isec
Engenharia

MESTRADO EM INFORMÁTICA E SISTEMAS

Data Anonymization: Algorithms, Techniques, and Tools

DEFINITIVO

Autor

Joana Carolina Pedroso Tomás

Orientador

Jorge Fernandes Rodrigues Bernardino

Co-Orientador

Deolinda Maria Lopes Dias Rasteiro

INSTITUTO POLITÉCNICO DE COIMBRA

INSTITUTO SUPERIOR DE ENGENHARIA DE COIMBRA

Coimbra, julho de 2022

DEPARTAMENTO DE INFORMÁTICA E SISTEMAS

Data Anonymization: Algorithms, Techniques, and Tools

Relatório de Trabalho de Projeto para a obtenção do grau de Mestre em Informática e Sistemas

Especialização em Desenvolvimento de Software

Autor

Joana Carolina Pedroso Tomás

Orientador

Jorge Fernandes Rodrigues Bernardino

Co-Orientador

Deolinda Maria Lopes Dias Rasteiro

INSTITUTO POLITÉCNICO DE COIMBRA

INSTITUTO SUPERIOR DE ENGENHARIA DE COIMBRA

Coimbra, julho de 2022

Acknowledgements

I would like to thank my parents for their unconditional support in carrying out this thesis and for never letting me give up. Thanks also to my sister and all my friends for their support.

A special thanks to Professor Jorge Bernardino and Professor Deolinda Rasteiro for all the patience they had with me during this thesis and for helping me to always move forward.

Thank you all.

Resumo

Nos últimos anos, o volume de informação online tem vindo a crescer exponencialmente. Os dados pessoais de cada indivíduo são utilizados de forma contínua pelo governo, por empresas ou por indivíduos, com a finalidade de criar dados estatísticos. Estes podem depois ser utilizados em campanhas de marketing, na previsão de tendências futuras, na ajuda em investigações ao nível da ciência e da medicina e muitos outros exemplos.

O maior problema com a utilização destes dados é que eles podem conter informação sensível e informação que permita identificar um indivíduo, podendo causar graves problemas a nível pessoal como, por exemplo, roubo de identidade, extração de dinheiro, etc., dependendo dos dados divulgados.

Para resolver este problema existe a anonimização de dados. Esta tem como finalidade alterar os dados de modo a ocultar informação sensível e que podem permitir a identificação de um indivíduo, tornando-os menos precisos.

Uma das maiores dificuldades perante a anonimização de dados é que ao mesmo tempo que se mantém a privacidade dos indivíduos, a utilidade dos dados deve permanecer e, para isto, é necessário ter em atenção as técnicas e os algoritmos que são utilizadas e a quantidade de vezes que estas são aplicadas.

Neste trabalho são estudadas as técnicas de anonimização mais comuns, como a generalização, a supressão, a anatomização, a permutação e a perturbação e também alguns dos algoritmos de anonimização mais conhecidos, como o k-anonimato e o l-diversidade.

Para a avaliação e a aplicação destas técnicas e algoritmos foram utilizadas as ferramentas open-source, ARX Data Anonymization Tool, UTD Anonymization Toolbox e Amnesia. Utilizando a metodologia OSSpal foi também realizada a avaliação de cada uma destas ferramentas.

A metodologia OSSpal tem como finalidade avaliar ferramentas open-source de forma a ajudar os utilizadores e as organizações a encontrar as melhores, recorrendo a um conjunto de categorias. No contexto desta tese, as categorias utilizadas foram a funcionalidade, as características funcionais do software, o suporte e os serviços, a documentação, os atributos da tecnologia do software, a comunidade e a adaptação e o processo de desenvolvimento.

Nesta tese, o trabalho experimental realizado consistiu na avaliação das três ferramentas de anonimização utilizando dois dataset reais. O UTD Anonymization Toolbox só foi utilizado com um dos datasets, o de menor tamanho, porque esta ferramenta requer a introdução manual dos elementos do dataset num ficheiro, o que pode originar erros.

Na avaliação das ferramentas é possível verificar que o ARX Data Anonymization Tool é a ferramenta que apresenta os dados de forma mais simples e que permite uma melhor visualização por parte do utilizador. O Amnesia é fácil de utilizar pois mostra ao utilizador todos os passos necessários para anonimizar um dataset, apesar de mostrar alguns erros, porém, o UTD Anonymization Toolbox foi a ferramenta que apresentou mais dificuldades na utilização devido ao facto de não ter uma interface gráfica, mas também porque a introdução dos dados tem de ser feita de forma manual.

Após a avaliação experimental é possível concluir que o ARX Data Anonymization Tool é a melhor ferramenta para ser usada na anonimização de dados, seguindo-se o Amnesia e, por último o UTD Anonymization Toolbox.

Palavras-Chave: Anonimização de dados, Privacidade, Generalização, Supressão, Anatomização, Permutação, Perturbação, K-anonymity, L-diversity, ARX, OSSpal

Abstract

In the past few years, the volume of online information has increased exponentially.

The personal data of everyone is used incessantly by the government, organisations and/or individuals for the purpose of creating statistical data. These can then be used in marketing companies, forecasting future trends, helping in scientific and medical research and many other examples.

The biggest problem with the use of this data is that it can have sensitive information and information that allows an individual to be identified, which can cause serious problems at the personal level, such as identity theft, money extraction, etc, depending on the data disclosed.

To solve this problem exists data anonymization. This has the intend to change data in order to hide sensitive information and substitute data that may allow the identification of an individual, making them less accurate.

One of the main problems with data anonymization is that while keeping the privacy of individuals, the data utility should remain and for this, it is necessary to consider the used techniques and algorithms and the number of times each one of them is applied to the data.

In this work, the most common anonymization techniques are studied, like generalization, suppression, anatomization, permutation, and perturbation, as well as some of the most well-known anonymization algorithms, like k-anonymity and l-diversity.

To assess the application of these techniques and algorithms are used some open-source tools: ARX Data Anonymization Tool, UTD Anonymization Toolbox and Amnesia. Using the OSSpal methodology, the evaluation of each of these tools was also carried out.

OSSpal methodology has the purpose of evaluating open-source tools to help users and organisations to find the best solutions using a set of categories. In the context of this thesis, the categories used are functionality, operational software characteristics, support and services, documentation, software technology attributes, community and adaption and development process.

In this thesis, the experimental work made consists in the assessment of the three anonymization tools using two real datasets. UTD Anonymization Tool was only used with one dataset because this has the smaller size and the tool requires the manual introduction of the dataset elements in a file, which can originate some errors.

In the evaluation of the tools, it is possible to verify that the ARX Data Anonymization Tool is the tool that presents the data in a simpler way and that allows a better visualization by the user. Amnesia is easy to use as it shows the user all the necessary steps to anonymize a dataset even though it showed some errors, however, the UTD Anonymization Toolbox was the tool that presented more difficulties in use due to the fact that it does not have a graphical interface but also because the data entry has to be done manually.

After this assessment it is possible to conclude that ARX Data Anonymization Tool is the best tool to use in data anonymization process, followed by Amnesia and, finally, UTD Anonymization Toolbox.

Keywords: Data anonymization, Privacy, Generalization, Suppression, Anatomization, Permutation, Perturbation, K-anonymity, L-diversity, ARX, OSSpal

Index

Figure Index

Table Index

Symbols and Abbreviations

EU – European Union

GDPR – General Data Protection Regulation

UTD – University of Texas at Dallas

1. Introduction

In the current days, the information online is increasing exponentially, and personal data is used by the government and organizations with the purpose of extracting value. This data can be used to make studies, statistics, scientific and medical investigations, forecasting future trends, and others.

One of the problems is that this data can contain sensitive information that conducts to individual identification and therefore misused. For example, information like address, age, bank account number, health and many others are considered sensitive information and should not be exposed or even accessible to people that are not allowed to it. The issue is that no one wants their personal data available online without consent, where everyone can have access to it and use it, sometimes not in the best way, causing serious problems like identity deft, money extraction and others, depended on the data disclosed.

Data anonymization exists to solve these problems by changing the original data to hide or modify the sensitive information. The result is a dataset less accurate, and it is not possible to identify an individual from the anonymized data. One of the issues with the data anonymization is the data utility loss. It is necessary to consider the number of used techniques and algorithms and the number of times each one of them is applied to the dataset. If it is applied more than the necessary, the data utility is lost and, therefore, the dataset does not have the correct information to be used by the government, organizations and/or individuals.

Some countries create legislation that helps people to keep their sensitive information secret by the companies. In Europe, exists the European General Data Protection Regulation (EU GDPR), which is the regulation present in European Union to protect personal data. This regulation lays down rules regarding the protection of individuals with regard to the processing of personal data and their free movement.

In this work, it is possible to verify that data anonymization is an important technique to protect personal data and to avoid attackers from having access to it. Some open-source tools help the users to anonymize their information and those tools are studied in this thesis. Anonymization algorithms like k-anonymity and l-diversity are presented as well as the most used anonymization techniques like generalization, suppression, anatomization, permutation, and perturbation. Some popular open-source tools, like ARX Data Anonymization Tool, Amnesia and UTD Anonymization Toolbox, are also studied to help people to have different options in the process of data anonymization and to know which one of them is the best tool.

An assessment is done for each tool using the OSSpal methodology. OSSpal methodology, based on a set of categories, has the purpose to evaluate open-source tools to help users and organizations to find which one is the best. The categories used in this thesis are functionality, operational software characteristics, support and services, documentation, software technology attributes, community and adaption and development process. Using the OSSpal evaluation, the ARX Data Anonymization Tool has the better evaluation, followed by Amnesia and finally by UTD Anonymization Toolbox.

Each one of the tools was also experimentally evaluated using two datasets to perform the anonymization. The first dataset has information available for research that is useful to predict a possible heart disease based on the presented features. The second dataset has the most recent tweets about Pfizer & BioNTech vaccine. In the dataset anonymization done by each tool is possible to conclude that Amnesia has the simplest anonymization process, but the results retrieved are confusing because the presented solutions graph has a solution for all the attributes, which make it big and hard to find the solutions and analyse the graph. ARX Data Anonymization Tool, besides not having an anonymization process so simplified, its results are clearer since it only shows the results in the solutions for the quasi-identifier attributes. UTD Anonymization Toolbox is the most difficult tool to use and the one where errors are most likely to occur since the tool depends on data manual insertion by the user. All the values that are string type need to be inserted manually in a specific document, which is a very difficult process to be done manually and can lead to errors. Also, the results retrieved by this tool are not the clearest. Thus, the ARX Data Anonymization Toolbox is the best tool, in our opinion, to use to anonymize a personal dataset.

The main contributions of this thesis are the following:

- Data anonymization tools assessment using OSSpal methodology;
- Data anonymization tools experimental evaluation using two public datasets;
- Best tool to use according to dataset characteristics;
- Main weaknesses and difficulties in practical use of each tool.

The rest of this thesis is organized as follows. Chapter 2 explains the main Concepts and Background about data anonymization. Chapter 3 reports some works related to anonymization techniques and anonymization tools. Chapter 4 presents the Anonymization Techniques. Chapter 5 presents the Anonymization Algorithms. Chapter 6 describes the Anonymization Tools and Chapter 7 presents the Assessment of the Anonymization Tools using OSSpal methodology. Chapter 8 describes the datasets used. Chapter 9 presents the experimental evaluation of the tools using the datasets. Finally, Chapter 10, presents the conclusions and future work.

2. Concepts and Background

In this chapter, some important concepts about Data Anonymization will be explained to help the understanding of the topic. First, is important to mention that exists a regulation to protect personal data that is titled "General Data Protection Regulation"[1] or GDPR, as short. The one used in this master thesis was created by the European Union to protect the people with the use of their personal data and the free movement of it. This document defines the *personal data* as "any information relating to" someone "who can be identified, directly or indirectly, by reference to an identifier (…), an online identifier or to one or more factors specific to the physical, physiological, genetic, mental, economic, cultural, or social identity of that natural person." (Union, 2016).

Another important concept to retain is *attribute types*. In a dataset with personal information, it is possible to identify four different types of attributes, which are the following (Sharma, Choudhary, & Jain, 2019):

- Identifier attributes – This attribute has information like the name, address, identification number, etc, which means that it is possible to identify a person just with this attribute.
- Quasi-identifier attributes – A quasi-identifier, on its own, does not allow to identify someone, but combined with other attributes is possible to identify a person. A quasi-identifier attribute could be age, job, gender, etc.
- Sensitive attributes – This type of attribute has sensitive information that an individual does not want to turn public, like information about a disease, Social Security Number, Salary, etc.
- Non-sensitive attributes – The non-sensitive attributes are all the attributes that do not belong to any of the previous types.

Data anonymization is a process where the personal information of an individual is removed from the dataset, so that the person cannot be identified in the dataset and the sensitive attributes cannot be matched to a specific person (Ren, Wang, Choo, & Xhafa, 2019).

Another important definition that should be retained is the difference between anonymization technique and anonymization algorithm. An *anonymization technique* replaces the original data by other characters or by more generalized values, depending in the anonymization technique that is used. For example, in the generalization technique, the values are replaced by more generic values and in the suppression technique, the original values are replaced by special characters. An *anonymization algorithm* is used to modify the original dataset without hiding the original values, just replacing them by the other values already existing in the dataset

[1] https://eur-lex.europa.eu/eli/reg/2016/679/oj

for that attribute. It also can use anonymization techniques to hide some values in the quasi-identifier attributes. The anonymization techniques are described in detail in Chapter 4 and anonymization algorithms in Chapter 5.

An *anonymization tool* is a software used to apply the anonymization techniques and algorithms to a dataset to simplify the process to the user, avoiding human errors. In some anonymization tools it is also possible to verify the data utility in the anonymized dataset. Another important term is *information loss*, which is the “number of information that is reduced due to data modification” (Gunawan & Mambo, 2018). If the value of the information loss is high, the usefulness of the data is lower, which means that the data utility for the other companies, organizations and/or individuals is not significant, and the values windrowed from the anonymized dataset (for example, statistical data) are not close to the reality.

The anonymization tools are described in detail in Chapter 6. *OSSpal methodology* is used to evaluate open-source tools to help users and organizations to find the best open-source tools using a set of categories. This is explained in detail in section 7.1 of Chapter 7.

3. Related Work

This chapter presents some of the most important works related to anonymization techniques and anonymization tools.

In the work of Prasser, Eicher, Spengler, Bild, & Kuhn (2020), the authors were required to extend the open-source ARX Data Anonymization Tool in a way that it could apply a full-domain generalization algorithm in different subsets. The idea was to implement horizontal and vertical partitioning strategies so that the tool could work with more flexible transformation models while preserving scalability. With these modifications, the ARX Data anonymization Tool was able to support four different transformation methods: generalization, suppression, sampling, and microaggregation. Inside the generalization type, the full-domain generalization, top-and-bottom-codding, and the categorization were implemented, which added a multi-dimensional generalization. The suppression technique was applied to attribute-level and record-level, and they extended this implementation to the cell level. The sampling type was already implemented and the microaggregation type is new for the tool, which has four different implementations: arithmetic and geometric mean, median and mode, set, and interval. The authors tested these modifications using six real-world datasets and compared the results with UTD Anonymization Toolbox, which is explained in section 6.3 from chapter 6 of this thesis, with the Mondrian algorithm and against the Sánchez et al. algorithm. The goal of this comparison is to prove that ARX Data Anonymization Tool has better results for scalability and data utility when using the local generalization with the horizontal and vertical partitioning strategies enabled. In comparison with UTD Anonymization Toolbox, the ARX Data Anonymization Tool has better results in data utility and scalability. Comparing the new implementation with the Sánchez et al. algorithm, the ARX Data Anonymization Tool has better results in terms of data utility but in terms of scalability, the Sánchez et al. algorithm has better results. This could be justified by the fact that the new algorithm always guarantees identical records in input data but also the output data, so this algorithm is less flexible.

He & Naughton (2009) realized that the existing data anonymization tools were not the best options to anonymize set-valued data. Facing this problem, the authors propose the top-down partitioning approach to anonymize the set-valued data. According to the authors, this approach has satisfactory results in terms of information loss. They based this implementation on the k-anonymity algorithm, but their definition of k-anonymity for set-valued data anonymization is different from the original one because in their approach “every transaction in the database occurs at least k times or the size of each equivalence class in the database is at least k” while in the original approach this is only performed for the quasi-identifier attributes.

They used three real-world datasets to test the new approach: BMS-WebView1, BMS-WebView2, and BMS-POS, evaluating the efficiency and effectiveness of the algorithm and measuring the execution time. They compared the results provided by their top-down partition-based anonymization with the bottom-up Apriori-based anonymization. In terms of search query logs, their algorithm is efficient to be applied to non-trivial real-world data but, despite the results were not bad for information loss, they were expecting better.

Gunawan & Mambo (2018) realized that, with the usage of the existing data anonymization algorithms, when several modifications were applied to the database, the data utility and data property were reduced a lot. Therefore, they tried to create a new schema to anonymize the data called subling suppression. The main goal of this anonymization approach is to reduce the data utility lost and maintain the data properties, like database size and the number of records. This schema is applied in two steps: in the first one, the records need to be grouped based on the adversary knowledge. In the second step, the items grouped before need to be replaced by a surrogated item, which is an item of the same category of the items in the adversary knowledge based on the hierarchy tree. To test this new approach, they used a real-world dataset (BMS-WebView2), the Normalized Certainty Penalty (NCP) metric to measure the information loss and the dissimilarity metric to measure the difference between the original database and the anonymized one so that they could compare the data property results. The results were positive for this new schema: the dissimilarity was zero, so, the number of records of the original database is the same as the number of records of the anonymized database since this schema works in the selection and replacement of the items. They also have low values in the result for information loss, which means they did not lose data in the anonymization process. Another advantage they refer to in the paper is that the data utility is also preserved in data mining tasks.

Ghinita, Karras, Kalnis, & Mamoulis (2007) proposed a framework to work as a solution to the inefficient anonymization process in terms of computation and I/O costs. They refer that the information loss metrics are not easy to understand, and that the l-diversity is solved by the techniques implemented for k-anonymity. Their framework addresses these issues with an efficient privacy preservation approach. They first implement the framework for a one-dimensional quasi-identifier with the running time linear to the size of the dataset. For the most complex l-diversity problem, they proposed an efficient heuristic algorithm with linear-time complexity. Finally, they extended the algorithms to the multi-dimensional quasi-identifiers, using the space-mapping techniques. The authors are studying optimal solutions for k-anonymity and l-diversity based on meaningful information loss metrics. For the tests, they used two datasets: CENSUS and ADULT and consider mapping based on the Hilbert space-filling curve and iDistance, which was used to solve the multi-dimensional problem. To measure the information loss, the authors used the GPC metric. The results were satisfactory in terms of the execution time and information loss, compared with the existing state-of-art.

Almokbily & Rauf (2018) identify that most of the existing anonymization methods do not prevent membership, identity disclosure, homogeneity attack and semantic similarity attack while the utility of data was guaranteed. They proposed a hybrid approach where the bucketization of (l, e) diversity and generalization and suppression of k-anonymity algorithms was used. In this new approach, they started to remove all explicit identifiers and then apply some steps of (l, e) diversity algorithms but with some modifications. First, the tuples should be ordered based on their quasi-identifier attributes so that it could be possible to group these tuples to form buckets. These buckets are created by recursively selecting the ‘l’ closest tuples from ‘l’ largest semantic groups. Finally, for each remaining tuple, compute the Normalized Certainty Penalty metric. This new approach ensures that in each bucket, the semantic similarity between at least ‘l’ sensitive values should be more than ‘e.’ After performing these steps, the generalization and suppression techniques are applied for each bucket. They have done some tests using the dataset Adult from the UCI machine learning data repository and compared the proposed technique with the klredInfo and (l, e) diversity. The proposed algorithm has increased the diversity degree, reduced the discernibility penalty, which means that they could preserve the data utility more effectively as klredInfo and (l, e) diversity.

Murthy, Bakar, Rahim, & Ramli (2019) compared some anonymization techniques using the same dataset. The purpose of the study was to review the strengths and weaknesses of each one of the techniques. The techniques studied were generalization, suppression, distortion, swapping, and masking. They associate the best techniques that should be used for each one of the attributes of the dataset, taking into account the data type of the attribute. They verified that the masking and distortion techniques were the ones that were ideal for all types of data; suppression has the same results as masking but with more efficiency and also that the distortion technique makes the data become unrecognized but can be reverted to the original data by removing the noise. The conclusion is that many techniques can be applied to any data.

Arora, Bansal, & Sofat (2014) analysed k-anonymity, l-diversity and t-closeness applications in high dimensional databases on the basis of privacy and performance. The privacy metric used by the authors is information loss. The datasets chosen were transportation system dataset, census marriage dataset and Crime state by state dataset. The conclusions they retained from the study is that t-closeness has less information loss than l-diversity and k-anonymity but all the techniques still lead to a big information loss values and also that as the number of attributes increases in the dataset, the information loss also increases.

The work done in this thesis differs from the work of the other authors presented in this chapter because we are analysing the different tools and comparing them in terms of usage, functionalities, number of algorithms, and other features while the existing work analyses ARX Data Anonymization Toolbox and presents some new approaches regarding anonymization techniques and/or anonymization algorithms.

4. Data Anonymization Techniques

Some anonymization techniques are mostly used than others and combined, they can provide higher levels of anonymization and decrease the probability of deidentification.

In this chapter, it is presented some of the most popular techniques. The anonymization techniques are generalization, suppression, anatomization, permutation, and perturbation. Each one of them is explained in the following sub-chapters as their applications to an original dataset and the result: the anonymized dataset.

4.1. Generalization

The main goal of this technique is to turn the identification of an individual more difficult, replacing the original values with others less specific but semantically similar (Rao & Satyanarayana, 2018). This technique is applied to the quasi-identifier attributes and their values can be categorical or numerical (Sharma, Choudhary, & Jain, 2019) (Brito & Machado, 2017).

Generalization could be divided into two types (Brito & Machado, 2017):

- Global generalization – every attribute with the same value is generalized to the value of the same hierarchy level;
- Local generalization – different values for the same attribute can be generalized to different values, even if they are not at the same hierarchy level.

The inverse operation of generalization is specialization (Gunawan & Mambo, 2018).

Different schemas can be used to apply the generalization technique to a dataset: full-domain generalization scheme, sub-tree generalization scheme, sibling generalization scheme, and cell generalization scheme. These different schemas are analysed in the next sub-chapters as well as an example for each one.

The dataset in Table 1 is used as the data input for the anonymization schemas presented below. Figure 1, Figure 2, and Figure 3 are the hierarchy trees for the quasi-identifier attributes of the dataset in Table 1.

Table 1 - Original Dataset

Job	Sex	Age	Disease
Writer	Male	25	Diabetes
Engineer	Female	32	Cancer
Engineer	Female	22	Hypertension
Musician	Male	38	Diabetes
Economist	Male	24	Hypertension

The attributes Job, Sex, and Age are quasi-identifiers, and disease is the sensitive attribute.

From the previous table, the data can be divided into three trees since it exists three quasi-identifiers. Consequently, Figure 1 is the hierarchy tree for the quasi-identifier attribute Job, Figure 2 is the hierarchy tree for quasi-identifier attribute Age, and Figure 3 is the hierarchy tree for the quasi-identifier attribute Sex.

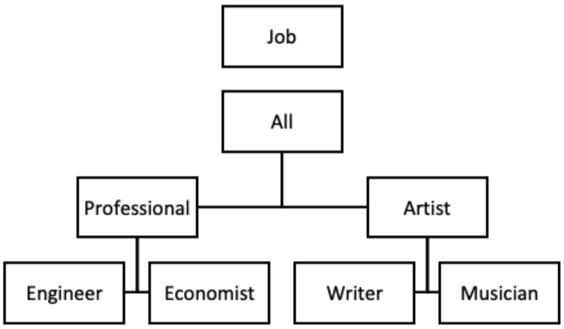


Figure 1 - Hierarchy Tree for Job Attribute

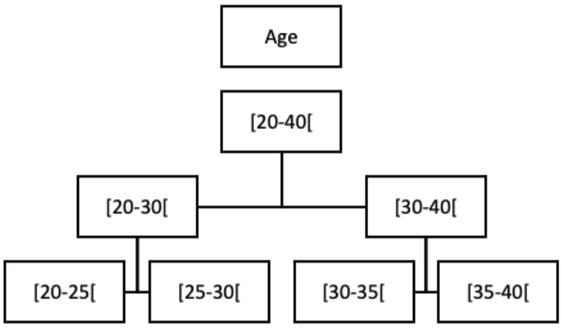


Figure 2 - Hierarchy Tree for Age Attribute

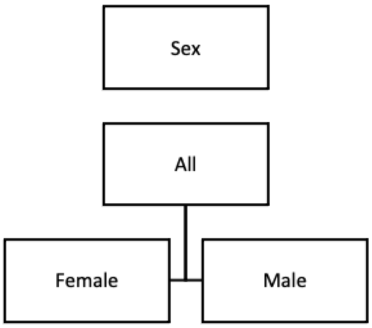


Figure 3 - Hierarchy Tree for Sex Attribute

4.1.1.Full-domain generalization scheme

In this generalization scheme, all the attribute values are generalized for the same level of the hierarchy in the tree. This scheme is the one that needs to search in less space, but it has a higher distortion from the original dataset because all the values are going to be generalized (Fung, Wang, Chen, & Yu, 2010).

Following the tree in Figure 1, using this generalization scheme, the values Engineer and Economist are going to be replaced by Professional, but the other values in the same hierarchy level in the tree need to be also anonymized, which means, the values Writer and Musician are going to be replaced by Artist. So, this attribute, in the input table, is going to be like what is represented in Table 2:

Table 2 - Full-Domain Generalization Result

Job	Sex	Age	Disease
Artist	Male	25	Diabetes
Professional	Female	32	Cancer
Professional	Female	22	Hypertension
Artist	Male	38	Diabetes
Professional	Male	24	Hypertension

4.1.2.Sub-tree generalization scheme

The sub-tree generalization scheme requires that just the values that belong to the same sub-tree need to be generalized to the same hierarchy level. If the value that is going to be anonymized is in different sub-trees, the ones for the different sub-tree can remain unchanged (Fung, Wang, Chen, & Yu, 2010).

Using again the tree in Figure 1, the values Engineer and Economist can be generalized to Professional. Although Writer and Musician are in the same hierarchy level in the tree they do not need to be generalized to Artist because they are in a different sub-tree. In Table 3, it is possible to see this change, in the attribute Job.

Table 3 - Sub-Tree Generalization Result

Job	Sex	Age	Disease
Writer	Male	25	Diabetes
Professional	Female	32	Cancer
Professional	Female	22	Hypertension
Musician	Male	38	Diabetes
Professional	Male	24	Hypertension

4.1.3.Sibling generalization scheme

This generalization scheme is like the sub-tree generalization scheme. The difference is that not all the values in the same sub-tree need to be generalized, some of them could maintain themselves unchanged. This scheme provides less distortion than the previous one since there is no need to generalize all the values (Fung, Wang, Chen, & Yu, 2010).

In the tree used in Figure 1, Engineer can be generalized to Professional, but the remaining values will be kept unchanged. This generalization scheme will change everything to Professional except Economist. Table 4 is the result of the application of the sibling generalization scheme to the input table, which is, Table 1.

Table 4 - Sibling Generalization Result

Job	Sex	Age	Disease
Artist	Male	25	Diabetes
Professional	Female	32	Cancer
Professional	Female	22	Hypertension
Artist	Male	38	Diabetes
Economist	Male	24	Hypertension

4.1.4.Cell generalization scheme

This is the scheme that causes less distortion in comparison with all the other ones because there is no need to generalize all the instances of the same value, some of them could remain unchanged. This will affect the data utility, which causes, consequently, a data exploration problem (Fung, Wang, Chen, & Yu, 2010).

As it was previously explained, the full-domain generalization is the one that requires a smaller search space, but at the same time, the one that results in the highest distortion of the data. Unlike what happens with the cell generalization scheme, in this one, the search space is the biggest one, but it does not cause such a huge distortion of the data. Consequently, it can affect the data utility, as mentioned previously.

Using Table 1 as the input table, in this case, the attribute Job in the second record of the table is the one that is going to be anonymized. This means that, with this anonymization schema, there is no need to anonymize any other record of the table and the result can be verified in Table 5.

Table 5 - Cell Generalization Result

Job	Sex	Age	Disease
Writer	Male	25	Diabetes
Professional	Female	32	Cancer
Engineer	Female	22	Hypertension
Musician	Male	38	Diabetes
Economist	Male	24	Hypertension

4.2. Suppression

The main goal of this technique is to replace the values in the quasi-identifier attributes for any special character, like, for example, '*' or '?'. This technique, like the previous one, is applied to the quasi-identifier attributes and they can be categorical or numerical (Sharma, Choudhary, & Jain, 2019).

Like generalization, suppression can be divided into two groups (Brito & Machado, 2017):

- Global Suppression – where all the instances of a value are suppressed;
- Local Suppression – In this type of suppression, just some instances of a value are suppressed.

There are also some schemes related to suppression and they are record suppression, value suppression and cell suppression, and each one of them is explained in the following subsections.

4.2.1.Record Suppression

In this suppression scheme, a column is entirely suppressed, and it is not possible to ever find which values it had initially (Brito & Machado, 2017).

If the suppression scheme is applied to the Sex attribute in Table 1, Male and Female need to be replaced by a special character. This change is represented in Table 6.

Table 6 - Record Suppression Result

Job	Sex	Age	Disease
Writer	*	25	Diabetes
Engineer	*	32	Cancer
Engineer	*	22	Hypertension
Musician	*	38	Diabetes
Economist	*	24	Hypertension

4.2.2.Value Suppression

The value suppression scheme allows to remove or replace all the instances of a value in a table with a special character (Brito & Machado, 2017).

In this scheme, just a value from the attribute Sex needs to be suppressed, which means that if the value Male is the one to suppress, all the equal values for this attribute need to be suppressed, but the value Female remains unchanged. Table 7 shows the application of the Value Suppression scheme to the input table, Table 1.

Table 7 - Record Suppression Result

Job	Sex	Age	Disease
Writer	*	25	Diabetes
Engineer	Female	32	Cancer
Engineer	Female	22	Hypertension
Musician	*	38	Diabetes
Economist	*	24	Hypertension

4.2.3.Cell Suppression

This suppression scheme is remarkably similar to the previous one, but in cell suppression not all the instances of a value are removed or replaced by a special character, but just some of them. This is not the best approach since it can lead to inconsistent values (Brito & Machado, 2017).

Using one more time Table 1 as the input dataset and the attribute Sex, just one of the values can be chosen to suppress. In Table 8, just one value was chosen to be suppressed.

Table 8 - Cell Suppression Result

Job	Sex	Age	Disease
Writer	Male	25	Diabetes
Engineer	Female	32	Cancer
Engineer	*	22	Hypertension
Musician	Male	38	Diabetes
Economist	Male	24	Hypertension

All the remaining values can be kept unchanged.

4.3. Perturbation

This anonymization technique is known for being simple, efficient and the one that has the best ability to preserve the statistical information calculated on the original data, since the statistic calculation results from the syntactic values are not so distant from the statistic calculation results from the original values (Brito & Machado, 2017).

The main purpose of the perturbation is to replace the original data with synthetical data, which means that the values from the quasi-identifier attributes are replaced by some fictitious values.

The preservation of the statistical results is one of the advantages of this anonymization technique. Another important advantage is that the attacker cannot identify an individual by the linkage of the quasi-identifier attribute values or having access to the individuals' sensitive information since the values are synthetic and it is not published the original information about an individual (Fung, Wang, Chen, & Yu, 2010).

The perturbation can be applied to the original data by some methods, like additive noise, data swapping and synthetic data generation, as it is explained in the following subsections.

4.3.1.Additive Noise

In additive noise, the main purpose is to replace the original value with a new one. If the original value is "v", then a value "r" is chosen, called noise. The noise is added or multiplied by the original value and the result is the perturbed value. This method is usually applied to sensitive numerical data (Fung, Wang, Chen, & Yu, 2010).

4.3.2.Data Swapping

Data swapping is a process where two values of the same attribute are exchanged between each other. This method can be used to protect numerical and categorical attributes, but it can also generate meaningless records (Fung, Wang, Chen, & Yu, 2010).

4.3.3.Synthetic Data Generation

In synthetic data generation, a new statistical model is generated, based on the original one, and then synthetical data is generated. This synthetical data follows the original model and this is the one that should be released to the public. This model is the one that best preserves the statistical information, but it can also generate meaningless information (Fung, Wang, Chen, & Yu, 2010).

4.4. Anatomization

This anonymization technique separates the attributes into two tables: one for the quasi-identifier attributes and another one for sensitive attributes. These tables are linked together by a new column that is equal in the quasi-identifier table and the sensitive table, the GroupID column (Almokbily & Rauf, 2018).

This technique should not be applied to continuous data publishing and its major advantage is that the data in the quasi-identifier table and the sensitive table is not modified.

Using a new table as a dataset, Table 9, the anatomization technique is applied to all attributes. In Table 9, the attributes Age and Sex are quasi-identifiers, and the attribute disease is a sensitive attribute. The rows with the same values will be joined together and then, in the anatomized table, will have a table, named 'Count', where it is written the number of times that each record was shown in the original table (Fung, Wang, Chen, & Yu, 2010).

Table 9 - Original Dataset for Anatomization

Sex	Age	Disease
Male	30	Hepatitis
Male	30	Hepatitis
Male	30	HIV
Male	32	Hepatitis
Male	32	HIV
Male	32	HIV
Female	36	Flu
Female	38	Flu
Female	38	Heart
Female	38	Heart

After applying the anatomization technique, the result is a table with the quasi-identifier attributes (Table 10) and a table with the sensitive attributes (Table 11).

Table 10 - Quasi-Identifier Attribute Table

Sex	Age	GroupID
Male	30	1
Male	30	1
Male	30	1
Male	32	1
Male	32	1
Male	32	1
Female	36	2
Female	38	2
Female	38	2
Female	38	2

The Count column with value 3 in the first record in Table 11 means that exists 3 persons, with GroupID 1 that has Hepatitis. Connecting Table 11 with Table 10 it is possible to conclude that 3 male individuals have hepatitis.

Table 11 - Sensitive Attribute Table

GroupID	Disease	Count
1	Hepatitis	3
1	HIV	3
2	Flu	2
2	Heart	2

4.5. Permutation

The permutation technique is an improved version of anatomization. While anatomization just divides the original data into two tables, one with quasi-identifier attributes and the other with sensitive attributes, the permutation applies a random permutation in the values before releasing the data. This technique provides strong privacy preservation and good data utility.

As permutation is similar to anatomization, the anatomized dataset is going to be used as the input example for the permutation, which means that Table 9 is going to be used as the original dataset.

After the division of the main dataset into two tables, it is needed to shuffle the sensitive attributes in the sensitive table, so the data can be released (Li, Xianmang, Cao, & Chen, 2015).

Taking Table 11, which is the table with the sensitive attributes, the next step is to shuffle them, within each GroupID, of the sensitive attribute Disease. Table 12 represents the application of this permutation technique.

Table 12 - Permutation Result

GroupID	Disease	Count
1	HIV	3
1	Hepatitis	3
2	Heart	2
2	Flu	2

5. Data Anonymization Algorithms

As mentioned in Chapter 2, an anonymization algorithm modifies the dataset without hiding the data. The anonymization algorithms can be used combined with the anonymization techniques to achieve a higher level of anonymization. In this case, an anonymized dataset will have some values that were replaced by a special character or generalized, due to the use of the anonymization techniques, but also modified values by another that were already present in the dataset, for the attribute used to be anonymized.

The aim of this thesis is not to study, in depth, the presented algorithms. Nevertheless, their comprehension and knowledge should be acquired in order to better understand the tools that were development on top of them.

In this chapter are presented some of the most important and more known anonymization algorithms.

5.1. K-anonymity

K-anonymity more than an algorithm is a data property inherent to its anonymization.

Samarati & Sweeney (1998) proposed an algorithm that prevents the identity linkage, computing the minimal generalization for a given table. The minimal generalization is explained by the authors as just generalizing the data as minimal as possible.

The k-anonymity algorithm is used with generalization and/or suppression techniques. In the work performed by Samarati & Sweeney (1998) k-anonymity is defined in the following way:

*Definition 1 (*k-anonymity*) (Samarati & Sweeney, 1998):*

"Let T $(A_{1,} \dots, A_n)$ *be a table with n attributes and* QI_T *be the quasi-identifiers associated with it. T is said to satisfy k-anonymity if for each quasi-identifier* $QI \in QI_T$ *each sequence of values in T[QI] appears at least with k occurrences in T[QI]".*

Satisfying this definition means that each record cannot be linked to a specific individual with a probability higher than 1/k.

The value k defines the level of privacy and is related to the information loss, thus, the higher the k value is, the higher the privacy is and less is data utility (Brito & Machado, 2017). This anonymization algorithm prevents identity linkage, but it does not prevent attribute disclosure (Israni, Chopra, & Jewani, 2017).

Table 13 will be used as an original dataset to better explain the application of the k-anonymity technique. According to the definition, the k-anonymity algorithm is applied to the quasi-identifier attributes, performing on them the generalization and/or suppression techniques. Both these techniques are used in Table 14, the suppression is used in ZipCode and Sex attributes and the generalization technique is used in Age attribute. As k-anonymity is only used in quasi-identifier attributes, the Disease attribute is not modified in Table 14 because it is a sensitive attribute.

Table 13 - Original dataset for k-anonymity

Quasi-identifier Attribute			Sensitive Attribute
ZipCode	**Sex**	**Age**	**Disease**
3090-461	Male	30	Cancer
3090-362	Male	35	Hepatitis
3090-432	Female	39	Flu
3090-351	Male	32	Heart
3080-460	Female	51	Heart
3080-320	Female	50	Heart
3080-200	Female	58	Heart
3080-020	Female	56	Heart
3040-321	Male	22	Hepatitis
3040-100	Male	25	Cancer
3040-200	Female	27	Heart
3040-300	Male	21	Flu

This table will be 4-anonymous, which means that the sequence of values in the quasi-identifier attributes appears at least with 4 occurrences in the table. This division can be seen in Table 14, where the three groups where the quasi-identifier attribute values repeat four times are presented, therefore this table is 4-anonymous. These different values can be verified in Table 14 where the Age attribute just has three different values: [30-40[, [50-60[and [20-30[and this also can be verified in the ZipCode attribute, where the only values that are present in Table 14 are “3090-***”, “3080-***” and “3040-***”.

Table 14 - 4-Anonymous data table

Quasi-identifier Attribute			Sensitive Attribute
ZipCode	**Sex**	**Age**	**Disease**
3090-***	*	[30-40[	Cancer
3090-***	*	[30-40[	Hepatitis
3090-***	*	[30-40[	Flu
3090-***	*	[30-40[	Heart
3080-***	*	[50-60[	Heart
3080-***	*	[50-60[	Heart
3080-***	*	[50-60[	Heart
3080-***	*	[50-60[	Heart
3040-***	*	[20-30[	Hepatitis
3040-***	*	[20-30[	Cancer
3040-***	*	[20-30[	Heart
3040-***	*	[20-30[	Flu

Aggarwal, et al. (2005) describes with extensive detail how k-anonymity algorithms are constructed and also proves the k-anonymity problem NP-hardness (a problem is said to be NP-hard if its complexity is not linear).

5.2. L-diversity

Machanavajjhala, Kifer, Gehrke, & Venkitasubramaniam, (2007) proposed an algorithm to mitigate some problems existing with k-anonymity. According to these authors, there are at least k records in a k-anonymous table that share the combination of values in the quasi-identifiers. With the combination of these values, if the sensitive attribute is the same, it is still possible to link the anonymous information to a specific individual. The main difference between the k-anonymity algorithm and the l-diversity algorithm is that the first one works on the quasi-identifier attributes while the second one works on the sensitive attributes (Sharma, Choudhary, & Jain, 2019).

The l-diversity algorithm, based on l-diversity principle, requires that all the quasi-identifier (QID) group contains, at least, l “well-represented” values for the sensitive attributes. The clarification of what “well-represented” values is made on top of defining entropy l-diversity based on the number of sensitive attributes that are equal. The reader is invited to Machanavajjhala, Kifer, Gehrke, & Venkitasubramaniam, (2007), for a deeper acquaintance with this type of algorithm.

With this algorithm, the data publisher does not need to have the same knowledge as the adversary (Machanavajjhala, Kifer, Gehrke, & Venkitasubramaniam, 2007).

One problem of the l-diversity algorithm is that it cannot prevent probabilistic inference attacks because some sensitive attributes are more frequent than others in the same QID group (Fung, Wang, Chen, & Yu, 2010).

If the l-diversity algorithm is well used, the optimal solution will reduce the information loss (Ghinita, Karras, Kalnis, & Mamoulis, 2007), otherwise, when the sensitive attributes are very similar, the information loss can be very high (Fung, Wang, Chen, & Yu, 2010).

The attacks that cannot be prevented with the use of l-diversity are skewness attacks and similarity attacks (Sharma, Choudhary, & Jain, 2019).

Using Table 14 as an example, it is possible to verify that if the attacker knows someone from the area where the ZipCode starts with 3080 and has an age between 50 and 59 (inclusive) s/he will discover that the individual suffers from a disease in the heart.

Therefore, the l-diversity algorithm is needed. Creating a 3-diverse table, the groups where the sensitive attribute has the same value for all instances will now have three different values. This algorithm application to Table 14 can be verified in Table 15.

Table 15 - 3-diverse data table

Quasi-identifier Attribute			**Sensitive Attribute**
ZipCode	**Sex**	**Age**	**Disease**
3090-***	*	[30-40[	Cancer
3090-***	*	[30-40[	Hepatitis
3090-***	*	[30-40[	Flu
3090-***	*	[30-40[	Heart
3080-***	*	[50-60[	Heart
3080-***	*	[50-60[	Hepatitis
3080-***	*	[50-60[	Heart
3080-***	*	[50-60[	Cancer
3040-***	*	[20-30[	Hepatitis

Quasi-identifier Attribute			**Sensitive Attribute**
ZipCode	**Sex**	**Age**	**Disease**
3040-***	*	[20-30[	Cancer
3040-***	*	[20-30[	Heart
3040-***	*	[20-30[	Flu

Observing now Table 15 one may verify that, if the attacker knows someone from the ZipCode 3080-*** and has an age between 50 and 60 years, s/he cannot know exactly the disease that the person suffers.

5.3. T-closeness

Li, Li, & Venkatasubramanian, (2007) noticed that the l-diversity algorithm had some vulnerabilities, and that l-diversity was not even required to prevent attribute disclosure. Therefore, they propose a new privacy notion and consequently a new algorithm called t-closeness.

The algorithm is deduced and constructed based on the following definition:

Definition 2 (T-closeness) (Li, Li, & Venkatasubramanian, 2007):

"An equivalence class is said to have t-closeness if the distance between the distribution of a sensitive attribute in this class and the distribution of the attribute in the whole table is no more than a threshold t. A table is said to have t-closeness if all equivalence classes have t-closeness."

This means that the distribution of an attribute in an equivalence class and its distribution in the whole table is required to be close. To measure this distance between both distributions, Li, Li, & Venkatasubramanian, (2007) used the Earth Mover Distance (EMD) function.

The EMD function is defined in the following way:

Definition 3 (EMD function) (Li, Li, & Venkatasubramanian, 2007)*:*

Having two probabilistic distributions $P = (p_1, p_2, \dots, p_m)$ *and* $Q = (q_1, q_2, \dots, q_m)$*, we want to find the flow of mass from element i of P to element j of Q that minimizes the overall work defined as:*

$$WORK(P,Q,F) = \sum_{i=1}^{m} \sum_{j=1}^{m} d_{ij} f_{ij},$$

where d_{ij} is the ground distance between element i of P and element j of Q. The minimization of the overall work must be performed subjected to three constraints that will guarantee that P is transformed to Q by the mass flow F. Those constraints are:

(1) $f_{ij} \geq 0, 1 \leq i \leq m; 1 \leq j \leq m$
(2) $p_i - \sum_{j=1}^{m} f_{ij} + \sum_{j=1}^{m} f_{ji} = q_i, \; 1 \leq i \leq m$
(3) $\sum_{i=1}^{m} \sum_{j=1}^{m} f_{ij} = \sum_{i=1}^{m} p_i = \sum_{i=1}^{m} q_i = 1$

The set of the above three restrictions ensure that: the mass flow between the attribute equivalence class and its table representation is positive (restriction 1); there is no loss of mass in the process since the mass that flows from one part to the other obeys to a mass conservation law) (restriction 2); and that the flow $f_{ij}, i = 1,\ldots,m; j = 1,\ldots,m$, the $, p_i, i = 1,\ldots,m$ and $q_i, i = 1,\ldots,m$, are, all three, probability distributions (restriction 3). To calculate the values $d_{ij}, i = 1,\ldots,m; j = 1,\ldots,m$, the reader is invited to follow the steps detailed on section 4 of Li, Li, & Venkatasubramanian, (2007) paper.

Once the transportation problem is solved, the EMD is defined to be the total work:

$$D[P,Q] = WORK\ (P,Q,F) = \sum_{i=1}^{m} \sum_{j=1}^{m} d_{ij} f_{ij}$$

The usability of the t-closeness algorithm has some limitations, such as:

- The difficult to specifying different protection levels for different sensitive values;
- EMD function is not capable to protect against attribute linkage in numerical attributes;
- It could generate a huge degradation in the data utility since the same distribution needs to be applied to every QID group.

5.4. δ-disclosure privacy

According to Brickell & Shmatikov (2008), sensitive attribute disclosure occurs when the adversary learns information about an individual's sensitive attribute(s). Therefore, these authors observed that the k-anonymity algorithms do not guarantee strong privacy to the users. They defined a table as δ-disclosure privacy if the distribution of the values in the sensitive attributes within each quasi-identifier is almost the same as their distribution in all tables. The algorithm is based on privacy metrics defined considering syntactic properties of the anonymised databases. The metrics ‘definitions and experiments performed with known, for being used on this type of research, databases are well described in the referred paper.

5.5. δ-presence

Nergiz, Atzori, & Clifton, (2007) showed in their paper that their algorithm is a good solution when the already existing algorithms are not the most appropriate.

They derive the algorithm based on the definition of what is intend by a table holding δ-presence. This definition states that:

Having a public table P and a private table T, δ-presence holds for a generalization T* of T, with $\delta = (\delta_{min}, \delta_{max})$ if

$$\delta_{min} \leq P(t \in T|T *) \leq \delta_{max}$$

With this definition a tuple $t \in P$ is said to be δ- present in T and $\delta = (\delta_{min}, \delta_{max})$ is a range of acceptable probabilities for $P(t \in T|T *)$.

Nergiz, Atzori, & Clifton, (2007) presents in their paper two types of δ- presence algorithms: Single-Dimensional Presence and Multi-Dimensional Presence. Both types are described and exemplified in the referred paper, therefore we excuse us to present an example for this type of algorithms.

6. Data Anonymization Tools

Data anonymization tools are used to help people with the anonymization of data. They help to anonymize a big amount of data in a faster and simpler way and can prevent users from introducing errors in the anonymized dataset.

The anonymization tools presented in this chapter are open-source tools that could be used by everyone. These tools have a lot of different algorithms, techniques and could have performance discrepancies in the same environment.

Sartor (2019), together with the company Aircloak GmbH, a company whose major concern is the responsible use of personal data, analysed the top 5 data anonymization tools in 2019 and appoints ARX Data Anonymization Tool, Amnesia, sdcMicro and µ-ARGUS as the best anonymization tools. These are not the solution for big companies since data anonymization is a very complex process and it should be done by use case and not dataset by dataset.

In this chapter, the following tools are going to be studied: ARX Data Anonymization Tool, Amnesia and UTD Anonymization Toolbox.

To help to learn a bit more about each tool, a small dataset was created to upload for each tool and try to obtain some results. The small dataset created has only twenty records and seven attributes, as can be verified in Table 16. The attributes are ID, Name, Age, Sex, Job, Social_Security_Number, and Disease.

Table 16 - Small Dataset to Test Different Anonymization Tools

ID	Name	Age	Sex	Job	Social_Security_Number	Disease
1	Joana	26	Female	Teacher	273649999	Cancer
2	João	27	Male	Scientist	537957333	Hypertense
3	Carolina	22	Female	Engineer	658658224	Diabetes
4	Ana	21	Female	Accountant	125896345	Aneurysm
5	Paulo	10	Male	Teacher	546895665	Cholesterol
6	Maria	56	Female	Scientist	756985321	Cancer
7	Pedro	45	Male	Engineer	120365120	Hypertense
8	Ricardo	76	Male	Accountant	125639856	Diabetes
9	Francisca	34	Female	Teacher	125630235	Aneurysm
10	Maria	36	Female	Scientist	896451230	Cholesterol
11	Teresa	44	Female	Engineer	890246879	Cancer
12	Vera	23	Female	Accountant	215634895	Hypertense
13	Rui	19	Male	Teacher	123456789	Diabetes
14	Tiago	67	Male	Scientist	120365201	Aneurysm
15	Henrique	43	Male	Engineer	890560235	Cholesterol
16	Carlos	89	Male	Accountant	452782982	Cancer

ID	Name	Age	Sex	Job	Social_Security_Number	Disease
17	Rita	34	Female	Teacher	325658985	Hypertense
18	Catarina	17	Female	Scientist	415756324	Diabetes
19	Micaela	21	Female	Engineer	120120120	Aneurysm
20	Micael	42	Male	Accountant	321654987	Cholesterol

The dataset is a CSV file, and each field is separated by commas.

6.1. ARX Data Anonymization Tool

ARX Data Anonymization Tool is a free open-source tool used to anonymise sensitive personal data. It is used in different contexts, like commercial big data analytics platforms, research projects, clinical trial data sharing and training (ARX, s.d.).

Analysing the tools in Sartor (2019) in terms of the number of algorithms and techniques, it is possible to conclude that ARX Data Anonymization is one of the most complete tools in terms of the usage of anonymization algorithms and techniques.

The anonymization algorithms supported by ARX are the following (ARX, s.d.):

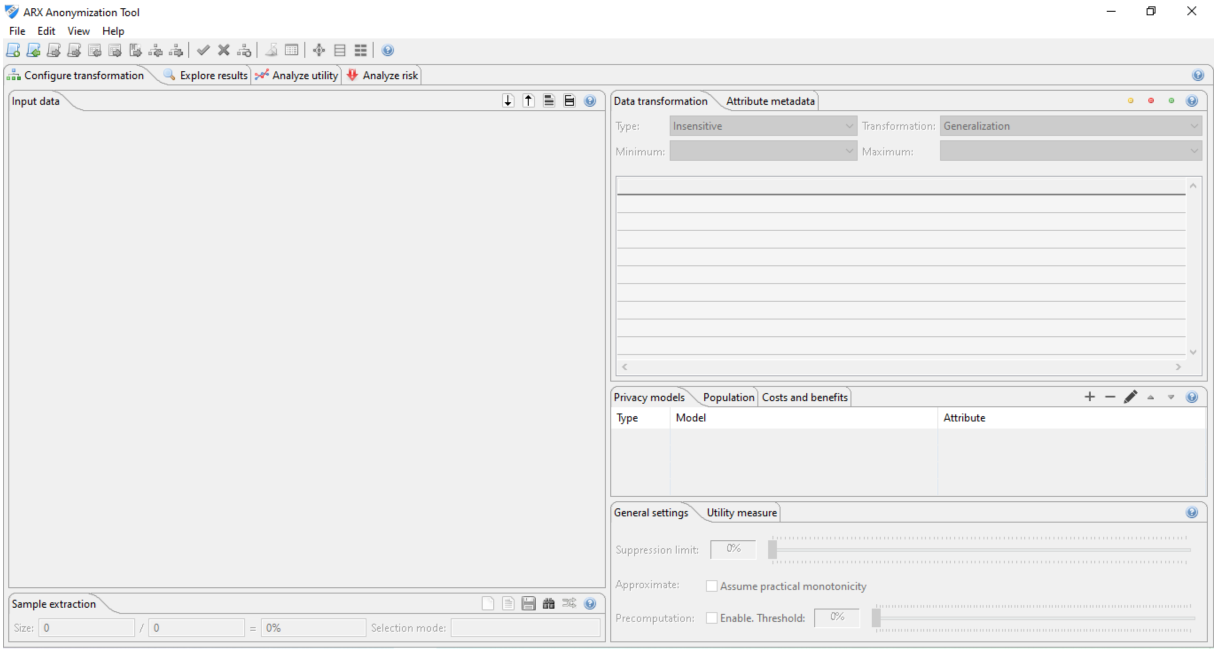
- k-Anonymity
- k-Map
- l-Diversity
- t-Closeness
- δ-Disclosure privacy
- β-Likeness
- δ-Presence
- (ε,δ)-differential privacy

The anonymization techniques supported by ARX are the following (ARX, s.d.):

- Global and local transformation schemes
- Random sampling
- Generalization
- Record, attribute, and cell suppression
- Microaggregation
- Top and bottom coding
- Categorization

To use the tool, the first necessary thing is to create a new project in the tool and upload a dataset. After the upload it is needed to choose the attribute types, if they are insensitive, sensitive, quasi-identifying or identifying. Then, set the correct values for attribute metadata, where the different levels for which the data can be anonymized also needs to be set. The next step is to select in the tool a privacy model for each attribute. After everything is set, the anonymization of the dataset can be done, and the output results are delivered so that they can be analysed.

