Privacidade de Dados - 2025-2

Trabalho 1 - Tem alguém aí?

Javam Machado - Malu Maia

1 Objetivo

O trabalho consiste em implementar um algoritmo capaz de realizar um ataque de ligação entre diferentes bases de dados públicas disponibilizadas em portais de transparência do país. Os conjuntos contém dados da área da saúde e esfera política. O intuito do exercício é demonstrar, em caráter acadêmico e experimental, os riscos de reidentificação de indivíduos. A partir desse exercício, pretende-se evidenciar como informações sensíveis, como possíveis condições de saúde de pessoas públicas, podem ser inferidas quando dados abertos não recebem tratamento adequado de anonimização.

2 Especificação

Considere o conjunto de dados "**dados_covid-ce_trab02.csv**". Você deve recuperá-lo por meio do classroom da disciplina. Este dataset contém vários atributos relacionados a saúde de um indivíduo cujos identificadores explícitos são anonimizados.

Os atributos de interesse contém **informações demográficas** (`municipioCaso`, `bairroCaso`, `sexoCaso`, `dataNascimento`, `racaCor`) de município, bairro, gênero e data de nascimento do indivíduo, respectivamente, e **informações sensíveis** do estado de saúde do indivíduo (14 atributos de que identificam se o paciente possui determinada comorbidade e `resultadoFinalExame`, que identifica se o paciente teve estava com COVID na data da coleta).

Os atributos são categorizados da seguinte maneira:

- Atributos semi-identificadores: `municipioCaso`, `bairroCaso`, `sexoCaso`, `dataNascimento`, `racaCor`;
- Atributos sensíveis: `resultadoFinalExame` e 14 atributos que iniciam com `comorbidade`.

Considere também o conjunto de dados públicos "**consulta_cand_2020_CE.csv**" acerca dos candidatos da eleição de 2020 do estado do Ceará. Você deve recuperá-lo através do classroom da disciplina. Este dataset contém informações acerca dos indivíduos que se candidataram politicamente no ano de 2020.

Os atributos de interesse são os **identificadores explícitos** (`NM_CANDIDATO`, `NR_CPF_CANDIDATO`), que representam o nome e CPF do candidato, respectivamente, e **informações demográficas** (`NM_UE`, `DT_NASCIMENTO`, `DS_GENERO`, `DS_COR_RACA`) com nome do município, data do nascimento, gênero e raça, respectivamente.

Os atributos são categorizados abaixo:

- Atributos identificadores: `NM_CANDIDATO`, `NR_CPF_CANDIDATO`;

- Atributos semi-identificadores: `NM_UE`, `DT_NASCIMENTO`, `DS_GENERO`, `DS_COR_RACA`.

Sua tarefa é fazer um cruzamento entre os dados disponibilizados publicamente utilizando os atributos semi-identificadores disponíveis em ambos os conjuntos de dados (data de nascimento, raça/cor, gênero e município) e criar um novo conjunto de dados com possíveis informações coletadas a partir do cruzamento desses dados.

Obs.: É importante salientar que esses resultados não garantem 100% de certeza no ataque visto que não temos o conhecimento externo de que os candidatos fizeram o teste de COVID no sistema público de saúde e contém características únicas na combinação dos semi-identificadores. O resultado coletado a partir do ataque é uma inferência.

3 Requisitos

- Linguagens: C++ ou Python;
- Trabalho em duplas;
- Comentar cada função implementada. Escreva um *Readme.txt* descrevendo o projeto;
- Comprimir o seu projeto (código-fonte e executável), o dataset *resultado_ataque.csv* obtido através do ataque e o *Readme.txt* em um único pacote e submeter via **Classroom**;
- O trabalho deverá ser entregue até as 10h da segunda-feira, dia 24/09/2025.

4 Avaliação

Na avaliação serão considerados os seguintes indicadores:

- **Corretude** do programa;
- **Pontualidade** e **documentação/qualidade** do código-fonte.