[4~6] 다음 글을 읽고 물음에 답하시오.

금융, 마케팅, 의료 등 다양한 분야에서 생성되는 빅데이터는 많은 경우 개인정보를 포함하고 있어 데이터를 활용하는 과정에서 민감한 개인정보가 유출될 가능성이 있다. 따라서 빅데이터 구축 과정에서 개인정보의 전부 또는 일부를 삭제하거나 대체함으로써 개인의 신원이 드러나지 않도록 하면서도 해당 데이터의 활용성을 최대한 유지할 수 있도록 하는 개인정보 비식별화 기술을 사용한다.

데이터 집합에서 정보를 표현하는 최소 단위를 속성이라고 하고 다양한 속성들의 조합으로 표현된 하나의 정보를 레코드라고 한다. 데이터 집합은 이 레코드들의 집합이다. 비식별화 기술은 속성을 식별자, 준식별자, 일반속성, 민감속성으로 구분한다. 주민번호와 같이 그 자체만으로도 누구인지 식별 가능한 속성이 식별자이다. 반면에 성별, 연령, 주소와 같이 개인에 대한 직접적인 식별은 불가능하지만 이들 속성이 결합하면 개인에 대한 식별이 가능해지는 속성을 준식별자라고 한다. 성별, 이름, 연령으로 구성되어 있는 원본 데이터 집합이 있을 때, 이름에서 성씨만을 남겨 비식별 데이터 집합을 만들었다고 하자. 비록 이름은 성만 남기고 가려져 있지만 '남성'이 유일하거나, 성이 '이씨'이면서 '35세'인 사람이 유일하다면, 원본에 이 두 사람이 포함된 사실을 알면서 이들 각자의 유일한 속성값 조합을 미리 알고 있는 사람은 특정 개인을 재식별할 수 있다. 일반적으로 개인정보는 개인의 여러 속성과 결합하여 사용된다. 익명 데이터라도 여러 속성과 결합하면 유일한 속성값 조합이 새로 생기게 되며 이에 따라 특정 개인이 재식별되는 불완전한 비식별 데이터 집합이 된다.

k-익명성은 특정 개인을 추정할 가능성을 1/k 이하로 낮추는 비식별화 기술로 원본 데이터 집합의 식별자나 준식별자 속성에 대해서만 마스킹, 범주화 등을 수행하여 유사한 준식별자 속성값들을 동일하게 만드는 작업을 수행한다. 마스킹은 '홍길동'을 '홍**'로 바꾸는 것이고 범주화는 '35세'를 '30대'로 바꾸는 식이다. 이렇게 만든 비식별 데이터 집합에서 준식별자 속성값들이 모두 동일한 레코드들의 집합을 동질집합이라고 하며 이때 레코드들의 수를 동질집합의 크기라고 한다. k-익명성은 비식별 처리로 만들어진 동질집합의 크기가 k개 미만인 동질집합을 모두 삭제하여 동질집합의 크기가 k개 이상 될 수 있도록 만든다. $k \geq 2$일 때 원본 데이터 집합에 있는 특정 개인의 준식별자를 미리 알고 있어도 비식별 데이터 집합만을 보고 원본의 특정 개인을 재식별하는 것은 불가능하다. 그러나 개인 추정 가능성은 존재한다. 즉 특정하고자 하는 개인이 속한 동질집합의 크기가 k일 때 이 특정 개인이 k명 중의 한 명임을 추정할 수 있으므로 1/k의 확률로 개인 추정이 가능하다.

k-익명성은 한 동질집합에 속하는 모든 레코드에서 준식별자 속성이 아닌 민감속성의 값이 모두 동일할 경우 해당 정보가 유출되는 단점이 있다. 민감속성은 병명, 수입 등 개인의 사생활과 관련된 속성을 의미한다. 예를 들어 동질집합이 3명의 레코드를 갖고 있고 이 3명이 모두 위암이라면, 홍길동이 동질집합의 3명 중 한 명이라는 사실을 아는 사람은 그중 누가 홍길동인지는 몰라도 홍길동이 위암이라는 사실을 정확히 알 수 있다. 이러한 k-익명성의 단점을 보완하기 위해 ℓ-다양성을 추가로 적용한다.

