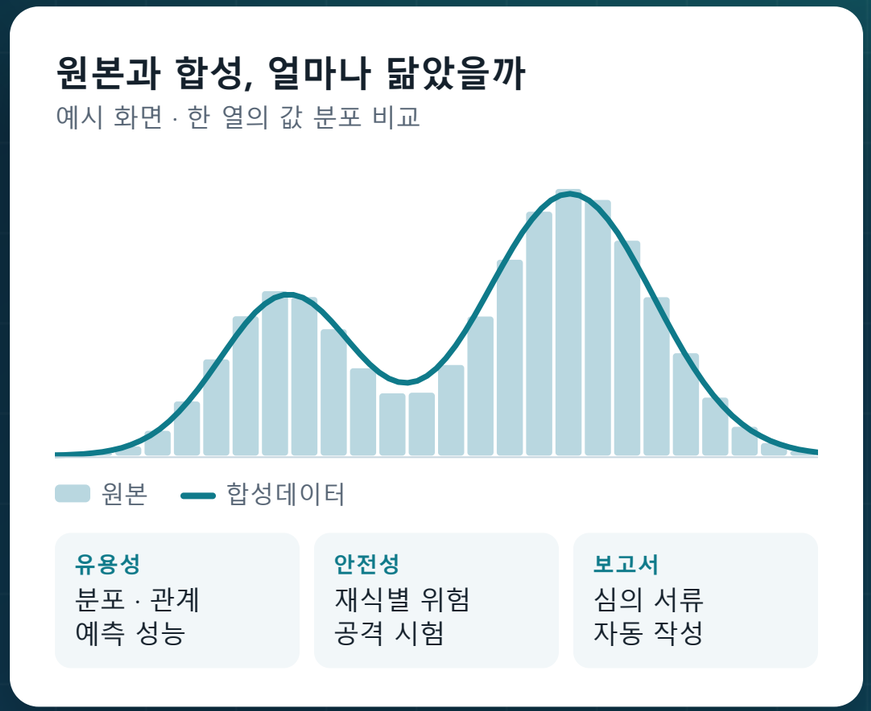
디큐SYN

합성데이터 구축 플랫폼 · 제품소개서

CSV · 엑셀만 올리면, 검증까지 끝난 합성데이터와 결과보고서를

데이터 분석부터 전처리, 생성, 유용성 · 안전성 검증, 결과보고서까지
코딩 없이 웹 화면에서 자동으로 이어지는 범용 합성데이터 플랫폼입니다.



업무 분야 무관 | 생성 모델 9종 자동 선택 | 목표 미충족 시 자동 개선 | 개인정보위 안내서 지표 반영 | 폐쇄망 설치

MITSOFT 개발 · 공급 ㈜엠아이티소프트

WHY 디큐SYN

01 데이터는 있는데, 개인정보 때문에 쓰지 못하고 있지 않으십니까

합성데이터는 원본의 통계적 특성은 살리고 실제 개인의 기록을 옮기지 않은 **새 데이터**를 만드는 방법입니다. 문제는 만들고 검증하는 과정입니다.

현장에서 자주 겪는 어려움	디큐SYN 이 바꾸는 것
합성데이터를 만들려면 데이터 과학 전문 인력과 코딩이 필요함	**파일을 올리면 분석부터 보고서까지 화면 안에서 자동 진행**
원본과 얼마나 비슷한지, 재식별 위험은 없는지 판단 기준이 막막함	**유용성 · 안전성을 같은 기준으로 계산해 용도별 사용 여부까지 제시**
개인정보위 안내서의 검증 지표와 서류를 손으로 맞춰야 함	**안내서 지표 · 체크리스트 19항목 · 자체 검토 결과서 초안을 자동 작성**
합계 · 순서 같은 업무 규칙이 깨진 데이터가 나와 다시 손봐야 함	**원본에서 업무 규칙을 찾아 생성 뒤 계산으로 맞추고 위반을 검사**
운영 데이터를 외부 클라우드 서비스에 올릴 수 없음	**기관 · 기업 서버에 설치해 원본이 밖으로 나가지 않음**

PRODUCT OVERVIEW

02 작업 하나를 만들면, 다섯 단계로 끝납니다

기본은 **자동**입니다. 전처리처럼 담당자 판단이 필요한 곳은 직접 정해서 확정할 수 있습니다.