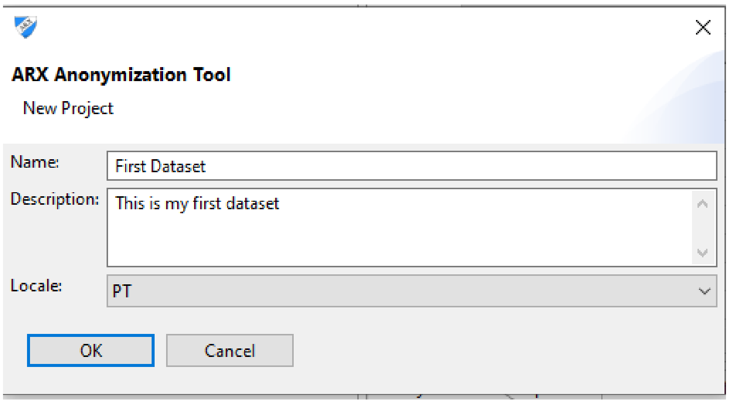
The first image the user gets when the tool is opened is the following one, shown in Figure 4.



Type: Insensitive
Transformation: Generalization
Minimum:
Maximum:
Suppression limit: 0%
Approximate: □ Assume practical monotonicity
Precomputation: □ Enable. Threshold: 0%
Size: 0 / 0 = 0% Selection mode:

Figure 4 - Initial Display of ARX

As mentioned before, the first thing is to create a new project. Going to the menu File and clicking on a new project, the pop-up window is displayed, which is shown in Figure 5.

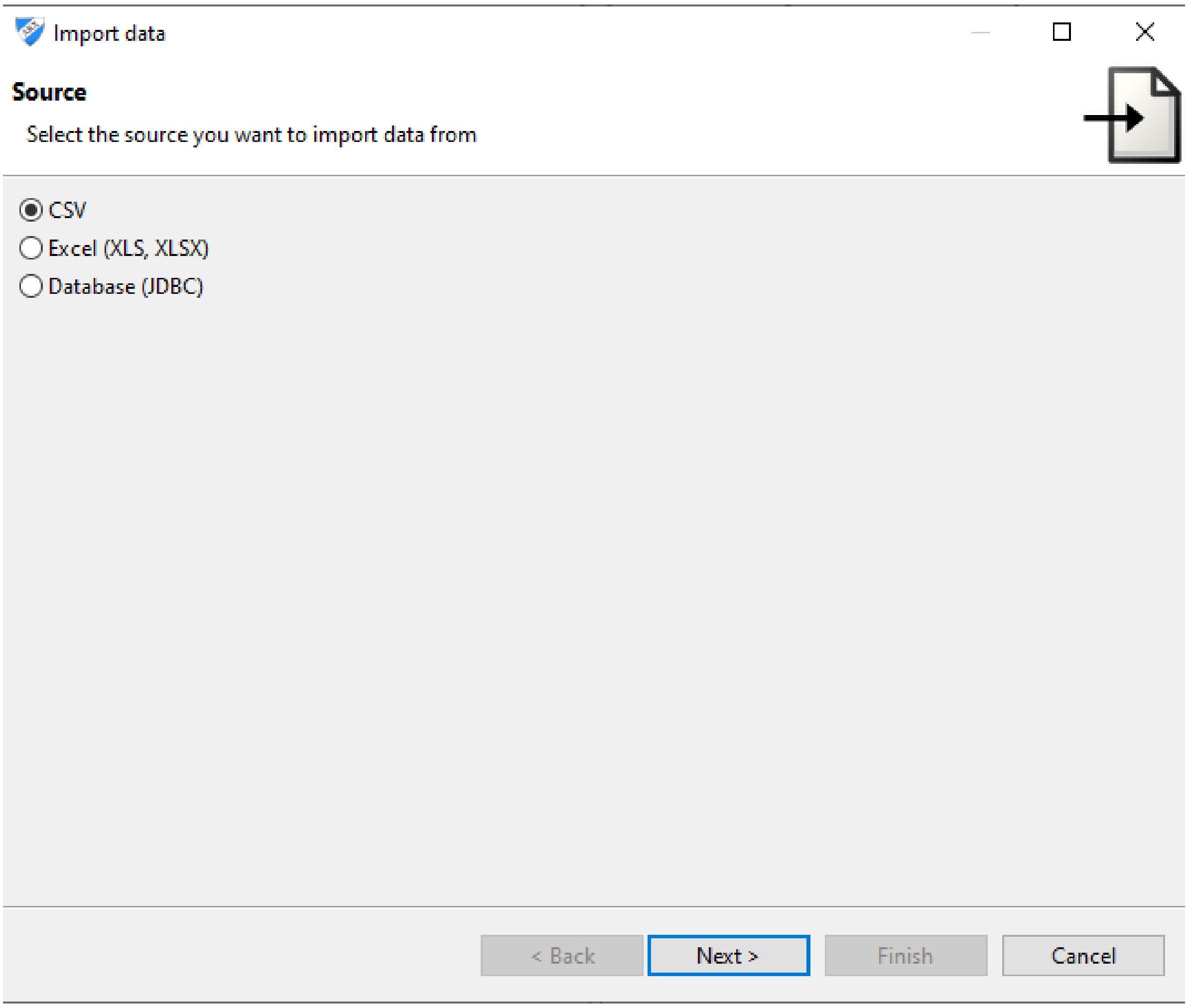


Name: First Dataset
Description: This is my first dataset
Locale: PT

Figure 5 - Create Project of ARX

In that pop-up window, a name and a small description for the project should be set. Then click OK and start working on the project. This project can be saved and used later if the user wants to save the definitions and the changes done for the uploaded dataset.

The next step is to insert the dataset. Going again to menu File, but this time click in “Import data”, a new window is presented. A new window is displayed where the data source type should be chosen. As the dataset was created in CSV format, the CSV type should be selected as shown in Figure 6.



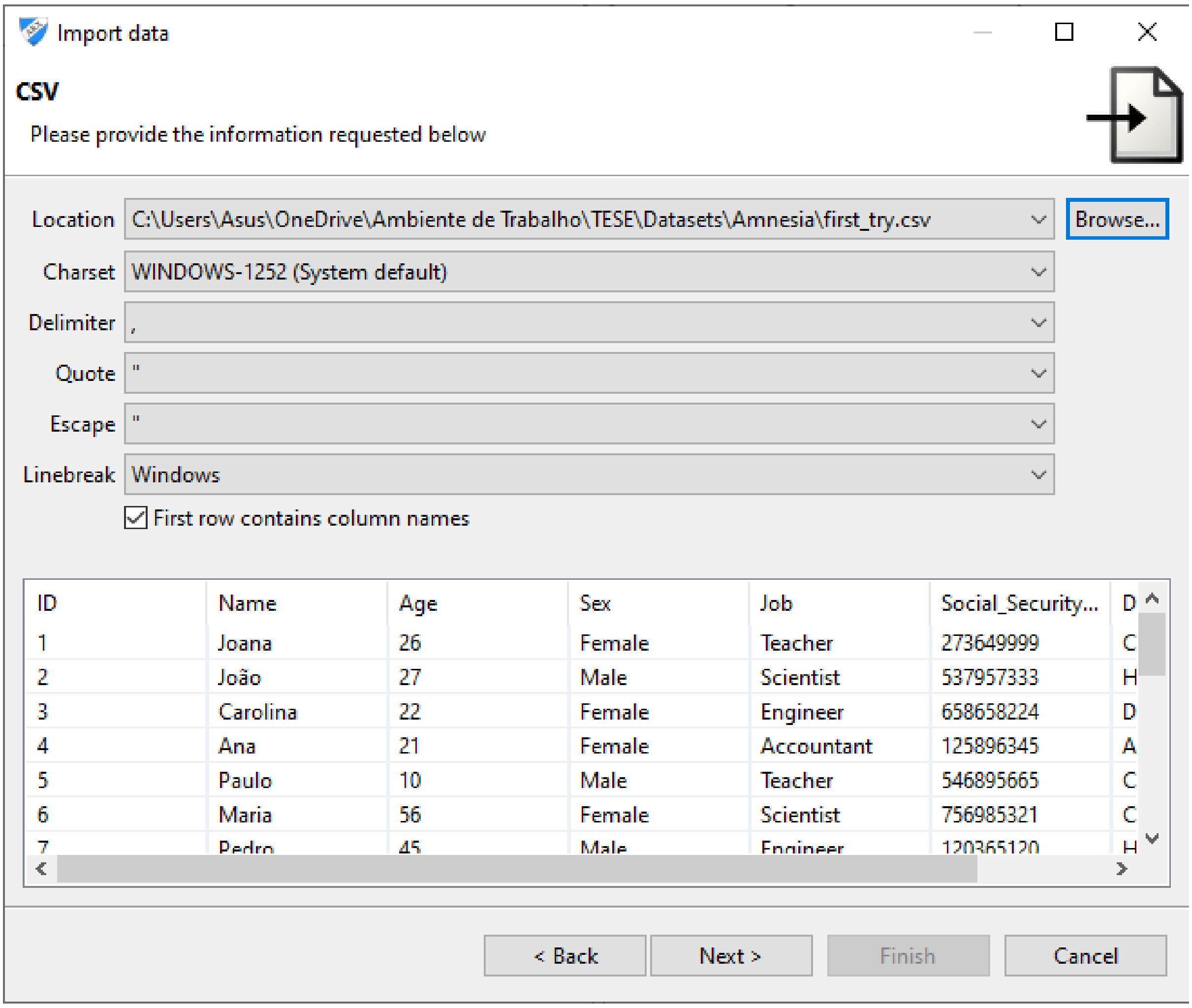
Source

Select the source you want to import data from

● CSV
○ Excel (XLS, XLSX)
○ Database (JDBC)

Figure 6 - Import data window – Source Type

After choosing the import type, some more information needs to be provided, like the location of the dataset, the delimiter, the linebreak type and other important definitions. The last step in this importation process is where a preview of the used dataset is displayed, as can be seen in Figure 7.



Location C:\Users\Asus\OneDrive\Ambiente de Trabalho\TESE\Datasets\Amnesia\first_try.csv

Charset WINDOWS-1252 (System default)

Delimiter ,

Quote "

Escape "

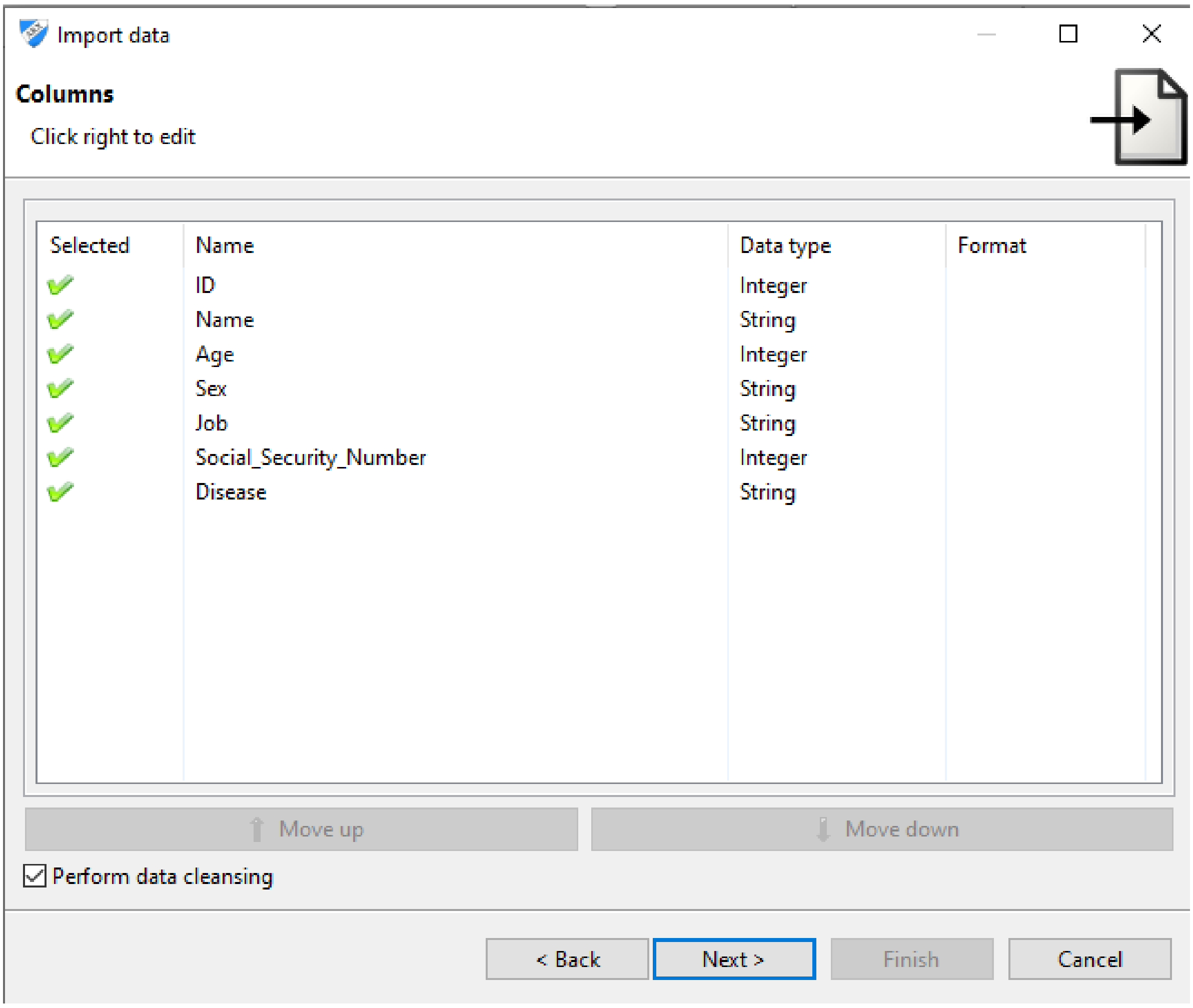
Linebreak Windows

☑ First row contains column names

ID	Name	Age	Sex	Job	Social_Security...	D
1	Joana	26	Female	Teacher	273649999	C
2	João	27	Male	Scientist	537957333	H
3	Carolina	22	Female	Engineer	658658224	D
4	Ana	21	Female	Accountant	125896345	A
5	Paulo	10	Male	Teacher	546895665	C
6	Maria	56	Female	Scientist	756985321	C
7	Pedro	45	Male	Engineer	120365120	H

Figure 7 - Import data window – Dataset Upload

The next window presented has the column types for each attribute of the dataset, as shown in Figure 8. These attribute types are already set, the user just needs to verify and correct if something is wrong.



Columns

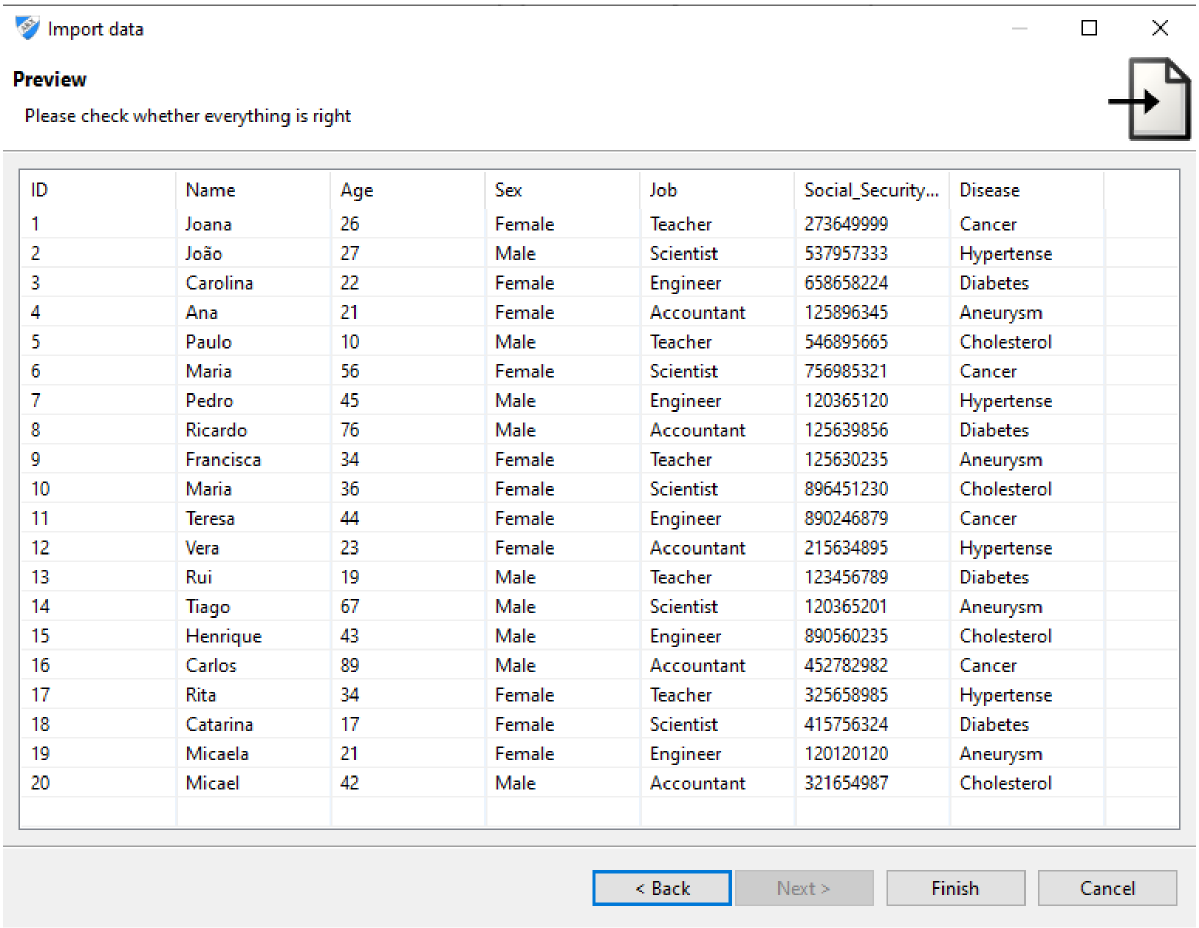
Click right to edit

Selected	Name	Data type	Format
✓	ID	Integer	
✓	Name	String	
✓	Age	Integer	
✓	Sex	String	
✓	Job	String	
✓	Social_Security_Number	Integer	
✓	Disease	String	

☑ Perform data cleansing

Figure 8 - Import data window – Column’s type

The last phase in this import data process is where the last preview of our dataset is shown to the user. An example of this preview is shown in Figure 9.



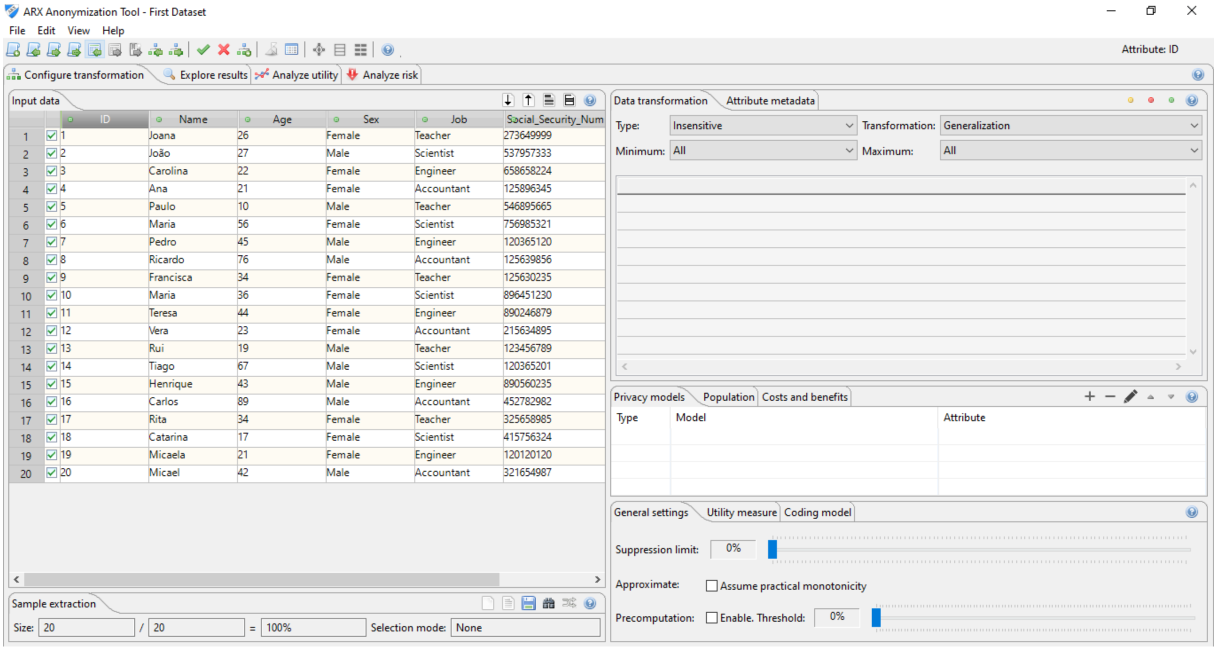
Preview

Please check whether everything is right

ID	Name	Age	Sex	Job	Social_Security...	Disease
1	Joana	26	Female	Teacher	273649999	Cancer
2	João	27	Male	Scientist	537957333	Hypertense
3	Carolina	22	Female	Engineer	658658224	Diabetes
4	Ana	21	Female	Accountant	125896345	Aneurysm
5	Paulo	10	Male	Teacher	546895665	Cholesterol
6	Maria	56	Female	Scientist	756985321	Cancer
7	Pedro	45	Male	Engineer	120365120	Hypertense
8	Ricardo	76	Male	Accountant	125639856	Diabetes
9	Francisca	34	Female	Teacher	125630235	Aneurysm
10	Maria	36	Female	Scientist	896451230	Cholesterol
11	Teresa	44	Female	Engineer	890246879	Cancer
12	Vera	23	Female	Accountant	215634895	Hypertense
13	Rui	19	Male	Teacher	123456789	Diabetes
14	Tiago	67	Male	Scientist	120365201	Aneurysm
15	Henrique	43	Male	Engineer	890560235	Cholesterol
16	Carlos	89	Male	Accountant	452782982	Cancer
17	Rita	34	Female	Teacher	325658985	Hypertense
18	Catarina	17	Female	Scientist	415756324	Diabetes
19	Micaela	21	Female	Engineer	120120120	Aneurysm
20	Micael	42	Male	Accountant	321654987	Cholesterol

Figure 9 - Import data window – Dataset Preview

Finally, the last step is to click on Finish and the dataset is uploaded in the tool. The uploaded dataset appears in the Import Data separator, as shown in Figure 10.



		ID	Name	Age	Sex	Job	Social_Security_Num
1	☑	1	Joana	26	Female	Teacher	273649999
2	☑	2	João	27	Male	Scientist	537957333
3	☑	3	Carolina	22	Female	Engineer	658658224
4	☑	4	Ana	21	Female	Accountant	125896345
5	☑	5	Paulo	10	Male	Teacher	546895665
6	☑	6	Maria	56	Female	Scientist	756985321
7	☑	7	Pedro	45	Male	Engineer	120365120
8	☑	8	Ricardo	76	Male	Accountant	125639856
9	☑	9	Francisca	34	Female	Teacher	125630235
10	☑	10	Maria	36	Female	Scientist	896451230
11	☑	11	Teresa	44	Female	Engineer	890246879
12	☑	12	Vera	23	Female	Accountant	215634895
13	☑	13	Rui	19	Male	Teacher	123456789
14	☑	14	Tiago	67	Male	Scientist	120365201
15	☑	15	Henrique	43	Male	Engineer	890560235
16	☑	16	Carlos	89	Male	Accountant	452782982
17	☑	17	Rita	34	Female	Teacher	325658985
18	☑	18	Catarina	17	Female	Scientist	415756324
19	☑	19	Micaela	21	Female	Engineer	120120120
20	☑	20	Micael	42	Male	Accountant	321654987

Size: 20 / 20 = 100% Selection mode: None

Type: Insensitive Transformation: Generalization
Minimum: All Maximum: All

Type	Model	Attribute

Suppression limit: 0%
Approximate: ☐ Assume practical monotonicity
Precomputation: ☐ Enable. Threshold: 0%

Figure 10 - ARX Display with Dataset Uploaded

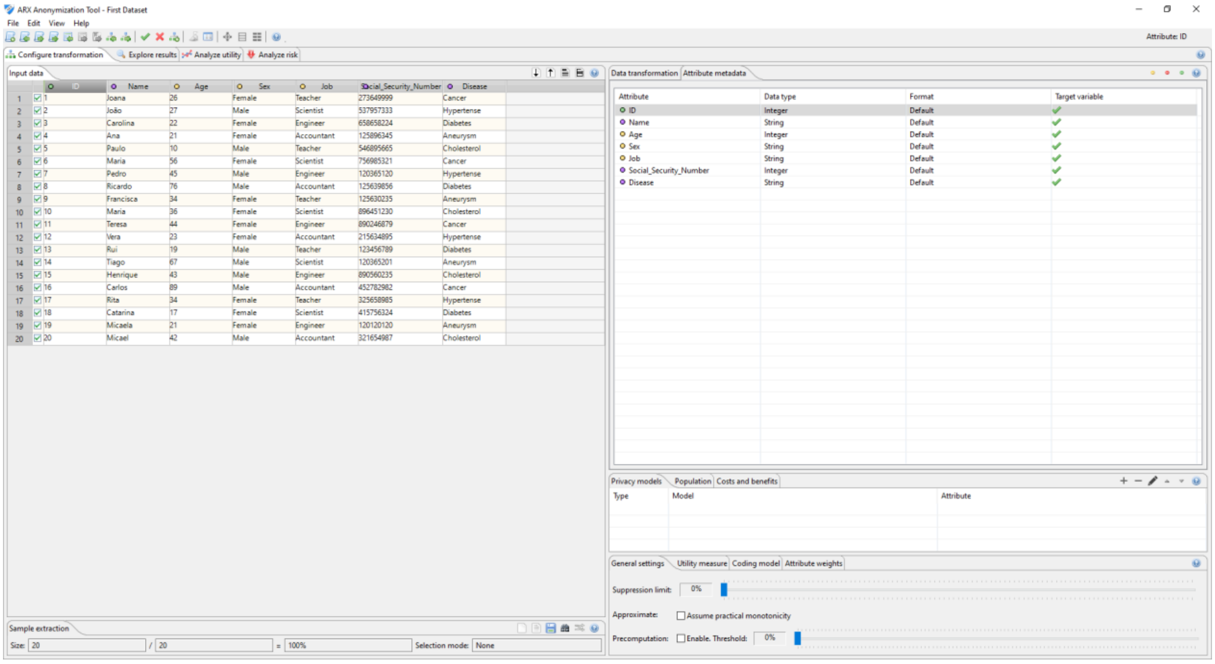
The next step is to set the values needed for data transformation, and they are the attribute types, transformation, minimum, and maximum. As it was mentioned before, the attribute types are insensitive, sensitive, quasi-identifying or identifying. For the uploaded dataset, the attribute types were set in the following way:

- ID -> insensitive
- Name -> Sensitive
- Age -> Quasi-Identifying
- Sex -> Quasi-Identifying
- Job -> Quasi-Identifying
- Social_Security_Number -> Sensitive
- Disease -> Sensitive

For the transformation, the user can choose between generalization, microaggregation, clustering and microaggregation techniques. The value chosen was generalization because it was the same anonymization technique used in other tools.

The parameters Minimum and Maximum only allow the value All.

After selecting these values for the Data transformation, the attribute metadata values need to be verified. In this tab, is displayed the data type for each attribute and mark it as a target variable or not. It is difficult to know what the developers mean as Target Variable. Clicking on the help icon it is not mentioned in the Target Variable column but a Response Variable instead. It is not possible to know for sure that Target Variable are the same thing because it is not explicit in the tool documentation. For this example, a check was chosen for the value for each attribute in the Target Variable column, as can be verified in Figure 11.



		ID	Name	Age	Sex	Job	Social_Security_Number	Disease
1	☑	1	Joana	26	Female	Teacher	273649999	Cancer
2	☑	2	João	27	Male	Scientist	537957333	Hypertense
3	☑	3	Carolina	22	Female	Engineer	658658224	Diabetes
4	☑	4	Ana	21	Female	Accountant	125896345	Aneurysm
5	☑	5	Paulo	10	Male	Teacher	546895665	Cholesterol
6	☑	6	Maria	56	Female	Scientist	756985321	Cancer
7	☑	7	Pedro	45	Male	Engineer	120365120	Hypertense
8	☑	8	Ricardo	76	Male	Accountant	125639856	Diabetes
9	☑	9	Francisca	34	Female	Teacher	125630235	Aneurysm
10	☑	10	Maria	36	Female	Scientist	896451230	Cholesterol
11	☑	11	Teresa	44	Female	Engineer	890246879	Cancer
12	☑	12	Vera	23	Female	Accountant	215634895	Hypertense
13	☑	13	Rui	19	Male	Teacher	123456789	Diabetes
14	☑	14	Tiago	67	Male	Scientist	120365201	Aneurysm
15	☑	15	Henrique	43	Male	Engineer	890560235	Cholesterol
16	☑	16	Carlos	89	Male	Accountant	452782982	Cancer
17	☑	17	Rita	34	Female	Teacher	325658985	Hypertense
18	☑	18	Catarina	17	Female	Scientist	415756324	Diabetes
19	☑	19	Micaela	21	Female	Engineer	120120120	Aneurysm
20	☑	20	Micael	42	Male	Accountant	321654987	Cholesterol

Attribute	Data type	Format	Target variable
ID	Integer	Default	✓
Name	String	Default	✓
Age	Integer	Default	✓
Sex	String	Default	✓
Job	String	Default	✓
Social_Security_Number	Integer	Default	✓
Disease	String	Default	✓

Type	Model	Attribute

Suppression limit: 0%

Approximate: □ Assume practical monotonicity

Precomputation: □ Enable. Threshold: 0%

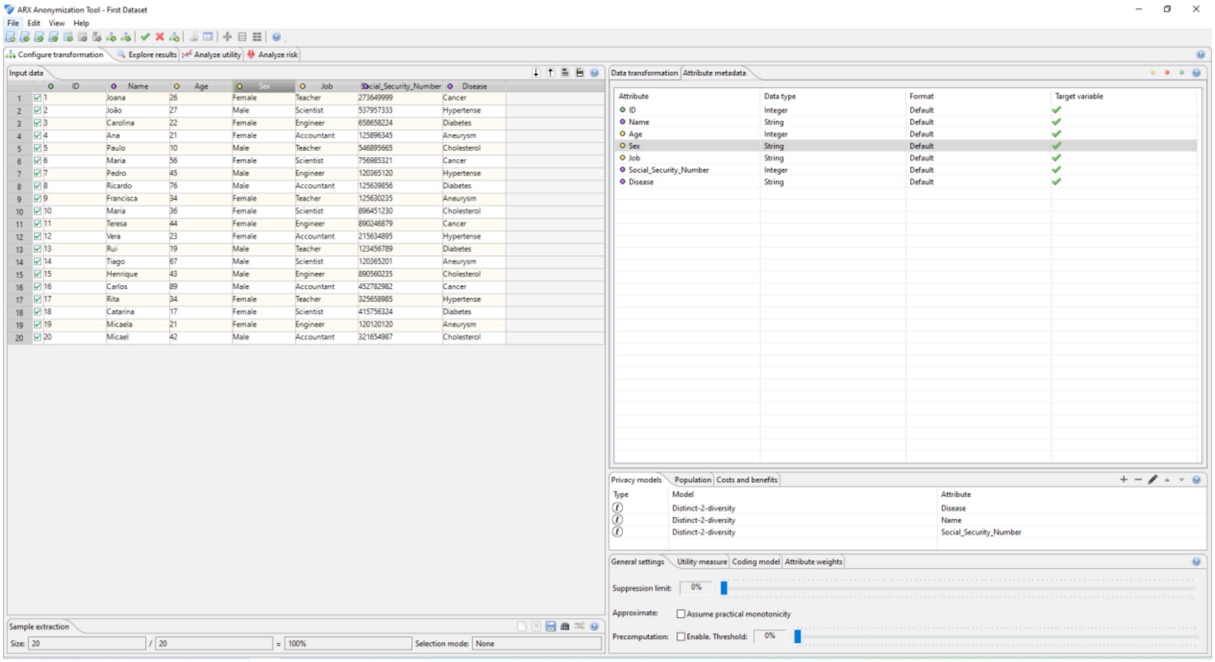
Size: 20 / 20 = 100% Selection mode: None

Figure 11 - ARX Display with Attribute Metadata Set

Now, the next step is defining the privacy models for the sensitive attributes. It should be noted that the attributes Name, Social_Security_Number, and Disease need a privacy model. The privacy model that was chosen is the k-anonymity algorithm for the sensitive attributes since it was the algorithm used in the other tools. But at the time to add a privacy model, it is shown a table with all privacy models and the attributes already linked to the privacy model. Here, the user could choose between different options for the same algorithm, but it does not allow to pick another privacy model if it is not already linked to the attribute. The existing privacy models are the following:

- (ε, δ) – Differential privacy
- K-Anonymity
- K-Map
- l-Diversity
- δ–presence
- t–Closeness
- δ–Disclosure privacy
- β–Likeness

All the dataset sensitive attributes were linked to the l–Diversity, t–Closeness, δ–Disclosure privacy, and β–Likeness. As the l–Diversity algorithm is the most similar one with k-anonymity this was the algorithm chosen, with l = 2. As can be verified in Figure 12, all the needed parameters are fields with values.



		ID	Name	Age	Sex	Job	Social_Security_Number	Disease
1	☑	1	Joana	26	Female	Teacher	273649999	Cancer
2	☑	2	João	27	Male	Scientist	537957333	Hypertense
3	☑	3	Carolina	22	Female	Engineer	658658224	Diabetes
4	☑	4	Ana	21	Female	Accountant	125896345	Aneurysm
5	☑	5	Paulo	10	Male	Teacher	546895665	Cholesterol
6	☑	6	Maria	56	Female	Scientist	756985321	Cancer
7	☑	7	Pedro	45	Male	Engineer	120365120	Hypertense
8	☑	8	Ricardo	76	Male	Accountant	125639856	Diabetes
9	☑	9	Francisca	34	Female	Teacher	125630235	Aneurysm
10	☑	10	Maria	36	Female	Scientist	896451230	Cholesterol
11	☑	11	Teresa	44	Female	Engineer	890246879	Cancer
12	☑	12	Vera	23	Female	Accountant	215634895	Hypertense
13	☑	13	Rui	19	Male	Teacher	123456789	Diabetes
14	☑	14	Tiago	67	Male	Scientist	120365201	Aneurysm
15	☑	15	Henrique	43	Male	Engineer	890560235	Cholesterol
16	☑	16	Carlos	89	Male	Accountant	452782982	Cancer
17	☑	17	Rita	34	Female	Teacher	325658985	Hypertense
18	☑	18	Catarina	17	Female	Scientist	415756324	Diabetes
19	☑	19	Micaela	21	Female	Engineer	120120120	Aneurysm
20	☑	20	Micael	42	Male	Accountant	321654987	Cholesterol

Attribute	Data type	Format	Target variable
ID	Integer	Default	✓
Name	String	Default	✓
Age	Integer	Default	✓
Sex	String	Default	✓
Job	String	Default	✓
Social_Security_Number	Integer	Default	✓
Disease	String	Default	✓

Type	Model	Attribute
ℓ	Distinct-2-diversity	Disease
ℓ	Distinct-2-diversity	Name
ℓ	Distinct-2-diversity	Social_Security_Number

Suppression limit: 0%

Approximate: □ Assume practical monotonicity

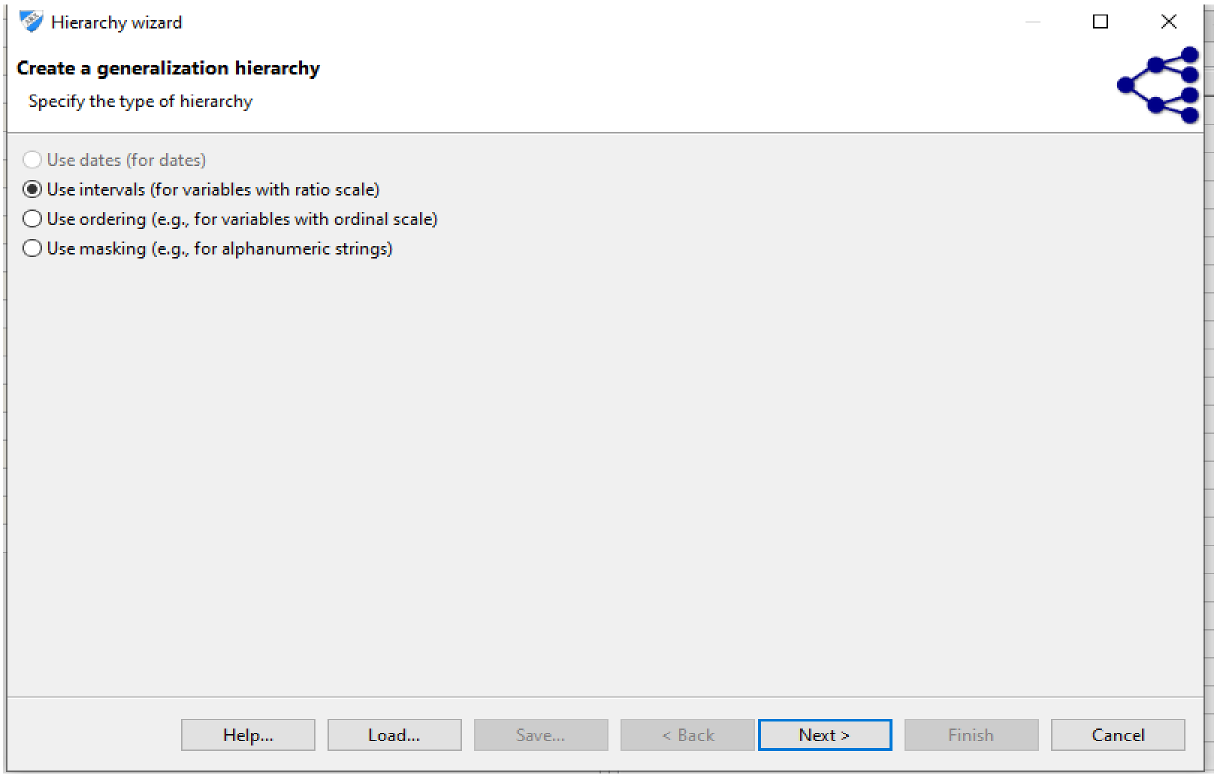
Precomputation: □ Enable. Threshold: 0%

Size: 20 / 20 = 100% Selection mode: None

Figure 12 - ARX Display with All Values Set

The next step is to create the hierarchies for the other attributes.

The first hierarchy created is for attribute Age. To create this hierarchy, the user needs to click on the attribute for which s/he wants and then in menu Edit -> Create hierarchy and the following window is opened. With these steps, the hierarchy type is automatically suggested according to the attribute type as illustrated in Figure 13.



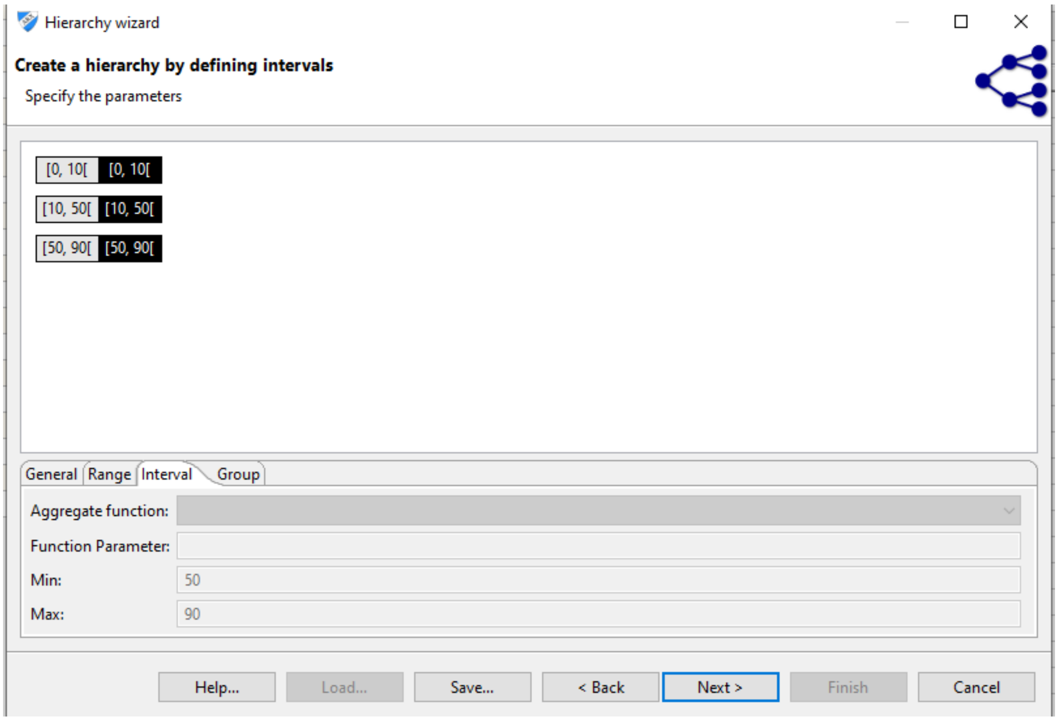
Create a generalization hierarchy

Specify the type of hierarchy

○ Use dates (for dates)
● Use intervals (for variables with ratio scale)
○ Use ordering (e.g., for variables with ordinal scale)
○ Use masking (e.g., for alphanumeric strings)

Figure 13 - Create Hierarchy – Age Attribute Hierarchy Type

The Use Intervals options were selected because the goal is to have the same hierarchy as used in other tools. Three levels to the Age hierarchy were created: [0,10[, [10,50[and [50,90[and they are represented in Figure 14.



Create a hierarchy by defining intervals

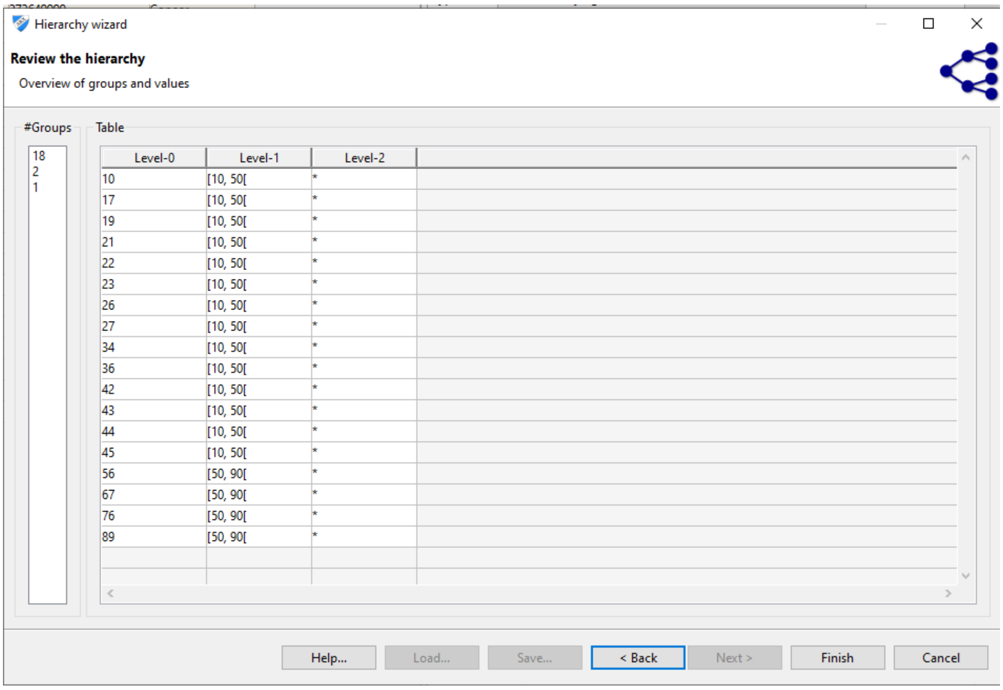
Specify the parameters

[0, 10[[0, 10[
[10, 50[[10, 50[
[50, 90[[50, 90[

Aggregate function:
Function Parameter:
Min: 50
Max: 90

Figure 14 - Create Hierarchy – Age Attribute Hierarchy Intervals

After setting the intervals, an overview of the groups and values for the age attribute is displayed in the tool, as shown in Figure 15.



Review the hierarchy

Overview of groups and values

#Groups

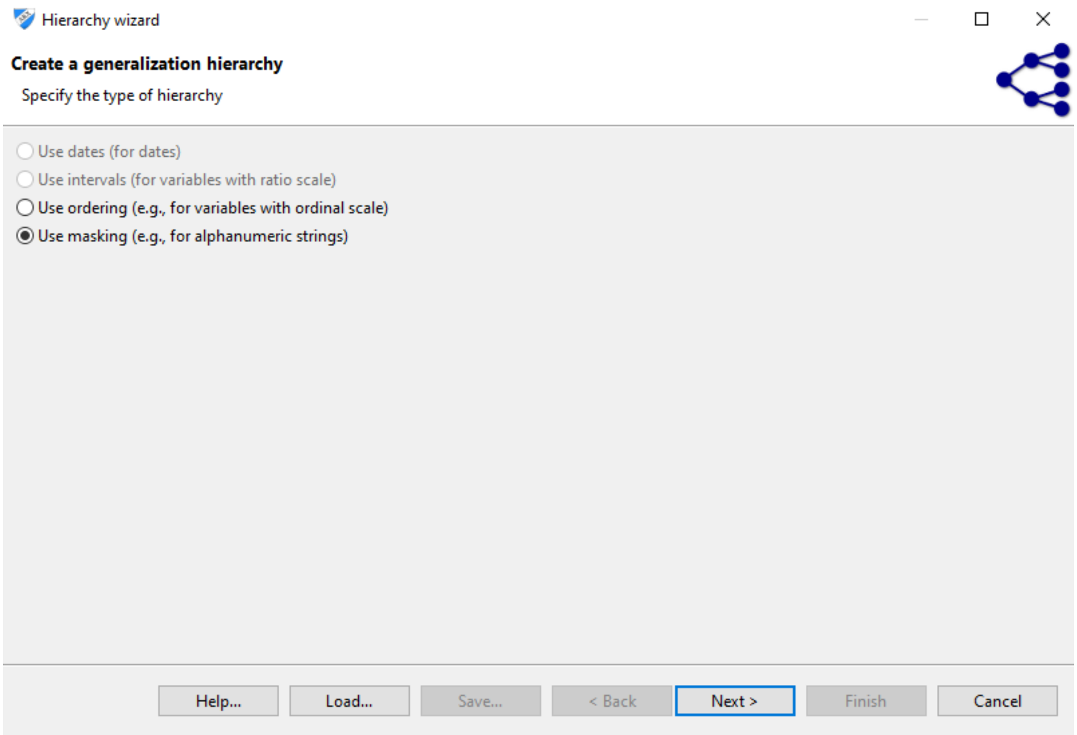
18
2
1

Table

Level-0	Level-1	Level-2
10	[10, 50[	*
17	[10, 50[	*
19	[10, 50[	*
21	[10, 50[	*
22	[10, 50[	*
23	[10, 50[	*
26	[10, 50[	*
27	[10, 50[	*
34	[10, 50[	*
36	[10, 50[	*
42	[10, 50[	*
43	[10, 50[	*
44	[10, 50[	*
45	[10, 50[	*
56	[50, 90[	*
67	[50, 90[	*
76	[50, 90[	*
89	[50, 90[	*

Figure 15 - Create Hierarchy – Age Attribute Hierarchy Preview

The next hierarchy created was for the Job attribute. The hierarchy type now used is the Masking type as shown in Figure 16.



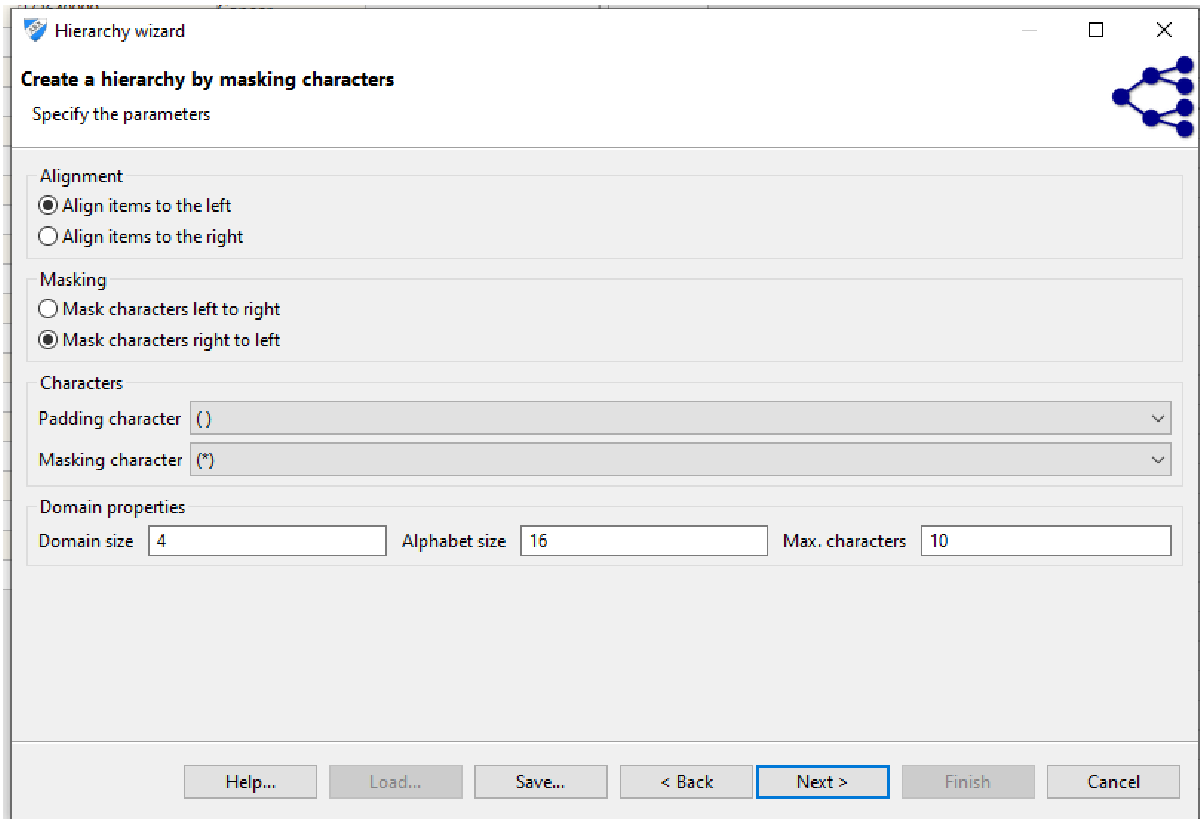
Create a generalization hierarchy

Specify the type of hierarchy

○ Use dates (for dates)
○ Use intervals (for variables with ratio scale)
○ Use ordering (e.g., for variables with ordinal scale)
● Use masking (e.g., for alphanumeric strings)

Figure 16 - Create Hierarchy – Job Attribute Hierarchy Type

With this hierarchy type, some characters of each attribute value are going to be suppressed, according to the hierarchy level where the user is working.



