2022 PIDICON

개인정보 가명·익명처리 기술 경진대회

대회접수	2022. **9.13**(화) ~ **10.6**(목) 16:00	온라인	
예 선	2022. **10.14**(금)	온라인	
본 선	기술경연	2022. **11.8**(화)~ **11.9**(수)	온라인
	발표평가	2022. **11.12**(토)	오프라인

주최 과학기술정보통신부 주관 KISA 한국인터넷진흥원 운영기관 CM (주)컬처메이커스

2022 PIDICON

개인정보 가명·익명처리 기술 경진대회

CONTENTS

주최 과학기술정보통신부 주관 KISA 한국인터넷진흥원 운영기관 CM culture makers (주)컬처메이커스

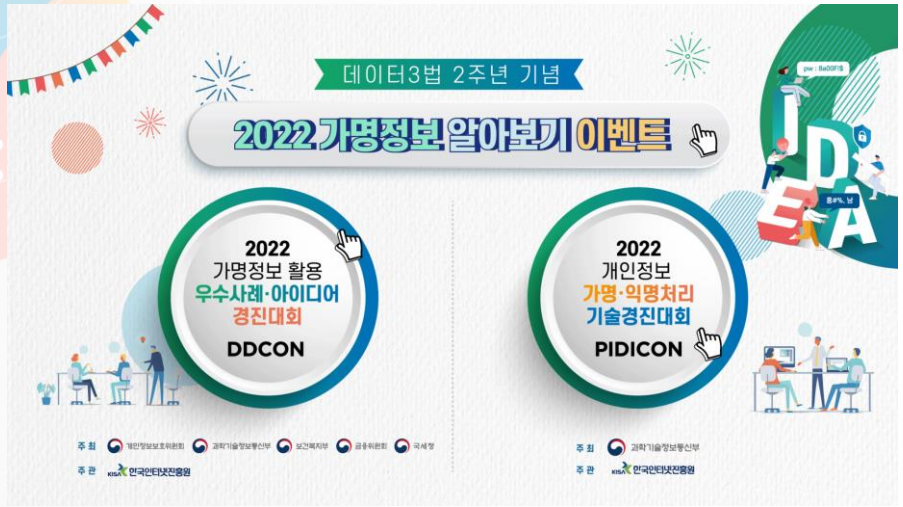
대회개요

대 회 명	2022 개인정보 가명·익명 기술경진 대회
주최/주관	과학기술정보통신부 / 한국인터넷진흥원
운영기관	㈜컬처메이커스
일정/장소	2022년 9월 13일(화) ~ 11월 16일(수)
목 적	안전한 데이터 활용을 위한 가명·익명처리 기술 발굴 및 저변확대
참가대상	대한민국 국민 누구나 참가 가능(학생부/일반부) *개인 또는 팀(4인 이하) 형식의 참여

