데이터	-
데이터 유형	정형데이터	csv 파일
데이터 규모	약 100,000건	-
특이사항		

2) 정보 유형 분류표

2-1) 정보 개요

연번	정보영역	수량	비중	정보영역설명
1	고객 정보	2	28.6%	쿠폰을 사용한 고객의 정보
2	멤버십 사용내역정보	5	71.4%	발행 및 사용 쿠폰 정보
전체	-	7	100	-

2-1) 정보 상세

연번	자료형태	정보영역	항목명	항목 설명
1	범주형	쿠폰정보	컨텐츠유형명	쿠폰 명
2	수치형	고객정보	고객연령	7세 ~ 91세 분포
3	범주형	고객정보	고객성별	FEMALE / MALE
4	범주형	쿠폰정보	주소	시도 수준의 주소 정보
5	수치형	쿠폰정보	발행요일	쿠폰 발행 요일
6	수치형	쿠폰정보	발행월	쿠폰 발행 월
7	수치형	쿠폰정보	발행시각	쿠폰 발행 시간

3) 합성데이터 예시

컨텐츠유형명	고객연령	고객성별	주소	발행요일	발행월	발행시각
영화 할인 쿠폰	29	FEMALE	경기도	saturday	6	15
조회쿠폰	16	MALE	부산시	monday	12	8
멤버십카드	36	MALE	대전시	thursday	11	21
조회쿠폰	28	FEMALE	경기도	wednesday	5	17
조회쿠폰	46	MALE	경기도	thursday	4	23
영화 할인 쿠폰	57	MALE	서울시	tuesday	7	21
~						

합성데이터 안전성 자체 검토 결과서

1) 측정 방법

※ 먼저 안전성 기준을 충족시킨 뒤, 그 중 유용성이 높은 것을 선택
※ 여러 가지 다양한 생성모델과 다양한 방식으로 생성한 합성데이터셋 6가지 중, 목표한 안전성 기준을 모두 충족하는 합성데이터 셋을 선별하고 이 중 유용성이 가장 높은 합성데이터 셋을 선택하고자 함

2) 측정지표 선정

연번	구분	지표명	지표 설명 및 측정 방법
1	안전성	구별 위험도	※ 합성데이터의 레코드가 원본데이터 내에 동일하게 존재할 확률을 측정 ※ 합성데이터 내 원본데이터와 같은 레코드를 삭제하면 식별 위험을 줄일 수 있음 ※ 계산식은 다음과 같고, 값이 0에 가까울수록 안전성 높음 * 식별 위험도 = $\frac{1}{n_s}\sum_{i=1}^{n_s} I(S_i = R_j)$
		추론 위험도	※ 합성데이터에 원본데이터와 같은 레코드가 존재하지는 않지만, 확률적으로 추론할 가능성이 높은지를 측정 ※ 합성 레코드가 자신과 가장 가까운 레코드를 원본에서 찾았을 때의 거리(dS)와 해당 원본 레코드와 가장 가까운 원본 레코드와의 거리(dO)를 비교하는 개념의 독창성 지표로, 전자가 더 가까운(dS<dO) 비율을 지표로 사용 * 추론 위험도= $\frac{1}{n}\sum_{i=1}^{n_s} I(d_S < d_O)$
2	유용성	2차원 관계 유사성 (상관계수 표준편차)	※ 원본데이터 각 컬럼 간 존재하는 상관관계의 유사성을 검증 ※ 원본과 합성데이터의 컬럼 간 상관계수 행렬 차이값의 표준편차를 계산 ※ 상관계수 행렬이 유사한지, 유의미한 상관계수값의 부호가 같은지 관찰하여 판단
		구별 불가능성 (pMSE)	※ 원본데이터와 합성데이터를 섞은 데이터셋에 대해 원본인지 합성데이터인지 구분하는 머신러닝 모형을 만들고 이를 정확히 구분할 확률을 유사성으로 계산 ※ 계산식은 다음과 같고, 값이 0에 가까울수록 유용성 높음 * 구별 불가능성 = $\frac{1}{2n}\sum_{i=1}^{2n}\left(p_i - \frac{1}{2}\right)^2$

3) 지표별 임계값

연번	구분	지표명	임계값
1	안전성	구별 위험도	4-1) 측정결과 참고
		추론 위험도	4-1) 측정결과 참고
2	유용성	2차원 관계 유사성(상관계수 표준편차 비교)	4-1) 측정결과 참고
		구별 불가능성(pMSE)	4-1) 측정결과 참고

4) 안전성 자체 측정 수행

※ 3가지 생성 모델(❶GMM, ❷synthpop-CART, ❸CTGAN)을 활용하여 합성데이터를 각각 두 가지 방식으로 생성
(합성데이터셋 A) 원본과 같은 수의 합성데이터 생성
(합성데이터셋 B) 원본의 2배수의 합성데이터를 생성 후, 식별 위험도가 큰 레코드(동일 레코드)를 삭제하고, 데이터 수를 원본과 같은 수로 줄여 생성
=> 3가지 생성모델을 활용해 만든 A, B 방식의 합성데이터 세트 총 6가지 생성 후, 안전성 기준을 충족시킨 데이터 셋 중 유용성이 높은 것을 선택

