

개인정보 가명·익명 처리 (실습)

김 승 환

인하대학교 데이터 사이언스학과

swkim4610@inha.ac.kr

이 노트는 라곰 소프트 오형섭 이사 강의자료를 편집
하였음을 밝혀 둡니다.

PART 1. 개인정보 가명·익명처리 방법론

PART 2. 개인정보 가명처리 실습

PART 1.

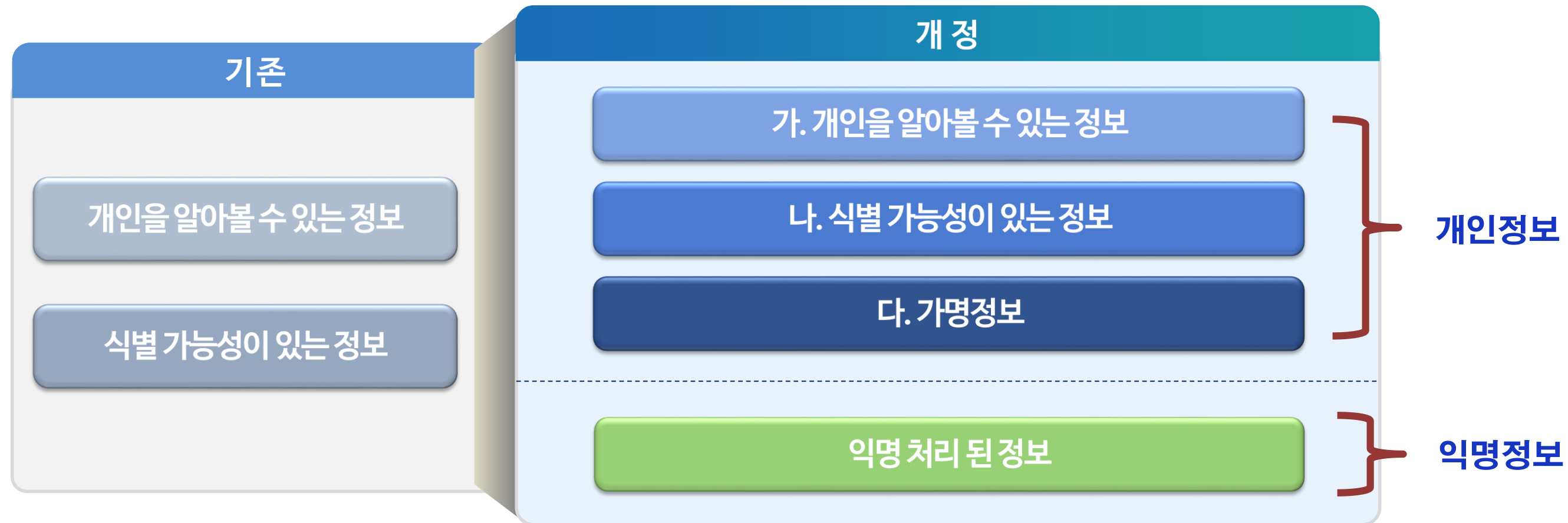
개인정보 가명/익명처리 방법론



1. 개인정보 가명·익명처리 방법론

■ 개인정보의 범위

가명정보를 도입하여, 개인정보를 개인을 알아볼 수 있는 정보, 식별 가능성이 있는 정보, 가명정보로 구분



법 제58조의2(적용제외) 이 법은 시간·비용·기술 등을 합리적으로 고려할 때 다른 정보를 사용하여도 더 이상 개인을 알아볼 수 없는 정보에는 적용하지 아니한다. [본조신설 2020. 2. 4.]

1. 개인정보 가명·익명처리 방법론

■ 개인정보 속성 분류

가명화 기준 !! 어떻게 정할 것인가?

| 개인정보 속성 분류 | 식별 가능성 | 프라이버시 침해 가능성 | 가명처리 |
|------------------------------------|---|--------------|--------|
| 식별자 (Unique Identifier) | 매우 높음 | 일부 있음 | 필수 |
| 준식별자 (Quasi Identifier) | Unique population issue 높음 특이정보 | 일부 있음 | 익명화 처리 |
| 민감정보 (Sensitive Attribute) | 거의 없음 (단, 특이치의 경우 높음) | 매우 높음 | 특이치 처리 |
| 일반정보 (Non Sensitive Attribute) | 거의 없음 (단, 특이치의 경우 높음) | 거의 없음 | 특이치 처리 |

➤ 준식별자, 민감정보 및 일반정보는 데이터 셋의 구성, 특성 및 가공/활용 목적 등에 따라 상대적인 성질로 변경될 수 있음

예) 암 환자 대상의 임상데이터 연구분석을 위한 처리에서 ‘암’ 자체는 민감정보로 구분하지 않을 수 있음

※ [참고자료] 금융보안원, 전자금융과 금융보안 제20호

1. 개인정보 가명·익명처리 방법론

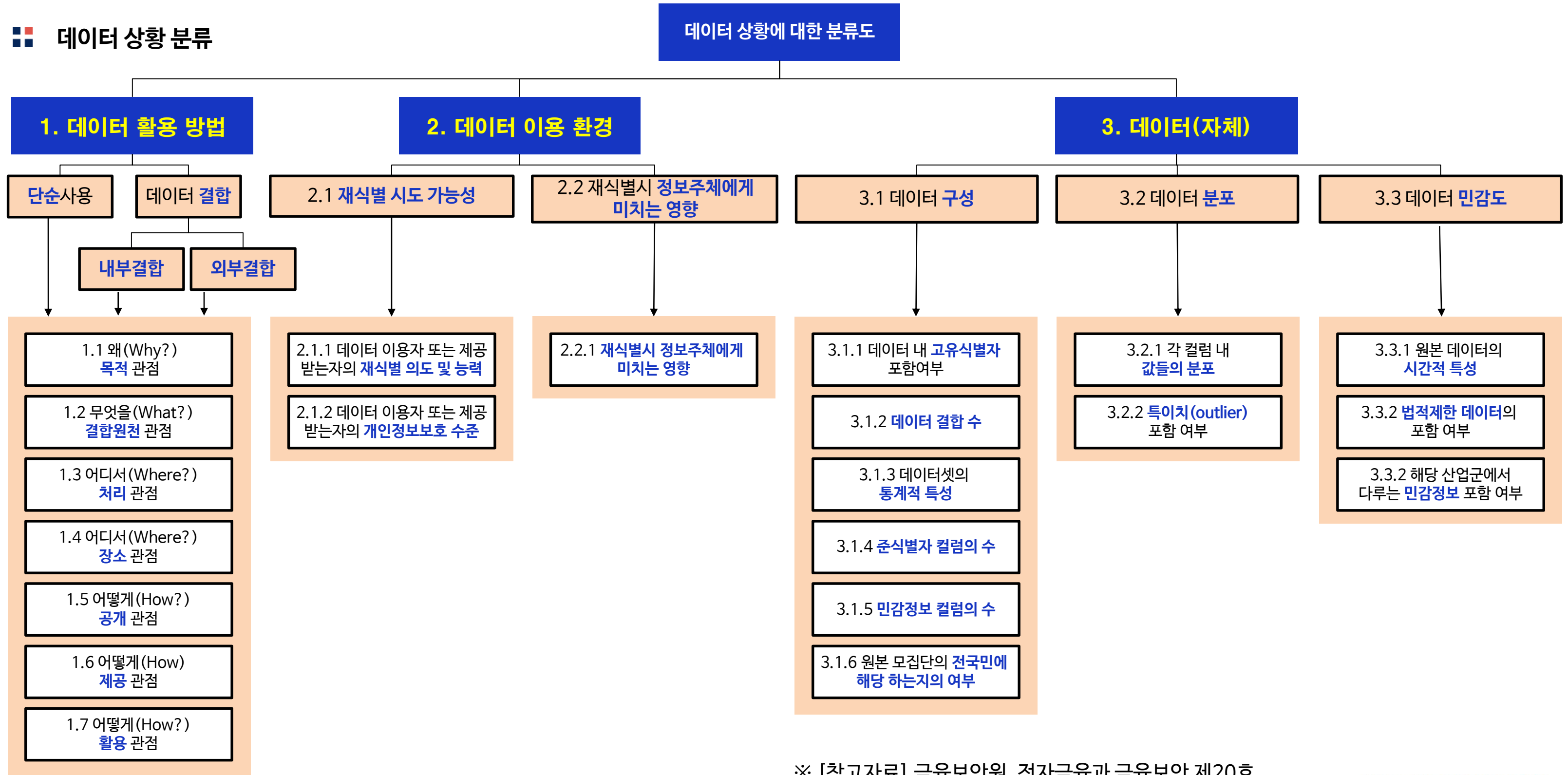
■ 개인정보 : 식별자 및 준식별자

세이프 하버 방식 : 18가지 식별자 및 준식별자 항목

- | | |
|--|---------|
| 1. 이름 | |
| 2. 주(state)보다작은지리적 단위(시, 군, 구역), 우편번호 및 이와 상응하는 지역코드 | |
| 3. 개인과 직접적으로 연관되는 모든 날짜로 생일, 입원일, 퇴원일, 사망일 등 | |
| 4. 전화번호 | |
| 5. 팩스번호 | 주민등록번호 |
| 6. 이메일 주소 | 운전면허번호 |
| 7. 사회보장번호 | 여권번호 |
| 8. 의무기록번호 / 환자번호 | 외국인번호 |
| 9. 건강보험 등록번호 | 각종 회원번호 |
| 10. 계좌번호 | 각종 아이디 |
| 11. 수료증 혹은 자격증 번호 | 군번 |
| 12. 자동차 번호판을 포함한 차량번호 | 자격증번호 |
| 13. 기기 아이디 및 시리얼 넘버 | 사원번호 등 |
| 14. 웹페이지 주소(URL) | |
| 15. 인터넷 IP 주소 | |
| 16. 신체특성 지표자, 지문 및 음성 포함 | |
| 17. 얼굴 정면 사진 및 이에 상응하는 사진 | |
| 18. 기타 모든 고유한 숫자, 문자, 코드 | |

