University of Tartu
Institute of Computer Science
Cybersecurity Curriculum

Joosep Parts

Homework 4: Data anonymization / pseudonymization

Privacy-preserving Technologies LTAT.04.007

Tartu 2023

List of Abbreviations and Terms

DA Data Attribute

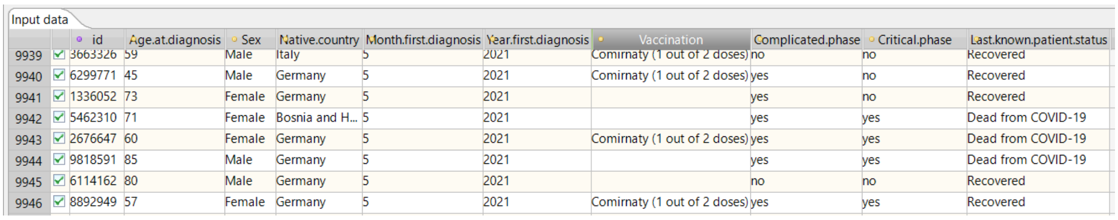
DS Data Subject

SA Sensitive Attribute

Contents

1 Analysis

Before choosing any options the given set was pre-analyzed using analyze utility to determine the Sensitive Attribute (SA) of the set. Anonymization options were redefined during the process many times to yeild desired result, but for the baseline settings the first chosen settings were synthesised quite subjectively. Data set is a sample set of COVID-19 health related data, see Figure 1 for example.



Input data

		id	Age.at.diagnosis	Sex	Native.country	Month.first.diagnosis	Year.first.diagnosis	Vaccination	Complicated.phase	Critical.phase	Last.known.patient.status
9939	☑	3663326	59	Male	Italy	5	2021	Comirnaty (1 out of 2 doses)	no	no	Recovered
9940	☑	6299771	45	Male	Germany	5	2021	Comirnaty (1 out of 2 doses)	yes	no	Recovered
9941	☑	1336052	73	Female	Germany	5	2021		yes	no	Recovered
9942	☑	5462310	71	Female	Bosnia and H...	5	2021		yes	yes	Dead from COVID-19
9943	☑	2676647	60	Female	Germany	5	2021	Comirnaty (1 out of 2 doses)	yes	yes	Dead from COVID-19
9944	☑	9818591	85	Male	Germany	5	2021		yes	yes	Dead from COVID-19
9945	☑	6114162	80	Male	Germany	5	2021		no	no	Recovered
9946	☑	8892949	57	Female	Germany	5	2021	Comirnaty (1 out of 2 doses)	yes	yes	Recovered

Figure 1. Preview of the sample set

1.1 Levels of generalisation

Before choosing levels of generalisation all fields distribution was observed to determine possible ways to generalisation. For example countries distribution was observed in Figure 3 and it was seen that smaller countries should be concatenated due to their small sample size. And from observing age (see Figure 2), I saw that most records existed for older ages, so it would make sense to group younger and middle-aged people wider than the older ages. Few fields were not generalised. If we transformed all the values of the given set we would lose the context greatly making the output data quite useless. Since we dont know what the Data Subject (DS) may not wish to be disclosed explicitly, I have not generalised some Data Attribute (DA), see Table 1. But for other DA following suggestions for data anonymization as relevant guidelines suggest on doing [1].

As seen in Table 1, for the “id“ field, a single level of generalisation is applied to suppress or replace all unique identifiers, thus providing strong privacy protection. Age at diagnosis, a quasi-identifier, is divided into 5-year intervals with 6 levels to maintain data utility while reducing the risk of re-identification. A 5-year range is a common choice in literature [2, p. 672], as it achieves a balance between privacy and utility. I would have wanted to group lower and middle aged people in a wider range than older ages, but I could not find the options to do so. To minimize the risk of re-identification for the “Sex“ attribute without significantly affecting data utility, we group genders while preserving their distribution.