5) 최종안전성 자체 측정결과 및 평가

5-1) 측정결과

* 괄호 안의 값은 각 데이터 셋별 임계값을 의미

모형		❶ GMM		❷ synthpop-CART		❸ CTGAN	
데이터셋 구분		❶-(A)	❶-(B)	❷-(A)	❷-(B)	❸-(A)	❸-(B)
안전성 측정 결과	식별 위험도	0.0818 (0.17)	0 (0.11)	0.3917 (0.17)	0 (0.11)	0.0875 (0.17)	0 (0.11)
	추론 위험도	0.426 (0.602)	0.422 (0.56)	0.512 (0.602)	0.524 (0.56)	0.518 (0.602)	0.446 (0.56)
유용성 측정 결과	상관계수 표준편차	0.038 (0.005)	0.061 (0.11)	0.007 (0.005)	0.071 (0.11)	0.056 (0.005)	0.070 (0.11)
	pMSE	0.006 (0.00009)	0.01 (0.05)	0.00008 (0.00009)	0.001 (0.05)	0.037 (0.00009)	0.009 (0.05)

5-2) 결과 평가

총평
※ 6개의 합성데이터셋 중 안전성 검증 지표를 모두 만족하는 데이터셋은 ❶-(A), ❶-(B), ❷-(B), ❸-(A), ❸-(B) 등 5개 ※ 안전성 지표를 충족한 5개의 데이터셋 중 모든 유용성 검증 지표에 대해 임계값을 만족하고, pMSE가 가장 낮은(유용성이 가장 높은) ❷-(B) 합성데이터셋이 가장 적절한 합성데이터셋으로 평가

부록6. 비정형 합성데이터 생성 예시

시나리오 : 구강 이미지 합성데이터

▶ A 병원은 B 치과병원이 보유한 환자 구강 사진 데이터를 합성데이터로 생성하여 충치를 진단하고 예방하는 AI 솔루션을 개발하고자 함

- 합성데이터 생성 및 활용 목적
  - 충치 진단·예방관리 AI솔루션 개발을 위한 학습용 데이터셋 생성
- 원본데이터 특징
  - JPG 형식의 이미지 데이터이며, 의료 이미지에 해당함
  - 500명의 정보 주체로부터 수집된 촬영데이터이며, 정보 주체별로 상악 이미지1장, 하악 이미지 1장으로 구성된 총 1,000장의 이미지 파일
  - 환자번호 등 메타데이터는 포함되어 있지는 않음
- 데이터 활용 범위
  - 합성데이터를 생성 및 활용하는 주체는 A병원으로 한정되어 있음
  - A병원 내부에서 합성데이터를 활용하며, 환경은 데이터의 취급 시 통제가 강제되는 폐쇄적 환경임

① 사전준비 단계에서 합성데이터 활용 목적, 활용 범위(활용환경·형태), 생성·활용 주체 등을 정의하여 계획 마련

합성데이터 활용 계획서

참여자	조직/부서명	A병원 00과		
	담당자 직위		담당자 성명	홍길동
	전화번호		이메일 주소	-

활용 목적		A병원은 B치과병원의 환자 구강촬영 이미지를 바탕으로 AI 충치진단 솔루션을 개발하여 진단에 활용하고자 함 그러나, 환자의 구강촬영 정보는 개인정보로 AI 개발을 위한 2차적 활용이 어려워, 대신 합성데이터로 생성하여 학습용 데이터 셋을 확보하고자 함 B 치과병원은 합성데이터 생성 업무를 C 사에 위탁하고, C 사가 합성데이터를 생성, 익명 정보로 처리하면 합성데이터는 A병원에 전달되어 내부에서 활용될 예정		
활용 구분	활용형태	☐ 내부 이용	☐ 타 부서 이용(제공)	☑ 특정 제3자 제공
		☐ 불특정 3자 제공		
		☐ 데이터 완전 공개		
	생성/활용 주체	원본데이터 보유자	합성데이터 생성자	합성데이터 수요자
		B 치과병원	B 치과병원 (위탁) C사	A 병원
	이용 장소	☐ 완전개방 환경* ☐ 부분개방 환경** ☑ 폐쇄적 환경 * 인터넷이 연결된 다른 정보의 유입이 가능한 분석 환경 ** 폐쇄환경은 아니지만 취급자 또는 제공자의 통제가 가능한 내부의 분석환경을 말함		
	반복 제공 여부	☑ 1회 제공 ☐ 반복 제공 예정 (회 예정)		
	제공 방법	☐ 온라인 ☑ 오프라인		
기타	원본데이터 명칭	B 치과병원의 구강 촬영 이미지		
	원본데이터 내역	B 치과병원이 2022년 12월부터 2023년 10월까지 내원 환자 500명을 대상으로 상·하악치 각 1장씩 촬영한 구강사진(상악치 500장, 하악치 500장)		
	합성데이터 이용 기간	없음		