1. 개인정보 가명·익명처리 방법론

데이터 상황 분류



※ [참고자료] 금융보안원, 전자금융과 금융보안 제20호

1. 개인정보 가명·익명처리 방법론

■ 개인정보 가명·익명 처리 절차

| 구분 | 단계 | 내용 |
|--------------|-----------------|---|
| 현황분석 | 1. 데이터 특성 분석 | <ul style="list-style-type: none"> 데이터의 가명·익명처리 필요 여부, 데이터의 민감도 등을 판단하기 위해 데이터의 소스, 제공 스키마, 각 컬럼별 특성 등을 분석 |
| | 2. 데이터 활용 목적 분석 | <ul style="list-style-type: none"> 1단계의 데이터 민감도와 함께 가명·익명처리의 강도를 산정하기 위해 데이터 사용 목적, 데이터 사용처, 사용 방법, 기술적 제공 방법 등을 분석 |
| | 3. 개인정보 속성 분석 | <ul style="list-style-type: none"> 2단계의 데이터 사용 목적을 감안하여 각 컬럼에 대해 필요 여부와 개인정보 속성을 선정 (식별자, 준식별자, 민감정보, 일반속성) |
| 가명처리 방안검토 | 4. 데이터 전처리 | <ul style="list-style-type: none"> 데이터의 정제 작업 및 가명·익명처리에 적합하도록 변경하는 작업 진행, k-익명성 적용 시 하나의 테이블로 만드는 작업 필요 |
| | 5. 데이터 분포 검토 | <ul style="list-style-type: none"> 각 컬럼에 대한 특이치 처리 등 가명처리 방안 수립을 위한 컬럼별 데이터 분포 검토 |
| | 6. 가명처리 방안 수립 | <ul style="list-style-type: none"> 1~5단계의 분석 결과를 바탕으로 데이터 사용 목적에 적합한 가명·익명처리 방안을 수립, 동일 컬럼도 사용 목적에 따라 다른 기법의 적용이 필요하며 데이터 사용처와 지속적으로 협의 필요 |
| 방안수립 | 7. 가명처리 방안 적용 | <ul style="list-style-type: none"> 6 단계에서 수립된 가명·익명처리 방안에 의해 각 컬럼별 가명·익명처리 기술 적용, 기술 적용 시 여러가지 문제가 발생할 수 있으며 문제 발생 시 데이터 사용처와 협의하여 적절하게 대처 |
| | 8. 가명처리 결과 분석 | <ul style="list-style-type: none"> 모든 가명·익명처리 기법이 적용되면 프라이버시 보호모델을 적용하여 적용 결과와 재식별 위험에 대한 분석 실시 분석 결과 원하는 수준에 미달할 경우 다시 6단계부터 진행 |
| 데이터 활용 | | <ul style="list-style-type: none"> 재식별 위험 분석 결과 적절한 수준의 가명·익명처리가 적용되었으면 데이터를 활용절차에 따라 활용 |

1. 개인정보 가명·익명처리 방법론

가명정보 조치사항(예)

| No | 속성 | 조치 사항 예시 | 비고 |
|----|------|--|-----------------------------------|
| 1 | 성별 | - 이용 목적상 필요하다면 별도 조치 없이 사용 가능 | - |
| 2 | 나이 | - 필요시 상황에 따라 5세, 10세 간격 등으로 범주화 - 특정 나이 이상 또는 이하의 경우 단일 범주로 집계(상·하단 코딩) | 범주화, 상·하단코딩 등 |
| 3 | 주소 | - 세부주소의 경우 식별자에 해당하므로, 필요시 시·군·구 단위 등으로 범주화 - 특히 도서산간 등 일부 지역의 경우 읍·면·동 단위의 거주자가 매우 적을 수 있으므로 필요시 범주화 등 조치 필요 - 우편번호에 대해서도 동일한 기준 적용 | 범주화 등 |
| 4 | 직업 | - 국회의원, 연예인, 운동선수 등 일부 직업의 경우 개인 식별 가능성이 높아지므로, 필요 시 직업 분류에 명시적으로 드러나지 않도록 조치 | 일반화, 범주화 등 |
| 5 | 국적 | - 특정 집단 내에서 대다수가 동일 국적자인 경우, 그 외 국적자에 대한 개인 식별 가능성이 높아지므로 필요시 범주화 등 조치 필요 | 범주화 등 |
| 6 | 기념일 | - 결혼기념일 등 일부 기념일의 경우 개인 식별 가능성이 높아지므로 필요시 범주화 등 조치 | 범주화 등 |
| 7 | 기혼여부 | - 이용 목적상 필요하다면 별도 조치 없이 사용 가능 | 범주화 등 |
| 8 | 거래지점 | - 거래지점의 경우 거래자의 주요 활동지를 한정 지을 수 있으므로 필요시 범주화 등 조치 필요 - 필요시 거래지점명/거래지점코드 대신 거래지점이 위치한 구 단위, 동 단위 주소 등으로 대체 | 범주화 등 |
| 9 | 기타 | - 데이터 특성 상 다른 정보와 결합하여 개인을 식별할 가능성이 높은 속성이 존재한다면 개인 식별 가능성이 높은 개인식별가능 정보로 지정 - 개인식별가능정보가 데이터 특성, 이용 상황, 제3자 제공 여부 등에 따라 재식별 위험이 높다고 판단되는 경우 추가 조치 적용 | 일반화, 범주화, 잡음추가, 삭제, 상·하단코딩등 |

1. 개인정보 가명·익명처리 방법론

■ 가명처리(예시) : 내부 활용 목적, 타 데이터와 결합 없음

| ID | 성명 | 카드 번호 | 전화 번호 | 성별 | 생년월일 | 주소 | 근무처 | 연봉(만원) | 내부신용등급 | 연체잔고(만원) | 결제 기관 |
|-------|-----|------------------------|---------------|-----|------------|----------------------|------------|--------|--------|----------|-----------|
| 19342 | 홍길동 | 3779 4593 3043 3921 | 010-3355-0934 | 남 | 1972.9.9. | 서울시 강남구 역삼동 332-1 | OO 자동차 | 4,500 | 5 | 35 | 국민 은행 |
| 19354 | 김철수 | 4832 2332 2344 4399 | 02-531-9834 | 남 | 1980.4.16. | 서울시 마포구 송내동 334-1 | OO 은행 | 6,000 | 2 | 0 | 새마을 금고 |
| 19445 | 전지연 | 4523 3234 9843 0394 | 010-9290-3344 | 여 | 1979.5.23. | 경기도 광주시 송정동 786-1 | OO 공사 | 5,500 | 9 | 125 | 우리 은행 |
| 20221 | 박식별 | 4932 3453 5943 4321 | 010-2891-3322 | 여 | 1983.12.3. | 경기도 용인시 죽전동 33-11 | OO 법무법인 | 7,000 | 1 | 0 | 기업 은행 |
| ... | ... | ... | ... | ... | ... | ... | ... | ... | ... | ... | ... |

- 자체 보유 개인정보를 가명처리하여 분석하고자 하는 상황으로, 가명정보를 생성한 회사의 내부에서 가명정보가 활용되고 있어 가명정보를 불법적으로 재식별할 의도가 상대적으로 낮다고 볼 수 있음
- 가명정보에 대해 개인정보에 준하는 보안대책을 적용하여 내부통제하에 활용하고 있으므로 재식별 가능성도 낮게 판단됨

※ 가명처리의 수준을 결정하기 위한 가명정보의 재식별 위험도 측정시 해당 이용기관의 재식별 의도와 능력, 가명정보 보호능력, 업무수행 신뢰도 등 다양한 측면에서 평가하여 이를 종합적으로 고려