주최 과학기술정보통신부 주관 KISA 한국인터넷진흥원 운영기관 CM culture makers (주)컬처메이커스

참가방법

접수방법

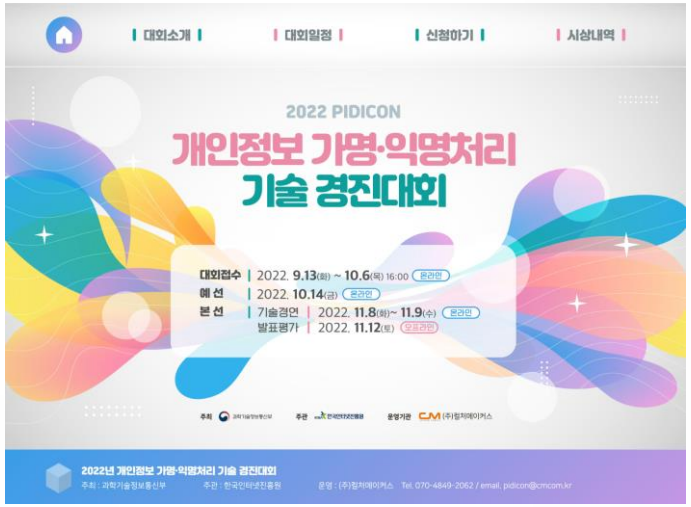


가명정보.kr

대회 홈페이지 접속

2022 개인정보 가명·익명처리 기술경진대회(PIDICON) 클릭

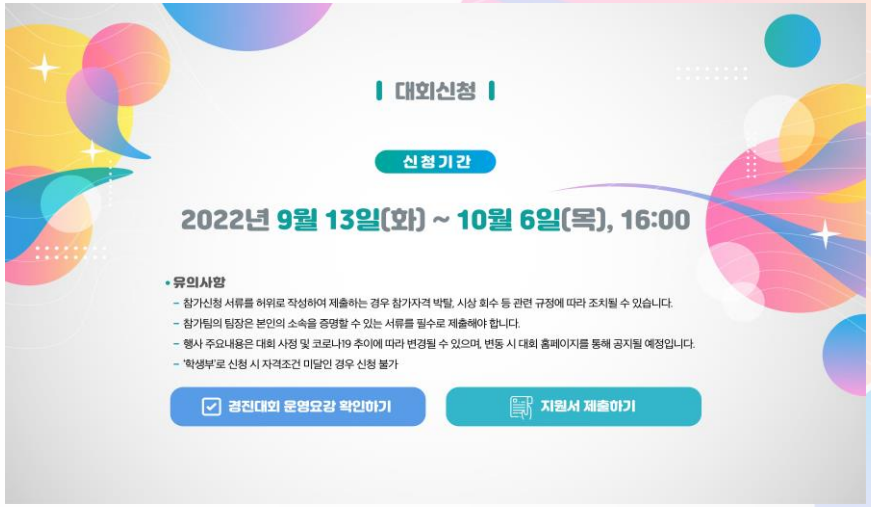
* 데이터3법 2주년 기념이벤트도 참가 해 보세요!



pidicon.kr

PIDICON 접속

대회소개, 대회일정, 신청하기, 시상내역 및
관련 일정을 확인



운영요강 확인 ▶ 지원서 제출하기

대회신청

경진대회 운영요강 확인, 참가자격을 증명할 수 있는
서류를 준비해서 지원서 제출하기

참가방법

운영절차

온라인 참가접수	예선 기술경연 (온라인)	본선 기술경연 (온라인)	본선 발표경연 (오프라인)	시상식
9.13(월)~10.6(목) 16:00까지	10.14(금) 10:00 ~ 22:00	11.8(화)~11.9(수)	11.12(토)	12.23(금) 예정

▷ 온라인 접수 : 대회 홈페이지를 통해 팀(개인)정보 등록, 참가자 서약 및 소속증빙서류 제출

▷ 예　　선 : 경연일정 및 시간에 맞춘 온라인 기술경연

▷ 본　　선 : 경연일정 및 시간에 맞춘 온라인 기술경연 및 오프라인 발표경연

※ 대회 동시 추진에 따른 부문별 중복참가 불가

주요일정

일정	내용	비고
9/15(목)	온라인 사전 설명회	경진대회 지표설명 및 사전교육
9/13(화)~10/6(목)	온라인 접수	16:00 까지 제출
10/14(금)	예선 기술경연(온라인)	10:00~22:00
11/8(화)~11/9(수)	본선 기술경연(온라인)	무박 해커톤 형식
11/12(토)	본선 발표경연(오프라인)	발표 10분 + 질의 및 평가 15분
11/16(수)	입상자 발표	대회 홈페이지 공고
12/23(금) 예정	시상식	-

주요일정

참가접수

▶ 일시 : ‘22. 9. 13.(화) ~ 10.5.(수) ~ 16:00 (기한내 접수건에 한해서만 인정)