첨부서류	① 합성데이터 처리 위탁계약서 ② 데이터 공급 협약서 ③ IRB 심의 결과통지서, 연구계획서 등

② 사전 준비단계의 원본데이터 특성 검토와 합성데이터 생성계획 등을 담은 합성데이터 생성 기초자료 명세서 작성

합성데이터 생성 기초자료 명세서

합성데이터명	구강 촬영 이미지 합성데이터
생성 기간	2024년 1월 1일 ~ 2024년 1월 30일

원본데이터 특성			
원본데이터명	B 치과병원의 구강 촬영 이미지		
데이터 유형	☑ 개인정보	☐ 가명정보	☐ 기타 ( )
데이터 업종	보건 의료 분야		
데이터 속성	* 비정형데이터(의료 이미지, JPG) * 내원 환자 500명의 구강사진 (상·하악치 각 1장씩) 상악치 500장, 하악치 500장 ('22.12.~'23.10. 촬영) * 환자번호 등 메타데이터는 제외하고 제공받으나, 파일이름에 환자이름, 촬영일자로 표시		
데이터 형식	☐ 정형데이터	☑ 비정형 데이터(이미지)	☐ 비정형 데이터(기타)
	* 해당 데이터는 JPG 형식의 비정형 이미지 데이터		

합성데이터 생성계획		
번호	구분	검토사항
1	환경 및 데이터 준비	• 합성데이터 생성은 인간대상연구에 해당하여, '생명윤리법'에 따라 기관생명윤리위원회(IRB) 심의 대상이므로, 합성데이터 생성 수행에 앞서 IRB 심의를 받음 • B 치과병원 내부 규정에 따라, 개인정보 책임자 관리하에 원본데이터를 추출, 메타데이터 제거하여 개인 식별성이 없는 상태로 추출하여 준비 • 데이터 분석과 합성데이터 생성을 위한 환경 구성은 개인정보 책임자 관리하에 합성데이터 생성자가 요구하는 환경을 B 치과병원 내에 직접 준비
2	데이터 이해	• 상악치, 하악치 등의 구강 영역 위주의 데이터이나 얼굴, 코, 턱, 수염 등 합성데이터 생성에 활용되지 않는 부분도 일부같이 촬영됨 • 메타데이터는 포함되어 있지 않으나, 원본데이터는 치과 내부정보를 통해 환자 개인식별이 가능하므로 개인정보에 해당한다고 판단되므로 비가역적인 알고리즘을 통한 완전합성 형태의 합성데이터 생성이 필요
3	합성데이터 생성	• B 치과병원 내부규정에 따라 원본데이터는 외부로 반출하지 않고, 일련의 합성데이터 생성 작업은 합성데이터 생성자(위탁처리-C사)가 개인정보처리자와 책임자에게 권한을 부여받고 관리하에 B 치과병원의 안전한 처리장소 내에서 진행 • 원본데이터의 크기를 고려하여 메모리 64GB 이상, GPU 4080Ti 12G 이상 고성능 컴퓨터 필요 • 원본데이터 특성을 고려하여 생성 기법은 딥러닝 기반의 GAN(Generative Adversarial Networks) 생성모델을 사용하고 알고리즘은 styleGAN3 선정, 상악치, 하악치 별로 모델학습을 진행하여 합성데이터 생성 • styleGAN3 알고리즘과 관련하여 명시적으로 알려진 위험은 없는 것으로 파악되나 합성 이미지 생성 후 지나치게 원본과 유사한 생성, 과적합(overfitting) 등이 발생하지 않았는지에 대해 안전성 검증 필요

③ 원본데이터 확보 후 탐색적 분석을 수행하는 과정에서 명세서를 작성하여 정리하고, 항목별 처리계획도 마련

원본데이터 명세서

1) 개요

항목	내용	비고
데이터명	구강 촬영 이미지 합성데이터	-
데이터 유형	비정형 이미지 데이터	jpg 파일
데이터 규모	이미지 1000장 (500명* 2장 - 상악치, 하악치)	-
특이사항	보건의료 데이터	

2) 개인정보 유형 분류표

2-1) 정보 개요

연번	정보영역	수량	비중	정보영역설명
1	상악치	500	50%	윗니 구강 촬영 이미지
2	하악치	500	50%	아랫니 구강 촬영 이미지
전체	-	1000	100.0%	-

2-2) 정보 상세

연번	항목명	구분	설명	예시
1	구강 촬영데이터	① 구개	입천장	-
		② 혀	구강의 바닥에 위치한 근육 조직	-
		③ 상악치	윗니	-
		④ 하악치	아랫니	-
		⑤ 충치 영역	절치와 구치 중 일부	-
		⑥ 보철 영역	절치와 구치 중 일부	-
		⑦ 그 외	구강 촬영 시, 같이 촬영된, 코, 턱, 인중, 수염, 배경 등	
2	구강 촬영데이터의 메타데이터	-	-	-