For the “Native.country“ field, we employ priority grouping with three descending levels based on factors such as geography or demographics to ensure sufficient generaliza-

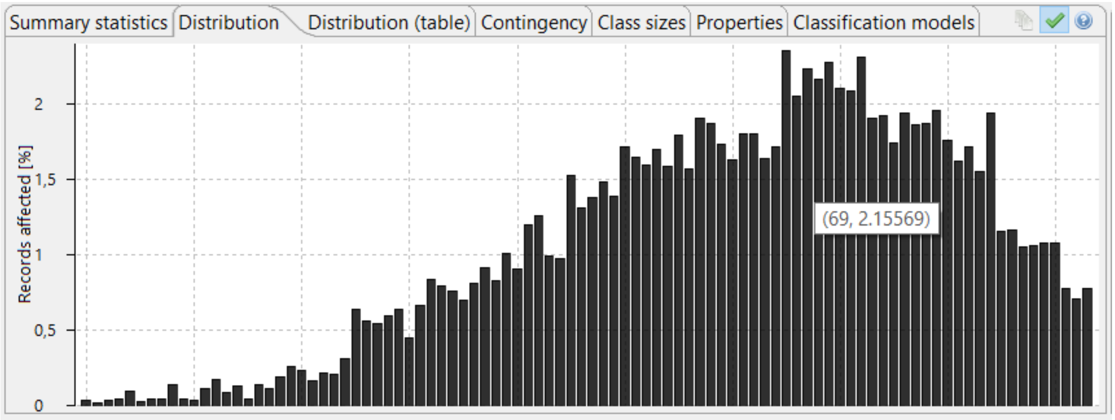


Figure 2. Initial age distribution

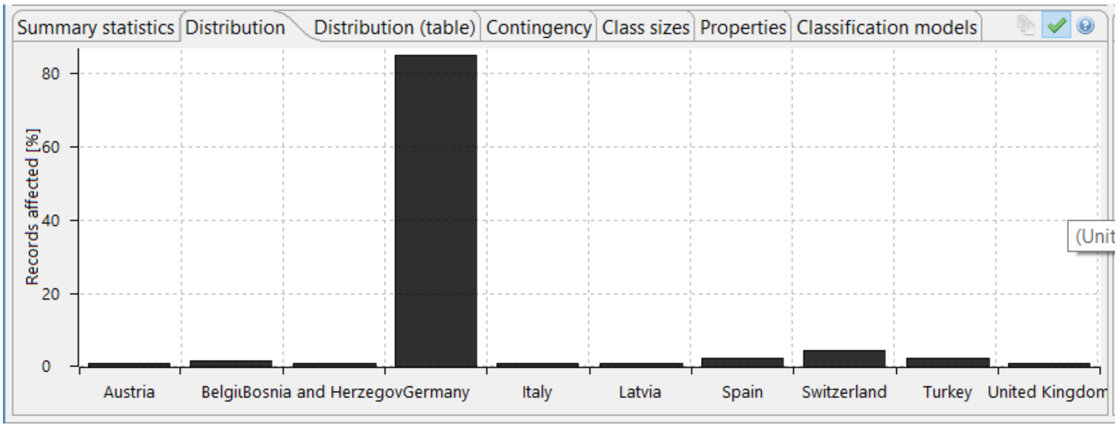


Figure 3. Initial countries distribution

tion without excessive information loss. The “Month.first.diagnosis“ attribute is not generalized, preserving its full utility under the assumption that the month of diagnosis alone does not significantly contribute to re-identification risk when combined with other anonymized attributes.