▶ 참가대상 : 대한민국 국민 누구나 참여 가능하며 학생부, 일반부로 구성해서 참가

※ 일반부: 대한민국 국민 누구나 참가 가능

※ 학생부: 청소년, 대학(원)생 및 취업준비생 *‘21년 1월~‘22년 2월 수료·졸업)으로 구성된 팀

▶ 학생부는 참가접수 유의사항 : 학생 경우 재학증명서 또는 수료·졸업예정 증명서 제출,
취업준비생*의 경우 건강보험자격득실확인서 제출(미취업 여부 확인 용)

* 취업준비생 : ‘21년 1월~’22년 8월 졸업생의 한해서만 인정됨

▶ 일반부 접수 유의사항 : 팀장 소속을 확인할 수 있는 증명서류 제출(재직증명서 등)

참가자격에 대한 증빙서류 제출 必

주요일정

예선

제공된 재현 데이터셋의 원본정보를 분석하여 익명 데이터셋 생성을 통해
가장 안전하면서도 유용성이 높은 익명 처리 기술경연 (Only 정량지표)

▶ 일시 : ‘22. 10. 14.(금), 10시~22시 (기한내 접수건에 한해서만 인정)

▶ 참가대상 : 참가접수 된 정보를 바탕으로 결격사유가 없는 부문별 모든 접수팀

*개인 또는 팀(4인 이하) 형식의 참여

▶ 경연시작 시 참가자 신분확인 진행예정 *신분증 준비 필요

▶ 제공문서 및 데이터셋, 경연 진행방식 등을 안내하는 경연지침 자료는 예선 전날 팀구성원들에게 발송

▶ 경연시작과 함께 데이터셋 암호를 팀구성원들에게 SMS로 일괄 배포

본선참가팀 안내 : ’22. 10. 19.(수), 대회 홈페이지 공고

주요일정

본선

제공된 재현 데이터셋의 원본정보를 분석하여 가명 및 익명 데이터셋 생성을 통해
가장 안전하면서도 유용성이 높은 가명 및 익명 처리 기술경연 (정성 + 정량지표)

▶ 일시 : (기술경연) ‘22. 11. 8.(화) 10시 ~ 11. 9.(수) 17시 (발표경연) ‘22. 11. 12.(토)

▶ 참가대상 : 예선을 통과한 부문별 본선진출팀

▶ 경연시작 시 참가자 신분확인 진행예정 *신분증 준비 필요(발표경연에도 신분증 지참 필요)

▶ 제공문서는 본선 진출팀 발표와 함께 팀구성원들의 메일로 발송되며
데이터셋, 경연 진행방식 등을 안내하는 경연지침 자료는 본선 전날 팀구성원들에게 발송

▶ 경연시작과 함께 데이터셋 암호를 팀구성원들에게 SMS로 일괄 배포

수상팀 안내 : ’22. 11. 16.(수), 대회 홈페이지 공고

시상내역

과학기술정보통신부장관상 2점, 한국인터넷진흥원장상 10점 / 총 2,800만원 상당 상금시상

일반부 · 학생부	대상(2팀)	과학기술정보통신부 장관상	500만원	부문별 1팀
	최우수상(2팀)	한국인터넷진흥원 원장상	300만원	부문별 1팀
	우수상(2팀)	한국인터넷진흥원 원장상	200만원	부문별 2팀
	장려상(2팀)	한국인터넷진흥원 원장상	100만원	부문별 2팀

경연방식

"안전한 데이터 활용을 위한
가명·익명처리 기술 발굴"

예선

▶ 익명처리 기술경연(정량평가)

(도전주제) A유통사의 유통정보를 이용하여 아래와 같은 목적을 달성하기 위해 필요한 익명처리를 하고자 한다.