항목별 개인정보 처리 계획

개인정보 처리계획						
연번	항목명	정보영역		처리방법		세부방법 및 처리수준
		정보영역	설명			
1	구강 촬영데이터	① 구개	입천장	전처리 시	유지	별도 처리하지 않음
		② 혀	-			
		③ 상악치	윗니			
		④ 하악치	아랫니			
		⑤ 충치 영역	-			
		⑥ 보철 영역	-			
		⑦ 그 외	-	전처리 시	삭제	구강 촬영 시 같이 촬영된, 코, 턱, 인중, 수염 배경 등은 활용 목적에 필요 없으므로 가장자리 부분을 cropping 처리하는 방식으로 제거
2	구강 촬영데이터의 메타데이터	-	-	-	-	-

④ 생성된 합성데이터 결과를 검토하고, 안전성 검증 등이 적절히 수행되었는지 종합적으로 검토

합성데이터 명세서

1) 개요

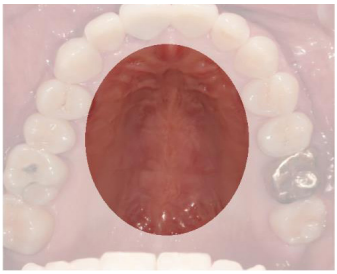
항목	내용	비고
데이터명	구강 촬영 이미지 합성데이터	-
데이터 유형	비정형 이미지 데이터	PNG 파일
데이터 규모	이미지 1000장	-
특이사항		

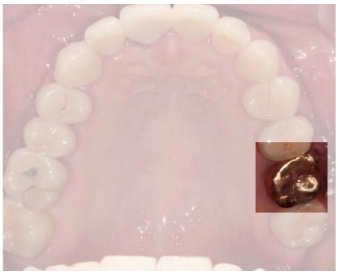
2) 정보 유형 분류표

2-1) 정보 개요

연번	정보영역	수량	비중	정보영역설명
1	상악치	500	50%	윗니 구강 촬영 이미지
2	하악치	500	50%	아랫니 구강 촬영 이미지
전체	-	1000	100.0%	-

2-2) 정보 상세

연번	항목명	구분	설명	예시
1	구강 촬영데이터	① 구개	입천장	
		② 혀	구강의 바닥에 위치한근육 조직	

연번	항목명	구분	설명	예시
1	구강 촬영데이터	③ 상악치	윗니	
		④ 하악치	아랫니	
		⑤ 충치 영역	절치와 구치 중 일부	
		⑥ 보철 영역	절치와 구치 중 일부	
2	구강 촬영데이터의 메타데이터	-	-	-

3) 합성데이터 예시

연번	항목명	상악치	하악치
1	구강 촬영 합성 데이터		
			
			
2	구강 촬영데이터의 메타 데이터	-	-

합성데이터 안전성 자체 검토 결과서

1) 측정 방법

※ 필요한 합성데이터 수량(500장)의 2배(1,000장) 이하의 합성데이터를 생성한 뒤, 유용성·안전성을 검증하고 기준 미달인 데이터는 삭제
※ 유용성·안전성이 동시에 만족하는 경우까지 삭제와 검증을 반복
※ 삭제를 반복하여 목표한 데이터 수량을 확보하지 못한 경우, 다시 합성데이터 생성 후 재검토

2) 측정지표 선정

연번	지표명	지표 설명 및 측정 방법
1	생성과정 검증	※ 원본데이터 복원이 어렵거나 불가능한 합성데이터 생성모델을 선택하고 적절한 데이터 전처리를 수행하였는지에 대해 생성과정을 검토 ① 생성 방법 검토 - 적절한 비가역적인 알고리즘(styleGAN 등) 사용 여부 확인 - 완전합성 방식의 이미지 생성 방법 사용 여부 확인 ② 데이터 전처리 검토 - 메타데이터 처리 여부(메타데이터 삭제 또는 비식별화) - 기타 합성에 필요하지 않은 데이터 영역 삭제 등
2	구조적·지각적 유사도 검증	※ 원본-합성데이터셋의 모든 이미지 조합간 유사성을 비교하여 구조적으로 과도하게 비슷하게 생성된 이미지가 있는지를 측정 ※ 이미지의 구도, 휘도, 대비를 하나의 품질 점수로 통합한 지표인 SSIM(Structural Similarity Index Measure) 지수 값 사용 ※ 계산식은 다음과 같고, 값이 0에 가까울수록 안전성 높음 $* LPIPS = \sum_l \frac{1}{H_l W_l} \sum_{h,w} \| w^l \odot (\hat{y}^l_{hw} - \hat{y}^l_{0hw}) \|_2^2$
		※ 원본-합성데이터셋의 모든 이미지간 유사성을 비교하여 지각적으로 과도하게 비슷하게 생성된 이미지가 있는지를 측정 ※ 사람의 시각인지능력 측면에서 원본-합성이미지 간 시각적 유사성을 측정 ※ LPIPS(Learned Perceptual Image Patch Similarity) 지수 값 사용, 단, 다른 지표와의 통일 및 점수산출 간편성을 위해 1에서 뺀 값을 사용(1-LPIPS). 이 경우 0에 가까울수록 안전성 높음 $* SSIM(x,y) = \frac{(2\mu_x\mu_y + C_1)(2\sigma_{xy} + C_2)}{(\mu_x^2 + \mu_y^2 + C_1)(\sigma_x^2 + \sigma_y^2 + C_2)}$
3	주관적 검증	※ 정량평가의 한계를 보완하기 위해 전문가·관계자가 합성이미지의 원본 유사성·개인식별성 등을 직접 검수(정성평가) ※ 비정형 데이터는 형태가 다양하고 예상치 못한 요인이 많아 정량적 지표로만 안전성을 검증하는 것은 한계가 존재하므로, 사람의 주관적 검증을 통해 최종 확인 ※ 유사성 검증지표(구조적 유사성, 지각적 유사성)가 비교적 높게 측정된 이미지 등에 대해서만 선별 검사하고, 적절하지 않다고 판단되는 이미지는 삭제