1. 개인정보 가명·익명처리 방법론

■ 가명처리(예시) : 내부 활용 목적, 타 데이터와 결합 없음

| 구분 | 속성 | 위험 | 가명처리 예시* |
|-------------|-----|--|-------------|
| ID | 식별자 | <ul style="list-style-type: none"> - 신용카드사에서 개인을 식별하기 위한 ID이므로 개인이 특정될 가능성이 있음. - 만약, ID를 키로 사용하여 한다면 ID를 1:1 암호화 가능함 | 암호화 |
| 성명 | | - 식별자로서 개인을 특정할 수 있음 | 삭제 |
| 카드번호 | | - 식별자이며 다른 사업자도 보유하고 있을 수 있으므로 개인이 식별될 위험이 존재 | 삭제 |
| 전화번호 | | - 식별자이며 다른 사업자도 보유하고 있을 수 있으므로 개인이 식별될 위험이 존재 | 삭제 |
| 성별 | | - 생년월일, 주소 등의 정보와 조합될 경우 개인이 식별될 수 있음 | 별도 조치 없이 활용 |
| 생년월일 | | - 주소, 성별 등의 정보와 조합하여 개인이 식별될 수 있음 | 별도 조치 없이 활용 |
| 주소 | | <ul style="list-style-type: none"> - 상세주소의 경우 그 자체로 식별자로 판단될 수 있으며, 다른 정보와 조합하여 개인이 식별될 위험이 존재 - 주소를 정확히 아는 것이 분석목적과 관계가 없으면 범주화 처리 가능 | 범주화 |
| 직업 (근무처) | | <ul style="list-style-type: none"> - 공인 등 일부 직업의 경우 다른 정보와 조합하여 개인이 식별될 위험이 존재 - 근무처 값의 빈도를 구해 저 빈도의 경우, 일반화 범주화 필요 | 부분 범주화 |
| 연봉 | | <ul style="list-style-type: none"> - 수입이 너무 많거나 적은 경우 다른 정보와 조합하여 개인이 식별될 위험이 존재 - 범주화 필요 | 범주화/상하단 코딩 |
| 내부신용 등급 | | - 이미 등급화된 정보이며 외부에 유출될 경우 개인에 민감한 정보 | 별도 조치 없이 활용 |
| 연체잔고 | | - 외부에 유출될 경우 개인에 민감한 정보(특이치 존재시 상·하단 코딩적용) | 별도 조치 없이 활용 |
| 결제기관 | | - 결제기관 정보로는 개인이 식별될 위험이 크지 않음 | 별도 조치 없이 활용 |

1. 개인정보 가명·익명처리 방법론

가명처리 기록 보존

| 날짜 | 가명처리 근거 | 정보항목 | 가명처리 사유 | 가명처리 방법 |
|--------------|--|--------|---|--|
| 2020. 9. 1. | ① 가명처리 목적 연령대별 신용등급에 따른 연체율 연구 ② 관련 문서 첨부(연구계획(안) 등) ③ 근거 규정 「신용정보법」제32조제6항 제9호의2 등 | 고객ID | 식별자 | 삭제 |
| | | 이름 | 식별자 | 이름, 휴대폰번호를 조합하여 해시함수(SHA-256, 솔트값 적용)로 ID생성 후 삭제 |
| | | 휴대폰번호 | 식별자 | |
| | | 나이 | 개인식별가능정보 (재식별가능성이있고연구목적상구체적인 나이는 불필요) | 연령대로 범주화 |
| | | 대출금액 | 개인식별가능정보 (연구목적상구체적인 수치는 불필요) | 만원 단위로 반올림 |
| | | 연체기록 | 개인식별가능정보 (구체적인 수치는 재식별우려가있음) | 연체여부(Y/N)만 표기 |
| | | 신용등급 | 개인식별가능정보 (신용등급은 이미 범주화된등급이어서재식별 우려가 거의 없음) | 별도 조치 없음 |
| | | ... | ... | ... |
| 2020. 11. 3. | ① 가명처리 목적 보험가입자 특성에 대한공동연구 (X사의 데이터와결합후분석) ② 관련 문서 첨부 (양사간 계약서, 공동연구계획(안)등) ③ 근거 규정 「신용정보법」제17조의2 제32조제6항 제9호의2, 동법시행령제14조의2등 | 고객ID | 식별자 | 삭제 |
| | | 이름 | 식별자 | 이름, 휴대폰번호를 조합하여 해시함수(SHA-256, 솔트값 적용)로 ID생성 후 삭제 |
| | | 휴대폰번호 | 식별자 | |
| | | 거래지점 | 개인식별가능정보 (구체적인 지점 정보는 재식별 우려가 있음) | 구 단위로 범주화 |
| | | 보험가입건수 | 개인식별가능정보 (구체적인 가입건수 수치는 재식별 우려가 낮음) | 특이치만 삭제하고 별도 조치 없이 활용 |
| | | 약관대출금액 | 개인식별가능정보 (연구목적 상 구체적인 수치는 불필요) | 십만원 단위로 반올림 |
| | | ... | ... | ... |

1. 개인정보 가명·익명처리 방법론

■ 가명정보 처리 관련 사항에 대한 개인정보처리방침 작성 후 공개

「개인정보보호법」 제30조(개인정보 처리방침의 수립 및 공개)

① 개인정보처리자는 다음 각 호의 사항이 포함된 개인정보의 처리 방침(이하 "개인정보 처리방침"이라 한다)을 정하여야 한다.

이 경우 공공기관은 제32조에 따라 등록대상이 되는 개인정보파일에 대하여 개인정보 처리방침을 정한다.

1. 개인정보의 처리 목적

2. 개인정보의 처리 및 보유 기간

3. 개인정보의 제3자 제공에 관한 사항(해당되는 경우에만 정한다)

3의2. 개인정보의 파기절차 및 파기방법(제21조제1항 단서에 따라 개인정보를 보존하여야 하는 경우에는 그 보존근거와 보존하는 개인정보 항목을 포함한다)

4. 개인정보처리의 위탁에 관한 사항(해당되는 경우에만 정한다)

5. 정보주체와 법정대리인의 권리·의무 및 그 행사방법에 관한 사항

6. 제31조에 따른 개인정보 보호책임자의 성명 또는 개인정보 보호업무 및 관련 고충사항을 처리하는 부서의 명칭과 전화번호 등 연락처

7. 인터넷 접속정보파일 등 개인정보를 자동으로 수집하는 장치의 설치·운영 및 그 거부에 관한 사항(해당하는 경우에만 정한다)

8. 그 밖에 개인정보의 처리에 관하여 대통령령으로 정한 사항

PART 2.