(분석목적) 예선 경연 당일 이용환경과 함께 공개

본선

▶ 가명·익명처리 기술경연(정량+정성평가)

(도전주제) A공공기관의 노동분야 관련 정보를 B기관에서 제공받아 아래와 같은 분석목적을 달성하기 위해 필요한 가명처리를 하고자 한다.

(분석목적) 본선 경연 당일 이용환경과 함께 공개

평가주안점

들어가며..

가명·익명 처리 기술 및 예시(2022 가명정보처리 가이드라인 참조)

분류	기술	세부기술	설명
개인정보 삭제	삭제기술	삭제(Suppression)	ㅇ 원본정보에서 개인정보를 단순 삭제
		부분삭제(Partial suppression)	ㅇ 개인정보 전체를 삭제하는 방식이 아니라 일부를 삭제
		행 항목 삭제(Record suppression)	ㅇ 다른 정보와 뚜렷하게 구별되는 행 항목을 삭제
		로컬 삭제(Local suppression)	ㅇ 특이정보를 해당 행 항목에서 삭제
		마스킹(Masking)	ㅇ 특정 항목의 일부 또는 전부를 공백 또는 문자(‘ * ’, ‘ _ ’ 등이나 전각 기호)로 대체
개인정보 일부 또는 전부 대체	통계도구	총계처리(Aggregation)	ㅇ 평균값, 최대값, 최소값, 최빈값, 중간값 등으로 처리
		부분총계(Micro aggregation)	ㅇ 정보집합물 내 하나 또는 그 이상의 행 항목에 해당하는 특정 열 항목을 총계처리. 즉, 다른 정보에 비하여 오차 범위가 큰 항목을 평균값 등으로 대체
	일반화 (범주화) 기술	일반 라운딩(Rounding)	ㅇ 올림, 내림, 반올림 등의 기준을 적용하여 집계 처리하는 방법으로, 일반적으로 세세한 정보보다는 전체 통계정보가 필요한 경우 많이 사용
		랜덤 라운딩(Random rounding)	ㅇ 수치 데이터를 임의의 수인 자리 수, 실제 수 기준으로 올림(round up) 또는 내림(round down)하는 기법
		제어 라운딩(Controlled rounding)	ㅇ 라운딩을 적용하는 경우 값의 변경에 따라 행이나 열의 합이 원본의 행이나 열의 합과 일치하지 않는 단점을 해결하기 위해 행이나 열이 맞지 않는 것을 제어하여 일치시키는 기법
		상하단코딩(Top and bottom coding)	ㅇ 정규분포의 특성을 가진 데이터에서 양쪽 끝에 치우친 정보는 적은 수의 분포를 가지게 되어 식별성을 가질 수 있음 ㅇ 이를 해결하기 위해 적은 수의 분포를 가진 양 끝단의 정보를 범주화 등의 기법을 적용하여 식별성을 낮추는 기법
		로컬 일반화(Local generalization)	ㅇ 전체 정보집합물 중 특정 열 항목(들)에서 특이한 값을 가지거나 분포상 의 특이성으로 인해 식별성이 높아지는 경우 해당 부분만 일반화를 적용하여 식별성을 낮추는 기법
		범위 방법(Data range)	ㅇ 수치 데이터를 임의의 수 기준의 범위(range)로 설정하는 기법으로, 해당 값의 범위 또는 구간(interval)으로 표현
		숫자데이터 범주화(Categorization of numeric data)	ㅇ 숫자로 저장된 정보에 대해 보다 상위의 개념으로 범주화하는 기법
		문자데이터 범주화(Categorization of character data)	ㅇ 문자로 저장된 정보에 대해 보다 상위의 개념으로 범주화하는 기법

평가주안점

들어가며..

