개인정보 비식별 기술 경진대회 본선 개최 계획(안)

< 개인정보비식별지원센터 '18.11.27 >

□ 개 요

o **일시** : '18. 11. 29(목) - 11. 30(금) 09:00~18:00

o **장소** : 서울 더 화이트베일 (서울시 서초구 소재, 3호선 남부터미널역 4번출구)
- 1일차 : 2층 에메랄드홀 / 2일차 : 2층 VIP홀

o **주최·주관** : 한국인터넷진흥원

o **후원** : 과학기술정보통신부, 행정안전부, 방송통신위원회

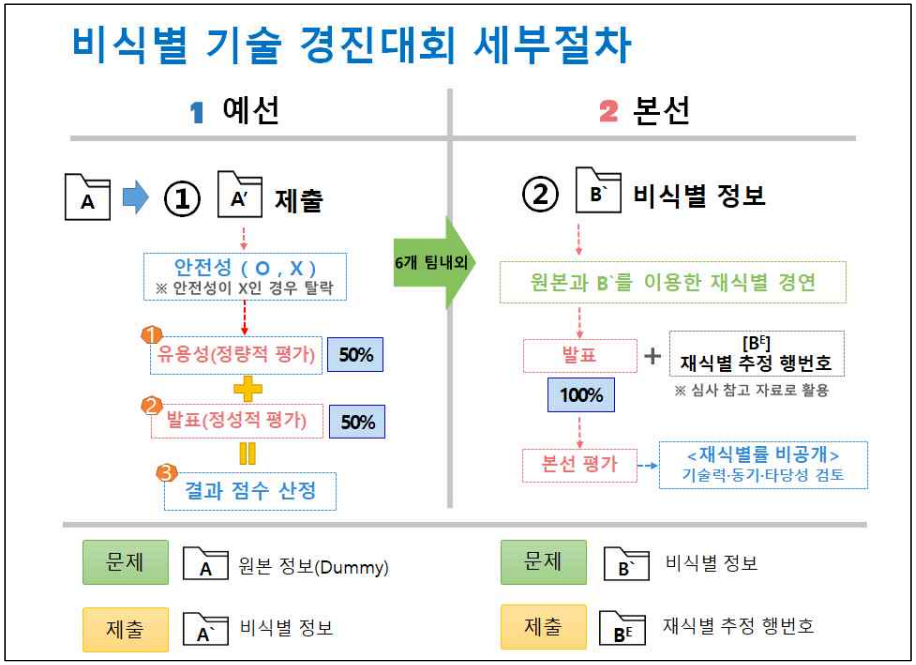
□ 프로그램(안)

일 자	시 간	주 요 내 용	비 고
1일차 11.29(목)	09:00 ~ 09:30	▪ **참가팀 참석 확인**	
	09:30 ~ 09:35	▪ **개회 및 행사안내**	
	09:35 ~ 09:40	▪ **환영사** – 정현철 개인정보보호본부장	
	09:40 ~ 09:50	▪ **경진대회 진행 및 운영규칙 안내** - 데이터셋 USB 배포 등	
	09:50 ~ 12:00	▪ **경진대회 본선 진행**	
	12:00 ~ 13:00	▪ **점심시간** ※ 2일차 팀별 발표 순서 제비뽑기	**도시락 제공**
	13:00 ~ 17:00	▪ **경진대회 본선 진행 (계속)**	**본선 경연 종료**
	17:00 ~ 18:00	▪ **재식별정보 및 발표자료 제출**	**발표자료 제출**
2일차 11.30(금)	09:00 ~ 09:30	▪ **참가팀 참석 확인 및 발표준비**	
	09:30 ~ 09:55	▪ 1번팀 발표 및 질의응답	
	09:55 ~ 10:20	▪ 2번팀 발표 및 질의응답	
	10:20 ~ 10:45	▪ 3번팀 발표 및 질의응답	
	10:45 ~ 11:10	▪ 4번팀 발표 및 질의응답	
	11:10 ~ 11:35	▪ 5번팀 발표 및 질의응답	
	11:35 ~ 12:00	▪ 6번팀 발표 및 질의응답	
	12:00 ~ 12:50	▪ **평가위원 평가 진행**	**13시 식사장소 이동**

※ 2일차 평가시 팀별 15분 발표, 10분 질의응답

붙임2 개인정보 비식별 기술 경진대회 규칙 (Ver. 1.6)