Hierarchy wizard

Create a hierarchy by masking characters

Specify the parameters

Alignment

● Align items to the left

○ Align items to the right

Masking

○ Mask characters left to right

● Mask characters right to left

Characters

Padding character ( )

Masking character (*)

Domain properties

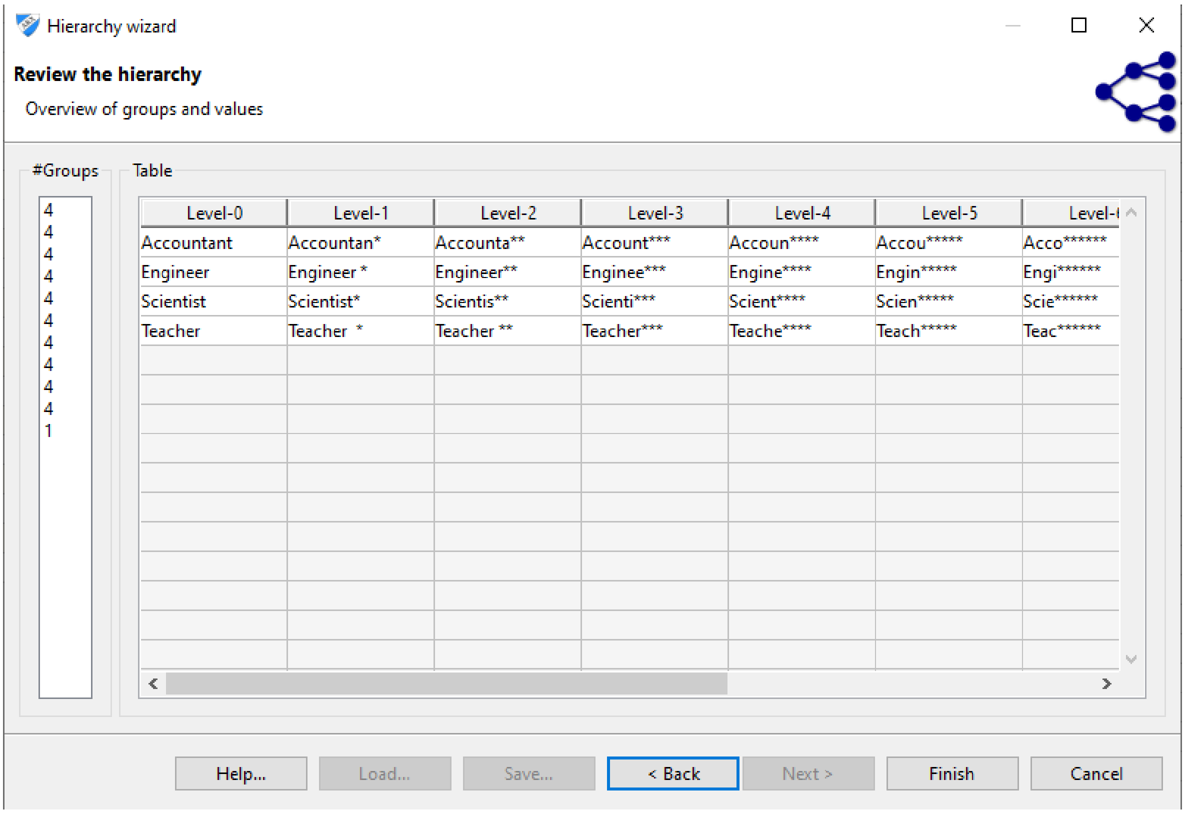
Domain size 4 Alphabet size 16 Max. characters 10

Figure 17 - Create Hierarchy – Job Attribute Hierarchy Masking Parameters

As it is displayed in Figure 17, the * is used as the character to mask and it is set to mask the values from the right to the left (therefore the option “Align items to the left” is selected.

Finally, the hierarchy results are shown in Figure 18.

In Figure 18 is possible to verify that the number of levels in this hierarchy, using the masking type, is the same as the number of characters of the biggest value of the original dataset for that attribute. In the smaller values of the attribute, some spaces are added to the value. For example, the value Accountant is the biggest value for Job attribute, with 10 characters, and Teacher is the smaller, with 7 characters. The Teacher value is filled with 3 spaces to perform the 10-value size attribute, therefore, in Level-1, Accountant has already started to be suppressed, but Teacher is still complete and has added spaces and, just at the end, has the special character *.



Review the hierarchy

Overview of groups and values

#Groups

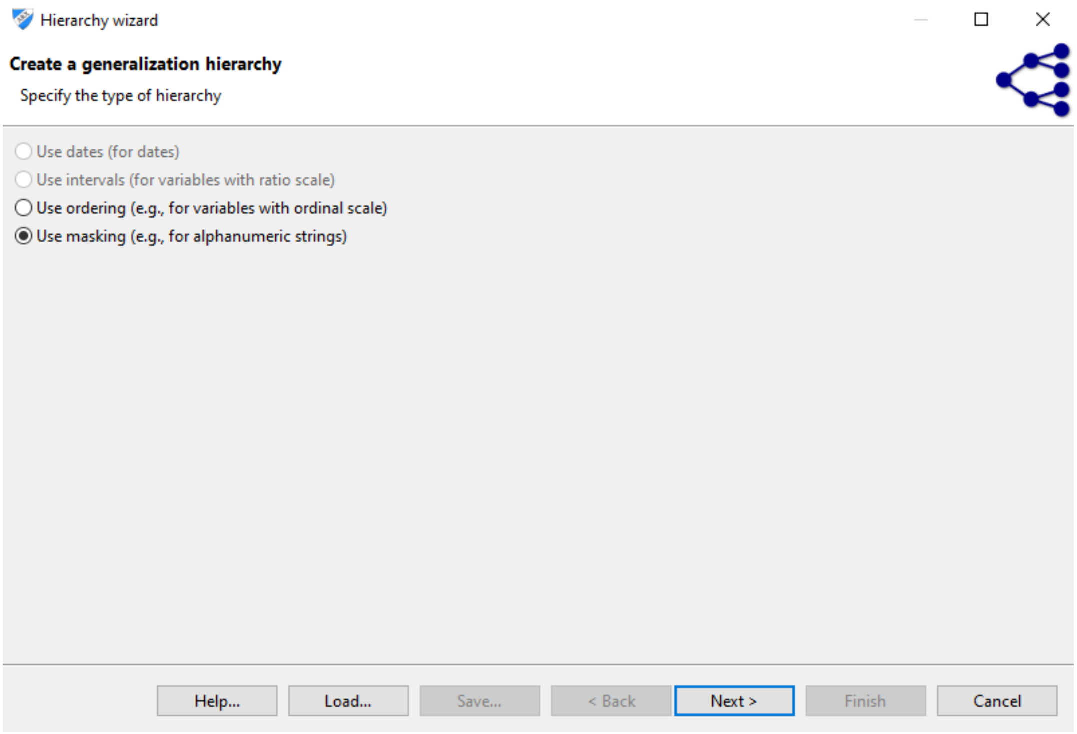
4
4
4
4
4
4
4
4
4
4
1

Table

Level-0	Level-1	Level-2	Level-3	Level-4	Level-5	Level-(
Accountant	Accountan*	Accounta**	Account***	Accoun****	Accou*****	Acco******
Engineer	Engineer *	Engineer**	Enginee***	Engine****	Engin*****	Engi******
Scientist	Scientist*	Scientis**	Scienti***	Scient****	Scien*****	Scie******
Teacher	Teacher *	Teacher **	Teacher***	Teache****	Teach*****	Teac******

Figure 18 - Create Hierarchy – Job Attribute Hierarchy Preview

Finally, the last attribute that needs a hierarchy is the Sex attribute. For this is also used the masking hierarchy type as shown in Figure 19.



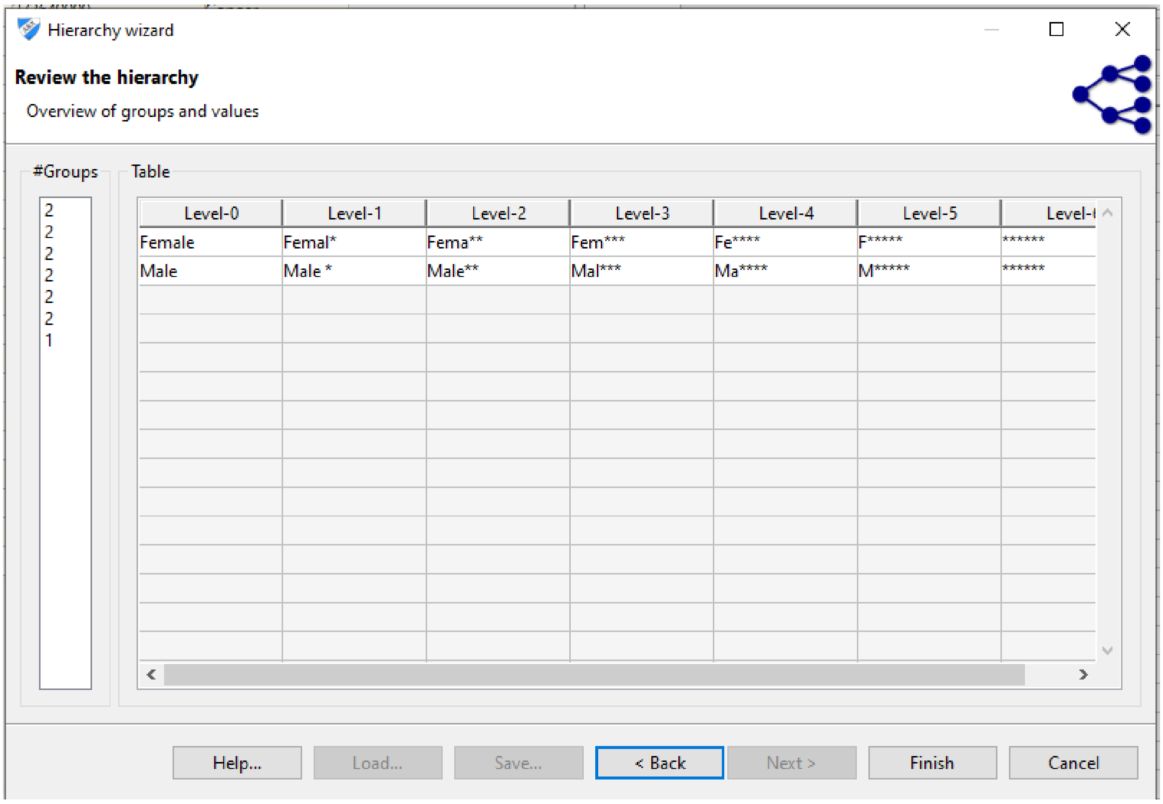
Create a generalization hierarchy

Specify the type of hierarchy

○ Use dates (for dates)
○ Use intervals (for variables with ratio scale)
○ Use ordering (e.g., for variables with ordinal scale)
● Use masking (e.g., for alphanumeric strings)

Figure 19 - Create Hierarchy – Sex Attribute Hierarchy Type

The parameters were set in the same way they were in the Job attribute hierarchy and the final hierarchy result is illustrated in Figure 20.



Hierarchy wizard

Review the hierarchy

Overview of groups and values

#Groups

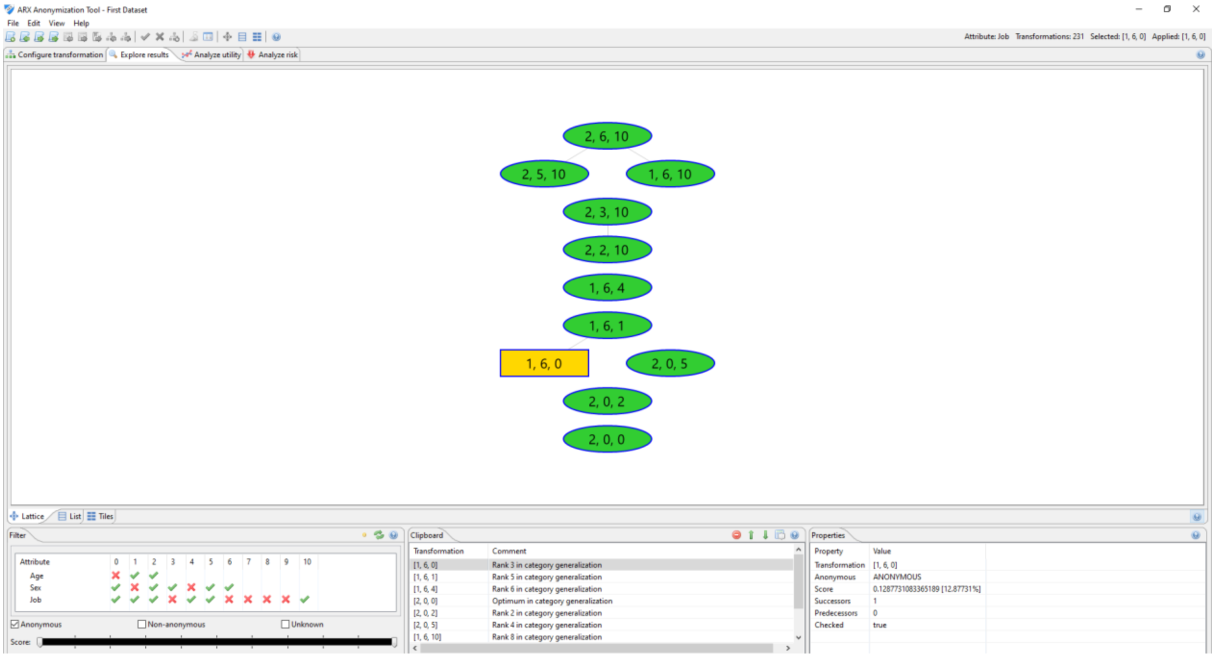
2
2
2
2
2
2
1

Table

Level-0	Level-1	Level-2	Level-3	Level-4	Level-5	Level-6
Female	Femal*	Fema**	Fem***	Fe****	F*****	******
Male	Male *	Male**	Mal***	Ma****	M*****	******

Figure 20 - Create Hierarchy – Sex Attribute Hierarchy Preview

After all the attributes have their anonymization properties set, the user can anonymize the uploaded dataset. The result for this anonymization process is a graph, where it is shown the different anonymization levels for the quasi-identifier attributes. The result is shown in Figure 21.



Attribute: Job Transformations: 231 Selected: [1, 6, 0] Applied: [1, 6, 0]

Attribute	0	1	2	3	4	5	6	7	8	9	10
Age	✗	✓	✓								
Sex	✓	✗	✓	✓	✗	✓	✓				
Job	✓	✓	✓	✗	✓	✓	✗	✗	✗	✗	✓

☑ Anonymous ☐ Non-anonymous ☐ Unknown

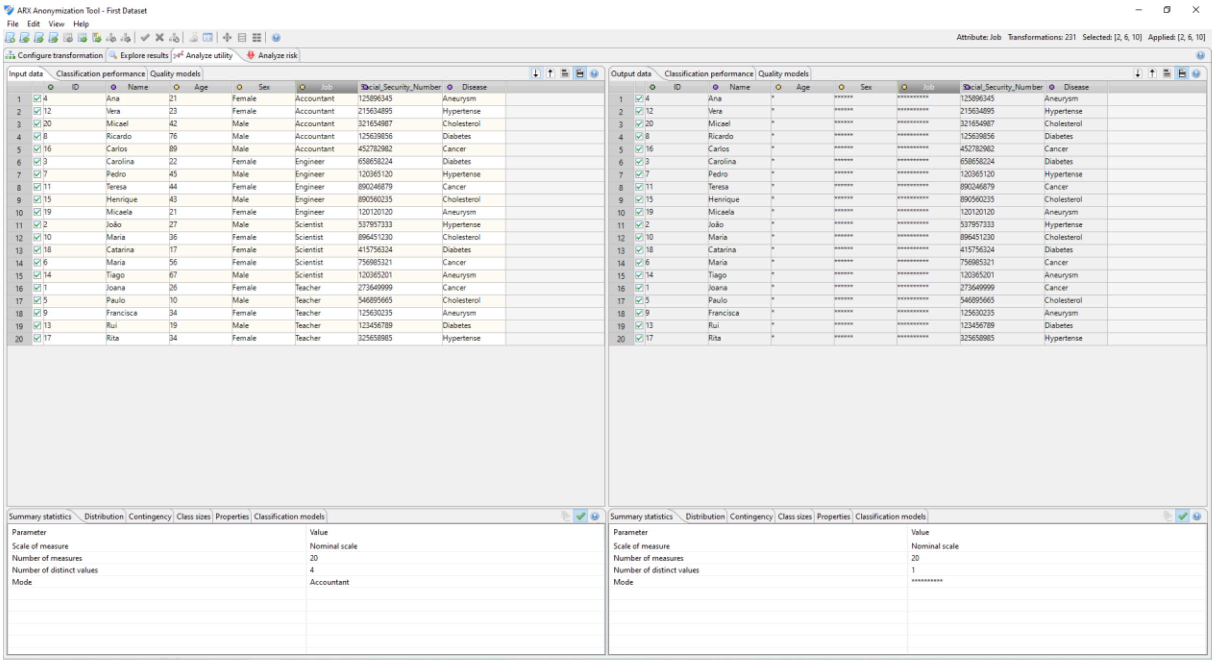
Transformation	Comment
[1, 6, 0]	Rank 3 in category generalization
[1, 6, 1]	Rank 5 in category generalization
[1, 6, 4]	Rank 6 in category generalization
[2, 0, 0]	Optimum in category generalization
[2, 0, 2]	Rank 2 in category generalization
[2, 0, 5]	Rank 4 in category generalization
[1, 6, 10]	Rank 8 in category generalization

Property	Value
Transformation	[1, 6, 0]
Anonymous	ANONYMOUS
Score	0.1287731083365189 [12.87731%]
Successors	1
Predecessors	0
Checked	true

Figure 21 - Anonymization Graph Results

For example, the first level, where the leaf has the values (2, 6, 10) means that Age stopped at anonymization level 2, Sex stopped at anonymization level 6 and Job at anonymization level 10. To see the results for this level, the user needs to Right-click in the green cell and choose the option “Apply transformation”.

Accessing the “Analyse utility” menu, the anonymized dataset is displayed, side-by-side with the original dataset, as it is displayed in Figure 22.



Attribute: Job Transformations: 231 Selected: [2, 6, 10] Applied: [2, 6, 10]

Input data

	ID	Name	Age	Sex	Job	Social_Security_Number	Disease
1	☑ 4	Ana	21	Female	Accountant	125896345	Aneurysm
2	☑ 12	Vera	23	Female	Accountant	215634895	Hypertense
3	☑ 20	Micael	42	Male	Accountant	321654987	Cholesterol
4	☑ 8	Ricardo	76	Male	Accountant	125639856	Diabetes
5	☑ 16	Carlos	89	Male	Accountant	452782982	Cancer
6	☑ 3	Carolina	22	Female	Engineer	658658224	Diabetes
7	☑ 7	Pedro	45	Male	Engineer	120365120	Hypertense
8	☑ 11	Teresa	44	Female	Engineer	890246879	Cancer
9	☑ 15	Henrique	43	Male	Engineer	890560235	Cholesterol
10	☑ 19	Micaela	21	Female	Engineer	120120120	Aneurysm
11	☑ 2	João	27	Male	Scientist	537957333	Hypertense
12	☑ 10	Maria	36	Female	Scientist	896451230	Cholesterol
13	☑ 18	Catarina	17	Female	Scientist	415756324	Diabetes
14	☑ 6	Maria	56	Female	Scientist	756985321	Cancer
15	☑ 14	Tiago	67	Male	Scientist	120365201	Aneurysm
16	☑ 1	Joana	26	Female	Teacher	273649999	Cancer
17	☑ 5	Paulo	10	Male	Teacher	546895665	Cholesterol
18	☑ 9	Francisca	34	Female	Teacher	125630235	Aneurysm
19	☑ 13	Rui	19	Male	Teacher	123456789	Diabetes
20	☑ 17	Rita	34	Female	Teacher	325658985	Hypertense

Output data

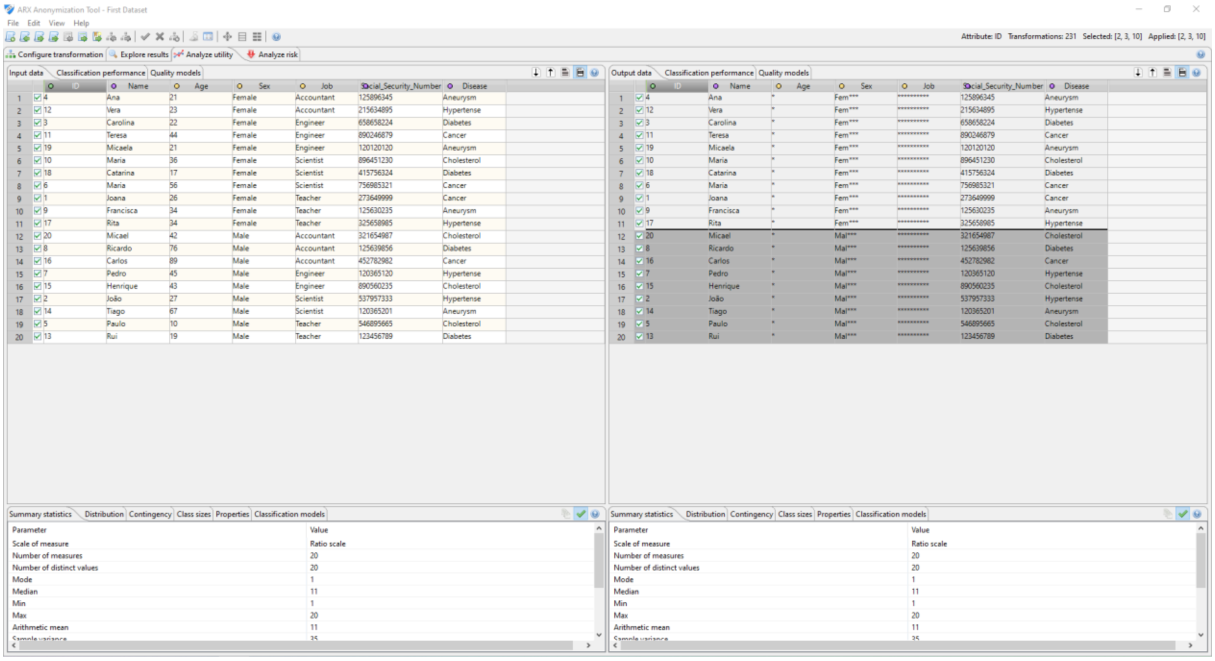
	ID	Name	Age	Sex	Job	Social_Security_Number	Disease
1	☑ 4	Ana	*	******	**********	125896345	Aneurysm
2	☑ 12	Vera	*	******	**********	215634895	Hypertense
3	☑ 20	Micael	*	******	**********	321654987	Cholesterol
4	☑ 8	Ricardo	*	******	**********	125639856	Diabetes
5	☑ 16	Carlos	*	******	**********	452782982	Cancer
6	☑ 3	Carolina	*	******	**********	658658224	Diabetes
7	☑ 7	Pedro	*	******	**********	120365120	Hypertense
8	☑ 11	Teresa	*	******	**********	890246879	Cancer
9	☑ 15	Henrique	*	******	**********	890560235	Cholesterol
10	☑ 19	Micaela	*	******	**********	120120120	Aneurysm
11	☑ 2	João	*	******	**********	537957333	Hypertense
12	☑ 10	Maria	*	******	**********	896451230	Cholesterol
13	☑ 18	Catarina	*	******	**********	415756324	Diabetes
14	☑ 6	Maria	*	******	**********	756985321	Cancer
15	☑ 14	Tiago	*	******	**********	120365201	Aneurysm
16	☑ 1	Joana	*	******	**********	273649999	Cancer
17	☑ 5	Paulo	*	******	**********	546895665	Cholesterol
18	☑ 9	Francisca	*	******	**********	125630235	Aneurysm
19	☑ 13	Rui	*	******	**********	123456789	Diabetes
20	☑ 17	Rita	*	******	**********	325658985	Hypertense

Parameter	Value
Scale of measure	Nominal scale
Number of measures	20
Number of distinct values	4
Mode	Accountant

Parameter	Value
Scale of measure	Nominal scale
Number of measures	20
Number of distinct values	1
Mode	**********

Figure 22 - Anonymization Dataset Results

As it can be verified in Figure 23, the attributes for which the hierarchies were defined were set to the last level of the hierarchy and the sensitive attributes remain unchanged. Choosing another leaf from the results graph, like the one with values (2, 3, 10) it can be verified that only the Age attribute was set to the last hierarchy level.



Attribute: ID Transformations: 231 Selected: [2, 3, 10] Applied: [2, 3, 10]

Input data

		ID	Name	Age	Sex	Job	Social_Security_Number	Disease
1	☑	4	Ana	21	Female	Accountant	125896345	Aneurysm
2	☑	12	Vera	23	Female	Accountant	215634895	Hypertense
3	☑	3	Carolina	22	Female	Engineer	658658224	Diabetes
4	☑	11	Teresa	44	Female	Engineer	890246879	Cancer
5	☑	19	Micaela	21	Female	Engineer	120120120	Aneurysm
6	☑	10	Maria	36	Female	Scientist	896451230	Cholesterol
7	☑	18	Catarina	17	Female	Scientist	415756324	Diabetes
8	☑	6	Maria	56	Female	Scientist	756985321	Cancer
9	☑	1	Joana	26	Female	Teacher	273649999	Cancer
10	☑	9	Francisca	34	Female	Teacher	125630235	Aneurysm
11	☑	17	Rita	34	Female	Teacher	325658985	Hypertense
12	☑	20	Micael	42	Male	Accountant	321654987	Cholesterol
13	☑	8	Ricardo	76	Male	Accountant	125639856	Diabetes
14	☑	16	Carlos	89	Male	Accountant	452782982	Cancer
15	☑	7	Pedro	45	Male	Engineer	120365120	Hypertense
16	☑	15	Henrique	43	Male	Engineer	890560235	Cholesterol
17	☑	2	João	27	Male	Scientist	537957333	Hypertense
18	☑	14	Tiago	67	Male	Scientist	120365201	Aneurysm
19	☑	5	Paulo	10	Male	Teacher	546895665	Cholesterol
20	☑	13	Rui	19	Male	Teacher	123456789	Diabetes

Output data

		ID	Name	Age	Sex	Job	Social_Security_Number	Disease
1	☑	4	Ana	*	Fem***	***********	125896345	Aneurysm
2	☑	12	Vera	*	Fem***	***********	215634895	Hypertense
3	☑	3	Carolina	*	Fem***	***********	658658224	Diabetes
4	☑	11	Teresa	*	Fem***	***********	890246879	Cancer
5	☑	19	Micaela	*	Fem***	***********	120120120	Aneurysm
6	☑	10	Maria	*	Fem***	***********	896451230	Cholesterol
7	☑	18	Catarina	*	Fem***	***********	415756324	Diabetes
8	☑	6	Maria	*	Fem***	***********	756985321	Cancer
9	☑	1	Joana	*	Fem***	***********	273649999	Cancer
10	☑	9	Francisca	*	Fem***	***********	125630235	Aneurysm
11	☑	17	Rita	*	Fem***	***********	325658985	Hypertense
12	☑	20	Micael	*	Mal***	***********	321654987	Cholesterol
13	☑	8	Ricardo	*	Mal***	***********	125639856	Diabetes
14	☑	16	Carlos	*	Mal***	***********	452782982	Cancer
15	☑	7	Pedro	*	Mal***	***********	120365120	Hypertense
16	☑	15	Henrique	*	Mal***	***********	890560235	Cholesterol
17	☑	2	João	*	Mal***	***********	537957333	Hypertense
18	☑	14	Tiago	*	Mal***	***********	120365201	Aneurysm
19	☑	5	Paulo	*	Mal***	***********	546895665	Cholesterol
20	☑	13	Rui	*	Mal***	***********	123456789	Diabetes

Parameter	Value
Scale of measure	Ratio scale
Number of measures	20
Number of distinct values	20
Mode	1
Median	11
Min	1
Max	20
Arithmetic mean	11
Sample variance	35

Parameter	Value
Scale of measure	Ratio scale
Number of measures	20
Number of distinct values	20
Mode	1
Median	11
Min	1
Max	20
Arithmetic mean	11
Sample variance	35

Figure 23 - Anonymization Dataset Result

Therefore, this is the process that which a dataset can be anonymized using ARX. The tool provides more information regarding some metrics and the risk analysis for the dataset, but the focus, for now, was just to understand how a simple dataset could be anonymized, using this tool.

6.2. Amnesia

Amnesia is a free open-source tool used to anonymize datasets. This tool goes beyond the GDPR guidelines and pseudo-anonymization. It offers high usability and flexibility, making it simple to use by everyone (Amnesia, s.d.).

The advantages of the tool are that it does not allow the information linkage, which means that the information from the anonymized dataset could not be linked to the information of the original dataset, it uses k-anonymity and k^m-anonymity algorithms, and it also allows the minimum reduction of information quality.

k^m-anonymity is defined by Amnesia as a weaker version of k-anonymity, where each combination of m quasi-identifiers appears k times in the anonymized dataset (Amnesia, 2020). This algorithm only uses generalization instead of generalization and/or suppression like k-anonymity (Terrovitis, Mamoulis, & Kalnis, 2008).

This tool can be used online or can also be downloaded to the desktop. It has available datasets for the user to start learning using the tool.

The first step to using this tool is providing a dataset. After uploading the dataset in the tool, hierarchy files for the quasi-identifier attributes need to be created, using the tool. It is also possible to create the hierarchy file outside the tool and then upload the files. This tool does not have a lot of anonymization algorithms. After the dataset is anonymized, a solution graph is displayed with which attributes were anonymized and the level of anonymization for each one.

The graphical user interface is shown in Figure 24.

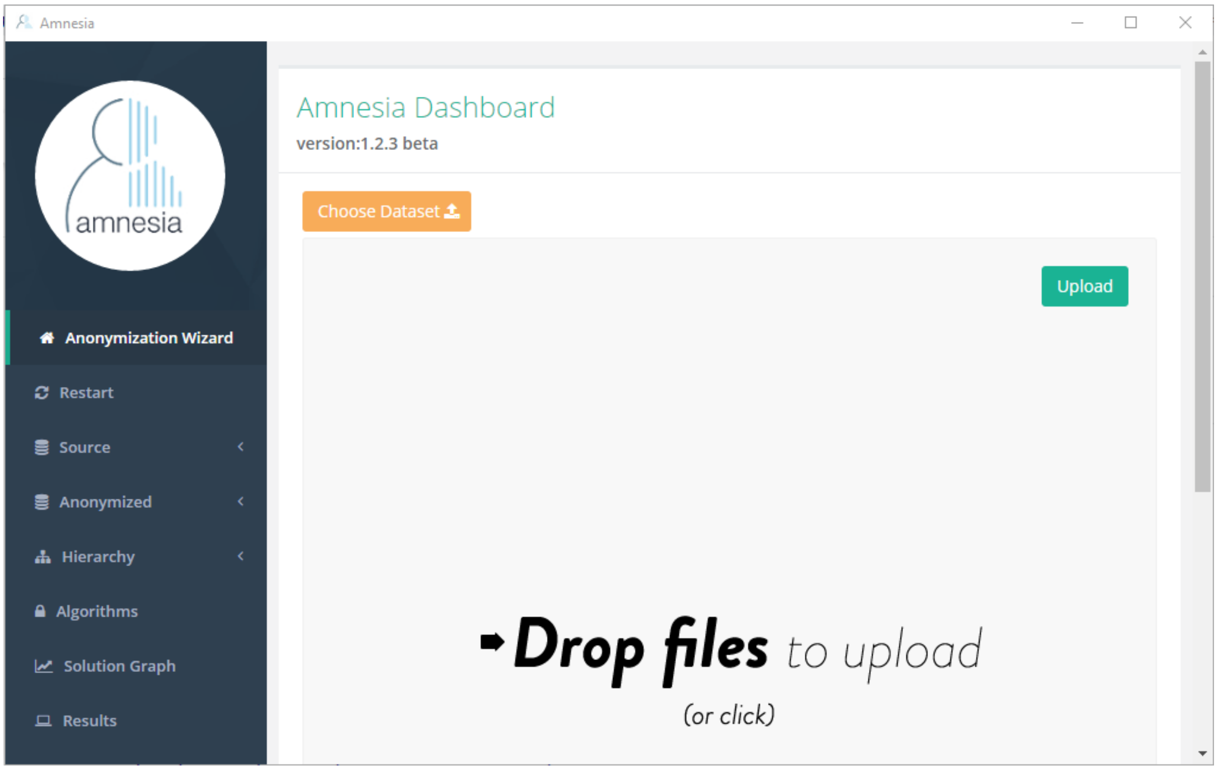
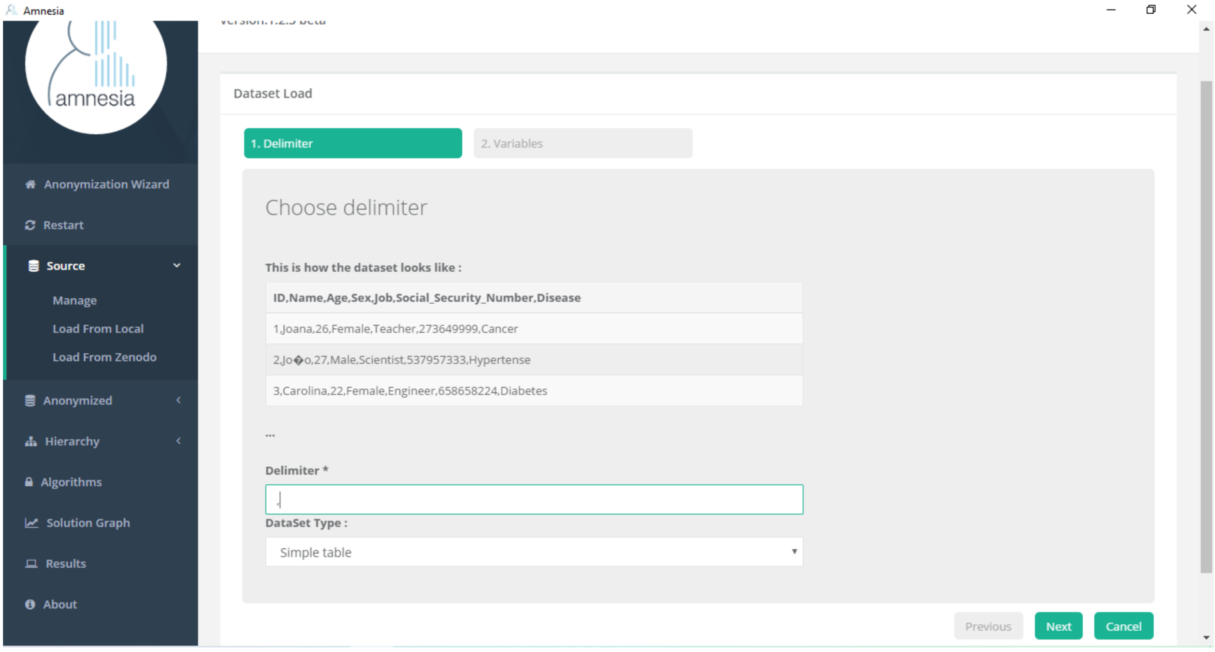


Figure 24 - Amnesia First Display

To learn more about how to use the tool some tests were done with the sample dataset. After choosing the dataset, the tool asks for the delimiter and the dataset type. The dataset type has some options like a simple table, sets of values, table with a set-valued attribute and disk-based simple table. The dataset used is a small table in a CSV file, so the type chosen was a simple table and the delimiter is a comma (,), as shown in Figure 25.



Choose delimiter

This is how the dataset looks like :

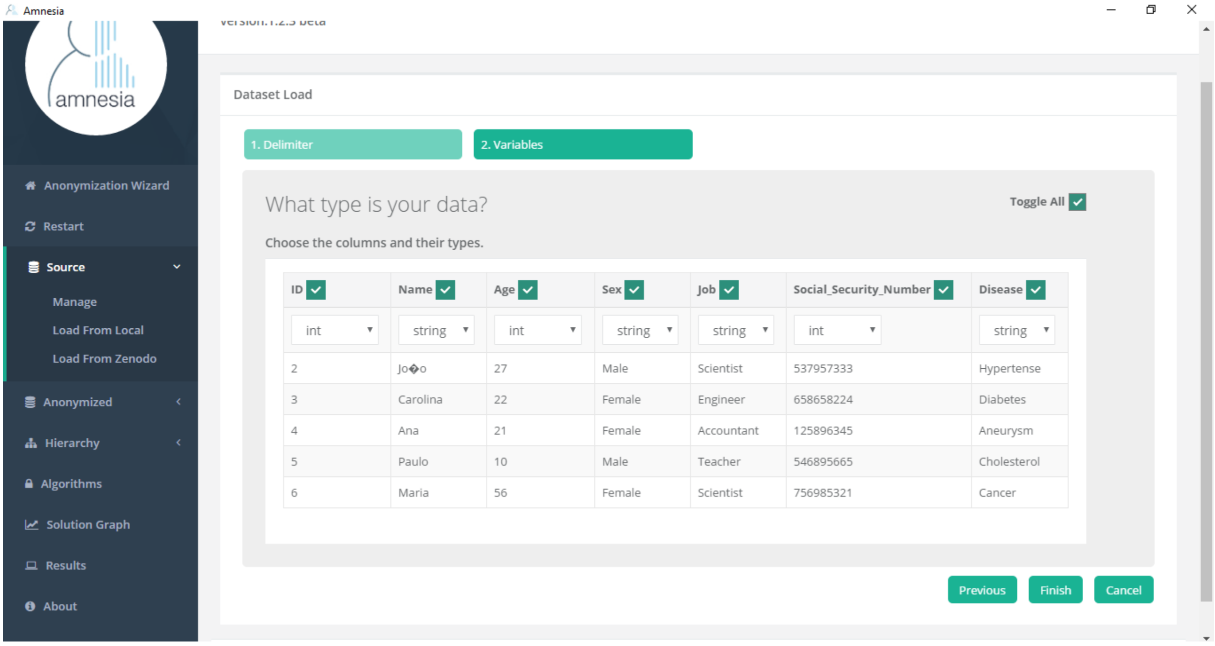
```
ID,Name,Age,Sex,Job,Social_Security_Number,Disease
1,Joana,26,Female,Teacher,273649999,Cancer
2,Jo�o,27,Male,Scientist,537957333,Hypertense
3,Carolina,22,Female,Engineer,658658224,Diabetes
...
```

Delimiter * ,

DataSet Type : Simple table

Figure 25 - Dataset Upload – Choose Delimiter

The next step is to confirm each column type. The tool already sets a type for each column, as shown in Figure 26, so it is just needed to confirm if each type is the correct one.



What type is your data?

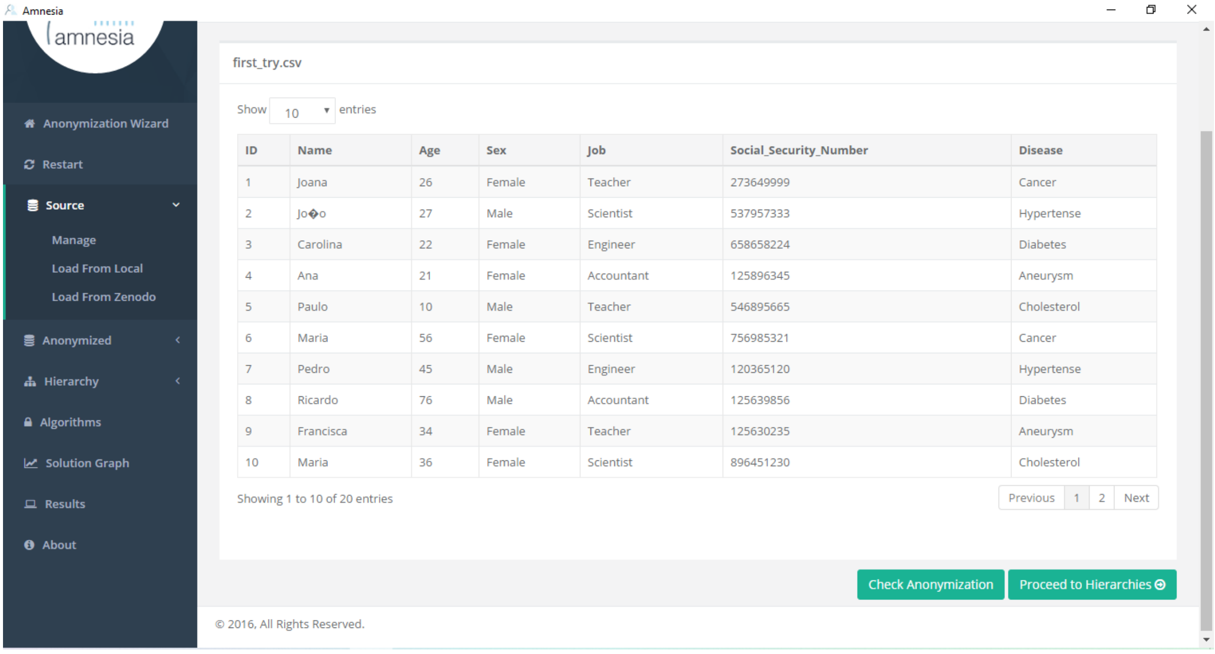
Toggle All ☑

Choose the columns and their types.

ID ☑	Name ☑	Age ☑	Sex ☑	Job ☑	Social_Security_Number ☑	Disease ☑
int	string	int	string	string	int	string
2	Jo�o	27	Male	Scientist	537957333	Hypertense
3	Carolina	22	Female	Engineer	658658224	Diabetes
4	Ana	21	Female	Accountant	125896345	Aneurysm
5	Paulo	10	Male	Teacher	546895665	Cholesterol
6	Maria	56	Female	Scientist	756985321	Cancer

Figure 26 - Dataset Upload – Choose Attribute Types

After the type is chosen, the user just needs to click on Finish and the dataset upload is finished. The dataset is then presented in the tool as shown in Figure 27.



first_try.csv

Show 10 entries

ID	Name	Age	Sex	Job	Social_Security_Number	Disease
1	Joana	26	Female	Teacher	273649999	Cancer
2	Jo�o	27	Male	Scientist	537957333	Hypertense
3	Carolina	22	Female	Engineer	658658224	Diabetes
4	Ana	21	Female	Accountant	125896345	Aneurysm
5	Paulo	10	Male	Teacher	546895665	Cholesterol
6	Maria	56	Female	Scientist	756985321	Cancer
7	Pedro	45	Male	Engineer	120365120	Hypertense
8	Ricardo	76	Male	Accountant	125639856	Diabetes
9	Francisca	34	Female	Teacher	125630235	Aneurysm
10	Maria	36	Female	Scientist	896451230	Cholesterol

Showing 1 to 10 of 20 entries

Previous 1 2 Next

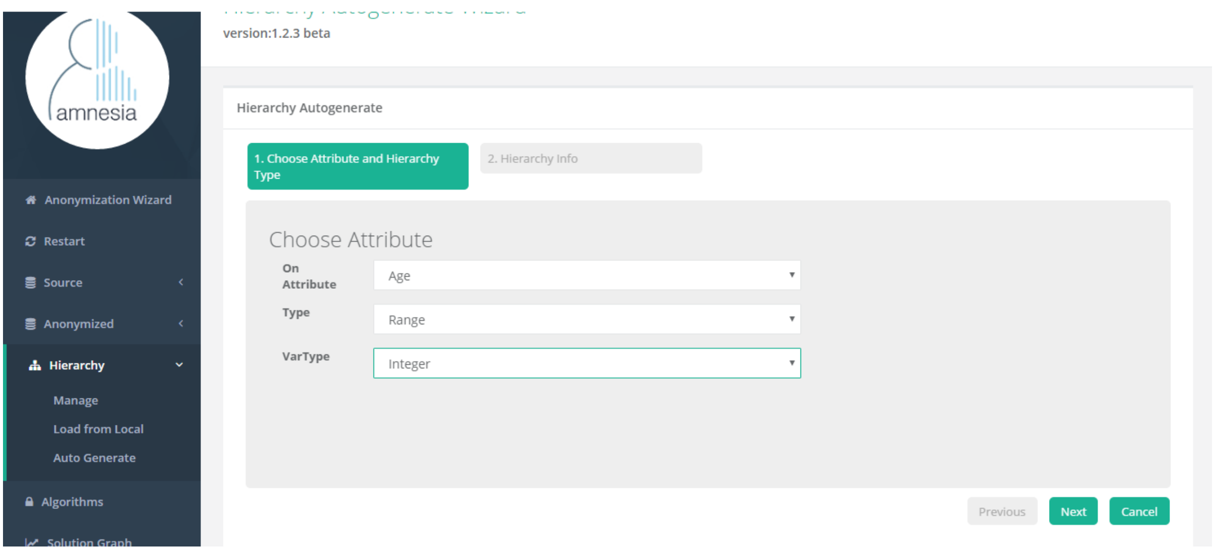
© 2016, All Rights Reserved.

Figure 27 - Dataset Visualization

The next step is to create hierarchies. The hierarchies will be used to anonymize the data for the pretended level. They should be created for each attribute where it is not supposed the original values to appear in the anonymized dataset.

For each attribute, is going to be created the respective hierarchy with the Autogenerate Hierarchy option.

The first hierarchy is going to be created for the Age attribute.



version:1.2.3 beta

Hierarchy Autogenerate

1. Choose Attribute and Hierarchy Type

2. Hierarchy Info

Choose Attribute

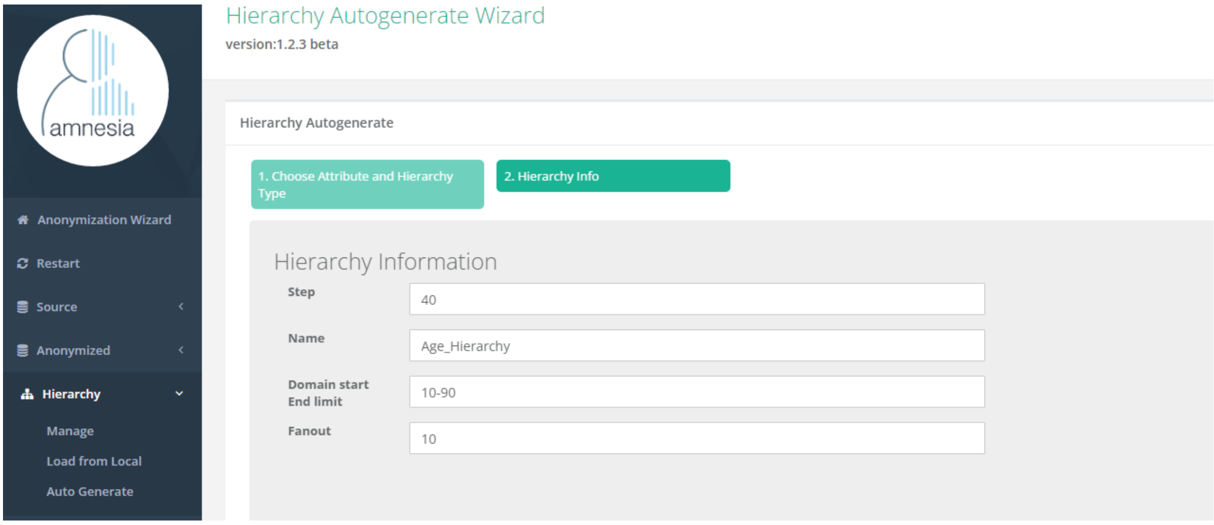
On Attribute: Age

Type: Range

VarType: Integer

Figure 28 - Create Age Attribute Hierarchy

The pretended result for the Age attribute is a different interval of values, and this is the way the Range type is the chosen one, as shown in Figure 28. Next is necessary to set the Hierarchy information, as the step between the intervals is pretended for this attribute as well as the start value and the end value. This information is displayed in Figure 29.



Hierarchy Autogenerate Wizard

version:1.2.3 beta

Hierarchy Autogenerate

Hierarchy Information

Step: 40

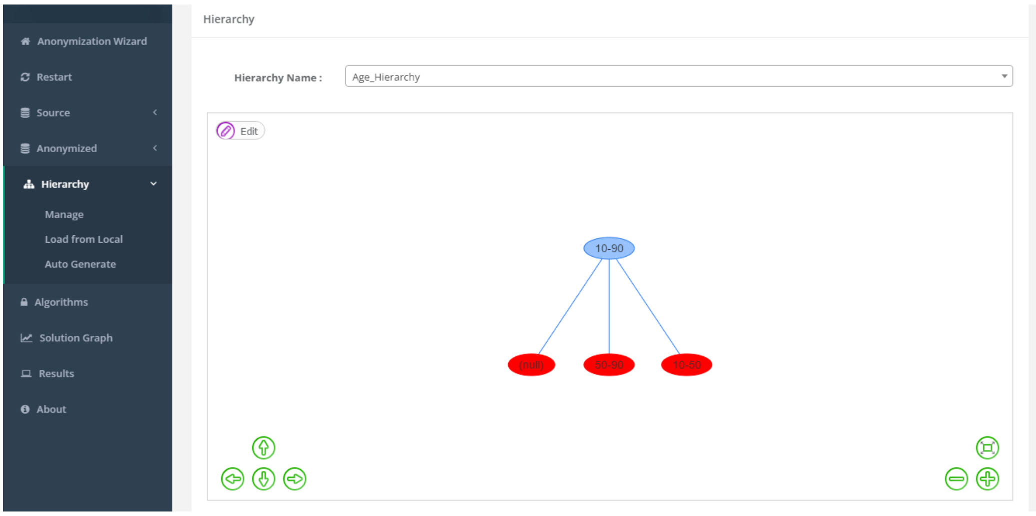
Name: Age_Hierarchy

Domain start End limit: 10-90

Fanout: 10

Figure 29 - Set hierarchy Information for Age Attribute

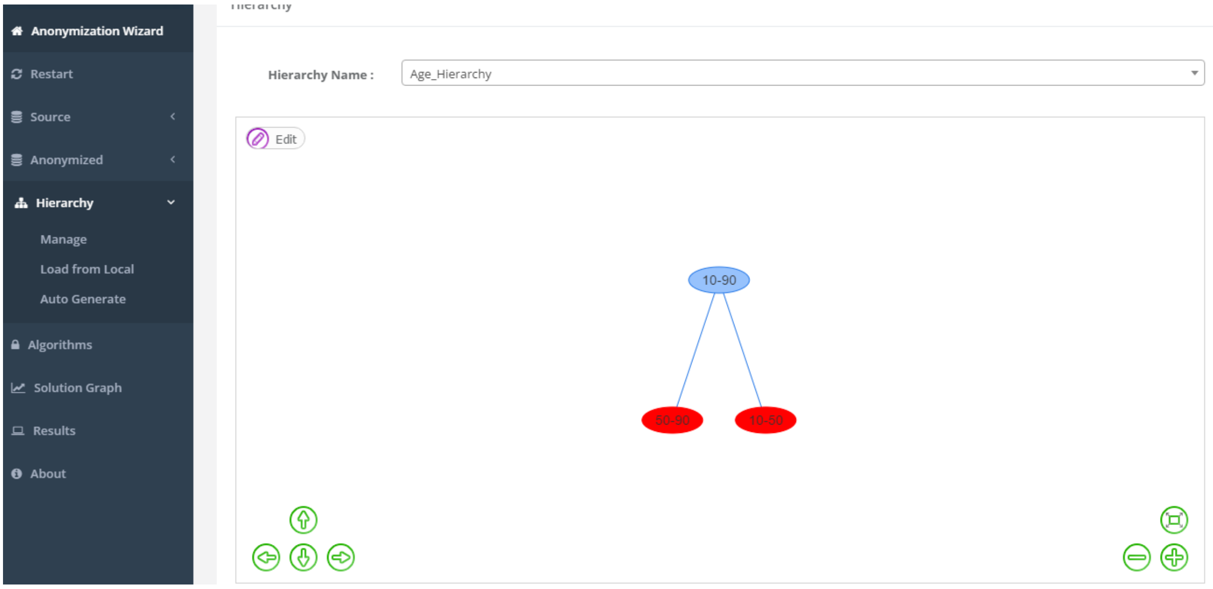
As this dataset is used just as training it is not necessary to have a lot of intervals of values for the Age attribute, this is the way the 40 value was chosen for the step (which means that each interval is going to contain 40 values). The Age values are between 10 and 89, so, to have complete sets, the value is chosen for the domain to start at 10 and finish at 90. With this information, the hierarchy generated is the following one, shown in Figure 30.