가명·익명 처리 기술 및 예시(2022 가명정보처리 가이드라인 참조)

분류	기술	세부기술	설명
개인정보 일부 또는 전부 대체	암호화	양방향 암호화 (Two-way encryption)	ㅇ 암호화 및 복호화에 동일 비밀키로 암호화하는 대칭키(Symmetric key) 방식과 공개키와 개인키를 이용하는 비대칭키(Asymmetric key) 방식으로 구분
		일방향 암호화 - 암호학적 해시함수 (One-way encryption - Cryptographic hash function)	ㅇ 키가 없는 해시함수(MDC, Message Digest Code), 솔트(Salt)가 있는 해시함수, 키가 있는 해시함수(MAC, Message Authentication Code)로 구분 ㅇ 암호화(해시처리)된 값에 대한 복호화가 불가능하고, 동일한 해시 값과 매핑(mapping)되는 2개의 고유한 서로 다른 입력값을 찾는 것이 계산상 불가능하여 충돌 가능성이 매우 적음
		순서보존 암호화 (Order-preserving encryption)	ㅇ 원본정보의 순서와 암호값의 순서가 동일하게 유지되는 암호화 방식 ㅇ 암호화된 상태에서도 원본정보의 순서가 유지되어 값들 간의 크기에 대한 비교 분석이 필요한 경우 안전한 분석이 가능
		형태보존 암호화 (Format-preserving encryption)	ㅇ 원본 정보의 형태와 암호화된 값의 형태가 동일하게 유지되는 암호화 방식 ㅇ 원본 정보와 동일한 크기와 구성 형태를 가지기 때문에 일반적인 암호화가 가지고 있는 저장 공간의 스키마 변경 이슈가 없어 저장 공간의 비용 증가를 해결할 수 있음 ㅇ 암호화로 인해 발생하는 시스템의 수정이 거의 발생하지 않아 토큰화, 신용카드 번호의 암호화 등에서 기존 시스템의 변경 없이 암호화를 적용할 때 사용
		동형 암호화 (Homomorphic encryption)	ㅇ 암호화된 상태에서의 연산이 가능한 암호화 방법으로 원래의 값을 암호화한 상태로 연산 처리를 하여 다양한 분석에 이용가능
		다형성 암호화 (Polymorphic encryption)	ㅇ 각 도메인별로 서로 다른 가명정보를 처리할 수 있도록 정보 제공 시 서로 다른 방식의 암호화된 가명처리를 적용함에 따라 도메인별로 다른 가명정보를 가지게 됨
	무작위화 기술	잡음 추가 (noise addition)	ㅇ 개인정보에 임의의 숫자 등 잡음을 추가(더하기 또는 곱하기)하는 방법
		순열(치환) (Permutation)	ㅇ 기존 값을 유지하면서 개인이 식별되지 않도록 데이터를 재배열하는 방법 ㅇ 개인정보를 다른 행 항목의 정보와 무작위로 순서를 변경하여 전체정보에 대한 변경 없이 특정 정보가 해당 개인과 연결되지 않도록 하는 방법
		토큰화 (Tokenisation)	ㅇ 개인을 식별할 수 있는 정보를 토큰으로 변환 후 대체함으로써 개인정보를 직접 사용하여 발생하는 식별 위험을 제거하여 개인정보를 보호하는 기술 ㅇ 토큰 생성 시 적용하는 기술은 의사난수생성 기법이나 암호화 기법을 주로 사용
		(의사)난수생성기 ((P)RNG, (Pseudo) Random Number Generator)	ㅇ 주어진 입력값에 대해 예측이 불가능하고 패턴이 없는 값을 생성하는 메커니즘으로 임의의 숫자를 개인정보에 할당

평가주안점

들어가며..