STEP 1

파일 올리기

- CSV · TXT · 엑셀 · JSON · Parquet
- 인코딩 · 구분자 · 머리글 · 합계 행 자동 판독
- 용도(목표) 선택

▶

STEP 2

데이터 확인

- 열 형식 · 분포 · 이상값
- 민감정보 · 준식별자
- 업무 규칙 · 테이블 관계
- 품질 진단 47개 기준

▶

STEP 3

전처리 확정

- 빈 값 · 특수값 · 이상값 · 중복 행 처리
- 자동 또는 직접 선택
- 처리 전 · 후 비교

▶

STEP 4

생성 · 검증

- 생성 모델 자동 선택
- 유용성 · 안전성 검증
- 목표 미충족이면 자동 개선

▶

STEP 5

결과 · 보고서

- 용도별 사용 여부 제시
- 담당자 검토 · 승인
- 최종 보고서 · 합성데이터 반출

9종
데이터에 맞춰 고르는 생성 모델

23종
자동 탐지하는 개인정보 유형

19항목
안내서 체크리스트 자동 판정

10개
한 작업에 올리는 파일(테이블) 수

3단계
용도별 목표 수준

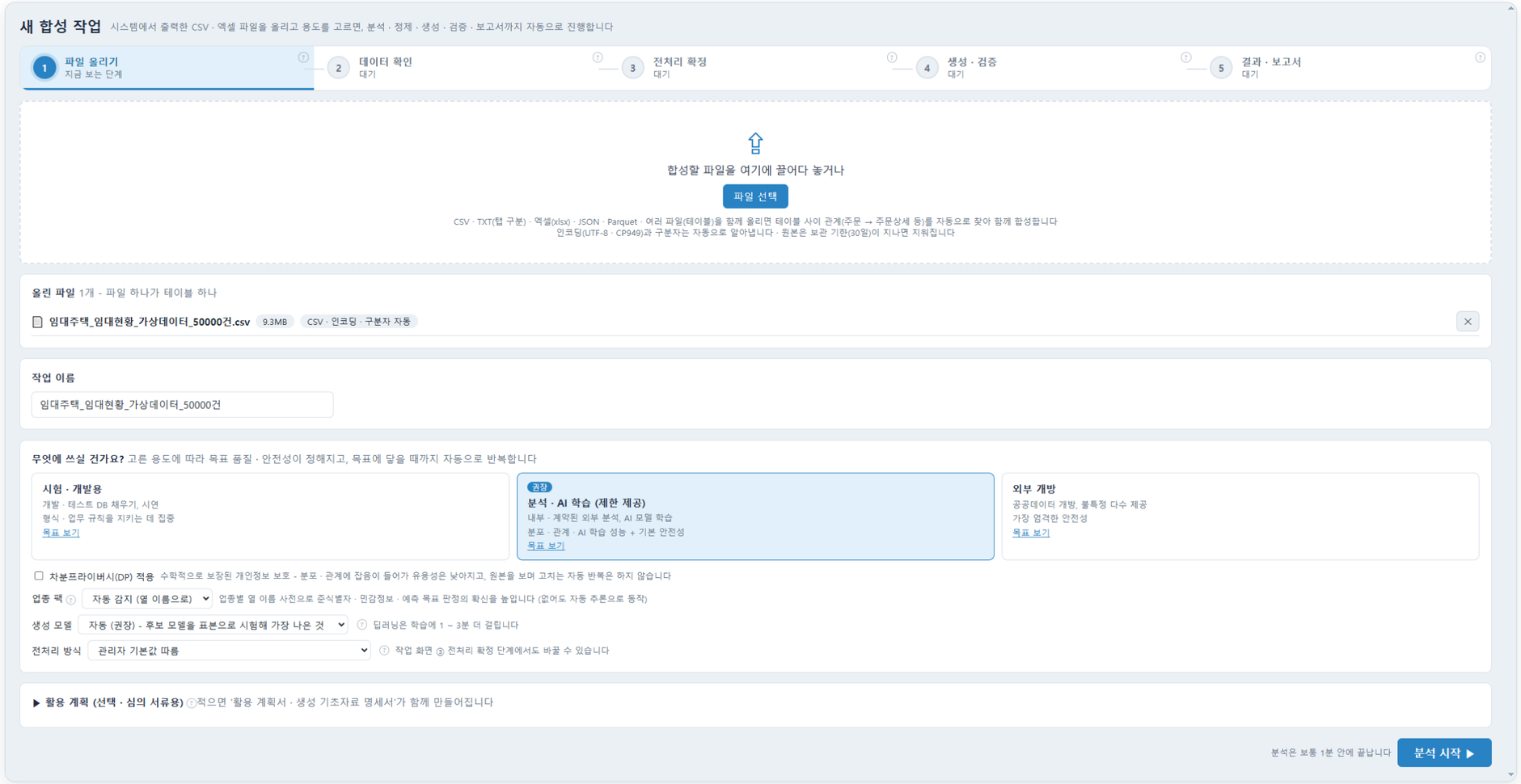
47개
데이터 품질 진단 기준

SCREEN · NEW JOB

03 파일을 올리고 용도만 고르면 시작합니다

작업 이름과 용도를 정하면 **목표 기준이 함께 정해지고**, 나머지는 자동으로 진행됩니다.

새 합성 작업



AUTO ANALYSIS · PREPROCESSING

04 올린 파일을 먼저 이해하고, 정리합니다

좋은 합성데이터는 원본을 정확히 아는 데서 시작합니다. 사람이 일일이 보지 않아도 **구조와 특성**을 찾아 정리합니다.

AUTO READ

파일 자동 판독

UTF-8 · CP949 인코딩, 구분자, 머리행, 병합 셀, 합계 행을 알아서 읽습니다.

EDA

탐색적 분석

열 형식을 추론하고 분포 · 결측 · 0 비율 · 이상값 · 특수값(예 : 9999999)을 찾아 그림으로 보여 줍니다.

PRIVACY

민감정보 탐지 23종

주민등록번호 · 휴대폰 · 계좌 · 이름 · 주소 등을 값의 형식 · 체크섬 · 열 이름으로 찾고, 준식별자 · 식별값(키)을 판별합니다.

DATA QUALITY

데이터 품질 진단 47개 기준

결측 · 중복 · 유사 중복 · 특수문자 · 표기 흔들림 · 잘못된 날짜 등을 심각도와 근거 기준으로 진단합니다.

PLAN

정제 계획 3등급

자동 · 권장 · 결정으로 나눠 사람이 볼 곳만 표시합니다. 결정을 미뤄도 보수적 기본값으로 먼저 진행합니다.

PREP

전처리 확정

보존 · 대체 · 경계값 자르기 · 행 삭제 중에서 고르고, 처리 전 · 후 통계를 비교한 뒤 확정합니다.

DATA QUALITY

05 47개 기준으로 데이터 상태를 먼저 진단합니다

문제마다 **심각도 · 근거 기준 · 건수 · 고치는 방법 · 전후 예시**를 제시하고, 항목별로 정제할지 고릅니다.

분야	진단 항목	항목 수
형식 표준화	문자 인코딩 · 구분자 · 엑셀 형식 · 머리글(열 이름) · 유니코드 표기 · 빈 값 표기 · 날짜 형식 · 숫자 형식 · 단위 혼재	9
정확성	닫히지 않은 따옴표 · 깨진 글자 · 잘못된 날짜 · 의미 없는 기본값 · 숫자 열의 문자 · 날짜 앞뒤 어긋남 · 통계적 이상치 · 날짜 범위 벗어남	8
완전성	여러 시트 · 칸이 밀린 행 · 빈 행 · 본문 없는 행 · 너무 짧은 본문 · 빈 열 · 단일값 열 · 결측률 높음 · 엑셀 셀 한도 초과 · 결측치 채우기	10
중복	완전 중복 행 · 본문 중복 · 유사 중복 · 같은 내용 열	4
노이즈	보이지 않는 문자 · HTML 태그 · 엔티티 · 줄바꿈 정리 · 연속 공백 · 앞뒤 공백	5
일관성	목록 값 정리 · 표기 흔들림	2
개인정보	개인정보(식별자) · 개인정보 열	2
학습 적합성	본문 길이 이상 · 드문 분류값 · 분류 불균형 · 식별 열(학습 변수 아님) · 치우친 분포 · 날짜 파생 열 · 1:1 대응 열	7

근거 기준

행안부·NIA 공공데이터 AI 친화적 관리 가이드라인 · NIA AI 데이터 품질관리 가이드라인 · 개인정보위 가명정보 처리 가이드라인

디큐브WEB 과 같은 기준

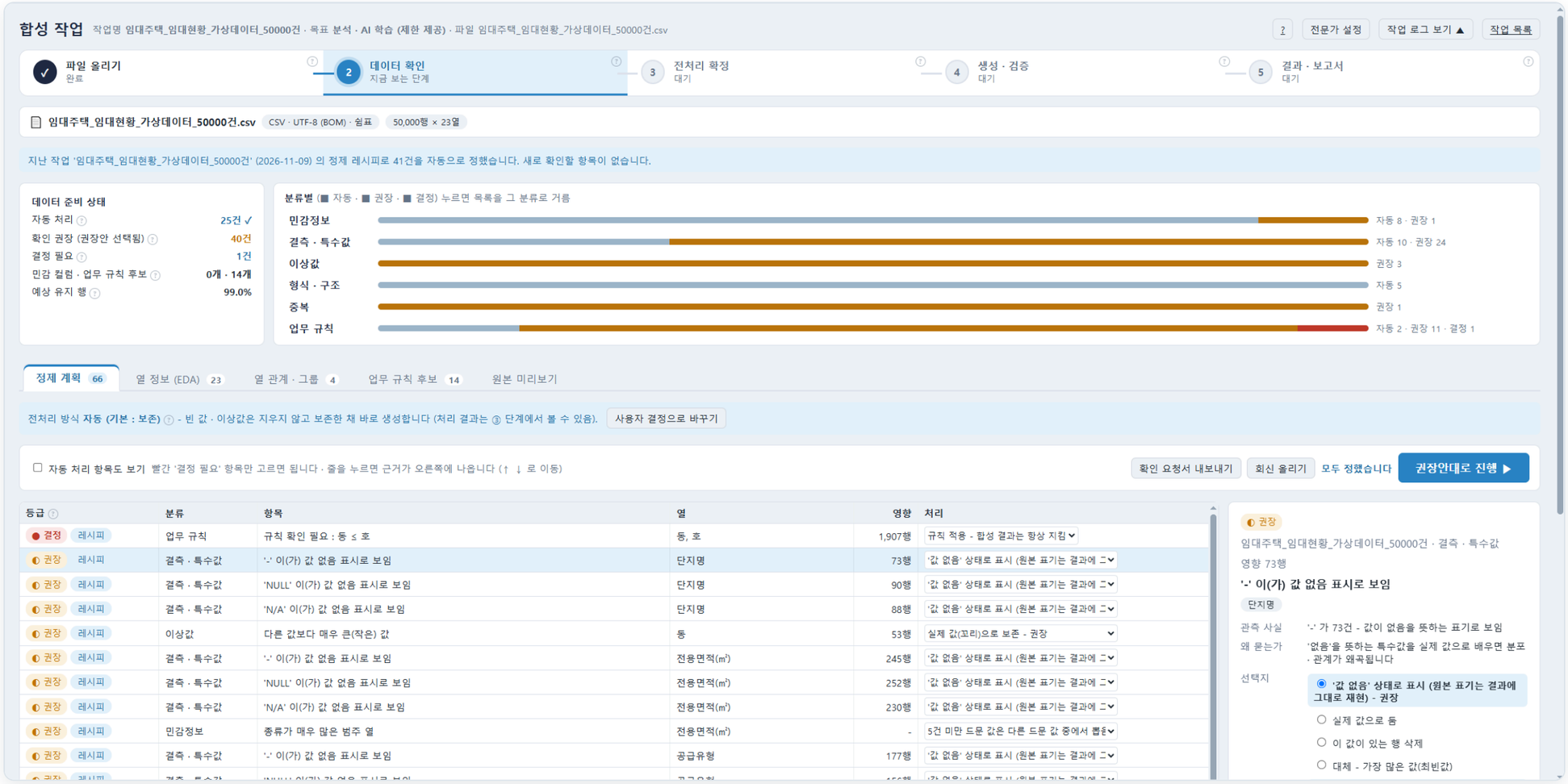
자사 데이터 품질 솔루션의 AI-Ready 정제와 같은 판정 규칙을 씁니다.