Hierarchy Name : Age_Hierarchy

Figure 30 - Age Hierarchy Display

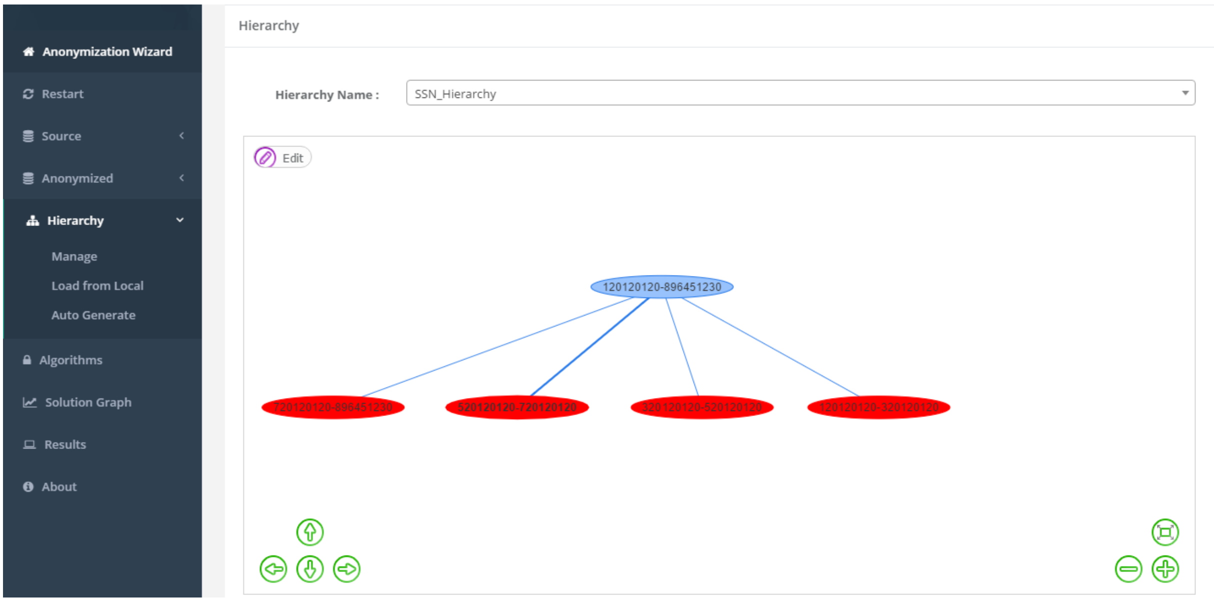
As the dataset does not have null values for age, the leaf with the null parameter can be removed from the hierarchy by editing it. And the final hierarchy is the following one, shown in Figure 31.



Hierarchy Name : Age_Hierarchy

Figure 31 - Final Age Hierarchy Display

The next hierarchy created is the one for the Social_Security_Number attribute. This one is also an attribute of type Integer, and it is created almost the same way as the Age_Hierarchy. The Range type was also chosen. The other option is the Distinct type, but with this one, in the end, what is presented is a hierarchy where the main node has the value 0 and all the leaves are the values present in the original dataset, and this information should not be retrieved in the anonymized dataset as it is. At this time, the domain starts and end limits are the lowest and the biggest values for this attribute, accordingly. When the hierarchy is created, the null leaf is also presented, and it was removed because the original dataset does not have null values. So, the hierarchy created is the one shown in Figure 32.



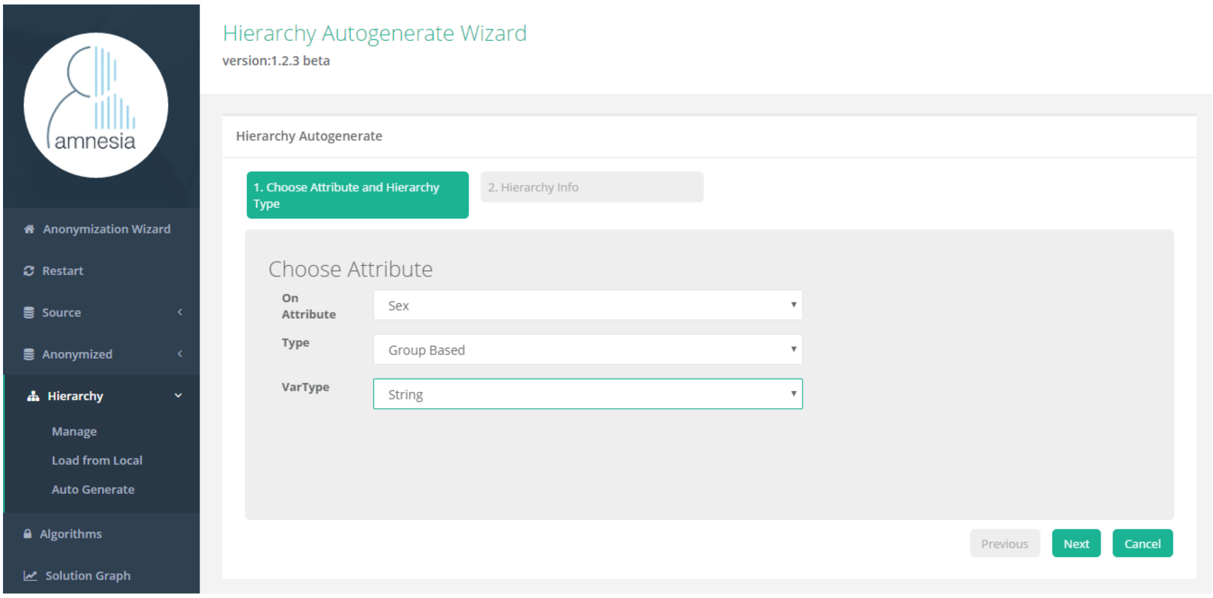
Hierarchy Name : SSN_Hierarchy

Figure 32 - Final Social_Security_Number Hierarchy Display

Now, all the attributes of type Integer have their hierarchies created.

For the Sex attribute, the entire dataset just has two values: Male or Female. As this is an attribute of type String, this hierarchy has some different fields to complete.

The options for the type given by the tool are Group-Based and Masking Based. For the Sex attribute, the Group Based type was selected. The Masking Based type suppresses almost all the letters of the attribute value, for example, instead of having the attribute anonymized to Sex, the anonymization results are M*** or F*****. The first board is completed as can be seen in Figure 33.



Choose Attribute

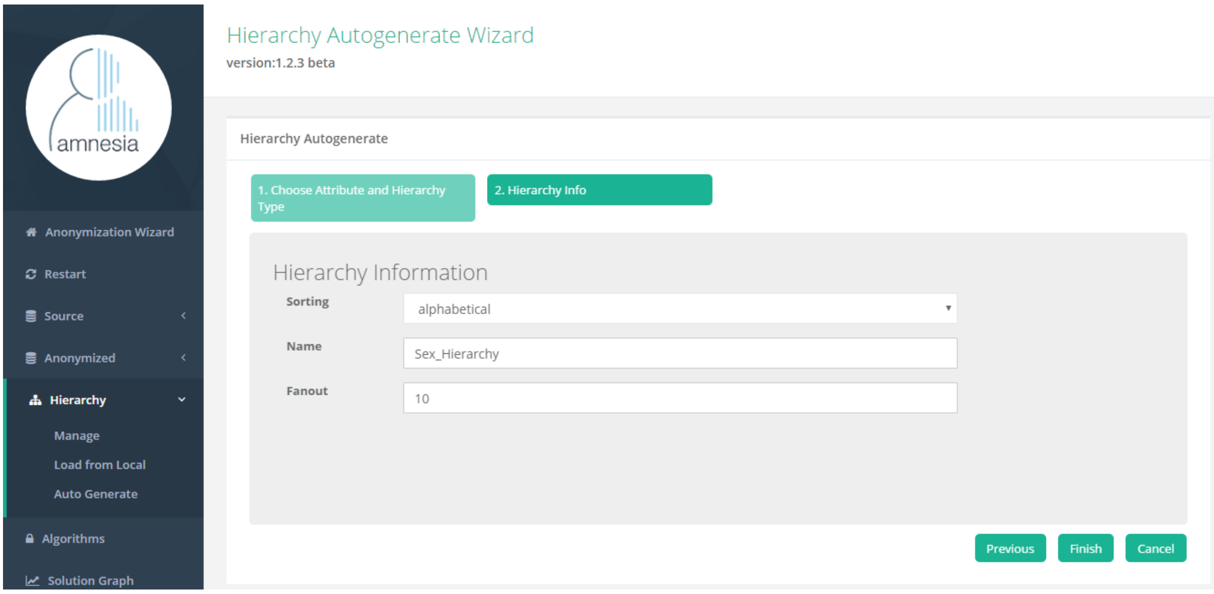
On Attribute: Sex

Type: Group Based

VarType: String

Figure 33 - Create Sex Attribute Hierarchy

In the next board, shown in Figure 34, the Sorting order can be selected between Random or Alphabetically options. The alphabetical order was the option chosen. The next board is completed as shown in Figure 34.



Hierarchy Autogenerate Wizard

version:1.2.3 beta

Hierarchy Autogenerate

1. Choose Attribute and Hierarchy Type

2. Hierarchy Info

Hierarchy Information

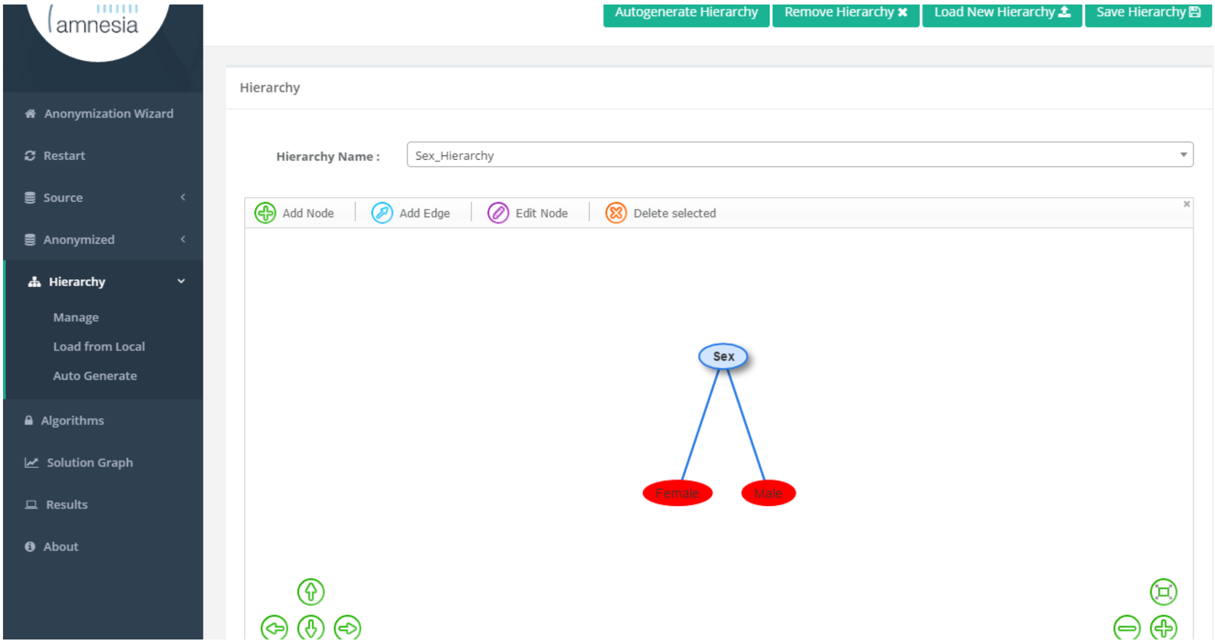
Sorting: alphabetical

Name: Sex_Hierarchy

Fanout: 10

Figure 34 - Set hierarchy Information for Sex Attribute

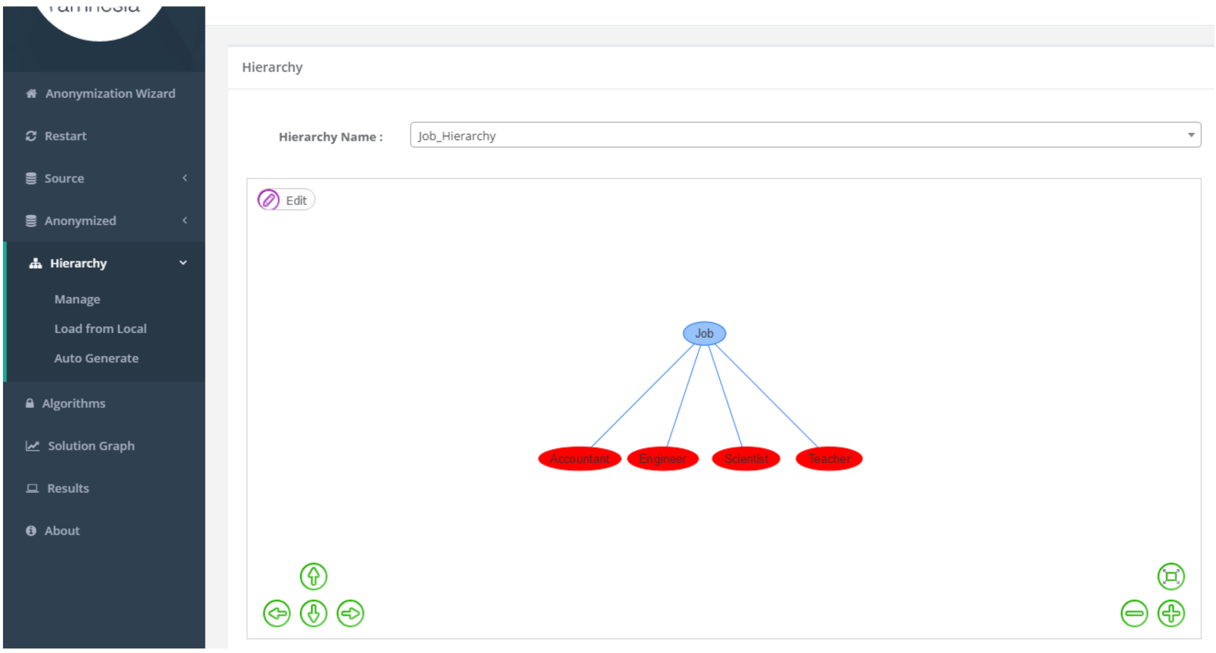
The hierarchy generated has also the null value, which is removed because, again, the dataset does not have null values. Another thing that was changed in the hierarchy generated was the main node. It was created with the value Random0 and it could be more explanatory, so it was changed to Sex, thus, the node was edited, and the result is the following one, shown in Figure 35.



Hierarchy Name : Sex_Hierarchy

Figure 35 - Final Sex Hierarchy Display

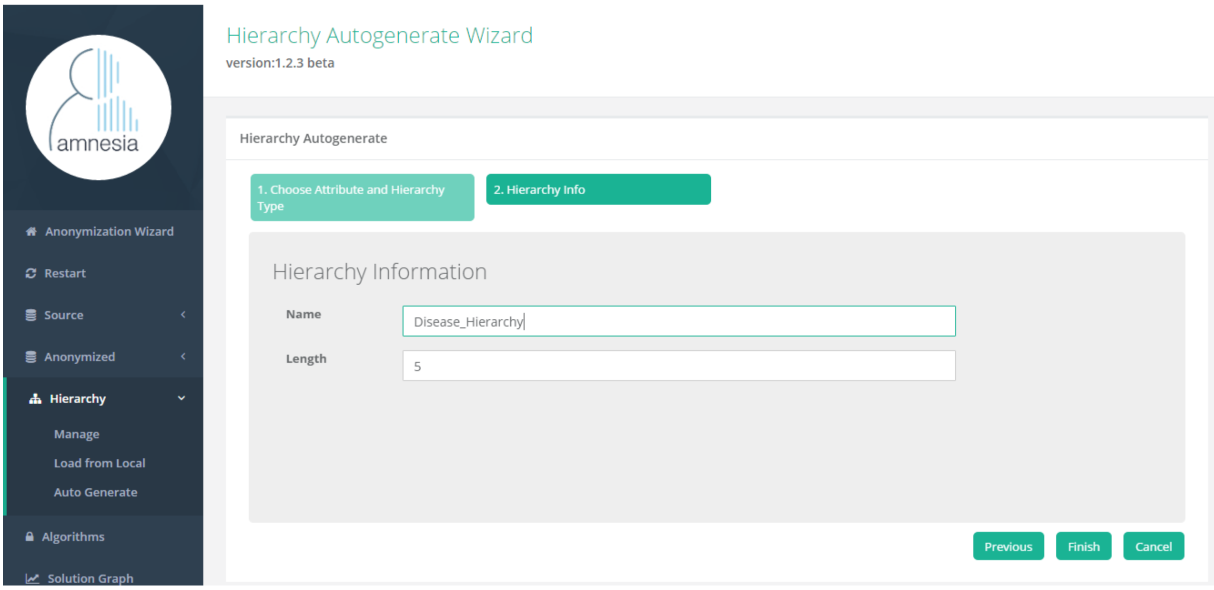
The hierarchy for Job attribute was created exactly has the Sex_Hierarchy, so the final view of the hierarchy created is shown in Figure 36.



Hierarchy Name : Job_Hierarchy

Figure 36 - Final Job Hierarchy Display

The last hierarchy created was for the Disease attribute. For this one, the steps are almost the same as the two before, but this time the Masking Based type was the chosen one, which changes things at the second board of the hierarchy. In the hierarchy information, it is needed to specify the length pretended for the leaf node, and the value chosen was 5, as shown in Figure 37. The smaller value for the Disease attribute is Cancer and it has 6 characters. The value 5 was chosen to verify that all the characters of the attribute values were replaced. It could be verified with any value smaller than 5, but if all the characters were not replaced, in this way, the result would be the first original character with the other replaced by the special character, like this C*****.



Hierarchy Autogenerate Wizard
version:1.2.3 beta

Hierarchy Autogenerate

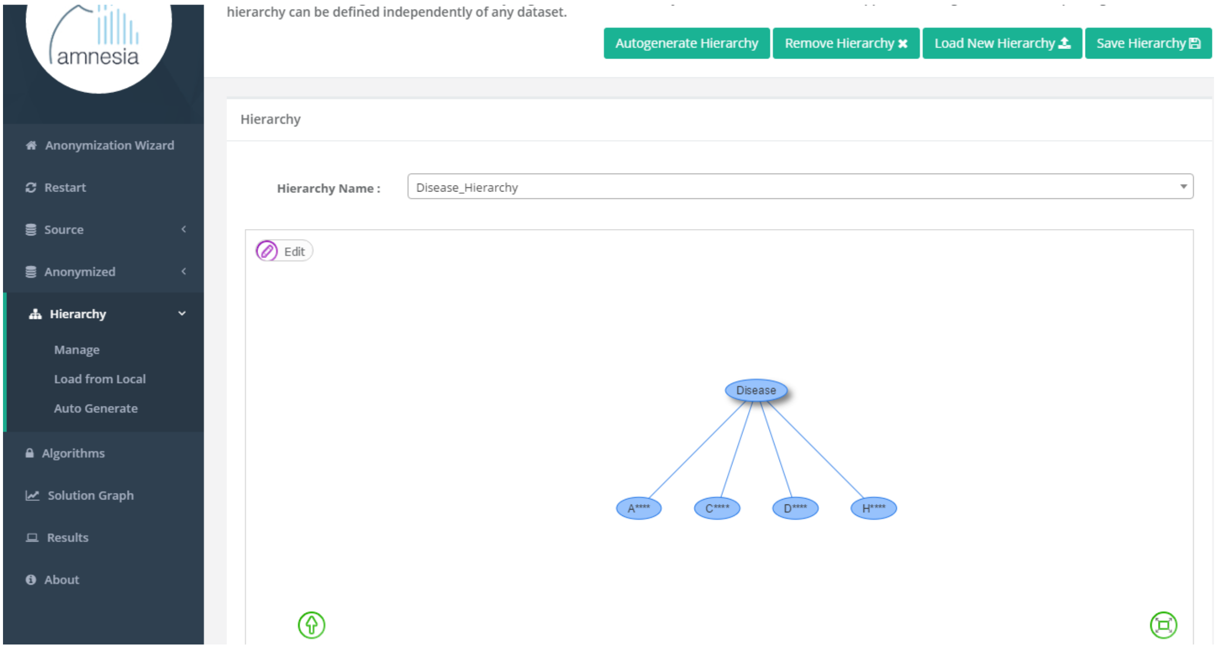
Hierarchy Information

Name: Disease_Hierarchy

Length: 5

Figure 37 - Set hierarchy Information for Disease Attribute

The null leaf was also removed and the main node, which had the value ***** was changed to Disease. The result is in Figure 38.



hierarchy can be defined independently of any dataset.

Hierarchy Name : Disease_Hierarchy

Figure 38 - Final Disease Hierarchy Display

After creating the hierarchies, they can be saved locally on your computer. This will create a TXT file with the necessary information of each board for the tool to create the hierarchy if the file is loaded to the tool.

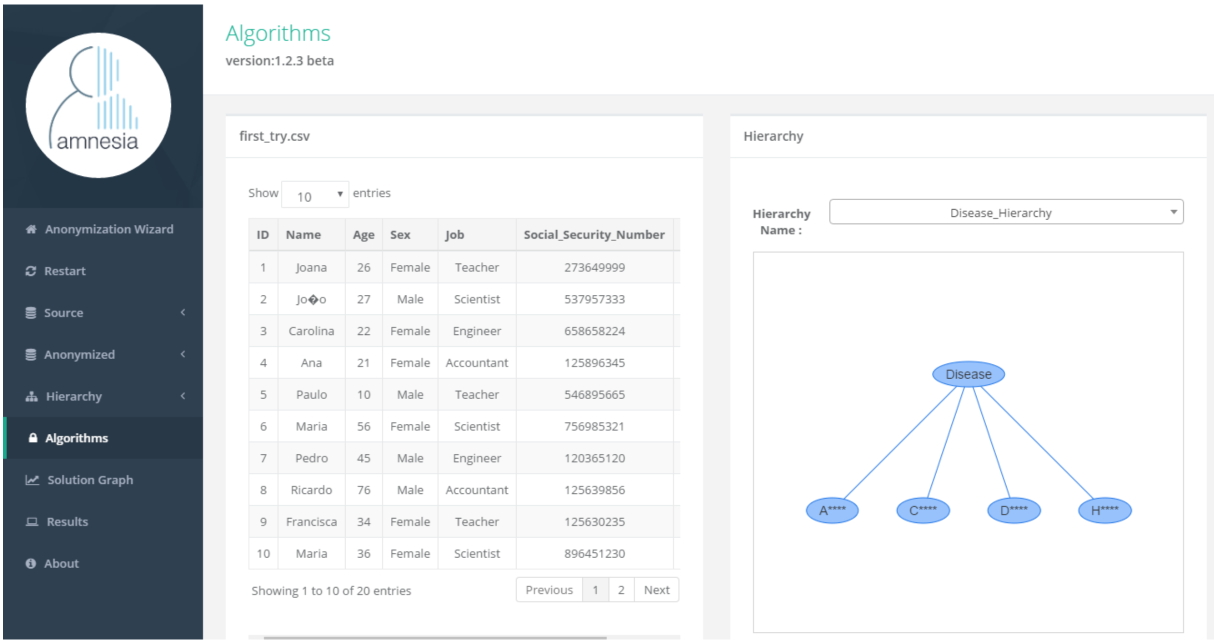
The content of the file for Age_Hierarchy is the following one:

```
range
name Age_Hierarchy
type int
height 2

0.0,90.0 has 0.0,50.0 50.0,90.0
```

This was just an example. The rest of the hierarchies can be verified in Annex A.

After creating the hierarchies, the next menu where information is needed is the Algorithms menu. At this view, the original dataset is displayed with all the hierarchies created, as shown in Figure 39.



Algorithms

version:1.2.3 beta

first_try.csv

Show 10 entries

ID	Name	Age	Sex	Job	Social_Security_Number
1	Joana	26	Female	Teacher	273649999
2	Jo�o	27	Male	Scientist	537957333
3	Carolina	22	Female	Engineer	658658224
4	Ana	21	Female	Accountant	125896345
5	Paulo	10	Male	Teacher	546895665
6	Maria	56	Female	Scientist	756985321
7	Pedro	45	Male	Engineer	120365120
8	Ricardo	76	Male	Accountant	125639856
9	Francisca	34	Female	Teacher	125630235
10	Maria	36	Female	Scientist	896451230

Showing 1 to 10 of 20 entries

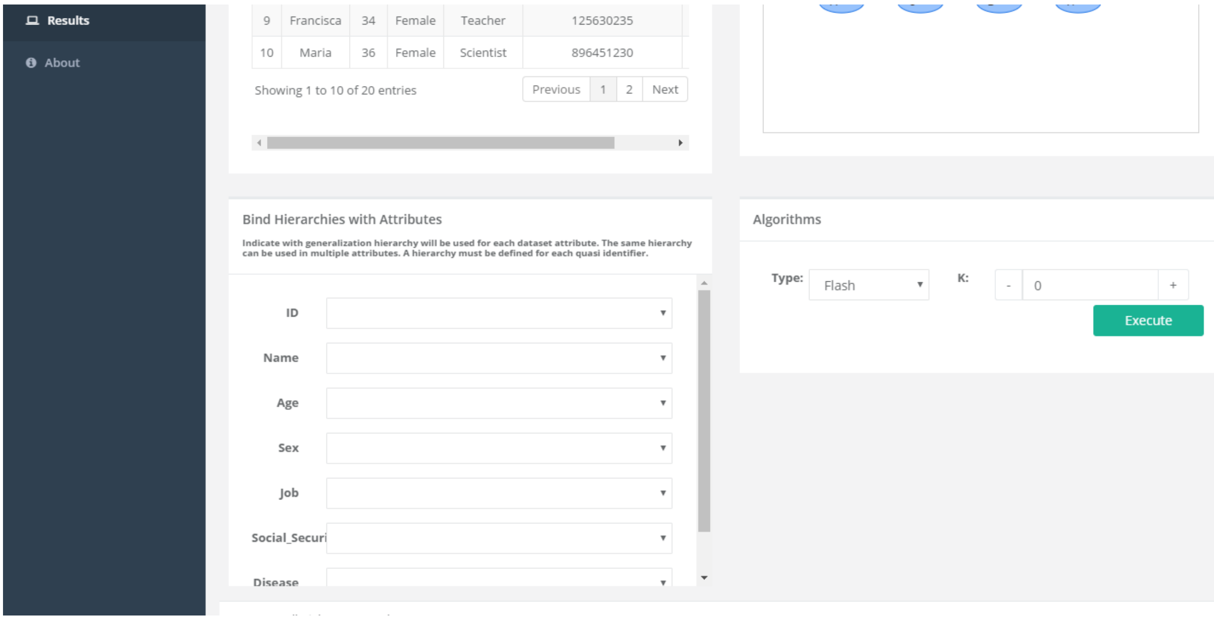
Previous 1 2 Next

Hierarchy

Hierarchy Name : Disease_Hierarchy

Figure 39 - Amnesia Algorithms Board – Dataset and Hierarchy

Scrolling down the image presented has two sections, as shown in Figure 40.



9	Francisca	34	Female	Teacher	125630235
10	Maria	36	Female	Scientist	896451230

Showing 1 to 10 of 20 entries

Previous 1 2 Next

Bind Hierarchies with Attributes

Indicate with generalization hierarchy will be used for each dataset attribute. The same hierarchy can be used in multiple attributes. A hierarchy must be defined for each quasi identifier.

ID

Name

Age

Sex

Job

Social_Securi

Disease

Algorithms

Type: Flash

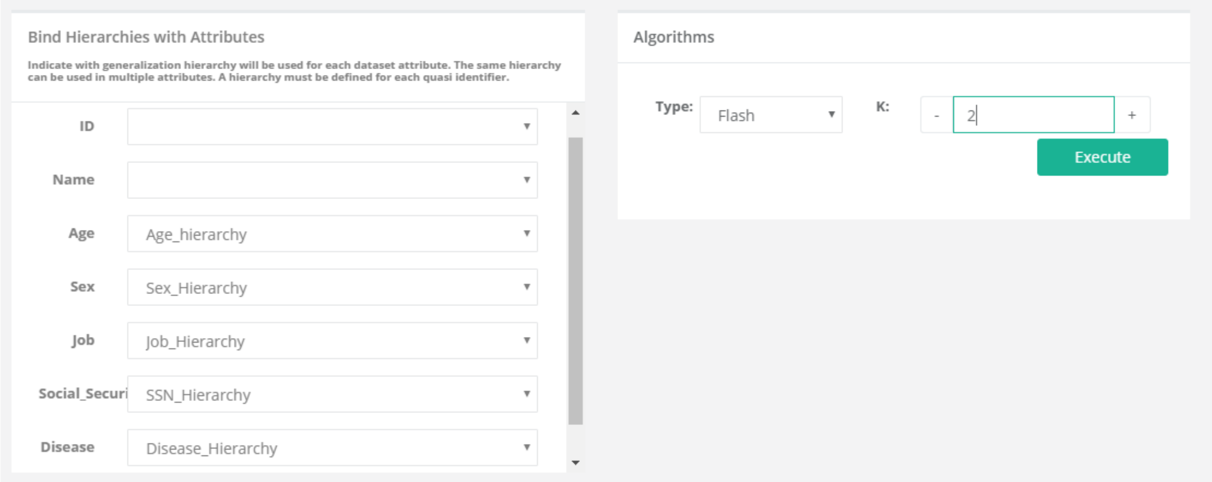
K: - 0 +

Execute

Figure 40 - Amnesia Algorithms Board – Hierarchies and Algorithms

In the section “Bind Hierarchies with Attributes”, the attributes existing in the original dataset can be linked to the hierarchies created in the previous step. In the Algorithms type, the only one that exists is the Flash-type, and only the value for the K variable can be set.

After linking the attributes with the respective hierarchies and giving a value to K, as it can be verified in Figure 41, it is only necessary to click on the Execute button and the dataset is anonymized.



Bind Hierarchies with Attributes

Indicate with generalization hierarchy will be used for each dataset attribute. The same hierarchy can be used in multiple attributes. A hierarchy must be defined for each quasi identifier.

ID:

Name:

Age: Age_hierarchy

Sex: Sex_Hierarchy

Job: Job_Hierarchy

Social_Securi: SSN_Hierarchy

Disease: Disease_Hierarchy

Algorithms

Type: Flash

K: 2

Figure 41 - Amnesia Algorithms Board – Hierarchies Linked and Algorithms

This button will lead the user to the Solution Graph menu, where a graph is displayed with all the anonymization levels possible in this case.

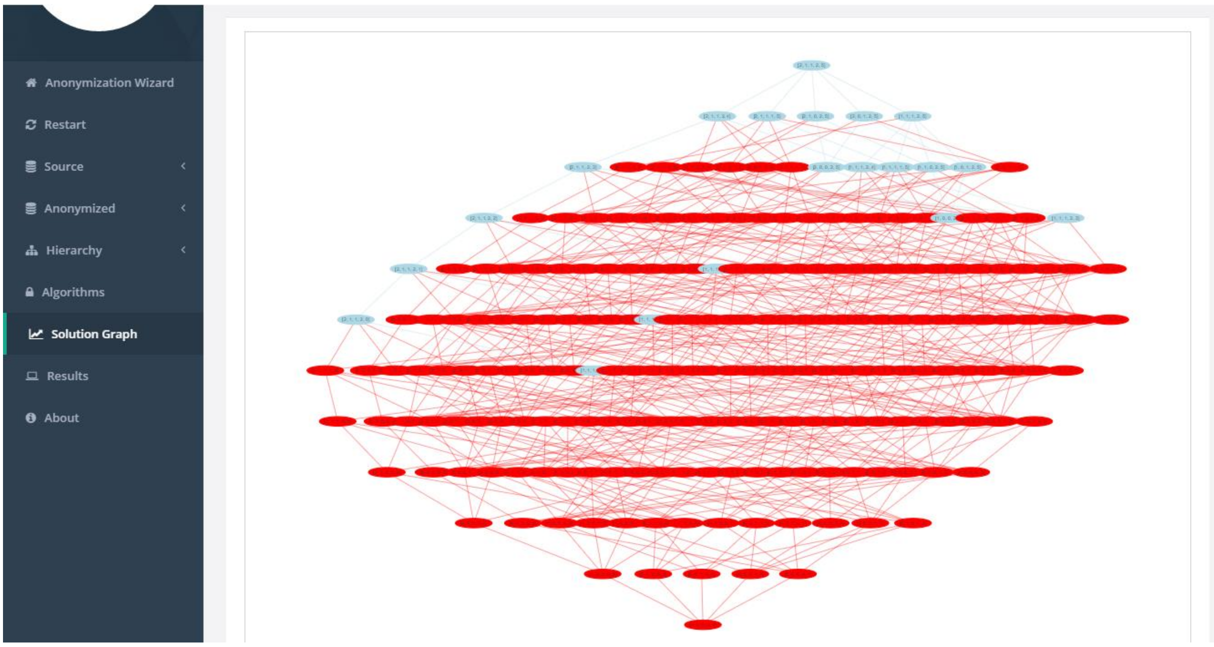


Figure 42 - Amnesia Solution Graph

In Figure 42 is possible to verify some blue and some red nodes. The safe solutions are the ones in blue. The ones in red, are the unsafe solutions. Unsafe solutions are the ones that “violates the desired k-anonymity guarantee just for a few records” (Amnesia, 2020) . These solutions can be transformed into safe solutions by using suppression, according to the tool, where the values that do not follow the anonymization performed by the k-anonymity algorithm are removed.

Each node has the following information, shown in Figure 43:

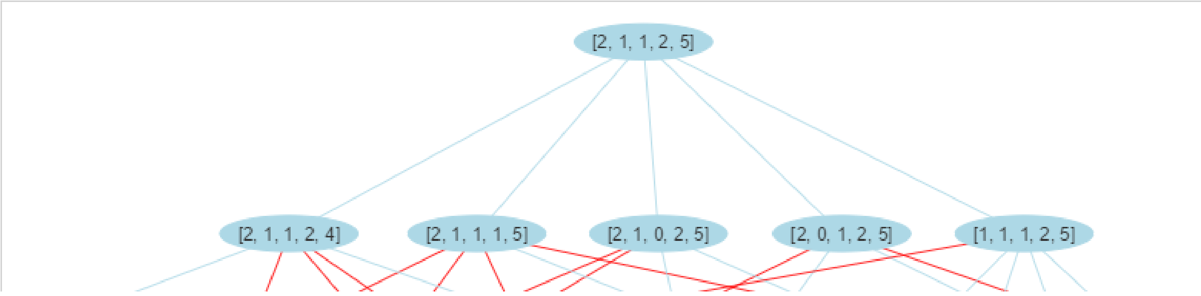


Figure 43 - Safe Solutions from Graph

For example, choosing the node with [1, 1, 1, 2, 5], means that Age was generalized to level 1, Sex was generalized to level 1, Job was generalized to level 1, Social_Security_Number was generalized to level 2 and Disease was generalized to level 5. After clicking on the node, the following pop-up is displayed, as shown in Figure 44.

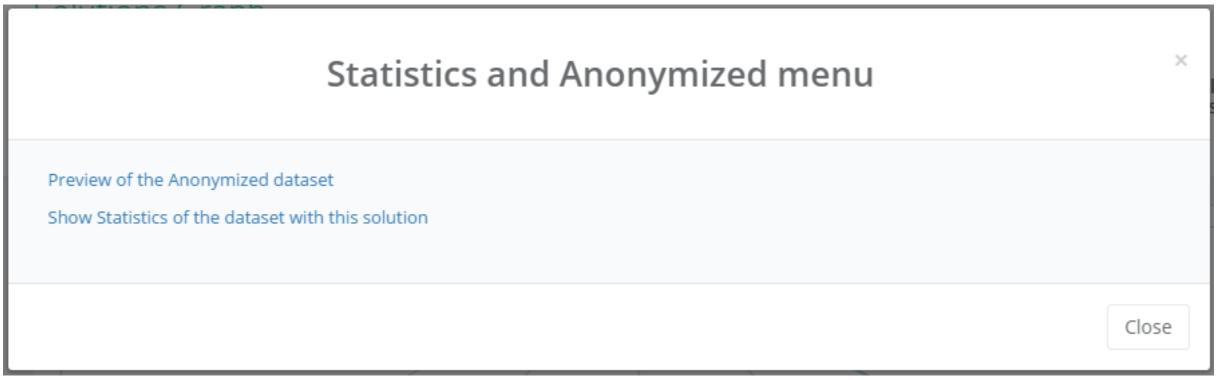
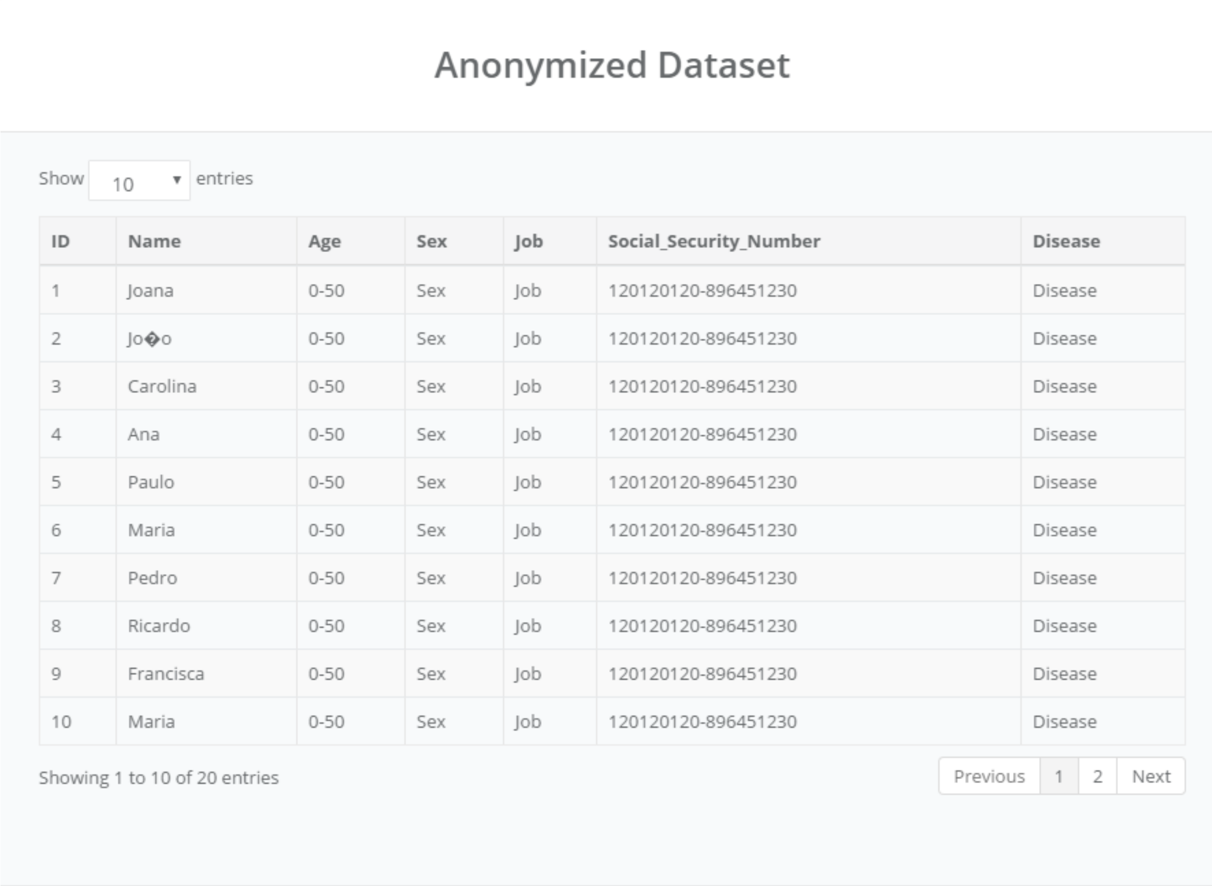

Statistics and Anonymized menu
×
Preview of the Anonymized dataset
Show Statistics of the dataset with this solution
Close

Figure 44 - Statistics and Anonymized Menu

First, the anonymized dataset for the node chosen needs to be verified. It is shown in Figure 45.



Anonymized Dataset

Show 10 ▾ entries

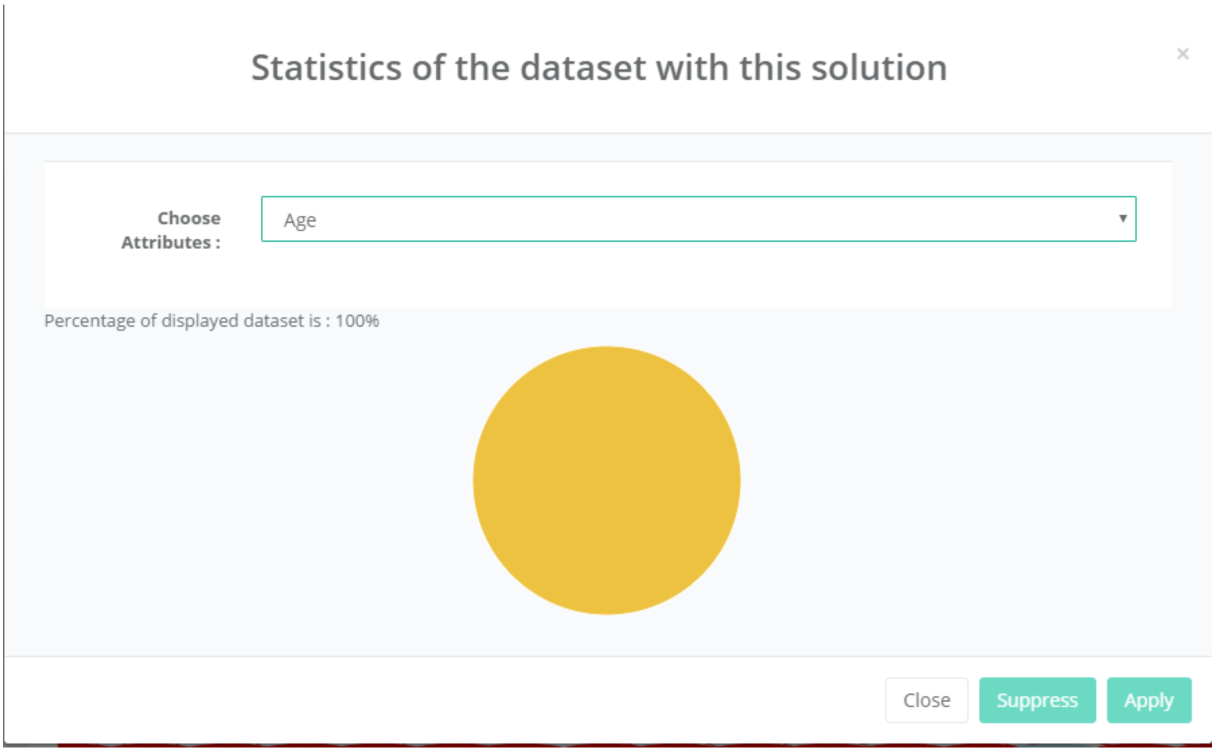
ID	Name	Age	Sex	Job	Social_Security_Number	Disease
1	Joana	0-50	Sex	Job	120120120-896451230	Disease
2	Jo�o	0-50	Sex	Job	120120120-896451230	Disease
3	Carolina	0-50	Sex	Job	120120120-896451230	Disease
4	Ana	0-50	Sex	Job	120120120-896451230	Disease
5	Paulo	0-50	Sex	Job	120120120-896451230	Disease
6	Maria	0-50	Sex	Job	120120120-896451230	Disease
7	Pedro	0-50	Sex	Job	120120120-896451230	Disease
8	Ricardo	0-50	Sex	Job	120120120-896451230	Disease
9	Francisca	0-50	Sex	Job	120120120-896451230	Disease
10	Maria	0-50	Sex	Job	120120120-896451230	Disease

Showing 1 to 10 of 20 entries

Figure 45 - Safe Solution - Anonymized Dataset

It is possible to verify that the attributes were anonymized for the main node of each hierarchy.

In Figure 46 it is possible to verify which are the statistics of the dataset.



Statistics of the dataset with this solution

Choose Attributes : Age

Percentage of displayed dataset is : 100%

Figure 46 - Safe Solution - Statistics of the Dataset

It shows to the user that all the dataset is displayed.

Now it is needed to verify what happens if a red node is chosen. The red nodes are represented in Figure 47.

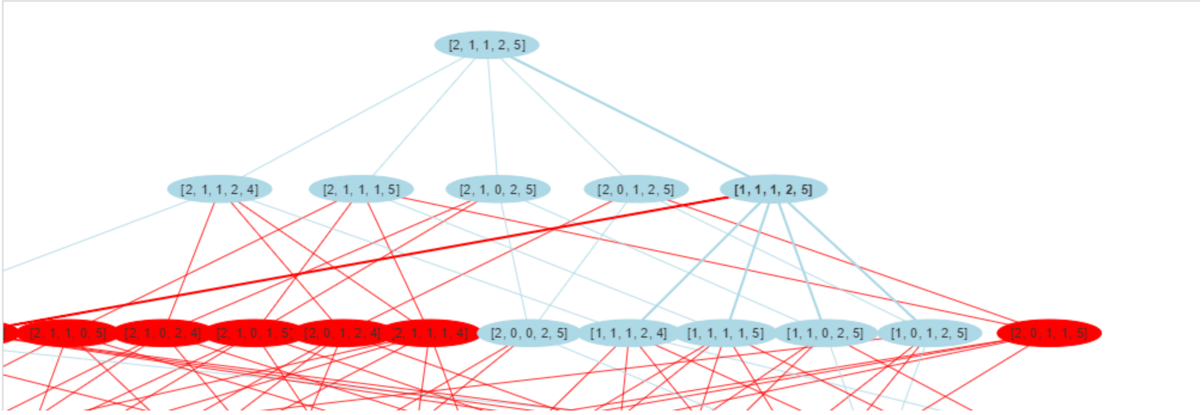
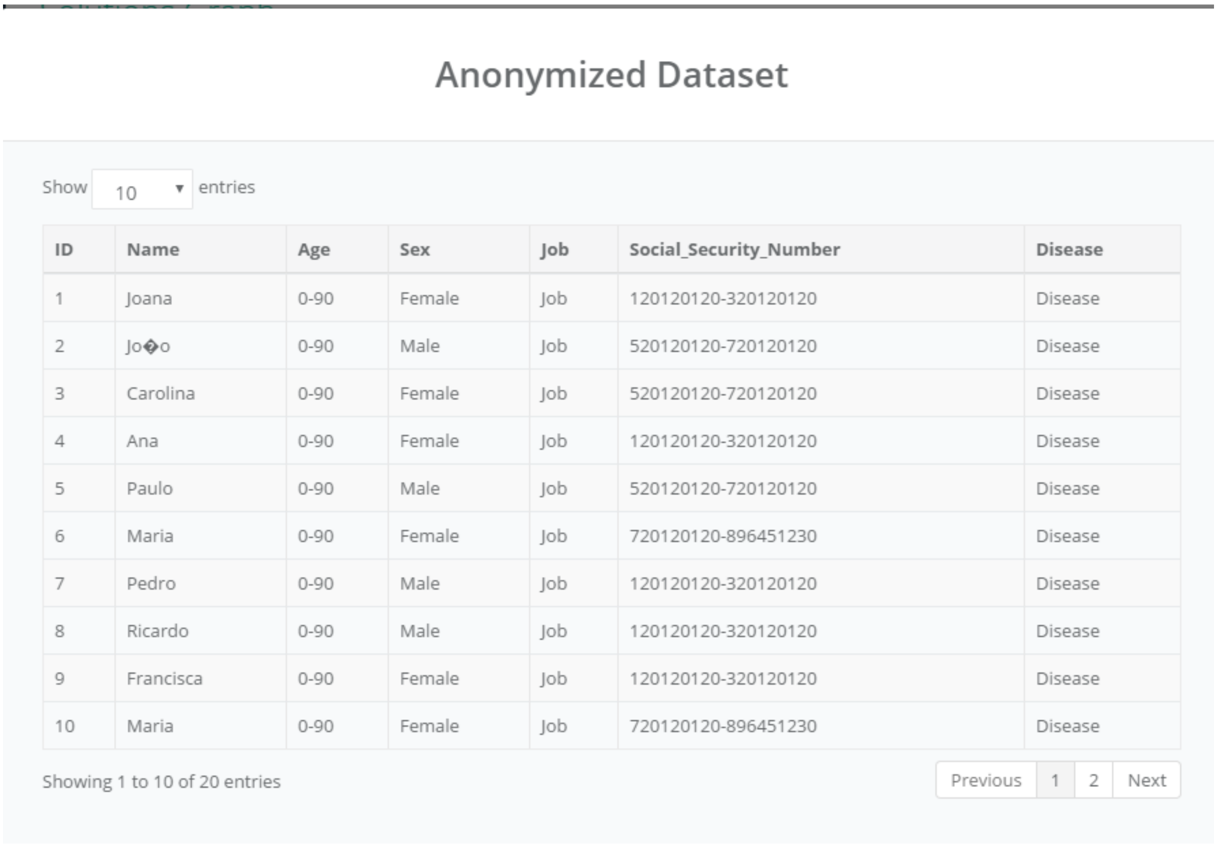


Figure 47 - Unsafe Solutions from Graph

The node picked is the one with [2, 0, 1, 1, 5], which means that Age was generalized to level 2, Sex was generalized to level 0, Job was generalized to level 1, Social_Security_Number was generalized to level 1 and Disease was generalized to level 5.

The anonymized dataset has different values compared with the previous results because the anonymization levels were different as shown in Figure 48.



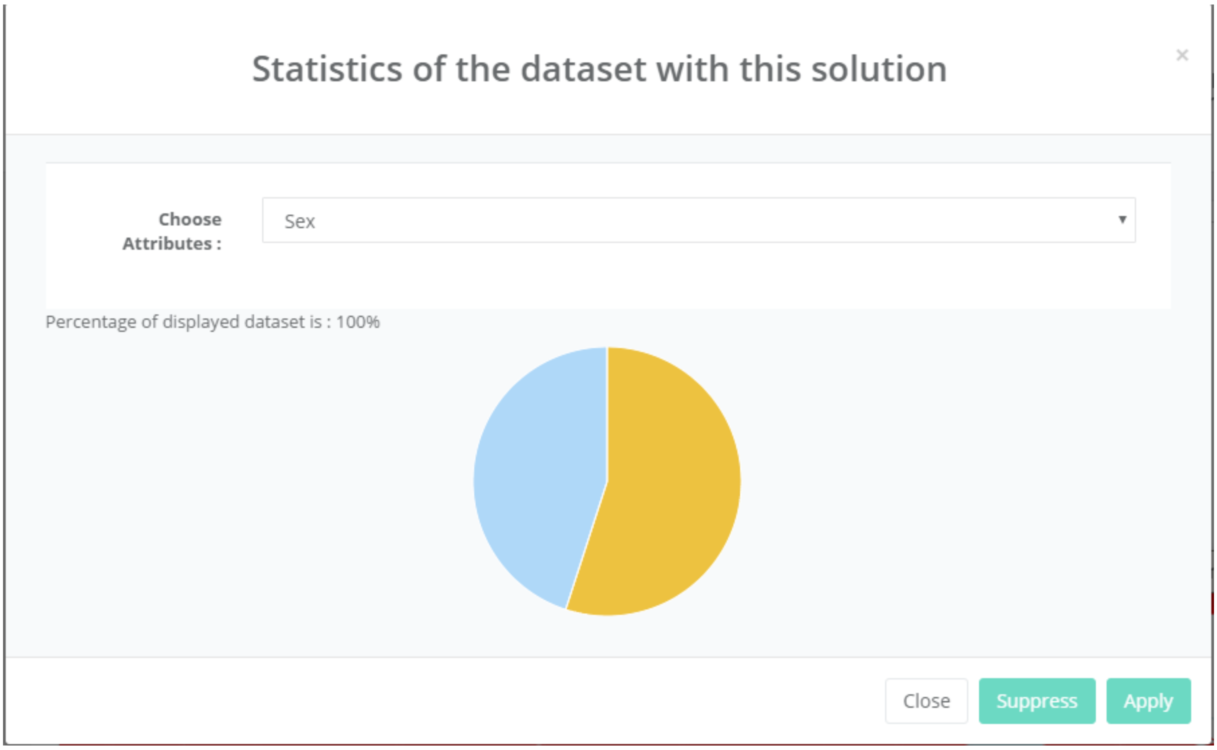
Anonymized Dataset

ID	Name	Age	Sex	Job	Social_Security_Number	Disease
1	Joana	0-90	Female	Job	120120120-320120120	Disease
2	Jo�o	0-90	Male	Job	520120120-720120120	Disease
3	Carolina	0-90	Female	Job	520120120-720120120	Disease
4	Ana	0-90	Female	Job	120120120-320120120	Disease
5	Paulo	0-90	Male	Job	520120120-720120120	Disease
6	Maria	0-90	Female	Job	720120120-896451230	Disease
7	Pedro	0-90	Male	Job	120120120-320120120	Disease
8	Ricardo	0-90	Male	Job	120120120-320120120	Disease
9	Francisca	0-90	Female	Job	120120120-320120120	Disease
10	Maria	0-90	Female	Job	720120120-896451230	Disease

Showing 1 to 10 of 20 entries

Figure 48 - Unsafe Solution - Anonymized Dataset

Checking the statistics, it is possible to verify that the Sex attribute has two different values, as shown in Figure 49. This is not the attribute that turned the solution into an unsafe solution because the percentage of records that do not respect the k-anonymity do not appear in the statistics display.