가명·익명 처리 기술 및 예시(2022 가명정보처리 가이드라인 참조)

분류	기술	세부기술	설명
가명·익명처리를 위한 다양한 기술	기타기술	표본추출 (Sampling)	ㅇ 데이터 주체별로 전체 모집단이 아닌 표본에 대해 무작위 레코드 추출 등의 기법을 통해 모집단의 일부를 분석하여 전체에 대한 분석을 대신하는 기법
		해부화 (Anatomization)	ㅇ 기존 하나의 데이터셋(테이블)을 식별성이 있는 정보집합물과 식별성이 없는 정보집합물로 구성된 2개의 데이터셋으로 분리하는 기술
		재현데이터 (Synthetic data)	ㅇ 원본과 최대한 유사한 통계적 성질을 보이는 가상의 데이터를 생성하기 위해 개인정보의 특성을 분석하여 새로운 데이터를 생성하는 기법
		동형비밀분산 (Homomorphic secret sharing)	ㅇ 식별자 또는 기타 속성정보를 메시지 공유 알고리즘에 의해 생성된 두 개 이상의 쉐어(share)*로 대체*기밀사항을 재구성하는데 사용할 수 있는 하위 집합

평가주안점

들어가며..

경진대회 유용성 평가지표

(* U6~U8 지표의 경우 프라이버시 보호모델인 k-익명성 적용시 해당)

지표명	지표 설명	순위
U3 : CS (Cosine Similarity)	○코사인 유사도로 원본과 비식별 동일 속성집합 간 벡터의 스칼라곱과 크기	내림차순
U2 : MC (Mean Correlation)	○지정된 2개 이상의 특정 속성쌍들에 대한 피어슨 상관계수에 대한 차이 (평균절대오차)	내림차순
U3 : MGD_CA (Mean Generalized Difference_Character Attribute)	○카테고리형 트리 구조를 갖는 문자(혹은 이들의 코드로 표기된) 속성집합들 간의 일반화 정도 차이(들에 대한 평균으로 계산)	내림차순
U4 : NED_SSE (Normalized Euclidian Distance_Sum of Squared Errors)	○정규화된 유클리디안 거리(NED, Normalized Euclidian Distance)를 이용한 제곱합오차(SSE, Sum of Squared Errors)	오름차순
U5 : SED_SSE (Standardized Euclidian Distance_Sum of Squared Errors)	○표준화된 유클리디안 거리(SED, Standardized Euclidian Distance)를 이용한 제곱합오차(SSE, Sum of Squared Errors)	오름차순
U6* : MD_ECM (Mean Distribution Equivalence Class Metric)	○동질집합별 속성들에 대한 평균 분포도(분산)	오름차순
U7* : NA_ECSM (Normalized Average Equivalence Class Size Metric)	○정규화된 동질집합들의 평균 크기	오름차순
U8* : NUEM (Non-uniform Entropy Metric)	○비균일 엔트로피 방법을 이용한 k-익명성 프라이버시 보호 모델에서의 정보손실 측도	오름차순
U9 : AR (Anonymisation Ratio)	○원본 데이터셋 대비 익명처리된 데이터셋의 정보량	내림차순

평가주안점

경진대회 평가기준

구분	지표		평가기준(주관적 정성, 객관적 정량평가)	배점	점수	종합점수	비고	평가 기초자료
예선	익명처리	안전성	익명정보의 안전성 계량화	100	50%	100%	정량	익명정보 익명처리 결과보고서
		유용성	익명정보의 유용성 계량화(U1~U9 9개 지표 반영)	100	50%			익명정보
본선	가명처리	안전성	가명처리 시 가명정보 자체만으로 특정 개인을 알아볼 수 있는지 여부 등 평가	O/X	적격 여부 판정 (심사위원 정성)			가명정보, 식별위험성 검토결과보고서, 항목별 가명처리 계획표
		유용성	목적달성가능성 계량화 (가명화율 포함, T1~T3 3개 지표 반영)	100	25%	25%	정량	가명정보
	익명처리	안전성	시간 · 비용 · 기술 등을 합리적으로 고려할 때 다른 정보를 사용하여도 더 이상 개인을 알아볼 수 없는지 여부 등 평가	O/X	적격 여부 판정 (심사위원 정성)			익명정보 항목별 익명처리 계획표
		유용성	익명정보의 유용성 계량화(U1~U9 9개 지표 반영)	100	25%	25%	정량	익명정보
		발표	가명·익명 조치 기법 선정 이유 및 타당성	20	10%	50%	정성	발표자료
			속성별 가명·익명 조치 적용 수준 및 방법의 우수성	50	25%			
			가명·익명 조치 적용 기술의 특징 및 장점, 차별성	30	15%			