□ 경기규칙 ver. 1.6. (`18.11.27)



o 본 경진대회에는 다음의 종류가 있다.
(1) 비식별 경진대회 예선(비식별 조치 경연)
(2) 비식별 경진대회 본선(재식별 경연)

o 사용하는 소프트웨어나 OS에는 제한이 없으며, 참가자는 자신의 실험 환경을 경진대회에 장소에 반입할 수 있다. 네트워크 연결은 가능하나 외부로의 정보유출은 금지한다.

o 경진대회 발표자료 작성을 위한 편집프로그램은 참석자가 개별적으로 준비하도록 한다.

o 다음의 규칙은 잠정적인 것으로 예선/본선은 본 규칙이 변경될 수 있으며 (1), (2)의 규칙은 다음과 같다.

1. **(참여자 분류)** 비식별 처리자, 재식별자, 평가자 등 3분류로 정한다. 이때 비식별 처리자와 재식별자는 동일하며 개인 또는 최대 5명 이하의 팀으로 구성할 수 있다.

2. **(비식별 처리자)** 비식별 처리자는 다음과 같은 경연을 수행한다.
① 비식별 처리자는 원본 정보 A를 활용하여 비식별 정보 A`를 생성하고, 생성한 비식별 정보 A`와 정성평가를 위한 발표 자료를 주최 측에 제출한다.

* 원본정보 A는 KISA에서 무작위로 생성한 분포도 있는 가상의 정보

② 비식별 처리자는 비식별 경연 다음날 비식별 조치 수행 전반에 대한 발표*를 수행한다.

* 발표에 포함되는 내용은 9. (평가 지표) 참조

3. **(재식별자)** 재식별자는 다음과 같은 경연을 수행한다.

① 재식별자는 예선에서 생성한 상대팀의 비식별 정보(또는 주최 측에서 제공하는 비식별 정보) B`와 원본 정보를 참조하여 재식별을 수행하며, 재식별 추정 행 번호와 발표 자료를 주최 측에 제출한다.

※ 재식별 시 제공되는 문제는 예선전 종료 후 평가자들의 검토 의견을 수렴하여 결정한다.

*** 재식별 시 제공되는 문제는 주최측에서 생성한 안전하고 유용성 있는 비식별 정보. 다만, 문제에는 재식별자가 가상의 원본정보를 추정 가능한 항목을 임의로 추가**

② 재식별자는 재식별 경연 다음날 재식별 수행 전반에 대한 발표*를 수행한다.

* 발표에 포함되는 내용은 9. (평가 지표) 참조

4. **(예선 및 본선 절차)** 예선과 본선 절차는 다음과 같다.

4-1. 예선 문제 데이터 셋(원본정보)을 팀별 USB를 통해 제공한다.(시나리오는 별도 제공)

4-2. 신청자들은 복사한 데이터 셋에 대한 무결성을 확인(프로그램 별도 배포)한 후 대회를 진행하며, 비식별 조치 후 정보 생성이 완료된 팀은 생성한 **①비식별 정보와 ②발표 자료**를 주최 측에서 제공한 USB에 복사 후 제출한다.

※ 비식별 시 범주화 표기의 정확한 표현을 위해 "(" 개구간(<), "[" 폐구간(<=)
(예 : [100,1000), [1000,10000))으로 통일하고 비식별시 CSV 파일로 생성하여 제출

4-3. 본선의 경우 예선에서 생성한 상대팀의 비식별 정보(또는 주최 측에서 마련한 비식별 조치된 정보)를 제공받아 재식별 경연을 실시한다.

4-4. 재식별 수행 후 **①결과(재식별 추정 원본 행번호)와 ②발표자료**는 주최측에서 제공한 USB에 복사 후 제출하여 대회를 종료한다.

4-5. (대회시간) 예선과 본선은 양일간 진행하기로 한다.

- 예선 : 첫날 10시부터 17시까지 비식별 경연 후, 18시 이전에 각 팀별로 생성한 비식별 정보와 발표 자료 제출(발표는 둘째날 10시부터 팀당 20분(10분 발표, 10분 질의응답)씩 발표 진행
- 본선 : 첫날 10시부터 17시까지 재식별 경연 후, 18시 이전에 각 팀별로 재식별을 추정한 행번호와 발표 자료를 제출(발표는 둘째날 10시부터 팀당 25분(15분 발표, 10분 질의응답)씩 발표 진행

※ 대회 시간 내 자료 미제출 팀은 실격으로 처리하며, 발표진행 순서에 대한 안내는 대회 첫날 공지 예정

※ 발표 질의응답 시 본선 대회에 다른 참가자의 비식별 정보를 사용한 경우, 해당 비식별 정보를 생성한 팀은 재식별 추정에 대한 의견(반론)을 제시할 수 있다.