Statistics of the dataset with this solution

Choose Attributes : Sex

Percentage of displayed dataset is : 100%

Figure 49 - Unsafe Solution - Statistics of the Dataset

In the tool, the information provided is that if the node is red the values can be suppressed, but for all the attributes this button is not available, thus a bug was found in the tool.

6.3. UTD Anonymization Toolbox

UTD Anonymization Toolbox is an open-source tool that was created to promote research in the data anonymization area. It was created in UT Dallas Data Security and Privacy Labs. UTD Anonymization Toolbox allows the usage of 6 anonymization algorithms:

- Datafly
- Mondrian Multidimensional k-anonymity
- Incognito
- Incognito with l-diversity
- Incognito with t-closeness
- Anatomy

The present open-source tool does not have a graphical interface, which may be a serious disadvantage.

To anonymize a dataset is necessary the following four different files:

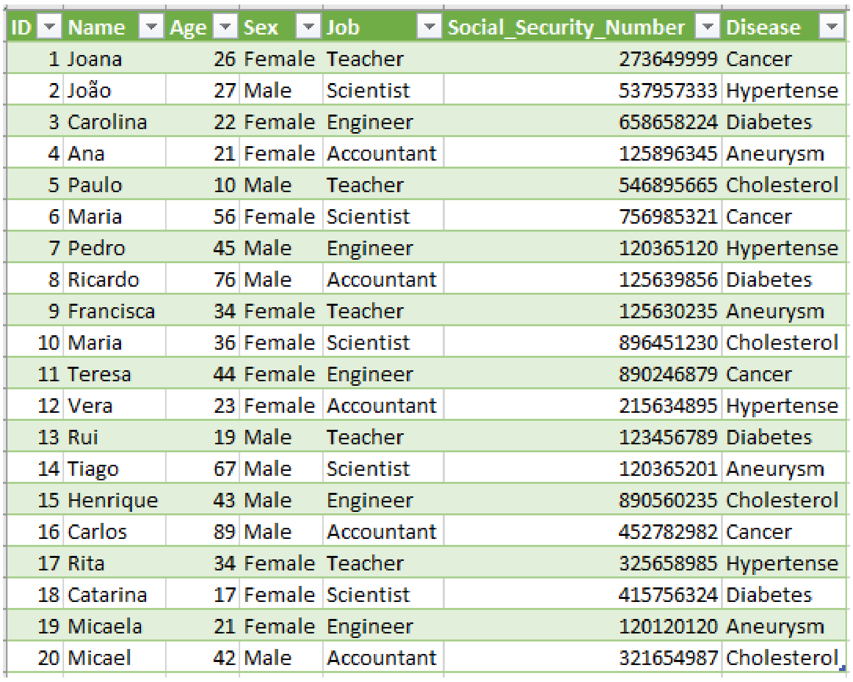
- Dataset, in the CSV format.
- File with .data extension, where the content is almost the same as the dataset. If the first row from the CSV file is removed, the .data file is created (just with the values)
- Header.txt, which is composed of all attributes present in the dataset (the columns of the CSV file) and every different value that every attribute can have
- Config.xml – This one is a configuration file. In this file, the algorithms to use in the dataset anonymization can be identified, as well as the name of the input file (original dataset) and the name of the output file (the one with the anonymized dataset). In this configuration file, the attribute types are also identified, the ones that are identifiers, the quasi-identifiers, and the sensitive attributes, with the corresponding hierarchy values (that are used when the data is anonymized).

After having all these files with the correct information, the tool can be run, and the anonymized dataset can be analysed. (UT Dallas Data Security and Privacy Lab, s.d.)

As in Amnesia, the same dataset was used to learn more about how to work with the tool and how to create the necessary files. The dataset is small because the main concern at this phase is to learn how the files should be created since the documentation is not so clear about these steps.

In the dataset folder, three files need to exist: the dataset in CSV format, a DATA format file and the header.txt file.

The used dataset has seven attributes (ID, Name, Age, Sex, Job, Social_Security_Number and Disease) and 20 rows, as presented in Figure 50.



ID	Name	Age	Sex	Job	Social_Security_Number	Disease
1	Joana	26	Female	Teacher	273649999	Cancer
2	João	27	Male	Scientist	537957333	Hypertense
3	Carolina	22	Female	Engineer	658658224	Diabetes
4	Ana	21	Female	Accountant	125896345	Aneurysm
5	Paulo	10	Male	Teacher	546895665	Cholesterol
6	Maria	56	Female	Scientist	756985321	Cancer
7	Pedro	45	Male	Engineer	120365120	Hypertense
8	Ricardo	76	Male	Accountant	125639856	Diabetes
9	Francisca	34	Female	Teacher	125630235	Aneurysm
10	Maria	36	Female	Scientist	896451230	Cholesterol
11	Teresa	44	Female	Engineer	890246879	Cancer
12	Vera	23	Female	Accountant	215634895	Hypertense
13	Rui	19	Male	Teacher	123456789	Diabetes
14	Tiago	67	Male	Scientist	120365201	Aneurysm
15	Henrique	43	Male	Engineer	890560235	Cholesterol
16	Carlos	89	Male	Accountant	452782982	Cancer
17	Rita	34	Female	Teacher	325658985	Hypertense
18	Catarina	17	Female	Scientist	415756324	Diabetes
19	Micaela	21	Female	Engineer	120120120	Aneurysm
20	Micael	42	Male	Accountant	321654987	Cholesterol

Figure 50 - Dataset for Experiences

The DATA file has the same content as the CSV file, except the first row, the one with the attribute names, which is removed in the DATA file, and content in this file has the following structure:

```
1,Joana,26,Female,Teacher,273649999,Cancer
2,João,27,Male,Scientist,537957333,Hypertense
```

The rest of the file can be seen in Annex B.

Now, in this folder, the missing file is the one named header.txt. This file has the name of the dataset file, the name of each attribute and the possible values for each attribute and finishes with a @DATA tag. The content of the file is the following:

```
@RELATION first_try
@ATTRIBUTE Name {Joana, João, Carolina, Ana, Paulo, Maria, Pedro, Ricardo,
Francisca, Teresa, Vera, Rui, Tiago, Henrique, Carlos, Rita, Catarina, Micaela,
Micael}
@ATTRIBUTE Age NUMERIC
@ATTRIBUTE Sex {Female, Male}
@ATTRIBUTE Job {Teacher, Scientist, Engineer, Accountant}
@ATTRIBUTE Social_Security_Number NUMERIC
@ATTRIBUTE Disease {Cancer, Hypertense, Diabetes, Aneurysm, Cholesterol}
@DATA
```

The most difficult file to complete is the config.xml file, where the algorithm is selected, the quasi-identifier attributes, the sensitive attributes and all the values allowed for the anonymization output.

The first element to set is the config method, where the algorithm to use is specified and the value for k. The value for k chosen was 2 because, as mentioned in chapter 5, the higher the value for k, the bigger is the information loss, although privacy is also higher, this anonymization process aims to maintain the data utility as high as possible. As 1 for the k value was a really low value, the value 2 was the best smaller option.

```
<config method = 'Datafly' k = '2'>
```

Then, the elements to fill are the input and the output file. In the input file element, the separator is also set, and for the output, the format should also be chosen.

```
<input filename='dataset/first_try.data' separator=','/>
<output filename='first_try_output.data' format ='genVals'/>
```

The next elements are the ones related to the attributes, where the identifiers, the quasi-identifiers and the sensitive attributes are defined.

At the identifier's attributes, the Name attribute was set because it is the one that immediately identifies the individual. Setting this attribute as an identifier causes this attribute to be removed from the output file.

```
<id>
        <att index='1' name='Name'/>
</id>
```

Then comes the quasi-identifier attributes, where, for each quasi-identifier attribute the mapping (for category values) or the ranges (for numerical values) that should be used to transform the data and anonymize it needs to be set.

The quasi-identifier attributes present in this dataset are Age, Sex and Job.

For Age, first is specified the biggest range of values for this attribute (the min value in the dataset and the max value also in the dataset). Then it is splitted into two sets, one that group the ages from 0 to 50, exclusive, and the other one from 50 (inclusive) to 90.

```
<att index='2' name ='Age'>
            <vgh value='[0:90)'>
                <node value='[0:50)'>
                    <node value='[10:25)'/>
                    <node value='[25:50)'/>
                </node>
                <node value='[50:90)'>
                    <node value='[50:65)'/>
                    <node value='[65:90)'/>
                </node>
            </vgh>
</att>
```

For the Sex attribute, it is needed to map the categorical value to a discrete numeric value. In the end, to set the interval of numeric values it is used to map the categoric values for the Sex attribute.

```
<att index='3' name='Sex'>
            <map>
                <entry cat='Female' int='0' />
                <entry cat='Male' int='1' />
            </map>
            <vgh value='[0:1]'
            </vgh>
</att>
```

The job attribute is created the same way as the Sex attribute. It is also needed to map the categorical values to numeric values.

```
<att index='4' name='Job'>
            <map>
                <entry cat='Teacher' int='0'
                <entry cat='Scientist' int='1' />
                <entry cat='Engineer' int='2' />
                <entry cat='Accountant' int='3' />
            </map>
            <vgh value='[0:3]'>
            </vgh>
</att>
```

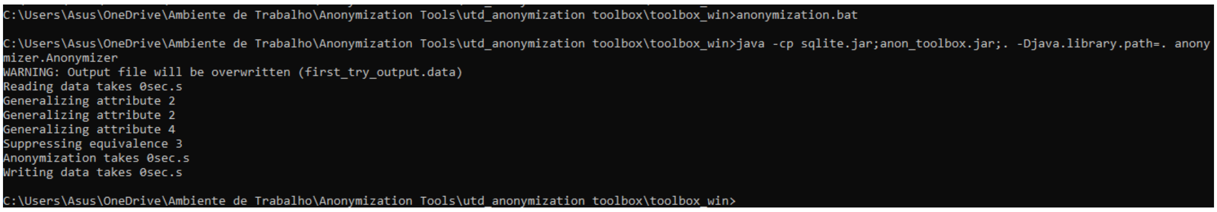
Finally, just missing to set the sensitive attributes. For these attributes, a mapping between the original values and the other numeric values should be done, as it was done for the quasi-identifier attributes.

```
<sens>
        <att index='5' name='Social_Security_Number'>
            <map>
                <entry cat='273649999' int='0' />
```

```
                        <entry cat='537957333' int='1' />
                <map>
        </att>
        <att index='6' name='Disease'>
                <map>
                        <entry cat='Cancer' int='0' />
                        <entry cat='Hypertense' int='1' />
                        <entry cat='Diabetes' int='2' />
                        <entry cat='Aneurysm' int='3' />
                        <entry cat='Cholesterol' int='4' />
                </map>
        </att>
</sens>
```

And this is all the configuration needed to anonymize our small dataset.

For this, the script anonymization.bat need to run in the command line. Figure 51 shows the logs from the script execution.



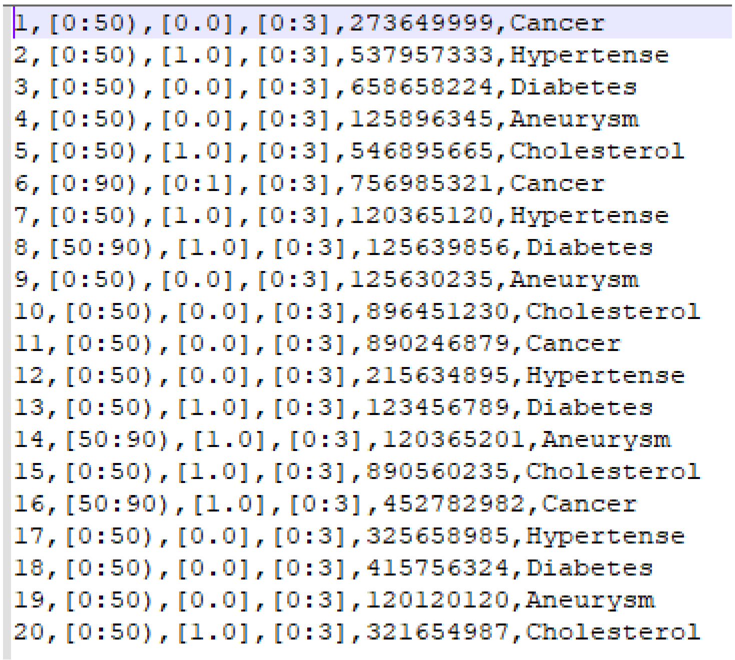
```
C:\Users\Asus\OneDrive\Ambiente de Trabalho\Anonymization Tools\utd_anonymization toolbox\toolbox_win>anonymization.bat

C:\Users\Asus\OneDrive\Ambiente de Trabalho\Anonymization Tools\utd_anonymization toolbox\toolbox_win>java -cp sqlite.jar;anon_toolbox.jar;. -Djava.library.path=. anony
mizer.Anonymizer
WARNING: Output file will be overwritten (first_try_output.data)
Reading data takes 0sec.s
Generalizing attribute 2
Generalizing attribute 2
Generalizing attribute 4
Suppressing equivalence 3
Anonymization takes 0sec.s
Writing data takes 0sec.s

C:\Users\Asus\OneDrive\Ambiente de Trabalho\Anonymization Tools\utd_anonymization toolbox\toolbox_win>
```

Figure 51 - anonymization.bat script logs

After this, it is possible to verify the output file. This file has the content, shown in Figure 52.



```
1,[0:50),[0.0],[0:3],273649999,Cancer
2,[0:50),[1.0],[0:3],537957333,Hypertense
3,[0:50),[0.0],[0:3],658658224,Diabetes
4,[0:50),[0.0],[0:3],125896345,Aneurysm
5,[0:50),[1.0],[0:3],546895665,Cholesterol
6,[0:90),[0:1],[0:3],756985321,Cancer
7,[0:50),[1.0],[0:3],120365120,Hypertense
8,[50:90),[1.0],[0:3],125639856,Diabetes
9,[0:50),[0.0],[0:3],125630235,Aneurysm
10,[0:50),[0.0],[0:3],896451230,Cholesterol
11,[0:50),[0.0],[0:3],890246879,Cancer
12,[0:50),[0.0],[0:3],215634895,Hypertense
13,[0:50),[1.0],[0:3],123456789,Diabetes
14,[50:90),[1.0],[0:3],120365201,Aneurysm
15,[0:50),[1.0],[0:3],890560235,Cholesterol
16,[50:90),[1.0],[0:3],452782982,Cancer
17,[0:50),[0.0],[0:3],325658985,Hypertense
18,[0:50),[0.0],[0:3],415756324,Diabetes
19,[0:50),[0.0],[0:3],120120120,Aneurysm
20,[0:50),[1.0],[0:3],321654987,Cholesterol
```

Figure 52 - Anonymized Dataset

Some of the values were anonymized but the social_security_number and the disease attributes were not anonymized, and it should be because they are sensitive attributes. For the dataset to be anonymized at a superior level, a bigger value of K should be given in the config element in the config.xml file, with the risk of increasing the information loss and, consequently, the data utility of the anonymized dataset.

7. Assessment of the Anonymization Tools

OSSpal methodology has the aim to evaluate open-source tools in order to help users and organizations to find the best ones, using a set of categories, as better explained in section 1 of this chapter.

Pereira, Sousa, Santos, & Bernardino (2018) evaluated the three of the most used data mining tools using OSSpal. The tools they evaluate were Knime, RapidMiner and Weka. They have used seven categories to evaluate the tools: functionality, operational software characteristics, support and services, documentation, software technology attributes, community and adaption and development process. For each category, they applied a weight, and then, for the evaluation, they applied scores, from 1 to 5 for each category. In the end, multiplying the weight for each functionality by the score, they have the final score for each category. The sum of each score is the final evaluation value for each tool. To evaluate the tools, besides the experience they had with them, they collect technical documentation and use the website of each tool. The RapidMiner was the tool with the best score, which means that this one is the best open-source tool (in the three they studied), which can also be proved by the higher number of users this tool has related to the others.

In this chapter, the evaluation with OSSpal is done to the anonymization tools used: ARX Data Anonymization Tool, Amnesia and UTD Anonymization Toolbox.

The metrics used to compare the tools are the number of anonymization algorithms; the number of anonymization techniques; the existence of a graphical interface; the facility in the dataset anonymization; the tool usability; and the number of bugs present in the tool.

The algorithms and the techniques that are used in all tools are the ones that are going to be explained in this thesis. The tools are evaluated using the OSSpal methodology, which is explained in the next subchapter.

7.1. Evaluating Data anonymization Tools with OSSpal

The OSSpal methodology was created in the Business Readiness Rating (BRR) project, which began in 2005. The goal of this work was to help users and organizations to find the best free open-source software, based on a set of categories that are evaluated in each software.

The categories used to evaluate the software are the following ones:

- Functionality
- Operational Software Characteristics

- Support and Services
- Documentation
- Software Technology Attributes
- Community and Adaption
- Development Process

The assessment process of OSSpal methodology for all categories, except for the functionality category, is composed of four phases:

- **The first phase** is where the software components are identified and selected to be evaluated against some criteria.
- **In the second phase** are assigned the weights for the criteria and for the measures, where each criterion has a percentage. The total percentages of all criteria must be 100% and then, for each measure within a category, it is necessary to rank the measure following its importance and assign it.
- **In the third phase** is where some data is collected to help the user to calculate the weight in a range between 1 to 5 (1- Unacceptable, 2- Poor, 3- Acceptable, 4- Very Good, 5- Excellent).
- Finally, **in the fourth phase**, the OSSpal final score is calculated.

The 'Functionality' category is calculated differently than the others. First, the characteristics to be evaluated are chosen and evaluated in scoring from 1 to 3 (less important to most important) and the characteristics are classified in this range. Then the first results should be standardized to a scale from 1 to 5. This category will have the following scale:

- Under 65% - score = 1 (Unacceptable)
- 65% to 80% - score = 2 (Poor)
- 80% to 90% - score = 3 (Acceptable)
- 90% to 96% - score = 4 (Very Good)
- 96% to 100% - score = 5 (Excellent)

To assign a percentage for each category is important to know exactly what each one of them represents (Pereira, Sousa, Santos, & Bernardino, 2018):

- Functionality – As functionality, is meant if the software is according to the user requirements.
- Operational Software Characteristics – evaluates if the software is saved, has a good performance, the user interface exists or works correctly, is easy to use, install, configure, maintain, and deploy.

- Support and Services – verify if the software has good community or commercial support. It also evaluates if the organization gives some training to help with the software usage.
- Documentation – Evaluates if the existing documentation is good enough to help with the tool.
- Software Technology Attributes – In this criterion is verified if the software architecture is good enough, if the software is portable, extensible, open, and easy to integrate, the code, design and tests have the needed quality, if it is bug-free or if it is complete.
- Community and Adaption – Measures if the software community is active and if the component is well accepted by the community.
- Development Process – Rates how good is the professionalism at the development point and project organization.

For each category, the following percentages were defined:

- Functionality – 30%
- Operational Software Characteristics – 20%
- Support and Services – 10%
- Documentation – 105%
- Software Technology Attributes – 15%
- Community and Adaption – 5%
- Development Process – 10%

Functionality is the category with a higher percentage (35%) because the software must comply with the user's needs. After Functionality comes Operational Software Characteristics (20%). A graphical user interface should be a mandatory requirement in a tool due to the inherent complexity of the subject. Therefore, a graphical interface is very helpful in the process of anonymizing a dataset.

Software technology attributes have a value of 15% because is important that it does not have bugs and it should be as complete as possible.

Documentation has a percentage of 10%. In these tools, especially the ones that do not have a graphical user interface is mandatory to have good documentation, otherwise, it is difficult for the user to have the expected results from the tool.

Support and Services has the percentage of 10% because, in an open-source tool, the users do not expect much help or support from the developers and from the organization that developed the tool.

The development process has 10% because it is an important category in the evaluation of the tools. In article 30 of the GDPR, some guidelines should be used by anonymization tools to agree with the regulation. A record in the tool should be maintained by a controller or its representative and should contain: the name and contacts of the controller, the controller's representative and the Data Protection Officer, the purposes of the processing, the categories of the data subject, the time limits, and the security (Cantiello, Mastroianni, & Rak , 2021).

Finally, the Community and Adaption has the less percentage of all categories with 5%. This is because, we consider the other categories more important, such as to have a good documentation

The functionality criteria need to be classified with the help of some characteristics. Table 17 presents the characteristics and the weight that each one has for the evaluation of these tools.

Table 17 - Functionality Category – Characteristics and Weights

Characteristics	Weight
Number of algorithms	2
Anonymized data visualization	1
Algorithm application	3
Anonymization process	3

The number of algorithms is important because it makes the tool more complete. The visualization of the anonymized data is also an important functionality. After anonymizing our dataset, an important thing to do is to check if the data is well anonymized and, verify some metrics after the anonymization. If the visualization of this data is not easy, it is more difficult to understand if the results are as expected.

The application of the algorithm to a dataset and how easy it is to anonymize a dataset are the main characteristics of these tools. The anonymization process is already too difficult, and it does not need a tool to complicate the process.

For each tool evaluation, first is given a value, between 1-5 for each Functionality criteria. Then, the other criteria are also evaluated from 1 to 5.

In the next sections are presented the evaluation of the used tools using OSSpal methodology according to the percentages and weights given.

7.2. Evaluating ARX Data Anonymization Tool with OSSpal

ARX Data Anonymization Tool was the simplest to use. It has a graphical interface that helps with the dataset upload and its anonymization. One of the biggest problems with this anonymization tool is that the help icon in some menus is not updated with the last version of the tool and sometimes they do not have the information needed to perform some actions. It is also difficult to understand what needs to be done in some windows because of the documentation not being so understandable.

For the functionality criteria, the values given are represented in Table 18.

Table 18 - Functionality Category – Weights ARX Data Anonymization Toolbox

Characteristics	Weight	Value
Number of algorithms	2	2
Anonymized data visualization	1	1
Algorithm application	3	3
Anonymization process	3	0

The tool has several algorithms that could be used to anonymize our dataset, but it does not allow the user to choose the algorithm s/he wants. For example, it shows a table with some algorithms linked to our sensitive attributes and other algorithms that do not have a sensitive attribute linked. If the user wants to choose one of these last algorithms the tool does not allow it, and therefore the value 2 was given. For the anonymized data visualization, the value given was 1 because the results appear in a table side-by-side with the original dataset, so it is easy to compare both versions of the dataset (the original one and the anonymized one). The algorithm application has a value of 3 because it is easy to apply an algorithm to the sensitive attributes, and besides the hierarchy creation is not simple, the tool gives different options considering the attribute type to create the hierarchy. The anonymization process has 0 because sometimes it is a considerable amount of information to deal with in the tool and it can be confusing instead of helping the user with the process.

Converting these values with the weight, the operation that needs to be performed is as follows: (6 x 100) / 9 = 66,7%, which means, on the scale for the Functionality criteria, that the value for this category is 2.

The values given for all the categories are the following ones:

- Functionality – 2 * 30% = 0.6
- Operational Software Characteristics – 5 * 20% = 1
- Support and Services – 4 * 10% = 0.4
- Documentation – 3 * 10% = 0.3
- Software Technology Attributes – 4 * 15% = 0.6
- Community and Adaption – 4 * 5% = 0.2
- Development Process – 0 * 10% = 0

The score given to Operational Software Characteristics is 5 due to the great graphical user interface which significantly helps the users to anonymize their data. The tool is also easy to install, easy to configure and easy to use. Support and Services and Community and Adaption have a score of 4 because there are many articles and much information on the Internet about this tool. Some of these articles are mentioned on their website. These last points were also important to the Documentation criteria, but the way that the documentation is used in the tool is not good and it makes it difficult for the users to have a valuable experience with it, which is why the score 3 was defined. Software Technology Attributes has 4 because the architecture seems good, and the tool does not have bugs. Finally, the Development Process has value of 0, because the tool does not maintain any records about the controller.

The result of all criteria values is: 0.6 + 1 + 0.4 + 0.3 + 0.6 + 0.2 + 0 = 3.1

So, the result given for ARX Data Anonymization Tool is 3.1 out of 5, between Acceptable and Very Good.

7.3. Evaluating Amnesia with OSSpal

Amnesia was not so difficult to use. The bigger problem with this anonymization tool is that sometimes it takes too long to perform an action (for example, to anonymize a small and simple dataset).

For Amnesia, the weights given for the functionality criteria characteristics are represented in Table 19.

Table 19 - Functionality Category – Weights Amnesia

Characteristics	Weight	Value
Number of algorithms	2	0
Anonymized data visualization	1	1
Algorithm application	3	0
Anonymization process	3	3

The tool has not had many anonymization algorithms for the user to choose the best one to apply, this is why the characteristic for the number of algorithms has the value of 0. For the anonymized data visualization, the value is 1 because, after doing a study with 7 attributes, the solution graph with the anonymization results was huge and it was very difficult to verify the best result for the dataset used, but then, after choosing a solution, it is possible to visualize the anonymized dataset. The algorithm application has a value of 0 because of the offer given by the tool for the algorithms to apply for the anonymization process. The only step where the user could have some doubts is in the hierarchy creation, but the rest of the process is easy to perform. The anonymization process has 3 because sometimes it could take more time than expected, but the results are retrieved.

To calculate the value for the Functionality criteria, it is needed to perform the following operation: (4 x 100) / 9 = 44,4%, which means, in the scale for the Functionality criteria, that the value for this category is 1.

The values given for all the categories are the following ones:

- Functionality – 1 * 30% = 0.3
- Operational Software Characteristics – 4 * 20% = 0.8
- Support and Services – 4 * 10% = 0.4
- Documentation – 3 * 10% = 0.3
- Software Technology Attributes – 3 * 15% = 0.45
- Community and Adaption – 4 * 5% = 0.2
- Development Process – 0 * 10% = 0

The operational software characteristics have a 4 because in this case, the tool has a graphical user interface, and the visual solution perception helps a lot in the tool usage. This tool is easy to install, easy to configure and easy to use, the only issue is when it takes a little longer to anonymize the dataset. Support and services and community and adaption have 4 because the website has some different spaces where the user can reach the organization and it also has a Twitter account where the users could make some comments on the tool, which means that the organization want to be close to the users. The documentation for the tool is not bad, it shows all the things that the user can make at each step in the anonymization process. It could be more helpful in the creation of hierarchy files, for example, which justifies the value 3 on the documentation criteria. Software Technology Attributes has 3 because the architecture seems good. The tool has a small number of bugs, and it could be more complete with other anonymization algorithms. Finally, the development process has a value of 0, because the tool does not save any record about the controller.

The result of all criteria values is: 0.3 + 0.8 + 0.4 + 0.3 + 0.45 + 0.2 + 0 = 2.45

So, the result given for Amnesia is 2.45 in 5, between Poor and Acceptable.

7.4. Evaluating UTD Anonymization Toolbox with OSSpal

UTD Anonymization Toolbox was the most difficult tool to use. A large part of the process needs to be done manually, which is a way of incorporating some errors performed by the user leading to more difficulties in the data anonymization.

Table 20 represents the values given for weight to each characteristic of the Functionality criteria.

Table 20 - Functionality Criteria – Weights UTD Anonymization Toolbox

Characteristics	Weight	Value
Number of algorithms	2	1
Anonymized data visualization	1	0
Algorithm application	3	2
Anonymization process	3	0

The number of algorithms in the tool is adequate, that is why the value given is 1, because, compared with the other tools, it could have more algorithms. The anonymized data visualization is hard for some algorithms because the output file is a .txt file, without the column names, and when the anonymization algorithm is more complex, the output results are also more difficult to understand, this is why the value given is 0.

The application of the algorithm, although it is done manually, there just needs to be selected one algorithm from the existing ones and insert in the correct element in the XML file, this is why the value given is 2. As all the data need to be inserted manually, this makes the anonymization process more difficult, and it has a higher probability of introducing errors. Besides this, all the information needed to anonymize each attribute need also to be inserted manually, so the anonymization process is slow and difficult, and these are the reasons why the Anonymization Process characteristic has the value of 0.

To calculate the value for the Functionality criteria, it is needed to perform the following operation: (3 x 100) / 9 = 33,3%, which means, in the scale for the Functionality criteria, that the value for this category is 1.

The values given for all the categories are the following ones:

- Functionality – 1 * 30% = 0.3
- Operational Software Characteristics – 3 * 20% = 0.2
- Support and Services – 2 * 10% = 0.2
- Documentation – 2 * 10% = 0.2
- Software Technology Attributes – 3 * 15% = 0.45

- Community and Adaption – 1 * 5% = 0.05
- Development Process – 2 * 10% = 0.2

For operational software characteristics, the value given was 3 because the graphical user interface should be given to help the users with the tool. The lack of a graphical user interface makes it difficult to use and the configuration files are not easy to create.

For support and services and community and adaption, was not given a good mark because it was not found any forum where some questions could be posted. It seems like there is no community for this tool.

The documentation also has 2 because, according to the complexity of this tool, the documentation is not complete. It could be simple for someone that already knows or developed the tool, but for another user, it is difficult and hard to use it.

Software technology attributes have 3 because the logging is not good, although the tool does not have many errors. The difficulty here is when one is trying to anonymize the dataset and have an error (because of the configuration file, for example) and it does not have a friendly error message for the user that helps him to understand what is wrong.

Finally, the development process has a value of 0, because the tool does not save any record about the controller.

The result of all criteria values is: 0.3 + 0.2 + 0.2 + 0.2 + 0.45 + 0.05 + 0.2 = 1.6

Therefore, the result given for UTD Anonymization Toolbox is 1.6 in 5 which situates this tool between Unacceptable and Poor.

7.5. OSSpal Evaluation Summary

In the previous subsections, an evaluation of each open-source tool was done. As it can be verified in each subsection, the results for each tool are:

- ARX Data Anonymization Tool: 3.1 in 5
- Amnesia: 2.45 in 5
- UTD Anonymization Toolbox: 1.6 in 5

This means that ARX Data Anonymization Tool is the one that has the best evaluation. One thing that proves this is the fact that this is one of the most complete and most used tools. There are a lot of papers and discussions all over the internet regarding this tool. UTD is the one that is not so recognized unless the papers about the tool are read, but this one is not so discussed in the interned.

Consequently, it is possible to conclude that ARX Data Anonymization is the best of the three tools and UTD Anonymization Toolbox is the one that has the worst evaluation.

8. Datasets Used in the Experiments

To analyse the tools mentioned in the previous chapter some datasets are used. As it should, the original datasets are not published on the Internet because it is not allowed to publish personal information without the individual agreement. This could represent a violation of privacy. Therefore, in this thesis, the real datasets provided online are used but some new columns are added to the datasets, to turn the data into something more non-anonymized.

The datasets used are explained in the next subsections.

8.1. Heart Failure Prediction Dataset

Cardiovascular diseases are the number one cause of death, globally, and heart failure is caused by cardiovascular diseases.

This dataset, which is a combination of five different datasets in eleven common features, making it the biggest dataset about heart disease available for research purposes, is useful because it can be used to predict a possible heart disease based on the presented features.

This dataset is composed of 12 attributes, which are represented in the next table as well as the explanation of the attributes (Fedesoriano, 2021).

The dataset has the original size of 35921 bytes but with the changes performed it stays with 52067 bytes and has 918 rows in the table.

Kaggle is the website from where the dataset was downloaded. This is a subsidiary of Google LLC where the users can download and upload datasets.

Table 21 - Description of the Dataset Attributes

Attribute	Description
Age	Age of the patient in years
Sex	Sex of the patient. The values could be Male or Female.
ChestPainType	The type of chest pain. This could be: • TA – Typical Angina • ATA – Atypical Angina • NAP – Non-Anginal Pain • ASY – Asymptomatic
RestingBP	Resting Blood Pressure in [mm Hg]

Cholesterol	Serum Cholesterol in [mm/dl]
FastingBS	Fasting Blood Sugar. The values for this attribute are: • 1, if FastingBS > 120mg/dl • 0, otherwise
RestingECG	Resting Electrocardiogram Results. This attribute allows the following values: • Normal – Normal • ST – having ST-T wave abnormality • LVH – showing probable or definite left ventricular hypertrophy by Estes' criteria
MaxHR	The Maximum Heart Rate achieved is a numeric value between 60 and 202
ExerciseAngina	Exercise-induced Angina. This could be Y: Yes or N: No.
Oldpeak	Oldpeak = ST which is the numeric value measured in depression
ST_Slope	The slop of the peak exercise ST-segment allows the values: • Up – upsloping • Flat – Flat • Down – Downsloping
Heart Disease	This is the output result, and it could be 1 in case of heart disease or 0 if the patient has normal results.

This dataset does not have an identifier attribute, so it is going to be added manually to the dataset. The attributes added are Name and Social_Security_Number. The name is going to be the patient's name and the Social_Security_Number is a number with nine digits that identify the patient in the national health service.

In the end, this dataset has 14 attributes. In the Table 22 is identified the attribute type for each attribute, which will help the anonymization in the different tools.

Table 22 - Attribute's type

Attribute	**Attribute Type**
Name	Identifier
Age	Quasi-Identifier
Sex	Quasi-identifier
Social_Security_Number	Identifier
ChestPainType	Non-sensitive
RestingBP	Non-sensitive
Cholesterol	Non-sensitive
FastingBS	Non-sensitive
RestingECG	Non-sensitive

MaxHR	Non-sensitive
ExerciseAngina	Non-sensitive
Oldpeak	Non-sensitive
ST_Slope	Non-sensitive
HeartDisease	Sensitive

8.2. Pfizer Vaccine Tweets Dataset

This dataset has the most recent tweets about Pfizer & BioNTech vaccine. This data was collected from Twitter using a python package that accesses Twitter API and collects the data (Preda, 2021). This dataset was also downloaded from Kaggle.

This dataset needed to be changed in the username attribute because it had symbols and the tool does not allow the upload of this type of character. The size of the dataset is 3488606 bytes, and it has 11021 registries. This dataset is composed by 16 attributes, which are descripted in Table 23.

Table 23 - Description of the Dataset Attributes

Attribute	Description
id	Id of the tweet
username	The username from the person who posted the tweet
userlocation	The location from the person who posted the tweet
userdescription	The description from the person who posted the tweet
usercreated	The date when the user joins Twitter
userfollowers	Number of the followers the user has
userfriends	Number of the friends the user has
userfavourites	Number of favourites the user has
userverified	Boolean value that says if the user is verified
date	Date of the post
text	The text posted on Twitter
hashtags	The hashtags inserted by the user
source	From where the data was obtained
retweets	Number of times the post was retweeted

Attribute	Description
favorites	Number of favourites made on the post
isretweet	If the post is a retweet from another user's post

The next step is identifying the attribute types. As it was mentioned before, the attribute types can be sensitive, quasi-identifier, identifier, and non-sensitive attributes. In Table 24 are the types set for each one of the attributes.

Table 24 - Attribute's type

Attribute	Attribute type
id	Sensitive
username	Identifier
userlocation	Quasi-identifier
userdescription	Quasi-identifier
usercreated	Non-sensitive
userfollowers	Non-sensitive
userfriends	Non-sensitive
userfavourites	Non-sensitive
userverified	Non-sensitive
date	Non-sensitive
text	Non-sensitive
hashtags	Non-sensitive
source	Non-sensitive
retweets	Non-sensitive
favorites	Non-sensitive
isretweet	Non-sensitive

Now, this dataset is going to be inserted in the different tools to be anonymized. The results can be seen in the next chapter.

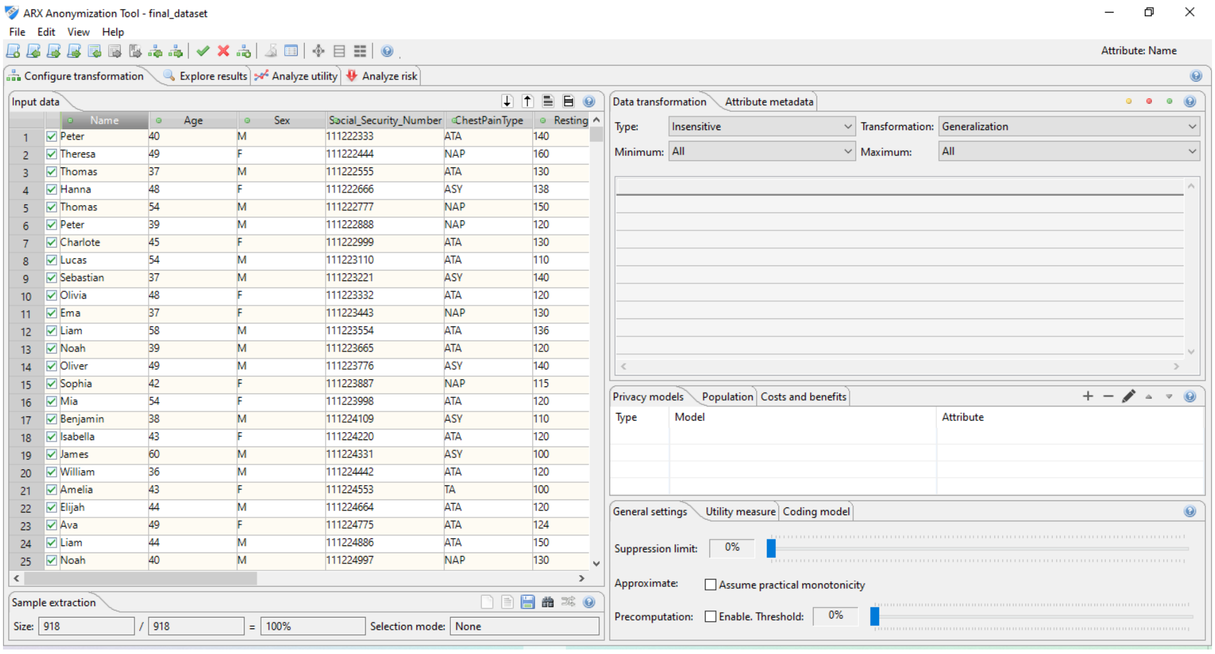
As the UTD is the tool with the worst results and needs all the strings to be introduced manually it is not going to be used in this anonymization process. The dataset is going to be uploaded in Amnesia and ARX Data Anonymization Tool.

9. Tools Assessment with the Datasets

In this chapter, the datasets of the previous chapter are introduced in the data anonymization tools. It is described the important steps performed in each tool and the respective results.

9.1. Heart Failure Prediction Dataset in ARX Data Anonymization Tool

Finally, the dataset is inserted in ARX Data Anonymization Tool. To use this tool, it is needed to create a project first and then import the dataset. As was previously explained, to import the data, it is needed to select the input file format, which is CSV in this case, then the delimiter, and in this dataset, the semi-comma is selected, then the data types are automatically assigned to each attribute, but they can be changed if the user wants to correct any of the attribute types. Finally, it is shown a preview of the input dataset. After all these steps the dataset is uploaded to the tool, and it can be seen in the "Input data" separator, as is shown in Figure 53.



		Name	Age	Sex	Social_Security_Number	ChestPainType	Resting
1	☑	Peter	40	M	111222333	ATA	140
2	☑	Theresa	49	F	111222444	NAP	160
3	☑	Thomas	37	M	111222555	ATA	130
4	☑	Hanna	48	F	111222666	ASY	138
5	☑	Thomas	54	M	111222777	NAP	150
6	☑	Peter	39	M	111222888	NAP	120
7	☑	Charlote	45	F	111222999	ATA	130
8	☑	Lucas	54	M	111223110	ATA	110
9	☑	Sebastian	37	M	111223221	ASY	140
10	☑	Olivia	48	F	111223332	ATA	120
11	☑	Ema	37	F	111223443	NAP	130
12	☑	Liam	58	M	111223554	ATA	136
13	☑	Noah	39	M	111223665	ATA	120
14	☑	Oliver	49	M	111223776	ASY	140
15	☑	Sophia	42	F	111223887	NAP	115
16	☑	Mia	54	F	111223998	ATA	120
17	☑	Benjamin	38	M	111224109	ASY	110
18	☑	Isabella	43	F	111224220	ATA	120
19	☑	James	60	M	111224331	ASY	100
20	☑	William	36	M	111224442	ATA	120
21	☑	Amelia	43	F	111224553	TA	100
22	☑	Elijah	44	M	111224664	ATA	120
23	☑	Ava	49	F	111224775	ATA	124
24	☑	Liam	44	M	111224886	ATA	150
25	☑	Noah	40	M	111224997	NAP	130

Size: 918 / 918 = 100% Selection mode: None

Type: Insensitive Transformation: Generalization
Minimum: All Maximum: All

Type	Model	Attribute

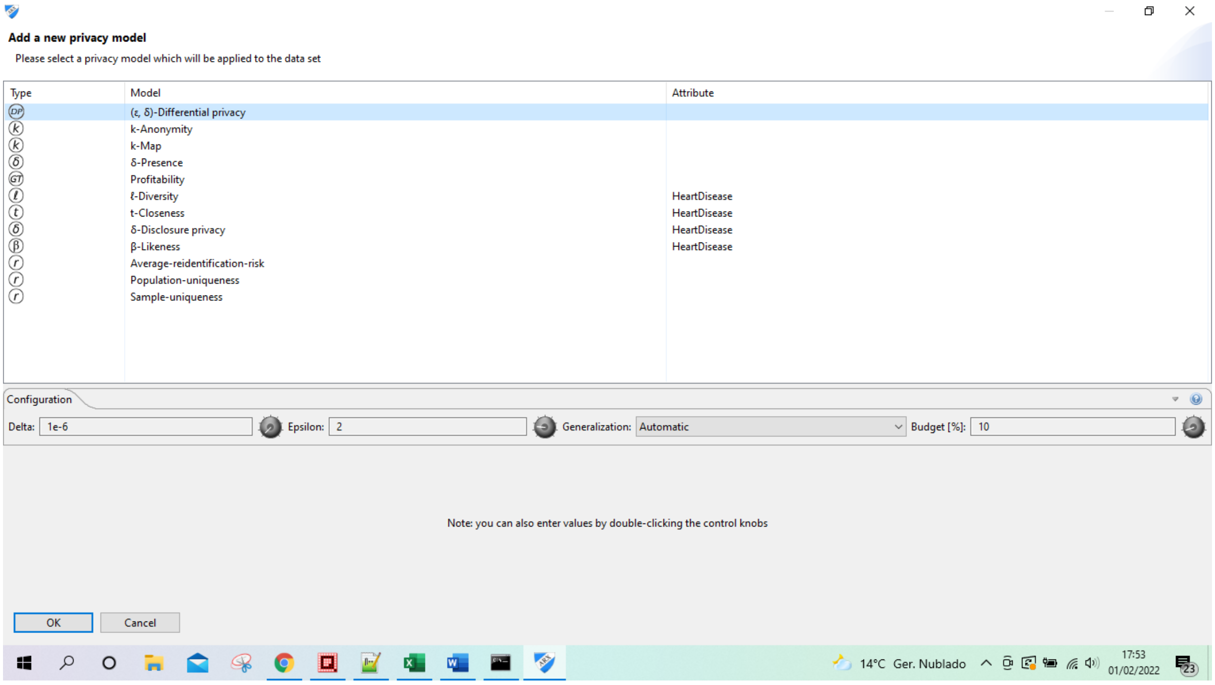
Suppression limit: 0%
Approximate: □ Assume practical monotonicity
Precomputation: □ Enable. Threshold: 0%

Figure 53 - ARX Display with Dataset Uploaded

The next step is to assign the type for each attribute. The types defined for this dataset are the same that was mentioned in Table 22.

The variables marked as target variables are Name, Age, Sex, Social_Security_Number and HeartDisease.

The next step is to set the privacy models to the sensitive attributes. This tool gives the options that are available for the sensitive attribute, and, in this case, the options given were l-Diversity, t-Closeness, δ-Disclosure privacy and β-Likeness, as it is possible to verify in Figure 54.



Add a new privacy model

Please select a privacy model which will be applied to the data set

Type	Model	Attribute
DP	(ε, δ)-Differential privacy	
k	k-Anonymity	
k	k-Map	
δ	δ-Presence	
GT	Profitability	
ℓ	ℓ-Diversity	HeartDisease
t	t-Closeness	HeartDisease
δ	δ-Disclosure privacy	HeartDisease
β	β-Likeness	HeartDisease
r	Average-reidentification-risk	
r	Population-uniqueness	
r	Sample-uniqueness	

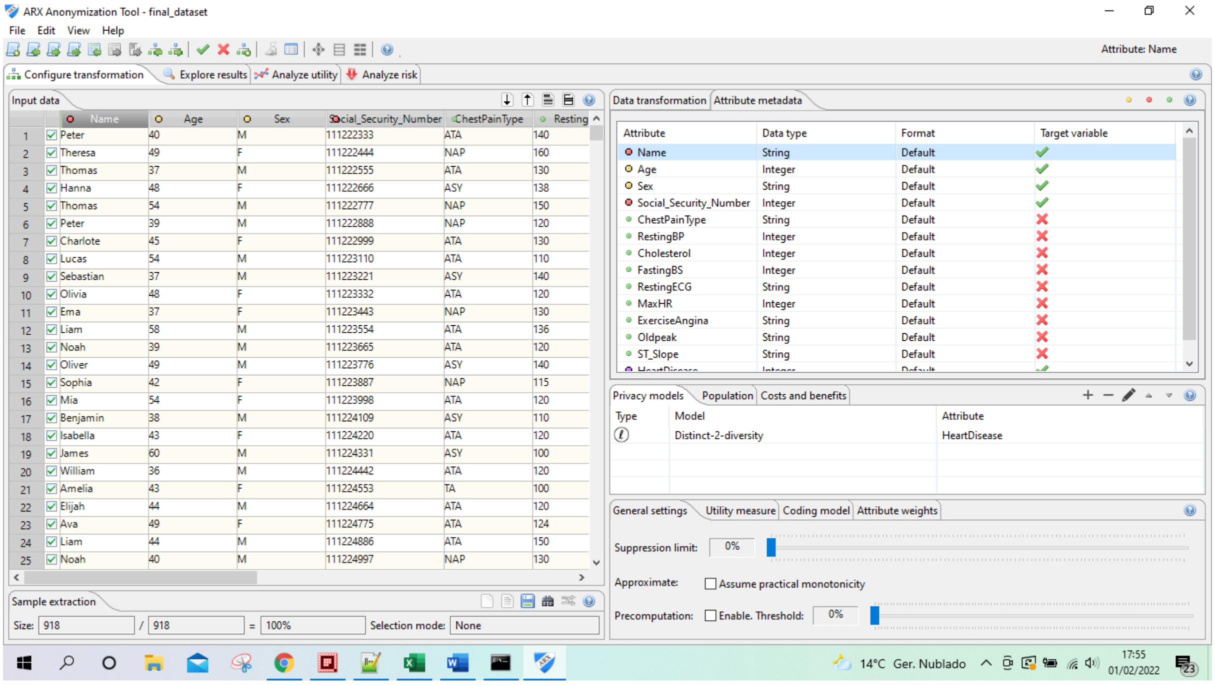
Configuration

Delta: 1e-6 Epsilon: 2 Generalization: Automatic Budget [%]: 10

Note: you can also enter values by double-clicking the control knobs

Figure 54 - ARX Allowed Privacy Models

As the l-Diversity is the closest algorithm to k-anonymity, this is the one chosen because the other tools also use k-anonymity. The l value selected is 2 also because this was the value used in the other tools. Figure 55 shows the privacy model created.



		Name	Age	Sex	Social_Security_Number	ChestPainType	Resting
1	☑	Peter	40	M	111222333	ATA	140
2	☑	Theresa	49	F	111222444	NAP	160
3	☑	Thomas	37	M	111222555	ATA	130
4	☑	Hanna	48	F	111222666	ASY	138
5	☑	Thomas	54	M	111222777	NAP	150
6	☑	Peter	39	M	111222888	NAP	120
7	☑	Charlote	45	F	111222999	ATA	130
8	☑	Lucas	54	M	111223110	ATA	110
9	☑	Sebastian	37	M	111223221	ASY	140
10	☑	Olivia	48	F	111223332	ATA	120
11	☑	Ema	37	F	111223443	NAP	130
12	☑	Liam	58	M	111223554	ATA	136
13	☑	Noah	39	M	111223665	ATA	120
14	☑	Oliver	49	M	111223776	ASY	140
15	☑	Sophia	42	F	111223887	NAP	115
16	☑	Mia	54	F	111223998	ATA	120
17	☑	Benjamin	38	M	111224109	ASY	110
18	☑	Isabella	43	F	111224220	ATA	120
19	☑	James	60	M	111224331	ASY	100
20	☑	William	36	M	111224442	ATA	120
21	☑	Amelia	43	F	111224553	TA	100
22	☑	Elijah	44	M	111224664	ATA	120
23	☑	Ava	49	F	111224775	ATA	124
24	☑	Liam	44	M	111224886	ATA	150
25	☑	Noah	40	M	111224997	NAP	130

Size: 918 / 918 = 100% Selection mode: None

Attribute	Data type	Format	Target variable
Name	String	Default	✓
Age	Integer	Default	✓
Sex	String	Default	✓
Social_Security_Number	Integer	Default	✓
ChestPainType	String	Default	✗
RestingBP	Integer	Default	✗
Cholesterol	Integer	Default	✗
FastingBS	Integer	Default	✗
RestingECG	String	Default	✗
MaxHR	Integer	Default	✗
ExerciseAngina	String	Default	✗
Oldpeak	String	Default	✗
ST_Slope	String	Default	✗
HeartDisease	Integer	Default	✓

Type	Model	Attribute
ℓ	Distinct-2-diversity	HeartDisease

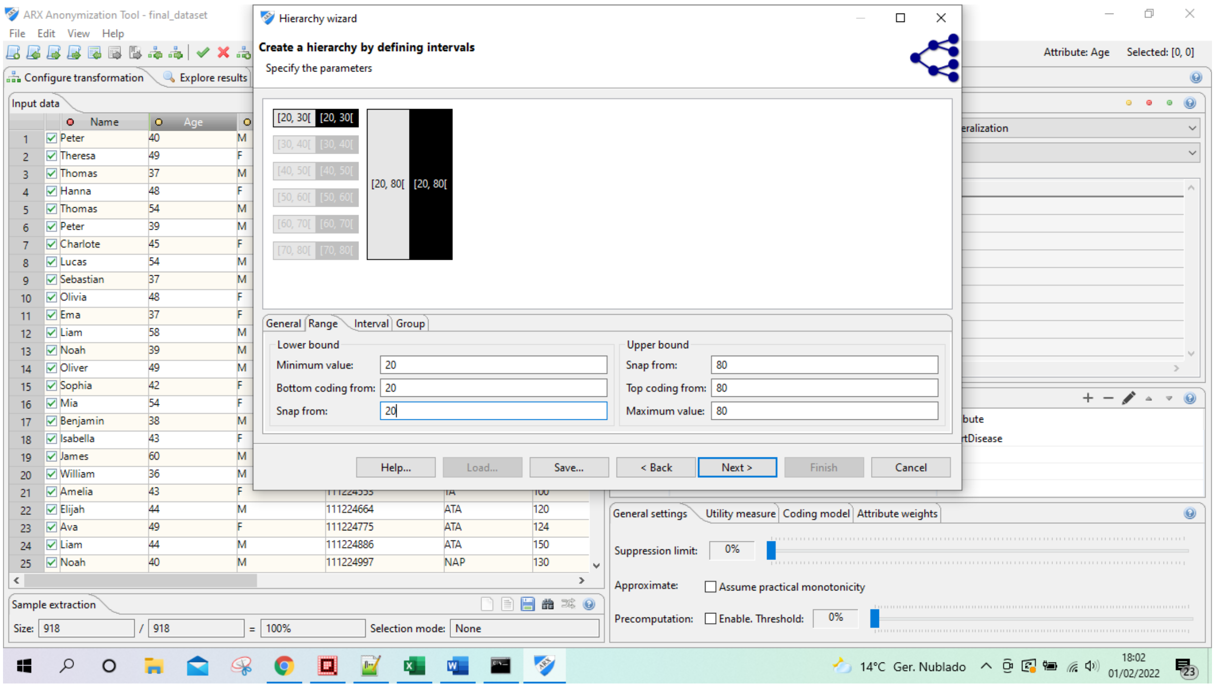
Suppression limit: 0%

Approximate: □ Assume practical monotonicity

Precomputation: □ Enable. Threshold: 0%

Figure 55 - ARX Target Variables Display

The quasi-identifier attributes need hierarchies. These were created for Sex and Age attributes. For the Age attribute, the type of the hierarchy selected is “use intervals”, and the intervals were set as they were created in the other tools, as is pictured in Figure 56.