3) 지표별 임계값

연번	지표명	임계값 설정		설정 근거 / 방법
		①상악치	②하악치	
1	생성과정 검증	정성평가이며 별도의 임계값 없음		-
2	구조적 유사도 검증	0.5450	0.4615	1) 원본데이터를 50%, 50% 크기로 임의분할하여 A, B 데이터셋으로 재구성 2) A를 원본, B를 합성데이터로 가정한 후 SSIM과 LIPIPS 를 측정하는 과정을 100회 이상 반복 3) 누적된 측정값의 99분위수를 정해 임계값으로 산출
	지각적 유사도 검증	0.5550	0.5392	
3	주관적 검증	별도의 임계값 없음		구조적·지각적 유사성 측정값이 비교적 높게 측정된 이미지에 대해 합성 이미지의 원본유사성·개인식별성 등을 추가로 선별검수하여 안전성을 검토

4) 안전성 자체 측정 수행

※ 상악치, 하악치에 대해 각각 따로 생성모델을 구축하고, 생성하였으므로 측정과 임계값 산출은 상악치, 하악치에 대해서 아래와 같이 각각 수행

① 생성과정 검증 : 수행 여부를 검토할 수 있는 체크리스트를 구성하여 평가

② 구조적·지각적 유사도 검증 : 각각의 합성데이터에 대해서 모든 원본 이미지에 대해 1:1로 이미지 유사성 비교 점수를 측정. 합성데이터 하나와 원본 전체를 비교 후 평균을 내는 식으로 전수조사

③ 주관적 검증 : 전체 합성데이터셋의 측정 평균은 모두 임계값 이하로 측정되나, 일부 이미지가 임계값보다 높게 측정된 합성이미지는 안전성에 대해 주관적 검증을 별로로 수행

5) 최종 안전성 자체 측정결과 및 평가

5-1) 측정결과

① 생성과정 검증

구분	검토사항	생성과정	평가	
			상악치	하악치
생성 방법 평가	적절한 비가역적인 알고리즘 사용여부	GAN(styleGAN3)	적절	적절
	완전 합성 방식의 이미지 생성 방법 사용여부	image generation (완전 합성방식)	적절	적절
데이터 전처리 평가	메타데이터 처리 여부	메타데이터 사용하지 않음	적절	적절
	합성에 필요하지 않은 데이터 영역 삭제 여부	치아 외 영역 삭제	적절	적절

② 구조적·지각적 유사도 평가

구분		임계값 (99분위 수)	측정 결과	임계값 비교
상악치	구조적 유사성	0.5450	0.4519	적절
	지각적 유사성	0.5550	0.5045	적절
하악치	구조적 유사성	0.4615	0.3670	적절
	지각적 유사성	0.5392	0.4704	적절

③ 주관적 검증

▶ 정량평가의 한계를 보완하기 위하여, 유사도 검증 자료를 기반으로 측정값이 임계값보다 높게 측정된 합성 이미지에 대해 원본유사성·개인식별성 등을 추가로 선별 검수하여 안전성을 검토

5-2) 결과 평가

총 평	※ '완전합성방식', 'GAN', '식별성 높은 부위 제거' 등의 방법을 통해 합성데이터 생성 시 개인이 식별될 위험을 충분히 낮추었음 ※ 구조적·지각적 유사도 평가결과 측정 평균은 모두 임계값 이하로 측정됨. 단 일부 임계값보다 높게 측정된 합성이미지가 존재하여 해당 이미지에 대해서는 안전성에 대해 주관적 검증을 추가 수행함 ※ 해당 샘플에 대해 원본-합성데이터 간 육안 검수를 통해 구조적·지각적 유사도가 높게 측정된 합성 이미지의 안전성에 대해 주관적 검증을 하였으며, 검증 결과 원본과 유사하거나 개인이 식별될 만한 요인은 없는 것으로 검토됨