5. **(본선 진출자 선정)** 비식별 정보에 대한 안전성을 유지하고, 가장 유용성이 높은 비식별 데이터를 제출한 처리자를 본선 진출자로 한다. 본선 진출자는 비식별 경연 시 생성한 비식별 정보를 대상으로 정량지표인 유용성 점수 50%와 평가자 최소 5인 이상의 외부 전문가 대상의 발표 점수(목적, 기술력, 타당성, 특장점 등) 50%를 종합하여 6개 팀 내외로 결정한다. 다만, **안전성에서 부적격을 받은 팀은 탈락** 처리한다.

6. **(최종 우승자 선정)** 비식별 경연 결과점수(70%)와 재식별 경연 결과 점수(30%)를 합산하여 최종 우승자를 선정한다.

7. **(유용성의 정의)** 비식별 데이터 A의 유용성은 유용성 지표 U1~U3 의 합계 순위로 정하며, 각 지표별로 원본 데이터셋 내에 정수값을 임의의 문자열로 변환하는 것은 금지한다.

비식별 기술 경진대회 계량지표(안)

지표명	지표 설명	기준
U1 : MA (Mean Attribute)	☑ 원본정보(X)와 비식별 정보(Y)간 특정(지정) 속성값에 대한 평균값 비교를 통해 비식별 정보의 유효성을 검증 ☑ 특정 속성 지정은 한 개 혹은 복수개로 지정 가능	오름차순
U2 : MC (Mean Correlation)	☑ 속성자 내에 모든 속성 쌍(xi, yj)들의 피어슨의 상관계수에 대한 x와 y간 MAE(평균절대오차) ☑ 피어슨의 상관계수 : 상관계수의 한 형태, 변인 X와 변인 Y 간의 선형 관계성의 정도를 0에서 1.00, 혹은 0에서 -1.00의 척도 상에서 기술해 주는 통계치 ☑ X와 Y간에 값의 상관관계를 통해 X와 Y간에 어느정도 차이가 있는지를 검증	
U3 : MD (Mean Distribution)	☑ 원본데이터와 비식별 정보와의 분포도 차이	

8. **(안전성 기준 및 평가)** 안전성 기준 및 정성평가 가이드는 다음과 같다.

8-1. 안전성 기준*

- 재식별 공격자로부터 아래의 위험요소를 감소시켜야 한다.

(1) 식별가능성(Single out) : 데이터 주체를 고유하게 식별하기 위해 데이터 셋의 속성 집합을 관찰하여 해당 데이터 주체에 속한 레코드를 결리(isolation)하는 행위

(2) 연결가능성(Linkability) : 동일한 데이터 주체 혹은 데이터 주체 그룹과 관련된 레코드를 별도의 데이터 셋과 연결하는 행위

(3) 추론가능성(Inferenceability) : 무시할 수 없는 확률로 다른 속성 집합의 값에서 해당 속성의 값을 추론하는 행위

* 안전성 정의에 관한 자세한 사항은 ISO FDIS 20889 참조

8-2. 안전성 정성평가 가이드(평가위원 참고자료 제공)

- 각 평가자들은 정성평가 시 아래 사항들을 참고하여 종합적으로 판단, 최종 점수를 산정한다.

(1) 준식별자에 대한 개인식별(Single-Out) 여부 평가

(2) 한 개 이상의 레코드나 혹은 속성들을 대상으로 비식별 처리한 기법이나 기술 평가(원본과 비식별 정보의 분포표 제공)

(3) k-익명성에 대한 검증(동질집합의 최대/최소 개수)

(4) 기타 안전성 정량지표(참고용)

9. **(평가 지표)** 예선 및 본선 평가 지표는 다음과 같다.

(1) 발표 점수 및 감점, 실격 내용은 외부에 공개하지 않으며, 본선 순위 결정 및 종합점수 산정 시에만 활용

(2) 예선 참가자들은 비식별 조치를 수행하여 생성한 비식별 정보와 발표자료를 반드시 제출하여야 한다.

(3) 발표 자료에는 다음과 같은 사항이 포함되어야 하며, 해당 사항이 누락되었을 경우 감점될 수 있다.

