LINMA2472 – Algorithms in Data Science

HW 3 – “Privacy”

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

Assignment:

As privay officers of a large british hospital, you are tasked with providing two different datasets for two different use cases (detailed below) in such a way that they preserve the anonymity of the subjects whose data is contained in the original dataset while still providing a decent source of information for the use cases.

Specifically, we want you to design anonymisation 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 guarantees 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 these guarantees (suppression, generalisation, *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* anonymisation method or propose *one* anonymisation 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 an analysis of the data***. For each of the anonymisation 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 anonymised 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, say, the distribution of the equivalence classes before and after the anonymisation process, the distribution of the sensitive attributes in the equivalence classes, or the difference of entropy between the original and the anonymised 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), *diseas.*

The 2 use cases for the dataset are:

- For sociologists, to study the impact of stress and high-pressure environments on one’s health. They are looking for factors in people’s lives that could correlate to certain diseases or conditions.
- For the US department of Health and Human services chairman, to decide where to build new hospitals, and which departments (radiology, neurology, pulmonology, oncology…) to have in these, as the US federal government has just allocated money to build 5 new hospitals across the country.

Notes:

-You are not asked to provide solutions for the use cases! You just need to make sure that the anonymisation 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.

-Remember that the two datasets will be simultaneously available and that potential attackers could use the data from both datasets and make some aspects of your anonymisation work redundant. You should keep this in mind while you proceed with anonyimising your data.

Instructions:

You should hand in one report (pdf format, no more than ten pages with figures included, annexes are allowed but you will be judged based upon the material in the main part of the report), one code file (jupyter notebook preferably) as well as one dataset (in .csv format), all zipped in one file under the name "group_*x*_project3-*y_1*-*y_2*-*y_3*.zip", where *x* is your group number and *y_i* are your respective family names. Please also make sure that every member of the group is signed up to the same group on the group choice for homework 3 on moodle otherwise ou may get a 0 overall.