부록7. 합성데이터 생성 참고 양식

■ 양식 1-1) 합성데이터 활용 계획서

▶ 합성데이터 생성 목적과 활용형태, 생성에 참여하는 각 주체를 명시하는 문서

합성데이터 활용 계획서

참여자	조직/부서명			
	담당자 직위		담당자 성명	
	전화번호		이메일 주소	

활용 목적				
활용 구분	활용형태	☐ 내부 이용	☐ 타 부서 이용(제공)	☐ 특정 제3자 제공
		☐ 불특정 3자 제공		
		☐ 데이터 완전 공개		
	생성/활용 주체	원본데이터 보유자	합성데이터 생성자	합성데이터 활용자
	활용 환경	☐ 완전개방 환경* ☐ 부분개방 환경** ☐ 폐쇄적 환경 * 인터넷이 연결된 다른 정보의 유입이 가능한 분석 환경 ** 폐쇄환경은 아니지만 취급자 또는 제공자의 통제가 가능한 내부의 분석환경을 말함		
	반복 제공 여부	☐ 1회 제공 ☐ 반복 제공 예정 (회 예정)		
	제공 방법	☐ 온라인 ☐ 오프라인		
기타	원본데이터 명칭			
	원본데이터 내역			
	합성데이터 이용 기간			

첨부서류	

■ 양식 1-2) 합성데이터 생성 기초자료 명세서

▶ 원본데이터를 통해 합성데이터를 생성하는 과정을 명시

합성데이터 생성 기초자료 명세서	
합성데이터명	
생성 기간	

원본데이터 특성			
원본데이터명			
원본데이터 유형	☐ 개인정보	☐ 가명정보	☐ 기타 ( )
데이터 업종			
데이터 속성			
데이터 형식	☐ 정형데이터	☐ 비정형 데이터(이미지)	☐ 비정형 데이터(기타)

합성데이터 생성계획		
번호	구분	검토사항
1	환경 및 데이터 준비	
2	데이터 이해	
3	합성데이터 생성	

■ 양식 1-3) 원본데이터/합성데이터 명세서 작성 예시(두 양식 모두 동일)

▶ 합성데이터/원본데이터의 세부 정보를 정리한 데이터 명세

합성데이터 명세서(정형데이터)

1) 개요

항목	내용	비고
데이터명		
데이터 유형		
데이터 규모		
특이사항		

2) 정보 유형 분류표

2-1) 정보 개요

연번	정보영역	수량	비중	정보영역설명

2-2) 정보 상세

연번	자료형태	정보영역	항목명	항목 설명

3) 합성데이터 예시

컨텐츠유형명	고객연령	고객성별	주소	발행요일	발행월	발행시각

합성데이터 명세서 (비정형데이터)

1) 개요

항목	내용	비고
데이터명		
데이터 유형		
데이터 규모		
특이사항		

2) 정보 유형 분류표

2-1) 정보 개요

연번	정보영역	수량	비중	정보영역설명

2-2) 정보 상세

연번	항목명	구분	설명	이미지 예시

3) 합성데이터 예시

연번	항목 설명	이미지 예시

■ 양식 1-4) 항목별 개인정보처리 계획서

▶ 원본데이터 세부속성 파악·분석 후 데이터 내 항목별로 처리계획 마련 가능
▶ 데이터 전처리 시 참고 가능

항목별 개인정보 처리 계획

개인정보 처리계획						
연번	항목명	정보영역		처리방법		세부방법 및 처리수준
		정보영역	설명			

양식 1-5) 합성데이터 안전성 심의 검토 결과서

▶ 생성된 합성데이터를 익명정보로 평가받기 위해 심의위원회 평가 등을 거칠 경우 사용

합성데이터 안전성 심의 검토 결과서

검토위원 정보	성명		소속		직위	

검토 대상	☐ 신규 ☐ 보완	
최종검토결과	☐ 적정(승인) ☐ 조건부 승인 ☐ 부적정(반려)	
세부결과	합성데이터 활용 목적 설정 적정성	☐ 적합 ☐ 미흡
	식별 위험성에 관한 결과 적정성	☐ 적합 ☐ 미흡
	항목별 합성데이터생성 적정성	☐ 적합 ☐ 미흡
	안전성 평가 적정성	☐ 적합 ☐ 미흡
	활용환경의 기술적/관리적 보호조치 적정성	☐ 적합 ☐ 미흡 ☐ 해당없음
종합검토의견		

위와 같이 안전성 검토 결과를 통지합니다.

년 월 일

서명란	

양식 1-6) 합성데이터 관리대장

▶ 합성데이터의 제 3자 전달과 같은 안전한 관리 시에 사용

합성데이터 관리대장

구분		내용	비고
필수	기간		
	항목		
	합성데이터 이용자		
	합성데이터 생성자		
상세	목적 / 사유 / 근거		
	제한사항		
	기타 (추가 정보)		

년 월 일

책임자	

부록8. 자주 묻는 질문 [FAQ]

Q1 | 합성데이터를 언제 사용하는 것이 좋은가요?