The “Year.first.diagnosis“ attribute is generalised using intervals with two levels, while the “Vaccination“ and “Last.known.patient.status“ fields are subjected to priority grouping with three levels, ensuring a balance between privacy protection and data utility. Finally, no generalization is applied to the “Complicated.phase“ and “Critical.phase“ attributes, as they are deemed not to significantly increase re-identification risk in combination with other anonymized attributes

Table 1. Chosen Generalizations

Generalization	Levels	Field
–	Level 1	id
Intervals (5y)	6 levels	Age.at.diagnosis
Ordering	1 levels	Sex
Priority grouping descending	3 levels	Native.country
–	–	Month.first.diagnosis
Intervals	2 levels	Year.first.diagnosis
Priority grouping	3 levels	Vaccination
–	–	Complicated.phase
–	–	Critical.phase
Priority grouping	3 levels	Last.known.patient.status

1.2 Guarantees the privacy models offer

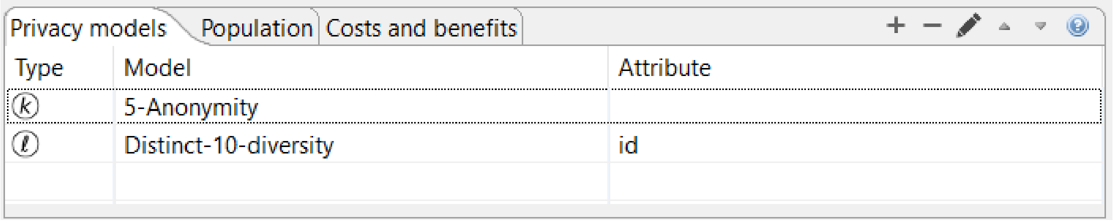
When analyzing data to anonymize using the ARX tool, I chose to use k-anonymity with k=5 and l-distinct 10-diversity to ensure privacy and data utility.

K-anonymity is a privacy model that aims to protect individual identities in a dataset by ensuring that each record cannot be uniquely distinguished from at least k-1 other records [3]. When k-anonymity is set to 5, it means that for any given combination of quasi -identifiers (attributes that can be used to identify individuals indirectly) in the dataset, there are at least five records with the same values. L-distinct diversity is a privacy model that extends k-anonymity by ensuring diversity in sensitive attributes within each group of records sharing the same quasi-identifiers. This helps protect against attribute disclosure, where an attacker could infer sensitive information about an individual even if their identity is not revealed [3]. Applying both helps to prevent the re -identification of individuals in the dataset. Because we had id in out set L-distinct diversity was required by ARX.

A higher value of k-anonymity or L-distinct diversity would provide more privacy, but lead to a significant loss of data utility due to generalisation or suppression while lower levels didn't provide enough protection. So I tested with ranges from 2..10 and ended up somwhere in middle of k=5 and l=10.

1.3 Chosen transformations and levels

"id" is a sensitive attribute because it represents a unique identifier for each individual in the dataset. Protecting this attribute is critical, as it can directly lead to the re-identification of individuals. Other attributes on their own might not be sufficient to uniquely identify individuals, but when combined, they can increase the risk of re-identification. By treating



Type	Model	Attribute
ⓚ	5-Anonymity	
ⓛ	Distinct-10-diversity	id

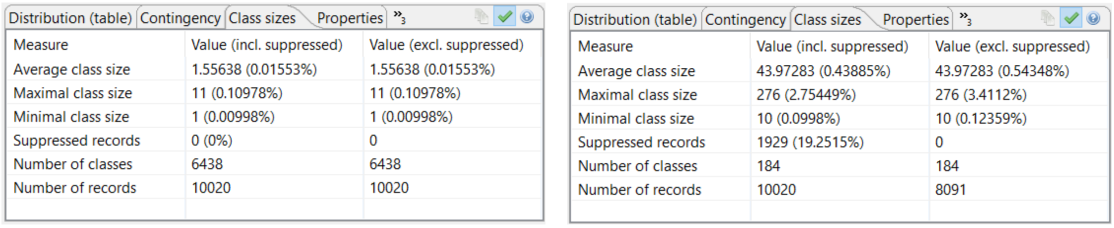
Figure 4. Applied privacy models

them as quasi-identifiers, I reduce this risk and ensure that the ARX tool considers them when applying privacy models like k-anonymity and l-diversity [4]. This approach allows me to apply different levels of privacy protection.

