COENCOM0108 - Privacidade de Dados
Trabalho 01 - k-Anonimato + l-Diversidade
Semestre: 2025.1

Prof. André Luís
andre.luis@unilab.edu.br

18 de junho de 2025

1 Objetivo

O trabalho consiste em implementar um algoritmo que anonimize um conjunto de dados contra ataques de ligação ao registro e atributo, atendendo, ao mesmo tempo, os modelos de privacidade k-anonimato e l-diversidade. O trabalho deverá implementar o modelo k-anonimato por meio da **generalização** de valores de atributos, como descrito no artigo *SWEENEY, Latanya. k-anonymity: A model for protecting privacy. International journal of uncertainty, fuzziness and knowledge-based systems, v. 10, n. 05, p. 557-570, 2002*[1] [3]. Você deve escolher uma técnica de construção das classes de equivalência para fazer a anonimização e seguir as instruções abaixo para aderir ao k-anonimato na geração dos datasets anonimizados. Especificamente sobre o modelo l-diversidade, o aluno deverá implementar o modelo por meio da **diversidade** de valores do atributo sensível, como descrito no artigo *MACHANAVAJJHALA, Ashwin et al. l-diversity: Privacy beyond k-anonymity. ACM Transactions on Knowledge Discovery from Ddata (TKDD), v. 1, n. 1, p. 3-es, 2007*[2] [1]. A técnica de diversidade complementa o modelo k-anonimato, portanto, deve ser aplicada sobre um conjunto de dados k-anonimizado. Você deve entender o princípio da l-diversidade e aplicar a técnica para cada uma das classes de equivalência. Para fazer a anonimização, você deve seguir as instruções abaixo para aderir à l-diversidade na geração dos datasets anonimizados.

2 Especificação

Considere o conjunto de dados "**dados_covid-ce.csv**". Você deve recuperá-lo por meio do link: `https://tinyurl.com/unilab-priv-dados-covid-ce`. Este dataset contém os atributos, `nome`, `cpf`, `localidade`, `data_nascimento` e `raca_cor`, os quais representam o nome, CPF (Cadastro de Pessoa Física), localidade, data de nascimento e raça do indivíduo, respectivamente. Os atributos são categorizados da seguinte maneira:

- **Identificadores explícitos:** `nome`; `cpf`.
- **Semi-identificadores:** `localidade` (bairro/cidade/estado); `data_nascimento` (dia/mês/ano, no formato: dd/mm/aaaa).

[1] `https://www.worldscientific.com/doi/pdf/10.1142/S0218488502001648`
[2] `https://personal.utdallas.edu/~muratk/courses/privacy08f_files/ldiversity.pdf`

- **Sensíveis:** `raca_cor` (PARDA, AMARELA, BRANCA, PRETA ou INDÍGENA).

O aluno deverá construir duas hierarquias de generalização: (1) para o atributo `data_nascimento` e (2) para o atributo `localidade`. A hierarquia (1) deve ter três níveis, do mais específico para o mais geral, a saber: dia/mês/ano, mês/ano e ano. A hierarquia (2) deve ter três níveis, do mais específico para o mais geral, a saber: bairro/cidade/estado, cidade/estado e estado.

O trabalho terá duas etapas: (1) aplicação da técnica de k-anonimato e (2) aplicação da técnica de l-diversidade.

O programa implementado deverá receber como entrada os valores de k e l, podendo ser $k = \{2, 4, 8\}$ e $l = \{2, 3, 4\}$. Para cada uma das configurações de k e l, deverá ser gerado um dataset anonimizado com o nome "**dados_covid-ce_k_l.csv**". Por exemplo: dados_covid-ce_2_2 e dados_covid-ce_4_3.csv e dados_covid-ce_8_4.csv para os valores de (k, l) igual a (2, 2), (4, 3) e (8, 4), respectivamente. A mesma lógica vale para as demais combinações. Cada dataset anonimizado deve conter os mesmos atributos do dataset original, respeitando a mesma ordem: (`nome`, `cpf`, `localidade`, `data_nascimento`, `raca_cor`) e seus respectivos registros anonimizados. Lembre-se que o valor de l sempre deverá ser menor ou igual ao valor de k. Por exemplo, para $k = 2$, o único valor possível de l é 2. Para $k = 4$, l poderá assumir os valores 2, 3 e 4, e assim sucessivamente.

2.1 k-Anonimato

Construa o reticulado de generalizações formado através das hierarquias de generalização de domínio de cada atributo semi-identificador. Para gerar as classes de equivalência, generalize os atributos baseando-se no reticulado, de forma que seja necessária a mínima generalização possível para gerar classes de equivalência de tamanho k.