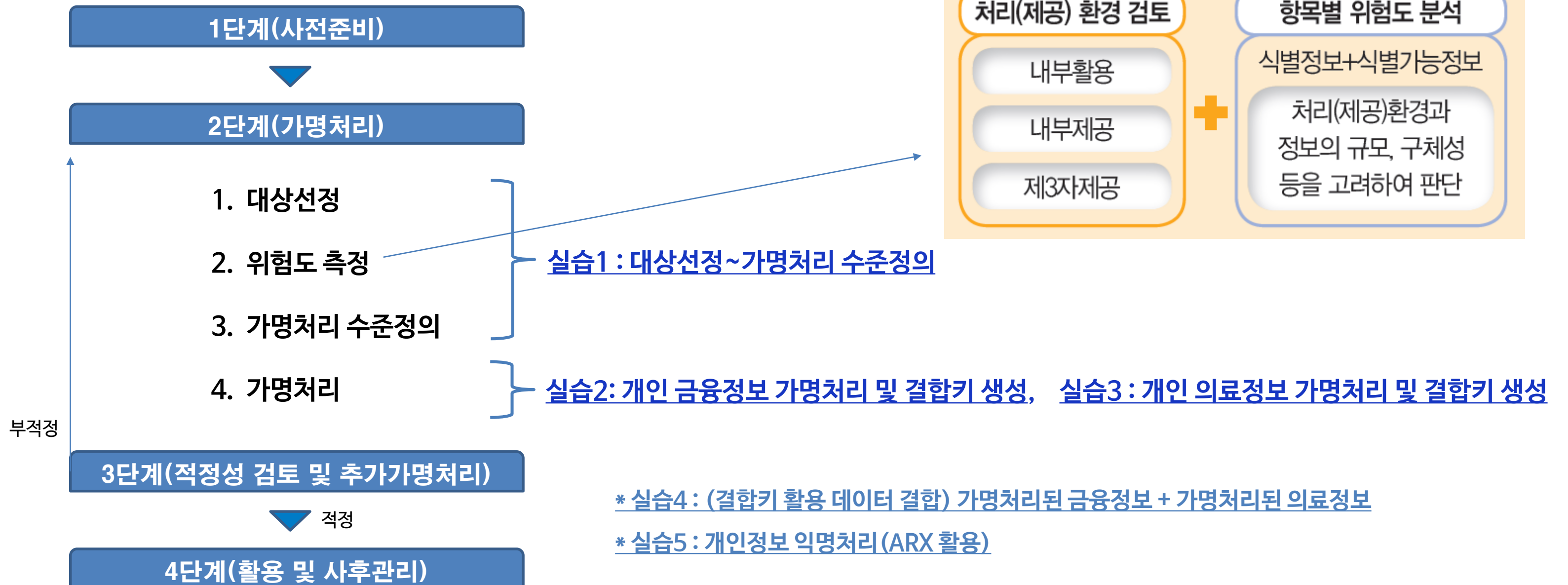
개인정보 가명처리 실습



2. 개인정보 가명·익명처리 실습

■ 실습 개요

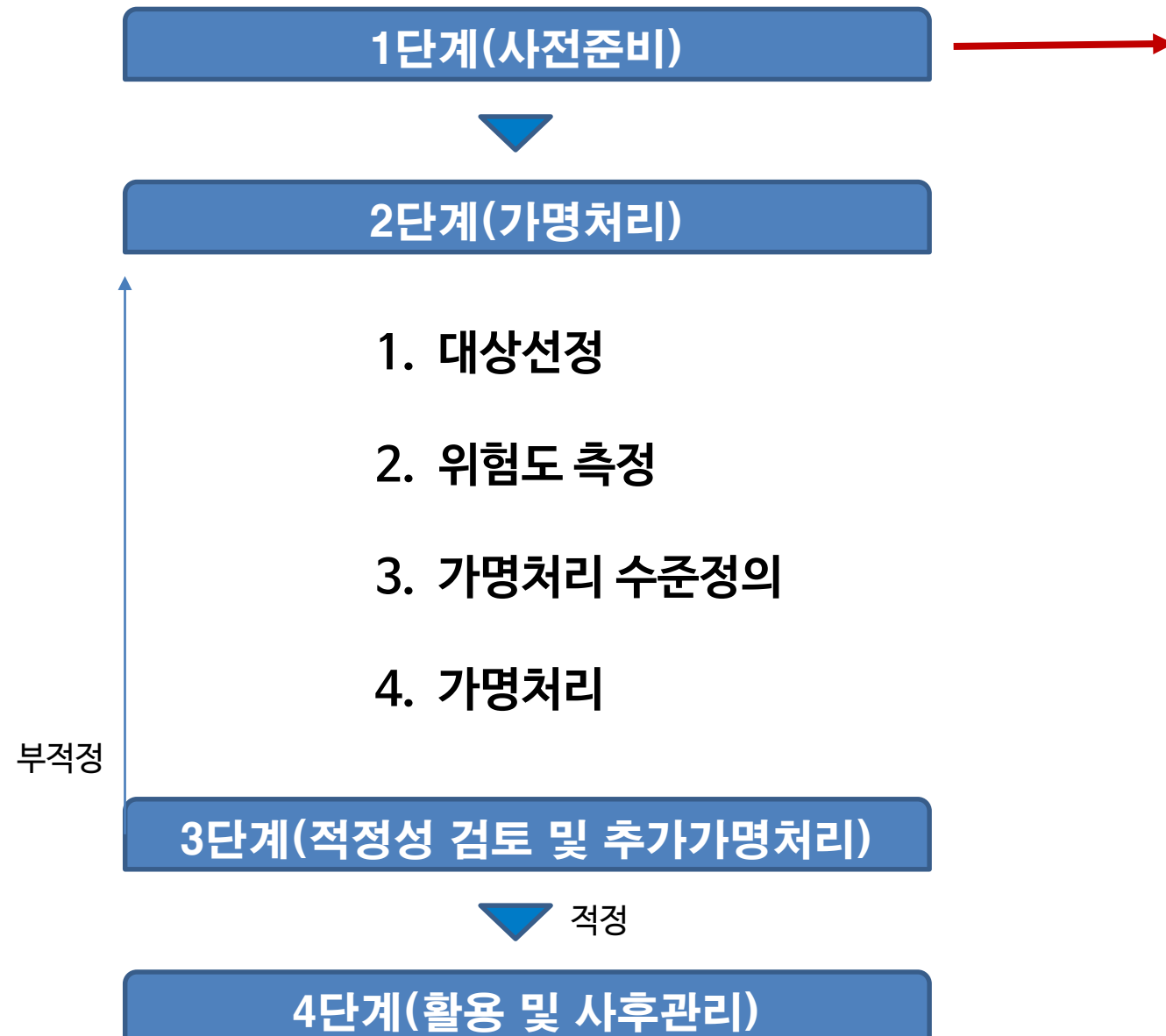
< 가명처리 세부절차 >



2. 개인정보 가명·익명처리 실습

■ 실습 개요

< 가명처리 세부절차 >



단계1 사전준비 : 처리 목적의 적합성 검토 및 가명처리 준비

가명처리를 위한 사전준비 단계에서는 가명정보 활용 목적을 명확히 하고 가명처리를 수행할 것인지를 결정하여야 하며, 가명처리하기로 결정한 경우 처리를 위하여 필요한 서류를 작성할 수 있음

- 가명정보의 처리목적 명확화 : 법률에서 허용하는 목적* 내에서 가명정보를 처리하는 목적을 최대한 명확히 작성하여야 함

* 통계작성, 과학적 연구, 공익적 기록보존 등에 한 함

- (적절하지 않은 예시) 신제품 개발을 위한 과학적 연구 수행
※ 목적이 구체적으로 명시되지 않아 적절하지 않음
- (적절한 예시) ○○제품의 성능 개선을 위해 개인별 ○○○특성에 대한 설문조사를 토대로 개인별 특성과 성능 요인의 연관성에 대한 과학적 연구

- 가명처리 적합성 검토(개인정보 보유부서 또는 전담부서) : 개인정보의 수집 목적 및 성격, 가명정보 활용 목적 등을 고려하여 가명처리 여부를 결정할 수 있음

※ 필요 시 심의위원회 구성 또는 외부전문가 평가 등을 통해 결정할 수 있음

- 필요서류 작성 : 가명정보의 처리 또는 가명처리를 위탁(보호법 제26조에 따라 수행)하거나 제3자에게 제공하는 경우 필요에 따라 재식별 금지에 관한 사항, 기타 처리에 있어 유의해야 할 사항* 등을 포함한 계약서를 작성할 수 있음

* (예시) 가명정보의 재제공 금지, 가명정보 재식별 금지, 가명정보의 안전성확보조치, 가명정보의 처리기록 작성 및 보관, 가명정보의 파기, 재식별 시 손해배상 등

- 가명정보 처리에 관한 내부관리계획이 없는 경우, 계획 수립 필요

[V.가명정보의 안전한 관리] 참조

2. 개인정보 가명·익명처리 실습

■ 실습1 (대상선정, 위험도 측정, 수준정의, 가명처리)

1. 대상선정 : 가명처리 목적 달성을 위하여 개인정보파일에서 가명처리에 필요한 항목을 추출

➤ 목적달성에 필요한 최소 항목을 처리하여야 함

개인
금융정보

| ID | Name | BirthDate | Age | Gender | Job | Income | Nodebt | Overdue | CreditScore |
|---------|------|-----------|-----|--------|---------|--------|--------|---------|---------------------|
| 2000002 | 이해성 | 19890927 | 28 | 여자 | 자영업_서비스 | 15200 | 7 | 530 | 1.1998485 |
| 2000003 | 장현태 | 19590529 | 58 | 여자 | 급여_공공 | 12000 | 12 | 80 | 2.1823490000000003 |
| 2000004 | 김서현 | 19801223 | 37 | 여자 | 급여_일반 | 7400 | 8 | 760 | 0.42755649999999995 |
| 2000005 | 김민섭 | 19870904 | 30 | 남자 | 정보없음 | 1700 | 23 | 1070 | 0.626918 |
| 2000006 | 조준이 | 19860929 | 31 | 남자 | 정보없음 | 5500 | 27 | 190 | 0.5381275 |
| 2000007 | 김효정 | 19950828 | 22 | 여자 | 자영업_서비스 | 6800 | 11 | 1670 | 4.4238285 |

➤ 대상선정 결과

- 가명처리 목적 : Credit Score와 나이, 성별, 직업, 수입 등의 관계를 모형화 하기 위해 개인정보 필요
- 가명처리 대상 : 아이디, 이름, 생년월일, 나이, 성별, 직업, 수입, 부채건수, 부채액, 신용평점
- 결합키 항목 : ID+이름

2. 개인정보 가명·익명처리 실습

■ 실습1 (대상선정, **위험도 측정**, 수준정의, 가명처리)

2. 위험도 측정 : 가명정보처리자의 개인정보 보호수준 및 다른 정보 보유여부 등을 검토하고, 항목별 위험도 분석을 통해
위험도 측정

가. 처리(제공)환경 검토 : 처리 목적에 따라 처리(제공)환경과
제공받는 자의 개인정보 보호수준 및 다른 정보 보유여부 등을 검토



※ 불특정 제3자(공개 등)에게 제공하는 경우
익명정보로 처리하는 것을 원칙으로 함

나. 항목별 위험도 분석

: 추출한 결과 정보의 항목별 위험도를 분석하여야 함



➤ 개인식별 가능성이 높은 항목을 분류하여 가명처리 방법 및 수준을
결정하는데 참고할 수 있도록 항목별 위험도를 분석

2. 개인정보 가명·익명처리 실습

■ 실습1 (대상선정, 위험도 측정, 수준정의, 가명처리)

가. 처리(제공)환경 검토 : 처리 목적에 따라 처리(제공)환경과 제공받는 자의 개인정보 보호수준 및 다른 정보 보유여부 등 검토

잘못된 내부활용 사례) ○○화장품 회사의 A팀은 화장품 판매 정보를 관리하는 팀으로서, 가명처리 접근권한을 분리하지 않고 해당 정보를 가명처리하여 신상품 수요조사 예측 모델 개발을 목적으로 활용

- (처리현황) A팀은 판매정보 내 개인식별 가능성이 있는 이름, 성별, 승인번호를 가명처리하고, 희귀 지역의 판매내역을 삭제하여 A팀 가명정보 분석담당자에게 제공
 - ✓ 가명정보 분석담당자는 A팀의 판매정보 관리 업무를 병행하여 업무를 수행하고 있음
- (문제점) 가명정보 분석담당자는 가명정보 분석을 통해 최고가 화장품의 금액과 판매지역을 파악할 수 있으며, 판매정보 관리 업무를 병행하고 있어 해당 금액과 지역을 통해 특정개인을 식별할 가능성이 있음
- (해결방안) 가명정보처리자는 가명정보를 처리(분석)하는 담당자가 다른 정보(위의 예시에서는 판매정보 관리)에 동시에 접근할 수 없도록 접근권한을 명확히 분리하고, 접근통제를 실시하여야 함