SCREEN · DATA CHECK

06 분석이 끝나면 확인할 것만 골라서 보여 줍니다

자동 처리 · 확인 권장 · 결정 필요로 나누고, 항목을 누르면 **근거와 선택지**가 옆에 나옵니다.

데이터 확인

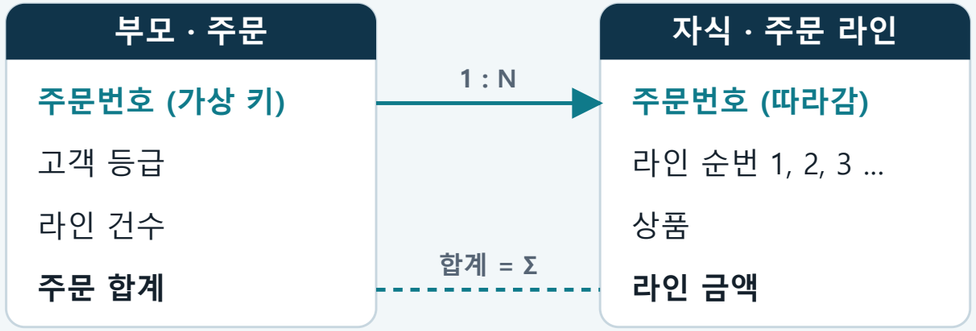


BUSINESS RULES · MULTI TABLE

07 업무 규칙과 테이블 관계까지 그대로 살립니다

합계가 맞지 않거나 부모 없는 자식 행이 있는 데이터는 쓸 수 없습니다. 규칙을 **원본에서 자동으로 찾아** 생성 단계에서 지킵니다.

여러 파일의 관계를 자동 발견 · 재현



부모를 먼저 만들고, 자식은 부모 값 · 자식 수를 조건으로 생성

참조무결성 · 부모당 자식 수 분포 · 교차 규칙 위반을 검증

자동 발견 키 포함률과 열 이름으로 부모 · 자식(1:N) 관계를 찾아 그림으로 표시

가상 키 키는 가상값으로 바꾸고 자식 행이 그대로 따라감

부모 단위 분할 학습 · 평가 · 홀드아웃을 부모 단위로 나눠 정보 섞임 방지

규칙 유형	예시	생성할 때
합계	공급가액 + 세액 = 합계금액	생성 뒤 계산으로 맞춤
곱 · 비율	단가 × 수량 = 금액, 세액 = 금액 × k	생성 뒤 계산으로 맞춤
순서	시작일 ≤ 종료일	차이로 바꿔 생성
종속	코드 → 코드명	대응 관계 유지
부모 합계	주문 합계 = Σ 라인 금액	계산으로 맞춤
부모 건수	라인 건수 = 자식 행 수	계산으로 맞춤
자식 순번	1, 2, 3 … 연속 번호	부모 안에서 다시 매김
식별값 · 키	이름 · 연락처 · 고객번호	원본과 겹치지 않는 가상값

발견한 규칙은 원본 성립률과 함께 보여 주며, 담당자가 화면에서 확인할 수 있습니다.

SCREEN · BUSINESS RULES

08 원본에서 찾은 규칙을 성립률과 함께 보여 줍니다

규칙마다 **원본 성립률과 위반 건수**를 근거로 자동 확정 · 권장 · 결정 필요를 나눕니다.

업무 규칙 후보

정제 계획 66 | 열 정보 (EDA) 23 | 열 관계 · 그룹 4 | **업무 규칙 후보 14** | 원본 미리보기

번호	규칙 (자동 발견)	유형	원본 성립	위반	판정
R01	계약시작일 ≤ 입주일	순서 (날짜)	100.00%	0	자동 확정
R02	계약시작일 ≤ 최종수정일	순서 (날짜)	100.00%	0	자동 확정
R03	입주일 ≤ 최종수정일	순서 (날짜)	99.79%	100	권장 (적용)
R04	계약시작일 ≤ 계약종료일	순서 (날짜)	99.70%	149	권장 (적용)
R05	퇴거일 ≤ 계약종료일	순서 (날짜)	99.66%	78	권장 (적용)
R06	단지코드 → 행정구역 (값이 정해지면 하나로 정해짐)	종속	99.60%	198	권장 (적용)
R07	계약시작일 ≤ 퇴거일	순서 (날짜)	99.58%	95	권장 (적용)
R08	단지명 → 단지코드 (값이 정해지면 하나로 정해짐)	종속	99.52%	239	권장 (적용)
R09	입주일 ≤ 계약종료일	순서 (날짜)	99.50%	240	권장 (적용)
R10	입주일 ≤ 퇴거일	순서 (날짜)	99.37%	140	권장 (적용)
R11	단지코드 → 주택유형 (값이 정해지면 하나로 정해짐)	종속	99.33%	334	권장 (적용)
R12	단지명 → 행정구역 (값이 정해지면 하나로 정해짐)	종속	99.22%	390	권장 (적용)
R13	단지명 → 주택유형 (값이 정해지면 하나로 정해짐)	종속	99.06%	465	권장 (적용)
R14	동 ≤ 호	순서 (수치)	96.19%	1,907	결정 필요

자동 확정

원본에서 100% 성립한 규칙은 생성 뒤 계산으로 맞추고 위반을 검사합니다.

권장 적용

99% 이상 성립하면 위반을 입력 오류로 보고 적용을 권장합니다.

결정 필요

그 미만은 업무 사정을 아는 담당자가 적용 여부를 정합니다.

SCREEN · PREPROCESSING

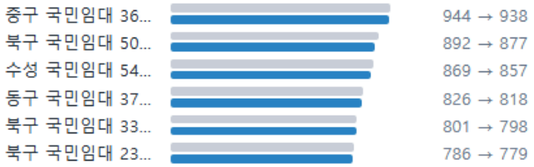
09 전처리 전후를 열마다 그림으로 비교합니다

빈 값 · 특수값 처리가 **분포를 어떻게 바꿨는지** 확정하기 전에 확인합니다.

전처리 결과 비교

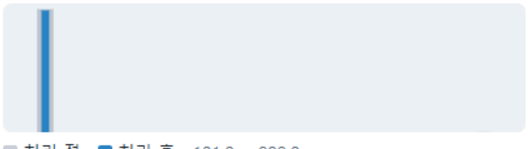
단지명 범주

빈 값 **249 → 494** · 행 50,000 → 49,500



동 정수

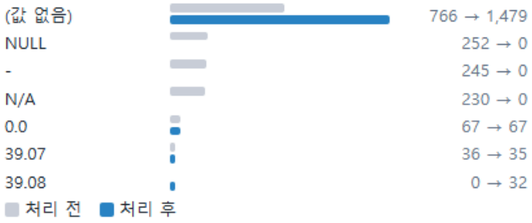
빈 값 **0 → 0** · 행 50,000 → 49,500



	평균	중앙값	표준편차	최소	최대
처리 전	106.4	105.0	29.3	101.0	999.0
처리 후	106.4	105.0	28.6	101.0	999.0