A. 합성데이터는 데이터가 부족하거나 데이터를 수집·이용하기 어려운 상황에서 사용할 수 있습니다. 데이터 증강이나 데이터 다양성 증가를 위해 합성데이터를 사용할 수 있고, 데이터 접근 및 공유가 어려운 상황에서도 합성데이터를 사용하여 해결할 수 있습니다. 특히, 합성데이터를 익명정보로 인정받은 경우에는 반복적으로, 목적에 상관없이, 다양한 이용자가 활용할 수도 있습니다.

Q2 | 개인정보로 만든 합성데이터를 외부 기관 간에 공유할 수 있나요?

A. 설정한 합성데이터의 활용 범위에 따라 공유 가능 여부가 달라집니다. 예를 들어, 활용 범위를 완전개방 및 외부공개로 설정한 경우 제약 없이 합성데이터를 공유할 수 있습니다. 이러한 경우 높은 수준의 임계값을 설정하여 안전성 검증을 하고 객관적인 평가 절차도 거치며 합성데이터를 생성해야 합니다. 또한, 당초에 개인정보 수집·이용이나 제3자 제공에 대한 정보주체의 동의를 별도로 받은 경우라면 동의받은 목적에 따라 외부기관 공유가 가능합니다.

Q3 | 안전성 및 유용성의 검증은 필수적인가요?

A. 합성데이터는 가상의 데이터이지만, 원본데이터에 있던 개인정보가 식별될 위험 등이 존재합니다. 그에 따라 개인정보가 포함된 데이터로 합성데이터를 생성하는 경우, 안전성 검증은 필수입니다. 유용성 검증은 활용 목적에 따라 선택적으로 수행하되, 유용성이 낮은 합성데이터가 잘못된 정보를 확대·재생산하는 부작용에는 유의할 필요가 있습니다.

Q4 | 합성데이터 생성 시, 완전 합성방식만 사용해야 되나요?

A. 합성데이터는 합성정도에 따라 완전합성, 부분합성 등으로 구분됩니다. 완전 합성데이터와 달리, 부분합성의 경우 합성데이터 내에 개인정보가 원본 그대로 존재할 가능성이 있습니다. 때문에 익명정보로 인정받기 위해서는 완전 합성방식이 권고되는 것입니다. 그 외의 경우에는 생성·활용 주체의 필요에 맞게 합성정도를 결정할 수 있습니다.

Q5 | 합성데이터 생성 후, 자체 익명성 검토 과정을 통해 안전하다고 판단된 경우, 익명 데이터로 사용할 수 있나요?

A. 합성데이터의 익명성을 자체적으로 검토하여, 안전하다고 판단하고 사용하는 것은 가능합니다. 안내서에서 소개된 심의위원회 평가는 합성데이터의 안전성을 객관적으로 증명하는 하나의 방법으로 볼 수 있습니다.

Q6 | 자체 개발한 알고리즘으로 합성데이터를 생성한 경우, 익명 데이터로 인정받을 수 있나요?

A. 자체 개발한 알고리즘을 사용하여 합성데이터를 생성할 수 있습니다. 단, 개발한 알고리즘의 안전성과 신뢰성이 인정되어야 합니다. 합성데이터 생성자는 합성데이터를 생성할 때 사용하였던 알고리즘과 이를 활용하여 생성한 과정 등을 심의위원회 평가 시 심의위원에게 같이 전달하여 평가받을 수 있습니다. 이때, 모델에 대한 전반적 설명, 알고리즘의 비가역성 등 안전성 여부를 평가받아야 합니다.

Q7 | 본 안내서에서 제시한 안전성, 유용성 검증지표 외 다른 지표도 사용 가능한가요?

A. 안전성과 유용성을 측정하는 방법은 안내서에서 제시하는 방법 외에 다양한 방법이 있으므로 가능합니다. 단, 사용한 안전성, 유용성 검증 방법론이 합성데이터의 안전성과 유용성을 측정하기에 타당함이 인정되어야 합니다. 안내서에서 소개한 심의위원회 평가 등을 거쳐 이를 인정받을 수 있습니다.

Q8 | 기관 내에서 합성하고 활용할 때도 심의위원회 평가가 필요한가요?

A. 심의위원회 평가는 합성데이터가 익명정보에 부합하는지를 평가하는 방법의 하나이므로, 합성데이터를 익명정보로 활용하지 않고, 동의받은 목적으로 기관 내에서만 사용하는 경우 심의위원회 평가를 거치지 않아도 됩니다. 동의받은 목적 외일지라도 통계작성, 과학적 연구, 공익적 기록보존 등의 목적에 한정하여서는 심의위원회 평가 없이도 기관 내 합성데이터 생성·활용이 가능합니다.

Q9 | 원본데이터의 극단값은 반드시 제거해야 하나요?

A. 극단값의 경우 식별성이 높으므로, 전처리 단계에서 원본데이터의 극단값을 삭제하거나 일정한 처리를 할 수 있지만 필수는 아닙니다. 극단값이 분석에 중요한 의미를 가진다고 판단할 경우 제거하지 않아도 됩니다. 단, 후처리 과정에서 극단값으로 인해 합성데이터의 식별 위험 증가 여부를 검토하고, 필요 시 모니터링 등으로 식별 위험을 관리할 수도 있습니다.