Após a ativação da técnica, calcule a **precisão**, cuja fórmula pode ser encontrada no artigo *SWEENEY, Latanya. Achieving k-anonymity privacy protection using generalization and suppression. International Journal of Uncertainty, Fuzziness and Knowledge-Based Systems, v. 10, n. 05, p. 571-588, 2002*[3] [2] ou vista na Equação 1, conforme abaixo:

$$\texttt{Precisão} = 1 - \texttt{PerdaInformação} \tag{1}$$

A definição de `PerdaInformação` é apresentada abaixo:

$$\texttt{PerdaInformação} = \frac{1}{N \cdot M} \sum_{i=1}^{N} \sum_{j=1}^{M} \frac{\mathit{nível}(A_{ij})}{\mathit{profundidade\_hierarquia}(A_j)},$$

onde:

- N: Número de registros (linhas) no conjunto de dados anonimizado;
- M: Número de atributos semi-identificadores;
- i: Índice do registro (de 1 até N);
- j: Índice do atributo semi-identificador (de 1 até M);
- A_{ij}: O valor generalizado do j-ésimo semi-identificador no i-ésimo registro;
- $\mathit{nível}(A_{ij})$: Representa o nível de generalização do valor A_{ij} na hierarquia de generalização do atributo A_j:

[3] `https://dataprivacylab.org/dataprivacy/projects/kanonymity/kanonymity2.pdf`

- Assumimos que o nível 0 é o mais específico (o valor original);
- O nível aumenta à medida que o valor se torna mais genérico;
- Se um valor é suprimido, o mesmo pode ser considerado do nível mais alto da hierarquia.

- *profundidade_hierarquia*(A_j): Representa o tamanho, ou profundidade, da hierarquia do atributo A_j. Ou seja, é o número total de níveis na hierarquia, do nível mais específico (nível 0) ao nível mais geral (nível mais alto).

Note que o valor da precisão varia no intervalo entre [0, 1], sendo 0 a precisão mínima, e 1 a precisão máxima. Em outras palavras, quanto mais próximo de 1 for a precisão, mais próximo do dado original o dado anonimizado estará e mais útil o dado anonimizado será.

Calcule também o tamanho médio das classes de equivalência através da razão entre a quantidade total de registros do dataset e a quantidade de classes de equivalência geradas pelo k-anonimato, para cada k.

Por fim, apresente os tamanhos das classes de equivalência para cada k. Para cada k, grave os resultados em uma planilha e complemente a apresentação com um histograma contendo as top-y classes de equivalência, onde y é a quantidade de maiores classes de equivalência que serão exibidas. Escolha um valor razoável para y, nem muito baixo para que as informações apresentadas no histograma não sejam muito superficiais, e nem muito alto para que o histograma não fique muito poluído.

2.2 l-Diversidade

Para cada classe de equivalência gerada pelo k-anonimato, você deve garantir que haverá pelo menos l diferentes valores para o atributo sensível. Construa um histograma onde o eixo x representa o número de valores distintos do atributo sensível nas classes de equivalência, enquanto que o eixo y representa a respectiva frequência dos valores de x.

3 Requisitos

- Linguagens: Python (recomendado), Java ou C/C++;
- Grupos de no máximo 3 pessoas;
- Preparar uma demonstração para explicar, mostrar o seu programa e os resultados durante a aula de apresentação;
- Comprimir o seu projeto, contendo: código-fonte, os datasets anonimizados, os gráficos (histogramas) e um arquivo *Readme.txt* (contendo a descrição do projeto e os integrantes do grupo).

4 Entrega e Apreentação

- A entrega deverá ser feita até a data limite (08:00h do dia 25/06/2025) através de uma tarefa cadastrada no SIGAA;
- A apresentação será realizada durante o horário de aula do dia 25/06/2025.

5 Avaliação

Na avaliação serão considerados os seguintes indicadores:

- **Corretude** do programa;
- **Precisão** pela comparação do dataset original com o dataset anonimizado;
- Clareza na **explicação** do programa durante a demonstração;
- **Pontualidade** da entrega e **documentação/qualidade** do código-fonte.

Referências

[1] Ashwin Machanavajjhala, Daniel Kifer, Johannes Gehrke, and Muthuramakrishnan Venkitasubramaniam. l-diversity: Privacy beyond k-anonymity. *Acm transactions on knowledge discovery from data (tkdd)*, 1(1):3–es, 2007.

[2] Latanya Sweeney. Achieving k-anonymity privacy protection using generalization and suppression. *International Journal of Uncertainty, Fuzziness and Knowledge-Based Systems*, 10(05):571–588, 2002.

[3] Latanya Sweeney. k-anonymity: A model for protecting privacy. *International journal of uncertainty, fuzziness and knowledge-based systems*, 10(05):557–570, 2002.