전용면적(㎡) 범주

빈 값 **766 → 1,479** · 행 50,000 → 49,500



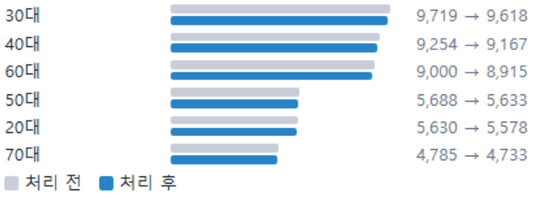
공급유형 범주

빈 값 **510 → 985** · 행 50,000 → 49,500



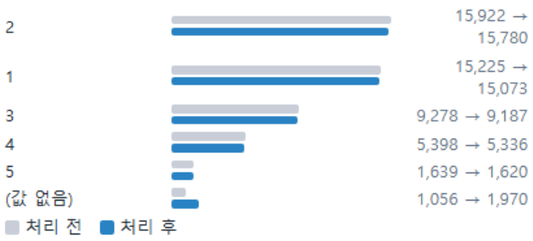
연령대 범주

빈 값 **1,222 → 2,461** · 행 50,000 → 49,500



세대원수 범주

빈 값 **1,056 → 1,970** · 행 50,000 → 49,500



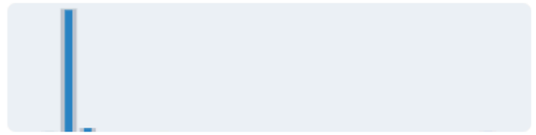
임대보증금(원) 정수

빈 값 **0 → 0** · 행 50,000 → 49,500

	평균	중앙값	표준편차	최소	최대
처리 전	34,168,116	25,460,000	154,152,159	0.000	20,845,000,00
처리 후	34,186,975	25,460,000	154,889,747	0.000	20,845,000,00

월임대료(원) 정수

빈 값 **526 → 982** · 행 50,000 → 49,500



	평균	중앙값	표준편차	최소	최대
처리 전	203,349	200,000	312,571	-469,000	10,132,000
처리 후	203,436	200,000	313,985	-469,000	10,132,000

GENERATION MODELS

10 데이터에 맞는 생성 모델을 스스로 고릅니다

데이터마다 잘 맞는 방식이 다릅니다. **통계 · 규칙 · 딥러닝 9종**을 갖추고, 어떤 모델을 왜 골랐는지 결과 화면에서 보여 줍니다.

통계 기반

가우시안 코풀라
열별 분포와 상관을 재현. 빠르고 안정적이며 설명하기 쉬움

GMM 코풀라
고객군처럼 무리가 나뉜 데이터에 강함

규칙 · 그래프 기반

CART 코풀라
조건 · 상호작용 관계에 강함 (synthpop 방식)

베이지안 네트워크
범주 간 의존이 뚜렷한 데이터에 적합

딥러닝 기반 (PyTorch · CPU 동작)

코풀라-VAE
비선형 · 범주 조합 관계 학습, 과적합 방지 조기 종료

CTAB-GAN 계열
통계 일치 · 열 예측 손실로 관계 강화

Transformer 계열
앞 열을 보고 다음 열을 예측하는 자기회귀

CTGAN 계열
생성망과 판별망의 경쟁 학습 (WGAN-GP)

TabDDPM 계열
잡음을 걷어 내며 행을 만드는 확산 모델

'계열' 모델은 공개 논문의 학습 방식을 디큐SYN 공통 구조에 맞춰 자체 구현한 것입니다.

자동 선택
표본으로 후보를 미리 학습해 점수로 비교합니다 (학습 300행 이상 · 자식 테이블 제외). 복잡한 모델은 기본 모델보다 나을 때만 선택합니다.

원본 분포 보존 구조
모든 모델이 열별 분포는 원본에서 그대로 배웁니다. 어떤 모델을 써도 값의 형태가 쉽게 흐트러지지 않습니다.

차분프라이버시 (선택)
ε 0.5 ~ 10 범위에서 잡음을 더하고(가우시안 코풀라), 사용 예산과 보장 범위를 결과와 보고서에 남깁니다.

VALIDATION

11 만든 다음이 더 중요합니다, 두 축으로 검증합니다

원본을 얼마나 닮았는지(**유용성**)와 원본 속 개인을 알아낼 위험이 기준 이내인지(**안전성**)를 따로 판정합니다.

유용성 원본처럼 쓸 수 있는가

형식 · 업무 규칙	확정한 규칙을 모두 지키는지
분포 유사도	수치 KS, 범주 JSD · 카이제곱
열 관계	스피어만 · 크래머 V · 상관비 η, 원본 분할 기준선과 비교
구별 불가능성	pMSE · PO50 · SPECKS · 판별 AUC
AI 학습 성능 (TSTR)	합성으로 학습한 모델을 실제 데이터로 평가 (분류 F1 · 회귀 R^2)
테이블 관계	참조무결성 · 자식 수 분포 · 교차 규칙
집단별 비교	집단을 나눠도 원본과 비슷한지 확인

안전성 원본을 드러내지 않는가

원본 행 복제	합성 행이 원본과 똑같은 비율, 식별값 잔존
근접 기록	DCR · NNDR : 원본을 외운 기록이 있는지
소속 추론 공격	MIA : 거리 공격 · 밀도비 공격 중 센 쪽으로 판정
연결 위험 (CAP)	준식별자로 민감값을 맞힐 수 있는지
공격 시험	단독 식별 · 연결 · 추론을 직접 시도 (Anonymeter 방식)
재식별 지표	준식별자 고유성 · I-다양성 · t-근접성
안내서 위험도	구별 · 연결 · 추론 위험도 (분할 반복 임계)



최종 판정은 분포 · 관계를 학습 데이터와 비교하고, AI 학습 성능 · 공격 시험을 학습에 쓰지 않은 **홀드아웃**으로 측정합니다. 1,000행 미만은 80 : 20 으로 나눕니다.

AUTO IMPROVEMENT

12 목표에 못 미치면, 사람이 손대기 전에 먼저 고칩니다

미충족 항목에 맞는 처방을 골라 다시 검증합니다. 전 과정은 **회차별 이력**으로 남아 보고서에 실립니다.

미충족 항목에 따른 4단계 처방

1	**레코드 교체**	원본에 지나치게 가까운 기록, 연결 위험이 높은 칸만 골라 다시 뽑음
2	**열 · 관계 보정**	분위 매핑으로 분포를 맞추고, 순위 재배열로 수치 열 관계를 보정
3	**모델 재학습**	상관 축소 · 평활 조정, 다른 난수로 재생성, 딥러닝 모델로 전환
4	**공개 정밀도**	준식별자 날짜의 정밀도를 주 · 월 단위로 낮춤

그래도 미충족이면 : 자동 반복 (선택)

더 오래 개선 → 공개 정밀도 낮추기 → 가장 덜 닮은 열 제외 → 목표 완화 순으로 적용해 다시 생성합니다.

과하게 고치지 않도록 안전장치를 둡니다

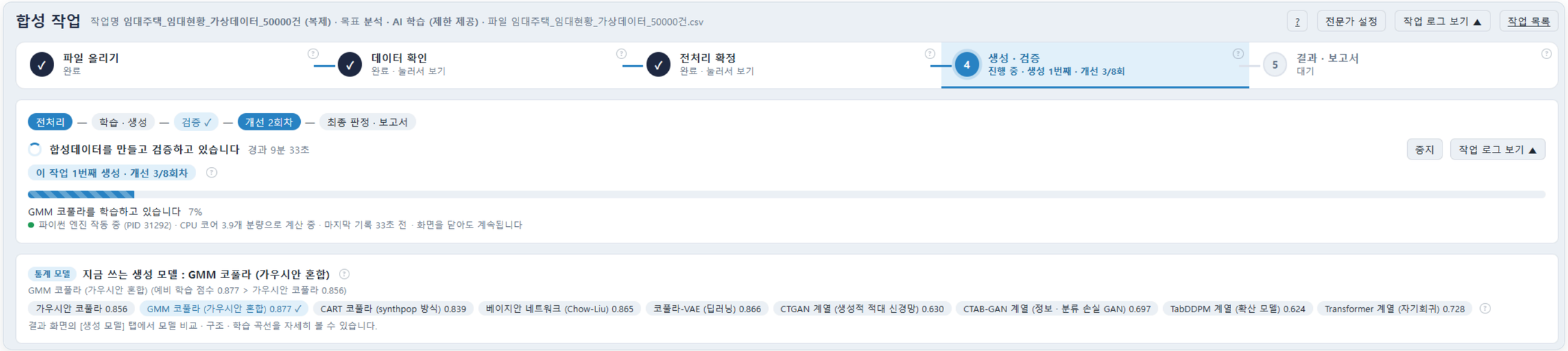
- ✓ **안전성 우선** : 안전성 미충족이 있으면 그것부터 처리합니다.
- ✓ **나빠지면 취소** : 관계 보정 후 다시 평가해, 결과가 나빠지면 버리고 다른 방법으로 넘어갑니다.
- ✓ **두 목표로 비교** : 안전성 미충족이 이어지면 후보 3개를 만들어 안전성과 유용성 양쪽에서 가장 나은 안을 고릅니다.
- ✓ **분리된 평가** : 반복 개선은 개선 평가 데이터로 하고, 최종 시험에는 홀드아웃을 씁니다 (1,000행 미만은 함께 씀).
- ✓ **차분프라이버시 보호** : 예산이 누적되는 자동 반복은 따로 켜야 하며 누적 ε 를 기록합니다.