잘못된 내부제공 사례) □ □공사는 A부서에서 관리하고 있는 고속도로 이용차량 빅데이터 분석 결과를 고속도로 통행요금을 관리하는 B부서에서 처리하는 개인정보를 확인하지 않고 교통서비스 개선을 위한 연구 목적으로 내부 제공

- (처리현황) A부서는 개인식별가능성이 있는 차량번호, 차종 등을 가명처리하고, 이동시간, 이동량, 사고정보 등의 정보를 B부서에 제공
 - ✓ B부서는 고속도로 통행요금 관리를 위해 고객번호와 차량번호, 톨게이트 입출시간 및 결제금액 정보를 보유하고 있음
- (문제점) B부서는 A부서에서 제공받은 정보의 이동시간 정보와 B부서가 보유한 톨게이트 입출시간을 활용하여 특정시간에 통과한 차량의 번호를 알 수 있으며, 해당 차량번호를 통해 특정 개인을 식별할 가능성이 있음
- (해결방안) 가명정보처리자는 가명정보 생성 시 가명정보를 처리할 부서에서 보유하고 있는 정보를 고려하여 가명처리를 수행하여야 함

잘못된 제3자제공 사례) ○○호텔에서는 최고급 객실을 이용한 VIP등의 특이정보를 삭제하지 않고 호텔 투숙 및 서비스 금액 등을 ○○분석회사에 제공하고, ○○분석회사는 해당 정보를 분석하여 시간에 따른 객실이용현황 및 서비스이용에 대한 조사 연구를 수행

- (처리현황) ○○분석회사는 온라인 SNS정보 및 다양한 기업의 정보를 수집하여 다양한 연구조사를 실시하는 회사로서 내부관리계획을 수립하고, 관리적·기술적 보호조치를 준수하고 있음
 - ✓ ○○호텔은 회원번호와 이름을 가명처리하고, 나이, 성별, 등급, 예약방법, 객실정보, 체크인, 체크아웃, 서비스 이용금액을 제공
- (문제점) ○○분석회사의 분석담당자는 특정일에 최고급 객실을 이용한 내용을 분석과정에서 인지할 수 있으며, 기존 업무(온라인 SNS정보 수집)를 수행하며 공개된 정보(예 : 개인이 SNS에 올리는 정보, 여행후기 등)를 통해 특정 개인을 식별할 가능성이 있음
- (해결방안) ○○호텔은 제공하는 가명정보에 포함된 특이정보(최고급 객실)를 삭제 또는 가명처리 등을 수행하여야 함

2. 개인정보 가명·익명처리 실습

■ 실습1 (대상선정, 위험도 측정, 수준정의, 가명처리)

나. 항목별 위험도 분석

: 추출한 결과 정보의 항목별 위험도를 분석하여야 함

<특이정보 예시>



* 규모(레코드, 항목의 크기), 구체성(정보의 정확성 수준)

개인식별 가능성이 높은 정보 <예시>

- 식별정보
 - 고유식별정보(여권번호, 외국인등록번호, 운전면허번호), 성명, 전화번호, 전자우편주소, 의료기록번호, 건강보험번호, 자동차 등록번호 등 외부 연계(식별)를 목적으로 생성된 정보 등
- 식별가능정보
 - 성별, 연령(나이), 국적, 혈액형, 신장, 몸무게, 직업, 위치정보, 세부주소 등 가명정보를 처리하는 자의 입장에서 개인을 알아볼 수 있는* 정보
 - * 개인을 '알아볼 수 있는'의 판단기준은 해당 정보를 처리하는 자(정보의 제공 관계에 있어서는 제공받은 자를 포함)
 - 특이정보
 - 국내 최고령, 최장신, 고액체납금액, 고액급여수급자 등 전체적인 패턴에서 벗어나 극단값이 발생 할 수 있는 정보
 - 희귀 성씨, 희귀 혈액형, 희귀 눈동자 색깔, 희귀 병명, 희귀 직업 등 정보 자체로 특이한 값을 가지고 있는 정보

사례1) 특정 지역에서 활동하고 있는 직업(국회의원) 정보의 경우 지역구 국회의원 명단 등을 통해 개인식별 가능성이 매우 높음

※ (예시) 특정 지역을 인접 지역에 병합*을 하거나, 직업을 일반화(정치인)하여 가명처리 수행

* 국가통계기관의 경우 세부 지역단위 통계 시 2천명이 되지 않는 경우 인접 지역에 병합

사례2) 호텔, 렌터카 회사에서 자사가 운영하고 있는 객실 이용정보 및 보유하고 있는 차량에 대한 이용내역을 가명처리하여 여행사에 제공하고, 여행사는 신규 여행상품 개발에 관한 연구를 수행하고자 함

- 호텔, 렌터카 등 여행업종에서 제공하는 객실이용정보, 차량이용정보는 특정 개인(공인 등)이 SNS등 온라인에 공개하는 정보와 결합되어 사생활 침해 우려가 매우 높음

※ (예시) 특정 차량(슈퍼카)의 이름을 일반화(스포츠카)하여 게시하거나, 호텔 최고급 객실 정보를 일반객실 정보로 가명처리 수행

사례3) 나이와 성별, 지역, 직업에 따른 질병의 분포 연구를 위해 진료정보를 제공하고자 함

- 희귀질병과 직업(공인)의 경우 해당 병명만으로 개인식별 가능성이 매우 높음

※ (예시) 희귀질병을 일반화(일반 질병명)하거나, 직업을 일반화(회사원)하여 가명처리 수행

사례4) 보훈대상 정보와 병원의 정보를 결합하여 보훈대상자에 대한 질병분포 연구를 수행하고자 함

- 특정 보훈대상(참전유공자 등)과 특정 종교, 직업 등의 정보가 결합될 경우 온라인 검색 등을 통해 특정 개인(공인 등)이 식별될 가능성이 매우 높음

※ (예시) 보훈대상의 빈도가 낮은 항목을 다른 분류로 병합을 하거나, 특정 종교, 질병명(알츠하이머) 등 특이항목이 발생하는 항목을 삭제 또는 일반화(뇌질환) 수행

2. 개인정보 가명·익명처리 실습

■ 실습1 (대상선정, 위험도 측정, 수준정의, 가명처리)

2. 위험도 측정 : 가명정보처리자의 개인정보 보호수준 및 다른 정보 보유여부 등을 검토하고, 항목별 위험도 분석을 통해
위험도를 측정

```
ID,Name,BirthDate,Age,Gender,Job,Income,Nodebt,Overdue,CreditScore
2000002,이해성,19890927,28,여자,자영업_서비스,15200,7,530,1.1998485
2000003,장현태,19590529,58,여자,급여_공공,12000,12,80,2.1823490000000003
2000004,김서현,19801223,37,여자,급여_일반,7400,8,760,0.42755649999999995
2000005,김민섭,19870904,30,남자,정보없음,1700,23,1070,0.626918
2000006,조준이,19860929,31,남자,정보없음,5500,27,190,0.5381275
2000007,김효정,19950828,22,여자,자영업_서비스,6800,11,1670,4.4238285
```

➤ 위험도 측정 결과

- 처리(제공)환경 : 내부 연구 목적으로 사용하기 때문에 가명처리를 원칙으로 하되 일부 민감한 컬럼에 대해 범주화, 일반화 등 처리
(재식별 가능성 최소화 처리)
- ID(삭제), 이름(제거), 생년월일(분석목적과 상관없으므로 삭제), 나이(범주화), 성별(유지), 직업(특이정보 식별),
수입액(범주화), 대출건수(유지), 대출금액(범주화), 신용등급(Round 처리)

2. 개인정보 가명·익명처리 실습

■ 실습1 (대상선정, 위험도 측정, 수준정의, 가명처리)

〈 가명처리 검토 결과보고서 작성 결과(샘플) 〉

| | 가명처리 검토 결과 보고서 |
|------------------|---|
| 가명정보 활용목적 | 신규고객의 신용평가 점수를 예측하기 위해 고객의 개인정보와 신용평가 점수간의 관계모형을 만들 고자 함. 분석 및 활용이 모두 사내에서 이루어지고 분석의 결과는 수학적 모형으로 산출되어 개인정 보 이슈는 거의 없으나 혹시 모를 개인정보 사고에 대비하여 불필요한 개인정보는 가명처리하고자 함 |
| 가명정보 항목 | 아이디, 이름, 생년월일, 나이, 성별, 직업, 수입액, 대출건수, 대출금, 신용평점 |
| 처리 환경 검토 | 처리환경: 사내에서 추출되고, 사내에서 분석됨 |
| 항목별 위험도 분석 | 식별자인 ID, 이름은 삭제처리하고 준식별자인 생년월일은 분석에 불필요항목으로 삭제함 준식별자인 나이(범주화), 성별(유지), 직업(특이정보 식별)은 익명화 처리하고, 민감정보인 수입액은 라운딩, 대출건수는 유지, 대출금액은 범주화, 신용등급은 Round 처리함 |
| 최종 검토 의견 | 내부 통계활용 목적으로 위험성이 높지 않아 가명처리가 가능하다고 판단하였음 가명처리 과정에서 재식별 등 위험성을 방지하기 위해 식별자는 삭제, 준식별자는 범주화, 특이값 처리등의 조치를 취하였고 민감정보 역시 익명화 수준의 처리를 시행하여 안전하게 활용하고자 함 |

■ 실습1 (대상선정, 위험도 측정, **수준정의**, 가명처리)