Q10 | 원본데이터 식별자는 반드시 제거해야 하나요?

A. 기본적으로 원본데이터의 불필요한 식별자는 삭제를 원칙으로 합니다. 정형 데이터의 경우, 레코드 연결을 위해 식별자가 필요하다면 식별자를 다른 일련번호 등으로 변환하여 식별성을 최대한 제거하여 사용할 수 있습니다. 비정형 이미지의 경우, 이미지에 포함된 식별자가 있다면 해당 부분을 비식별 처리 하여야 합니다.

Q11 | 이미지 육안검사의 경우, 전수조사해야 하나요?

A. 이미지 합성데이터를 전수조사를 통해 검증한 경우 매우 높은 안전성을 확보할 수 있습니다. 다만, 다양한 이유로 전수 검사가 불가능한 경우, 안전한 내부 폐쇄환경에서의 활용, 권한있는 부서 내부 활용 등 안전한 활용환경이 전제될 시 표본검수도 가능합니다.

Q12 | 개인정보를 사용하려면 정보 주체의 동의가 필요한데, 합성데이터로 만들어 활용하면 동의는 필요 없나요?

A. 합성데이터 생성을 위한 정보주체 동의는 별도로 받지 않아도 됩니다. 합성데이터 활용 시 정보주체 동의가 별도로 필요한지는 상황에 따라 다를 수 있습니다. 완전합성데이터는 개인정보 수집 목적 내에서 정보주체의 별도 동의 없이 활용할 수 있고, 수집 목적 외일지라도 익명정보로 인정되면 동의없이 활용이 가능합니다. 자세한 내용은 2장의 4절(합성데이터 생성·활용 시 고려사항)을 참고 바랍니다.

Q13 | 개인정보를 통해 합성데이터를 만드는 경우에는, 개인정보를 보유한 데이터 보유기관이 직접 합성데이터를 생성할 수밖에 없나요?

A. 합성데이터는 원본데이터 보유기관이 직접 생성할 수도 있고, 개인정보 처리 위탁계약을 통해 제3의 기관이 생성할 수도 있습니다(개인정보보호법 제26조 등 참고). 또한, 원본데이터의 제3자 제공 동의를 별도로 받은 경우(개인정보보호법 제17조) 제3의 기관에 원본데이터를 이전하여 합성데이터를 생성하는 것도 가능합니다. 원본데이터 보유기관이 가명처리를 하여 제3의 기관에 데이터를 반출하고, 제3의 기관은 가명정보로 합성데이터를 만들 수도 있습니다. 자세한 내용은 2장의 4절(합성데이터 생성·활용 시 고려사항)을 참고 바랍니다.

Q14 | 합성데이터 사용 시 준수해야 하는 개인정보보호법 외 다른 관련 법령은 어떤 것이 있나요?

A. 원본데이터의 특성에 따라 다음과 같은 법률을 준수해야 할 수도 있습니다. 보건의료데이터를 활용하여 합성데이터 생성 시, 해당 연구가 IRB(Institutional Review Board, 기관생명윤리위원회) 심의 대상인지 확인하여 심의 대상인 경우 별도의 심의가 필요합니다.(생명윤리법 제15조 제1항, 제2항) 신용정보회사 등이 개인신용정보를 합성데이터로 생성하였을 경우, 익명정보 수준으로 처리되었는지에 대해 데이터전문기관에 익명처리 적정성 평가를 요청할 수 있습니다.(신용정보법 제26조의4 제2항 제2호)

Q15 | 심의위원회 평가는 합성데이터를 익명정보로 인정받기 위한 절차인가요?

A. 심의위원회 평가는 당초 설정된 합성데이터의 활용목적, 범위 등에 따라 합성데이터가 안전하게 생성되었는지 종합적으로 심의받는 절차입니다. 합성데이터를 익명정보로 자유롭게 활용하기 위해서는 객관적인 인정이 필요함에 따라, 심의위원회 평가를 통해 이를 증명하고 합성데이터를 자유롭게 활용할 수 있습니다. 그러나 합성데이터를 익명정보로 인정받을 때만 심의위원회 평가를 거치는 것은 아니고, 필요에 따라 활용 환경/형태에 제약조건이 있음에도 심의위원회 평가를 받아 안전하게 활용할 수 있습니다.

발 행 일 2024년 12월
발 행 처 개인정보보호위원회
지원기관 한국인터넷진흥원
수행기관 에이아이티스토리(주) 김준오 부대표
남현수 프로
인하대학교 김승환 교수
디 자 인 에이프린트(☎02-6272-6000)

※ 최신자료는 “개인정보보호위원회 누리집(pipc.go.kr)”,
“개인정보 포털(privacy.go.kr)”에서 확인할 수 있습니다.

개인정보보호위원회

데이터의 안전한 활용을 위한

합성데이터 생성·활용 안내서

개인정보보호위원회