Table 2. Chosen data transformation types

Sensitive Attributes	**Quasi-identifiers**
id	Age.at.diagnosis Sex Native.country Month.first.diagnosis Year.first.diagnosis Vaccination Complicated.phase Critical.phase Last.known.patient.status

1.4 Minimal class size and number of records suppressed



Measure	Value (incl. suppressed)	Value (excl. suppressed)
Average class size	1.55638 (0.01553%)	1.55638 (0.01553%)
Maximal class size	11 (0.10978%)	11 (0.10978%)
Minimal class size	1 (0.00998%)	1 (0.00998%)
Suppressed records	0 (0%)	0
Number of classes	6438	6438
Number of records	10020	10020

(a) Class size unsuppressed

Measure	Value (incl. suppressed)	Value (excl. suppressed)
Average class size	43.97283 (0.43885%)	43.97283 (0.54348%)
Maximal class size	276 (2.75449%)	276 (3.4112%)
Minimal class size	10 (0.0998%)	10 (0.12359%)
Suppressed records	1929 (19.2515%)	0
Number of classes	184	184
Number of records	10020	8091

(b) Class size suppressed

Figure 5. Class size suppressions before and after

Figure 5 shows class sizes before and after anonymization. Prior to anonymization, the dataset comprised 10,020 records with no suppressed records as seen on Figure 5a. The average class size was 1 .55638 (0.01553%), with a maximal class size of 11 (0.10978%)

and a minimal class size of 1 (0.00998%). There were 6,438 classes in total. In Figure 5b, after anonymization process, the dataset contained 8,091 records, with 1,929 (19 .2515%) records suppressed. The average class size increased to 43.92283 (0.43885%), with a maximal class size of 226 (3.4112%) and a minimal class size of 10 (0.0998%). The total number of classes reduced to 184.

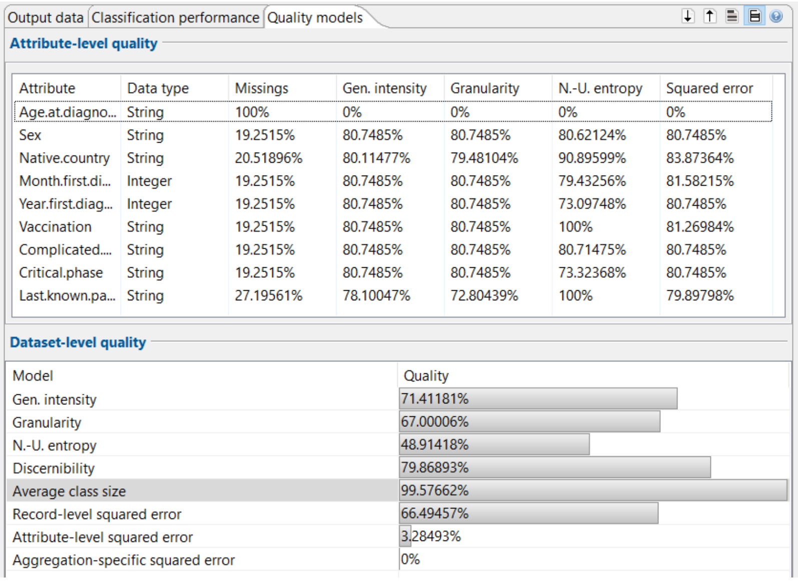
The comparison of the dataset before and after anonymization reveals a significant change in the distribution of records. The number of suppressed records increased to 1,929 (19.2515%), indicating a stronger emphasis on privacy protection. This increase in suppression, however, comes at the cost of data utility, as more records are removed from the dataset.

The average class size grew substantially from 1.55638 (0.01553%) to 43.92283 (0.43885%), reflecting the generalisation applied to the attributes. This generalisation ensures that each group of records sharing the same quasi-identifiers has a higher level of anonymity, but it may also lead to a loss of granularity in the data.