SCREEN · RUN · COMPARE

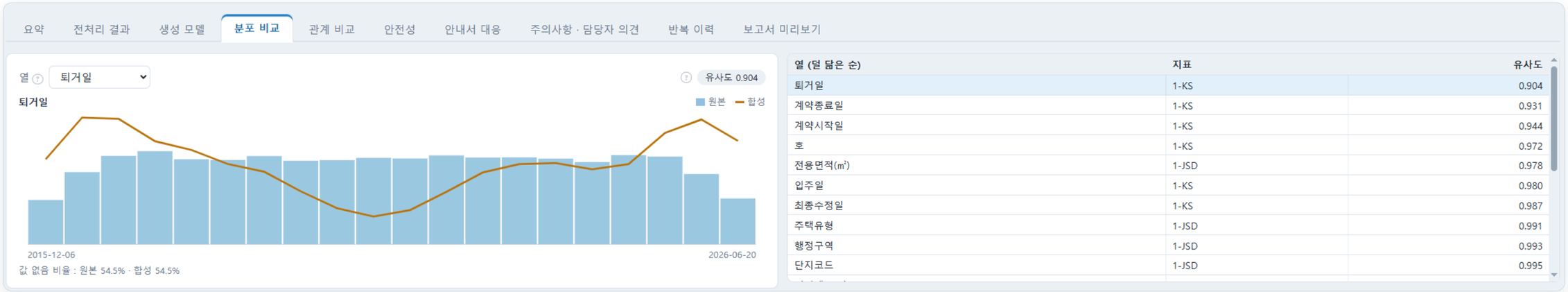
13 생성 중에도 진행 상황과 결과를 바로 확인합니다

지금 단계 · 쓰는 모델 · 후보 점수를 실시간으로 보고, 끝나면 **열마다 원본과 합성을 비교**합니다.

생성 · 검증 진행



분포 비교



TARGET PROFILES

14 쓰임새에 맞는 목표를 처음에 고릅니다

용도를 고르면 합격 기준이 함께 정해집니다. 기준값은 관리자가 **기관 정책에 맞게 조정**할 수 있습니다.

판정 항목	시험 · 개발용	분석 · AI 학습 (제한 제공)	외부 개방
대표 용도	개발 · 테스트 DB 채우기, 시연	내부 · 계약된 외부 분석, AI 학습	공공데이터 개방, 불특정 다수 제공
분포 유사도 (평균)	0.90 이상	0.95 이상	0.95 이상
분포 유사도 (최저 열)	-	0.90 이상	0.90 이상
열 관계 차이	-	원본 분할 기준선의 2배 이내	기준선 2배 이내 + 최대 차이 0.30 이하
AI 학습 성능 (TSTR)	-	0.90 이상	0.90 이상
원본 근접 기록 비율	-	8% 이하	6% 이하
소속 추론 공격 (MIA)	-	0.60 이하	0.55 이하
연결 위험 (CAP)	검증	검증	검증
준식별자 조합 고유성	-	-	검증
원본 행 복제 · 식별값 잔존 · 업무 규칙	검증	검증	검증

표의 값은 제품 기본값입니다. 작업마다 전문가 설정에서 목표값 · 생성 모델 · 반복 예산 · 결과에서 뺄 열을 바꿀 수 있으며, 바꾼 내용은 보고서에 기록됩니다.

VERDICT · USE CRITERIA

15 합격 · 불합격이 아니라, 어디에 쓸 수 있는지를 알려 드립니다

일부 항목이 미충족이어도 쓸 수 있는 용도가 있습니다. 담당자가 판단할 수 있도록 **미충족 항목과 용도를 연결**해 제시합니다.

안전성 충족 · 유용성 충족

유용성 · 안전성 목표를 모두 충족. 승인 후 반출합니다.

안전성 충족 · 유용성 부분충족

미충족 항목이 필요한 용도만 제한합니다. 쓸 수 있는 용도가 없으면 유용성 미충족입니다.

안전성 미충족 · 사용 불가

안전성은 필수 요건입니다. 버튼이 잠기고 마우스를 올리면 사유를 안내합니다.

미충족 항목	쓰면 안 되는 용도 (사용 불가)	영향 없는 용도 (사용 가능)
형식 · 업무 규칙	시스템 테스트 · 적재, 교육 · 시연	없음 (먼저 고쳐야 함)
분포 유사도	기초 통계 · 집계	시스템 테스트
열 관계	상관 · 회귀, 교차 분석	기초 통계, 시스템 테스트
AI 학습 성능 (TSTR)	AI 모델 학습 · 검증	기초 통계, 시스템 테스트
참조무결성 · 자식 수 · 교차 규칙	테이블 결합 분석	단일 테이블 분석

결과 화면에는 용도 · 사용 여부 · 이유(실제 결과와 기준)가 작업마다 표로 나오고, 행마다 [설명] 버튼으로 쉬운 해설을 볼 수 있습니다.

SCREEN · RESULT

16 용도별 사용 여부까지 한 화면에 제시합니다

판정 요약 아래에 **용도 · 사용 여부 · 이유**와 활용 범위별 기준이 함께 나옵니다.

결과 · 판정 요약

합성 작업 작업명 임대주택_임대현황_가상데이터_50000건 · 목표 분석 · AI 학습 (제한 제공) · 파일 임대주택_임대현황_가상데이터_50000건.csv

? | 전문가 설정 | 작업 로그 보기 ▲ | 작업 목록

✓ 파일 올리기 완료 — ✓ 데이터 확인 완료 · 눌러서 보기 — ✓ 전처리 확정 완료 · 눌러서 보기 — ✓ 생성 · 검증 완료 · 눌러서 보기 — 5 결과 · 보고서 지금 보는 단계

안전성은 충족했지만 유용성 관문 중 **관계 (평균 차이), AI 학습 성능 (TSTR · '계약상태' 예측)** 이(가) 목표에 못 미쳤습니다. 개인정보위 안내서상 유용성 검증은 활용 목적에 따라 선택할 수 있으나, 유용성이 낮으면 잘못된 정보를 확대 · 재생산할 수 있으니 활용 목적에 맞는지 확인한 뒤 승인하세요.

판정	유용성 관문 (활용 목적별 선택)	안전성 관문 (필수)	행 × 열	반복 · 소요
안전성 충족 · 유용성 미달	3 / 5 통과 1차원 분포 평균 0.985	5 / 5 통과	50,000 → 49,500 × 23	최대 8회 · 1,256초

최종 판정은 처음부터 떼어 둔 홀드아웃 데이터(원본의 약 15%)로 한 번 한 결과입니다. 개선 중 점수와 함께 보고서 7장에 적힙니다.