Hierarchy wizard

Create a hierarchy by defining intervals

Specify the parameters

Lower bound

Minimum value: 20

Bottom coding from: 20

Snap from: 20

Upper bound

Snap from: 80

Top coding from: 80

Maximum value: 80

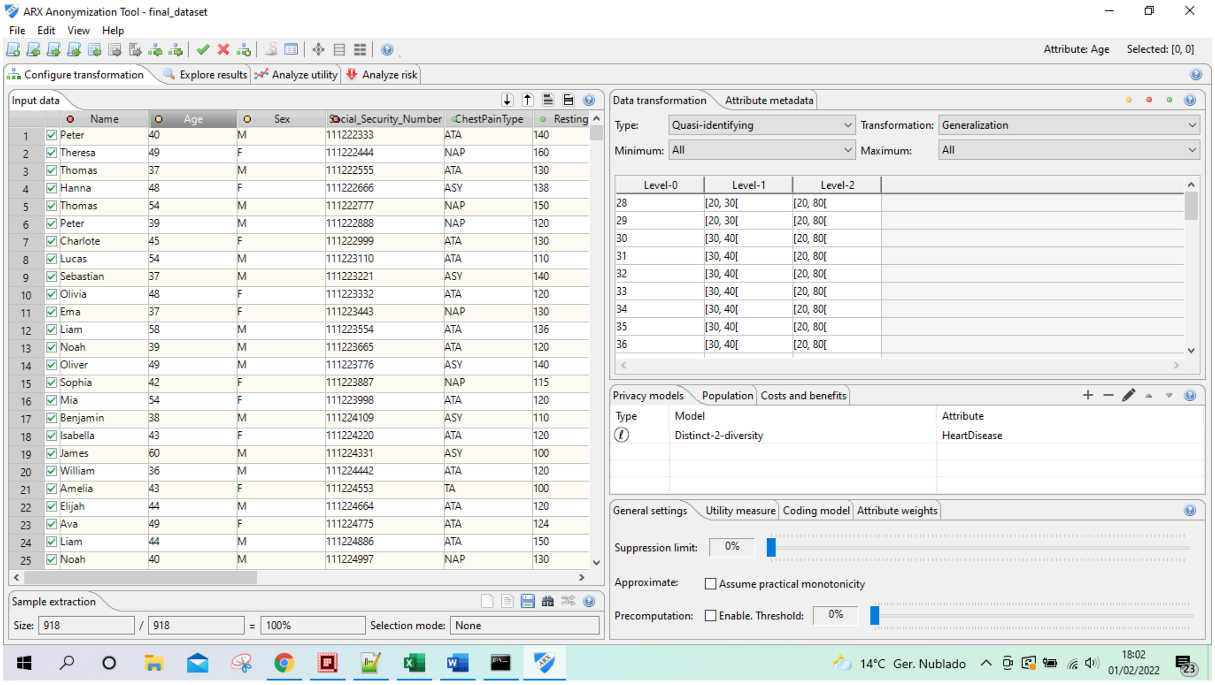
Suppression limit: 0%

Approximate: □ Assume practical monotonicity

Precomputation: □ Enable. Threshold: 0%

Size: 918 / 918 = 100% Selection mode: None

Figure 56 - Create Hierarchy - Age Attribute Hierarchy Intervals



ARX Anonymization Tool - final_dataset

Attribute: Age Selected: [0, 0]

		Name	Age	Sex	Social_Security_Number	ChestPainType	Resting
1	☑	Peter	40	M	111222333	ATA	140
2	☑	Theresa	49	F	111222444	NAP	160
3	☑	Thomas	37	M	111222555	ATA	130
4	☑	Hanna	48	F	111222666	ASY	138
5	☑	Thomas	54	M	111222777	NAP	150
6	☑	Peter	39	M	111222888	NAP	120
7	☑	Charlote	45	F	111222999	ATA	130
8	☑	Lucas	54	M	111223110	ATA	110
9	☑	Sebastian	37	M	111223221	ASY	140
10	☑	Olivia	48	F	111223332	ATA	120
11	☑	Ema	37	F	111223443	NAP	130
12	☑	Liam	58	M	111223554	ATA	136
13	☑	Noah	39	M	111223665	ATA	120
14	☑	Oliver	49	M	111223776	ASY	140
15	☑	Sophia	42	F	111223887	NAP	115
16	☑	Mia	54	F	111223998	ATA	120
17	☑	Benjamin	38	M	111224109	ASY	110
18	☑	Isabella	43	F	111224220	ATA	120
19	☑	James	60	M	111224331	ASY	100
20	☑	William	36	M	111224442	ATA	120
21	☑	Amelia	43	F	111224553	TA	100
22	☑	Elijah	44	M	111224664	ATA	120
23	☑	Ava	49	F	111224775	ATA	124
24	☑	Liam	44	M	111224886	ATA	150
25	☑	Noah	40	M	111224997	NAP	130

Size: 918 / 918 = 100% Selection mode: None

Type: Quasi-identifying Transformation: Generalization

Minimum: All Maximum: All

Level-0	Level-1	Level-2
28	[20, 30[	[20, 80[
29	[20, 30[	[20, 80[
30	[30, 40[	[20, 80[
31	[30, 40[	[20, 80[
32	[30, 40[	[20, 80[
33	[30, 40[	[20, 80[
34	[30, 40[	[20, 80[
35	[30, 40[	[20, 80[
36	[30, 40[	[20, 80[

Type	Model	Attribute
ℓ	Distinct-2-diversity	HeartDisease

Suppression limit: 0%

Approximate: □ Assume practical monotonicity

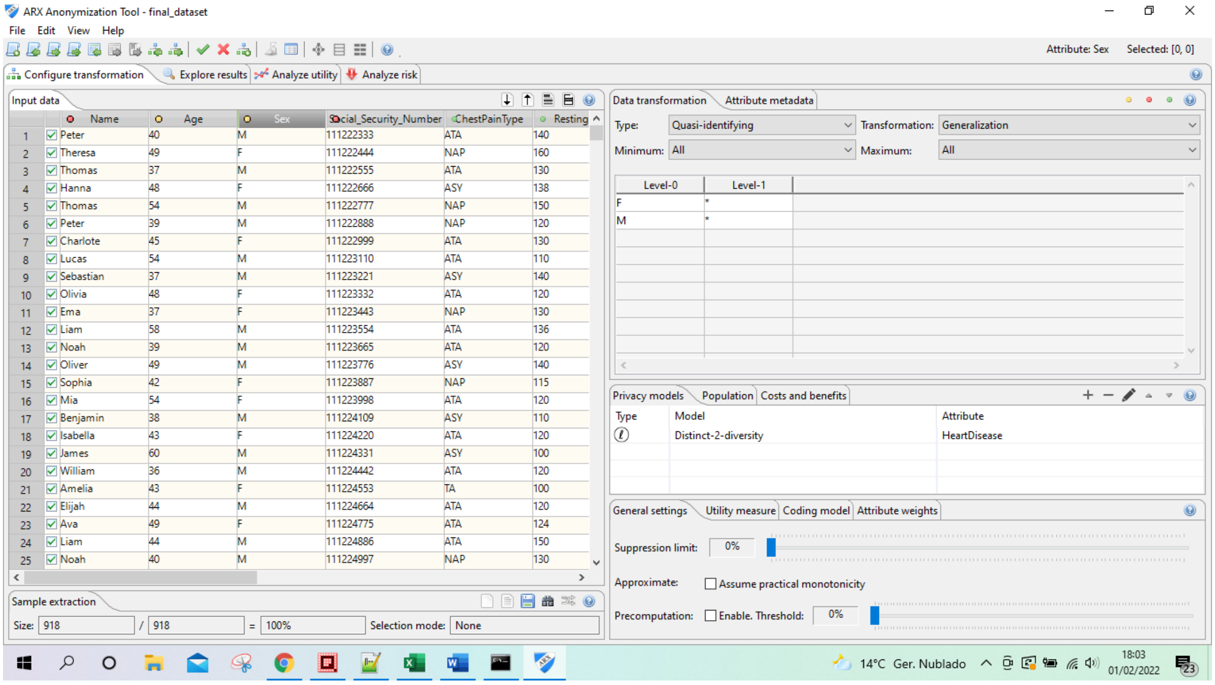
Precomputation: □ Enable. Threshold: 0%

14°C Ger. Nublado 18:02 01/02/2022

Figure 57 – Age Hierarchy Preview

Thus, the first level of the tree has the original values, the second level of the tree has the values in intervals of 10 and the third level of the tree has the value [20,80[. This is specified in Figure 57 with the hierarchy created.

For the Sex attribute, the type of the hierarchy is “Use masking”. The size set is 2 and the masking character chosen is *. The first level of the tree has the original values, and the second level of the tree has the masked values, as it is represented in Figure 58.



		Name	Age	Sex	Social_Security_Number	ChestPainType	Resting
1	☑	Peter	40	M	111222333	ATA	140
2	☑	Theresa	49	F	111222444	NAP	160
3	☑	Thomas	37	M	111222555	ATA	130
4	☑	Hanna	48	F	111222666	ASY	138
5	☑	Thomas	54	M	111222777	NAP	150
6	☑	Peter	39	M	111222888	NAP	120
7	☑	Charlote	45	F	111222999	ATA	130
8	☑	Lucas	54	M	111223110	ATA	110
9	☑	Sebastian	37	M	111223221	ASY	140
10	☑	Olivia	48	F	111223332	ATA	120
11	☑	Ema	37	F	111223443	NAP	130
12	☑	Liam	58	M	111223554	ATA	136
13	☑	Noah	39	M	111223665	ATA	120
14	☑	Oliver	49	M	111223776	ASY	140
15	☑	Sophia	42	F	111223887	NAP	115
16	☑	Mia	54	F	111223998	ATA	120
17	☑	Benjamin	38	M	111224109	ASY	110
18	☑	Isabella	43	F	111224220	ATA	120
19	☑	James	60	M	111224331	ASY	100
20	☑	William	36	M	111224442	ATA	120
21	☑	Amelia	43	F	111224553	TA	100
22	☑	Elijah	44	M	111224664	ATA	120
23	☑	Ava	49	F	111224775	ATA	124
24	☑	Liam	44	M	111224886	ATA	150
25	☑	Noah	40	M	111224997	NAP	130

Size: 918 / 918 = 100% Selection mode: None

Type: Quasi-identifying Transformation: Generalization
Minimum: All Maximum: All

Level-0	Level-1
F	*
M	*

Type	Model	Attribute
ℓ	Distinct-2-diversity	HeartDisease

Suppression limit: 0%

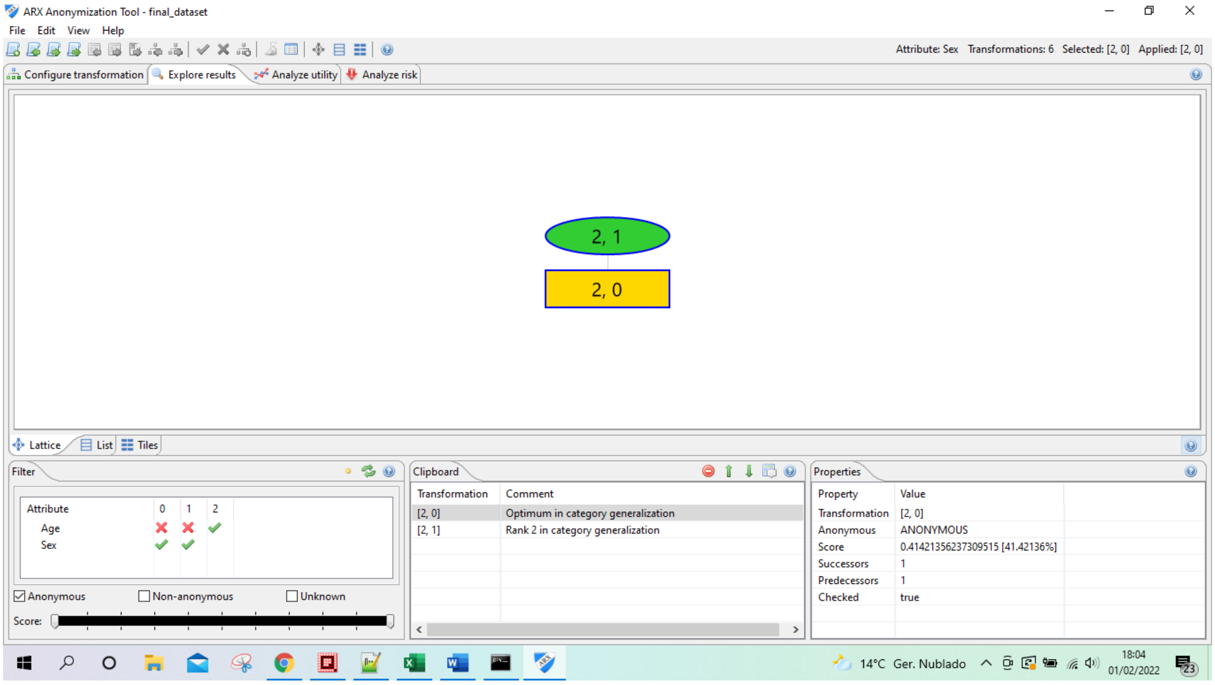
Approximate: □ Assume practical monotonicity

Precomputation: □ Enable. Threshold: 0%

Figure 58 - Sex Attribute Hierarchy Preview

Now is time to anonymize the dataset. Clicking in the button to anonymize and using the default values that appear in the pop-up window, the anonymized dataset is created. The results are just for quasi-identifier attributes that, in this case, are Age and Sex, as was mentioned before.

In the “Explore Results” tab, it is possible to verify the different anonymization levels created, and this is represented in Figure 59.



Attribute	0	1	2
Age	✗	✗	✓
Sex	✓	✓	

☑ Anonymous □ Non-anonymous □ Unknown

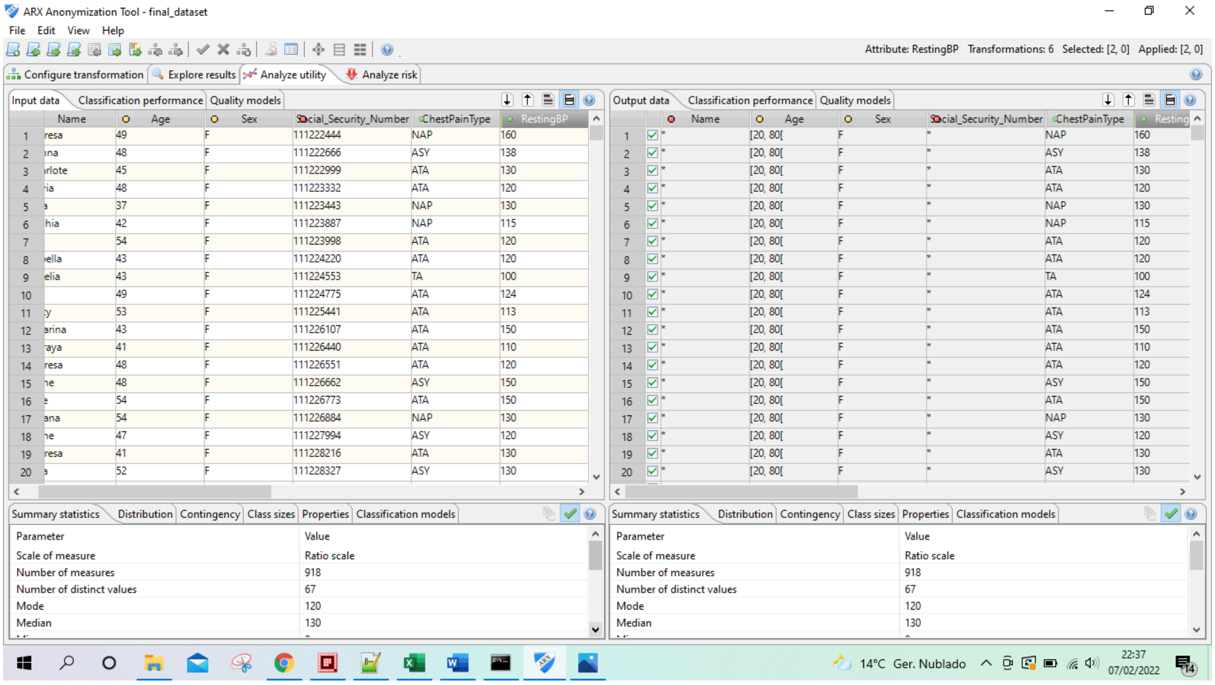
Score:

Transformation	Comment
[2, 0]	Optimum in category generalization
[2, 1]	Rank 2 in category generalization

Property	Value
Transformation	[2, 0]
Anonymous	ANONYMOUS
Score	0.41421356237309515 [41.42136%]
Successors	1
Predecessors	1
Checked	true

Figure 59 - Anonymization Graph Results

It appears two levels of generalization. According to the tool, the result with value [2, 0] is the one with “Optimum in category generalization”, so this is the one chosen to apply the transformation to the original dataset. Applying this transformation, it is possible to verify in the anonymized dataset that the Identifier attributes were modified to a special character (*). The Age attribute was modified for level-2 of the tree, the one with the values [20, 80[. As the value in the generalization level is 0 for Sex this means that this attribute did not have any modification compared to the original dataset, so the level-0 values of the hierarchy are the ones displayed in the anonymized dataset. A small part of the anonymized dataset is represented in Figure 60, and it is possible to export the complete anonymized dataset.



Attribute: RestingBP Transformations: 6 Selected: [2, 0] Applied: [2, 0]

	Name	Age	Sex	Social_Security_Number	ChestPainType	RestingBP
1	resa	49	F	111222444	NAP	160
2	ına	48	F	111222666	ASY	138
3	ırlote	45	F	111222999	ATA	130
4	ia	48	F	111223332	ATA	120
5	a	37	F	111223443	NAP	130
6	hia	42	F	111223887	NAP	115
7		54	F	111223998	ATA	120
8	ella	43	F	111224220	ATA	120
9	elia	43	F	111224553	TA	100
10		49	F	111224775	ATA	124
11	y	53	F	111225441	ATA	113
12	arina	43	F	111226107	ATA	150
13	aya	41	F	111226440	ATA	110
14	resa	48	F	111226551	ATA	120
15	ıe	48	F	111226662	ASY	150
16	e	54	F	111226773	ATA	150
17	ana	54	F	111226884	NAP	130
18	ıe	47	F	111227994	ASY	120
19	resa	41	F	111228216	ATA	130
20	a	52	F	111228327	ASY	130

		Name	Age	Sex	Social_Security_Number	ChestPainType	Resting
1	☑	*	[20, 80[	F	*	NAP	160
2	☑	*	[20, 80[	F	*	ASY	138
3	☑	*	[20, 80[	F	*	ATA	130
4	☑	*	[20, 80[	F	*	ATA	120
5	☑	*	[20, 80[	F	*	NAP	130
6	☑	*	[20, 80[	F	*	NAP	115
7	☑	*	[20, 80[	F	*	ATA	120
8	☑	*	[20, 80[	F	*	ATA	120
9	☑	*	[20, 80[	F	*	TA	100
10	☑	*	[20, 80[	F	*	ATA	124
11	☑	*	[20, 80[	F	*	ATA	113
12	☑	*	[20, 80[	F	*	ATA	150
13	☑	*	[20, 80[	F	*	ATA	110
14	☑	*	[20, 80[	F	*	ATA	120
15	☑	*	[20, 80[	F	*	ASY	150
16	☑	*	[20, 80[	F	*	ATA	150
17	☑	*	[20, 80[	F	*	NAP	130
18	☑	*	[20, 80[	F	*	ASY	120
19	☑	*	[20, 80[	F	*	ATA	130
20	☑	*	[20, 80[	F	*	ASY	130

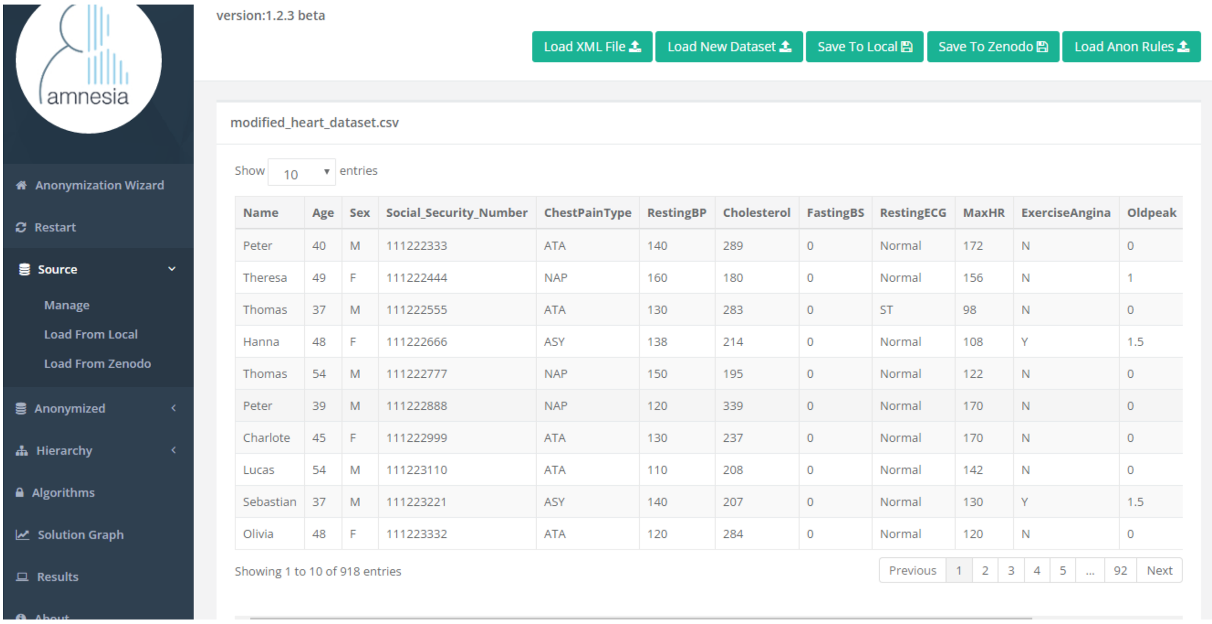
Parameter	Value
Scale of measure	Ratio scale
Number of measures	918
Number of distinct values	67
Mode	120
Median	130

Parameter	Value
Scale of measure	Ratio scale
Number of measures	918
Number of distinct values	67
Mode	120
Median	130

Figure 60 - Anonymized Dataset

9.2. Heart Failure Prediction Dataset in Amnesia

Now, using the original dataset chosen, it is going to be introduced in Amnesia Tool so that it is possible to anonymize the data. In Chapter 6 is explained the first steps that need to be performed to start the anonymization process that are the ones to upload the dataset. Figure 61 represents the dataset uploaded in the tool.



version:1.2.3 beta

modified_heart_dataset.csv

Name	Age	Sex	Social_Security_Number	ChestPainType	RestingBP	Cholesterol	FastingBS	RestingECG	MaxHR	ExerciseAngina	Oldpeak
Peter	40	M	111222333	ATA	140	289	0	Normal	172	N	0
Theresa	49	F	111222444	NAP	160	180	0	Normal	156	N	1
Thomas	37	M	111222555	ATA	130	283	0	ST	98	N	0
Hanna	48	F	111222666	ASY	138	214	0	Normal	108	Y	1.5
Thomas	54	M	111222777	NAP	150	195	0	Normal	122	N	0
Peter	39	M	111222888	NAP	120	339	0	Normal	170	N	0
Charlote	45	F	111222999	ATA	130	237	0	Normal	170	N	0
Lucas	54	M	111223110	ATA	110	208	0	Normal	142	N	0
Sebastian	37	M	111223221	ASY	140	207	0	Normal	130	Y	1.5
Olivia	48	F	111223332	ATA	120	284	0	Normal	120	N	0

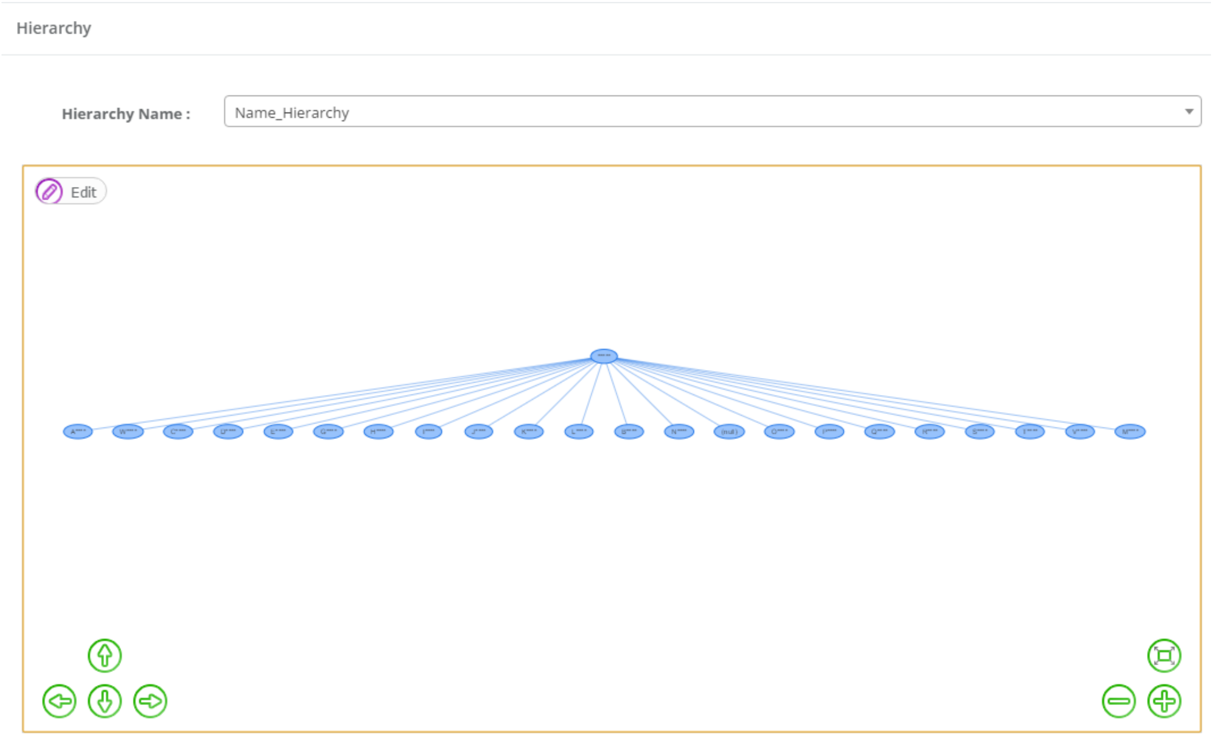
Showing 1 to 10 of 918 entries

Figure 61 - Dataset Uploaded in Amnesia

Now it is possible to proceed to the Hierarchy creation.

The hierarchies are going to be created for the quasi-identifier, identifier, and sensitive attributes. The identifier attributes are Name and Social_Security_Number, the sensitive attribute is Heart_Disease and the quasi-identifier attributes are Age and Sex.

For the Name attribute, the type selected is Masking Based with length 5 for not to have been a big value nor a small length for the variable. The attribute Name is Name_Hierarchy. The result of this hierarchy is represented in Figure 62. As it is possible to verify just the first letter of the name is shown up, and the other ones are replaced by a special character in the second level of the hierarchy tree. In the first level of the hierarchy tree, the name is going to be replaced by five special characters since this was the length specified in the hierarchy creation. This is advantage compared to the ARX Data Anonymization Tool because there are not so many levels in the hierarchy. The hierarchy also has the null value in the second hierarchy level in case there is some null value in the original dataset.

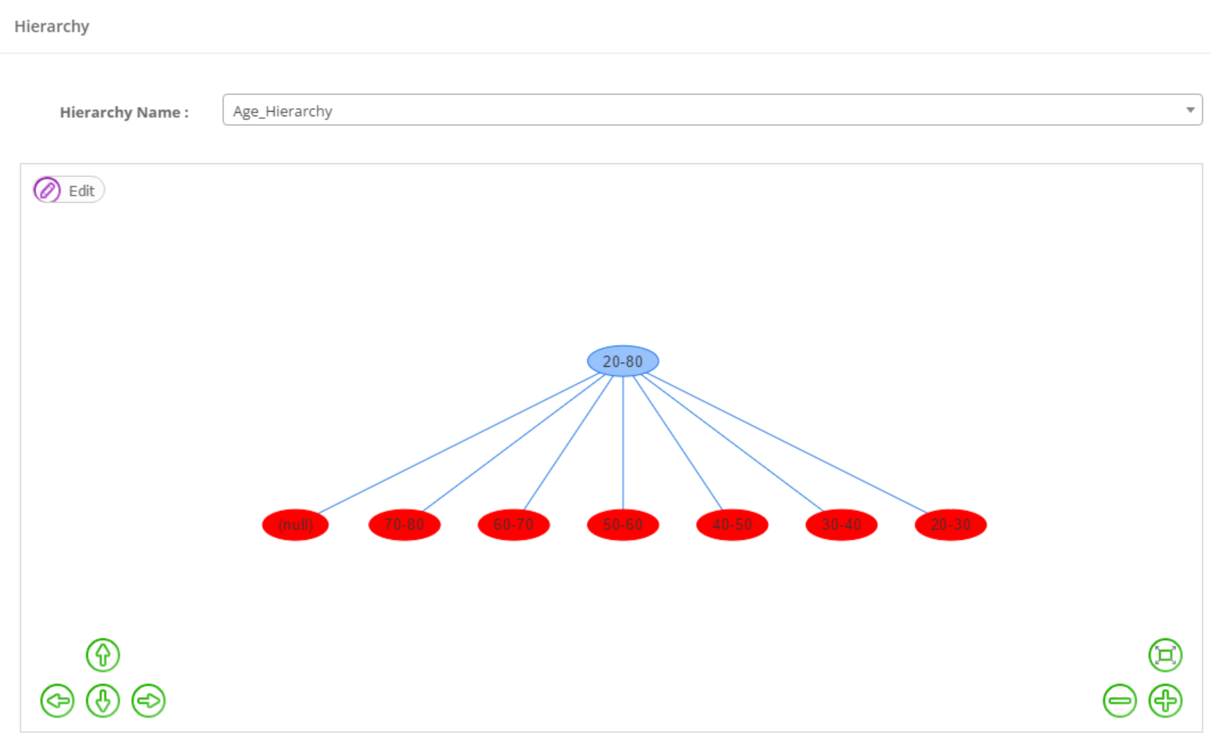


Hierarchy

Hierarchy Name : Name_Hierarchy

Figure 62 - Final Name Hierarchy Display

The next hierarchy to create is for the Age attribute. The type for this hierarchy is Range, with a step 10 and a Domain start end limits with the values 20-80, which means that the leaf nodes will have the multiple-10 values between 20 and 80. For the result of this hierarchy, the first level of the tree has a value of 20-80. The second level of the tree has the eight interval values and the null value because the dataset can have null values for this attribute. This result can be verified in Figure 63.

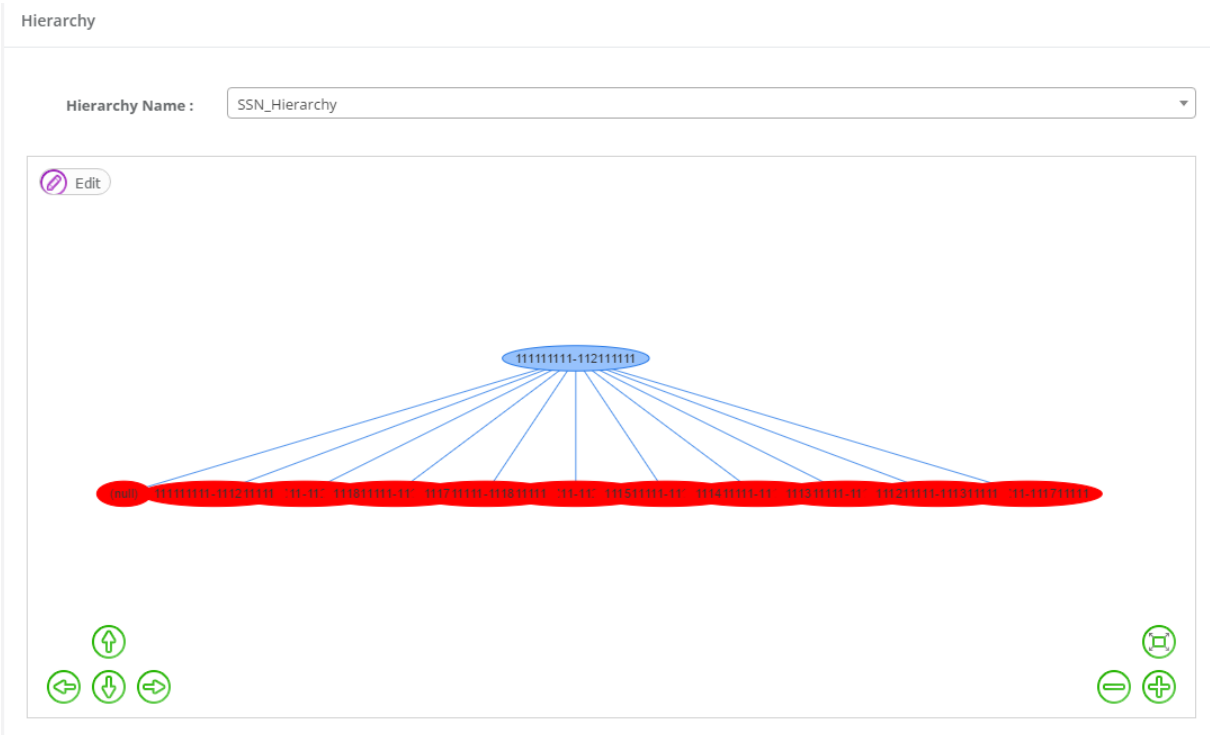


Hierarchy

Hierarchy Name : Age_Hierarchy

Figure 63 - Final Age Hierarchy Display

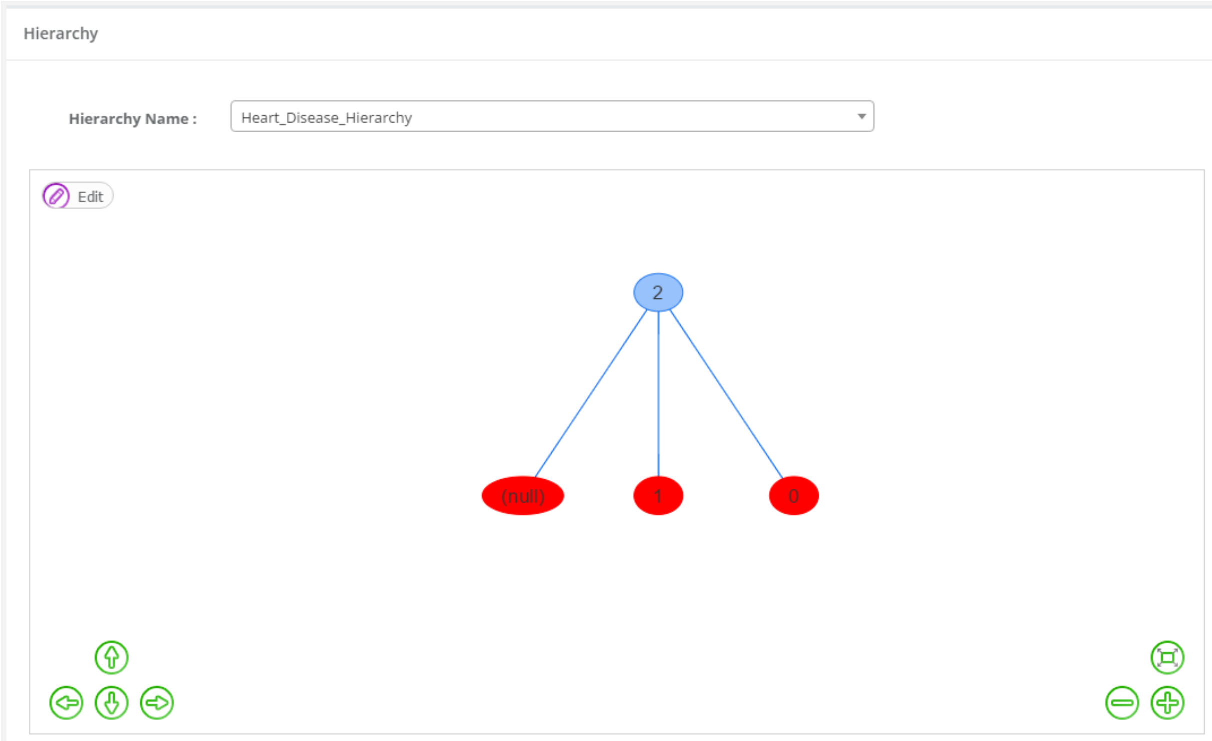
Social_Security_Number hierarchy is created as the Age hierarchy since both are attributes of type Integer. The difference between both is the step and the domain start end limits. In this case, the step is 100000 and the Domain start end limit is 111111111-112111111. The result for this hierarchy is similar to the previous one, just differing the values, as is verified in Figure 64.



Hierarchy Name : SSN_Hierarchy

Figure 64 - Final Social_Security_Number Hierarchy Display

Now the hierarchy is created for the HeartDisease attribute. This attribute has type Integer, so the hierarchy type is Distinct. The sorting type is numeric and the name for this attribute is HeartDisease_Hierarchy. The result for this hierarchy is a first node of the hierarchy with the value 2 and the leaf nodes with the values that exist for this attribute, and they are 0 and 1. It has also the null value in the leaf nodes because it can exist any null value in the original dataset. Figure 65 shows the diagram for this hierarchy.

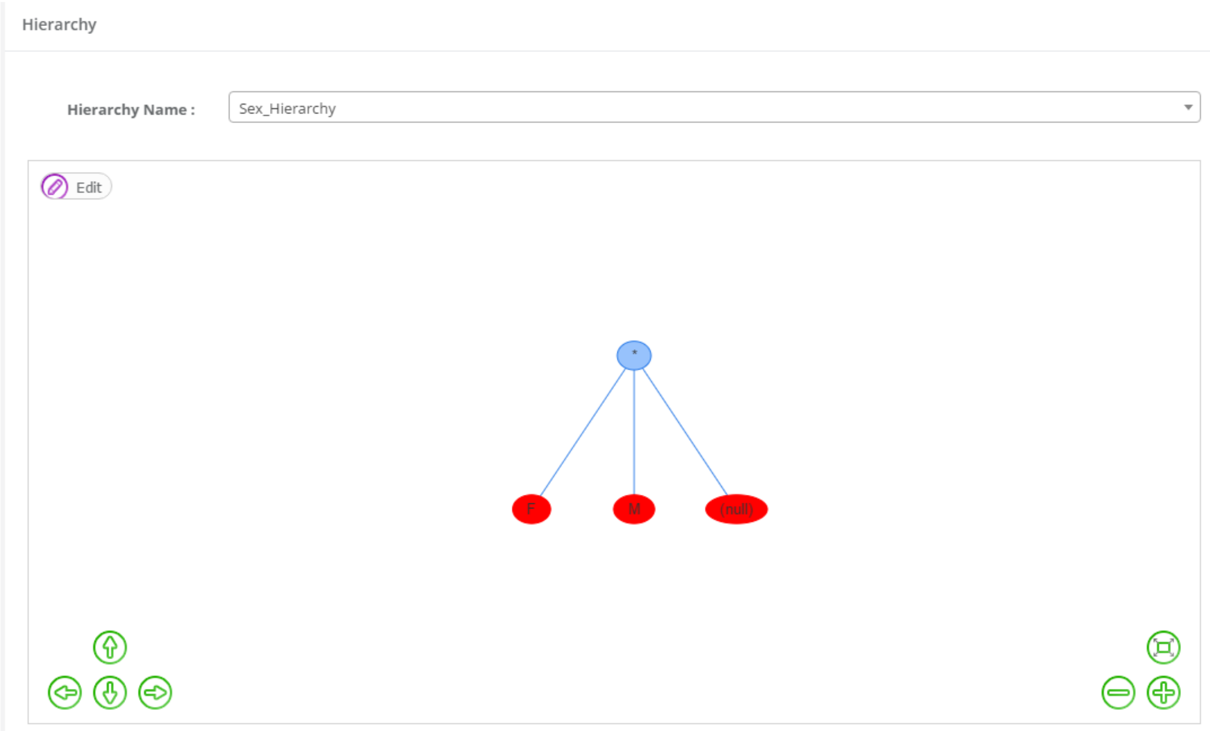


Hierarchy

Hierarchy Name : Heart_Disease_Hierarchy

Figure 65 - Final Heart_Disease Hierarchy Display

The last hierarchy to be created is for the Sex attribute. This attribute has type String, thus the hierarchy type chosen is Masking Based with length 1 and the name for this hierarchy is Sex_Hierarchy. The result, like it is possible to verify in Figure 66, has the first level node with just a special character (*) and the second level of the tree has the original values for this attribute (Male, Female) and the null value, in case it exists any null value in the original dataset.

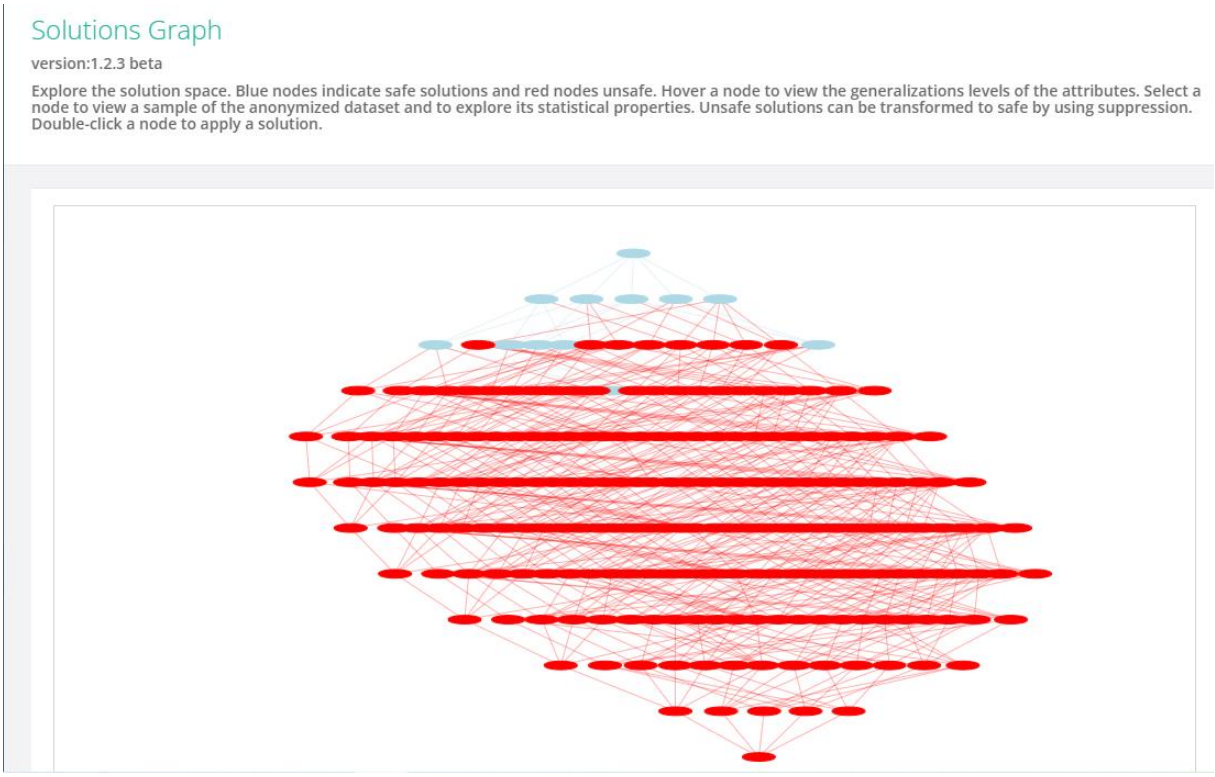


Hierarchy

Hierarchy Name : Sex_Hierarchy

Figure 66 - Final Sex Hierarchy Display

Now that all the hierarchies are created, the next step is to proceed to the algorithms page. On this page, the hierarchies created are linked to each attribute and the K value for the existing algorithm needs to be set. In this tool, the only choice for the algorithm is “Flash”, as was mentioned before. The K value chosen is 2 to avoid data loss, because if the k value is bigger, although the data has higher privacy the data utility is smaller. The non-sensitive attributes do not have a hierarchy to link. After executing the algorithm, a solution graph with all the possible combinations of anonymization levels is created. In the solution graph, all the blue nodes are safe solutions, and the red nodes are unsafe solutions. According to the information given by the tool, the unsafe solutions can be transformed into safe solutions throw suppression. Figure 67 shows the complete solution graph, where it is possible to verify that exists more unsafe than safe solutions.



Solutions Graph

version:1.2.3 beta

Explore the solution space. Blue nodes indicate safe solutions and red nodes unsafe. Hover a node to view the generalizations levels of the attributes. Select a node to view a sample of the anonymized dataset and to explore its statistical properties. Unsafe solutions can be transformed to safe by using suppression. Double-click a node to apply a solution.

Figure 67 - Amnesia Solution Graph

Observing closer the solution graph and verify the first levels of the graph with some of the safe solutions. Figure 68 shows the values [5, 2, 1, 2, 1] for the first node. These values are related to the generalization level for each attribute, which means that Name is generalized for level 5, Age is generalized for level 2, Sex is generalized for level 1, Social_Security_Number is generalized for level 2 and HeartDisease is generalized for level 1.

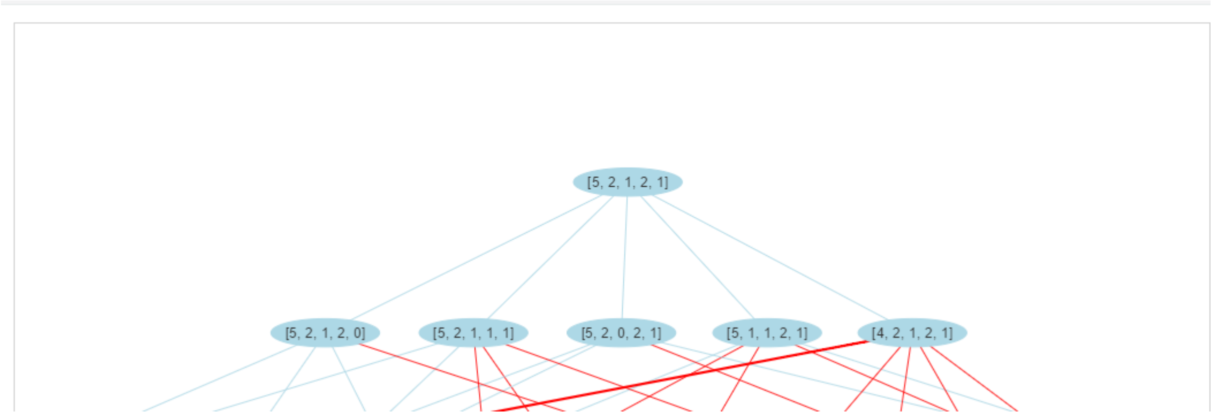
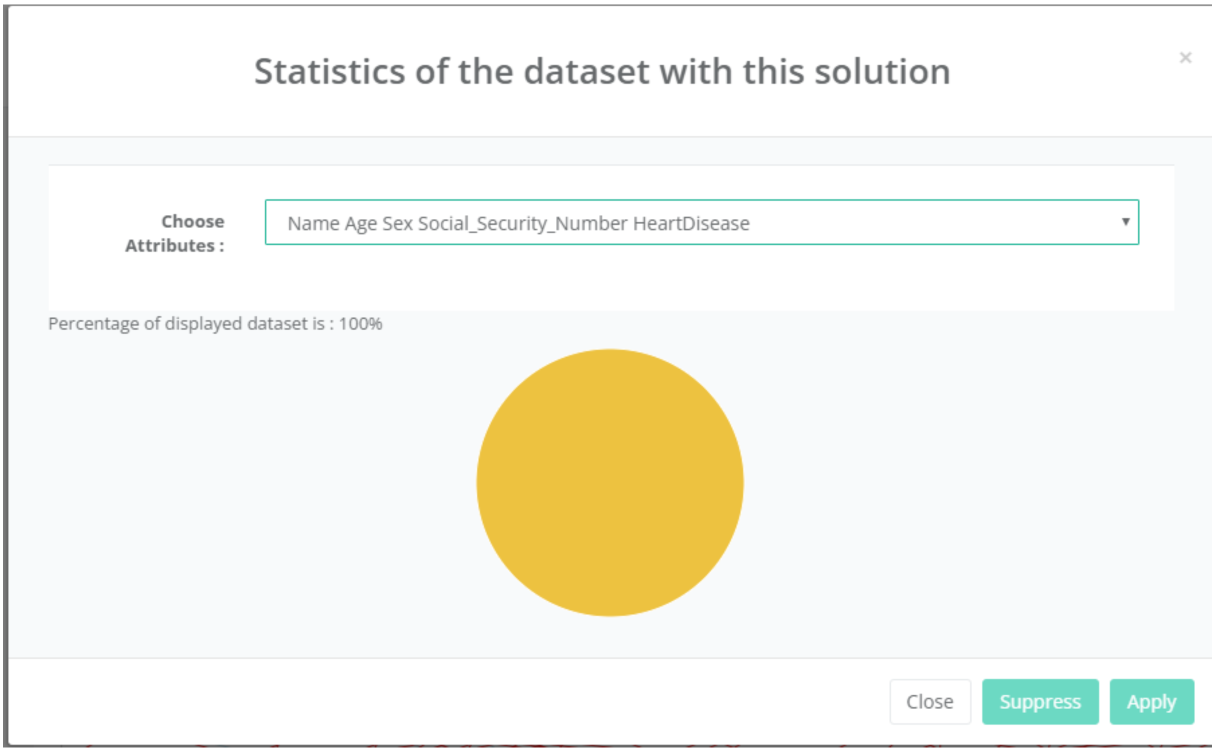


Figure 68 - Safe Solutions from Graph

Consulting the solutions graph for this solution, the one represented in Figure 69, it is possible to verify that all the dataset was anonymized because it does not appear any note with the percentage of suppression needed to make this a safe solution, which proves that this is a safe solution.