3. 수준정의 : ‘가명처리 검토 결과보고서’를 기반으로 활용목적 달성에 필요한 수준을 고려하여 가명처리 수준을 정의 함

| | 가명처리 수준 정의서 |
|------|--|
| 항목명 | 처리수준 |
| 아이디 | 삭제 |
| 이름 | 삭제 |
| 생년월일 | 삭제 |
| 나이 | 5세 단위 범주화(5세 단위로 범주화했을 때, 분석 정확도에 영향이 없다고 판단) |
| 성별 | 유지(분석 목적상 필요하고 재식별 위험 변수가 아님) |
| 직업 | 특이정보만 일반화/범주화 처리하고 나머지는 유지(분석 목적상 필요하나 특이정보의 식별성 존재) |
| 수입액 | 분석 정확도를 해치지 않는 범위에서 만원 단위 라운딩 |
| 대출건수 | 유지(분석 목적상 필요하고 재식별 위험 변수가 아님) |
| 대출금 | 범주화(분석 정확도를 해치지 않는 범위로 100만원 단위 범주화) |
| 신용평점 | 정수처리 라운딩(분석 정확도를 해치지 않는 범위) |

■ 실습1 (대상선정, 위험도 측정, **수준정의**, 가명처리)

3. 수준정의 : ‘가명처리 검토 결과보고서’를 기반으로 활용목적 달성에 필요한 수준을 고려하여 가명처리 수준을 정의 함

| | 가명처리 수준 정의서 |
|------|--|
| 항목명 | 처리수준 |
| 아이디 | 삭제 |
| 이름 | 삭제 |
| 생년월일 | 삭제 |
| 나이 | 5세 단위 범주화(5세 단위로 범주화했을 때, 분석 정확도에 영향이 없다고 판단) |
| 성별 | 유지(분석 목적상 필요하고 재식별 위험 변수가 아님) |
| 직업 | 특이정보만 일반화/범주화 처리하고 나머지는 유지(분석 목적상 필요하나 특이정보의 식별성 존재) |
| 수입액 | 분석 정확도를 해치지 않는 범위에서 만원 단위 라운딩 |
| 대출건수 | 유지(분석 목적상 필요하고 재식별 위험 변수가 아님) |
| 대출금 | 범주화(분석 정확도를 해치지 않는 범위로 100만원 단위 범주화) |
| 신용평점 | 정수처리 라운딩(분석 정확도를 해치지 않는 범위) |

2. 개인정보 가명·익명처리 실습

■ 실습1 (대상선정, 위험도 측정, 수준정의, 가명처리)

가명처리에서 주로 사용하는 방법은 컬럼 삭제, 특이값 처리, 일반화, 범주화로 볼 수 있다.
통계 소프트웨어 R을 사용하여 이 문제를 해결해 보자.



[Home]

Download

CRAN

R Project

The R Project for Statistical Computing

Getting Started

R is a free software environment for statistical computing and graphics. It compiles and runs on a wide variety of UNIX platforms, Windows and MacOS. To **download R**, please choose your preferred [CRAN mirror](#).

If you have questions about R like how to download and install the software, or what the license terms are, please read our [answers to frequently asked questions](#) before you send an email.

R은 프리웨어로 제공되는 통계 소프트웨어다.

SAS, SPSS, Minitab 등 다른 통계 소프트웨어와 다르게 객체지향 언어이므로 JAVA, C 등의 컴퓨터 언어와 결합이 쉽고, 자체적으로 라이브러리를 만드는 것도 용이하다.

오픈 소프트웨어로 풍부한 라이브러리를 가지고 있고, 전세계에서 가장 많이 사용되는 통계 소프트웨어다.

R은 <https://www.r-project.org/> 에서 다운로드 가능하다.

리눅스 OS에서 개발되었지만, 윈도우와 맥 OS에서도 사용가능하다.

한글 사용이나 윈도우 사용자명, 폴더명 등에 한글 사용제약이 존재한다.

2. 개인정보 가명·익명처리 실습

■ 실습1 (대상선정, 위험도 측정, 수준정의, 가명처리)

R 프로그램을 작성하고 실행하는 개발환경으로 RStudio를 많이 사용한다.

RStudio는 상용 소프트웨어지만, 무료버전을 제공하고 있다.

개인적 목적으로는 무료버전을 사용할 수 있고, 상업적 목적으로는 무료버전을 사용할 수 없다.

우리는 무료버전 중 데스크톱 버전을 사용할 것이다.

Download at [<https://www.rstudio.com/>]

| RStudio Desktop | RStudio Desktop | RStudio Server | RStudio Server Pro |
|---------------------|--------------------|---------------------|-------------------------|
| Open Source License | Commercial License | Open Source License | Commercial License |
| Free | \$995 | Free | \$4,975 |
| | /year | | /year |
| | | | (5 Named Users) |
| DOWNLOAD | BUY | DOWNLOAD | BUY |
| Learn more | Learn more | Learn more | Evaluation Learn more |

2. 개인정보 가명·익명처리 실습

■ 실습1 (대상선정, 위험도 측정, 수준정의, 가명처리)

```
setwd("D:/NaverCloud/Lecture/KICT/data")
dsn = read.csv("금융정보.csv", header = T, fileEncoding = "utf-8")
library("dplyr")

# 컬럼 삭제
dsn = select(dsn, -c(ID, Name, BirthDate))

# 나이 범주화
dsn$AgeGrp <- NA
dsn$AgeGrp <- ifelse(dsn$Age < 20, 0, dsn$AgeGrp)
dsn$AgeGrp <- ifelse(dsn$Age >= 20 & dsn$Age <25, 1, dsn$AgeGrp)
dsn$AgeGrp <- ifelse(dsn$Age >= 25 & dsn$Age <30, 2, dsn$AgeGrp)
dsn$AgeGrp <- ifelse(dsn$Age >= 30 & dsn$Age <35, 3, dsn$AgeGrp)
dsn$AgeGrp <- ifelse(dsn$Age >= 35 & dsn$Age <40, 4, dsn$AgeGrp)
dsn$AgeGrp <- ifelse(dsn$Age >= 40 & dsn$Age <45, 5, dsn$AgeGrp)
dsn$AgeGrp <- ifelse(dsn$Age >= 45 & dsn$Age <50, 6, dsn$AgeGrp)
dsn$AgeGrp <- ifelse(dsn$Age >= 50 & dsn$Age <55, 7, dsn$AgeGrp)
dsn$AgeGrp <- ifelse(dsn$Age >= 55 & dsn$Age <60, 8, dsn$AgeGrp)
dsn$AgeGrp <- ifelse(dsn$Age >= 60, 9, dsn$AgeGrp)
dsn$Age <- dsn$AgeGrp
dsn = select(dsn, -c(AgeGrp))
```

금융정보가명처리.R

2. 개인정보 가명·익명처리 실습

■ 실습1 (대상선정, 위험도 측정, 수준정의, 가명처리)

```
# 특이정보 처리(범주형변수)
table(dsn$Job)
dsn$Job = ifelse(dsn$Job == "국회의원", "급여_공공", dsn$Job)
table(dsn$Job)
# 특이정보 처리(연속형변수) --> 범주화로 처리 혹은 특이값이 아닌 최대값으로 변경 대체
boxplot(dsn$Nodebt)
boxplot.stats(dsn$Nodebt)$out
result <- boxplot.stats(dsn$Nodebt)
result
dsn$Nodebt <- ifelse(dsn$Nodebt > 30, 30, dsn$Nodebt)
boxplot(dsn$Nodebt)

# Rounding(천원단위)
boxplot(dsn$Income)
dsn$Income <- round(dsn$Income/10**3, digits = 0) * 10**3
boxplot(dsn$Overdue)
dsn$Overdue <- round(dsn$Overdue/10**2, digits = 0) * 10**2
boxplot(dsn$Income)
boxplot(dsn$Overdue)
# Rounding(정수화)
dsn$CreditScore <- round(dsn$CreditScore)

write.csv(dsn, file="금융정보_Pseudonymization.csv", row.names=FALSE, fileEncoding='utf-8')
```