용도	사용 여부	이유 (결과 / 기준)
시스템 개발 · 테스트 (적재 · 집계 · 화면 개발), 교육 · 시연	사용 가능	통과 : 형식 · 업무 규칙 위반 0건
열별 기초 통계 · 분포 파악	사용 가능	통과 : 1차원 분포 (평균) 0.985 / 목표 0.95 이상 · 1차원 분포 (최솟값) 0.904 ('퇴거일') / 목표 0.90 이상
상관 · 회귀 등 변수 간 관계 분석	부적합	미달 : 관계 (평균 차이) 0.028 / 기준 0.020 이하 (원본끼리 차이 0.007 × 2)
AI 모델 학습 · 검증	부적합	미달 : AI 학습 성능 (TSTR · '계약상태' 예측) 0.657 / 목표 0.90 이상 (합성 학습 macro F1 0.649 · 원본 학습 macro F1 0.988 · 찍기 0.250)
테이블 결합 분석 (부모 · 자식)	해당 없음	테이블이 하나입니다
개별 기록 · 개인 단위 추적	불가	모든 행이 가상 기록 · 가상 식별값이라 실제 개인과 대응되지 않습니다
극단값 · 희귀 사례 정밀 분석	주가 검증 필요	분포 관문은 전체 경향을 보며, 희귀한 조합은 안전성을 위해 그대로 재현되지 않을 수 있습니다
외부 개방	주가 검증 필요	외부 개방 목표로 다시 생성해 모든 관문을 통과해야 합니다

미달 항목에 따른 용도 구분

미달 항목	쓰면 안 되는 용도 (사용 불가)	영향 없는 용도 (사용 가능)	이 작업	설명
형식 · 업무 규칙	시스템 테스트 · 적재, 교육 · 시연	없음 (먼저 고쳐야 함)	통과	설명
1차원 분포	기초 통계 · 집계	시스템 테스트	통과	설명
열 관계	상관 · 회귀, 교차 분석	기초 통계, 시스템 테스트	미달	설명
AI 학습 성능 (TSTR)	AI 모델 학습 · 검증	기초 통계, 시스템 테스트	미달	설명
참조무결성 · 자식 수 · 교차 규칙	테이블 결합 분석	단일 테이블 분석	측정 안 함	설명

활용 범위별 기준

활용 범위	안내서 구분	안전성 기준	유용성 기준	절차	이 작업	설명
내부 이용	⑤ 내부 이용 (보통)	시험 · 개발용 목표 (T1) : 복제 · 식별값 잔존 · 연결 위험	해당 용도 항목	자체 검토	안전성 기준 충족	설명
타 부서 제공	④ 타 부서 이용	분석 · AI 학습 목표 (T2)	해당 용도 항목	자체 검토 · 제공 기록	안전성 기준 충족	설명
기관 간 공유 (받는 기관 특정)	③ 특정 제3자	분석 · AI 학습 목표 (T2) : 근접 기록 8% 이하, MIA 0.60 이하	받는 기관의 활용 목적 항목	심의 권장, 재식별 · 재제공 금지 약정	안전성 기준 충족	설명
기관 간 공유 (참여 기관 누구나 열람)	② 불특정 제3자	외부 개방 목표 (T3)	받는 기관의 활용 목적 항목	심의위원회 평가	외부 개방 목표 (T3)로 다시 생성 필요	설명
공공데이터 개방	① 외부 공개 (최고)	외부 개방 목표 (T3) : 근접 기록 6% 이하, MIA 0.55 이하, 준식별자 고유성	형식 · 분포는 필수, 관계 · TSTR 미달은 '활용 제한' 명시	심의위원회 평가, 잔여 위험 관리계획	외부 개방 목표 (T3)로 다시 생성 필요	설명

SHARING SCOPE

17 제공 범위가 넓어질수록 더 엄격한 기준을 안내합니다

개인정보위 안내서의 활용 범위 구분을 바탕으로, **내부 이용부터 공공데이터 개방까지** 필요한 안전성 목표와 절차를 제시합니다.

제공 범위	안내서 구분	필요한 안전성 목표	권장 절차
내부 이용	내부 이용	시험 · 개발용 목표	자체 검토
타 부서 제공	타 부서 이용	분석 · AI 학습 목표	자체 검토 · 제공 기록
기관 간 공유 (받는 기관 특정)	특정 제3자	분석 · AI 학습 목표	심의 권장, 재식별 · 재제공 금지 약정
기관 간 공유 (참여 기관 열람)	불특정 제3자	외부 개방 목표	심의위원회 평가
공공데이터 개방	외부 공개	외부 개방 목표	심의위원회 평가, 잔여 위험 관리계획

이 작업은 어디까지 가능한가

실제 적용한 목표값과 검증 결과로 범위별 상태를 표시합니다. 부족하면 어떤 목표로 다시 생성해야 하는지 알려 드립니다.

최종 판단은 기관의 몫입니다

디큐SYN 의 결과는 자체 검토와 심의를 돕는 근거 자료이며, 법적 판단이나 기관의 심의 절차를 대신하지 않습니다.

GUIDELINE ALIGNMENT

18 개인정보위 안내서의 지표와 서류 초안을 자동으로 만듭니다

개인정보보호위원회 「합성데이터 생성 · 활용 안내서」(2024.12) **부록의 산식**을 구현해 대응표로 보여 줍니다.

구분	지표	계산 방법
안전성	구별 위험도	원본과 같은 레코드 비율, 50:50 분할 반복 95분위 임계
	연결 위험도	레코드별 CAP
	추론 위험도	최근접 거리 비교, 적정 수준 0.5
유용성	분포 유사도	얀센-섀넌 발산 · KS · 카이제곱
	2차원 관계	피어슨 · 크래머 V · SSB/SST 행렬 차이의 표준편차
	구별 불가능성	pMSE · PO50 · SPECKS
	모형 성능	분류 4지표, 회귀 MSE · 회귀계수 신뢰구간 겹침

CHECKLIST

체크리스트 19항목

탐색 · 전처리 · 생성 · 후처리 단계를 충족 · 확인 · 추가 검토 · 기관 확인으로 자동 판정

SELF REVIEW

안전성 자체 검토 결과서

측정 방법 · 지표 · 임계 · 결과 · 총평을 안내서 서식으로 초안 작성

SPECIFICATION

원본데이터 명세서

개인정보 유형 분류표 · 항목별 처리 계획을 자동 정리

SCREEN · GUIDELINE

19 안내서 지표를 대응표로 보여 주고, 쉽게 설명합니다

지표 · 측정 방법 · 결과 · 임계값 · 판정을 한 표로 보고, **(?) 를 누르면 쉬운 해설**이 열립니다.

안내서 대응

요약 | 전처리 결과 | 생성 모델 | 분포 비교 | 관계 비교 | 안전성 | **안내서 대응** | 주의사항 · 담당자 의견 | 반복 이력 | 보고서 미리보기

구분	안내서 지표	DQ SYNTH 측정 방법	결과	임계값	판정	근거
안전성	**구별 위험도 (Single out)**	합성 3,000건 중 원본(같은 수 10,000건 임의 추출)에 같은 레코드가 있는 비율	0.0000 (원본 전체 대비 0.0000)	≤ 0.0000 (원본 50:50 분할 30회 95분위)	충족	6장 · 안전성 자체 검토 결과서
안전성	**연결 위험도 (CAP)**	준식별자 연령대, 계약시작일, 계약종료일 → 민감정보 계약상태, 공급유형, 원본 레코드별 CAP	CAP > 0.7 레코드 0.1% · 평균 CAP 0.001	레코드별 0.7 이하 (기준선 8.9%)	충족	6장 (관문 '기록별 연결 위험')
안전성	**추론 위험도 (Inference)**	합성 레코드의 원본 최근접 거리 dS < 그 원본의 다른 원본 최근접 거리 dO 인 비율 (수치 0~1 정규화 + 범주 원-핫 유클리드 거리)	0.224	≤ 0.497 또는 0.5 미만 (분할 30회 95분위)	충족	6장 · 안전성 자체 검토 결과서
안전성 (추가)	**준식별자 k-익명성**	준식별자 연령대, 계약시작일, 계약종료일, 입주일 조합이 5건 이상인 셀의 레코드 비율	합성 0.0% · 원본 0.0%	외부 개방(T3) 95% 이상 · 그 밖에는 참고 (합성 레코드는 실제 개인이 아니므로 연결 위험은 CAP 로 판정)	참고	6장
안전성 (추가)	**l-다양성 · t-근접성**	준식별자 칸마다 민감속성 '계약상태' 값의 종류(l)와 전체 분포와의 차이(t)	l ≥ 2 레코드 합성 0.0% (원본 1.5%) · t ≤ 0.2 합성 0.0% (원본 1.4%)	합성 ≥ 원본 (원본보다 나빠지지 않음)	충족	6장
안전성 (추가)	**최근접 거리 비율 (NNDR)**	가장 가까운 원본 ÷ 두 번째로 가까운 원본 거리. 실제 신규 기록 하위 5% 보다 작은 합성 비율	합성 중앙값 0.962 · 실제 신규 기록 0.934 · 하위 5% 미만 0.8%	하위 5% 미만 비율 10% 이하	충족	6장
안전성 (추가)	**공격 시험 (Anonymeter 방식)**	학습 원본 400건을 공격 대상, 홀드아웃을 통제군으로 단독 식별 · 연결 · 추론 공격을 시도	단독 식별(한 열 조건) 0.071 · 연결 0.000 · 추론('계약상태') 0.000 · 추론('공급유형') 0.000 (위험도, 95% 상한 최대 0.235)	위험도 0.1 이하 (공격 성공률이 통제군보다 거의 높지 않음)	충족	6장
유용성	**1차원 분포 유사성**	열마다 얀센-섀넌 발산(JSD) · KS 통계량(수치) · 카이제곱 검정(범주, 같은 크기 표본)	JSD 평균 0.0051 · 최대 0.0384 · KS 평균 0.0337 · 카이제곱 p ≥ 0.05 열 70%	관문 '1차원 분포' 목표 (1−KS · 1−JSD)	충족	5장 · 결과 화면 '분포 비교'
유용성	**2차원 관계 유사성**	피어슨(수치) · 크래머 V(범주) · ANOVA SSB/SST(수치×범주) 행렬 차이값의 표준편차	0.0663 (18개 열)	≤ 0.0444 (원본 분할 반복 95분위)	미달	5장 · 결과 화면 '관계 비교'
유용성	**구별 불가능성 (성향점수)**	원본 · 합성을 섞어 로지스틱 회귀 5겹으로 구분 : pMSE · PO50 · SPECKS	pMSE 0.0204 · PO50 17.6 · SPECKS 0.373	pMSE ≤ 0.0065 (원본 분할 반복 95분위)	미달	5장
유용성	**모형 성능 유사성 (분류)**	'계약상태' 예측 : 원본 학습 vs 합성 학습, 홀드아웃으로 평가	정확도 0.996→0.848 · 정밀도 0.985→0.640 · 재현율 0.991→0.673 · F1 0.988→0.649	F1 비율 목표 (관문 'AI 학습 성능')	미달	5장
유용성	**모형 성능 유사성 (회귀)**	'퇴거일' 선형 회귀 (설명 열 2개) : MSE 비교 · 회귀계수 95% 신뢰구간 겹침	MSE 2.678e+04→5.852e+04 (비율 2.19) · 신뢰구간 겹침 평균 0.000 (원본끼리 0.673)	겹침 0.5 이상 또는 원본끼리 겹침 - 0.1 이상, MSE 비율 1.2 이하	추가 검토	5장