ℓ-다양성은 동질집합에서 민감속성이 최소 ℓ개의 서로 다른 속성값들을 갖도록 한다. 이 조건을 만족하지 못하는 동질집합은 비식별 데이터 집합에서 삭제한다. 앞의 예에서 동질집합의 병명 속성은 모두 '위암' 값만을 가지므로 ℓ-다양성을 만족하지 못하기 때문에 이 동질집합은 삭제된다.

비식별화 기술은 개인 식별 가능성은 낮출 수 있지만 정보 손실을 유발하기 때문에 구축된 빅데이터를 활용하는 측에서는 데이터의 가치가 낮아진다. 원본 유사도는 비식별 데이터 집합의 활용성을 나타내는 지표이며 원본 데이터 집합과 이를 비식별 처리한 비식별 데이터 집합이 얼마나 유사한지를 나타낸다. 이 지표는 레코드 잔존율과 레코드 유사도로 측정한다. 레코드 잔존율은 원본 데이터 집합의 총 레코드 수 대비 비식별 데이터 집합의 총 레코드 수를 백분율로 나타낸 지표이다. 한편 레코드 유사도는 원본 데이터 집합의 한 원본 레코드가 비식별 데이터 집합에 남아 있을 경우 원본 레코드와 비식별 레코드 쌍 간의 통계적 유사성을 0과 1 사이의 값으로 표현한 지표이다.

4. 윗글의 내용과 일치하지 <u>않는</u> 것은?

① 휴대전화 번호는 일반적으로 식별자에 해당한다.

② 민감속성은 범주화와 마스킹으로 비식별 처리를 한다.

③ 레코드 유사도가 높을수록 개인정보 식별 가능성은 커진다.

④ 준식별자들의 조합만으로도 특정 개인이 식별되는 경우가 있다.

⑤ 레코드는 식별자와 준식별자 이외에도 다양한 속성으로 구성된다.

5. k-익명성에 대한 추론으로 가장 적절한 것은?

① k를 낮추면 재식별 가능성과 레코드 잔존율 모두 감소한다.

② k를 낮추면 동질집합의 수는 증가하고 동질집합은 서로 크기가 같아진다.

③ k를 높이면 재식별 가능성은 증가하고 동질집합의 레코드 수는 감소한다.

④ k를 높이면 동질집합의 수는 감소하고 동질집합의 민감속성값은 모두 같아진다.

⑤ k를 변경했더니 레코드 잔존율이 증가했다면 동질집합의 크기들 중 최솟값은 작아진다.

6. 윗글을 바탕으로 <보기>의 사례를 이해할 때, ㄱ~ㄷ 중 맞는 것만을 있는 대로 고른 것은?

<보 기>

다음 표는 한 쇼핑몰의 고객 관리 원본 데이터 집합이다. 여기서 우편번호, 연령, 성별은 준식별자이고, 구매 수준은 민감속성이다. (a)와 (b) 방식으로 각각 비식별화 기술을 적용하고자 한다.

No.	우편번호	연령	성별	구매 수준
1	15093	25	남	상
2	15002	28	남	상
3	15000	21	여	중
4	15090	22	남	중
5	13851	45	여	하
6	13852	42	남	상

(a) 우편번호를 1509*, 1385*, 1500*로 표시하고, 연령은 40세 미만과 40세 이상으로 나누고, 성별은 마스킹한 후 k-익명성과 ℓ-다양성을 적용한다.

(b) 우편번호를 150**, 138**로 표시하고, 연령은 40세 미만과 40세 이상으로 나누고, 성별은 마스킹한 후 k-익명성과 ℓ-다양성을 적용한다.

ㄱ. (a)보다 (b)의 레코드 잔존율이 크고 (a)와 (b)의 k 값이 같고 (a)와 (b)의 ℓ 값도 같다면, (a)의 동질집합의 수는 0이다.

ㄴ. (a)와 (b)의 레코드 잔존율이 100%라면, (a)와 (b)는 k 값이 같고 ℓ 값도 같으며 동질집합의 수도 같다.

ㄷ. 레코드 잔존율이 (a)는 100%이고 (b)는 50% 이상 100% 미만이라면, (a)의 k 값이 (b)의 k 값보다 작고, (a)와 (b)의 ℓ 값은 서로 같다.

① ㄱ ② ㄴ ③ ㄱ, ㄷ

④ ㄴ, ㄷ ⑤ ㄱ, ㄴ, ㄷ