2. 개인정보 가명·익명처리 실습

■ 실습2 병원 정보 가명처리

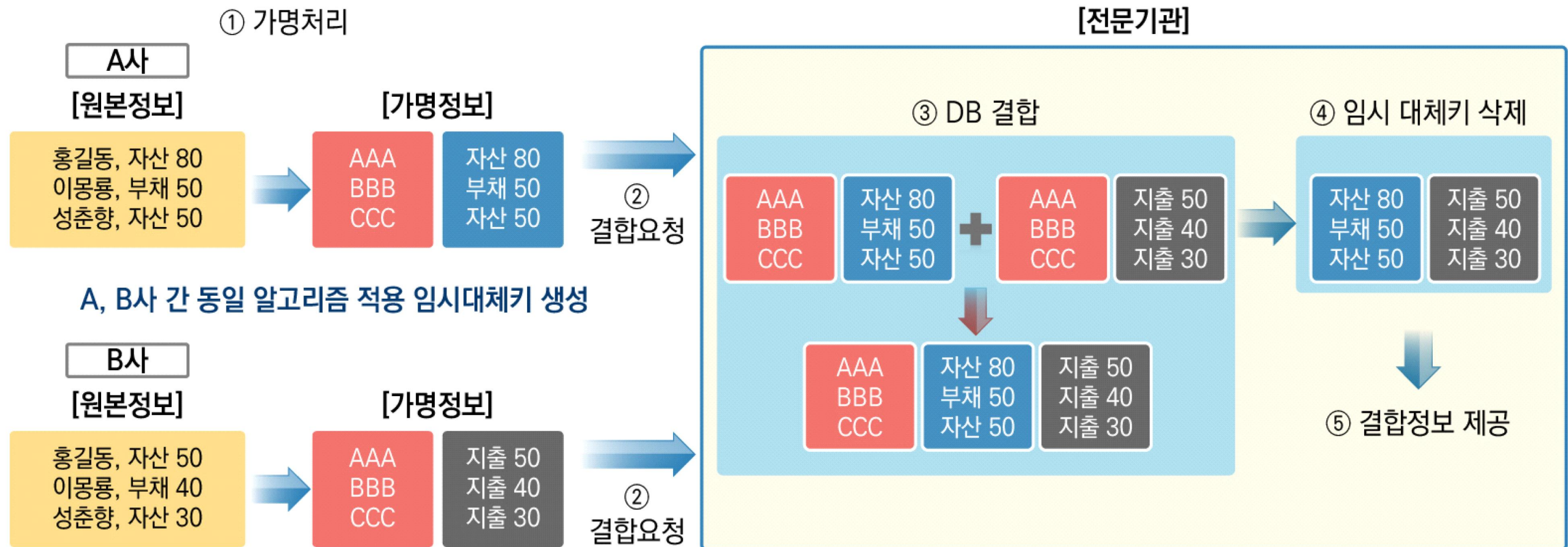
금융정보 가명 처리 예를 참조하여 병원 정보 가명처리를 수행한 후, 병원정보_Pseudonymization.csv 파일을 생성하시오.

실 습

2. 개인정보 가명·익명처리 실습

■ 실습3 가명정보 결합

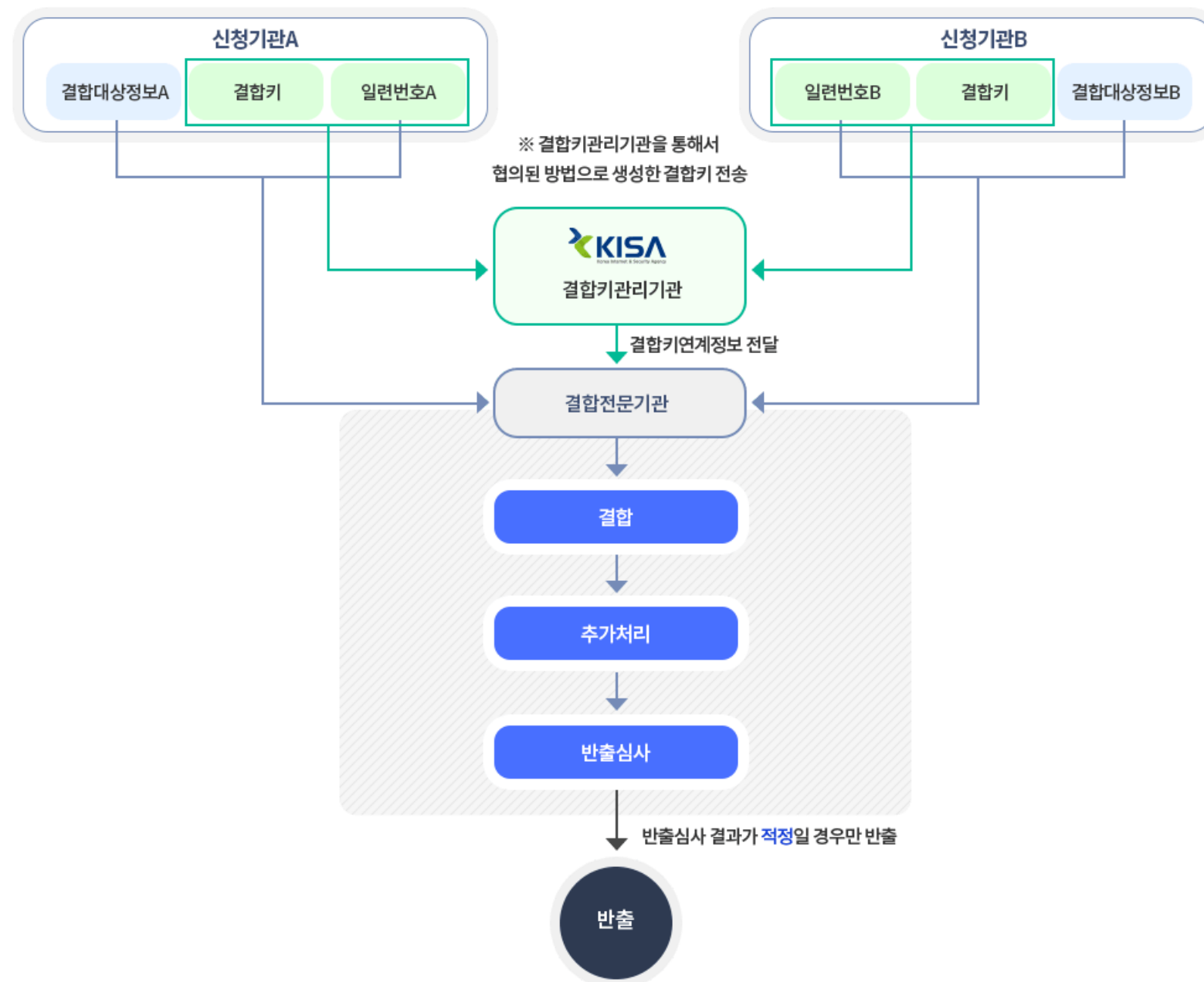
1. 대상선정: 양사가 보유한 정보에서 결합 키와 결합 데이터 항목을 협의하고 전문기관에 결합을 의뢰함



2. 개인정보 가명·익명처리 실습

■ 실습3 가명정보 결합

2. 결합키 생성: 신청기관 A, B는 KISA와 결합키 알고리즘을 협의하여 각자 보유 정보의 결합키를 암호화하여 KISA에 송부한다.



2. 개인정보 가명·익명처리 실습

■ 실습3 가명정보 결합

2. 결합키 생성: 신청기관 A, B는 KISA와 결합키 알고리즘을 협의하여 각자 보유 정보의 결합키를 암호화하여 KISA에 송부한다.

```
setwd("D:/NaverCloud/Lecture/KICT")
dsn1 = read.csv("금융정보가공.csv", header = T, fileEncoding = "utf-8")
dsn2 = read.csv("병원정보가공.csv", header = T, fileEncoding = "utf-8")

library(digest)

# 결합키 생성 예(sha256 알고리즘으로 협의했다고 가정)
dsn1$combine_key = sapply(paste(dsn1$Name, dsn1$BirthDate), digest, algo="sha256")
dsn2$combine_key = sapply(paste(dsn2$Name, dsn2$BirthDate), digest, algo="sha256")

dsn1$combine_key[1]
dsn2$combine_key[1]

library("dplyr")

# KISA에 송부할 결합키 파일 생성
dsn1_key = select(dsn1, c(ID, combine_key))
dsn2_key = select(dsn2, c(ID, combine_key))
write.csv(dsn1_key, file="금융정보_key.csv", row.names=FALSE, fileEncoding='utf-8')
write.csv(dsn2_key, file="병원정보_key.csv", row.names=FALSE, fileEncoding='utf-8')
```

KISA송부키생성.R

2. 개인정보 가명·익명처리 실습

■ 실습3 가명정보 결합

3. 결합요청: 결합을 원하는 정보를 결합기관에 송부한다. 이 때, 결합키는 제외하고 적절한 가명화를 수행해야 한다.

```
setwd("D:/NaverCloud/Lecture/KICT/data")
dsn = read.csv("금융정보.csv", header = T, fileEncoding = "utf-8")
library("dplyr")

# 컬럼 삭제
dsn = select(dsn, -c(ID, Name, BirthDate))
# 나이 범주화
dsn$AgeGrp <- NA
dsn$AgeGrp <- ifelse(dsn$Age < 20, 0, dsn$AgeGrp)
dsn$AgeGrp <- ifelse(dsn$Age >= 20 & dsn$Age < 25, 1, dsn$AgeGrp)
dsn$AgeGrp <- ifelse(dsn$Age >= 25 & dsn$Age < 30, 2, dsn$AgeGrp)
dsn$AgeGrp <- ifelse(dsn$Age >= 30 & dsn$Age < 35, 3, dsn$AgeGrp)
dsn$AgeGrp <- ifelse(dsn$Age >= 35 & dsn$Age < 40, 4, dsn$AgeGrp)
dsn$AgeGrp <- ifelse(dsn$Age >= 40 & dsn$Age < 45, 5, dsn$AgeGrp)
dsn$AgeGrp <- ifelse(dsn$Age >= 45 & dsn$Age < 50, 6, dsn$AgeGrp)
dsn$AgeGrp <- ifelse(dsn$Age >= 50 & dsn$Age < 55, 7, dsn$AgeGrp)
dsn$AgeGrp <- ifelse(dsn$Age >= 55 & dsn$Age < 60, 8, dsn$AgeGrp)
dsn$AgeGrp <- ifelse(dsn$Age >= 60, 9, dsn$AgeGrp)
dsn$Age <- dsn$AgeGrp
dsn = select(dsn, -c(AgeGrp))
```