개인정보위 「합성데이터 생성 · 활용 안내서」(2024.12) 부록 2 · 3 · 4 의 산식대로 계산했습니다. 구별 · 추론 위험도 · 2차원 · pMSE 임계는 원본 50:50 분할 반복의 95분위, 연결 위험도(CAP)는 절대 임계 0.7 입니다.

(?) 쉬운 해설

구분	안내서 지표	DQ SYNTH 측정 방법
안전성	**구별 위험도 (Single out)**	합성 3,000건 중 원본(같은 수 10,000건 임의 추출)에 같은 레코드가 있는 비율
안전성		약시작일, 계약종료일 → 민감정보 계약상태, 공급유형, 원본 레코드별 CAP
안전성		최근접 거리 dS < 그 원본의 다른 원본 최근접 거리 dO 인 비율 (수치 0~1 정 유클리드 거리)
안전성 (추가)		약시작일, 계약종료일, 입주일 조합이 5건 이상인 셀의 레코드 비율
안전성 (추가)		감속성 '계약상태' 값의 종류(l)와 전체 분포와의 차이(t)
안전성 (추가)		두 번째로 가까운 원본 거리. 실제 신규 기록 하위 5% 보다 작은 합성 비율
안전성 (추가)	**공격 시험 (Anonymeter 방식)**	학습 원본 400건을 공격 대상, 홀드아웃을 통제군으로 단독 식별 · 연결 · 추론 공격을 시도
유용성	**1차원 분포 유사성**	열마다 얀센-섀넌 발산(JSD) · KS 통계량(수치) · 카이제곱 검정(범주, 같은 크기 표본)

구별 위험도 (안내서)
합성 데이터 안에 원본과 똑같은 레코드가 있을 위험입니다.
어떻게 합성 레코드마다 원본에 같은 레코드가 있으면 1, 없으면 0으로 세어 평균을 냅니다.
보는 법 0에 가까울수록 안전하며, 원본을 반으로 나눠 반복 측정한 값의 95분위(임계값) 이하이면 충족입니다.

전문가가 아니어도 읽을 수 있게

지표마다 뜻 · 계산 방법 · 보는 법을 비전문가 눈높이로 안내합니다.

DELIVERABLES

20 심의와 반출에 필요한 자료가 함께 나옵니다

합성데이터만 주고 끝나지 않습니다. **어떻게 검증했는지**를 설명하는 문서와 재현 자료까지 한 번에 만듭니다.

REPORT

최종 결과보고서

- 요약 · 원본 분석 · 전처리 · 생성 방법
- 유용성 · 안전성 결과, 반복 개선 이력
- 활용 범위 · 안내서 대응
- Word · PDF (결과 화면과 같은 가로형)
- 기관 이름 · 로고 · 꼬리말 양식 지정

REVIEW

심의 · 검토 서류

- 안전성 자체 검토 결과서
- 원본데이터 명세서 · 개인정보 유형 분류표
- 활용 계획서 · 생성 기초자료 명세서
- 합성데이터 명세서
- 승인 전 문서에는 '검토용' 표시

DELIVERY

합성데이터 · 재현 자료

- CSV · 엑셀(테이블별 시트)
- JSON · SQL · Parquet
- 재현 패키지 ZIP
- 보고서 ID · SHA-256 해시로 위변조 확인
- 내려받기는 반출 대장에 자동 기록

검토부터 반출까지, 단계마다 사람이 확인합니다

결과 검토
판정 · 용도 · 담당자 의견

▶

승인
승인자가 결과를 확정

▶

최종 보고서 작성
완성되면 PDF 를 자동 저장

▶

반출 · 기록
누가 · 언제 · 무엇을

SCREEN · REPORT · AUDIT

21 보고서와 주요 작업 기록이 남습니다

결과보고서는 Word · PDF 로 저장하고, **누가 언제 무엇을 했는지**는 감사 기록으로 확인합니다.

최종 결과보고서

합성데이터 결과보고서

임대주택_임대현황_가상데이터_50000건 · 2026-11-09 · DQ SYNTH

1. 요약

작업명	임대주택_임대현황_가상데이터_50000건
원본 파일	임대주택_임대현황_가상데이터_50000건.csv (50,000행)
활용 목적 · 공개 수준	분석 · AI 학습 (제한 제공)
차분프라이버시	적용 안 함
최종 판정	안전성 충족 · 유용성 미달 (안전성(필수) 통과 · 유용성(활용 목적별 선택) 미달) - 유용성이 낮은 합성데이터는 잘못된 정보를 확대 · 재생산할 수 있으므로 활용 목적에 맞는지 확인 필요
합성 결과	임대주택_임대현황_가상데이터_50000건 49,500행 × 23열
테이블 관계	없음 (단일 테이블)
반복 개선	최대 8회 · 1,255.8초
생성 모델	GMM 코퓰라 (가우시안 혼합)
작성 · 승인	작성 admin · 승인 관리자 (admin)

보고서 ID	DQS-20261109-162144-8f7a
합성 파일 해시 (SHA-256 앞 16자리)	synthetic.csv c90363505fed7de9... · synthetic.json 93ba1468aca10361... · synthetic.parquet d9abe6bfa699d252... · synthetic.sql da9ecf6bbd5a97aa...
위변조 확인	내려받은 파일의 SHA-256 을 manifest.json · 이 표와 대조하면 같은 결과인지 확인할 수 있습니다

'임대주택_임대현황_가상데이터_50000건' 49,500행은 열별 분포 유사도 평균 0.985, 관계 평균 차이 0.028, '계약상태' 예측 성능 비율 0.66 로 원본의 통계 특성을 재현했습니다. 미달 관문 : 관계 (평균 차이), AI 학습 성능 (TSTR · '계약상태' 예측).

> 용어 · 지표의 쉬운 설명은 맨 뒤 '부록. 지표 해설'에 있습니다.