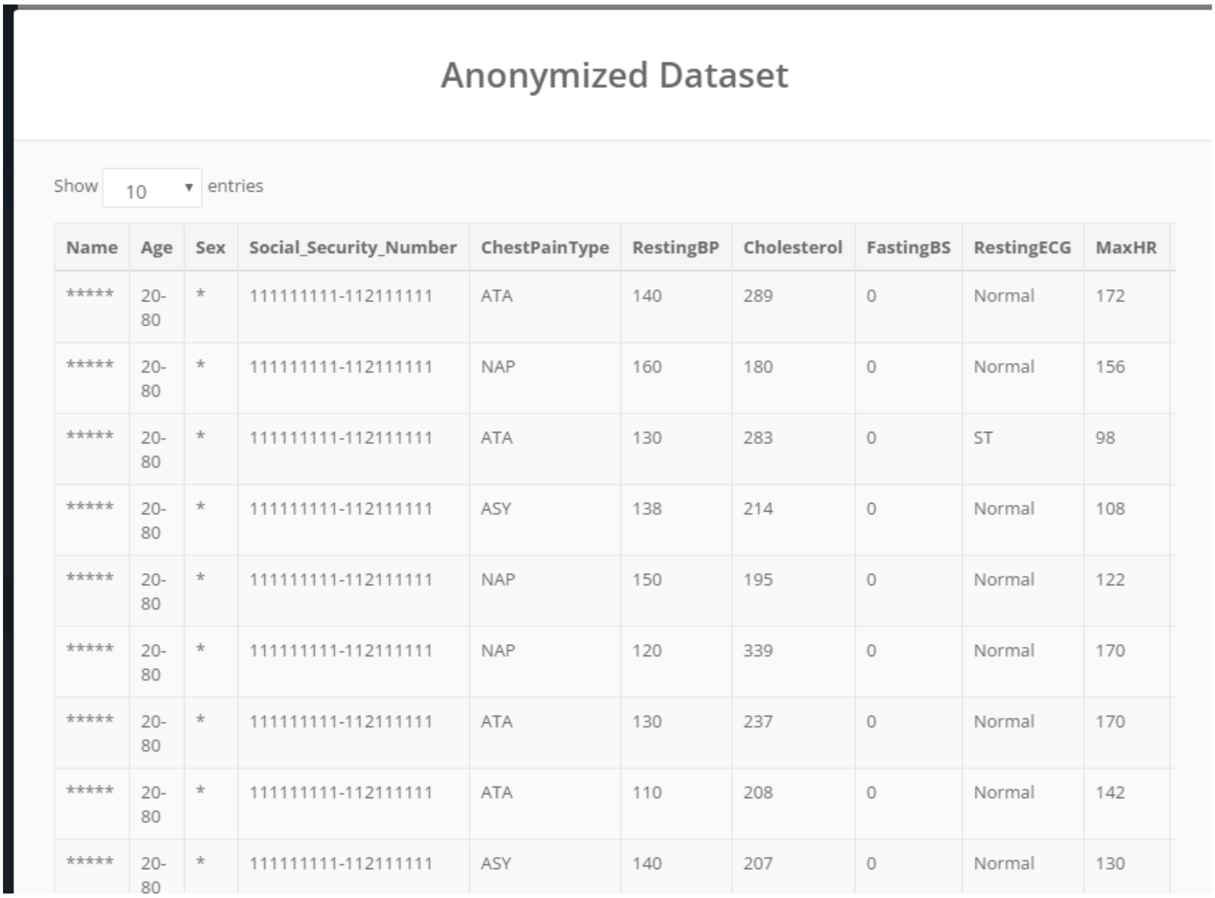
Statistics of the dataset with this solution

Choose Attributes : Name Age Sex Social_Security_Number HeartDisease

Percentage of displayed dataset is : 100%

Figure 69 - Statistics for Safe Solution

It is also possible to visualize the anonymized dataset and download it in a CSV format. In Figure 70 it is possible to verify that the data is not equal to the original one since some attributes were anonymized for the level represented in the solutions graph.



Anonymized Dataset

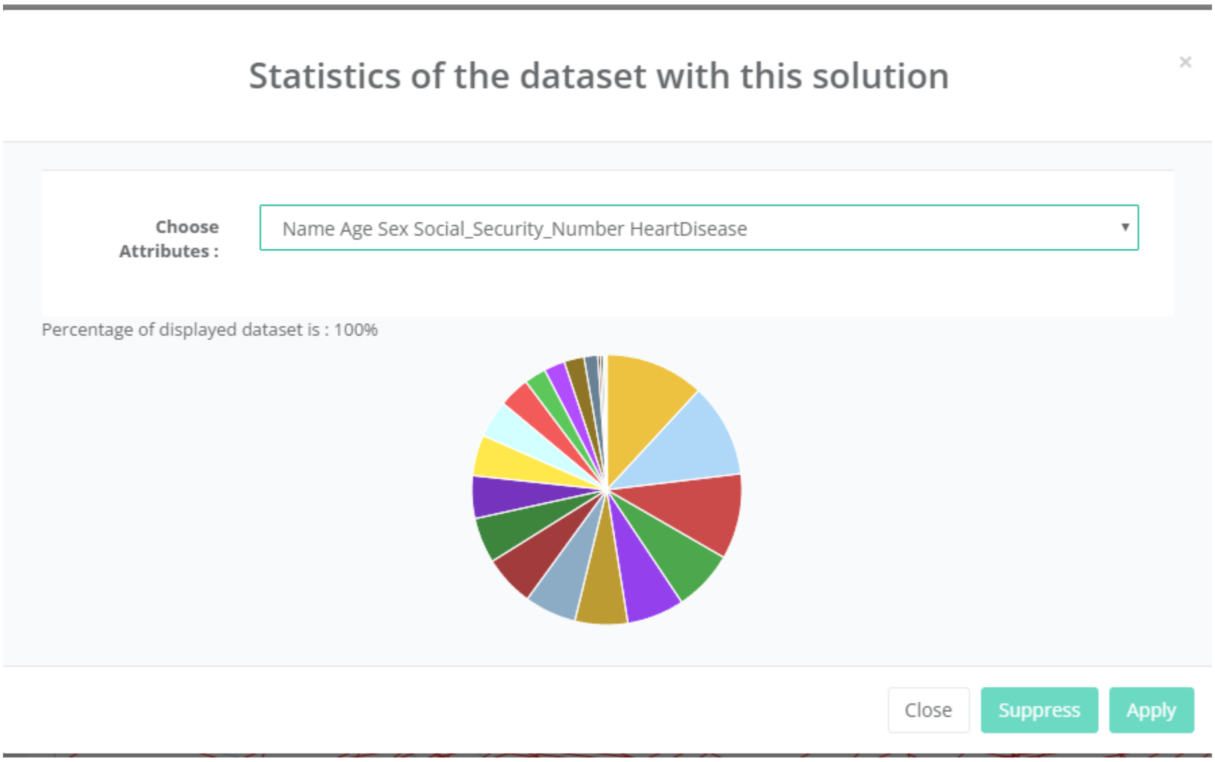
Show 10 ▾ entries

Name	Age	Sex	Social_Security_Number	ChestPainType	RestingBP	Cholesterol	FastingBS	RestingECG	MaxHR
*****	20-80	*	111111111-112111111	ATA	140	289	0	Normal	172
*****	20-80	*	111111111-112111111	NAP	160	180	0	Normal	156
*****	20-80	*	111111111-112111111	ATA	130	283	0	ST	98
*****	20-80	*	111111111-112111111	ASY	138	214	0	Normal	108
*****	20-80	*	111111111-112111111	NAP	150	195	0	Normal	122
*****	20-80	*	111111111-112111111	NAP	120	339	0	Normal	170
*****	20-80	*	111111111-112111111	ATA	130	237	0	Normal	170
*****	20-80	*	111111111-112111111	ATA	110	208	0	Normal	142
*****	20-80	*	111111111-112111111	ASY	140	207	0	Normal	130

Figure 70 - Safe Solution - Anonymized Dataset

Now, is time to choose another leaf from the solutions graph in Figure 68 to compare the different results. The node chosen is [4, 2, 1, 2, 1], which means that Name was generalized for level 4, Age was generalized for level 2, Sex was generalized for level 1, Social_Security_Number was generalized for level 2 and HeartDisease was generalized for level 1.

As this is also a safe solution, in the statistics windows it does not appear also any percentage of the dataset should be suppressed, as shown in Figure 71.



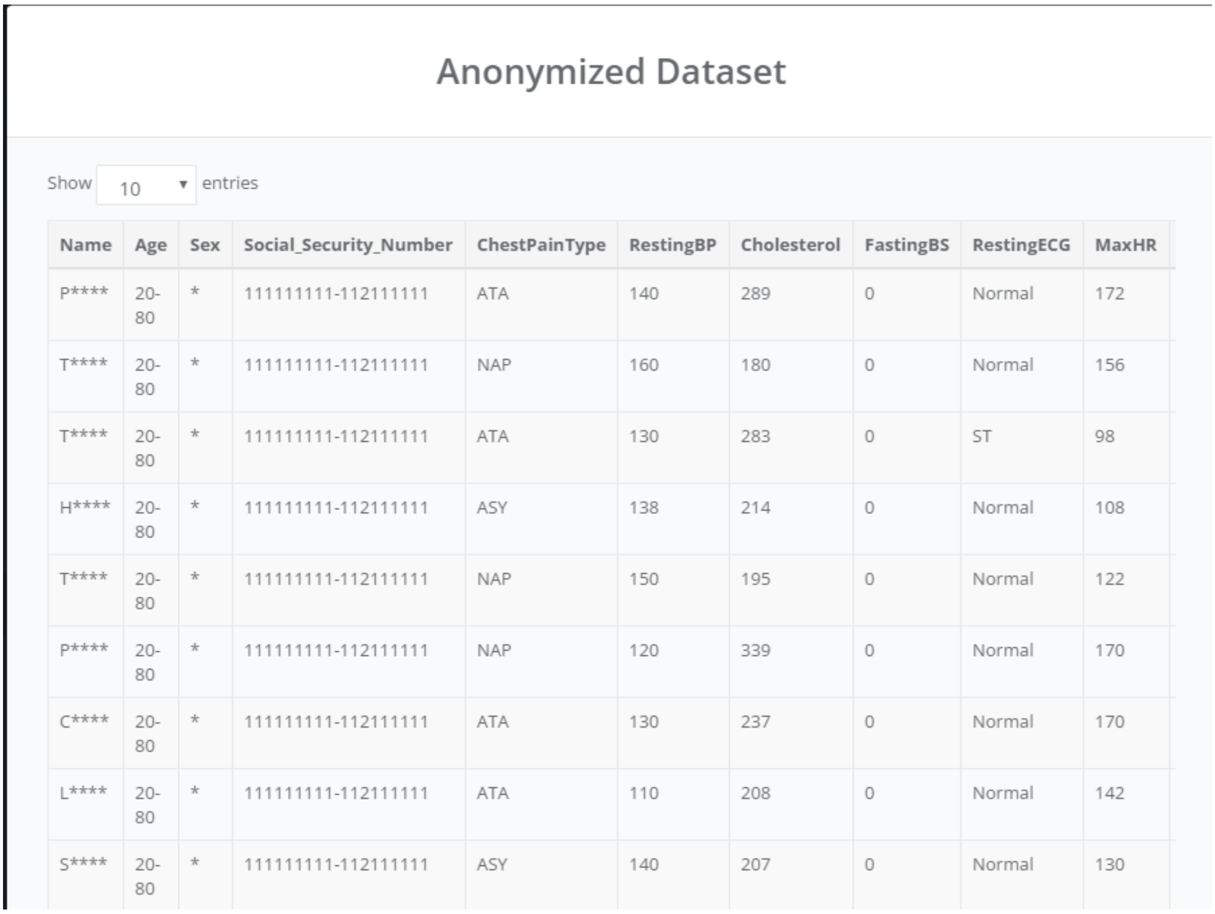
Statistics of the dataset with this solution

Choose Attributes : Name Age Sex Social_Security_Number HeartDisease

Percentage of displayed dataset is : 100%

Figure 71 - Statistics from Safe Solution

The anonymized results are represented in Figure 72, and it is also possible to download a CSV file with the anonymized dataset.



Anonymized Dataset

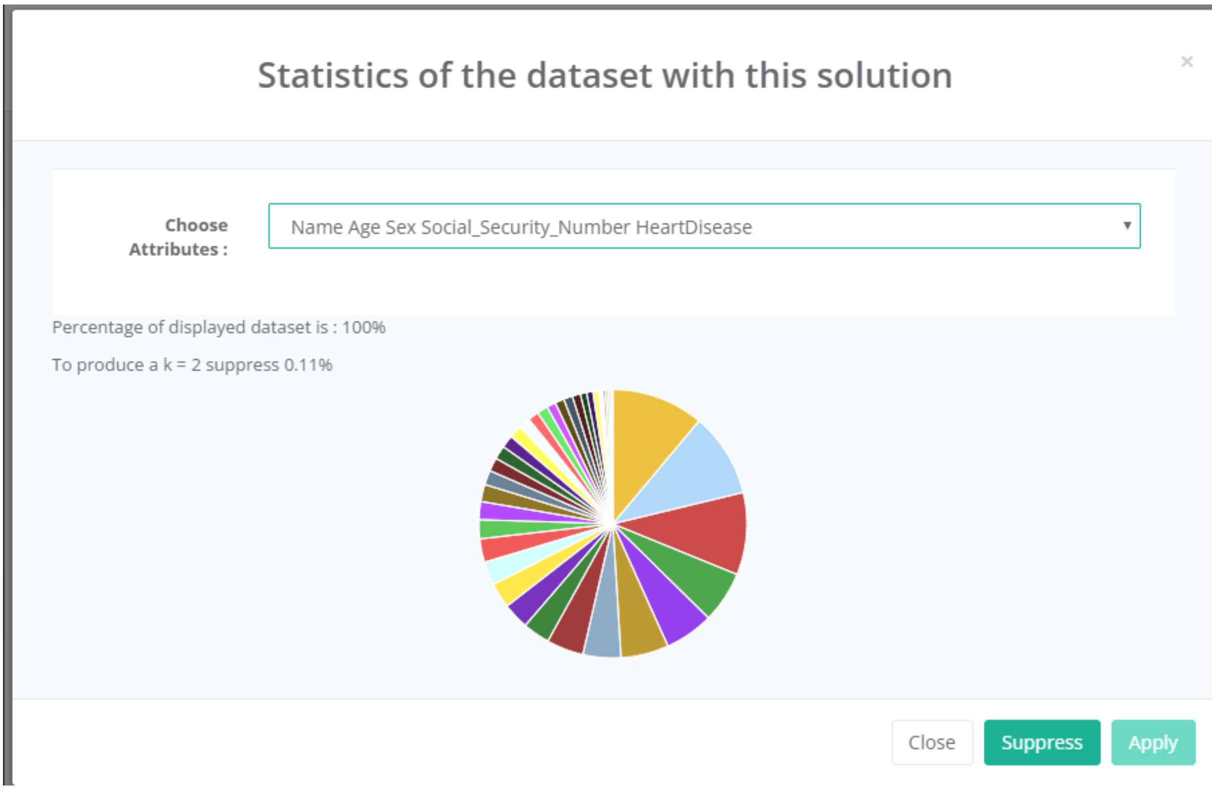
Show 10 entries

Name	Age	Sex	Social_Security_Number	ChestPainType	RestingBP	Cholesterol	FastingBS	RestingECG	MaxHR
P****	20-80	*	111111111-112111111	ATA	140	289	0	Normal	172
T****	20-80	*	111111111-112111111	NAP	160	180	0	Normal	156
T****	20-80	*	111111111-112111111	ATA	130	283	0	ST	98
H****	20-80	*	111111111-112111111	ASY	138	214	0	Normal	108
T****	20-80	*	111111111-112111111	NAP	150	195	0	Normal	122
P****	20-80	*	111111111-112111111	NAP	120	339	0	Normal	170
C****	20-80	*	111111111-112111111	ATA	130	237	0	Normal	170
L****	20-80	*	111111111-112111111	ATA	110	208	0	Normal	142
S****	20-80	*	111111111-112111111	ASY	140	207	0	Normal	130

Figure 72 - Safe Solution - Anonymized Dataset

Now, analysing an unsafe solution, the red node chosen from the solution graph in Figure 68 has the values [4, 2, 0, 1, 2], which means that Name was generalized for level 4, Age was generalized for level 2, Sex was generalized for level 0, Social_Security_Number was generalized for level 1 and HeartDisease was generalized for level 2.

This is an unsafe solution because the Sex attribute was not changed, and as this is a quasi-identifier attribute, it needs to be anonymized. In the statistics window now it is possible to see the percentage of suppression needed to make this a safe solution, which means that 0.11% of the records violate k-anonymity definition and should be removed to turn this into a safe solution, as is represented in Figure 73.



Statistics of the dataset with this solution

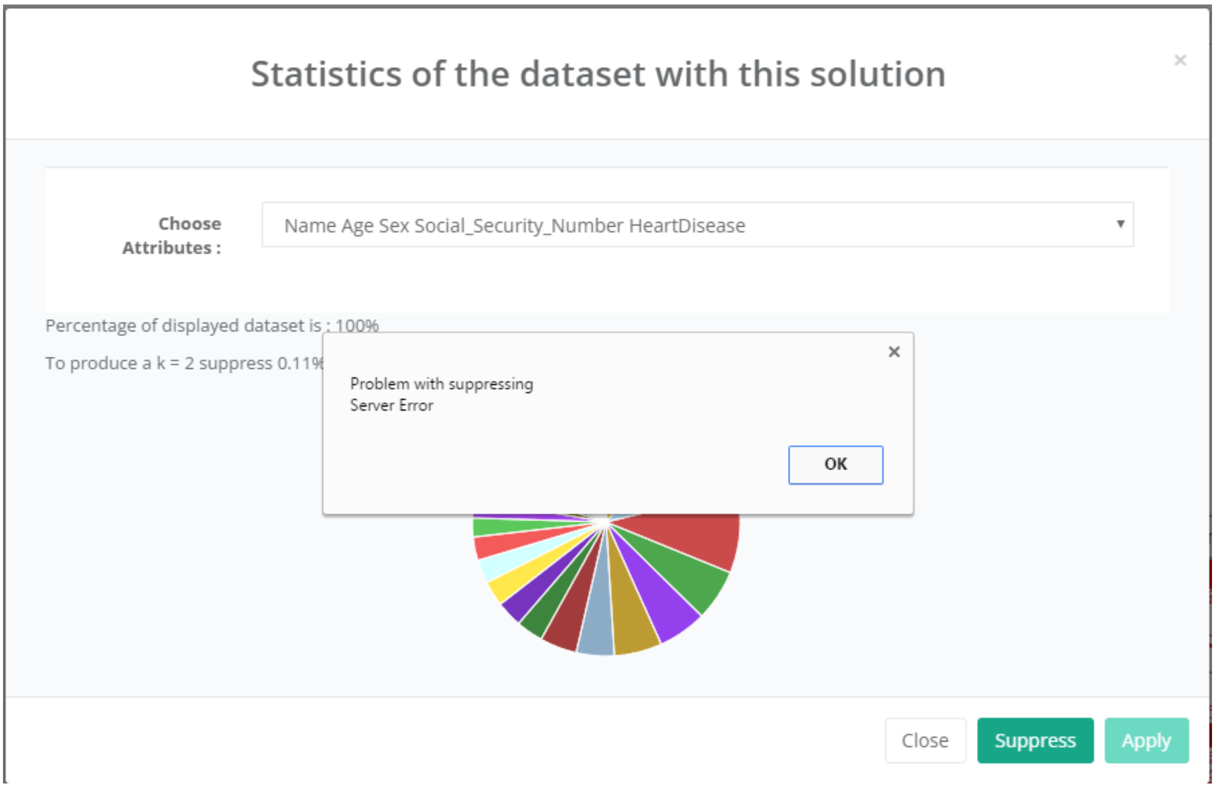
Choose Attributes : Name Age Sex Social_Security_Number HeartDisease

Percentage of displayed dataset is : 100%

To produce a k = 2 suppress 0.11%

Figure 73 - Statistics from Unsafe Solution

As it was mentioned in the solutions graph, the unsafe solutions could be suppressed to make this a safe solution, but if the Suppress button is used, the tool returns just an error, like the one in Figure 74, and does not turn this a safe solution, so this is a bug in the tool.



Statistics of the dataset with this solution

Choose Attributes : Name Age Sex Social_Security_Number HeartDisease

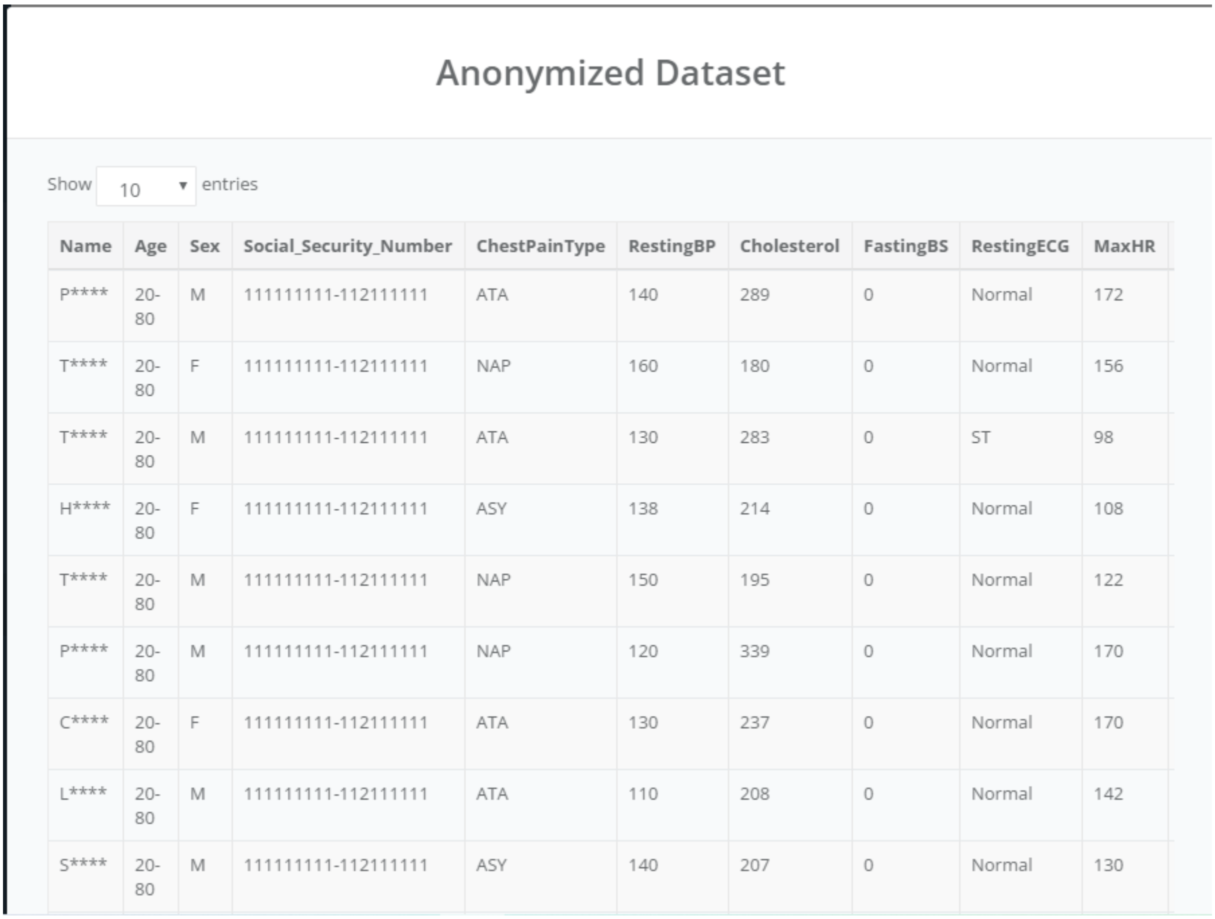
Percentage of displayed dataset is : 100%

To produce a k = 2 suppress 0.11%

Problem with suppressing
Server Error

Figure 74 - Supressing Error Statistics

It is also possible for this unsafe solution to visualize the anonymized dataset and to download it in a CSV format. The anonymized dataset preview is in Figure 75 and, as it is possible to confirm, the Sex attribute maintains its original values.



Anonymized Dataset

Name	Age	Sex	Social_Security_Number	ChestPainType	RestingBP	Cholesterol	FastingBS	RestingECG	MaxHR
P****	20-80	M	111111111-112111111	ATA	140	289	0	Normal	172
T****	20-80	F	111111111-112111111	NAP	160	180	0	Normal	156
T****	20-80	M	111111111-112111111	ATA	130	283	0	ST	98
H****	20-80	F	111111111-112111111	ASY	138	214	0	Normal	108
T****	20-80	M	111111111-112111111	NAP	150	195	0	Normal	122
P****	20-80	M	111111111-112111111	NAP	120	339	0	Normal	170
C****	20-80	F	111111111-112111111	ATA	130	237	0	Normal	170
L****	20-80	M	111111111-112111111	ATA	110	208	0	Normal	142
S****	20-80	M	111111111-112111111	ASY	140	207	0	Normal	130

Figure 75 - Unsafe Solution - Anonymized Dataset

Thus, this tool was quite simple to use but it is not possible to have a safe solution from an unsafe solution because the suppression button is not working as it should and the records that do violate the k-anonymity are not removed, so this tool has an error that should be corrected in the future.

9.3. Heart Failure Prediction Dataset in UTD Anonymization Toolbox

The tool now used to anonymize the dataset is UTD Anonymization Toolbox. It is needed to create some files before anonymizing the dataset. The first file created is the data file, where it is just needed to remove the name of the attributes at the beginning of the file and save it as a .data file. The next file to create is the header.txt file which is constituted by the attributes and the values possible for each attribute. After having these two files created, just the config file is missing. In this file, the method chosen was the Mondrian with k=2, because the k value used in the Amnesia was also 2. The output format selected is the genValsDist. In the ID element, the list of attributes set is the Name and the Social_Security_Number because these two attributes were the ones selected as identifier attributes in the dataset presentation. Age and Sex are quasi-identifier attributes and the HeartDisease attribute is a sensitive attribute. The Age attribute has the same values in each leaf as it was in Amnesia, i.e., is the nodes have the values [20:30), [30:40), [40:50), [50:60), [60:70) and [70:80), which make the first leaf interval [20:80). For the Sex attribute, the values are also the same as the ones used in Amnesia, so, F is generalized to 0 and M is generalized to 1. In the HeartDisease attribute, as this is a sensitive attribute, the values 0 and 1 should be anonymized to 2, as it is also possible to verify in the Amnesia attribute.

After preparing the config.xml file with all this information, the dataset can now be anonymized. Running the script anonymization.bat, the following information is a small part of what is written in the command line. The complete information is represented in Annex C.

```
Reading data takes 0sec.s
Processing EID = 1, [[[20:80)],[[0:1]]]
        Inserted 2 (left) and 3 (right)
Processing EID = 2, [[[20.0:54.0]],[[0:1]]]
        Inserted 4 (left) and 5 (right)
Processing EID = 3, [[(54.0:80.0)],[[0:1]]]
        Inserted 6 (left) and 7 (right)
Processing EID = 4, [[[20.0:47.0]],[[0:1]]]
        Inserted 8 (left) and 9 (right)
Processing EID = 5, [[(47.0:54.0]],[[0:1]]]
        Inserted 10 (left) and 11 (right)
Processing EID = 6, [[(54.0:60.0]],[[0:1]]]
        Inserted 12 (left) and 13 (right)
```

With this, also an output file is created with the result of the anonymization. This output file does not have the attribute names, as the DATA file. It is possible to verify in the output file that the identifier attributes were removed from the dataset, the quasi-identifier attributes were anonymized, and the sensitive attribute remains unchanged.

9.4. Conclusions Heart Failure Prediction Dataset

After anonymizing the same dataset in the different tools, it is possible to make some conclusions.

The first is about the usability of the tools. Since UTD Anonymization Toolbox does not have a graphical interface, this is the tool that has the worst usability, and the dataset is not simply anonymized because of all the files that need to be created. The output file retrieved by the tool is not also very friendly because it does not have the attribute names, so it is not easy to identify each one of the values. The output file is like the .data file created to anonymize the data.

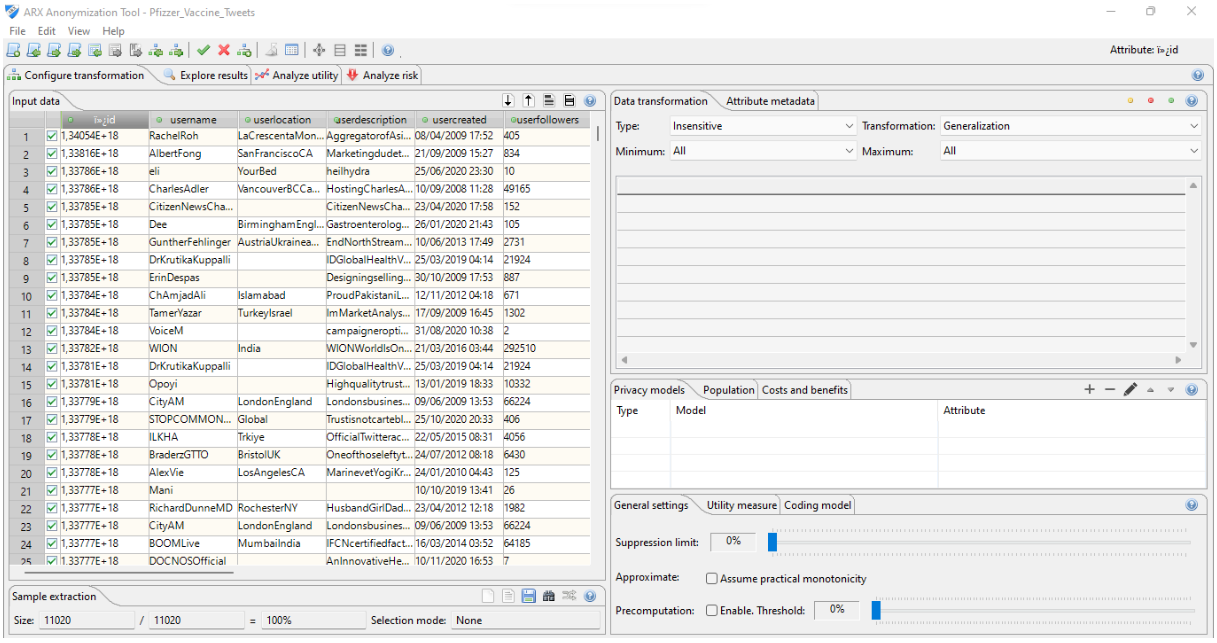
Amnesia is simpler to use, it does not have a good range of algorithms to choose to anonymize the data, the tool is always stopping, and it has some bugs, but as it does not have too much information happening it facilitates the usage. One thing that is also a plus for the tool is that every step has a button to guide the user to the next step that needs to be performed until the end of the anonymization process. The output file is easier to analyse since it is in CSV format. In Amnesia, all the attributes for which were created the hierarchies were anonymized according to the hierarchies created and the solution selected in the solution graph.

ARX Data Anonymization Tool is the most complete tool used in this comparison (and, as it is possible to verify in the OSSpal evaluation, it is the one with a better score). The anonymization process is simple, but it should allow the user to select the pretended algorithm to the sensitive attributes and not just the ones that the tool selects. It should help a little more the users in the anonymization process, guiding the user in every step of the anonymization process. The result is possible to export in the CSV format, which helps a lot in the analyse of the results. In the output, the sensitive attribute is not modified, the Age and Sex attributes are generalized according to the solution selected and the Name and Social_Security_Number attributes are modified to a special character * because they are identifying attributes.

So, the more complete results were retrieved by Amnesia, but the ARX Data Anonymization Tool is the most complete tool used.

9.5. Pfizer Vaccine Tweets Dataset in ARX Data Anonymization Tool

In this section we will demonstrate how the dataset “Pfizer Vaccine Tweets Dataset” was inserted into the ARX Data Anonymization Tool. After creating the project in the tool, the dataset can be uploaded. The format of the file chosen is CSV and the delimiter is a semi-comma. The verification of the data type of each attribute needs to be validated. A preview of the dataset is displayed, and the import is finished. It appears in the tool as shown in Figure 76.



Attribute: ï»¿id

		ï»¿id	username	userlocation	userdescription	usercreated	userfollowers
1	☑	1,34054E+18	RachelRoh	LaCrescentaMon...	AggregatorofAsi...	08/04/2009 17:52	405
2	☑	1,33816E+18	AlbertFong	SanFranciscoCA	Marketingdudet...	21/09/2009 15:27	834
3	☑	1,33786E+18	eli	YourBed	heilhydra	25/06/2020 23:30	10
4	☑	1,33786E+18	CharlesAdler	VancouverBCCa...	HostingCharlesA...	10/09/2008 11:28	49165
5	☑	1,33785E+18	CitizenNewsCha...		CitizenNewsCha...	23/04/2020 17:58	152
6	☑	1,33785E+18	Dee	BirminghamEngl...	Gastroenterolog...	26/01/2020 21:43	105
7	☑	1,33785E+18	GuntherFehlinger	AustriaUkrainea...	EndNorthStream...	10/06/2013 17:49	2731
8	☑	1,33785E+18	DrKrutikaKuppalli		IDGlobalHealthV...	25/03/2019 04:14	21924
9	☑	1,33785E+18	ErinDespas		Designingselling...	30/10/2009 17:53	887
10	☑	1,33784E+18	ChAmjadAli	Islamabad	ProudPakistaniL...	12/11/2012 04:18	671
11	☑	1,33784E+18	TamerYazar	TurkeyIsrael	ImMarketAnalys...	17/09/2009 16:45	1302
12	☑	1,33784E+18	VoiceM		campaigneropti...	31/08/2020 10:38	2
13	☑	1,33782E+18	WION	India	WIONWorldIsOn...	21/03/2016 03:44	292510
14	☑	1,33781E+18	DrKrutikaKuppalli		IDGlobalHealthV...	25/03/2019 04:14	21924
15	☑	1,33781E+18	Opoyi		Highqualitytrust...	13/01/2019 18:33	10332
16	☑	1,33779E+18	CityAM	LondonEngland	Londonsbusines...	09/06/2009 13:53	66224
17	☑	1,33779E+18	STOPCOMMON...	Global	Trustisnotcartebl...	25/10/2020 20:33	406
18	☑	1,33778E+18	ILKHA	Trkiye	OfficialTwitterac...	22/05/2015 08:31	4056
19	☑	1,33778E+18	BraderzGTTO	BristolUK	Oneofthoseleftyt...	24/07/2012 08:18	6430
20	☑	1,33778E+18	AlexVie	LosAngelesCA	MarinevetYogiKr...	24/01/2010 04:43	125
21	☑	1,33777E+18	Mani			10/10/2019 13:41	26
22	☑	1,33777E+18	RichardDunneMD	RochesterNY	HusbandGirlDad...	23/04/2012 12:18	1982
23	☑	1,33777E+18	CityAM	LondonEngland	Londonsbusines...	09/06/2009 13:53	66224
24	☑	1,33777E+18	BOOMLive	MumbaiIndia	IFCNcertifiedfact...	16/03/2014 03:52	64185
25	☑	1,33777E+18	DOCNOSOfficial		AnInnovativeHe...	10/11/2020 16:53	7

Size: 11020 / 11020 = 100% Selection mode: None

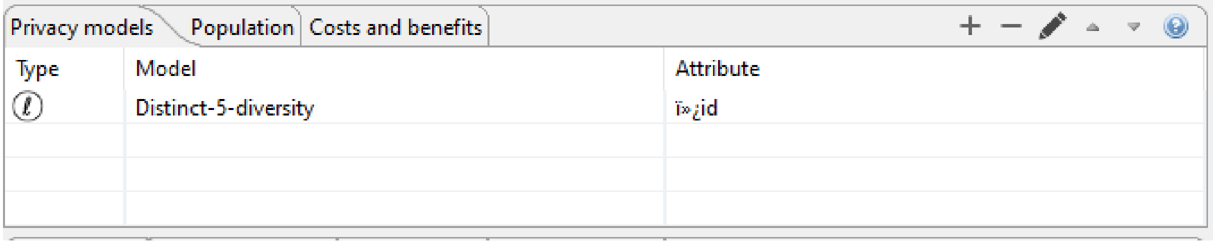
Type: Insensitive Transformation: Generalization
Minimum: All Maximum: All

Type	Model	Attribute

Suppression limit: 0%
Approximate: □ Assume practical monotonicity
Precomputation: □ Enable. Threshold: 0%

Figure 76 - ARX Dataset Display

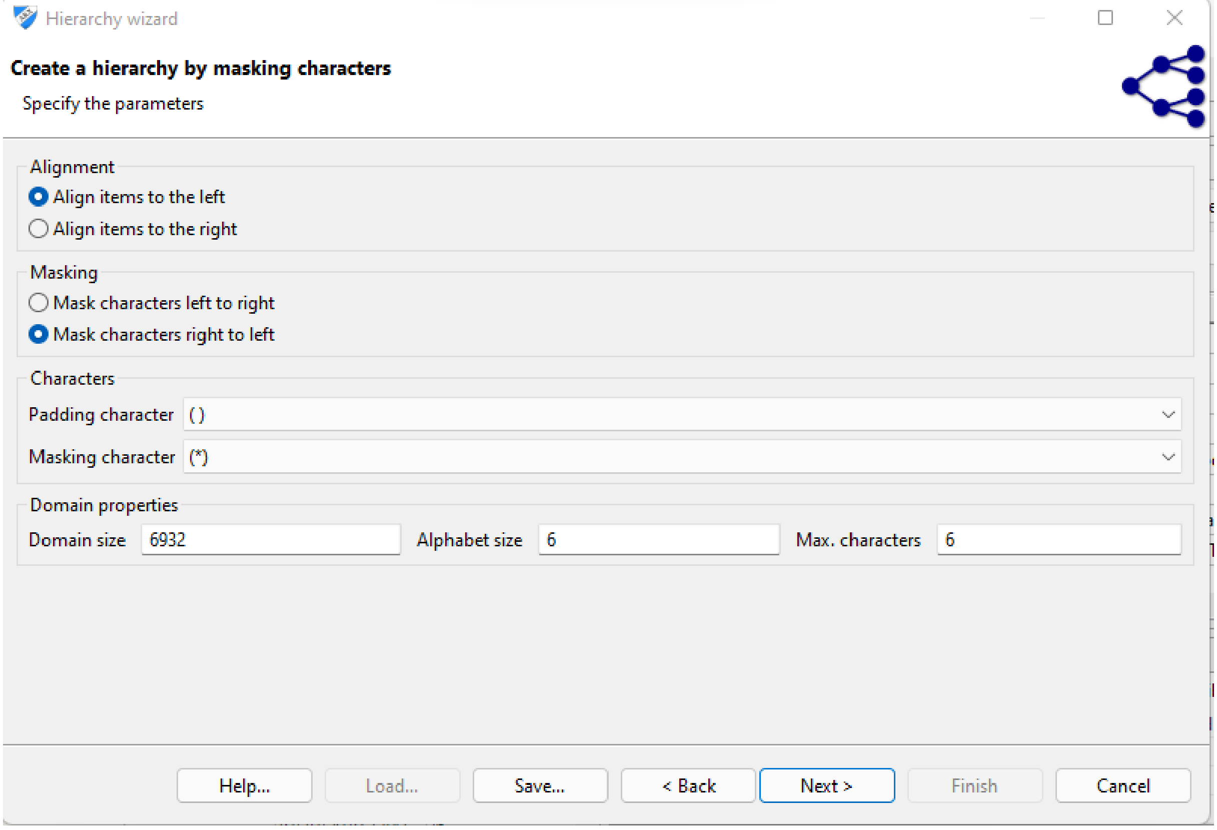
After uploading the dataset, the type of each attribute needs to be set according to Table 24. The transformation chosen is the generalization to be compliant with the generalization technique of the Amnesia tool and the target variables are set for the attributes id, username, userlocation and userdescription. The privacy mode for the sensitive attribute needs to set among the privacy models allowed by the tool for this attribute which are: l-Diversity, t-Closeness, δ-Disclosure privacy, and β-Likeness. l-Diversity is the one chosen, and the l value is 5, as the k value chosen in the Amnesia. The privacy model is shown in Figure 77.



Type	Model	Attribute
ℓ	Distinct-5-diversity	ï»¿id

Figure 77 - ARX - Privacy Model

Before continuing one needs to create the hierarchies for the identifier and quasi-identifier attributes. The first hierarchy that is going to be created is for the username hierarchy. The type of the hierarchy is masking, and the masking character is *, as it is shown in Figure 78.



Hierarchy wizard

Create a hierarchy by masking characters

Specify the parameters

Alignment

● Align items to the left

○ Align items to the right

Masking

○ Mask characters left to right

● Mask characters right to left

Characters

Padding character ( )

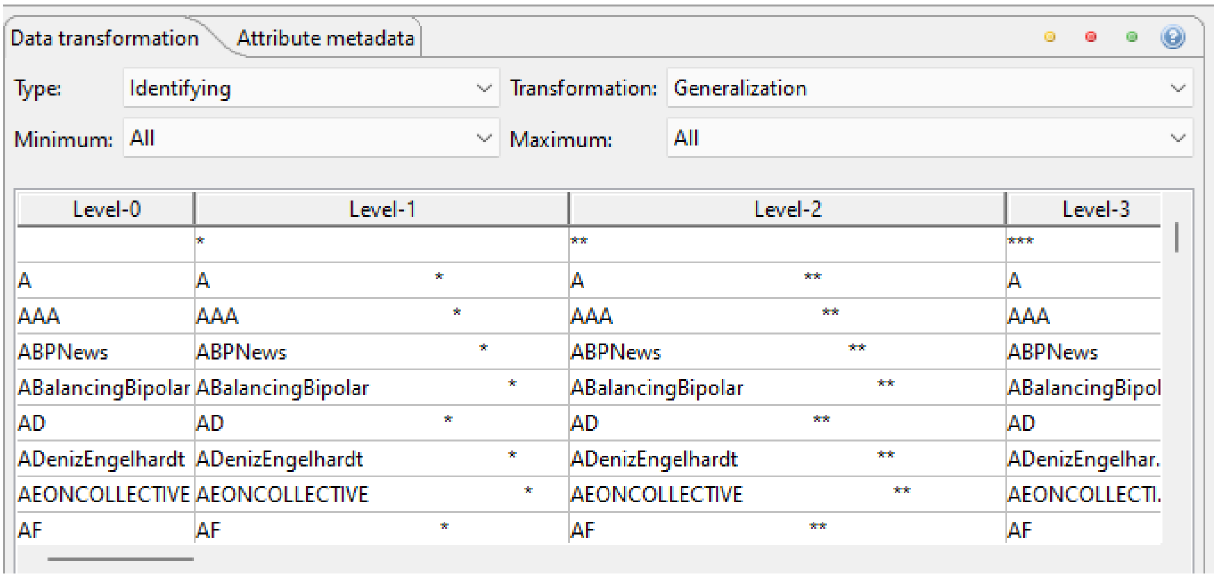
Masking character (*)

Domain properties

Domain size 6932 Alphabet size 6 Max. characters 6

Figure 78 - ARX - Create username Hierarchy

For this hierarchy was chosen the value 6 as for Max characters because it was the same size that was given in Amnesia. The result of this anonymization hierarchy is, in Level-0, the original value of the attribute and as the level rises, the value begins to be modified by special characters, starting in the right side to the left side of the value. This can be seen in Figure 79.

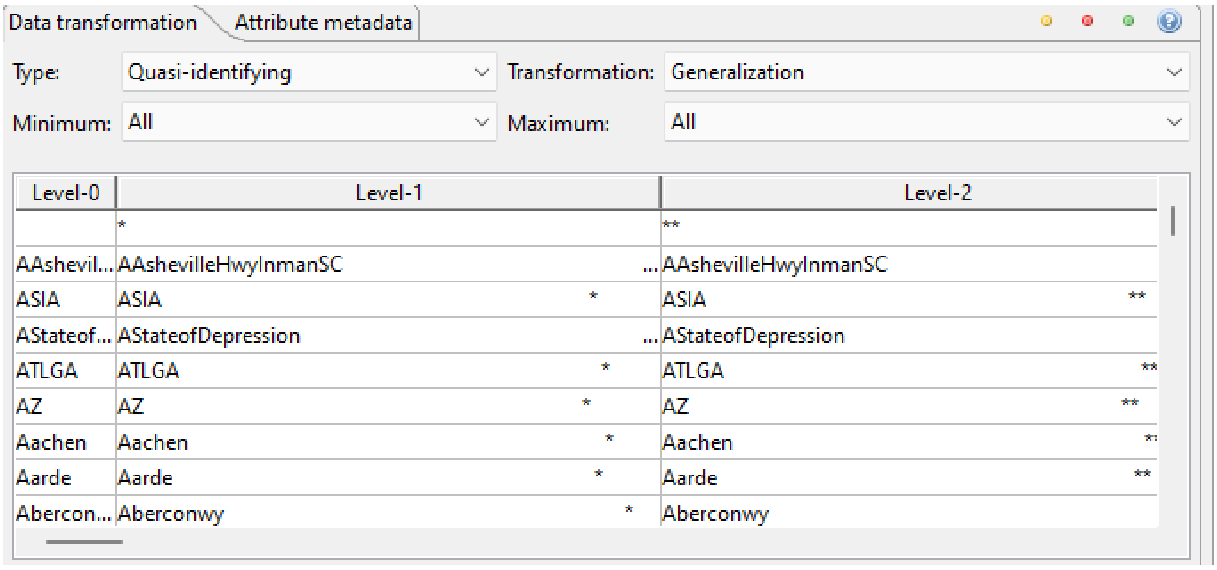


Type: Identifying
Transformation: Generalization
Minimum: All
Maximum: All

Level-0	Level-1	Level-2	Level-3
	*	**	***
A	A *	A **	A
AAA	AAA *	AAA **	AAA
ABPNews	ABPNews *	ABPNews **	ABPNews
ABalancingBipolar	ABalancingBipolar *	ABalancingBipolar **	ABalancingBipol
AD	AD *	AD **	AD
ADenizEngelhardt	ADenizEngelhardt *	ADenizEngelhardt **	ADenizEngelhar.
AEONCOLLECTIVE	AEONCOLLECTIVE *	AEONCOLLECTIVE **	AEONCOLLECTI.
AF	AF *	AF **	AF

Figure 79 - username Hierarchy

The hierarchy for userlocation attribute was created in the same way that the username was, so the result is going to be similar, as shown in Figure 80.

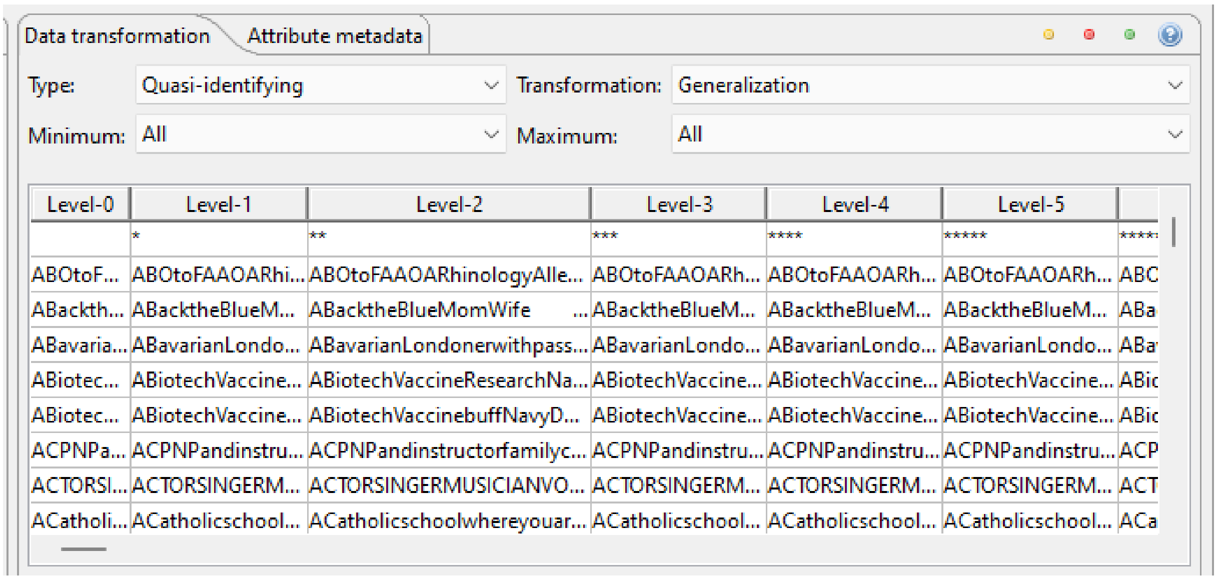


Type: Quasi-identifying
Transformation: Generalization
Minimum: All
Maximum: All

Level-0	Level-1	Level-2
	*	**
AAshevil...	AAshevilleHwyInmanSC ...	AAshevilleHwyInmanSC
ASIA	ASIA *	ASIA **
AStateof...	AStateofDepression ...	AStateofDepression
ATLGA	ATLGA *	ATLGA **
AZ	AZ *	AZ **
Aachen	Aachen *	Aachen **
Aarde	Aarde *	Aarde **
Abercon...	Aberconwy *	Aberconwy

Figure 80 - userlocation Hierarchy

The last hierarchy that is going to be created is for userdescription attribute, and it is going to be created also as the username attribute was. The result can be seen in Figure 81.



Data transformation | Attribute metadata

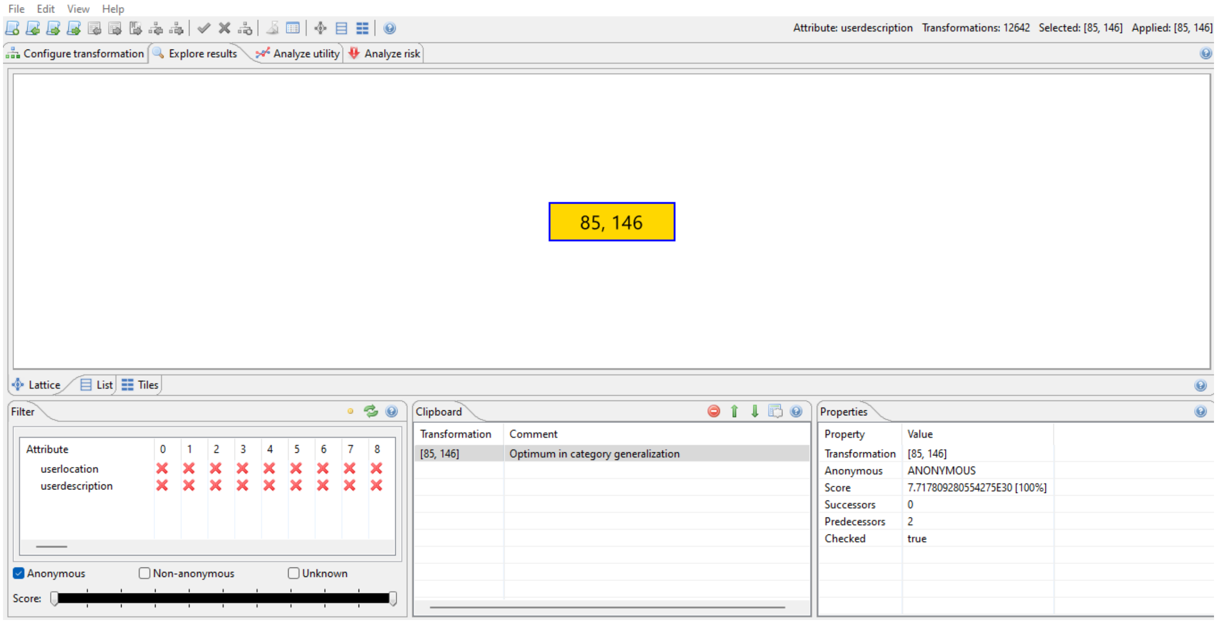
Type: Quasi-identifying — Transformation: Generalization

Minimum: All — Maximum: All

Level-0	Level-1	Level-2	Level-3	Level-4	Level-5	
	*	**	***	****	*****	****:
ABOtoF...	ABOtoFAAOARhi...	ABOtoFAAOARhinologyAlle...	ABOtoFAAOARh...	ABOtoFAAOARh...	ABOtoFAAOARh...	ABC
ABackth...	ABacktheBlueM...	ABacktheBlueMomWife ...	ABacktheBlueM...	ABacktheBlueM...	ABacktheBlueM...	ABa
ABavaria...	ABavarianLondo...	ABavarianLondonerwithpass...	ABavarianLondo...	ABavarianLondo...	ABavarianLondo...	ABa
ABiotec...	ABiotechVaccine...	ABiotechVaccineResearchNa...	ABiotechVaccine...	ABiotechVaccine...	ABiotechVaccine...	ABio
ABiotec...	ABiotechVaccine...	ABiotechVaccinebuffNavyD...	ABiotechVaccine...	ABiotechVaccine...	ABiotechVaccine...	ABio
ACPNPa...	ACPNPandinstru...	ACPNPandinstructorfamilyc...	ACPNPandinstru...	ACPNPandinstru...	ACPNPandinstru...	ACP
ACTORSI...	ACTORSINGERM...	ACTORSINGERMUSICIANVO...	ACTORSINGERM...	ACTORSINGERM...	ACTORSINGERM...	ACT
ACatholi...	ACatholicschool...	ACatholicschoolwhereyouar...	ACatholicschool...	ACatholicschool...	ACatholicschool...	ACa

Figure 81 - userdescription Hierarchy

Now that all the hierarchies and privacy models are created, it is possible to proceed to the anonymization of the dataset. ARX just returns one value in the solution graph, as shown in Figure 82.