금융정보가명처리.R

2. 개인정보 가명·익명처리 실습

■ 실습3 가명정보 결합

```
# 특이정보 처리(범주형 변수)
table(dsn$Job)
dsn$Job = ifelse(dsn$Job == "국회의원", "급여_공공", dsn$Job)
table(dsn$Job)
# 특이정보 처리(연속형 변수) --> 범주화로 처리 혹은 특이값이 아닌 최대값으로 변경 대체
boxplot(dsn$Nodebt)
boxplot.stats(dsn$Nodebt)$out
result <- boxplot.stats(dsn$Nodebt)
result
dsn$Nodebt <- ifelse(dsn$Nodebt > 30, 30, dsn$Nodebt)
boxplot(dsn$Nodebt)

# Rounding(천원 단위)
boxplot(dsn$Income)
dsn$Income <- round(dsn$Income/10**3, digits = 0) * 10**3
boxplot(dsn$Overdue)
dsn$Overdue <- round(dsn$Overdue/10**2, digits = 0) * 10**2
boxplot(dsn$Income)
boxplot(dsn$Overdue)
# Rounding(정수화)
dsn$CreditScore <- round(dsn$CreditScore)
write.csv(dsn, file="금융정보_결합요청.csv", row.names=FALSE, fileEncoding='utf-8')
```

2. 개인정보 가명·익명처리 실습

■ 실습3 가명정보 결합

```
setwd("D:/NaverCloud/Lecture/KICT/data")
dsn = read.csv("병원정보.csv", header = T, fileEncoding = "utf-8")
library("dplyr")

# 컬럼 삭제
dsn = select(dsn, -c(ID, Name, BirthDate))
# 나이 범주화
dsn$AgeGrp <- NA
dsn$AgeGrp <- ifelse(dsn$Age < 20, 0, dsn$AgeGrp)
dsn$AgeGrp <- ifelse(dsn$Age >= 20 & dsn$Age <25, 1, dsn$AgeGrp)
dsn$AgeGrp <- ifelse(dsn$Age >= 25 & dsn$Age <30, 2, dsn$AgeGrp)
dsn$AgeGrp <- ifelse(dsn$Age >= 30 & dsn$Age <35, 3, dsn$AgeGrp)
dsn$AgeGrp <- ifelse(dsn$Age >= 35 & dsn$Age <40, 4, dsn$AgeGrp)
dsn$AgeGrp <- ifelse(dsn$Age >= 40 & dsn$Age <45, 5, dsn$AgeGrp)
dsn$AgeGrp <- ifelse(dsn$Age >= 45 & dsn$Age <50, 6, dsn$AgeGrp)
dsn$AgeGrp <- ifelse(dsn$Age >= 50 & dsn$Age <55, 7, dsn$AgeGrp)
dsn$AgeGrp <- ifelse(dsn$Age >= 55 & dsn$Age <60, 8, dsn$AgeGrp)
dsn$AgeGrp <- ifelse(dsn$Age >= 60, 9, dsn$AgeGrp)
dsn$Age <- dsn$AgeGrp
dsn = select(dsn, -c(AgeGrp))
```

병원정보가명처리.R

2. 개인정보 가명·익명처리 실습

■ 실습3 가명정보 결합

```
# 특이정보 처리(범주형변수)
sort(table(dsn$Dept))
dsn$Dept = ifelse(dsn$Dept %in% c("BRES", "COGE","CORO","PAMD","PCHO"), "etc", dsn$Dept)

sort(table(dsn$DCode))

sort(table(dsn$Medicine))
dsn$Medicine = ifelse(dsn$Medicine %in% c(10180, 10206, 10567, 10695, 10808, 10826, 10838,
10957, 11096, 11148, 11441, 11592, 11706, 11725,
11826, 11832, 12140, 12525, 12529, 50398, 50773,
51277, 61379, 65036, 65055, 90928), 99999, dsn$Medicine)

write.csv(dsn, file="병원정보_결합요청.csv", row.names=FALSE, fileEncoding='utf-8')
```

실습3 가명정보 결합

3. 결합 신청서 작성: 신청기관은 결합목적 및 제공기관 정보를 결합전문기관장에게 제출해야 한다.

■ 「가명정보의 결합 및 반출 등에 관한 고시」 [별지 제3호서식]

결합 신청서

신청번호

| | | | |
|--|---|--|-------------------------------|
| 결합신청자 | | | |
| 기관명 | 사업자등록번호 또는 법인등록번호 | | |
| 주소 | 대표자명 | | |
| 담당자 | 담당자 연락처 (전화, e-mail) | | |
| 신청자 구분 | <input type="checkbox"/> 개인 <input type="checkbox"/> 공공기관 <input type="checkbox"/> 비영리법인 <input type="checkbox"/> 민간기관 | | |
| 가명정보 제공 | | | 해당없음 <input type="checkbox"/> |
| 파일명 | | | |
| 제공방법 | <input type="checkbox"/> 온라인 <input type="checkbox"/> 오프라인 | | |
| 제출예정일 | 년 월 일 | | |
| 전체 가명정보 제공 기관명(총수) | | | |
| 제공정보 요약 | 결합 수 | | |
| | 전체 레코드 수 | | |
| | 전체 파일 크기 | | |
| 결합 결과물 이용 | | | 해당없음 <input type="checkbox"/> |
| 시계열 분석 | <input type="checkbox"/> 해당없음 <input type="checkbox"/> 시계열(신규) <input type="checkbox"/> 시계열(추가, 결합접수번호 :) | | |
| 결합목적 | <input type="checkbox"/> 통계작성 <input type="checkbox"/> 과학적 연구 <input type="checkbox"/> 공익적 기록보존 등 | | |
| 세부결합목적 | | | |
| 「개인정보보호법」 제28조의3제1항 및 같은 법 시행령 제29조의3제1항에 따른 결합 을 위하여 결합전문기관에 결합신청서를 위와 같이 제출합니다. 년 월 일 신청인 (서명 또는 인) | | | |
| 결합전문기관의 장 귀하 | | | |
| 첨부 서류 | 1. 사업자등록증, 법인등기부등본 등 결합신청자 관련 서류 1부 2. 결합 대상 가명정보에 관한 서류(전체 항목명 가명처리 대상 항목명, 가명처리 기법 및 예시 등) 1부(해당 경우에 한함) * 결합키 생성에 사용된 속성 제외 3. 결합 목적을 증명할 수 있는 서류 1부 4. 안전조치계획 및 이를 증빙할 수 있는 서류 1부 | | |

- 17 -

[별지 제8의2호 서식] 정보집합물 결합신청서(제15조의2 제1항 관련)

정보집합물 결합 신청서

| | | | | | |
|---|-------------------------------|----------------------------------|---|-----|--|
| 접수번호 | | | | 접수일 | |
| 결합의뢰 기관명 | 기관명 | | 담당자(성명,직함) | | |
| | 담당부서 | | 전화번호 | | |
| 결합상대 기관명 | 소재지 | | 이메일 주소 | | |
| | 기관명 | | | | |
| (복수기재 가능) | 담당부서 | | | | |
| | 소재지 | | | | |
| 결합목적 | | 간략히 서술 (예사: 내부활용, 제3자 제공(이용기관명)) | | | |
| 정보집합물 주요내용 요약 | | 간략히 서술 (예사: 16년 9월 A병원의 진료내역) | | | |
| 결합 데이터 제공형태 | <input type="checkbox"/> 가명정보 | 이용목적 | <input type="checkbox"/> 통계작성 (상업적 목적 포함) <input type="checkbox"/> 연구 (산업적 연구 포함) <input type="checkbox"/> 공익적 기록보존 등 | | |
| | | 결합키 | <input type="checkbox"/> 삭제 <input type="checkbox"/> 대체 | | |
| | <input type="checkbox"/> 익명정보 | | | | |
| 「신용정보의 이용 및 보호에 관한 법률」 제17조의2제1항에 따라 위와 같이 신청합니다. 년 월 일 신청인(기관장명) (서명 또는 인) (전문기관명)장 귀하 | | | | | |
| 첨부서류 | 정보집합물의 데이터 명세서 | | | | |

2. 개인정보 가명·익명처리 실습

■ 실습3 가명정보 결합

4. 결합 수행: 두 기관에서 받은 정보를 결합한다.

```
setwd("D:/NaverCloud/Lecture/KICT/Data")

# KISA는 두 기관에서 받은 파일로 키와 ID 결합정보를 생성한다.
key1 = read.csv("금융정보_key.csv", header = T, fileEncoding = "utf-8")
key2 = read.csv("병원정보_key.csv", header = T, fileEncoding = "utf-8")
key = merge(key1, key2, by = 'combine_key')

# 결합처리는 두 기관에서 받은 정보를 KISA 결합키를 기준으로 병합한다.
dsn1 = read.csv("금융정보_결합요청.csv", header = T, fileEncoding = "utf-8")
dsn2 = read.csv("병원정보_결합요청.csv", header = T, fileEncoding = "utf-8")

library("dplyr")
dsn1 = rename(dsn1, ID.x = ID)
dsn2 = rename(dsn2, ID.y = ID)

dsn_combine = merge(key, dsn1, by = 'ID.x')
dsn_combine = merge(dsn_combine, dsn2, by = 'ID.y')
dsn_combine = select(dsn_combine, -c(ID.x, ID.y, combine_key))

write.csv(dsn_combine, file="결합결과정보.csv", row.names=FALSE, fileEncoding='utf-8')
```

가명정보결합.R

개인정보 가명·익명처리

End of Document

감사합니다.