> 최종 판정은 처음부터 떼어 둔 홀드아웃 데이터(원본의 약 15%, 관계가 있으면 부모 단위로 나눔)로 한 번 한 결과입니다. 반복 개선 중의 점수(개선 평가셋)는 7장에 함께 적었습니다.

감사 기록



심의와 감사 자료로 씁니다

작업 생성 · 실행 · 승인 · 내려받기가 사용자 · 일시와 함께 남고 CSV 로 제출할 수 있습니다. 승인 후 만든 보고서에는 '검토용' 표시가 없습니다.

EASE OF USE

22 전문가가 아니어도 끝까지 갈 수 있게 만들었습니다

결정은 줄이고, 결과는 쉽게 설명합니다. 담당자가 **코딩 없이 웹 화면만으로** 작업을 마칩니다.

NO CODE

설치 없이 브라우저로

서버에 한 번 설치하면 담당자들은 웹으로 접속해 함께 씁니다.

PLAIN WORDS

주요 지표마다 (?) 쉬운 해설

어려운 지표 옆마다 뜻과 읽는 법을 비전문가 눈높이로 안내합니다.

MINIMUM DECISION

결정할 곳만 빨갛게

나머지는 자동입니다. 정하기 어려우면 [먼저 진행]으로 임시 결과부터 봅니다.

REQUEST FORM

확인 요청서 주고받기

현업에게 물을 항목을 CSV 로 내보내고, 회신을 올리면 결정이 채워집니다.

RECIPE

한 번 정하면 재사용

승인한 작업의 결정을 저장해, 같은 구조 파일에 자동으로 적용합니다.

LIVE STATUS

지금 어디까지 왔는지

단계 · 예상 남은 시간 · 작업 로그를 실시간으로 보고, 완료되면 알림을 받습니다.

RESULT VIEW

원본과 합성을 나란히

열별 분포 · 관계 히트맵 · 회차별 점수 추이를 그림으로 비교합니다.

WHY THIS MODEL

모델 선택 과정 공개

후보 점수, 쓰인 과정, 딥러닝 구조와 학습 곡선을 그림으로 보여 줍니다.

FOR EXPERTS

전문가 설정은 따로

평소에는 숨겨 두고, 필요할 때 목표 · 모델 · 예산을 세밀하게 정합니다.

BEYOND GENERATION

23 합성데이터 생성 너머까지 한 플랫폼에 담았습니다

개발 · 시험 · 인공지능 학습 현장에서 **함께 필요한 기능**을 별도 도구 없이 제공합니다.

TEST DATA

테스트 데이터 생성

원본 없이 열 정의나 DDL 만으로 시험용 데이터를 만듭니다. 36종 생성 규칙과 오류 주입(정답표)을 지원합니다.

TEXT

텍스트 합성

문장 속 개인정보를 가리고 새 문장을 만들어 JSONL 로 내보냅니다. 외부 LLM 으로 보내지 않습니다.

VALIDATE

외부 합성데이터 검증

다른 도구로 만든 합성데이터도 올려서 같은 기준으로 검증하고 보고서를 받습니다.

DOMAIN PACK

업종 팩 5종

회계 · ERP · 인사 · 제조 · 유통의 열 이름 지식을 더해 자동 판정에 참고합니다.

BALANCE

재균형 · 증강

희귀 범주를 원하는 배수로 늘려 AI 학습에 맞게 균형을 잡습니다. 판정은 증강 전 결과로 합니다.

COMPARE

대시보드 · 작업 비교

여러 작업의 결과를 나란히 비교하고, 작업을 복제해 조건만 바꿔 다시 돌립니다.

SECURITY · GOVERNANCE

24 원본은 안에 두고, 주요 과정은 기록합니다

합성데이터 작업은 원본을 다루는 일입니다. 디큐SYN 은 **기관 · 기업의 서버 안에서만** 동작하고, 누가 무엇을 했는지 기록합니다.

역할 권한	작업자	승인자	관리자
자기 작업 수행	●	●	●
모든 작업 보기	-	●	●
결과 승인	-	●	●
사용자 · 감사 기록 관리	-	-	●

직무 분리 설정

승인자가 자기 작업을 승인하지 못하도록 설정할 수 있어, 만든 사람과 확정하는 사람을 나눌 수 있습니다.

계정 보안

비밀번호 규칙, 연속 5회 실패 잠금, 90일 주기 변경, 세션 · CSRF 보호를 적용했습니다.

폐쇄망 설치

외부 연결 없이 설치 · 운영합니다.

감사 기록

로그인 · 생성 · 실행 · 승인 · 내려받기 · 삭제를 기록, CSV 제출

반출 대장

누가 · 언제 · 어떤 파일을 받았는지 판정 · 승인자와 함께 기록

내려받기 차단

안전성 미충족 또는 판정 전 합성데이터는 서버에서 차단

원본 보관 기한

기본 30일 뒤 자동 삭제, 즉시 파기 기능 제공

민감 열 가림

미리보기에서 가리고, 열람하면 감사 기록에 남김

EXPECTED EFFECTS

25 묶여 있던 데이터를, 근거를 갖춰 쓸 수 있게 됩니다

개인정보 부담은 낮추고, 데이터 활용의 범위와 속도는 높입니다.

데이터 활용 범위 확대

개인정보 때문에 쓰지 못하던 데이터를 안전성 검증과 기관 심의를 거쳐 분석 · AI 학습 · 개발 시험 · 외부 제공에 활용할 수 있습니다.

구축 시간과 비용 절감

분석 · 생성 · 검증 · 문서 작업을 한 흐름으로 묶어 전문 인력의 반복 작업을 줄입니다.

개인정보 위험 감소

개발 · 시험 환경에서 운영 데이터 대신 합성데이터를 써서 원본 노출 경로를 줄입니다.

일관된 품질 기준

담당자가 바뀌어도 같은 지표와 목표로 판정해 결과의 편차를 줄입니다.

심의 · 감사 대응 수월

지표 · 체크리스트 · 결과서 · 감사 기록이 자동으로 쌓여 설명 근거가 됩니다.

AI 학습 데이터 확보

희귀 범주 증강과 예측 성능 검증으로 학습에 쓸 수 있는 데이터를 준비합니다.

데이터 · 개발 담당자

운영 원본을 개발 · 시험 환경으로 옮기지 않고 필요한 데이터를 준비합니다.

개인정보 · 보안 책임자

정량 지표와 기록을 근거로 제공 여부를 검토합니다.

경영 · 사업 부서

보유 데이터로 새로운 분석과 AI 과제를 시도할 수 있습니다.

DEPLOYMENT · CONTACT

26 설치 프로그램 하나로 시작합니다

보유 데이터로 직접 돌려 보고, **결과보고서로 효과를 확인**해 보십시오.

도입 환경

- **설치** : Windows 서버용 설치 프로그램 하나로 실행 환경 · DB · 엔진을 함께 설치
- **운영** : Windows 서비스로 자동 시작, 업그레이드 때 설정 · DB · 작업 폴더 유지
- **접속** : 웹 브라우저, 여러 담당자가 함께 사용
- **연산** : GPU 없이 CPU 만으로 동작
- **입력** : 시스템에서 내려받은 파일 (CSV · TXT · 엑셀 · JSON · Parquet)
- **컨설팅** : 데이터 품질 · 표준화 컨설팅 경험을 바탕으로 활용 목적 정의부터 심의 서류까지 지원

시범 적용(PoC) 절차

1. **상담**
   활용 목적과 대상 데이터 확인
2. **설치**
   기관 · 기업 내부 서버에 설치
3. **생성 · 검증**
   보유 데이터로 목표에 맞춰 실행
4. **결과 검토**
   결과보고서로 품질과 안전성 확인
5. **도입 결정**
   적용 범위와 운영 방안 협의

시범 적용 기관을 모집합니다

㈜엠아이티소프트 · 도입 문의 · 데모 요청

TEL 02-719-0380

E-MAIL sales@mitsoft.kr

WEB www.mitsoft.kr

출시 준비 중인 제품으로 화면과 기능은 버전에 따라 달라질 수 있습니다. 표지 그림은 이해를 돕기 위한 예시이고, 화면 이미지의 데이터는 가상 데이터이며, 품질 결과는 데이터와 목표에 따라 달라집니다.