The maximal and minimal class sizes also changed considerably after anonymization. The maximal class size increased from 11 (0.10978%) to 226 (3.4112%), while the minimal class size rose from 1 (0.00998%) to 10 (0.0998%). These changes highlight the impact of the anonymization techniques in creating a more uniform distribution of records across classes, enhancing privacy protection.

Quality models as seen in Figure 6 interestingly show that “Age.at.diagnosis“ was considered the most revealing and was completely suppressed in the output data while other DA remained in the near 20% range.

In summary, it can be reasonably concluded that appropriate measures have been taken to anonymize the dataset in question. Although it is essential to further evaluate the sufficiency of the generalisations applied, such an assessment is contingent upon the specific context in which the data is required, the stipulated data constraints, and any DA suppression that may be explicitly requested by the DS. Given the scope of this exercise, the output generated is adequate in demonstrating the effectiveness of the anonymization process. For a visual representation of the results, please refer to Figure 7.



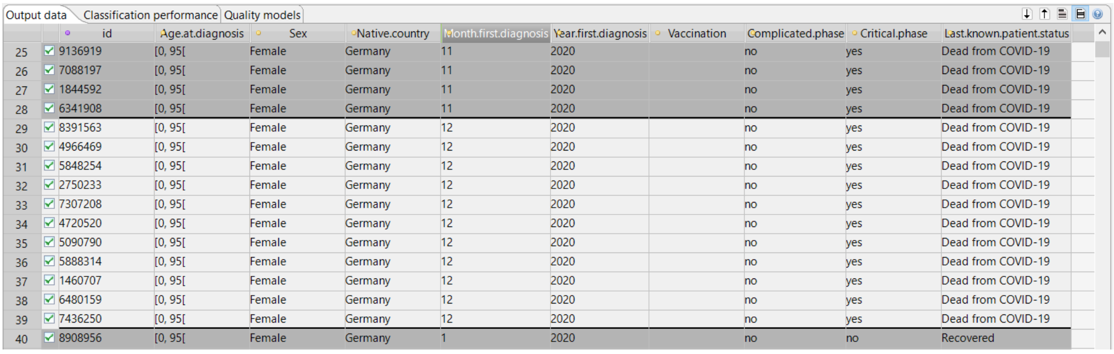
Attribute-level quality

Attribute	Data type	Missings	Gen. intensity	Granularity	N.-U. entropy	Squared error
Age.at.diagno...	String	100%	0%	0%	0%	0%
Sex	String	19.2515%	80.7485%	80.7485%	80.62124%	80.7485%
Native.country	String	20.51896%	80.11477%	79.48104%	90.89599%	83.87364%
Month.first.di...	Integer	19.2515%	80.7485%	80.7485%	79.43256%	81.58215%
Year.first.diag...	Integer	19.2515%	80.7485%	80.7485%	73.09748%	80.7485%
Vaccination	String	19.2515%	80.7485%	80.7485%	100%	81.26984%
Complicated....	String	19.2515%	80.7485%	80.7485%	80.71475%	80.7485%
Critical.phase	String	19.2515%	80.7485%	80.7485%	73.32368%	80.7485%
Last.known.pa...	String	27.19561%	78.10047%	72.80439%	100%	79.89798%

Dataset-level quality

Model	Quality
Gen. intensity	71.41181%
Granularity	67.00006%
N.-U. entropy	48.91418%
Discernibility	79.86893%
Average class size	99.57662%
Record-level squared error	66.49457%
Attribute-level squared error	3.28493%
Aggregation-specific squared error	0%