평가주안점

유통데이터

[대회 예선] 익명처리

운영 절차

문제

A

원본 정보

참가자 제출물

A’

익명 정보 + 익명처리 결과보고서

※ A’에는 원본 행(일련)번호를 포함

평가 절차

1	안전성(정량적 평가)	50%
+		
2	유용성(정량적 평가)	50%
=		
3	최종 점수 산출	100%

※ 안전성이 일정기준 이하인 경우 탈락될 수 있음

※ A는 완전 재현데이터이며, 제출물은 총 2종(익명정보 A’, 익명처리결과보고서)임

평가주안점

대회예선

▶ 익명처리 기술경연

(제공문서) 대회안내서(예선용), 익명처리 결과 보고서 양식

(제공데이터셋) 원본 데이터셋 1종(원본 행 일련번호 포함)

(제출문서) 익명처리된 데이터셋 1종(원본 행 일련번호 포함), 익명처리 결과 보고서

평가주안점

공공분야(노동)데이터

[대회 본선] 가명·익명처리

운영 절차

문제

A

원본 정보

참가자 제출물

A'

가명 정보 + 식별위험성 검토결과보고서 + 항목별 가명 처리계획표 +

A''

익명 정보 + 항목별 익명 처리 계획표 +

발표 자료

※ A', A''에는 원본 행(일련)번호를 포함

평가 절차

1	가명·익명정보 안전성(정성적 평가)	O/X
	+ ※ 안전성이 일정기준 이하인 경우 탈락될 수 있음	
2	가명정보 유용성(T1~T3, 정량적 평가)	25%
	+ ※ 특정 항목들에 대해 목적달성가능성으로 측정	
3	익명정보 유용성(U1~U9, 정량적 평가)	25%
	+ ※ 특정 항목들에 대해 유용성 지표를 이용, 측정	
4	발표(타당성, 우수성, 차별성 등 정성적 평가)	50%
	= ※ 가명처리 및 익명처리 모두를 대상으로 함	
5	최종 점수 산출	100%

※ A는 완전재현데이터이며, 제출물은 총 6종(가명정보 A', 식별 위험성 검토결과보고서, 항목별 가명처리 계획표, 익명정보 A'', 항목별 익명처리 계획표, 발표자료(가명 및 익명 포함))임

※ 식별 위험성 검토 결과보고서, 항목별 가명(익명)처리 계획표는 2022년 4월 개인정보보호위원회에서 발행한 '가명정보 처리 가이드라인 개정본' 참조

평가주안점

대회본선

▶ 가명·익명처리 기술경연

(제공문서) 대회 안내서(본선용), 가명처리 식별 위험성 검토 결과보고서 양식,
항목별 가명처리 계획표 양식, 항목별 익명처리 계획표 양식, 발표자료 양식

(제공데이터셋) 원본 데이터셋 1종(원본 행 일련번호 포함)

(제출문서) (가명) 가명처리 식별 위험성 결과보고서, 항목별 가명처리 계획표,
가명처리된 데이터셋 1종(원본 행 일련번호 포함),
(익명) 항목별 익명처리 계획표, 익명처리된 데이터셋 1종(원본 행 일련번호 포함)
(가명+익명 통합) 발표자료 1종

2022 PIDICON

개인정보 가명·익명처리 기술 경진대회

▣ https://가명정보.kr ▣ E-Mail. pidicon@cmcom.kr ▣ Tel. 070-4849-2062

주최 과학기술정보통신부 주관 KISA 한국인터넷진흥원 운영기관 CM culture makers (주)컬처메이커스