LINMA2472 – Algorithms in Data Science

HW 3 – module "Privacy"

Please prepare a written report with appropriate figures or tables based on the results from the assignment.

Assignment:

In this assignment, you will be carrying out data anonymization to reach a satisfying balance between privacy and utility on a dataset that is to be publicly released and used for three use cases. You will analyse in depth the level of privacy and utility that can be reached for these use cases, and make recommendations. We will also ask you to send us your anonymised dataset.

We are providing you a dataset and ask you to come up with and discuss an anonymisation strategy for 2 different use cases (see below). We expect you to write a report explaining your choice of methods and your analysis of the performances of each method on each of these use cases, both in terms of privacy and utility of the resulting data.

Specifically, we want you to design anonymization methods as seen in the lecture (or any other custom one that you find interesting) in order to protect the privacy of users in the dataset. By method, we mean the guarantee you're trying to reach (*l*-diversity, *k*-anonymity, *etc.*), the choice of quasi-identifiers and sensitive information, the choice of parameters (*l*, *k* , *etc.*) and of algorithm to reach this guarantee (suppression, generalization, *etc.*). You have complete freedom on the method you use, as long as you justify your choice. For example, you can decide to address *multiple* use cases with *one* anonymization method or propose *one* anonymization method for *each* use case.

We are mostly interested in your reasoning and how you reach a balance: protecting privacy *vs.* utility of the dataset. ***It is very important that every choice you make is clearly justified and supported by a thorough analysis***. For each of the anonymization method you design, you should think of any possible attack (*e.g.,* different auxiliary information known to the adversary) if you were to publish the anonymized version(s) of your dataset online, and evaluate the risk they represent. The level of utility for each use case should be considered and justified as well. You could look at, *e.g.*, the distribution of the equivalence classes before and after the anonymization process, the distribution of the sensitive attributes in the equivalence classes, or the difference of entropy between the original and the anonymized dataset; or any other metric that you deem appropriate.

The dataset contains 2000 records. The columns are: *id, gender, dob* (date of birth), *zipcode, education (level), employment_status, children, marital_status, ancestry,*

number_vehicles, commute_time (average daily commuting time, in hours)*, accommodation* (type of housing)*, disease*

The 2 use cases for the dataset are:

- To study the impact of stress and high-pressure environments on one's health.
- To decide where to build new hospitals, and which departments (radiology, neurology, pulmonology…) to open in these, as the UK government has just allocated the money to build 5 new hospitals in the UK.

Note: you are not asked to provide solutions for the use cases! You just need to make sure that the anonymization method you choose doesn't make the data completely useless for the specific use case. Remember that most of the time, when the government publishes data as open-data they don't know in advance what people (such as researchers, data scientists) might want to use the data for. In fact, researchers don't often know what analyses they are going to do in advance either.