Figure 6. Quality models after anaonmization



		id	Age.at.diagnosis	Sex	Native.country	Month.first.diagnosis	Year.first.diagnosis	Vaccination	Complicated.phase	Critical.phase	Last.known.patient.status
25	☑	9136919	[0, 95[	Female	Germany	11	2020		no	yes	Dead from COVID-19
26	☑	7088197	[0, 95[	Female	Germany	11	2020		no	yes	Dead from COVID-19
27	☑	1844592	[0, 95[	Female	Germany	11	2020		no	yes	Dead from COVID-19
28	☑	6341908	[0, 95[	Female	Germany	11	2020		no	yes	Dead from COVID-19
29	☑	8391563	[0, 95[	Female	Germany	12	2020		no	yes	Dead from COVID-19
30	☑	4966469	[0, 95[	Female	Germany	12	2020		no	yes	Dead from COVID-19
31	☑	5848254	[0, 95[	Female	Germany	12	2020		no	yes	Dead from COVID-19
32	☑	2750233	[0, 95[	Female	Germany	12	2020		no	yes	Dead from COVID-19
33	☑	7307208	[0, 95[	Female	Germany	12	2020		no	yes	Dead from COVID-19
34	☑	4720520	[0, 95[	Female	Germany	12	2020		no	yes	Dead from COVID-19
35	☑	5090790	[0, 95[	Female	Germany	12	2020		no	yes	Dead from COVID-19
36	☑	5888314	[0, 95[	Female	Germany	12	2020		no	yes	Dead from COVID-19
37	☑	1460707	[0, 95[	Female	Germany	12	2020		no	yes	Dead from COVID-19
38	☑	6480159	[0, 95[	Female	Germany	12	2020		no	yes	Dead from COVID-19
39	☑	7436250	[0, 95[	Female	Germany	12	2020		no	yes	Dead from COVID-19
40	☑	8908956	[0, 95[	Female	Germany	1	2020		no	no	Recovered

Figure 7. Output dataset after anonymization process

1.5 Risk of the input data and output data

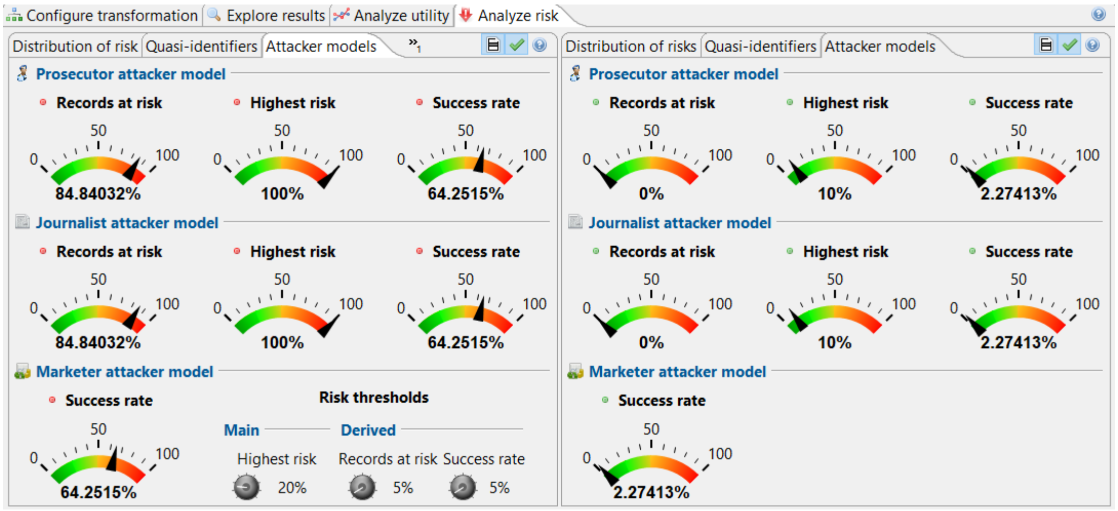
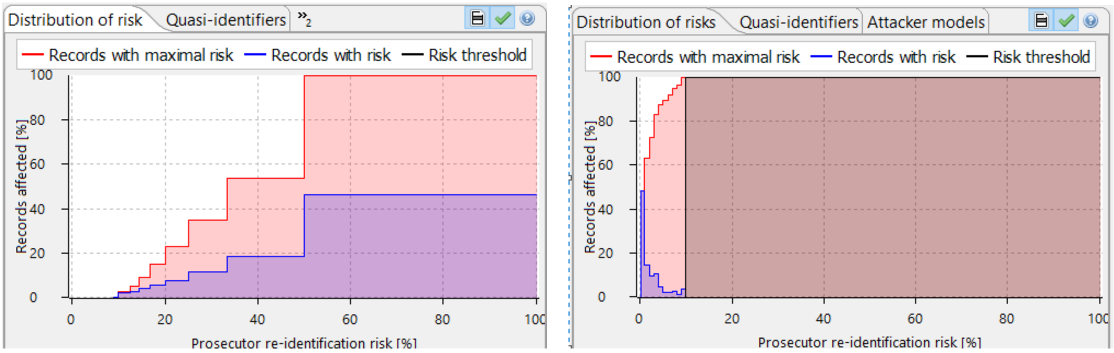


