Técnicas de anonimidad

Pensadas para publicar datasets sin exponer datos sensibles

Personal Identifiable Information (PII): Datos que se pueden usar para identificar a una pesona.

- **Identificadores**: Atributos únicos de individuos
- **Pseudo-Identificadores o Cuasi-identificadores**: Datos que de por sí no significan nada pero si se agrupan con más indormación pueden ser identificables

Aproximaciones típicas

- Data Masking: Se oculta o elimina cierta información para que no se puedan deducir los valores originales
  - Se tapa con asteriscos
  - Se mestran los últimos dígitos
- **Pseudoanonimización**: Se usan pseudónimos para dificultar la identificación de un individuo.
- **Generalización**: En vez de publicar el dato, se publica un rango, por ejemplo, en vez de poner 7 se pone el rango [1, 10]
- **Data Swapping**: Se intercambian filas y sus valores.
- **Perturbación de los datos**: Se aplica respuesta aleatorizada a algunos de los datos
- **Datos Sintéticos**: Se piublican unos dato sintéticos que representan el dataset simulando datos reales.

K-Anonimidad

Se forman grupos de K elementos de forma que las filas comparten k cuasi-identificadores. Se busca un equilibrio entre K y el nivel de aononimidad que se quiere obtener. Para formar los grupos, hay que mirar por que atrbutos podemos realizar las grupacoones. Hay que evitar que se formen pocos grupos o agrupar por atributos sensibles. Hay que identificar bien pseudoidentificadores y datos sensbibles.

- Los pseudoidentificadores deben ser anonimizados siempre
- Los datos críticos del dataset no deben ser modificados.

Si los datos están demasiado dispersos se pierde demasiada información ya que hay que añadir demasiado ruido. Si los datos están muy juntos tampoco es efectivo.

- Para cualquier registro dado (fila) hay al menos otros k-1 registros que comparten el mismo set de atributos cuasi-identificadores
- El Valor K se suele usar para medir la privacidad
  - Cuanto más grande sea K, más difícil es deanonimizar los datos
  - La utilidad de los datos disminuye cuando K crece.
- Los cuasi-identificadores y los atributos sensibles deben ser distinguibles de forma que no revelen información sobre atributos ya anonimizados.

K-Anonimidad: Ataques

- Homogeneidad: Explota el hecho de que todos los registros dentro del mismo grupo tienen el mismo valor para los datos sensibles.
- Conocimiento del Transfondo: Explota alguna información externa que se puede utilizar para identificar un individuo del dataset.

Ejemplo de K-Anonimidad

Imagina que tenemos el siguiente dataset:

Nombre	Codigo Postal	Edad	Género	Religión	Enfermedad
John	15846	15	M	C	Covid
Emily	25105	41	F	C	Gripe
Sarah	15834	18	F	M	Cancer
Jeremy	25504	25	M	C	Gripe
Carl	15894	22	M	C	Infección
Laura	15833	31	F	H	Cancer
Michael	25974	58	M	C	Covid
Hank	25785	29	M	B	Gripe
Kim	15874	62	F	H	Corazón

En este dataset podemos identificar los siguientes tipos de información:

- Identificadora:
  - Nombre
- Quasi-Identificador:
  - Código Postal
  - Edad
  - Genero
- Atributos sensibles
  - Religión
  - Enfermedad

Al clasificar los atributos podemos saber a cuales les podemos aplicar pseudoanonimización o supresión. En este caso, como es un dataset médico, el atributo sensible religión puede ser suprimido. Por otro lado debemos pseudoanonimizar los nombres, cambiando estos por IDs, obteniendo como resultado el siguiente dataset:

ID	Codigo Postal	Edad	Género	Religión	Enfermedad
1	15846	15	M	*	Covid
2	25105	41	F	*	Gripe
3	15834	18	F	*	Cancer
4	25504	25	M	*	Gripe
5	15894	22	M	*	Infección
6	15833	31	F	*	Cancer
7	25974	58	M	*	Covid