- **① 비식별 조치 기법 선정 이유, ② 각 속성별 비식별 조치 적용 수준 및 방법, ③ 비식별 조치 적용 기술의 특장점**

9-2. 본선은 발표 평가(정성, 100%)로 하며, 평가지표는 아래와 같다.

(1) 재식별 시도에 대한 추정 행번호와 발표자료를 제출하여야 한다.

(2) 발표 자료에는 다음과 같은 사항이 포함되어야 하며, 해당 사항이 누락되었을 경우 감점될 수 있다.

- **① 재식별 조치 기법 선정 이유, ② 각 속성별 재식별 시도 방법, ③ 비식별 정보의 재식별 위험성, ④ 재식별 방지를 위한 제언 및 특장점**

(3) 발표 질의응답 시 본선 대회에 다른 참가자의 비식별 정보를 사용한 경우 해당 비식별 정보를 생성한 팀으로부터 재식별 추정에 대한 의견을 청취할 수 있다.

10. **(종합 평가)** 예선 점수(70%)와 본선 점수(30%)를 합산하여 종합 평가를 실시한다.

11. **(공통 금지사항) 비식별 경진대회에 참석하는 모든 신청자들은 다음의 행위를 금지한다.**
    (1) 비식별 처리자 및 재식별자의 일체 부정행위
    (2) USB 등 보조기억매체, 반입 장비, 네트워크를 활용한 SNS 및 메신저, 메일 등 대회정보 유출 및 외부 도움을 시도하는 행위
    (3) 본 경진대회에 활용한 원본데이터 및 비식별 조치 정보에 대한 외부 유출 행위 및 활용

12. **(재식별자의 금지 사항)** 재식별자는 다음의 행위를 금지한다.
    (1) 다른 팀 비식별 처리자와의 결탁(행 번호 데이터 등을 가르쳐 달라고 하는)하는 행위
    (2) 본 경진대회에 활용한 원본데이터, 비식별 데이터 및 재식별 조치 정보에 대한 외부 유출 행위 및 활용

13. **(평가자의 금지 사항)** 평가자는 다음의 행위를 금지한다.
    (1) 비식별 처리자 또는 재식별자들과의 결탁(평가자의 권한에 의해 알게된 정보(행 번호 데이터 등) 알려주는)하는 행위
    (2) 비식별 처리자나 재식별자에게 경진대회에서 유리한 정보를 비공개로 제공하는 행위

14. **(대회 데이터셋 정의)**
    (1) 분포도 있는 더미데이터에 금융데이터를 조합하여 약 100만개의 데이터를 생성
    (2) 데이터셋 샘플 : 16개 컬럼의 약 100만개 데이터
    - 항목 : 일련번호, 이름, 성별, 생년월일, 우편번호, 주소, 나이, 통신사, 신용등급, 핸드폰번호, 월카드 이용금액, 카드한도, 총 대출잔액, 연체금액, 직업군, 월소득
    * 대회시 항목별 배치 순서는 바뀔 수 있음

일련번호	이름	성별	생년월일	우편번호	주소	나이	통신사	신용등급
13574922	이옥순	F	1963-03-19	20336	경기도 고양시 풍동	39	SKT	2
13575087	유만규	M	1964-12-31	36526	전라남도 순천시 승주읍	72	SKT	3
13575361	안미애	F	1995-10-03	23678	경기도 안성시 공도읍	50	LGT	4
13575715	김은미	F	1959-03-21	48386	경기도 고양시 성석동	49	KT	2
13577109	조원제	M	1978-05-31	20670	경기도 광주시 쌍령동	50	KT	7

핸드폰 번호	월카드 이용금액	카드 한도	총 대출잔액	연체금액	직업군	월소득
010-4534-4589	6,880,909	26,000,000	1,404,046	1,320,000	1	2,178,508
010-664-4578	708,707	7,500,000	1,041,667	4,512,000	9	5,335,354
010-123-7894	1,410,296	15,000,000	8,042,500	0	9	1,846,178
010-9658-7894	98,921	10,000,000	8,043,333	0	9	1,723,664
010-7544-4446	6,465,163	6,200,000	3,044,167	0	3	1,572,875

(3) 재식별 경연은 주최측에서 생성한 안전하고 유용성 있는 비식별 정보를 문제로 활용하며, 문제에는 재식별자가 가상의 원본정보를 추정 가능한 항목을 임의로 생성하였다.