Figure 8. Attcker model risk meters



(a) Risk analyzis before ananymization

(b) Risk analyzis after ananymization

Figure 9. Risk before and after

The risk analysis conducted before and after data anonymization provides valuable insights into the effectiveness of the anonymization process in mitigating privacy risks associated with various types of adversaries, such as prosecutors, journalists, and marketers. Before anonymization on Figure 9a and looking at risk metrics on Figure 8 we can see that about 60-80% of our records are at high risk and records are identifiable by different models. Which is not suprising none of the DA are supressed. Yet which is surprising, is that even with “id“ fields given the percent is not 100% which I found odd. In the other hand, after ananomyzation, we can see that on all attacker models out records are relatively safe with the highest risk being around 10% with marginal 2% success rate.

Most optimal model can be seen on Figure 10 with almost all DA supressed and only 3 records left. The risk is very low, less than 1

Property;Value
Transformation;[6, 1, 1, 0, 0, 0, 0, 0, 2]
Anonymous;ANONYMOUS
Score;0.15920452011443986 [2.63116%]
Successors;3
Predecessors;4
Checked;true

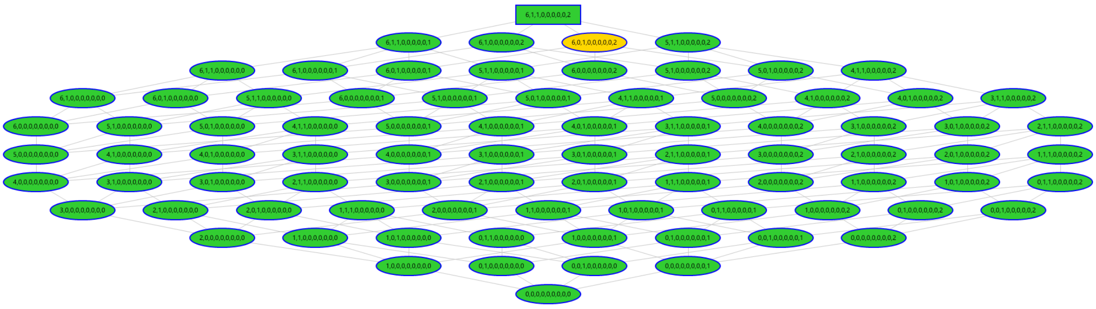


Figure 10. Output dataset after anonymization process

References

[1] [Online]. Available: http://www.operando.eu/upload/operando/moduli/D4.3Guidelinesfordataanonymizationreportv1.0_77_326.pdf.

[2] K. El Emam *et al.*, “A globally optimal k-anonymity method for the de-identification of health data,” *Journal of the American Medical Informatics Association*, vol. 16, no. 5, pp. 670–682, 2009.

[3] L. Sweeney, “K-anonymity: A model for protecting privacy,” *International journal of uncertainty, fuzziness and knowledge-based systems*, vol. 10, no. 05, pp. 557–570, 2002.

[4] [Online]. Available: https://arx.deidentifier.org/.