ID	Codigo Postal	Edad	Género	Religión	Enfermedad
8	25785	29	M	*	Gripe
9	15874	62	F	*	Corazón

Tras eso, procedemos a aplicar una K-Anonimidad donde K=2, por lo que debemos crear grupos de al menos tamaño 2 a los que aplicaremos generalización:

- Código Postal: Se suprimen los últimos números, reemplazándolos con asteriscos.
- Edad: Se suprime la edad y se cambia por los siguientes valores:
  - <30: Tiene menos de 30 años pero más de 25
  - >40: Tiene más de 40 años
  - <25: Tiene menos de 25 años
  - >30: Tiene más de 30 años, pero menos de 40
- Género: En los grupos donde se detecte que el género puede ser hasta cierto punto identificativo, se suprime.

Finalmente realizamos una agrupación de dos bloques en función al código postal (25xxx y 158xx) y luego realizamos pequeños grupos de elementos por edad. Como resultado obtenemos el siguiente dataset con 2-anonimidad:

ID	Codigo Postal	Edad	Género	Religión	Enfermedad	Agrupación
4	25xxx	<30	M	*	Gripe	Grupo 1
8	25xxx	<30	M	*	Gripe	
7	25xxx	>40	*	*	Covid	Grupo 2
2	25xxx	>40	*	*	Gripe	
5	158xx	<25	*	*	Infección	Grupo 3
1	158xx	<25	*	*	Covid	
3	158xx	<25	*	*	Cancer	
7	158xx	>30	F	*	Cancer	Grupo 4
9	158xx	>30	F	*	Corazón	

L-Diversidad

Se busca que en cada grupo, el atributo sutituible tenga al menos L valores diferentes. Cada uno de los grupos K-Anonimos debe tener al menos l valores diferentes del atributo sensible para que sea más robusto contra filtraciones de privacidad.

- Cuanto más grande sea el valor de L, más difícil es inferir cosas a partir de los datos de cada grupo.
- La L-Diversidad puede distorsionar la distribución de los datos, sesgando alguno de los grupos.

Teniendo en cuenta el ejemplo anterior para k-anonimidad, se cumple la L-Diversidad para L=2 en todos los grupos menos en el de las edades <30, donde podemos observar que los dos valores del campo sensible (gripe) que hay son iguales. Para cumplir con la 2-Diversidad sería necesario realizar cambios como juntar este grupo con otro o proceder a reagrupar.

OJO: Si no tenemos cuidado podemos acabar inutilizando el dataset.

Problemas de la L-Diversidad

- Puede filtrar datos si se tiene cierto conocimiento externo
- Vulnerable a ataques de skeweness debido a una distribución desbalanceada.

Ejemplo de L-Diversidad

Tomando como base el dataset anterior con 2-Anonimidad:

ID	Codigo Postal	Edad	Género	Religión	Enfermedad	Agrupación
4	25xxx	<30	M	*	Gripe	Grupo 1
8	25xxx	<30	M	*	Gripe	
7	25xxx	>40	*	*	Covid	Grupo 2
2	25xxx	>40	*	*	Gripe	
5	158xx	<25	*	*	Infección	Grupo 3
1	158xx	<25	*	*	Covid	
3	158xx	<25	*	*	Cancer	
7	158xx	>30	F	*	Cancer	Grupo 4
9	158xx	>30	F	*	Corazón	

Queremos aplicar una L-Diversidad donde L=2. Como mencionamos antes, esto se cumple para todos los grupos menos para el grupo 1, donde los valores sensibles son iguales. Para Hacer que este dataset cumpla con la 2-Diversidad vamos a juntar el grupo 1 con el grupo 2 realizando las siguientes operaciones adicionales:

- Suprimimos las Edades
- Suprimimos el género

De esta forma se crea un nuevo “Grupo 1-2” el cual va agrupado por el código postal en vez de por la edad:

ID	Codigo Postal	Edad	Género	Religión	Enfermedad	Agrupación
4	25xxx	*	*	*	Gripe	Grupo 1-2
8	25xxx	*	*	*	Gripe	
7	25xxx	*	*	*	Covid	
2	25xxx	*	*	*	Gripe	
5	158xx	<25	*	*	Infección	Grupo 3
1	158xx	<25	*	*	Covid	
3	158xx	<25	*	*	Cancer	
7	158xx	>30	F	*	Cancer	Grupo 4
9	158xx	>30	F	*	Corazón	

T-Proximidad

Queremos conseguir que la distribución cumpla con un umbral de distancia:

- La distancia entre las distribuciones debe ser menor o igual que T

$Dist(X,Y)≤t$

- Hay varias métricas que se pueden utilizar para medir la distancia entre 2 distribuciones. Una de las más usadas es EMD (Earths Mover's Distance)
- Dadas 2 distribuciones X e Y, con probabilidades X_i e Y_i para el elemento i de cada set, el EMD se puede definir de la siguiente forma para atributos categóricos:

$EMD(X,Y)= 1/2 ∑_{i=1}^m|X_i-Y_i|$

Un ejemplo del cálculo de T-Proximidad categórico sería el siguiente:

$Y = [Gripe, Covid, Gripe, Cancer, Gripe, Covid]$ → $Y = {Gripe, Covid, Cancer}$; Y={$\frac{1}{2}$,$\frac{1}{3}$, $\frac{1}{6}$}

$X = [Gripe, Covid, Cancer]$ → $X= {Gripe, Covid, Caner}$; X = {$\frac{1}{3}$, $\frac{1}{3}$, $\frac{1}{3}$}

$EMD(X,Y) = \frac{1}{2}[|\frac{1}{3} - \frac{1}{2}| + |\frac{1}{3} - \frac{1}{3}| + |\frac{1}{3} - \frac{1}{6}|] = \frac{1}{2}[\frac{1}{6} + \frac{1}{6}] = 0.1667$

Ejemplo de T-Proximidad

Para este ejemplo vamos a tomar la siguiente tabla como base:

ID	Codigo Postal	Edad	Género	Religión	Enfermedad
1	15846	15	M	*	Covid
2	25105	41	F	*	Gripe
3	15834	18	F	*	Cancer
4	25504	25	M	*	Gripe
5	15894	22	M	*	Infección
6	15833	31	F	*	Cancer
7	25974	58	M	*	Covid
8	25785	29	M	*	Gripe

ID	Codigo Postal	Edad	Género	Religión	Enfermedad
9	15874	62	F	*	Corazón

Comenzamos mirando la **distribución global de las enfermedades**:

- Gripe = $\frac{3}{9}$
- Covid = $\frac{2}{9}$
- Cancer = $\frac{2}{9}$
- Infección = $\frac{1}{9}$
- Corazón = $\frac{1}{9}$

Ralizamos **agrupaciones por rango de edades**:

- Grupo 1 (15 a 25 años):
  - Covid
  - Cancer
  - Gripe
  - Infección
- Grupo 2 (26 a 35 años)
  - Cancer
  - Gripe
- Grupo 3 (40 a 65 años)
  - Gripe
  - Covid
  - Corazón

Cálculo de la distribución de enfermedades por grupo:

- Grupo 1:
  - Gripe = $\frac{1}{4}$
  - Covid = $\frac{1}{4}$
  - Cancer = $\frac{1}{4}$
  - Infección = $\frac{1}{4}$
  - Corazón = $\frac{0}{4}$
- Grupo 2:
  - Gripe = $\frac{1}{2}$
  - Covid = $\frac{0}{2}$
  - Cancer = $\frac{1}{2}$
  - Infección = $\frac{0}{2}$
  - Corazón = $\frac{0}{2}$
- Grupo 3:
  - Gripe = $\frac{1}{3}$
  - Covid = $\frac{1}{3}$
  - Cancer = $\frac{0}{3}$
  - Infección = $\frac{0}{3}$
  - Corazón = $\frac{1}{3}$

Medir Distancia (EMD)

En este caso tenemos los siguientes datos:

- X_i = Probabilidad global

- Y_i = Probabilidad del grupo