Attribute: userdescription Transformations: 12642 Selected: [85, 146] Applied: [85, 146]

Attribute	0	1	2	3	4	5	6	7	8
userlocation	✗	✗	✗	✗	✗	✗	✗	✗	✗
userdescription	✗	✗	✗	✗	✗	✗	✗	✗	✗

☑ Anonymous ☐ Non-anonymous ☐ Unknown

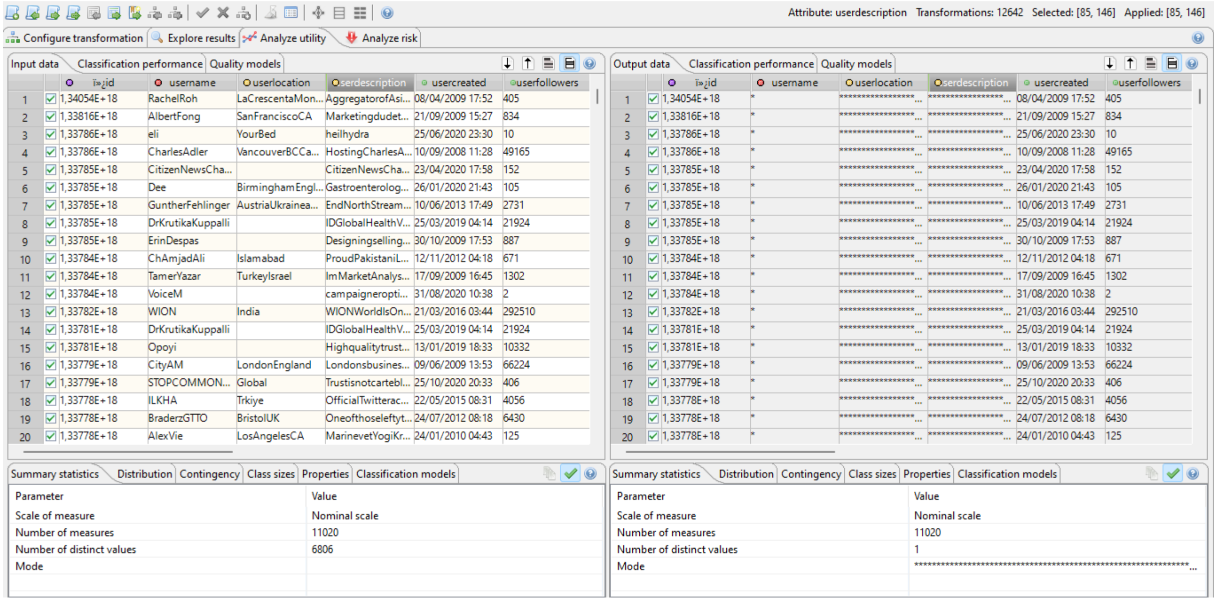
Score:

Transformation	Comment
[85, 146]	Optimum in category generalization

Property	Value
Transformation	[85, 146]
Anonymous	ANONYMOUS
Score	7.717809280554275E30 [100%]
Successors	0
Predecessors	2
Checked	true

Figure 82 - Solution Graph - One solution

Only the quasi-identifier attributes are represented in this graph, and they were 100% anonymized. This information is in the tool when the mouse is over the yellow rectangle. Applying this transformation, in the Analyse utility tab is possible to see the difference between the original and the anonymized dataset, as shown in Figure 83.



Attribute: userdescription Transformations: 12642 Selected: [85, 146] Applied: [85, 146]

Input data

		ï»¿id	username	userlocation	serdescription	usercreated	userfollowers
1	☑	1,34054E+18	RachelRoh	LaCrescentaMon...	AggregatorofAsi...	08/04/2009 17:52	405
2	☑	1,33816E+18	AlbertFong	SanFranciscoCA	Marketingdudet...	21/09/2009 15:27	834
3	☑	1,33786E+18	eli	YourBed	heilhydra	25/06/2020 23:30	10
4	☑	1,33786E+18	CharlesAdler	VancouverBCCa...	HostingCharlesA...	10/09/2008 11:28	49165
5	☑	1,33785E+18	CitizenNewsCha...		CitizenNewsCha...	23/04/2020 17:58	152
6	☑	1,33785E+18	Dee	BirminghamEngl...	Gastroenterolog...	26/01/2020 21:43	105
7	☑	1,33785E+18	GuntherFehlinger	AustriaUkrainea...	EndNorthStream...	10/06/2013 17:49	2731
8	☑	1,33785E+18	DrKrutikaKuppalli		IDGlobalHealthV...	25/03/2019 04:14	21924
9	☑	1,33785E+18	ErinDespas		Designingselling...	30/10/2009 17:53	887
10	☑	1,33784E+18	ChAmjadAli	Islamabad	ProudPakistaniL...	12/11/2012 04:18	671
11	☑	1,33784E+18	TamerYazar	TurkeyIsrael	ImMarketAnalys...	17/09/2009 16:45	1302
12	☑	1,33784E+18	VoiceM		campaigneropti...	31/08/2020 10:38	2
13	☑	1,33782E+18	WION	India	WIONWorldIsOn...	21/03/2016 03:44	292510
14	☑	1,33781E+18	DrKrutikaKuppalli		IDGlobalHealthV...	25/03/2019 04:14	21924
15	☑	1,33781E+18	Opoyi		Highqualitytrust...	13/01/2019 18:33	10332
16	☑	1,33779E+18	CityAM	LondonEngland	Londonsbusines...	09/06/2009 13:53	66224
17	☑	1,33779E+18	STOPCOMMON...	Global	Trustisnotcartebl...	25/10/2020 20:33	406
18	☑	1,33778E+18	ILKHA	Trkiye	OfficialTwitterac...	22/05/2015 08:31	4056
19	☑	1,33778E+18	BraderzGTTO	BristolUK	Oneofthoseleftyt...	24/07/2012 08:18	6430
20	☑	1,33778E+18	AlexVie	LosAngelesCA	MarinevetYogiKr...	24/01/2010 04:43	125

Output data

		ï»¿id	username	userlocation	serdescription	usercreated	userfollowers
1	☑	1,34054E+18	*	****************...	****************...	08/04/2009 17:52	405
2	☑	1,33816E+18	*	****************...	****************...	21/09/2009 15:27	834
3	☑	1,33786E+18	*	****************...	****************...	25/06/2020 23:30	10
4	☑	1,33786E+18	*	****************...	****************...	10/09/2008 11:28	49165
5	☑	1,33785E+18	*	****************...	****************...	23/04/2020 17:58	152
6	☑	1,33785E+18	*	****************...	****************...	26/01/2020 21:43	105
7	☑	1,33785E+18	*	****************...	****************...	10/06/2013 17:49	2731
8	☑	1,33785E+18	*	****************...	****************...	25/03/2019 04:14	21924
9	☑	1,33785E+18	*	****************...	****************...	30/10/2009 17:53	887
10	☑	1,33784E+18	*	****************...	****************...	12/11/2012 04:18	671
11	☑	1,33784E+18	*	****************...	****************...	17/09/2009 16:45	1302
12	☑	1,33784E+18	*	****************...	****************...	31/08/2020 10:38	2
13	☑	1,33782E+18	*	****************...	****************...	21/03/2016 03:44	292510
14	☑	1,33781E+18	*	****************...	****************...	25/03/2019 04:14	21924
15	☑	1,33781E+18	*	****************...	****************...	13/01/2019 18:33	10332
16	☑	1,33779E+18	*	****************...	****************...	09/06/2009 13:53	66224
17	☑	1,33779E+18	*	****************...	****************...	25/10/2020 20:33	406
18	☑	1,33778E+18	*	****************...	****************...	22/05/2015 08:31	4056
19	☑	1,33778E+18	*	****************...	****************...	24/07/2012 08:18	6430
20	☑	1,33778E+18	*	****************...	****************...	24/01/2010 04:43	125

Summary statistics (input)

Parameter	Value
Scale of measure	Nominal scale
Number of measures	11020
Number of distinct values	6806
Mode	

Summary statistics (output)

Parameter	Value
Scale of measure	Nominal scale
Number of measures	11020
Number of distinct values	1
Mode	**...

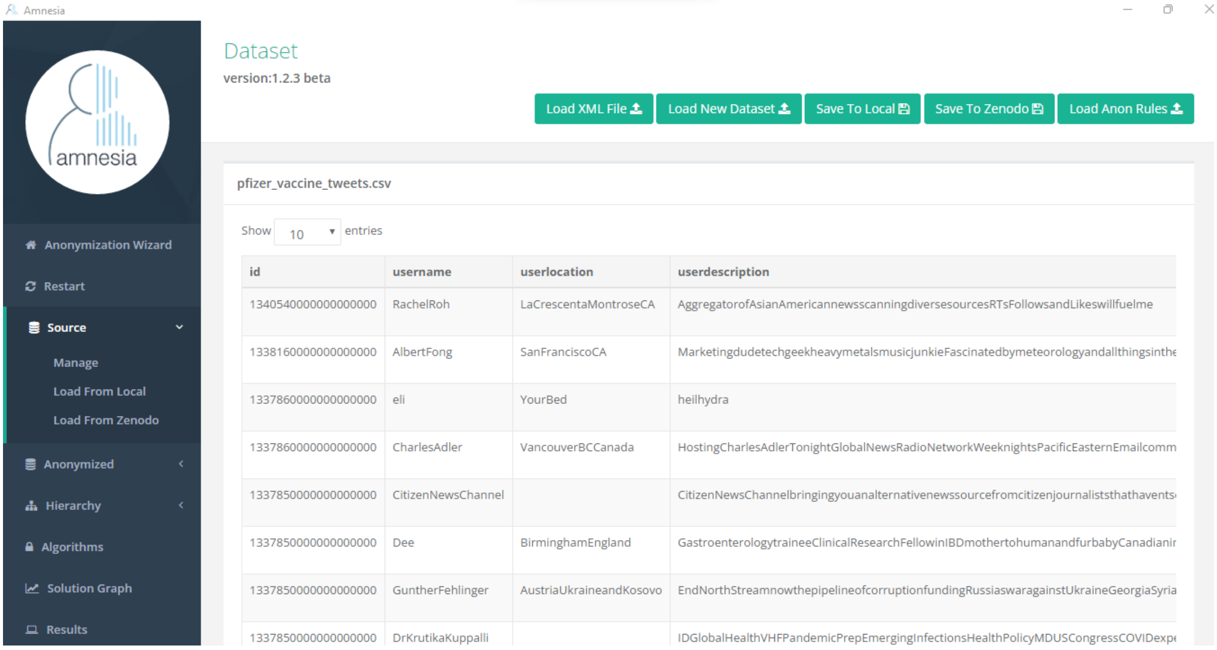
Figure 83 - Anonymization result

As it is possible to verify in Figure 83, all the values were anonymized to the biggest level of the hierarchy, because all of them have the special characters until the first character of the attribute value.

9.6. Pfizer Vaccine Tweets Dataset in Amnesia

Now, the original dataset is introduced in Amnesia Tool so that it is possible to anonymize the data.

At the moment of the dataset upload, the delimiter needs to be set. In this case, the delimiter is the semi-comma. The dataset upload can be seen in Figure 84.



version:1.2.3 beta

pfizer_vaccine_tweets.csv

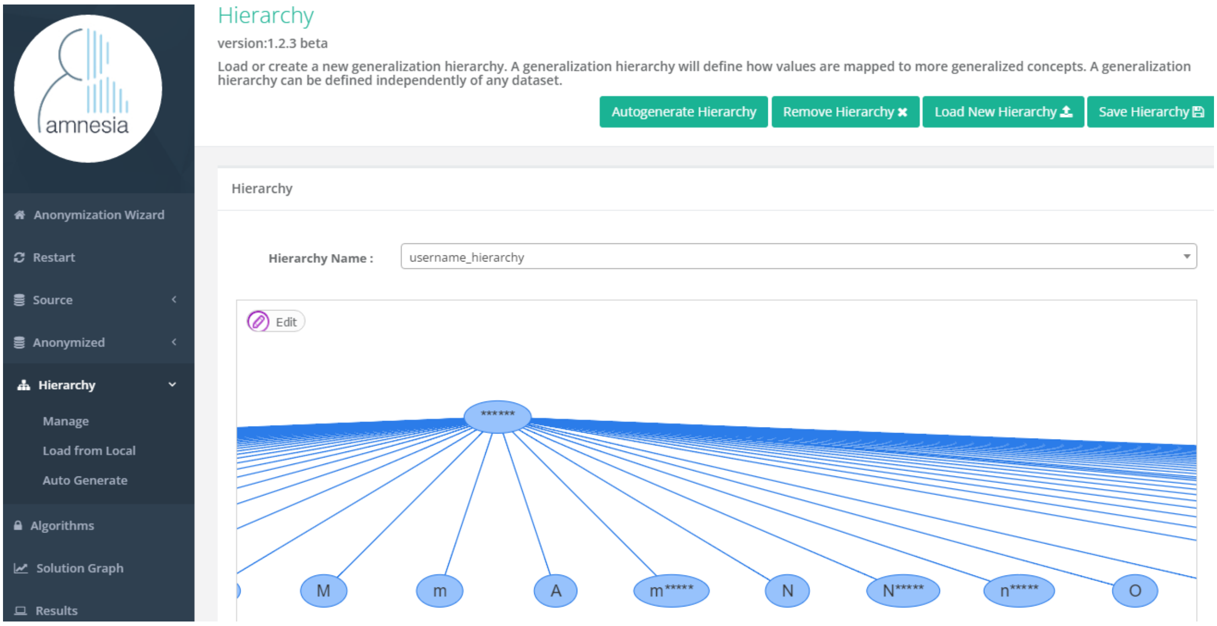
id	username	userlocation	userdescription
1340540000000000000	RachelRoh	LaCrescentaMontroseCA	AggregatorofAsianAmericannewsscanningdiversesourcesRTsFollowsandLikeswillfuelme
1338160000000000000	AlbertFong	SanFranciscoCA	MarketingdudetechgeekheavymetalsmusicjunkieFascinatedbymeteorologyandallthingsinthe
1337860000000000000	eli	YourBed	heilhydra
1337860000000000000	CharlesAdler	VancouverBCCanada	HostingCharlesAdlerTonightGlobalNewsRadioNetworkWeeknightsPacificEasternEmailcomm
1337850000000000000	CitizenNewsChannel		CitizenNewsChannelbringingyouanalternativenewssourcefromcitizenjournaliststhathavents
1337850000000000000	Dee	BirminghamEngland	GastroenterologytraineeClinicalResearchFellowinIBDmothertohumanandfurbabyCanadianir
1337850000000000000	GuntherFehlinger	AustriaUkraineandKosovo	EndNorthStreamnowthepipelineofcorruptionfundingRussiaswaragainstUkraineGeorgiaSyria
1337850000000000000	DrKrutikaKuppalli		IDGlobalHealthVHFPandemicPrepEmergingInfectionsHealthPolicyMDUSCongressCOVIDexpe

Figure 84 - Dataset Uploaded in Amnesia

After uploading the dataset, the hierarchies need to be created. The hierarchies are going to be created for the identifier, quasi-identifier, and sensitive attributes. According to Table 24, the identifier attribute is the username, the quasi-identifier attributes are userlocation, and userdescription and the sensitive attribute is the ID.

Starting with the identifier attribute, the first hierarchy to create is for the username.

The type of this hierarchy is Masking Based so that most of the characters are replaced by the special character *. The length assigned for this attribute is 6, which means that all the values are going to have 6 characters. In the first level of the hierarchy, the value is just the six special characters; in the second level of the hierarchy, the first character is the first letter of the username, and the other 5 characters are the special characters *. The null value is also present in the second level of the tree. The values that have just one character remains unchanged. This explanation can be verified in Figure 85.



Hierarchy

version:1.2.3 beta

Load or create a new generalization hierarchy. A generalization hierarchy will define how values are mapped to more generalized concepts. A generalization hierarchy can be defined independently of any dataset.

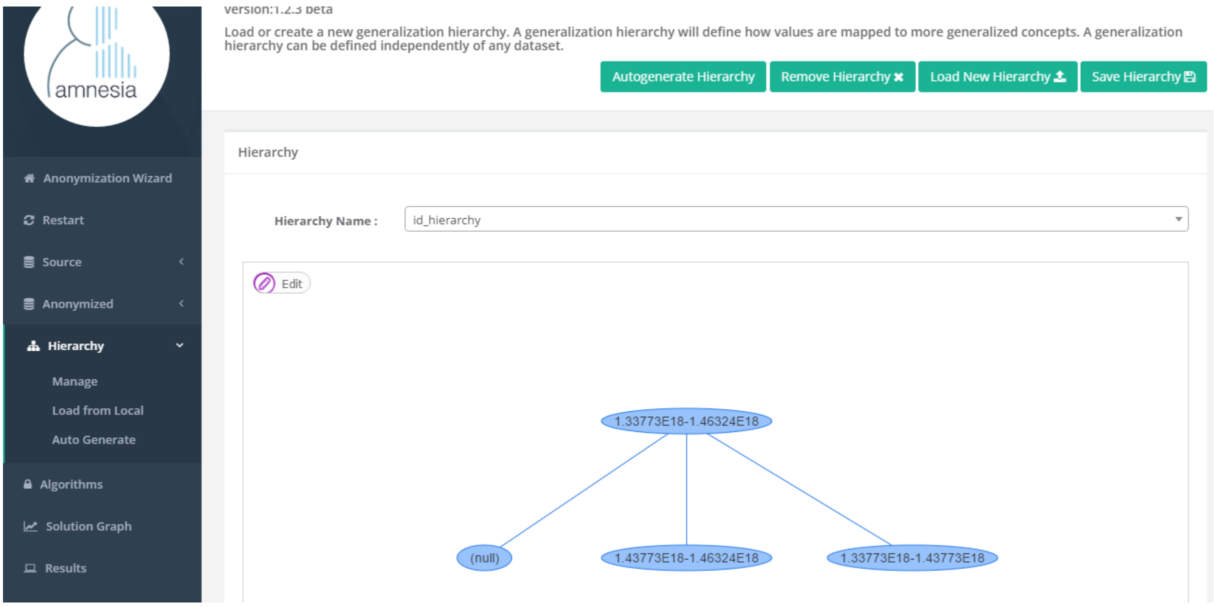
Hierarchy

Hierarchy Name : username_hierarchy

Figure 85 - username Hierarchy Amnesia

As there are no more identifying attributes now is time to proceed to the sensitive attribute (id).

The type chosen for this attribute is the range so that it is possible to have different intervals of values. The Domain start end limits interval is from 1337730000000000000 to 1463240000000000000 and the step given is 100000000000000000 so that there are not a huge number of values in the second level of the hierarchy. The result can be seen in Figure 86 and the null value is also present.



version:1.2.3 beta

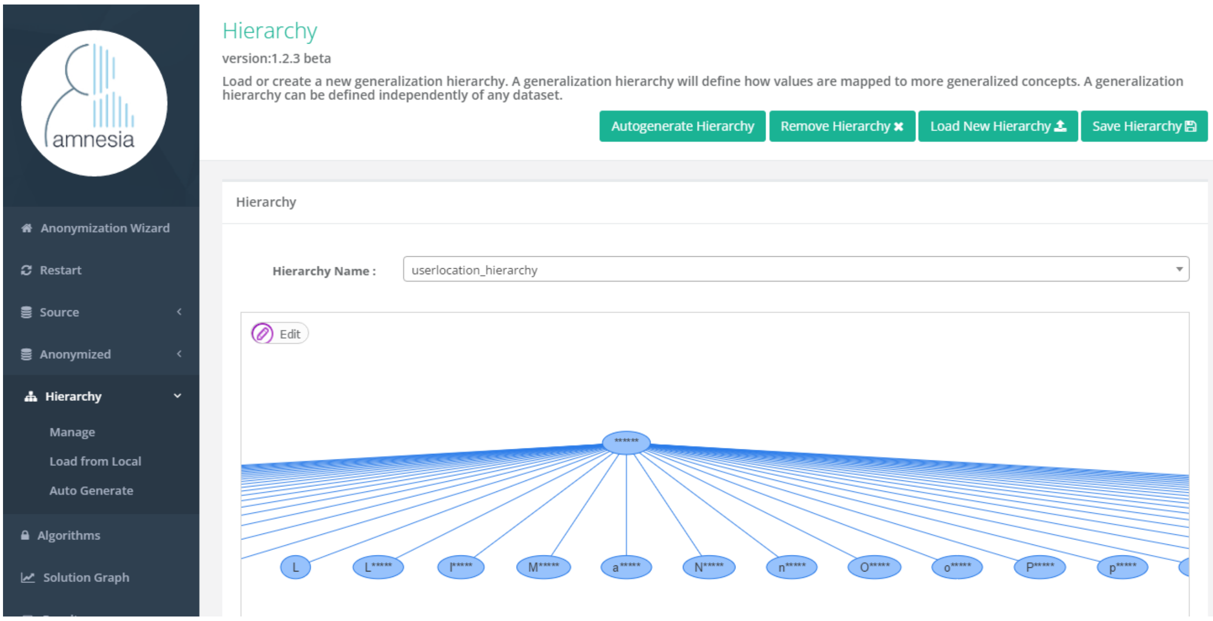
Load or create a new generalization hierarchy. A generalization hierarchy will define how values are mapped to more generalized concepts. A generalization hierarchy can be defined independently of any dataset.

Hierarchy

Hierarchy Name : id_hierarchy

Figure 86 - id Hierarchy Amnesia

As there are no more sensitive attributes now is time to create the hierarchies for the quasi-identifier attributes. The first one to create is for userlocation. The values given for the hierarchy are the same as the ones given for the username hierarchy, so the result is similar, as it is possible to verify in Figure 87. The first level of the tree also has the six special characters and the values from the second level of the hierarchy are the first original character with more than five special characters, besides the null value.



Hierarchy

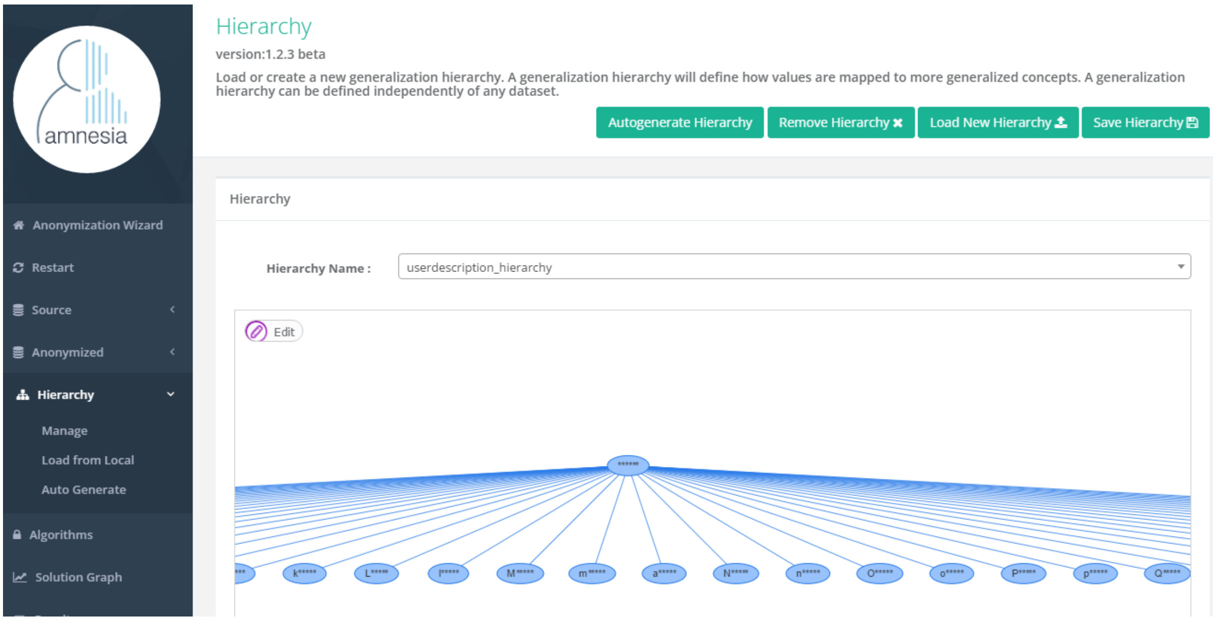
version:1.2.3 beta

Load or create a new generalization hierarchy. A generalization hierarchy will define how values are mapped to more generalized concepts. A generalization hierarchy can be defined independently of any dataset.

Hierarchy Name : userlocation_hierarchy

Figure 87 - userlocation Hierarchy Amnesia

Finally, the hierarchy is going to be created for the attribute userdescription. The information given to create the hierarchy is one more time the same that was used for the attributes userlocation and username. In Figure 88 is possible to verify that the result is also similar.



Hierarchy

version:1.2.3 beta

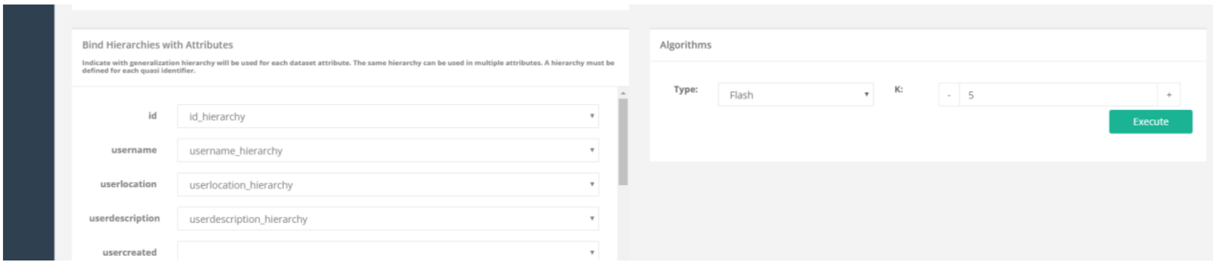
Load or create a new generalization hierarchy. A generalization hierarchy will define how values are mapped to more generalized concepts. A generalization hierarchy can be defined independently of any dataset.

Hierarchy

Hierarchy Name : userdescription_hierarchy

Figure 88 - userdescription Hierarchy Amnesia

Now that all the hierarchies are created, the tool has the button to conduct the user to the algorithms. Now is the time to connect each hierarchy to the attribute. As it was mentioned before, the tool only supports the algorithm type Flash. The K value set by the user is the value 5. As this is a bigger dataset, the value could be a little higher than the one used in the previous dataset. This can be verified in Figure 89.



Bind Hierarchies with Attributes

Indicate with generalization hierarchy will be used for each dataset attribute. The same hierarchy can be used in multiple attributes. A hierarchy must be defined for each quasi identifier.

id: id_hierarchy

username: username_hierarchy

userlocation: userlocation_hierarchy

userdescription: userdescription_hierarchy

usercreated:

Algorithms

Type: Flash

K: 5

Figure 89 - Algorithms Display

Figure 90 is the solution graph retrieved by the tool. As it is possible to verify there are only three safe solutions for this dataset.

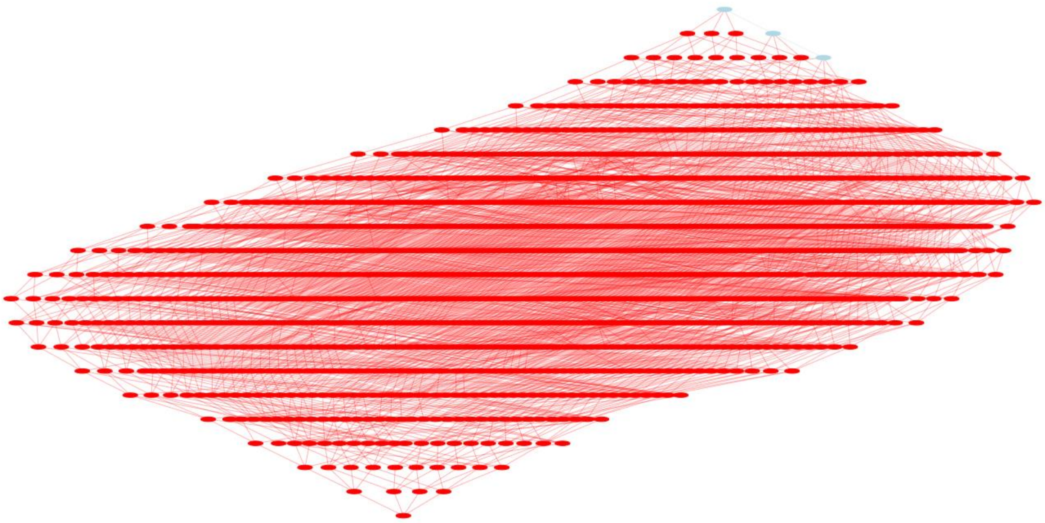


Figure 90 - Solution Graph Display

Observing closer the solution graph, it is possible to verify the values for the safe solutions, as shown in Figure 91.

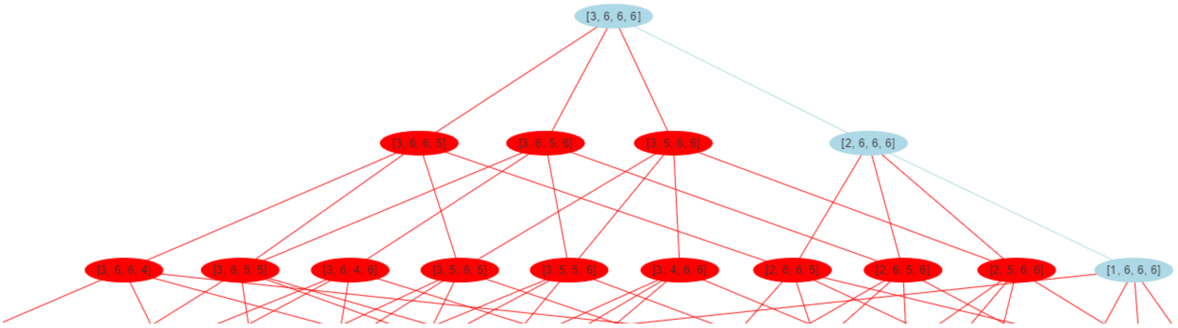


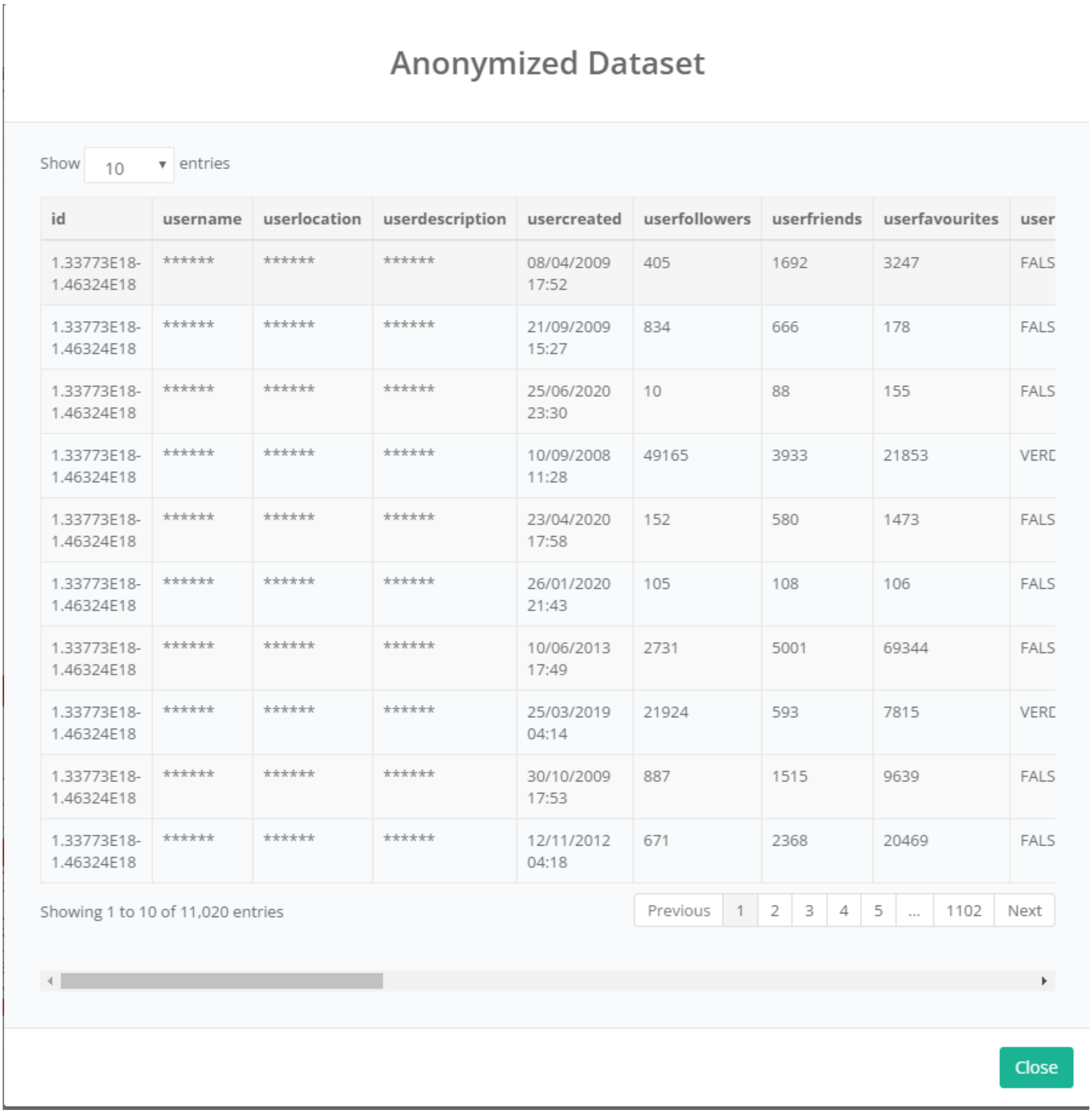
Figure 91 - Safe Solutions

The first level of the solution graph has the solution [3, 6, 6, 6], which means that the attribute ID was generalized to level 3, username was generalized to level 6, userlocation was generalized to level 6 and userdescription was also generalized to level 6.

In the second level of the solution graph, the solution [2, 6, 6, 6] is also a safe solution and means that id was generalized to level 2, username was generalized to level 6, userlocation was generalized to level 6 and userdescription was generalized to level 6.

Finally, the last safe solution in the graph, in the third level has the values [1, 6, 6, 6], which means that the attribute id was generalized to level 1, the attribute username was generalized to level 6, the attribute userlocation was generalized to level 6 and the attribute userdescription was generalized also to level 6.

The preview of the solution from the first level of the solution graph retrieves the following anonymized dataset, as shown in Figure 92.



Anonymized Dataset

Show 10 entries

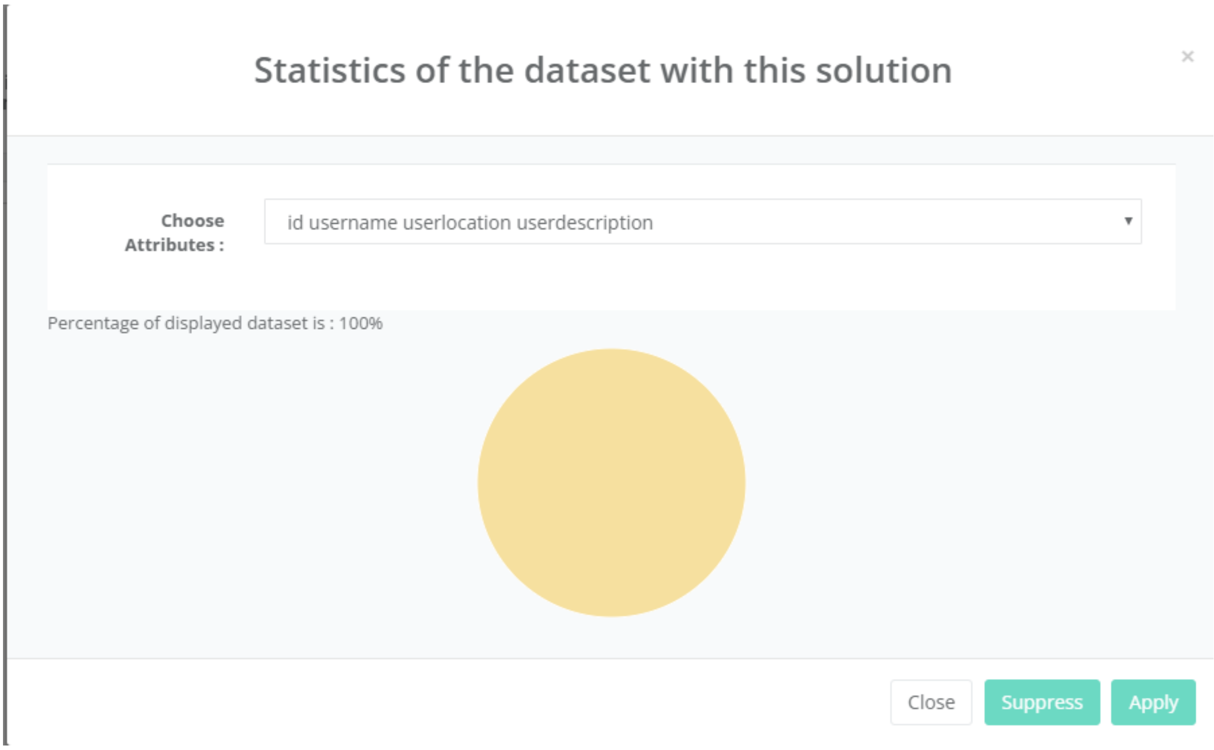
id	username	userlocation	userdescription	usercreated	userfollowers	userfriends	userfavourites	user
1.33773E18-1.46324E18	******	******	******	08/04/2009 17:52	405	1692	3247	FALS
1.33773E18-1.46324E18	******	******	******	21/09/2009 15:27	834	666	178	FALS
1.33773E18-1.46324E18	******	******	******	25/06/2020 23:30	10	88	155	FALS
1.33773E18-1.46324E18	******	******	******	10/09/2008 11:28	49165	3933	21853	VERD
1.33773E18-1.46324E18	******	******	******	23/04/2020 17:58	152	580	1473	FALS
1.33773E18-1.46324E18	******	******	******	26/01/2020 21:43	105	108	106	FALS
1.33773E18-1.46324E18	******	******	******	10/06/2013 17:49	2731	5001	69344	FALS
1.33773E18-1.46324E18	******	******	******	25/03/2019 04:14	21924	593	7815	VERD
1.33773E18-1.46324E18	******	******	******	30/10/2009 17:53	887	1515	9639	FALS
1.33773E18-1.46324E18	******	******	******	12/11/2012 04:18	671	2368	20469	FALS

Showing 1 to 10 of 11,020 entries

Figure 92 - Anonymized Dataset - First Level of Solution Graph

As it is possible to verify in Figure 92, the attribute id was generalized to a solution from the first level of the hierarchy, which means that this is the most generalized anonymization level. The other attributes were also anonymized to the first level of each hierarchy, so this solution is the one with the most anonymized data.

The statistics for this dataset, shown in Figure 93, shows that all the dataset was anonymized which is why the percentage of the displayed dataset is 100%.



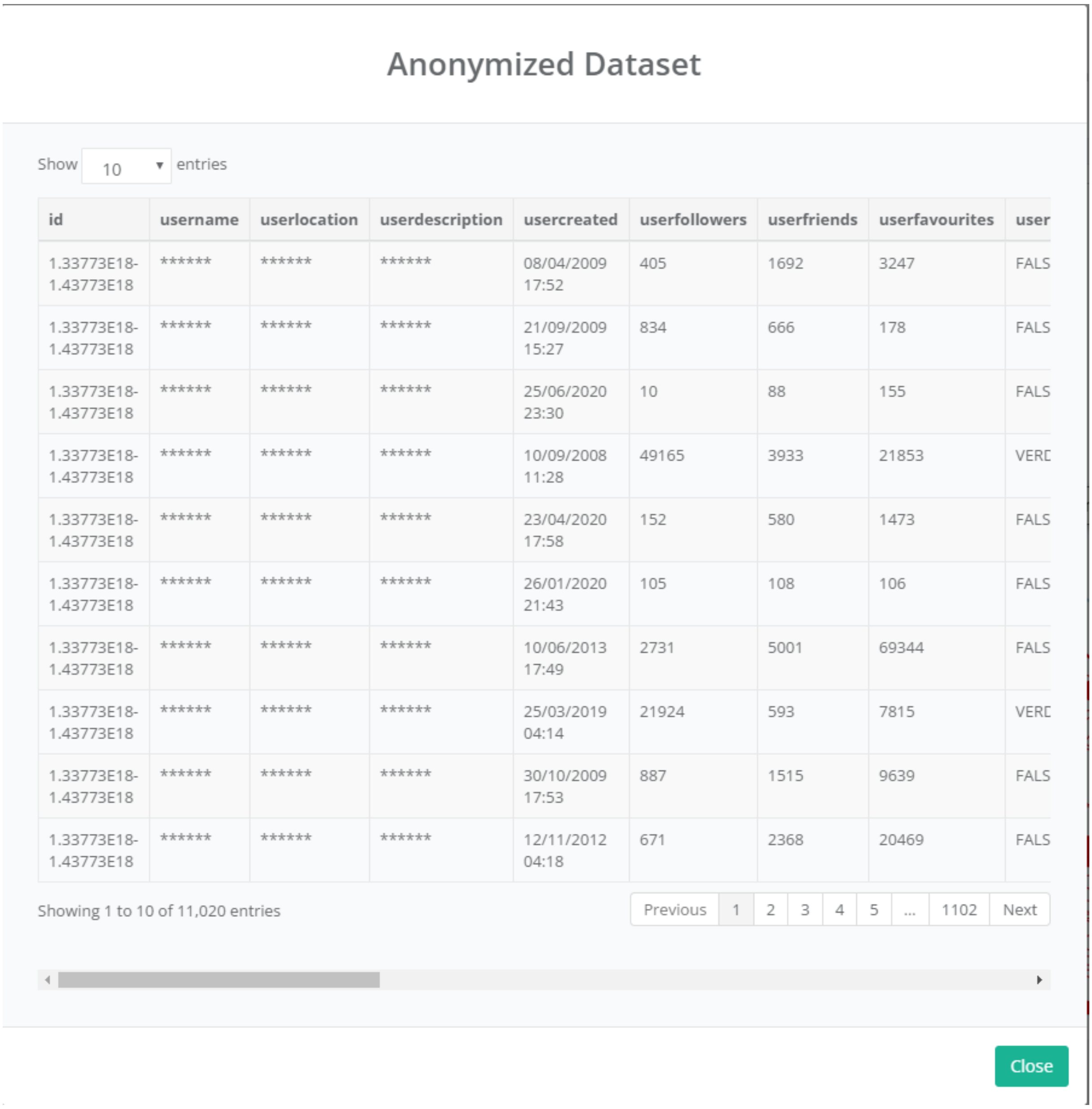
Statistics of the dataset with this solution

Choose Attributes : id username userlocation userdescription

Percentage of displayed dataset is : 100%

Figure 93 - Statistics - First Level of Solution Graph

Analysing now the safe solution from the second level of the solution graph, the result of the anonymized dataset is shown in Figure 94.



Anonymized Dataset

Show 10 entries

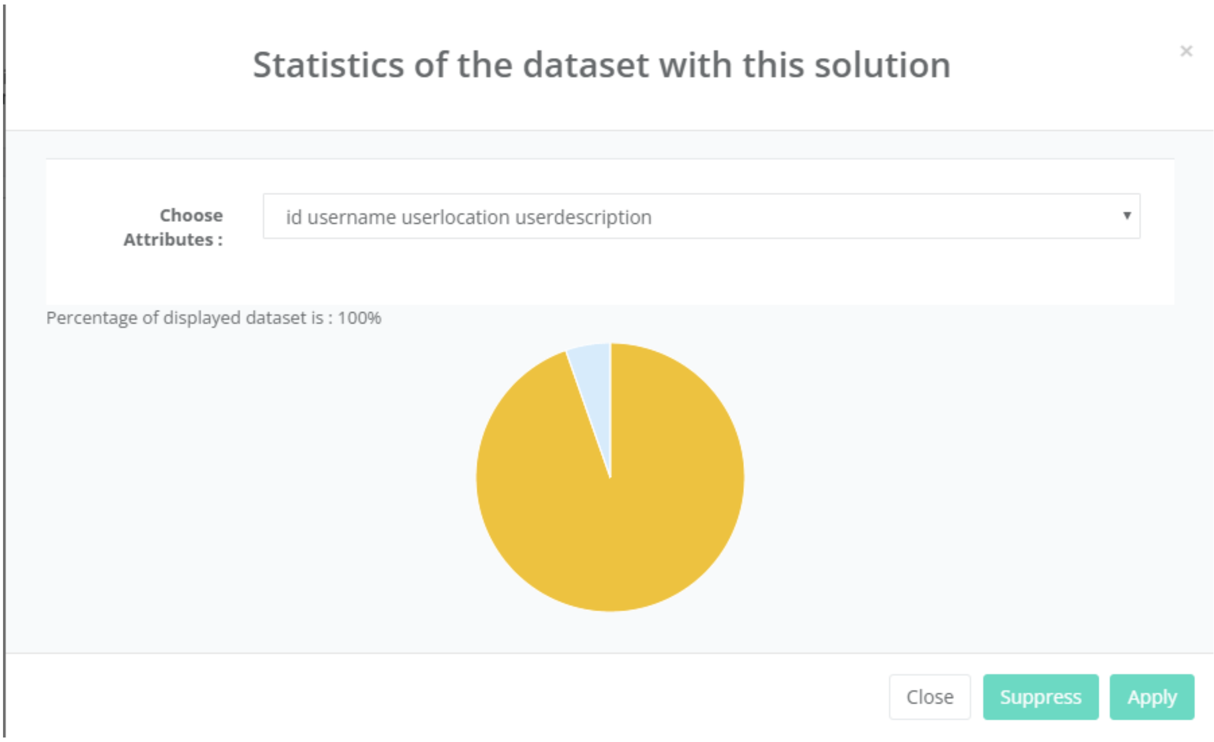
id	username	userlocation	userdescription	usercreated	userfollowers	userfriends	userfavourites	user
1.33773E18-1.43773E18	******	******	******	08/04/2009 17:52	405	1692	3247	FALS
1.33773E18-1.43773E18	******	******	******	21/09/2009 15:27	834	666	178	FALS
1.33773E18-1.43773E18	******	******	******	25/06/2020 23:30	10	88	155	FALS
1.33773E18-1.43773E18	******	******	******	10/09/2008 11:28	49165	3933	21853	VERD
1.33773E18-1.43773E18	******	******	******	23/04/2020 17:58	152	580	1473	FALS
1.33773E18-1.43773E18	******	******	******	26/01/2020 21:43	105	108	106	FALS
1.33773E18-1.43773E18	******	******	******	10/06/2013 17:49	2731	5001	69344	FALS
1.33773E18-1.43773E18	******	******	******	25/03/2019 04:14	21924	593	7815	VERD
1.33773E18-1.43773E18	******	******	******	30/10/2009 17:53	887	1515	9639	FALS
1.33773E18-1.43773E18	******	******	******	12/11/2012 04:18	671	2368	20469	FALS

Showing 1 to 10 of 11,020 entries

Figure 94 - Anonymized Dataset - Second Level of Solution Graph

The attribute id was generalized mostly to a solution from the second level of the hierarchy, which is possible to verify in Figure 94. This means that the interval of values for this attribute is smaller than the ones from the previous solution. The other attributes were, again, anonymized for the first level of each hierarchy since the generalization levels for these is the same as for the ones from the previous solution.

Analysing the statistics for the second solution, in Figure 95, it is possible to see that the circle is not all the same colour because the id attribute had different results in the anonymization process.



Statistics of the dataset with this solution

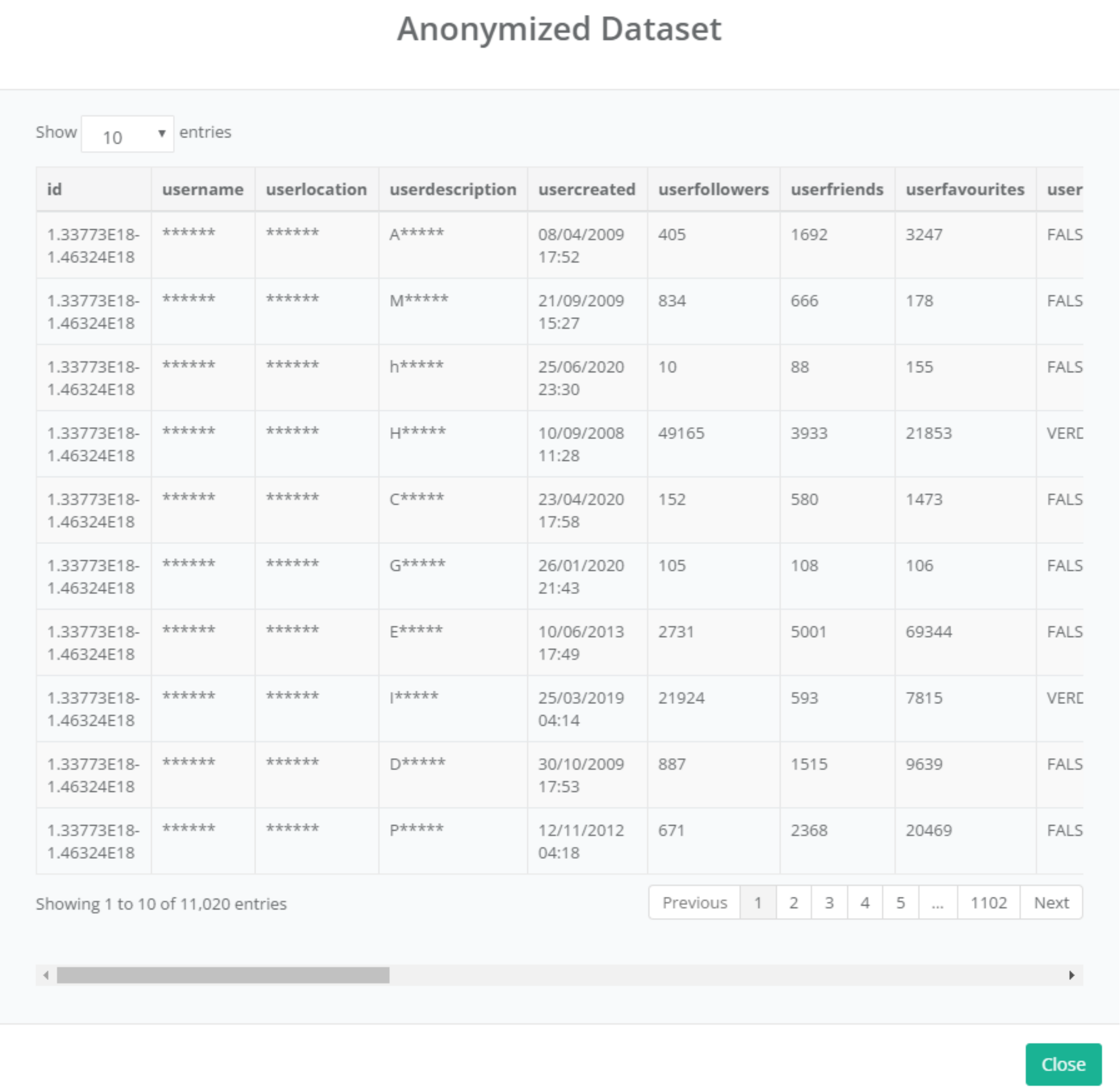
Choose Attributes : id username userlocation userdescription

Percentage of displayed dataset is : 100%

Figure 95 - Statistics - Second Level of Solution Graph

When the mouse is over the blue piece, it displays that 5% of the solution has the values [1.43773E18-1.46324E18, ******, ******, ******] and 95% of the solution has the values [1.33773E18-1.43773E18, ******, ******, ******]. This shows to the user the different value that the anonymized dataset has for the id attribute.

Analysing now the unsafe solution [3, 6, 6, 5] from the solution graph from Figure 91, which means that the id attribute was generalized to level 3, the username attribute was generalized to level 6, the userlocation attribute was generalized to level 6 and the userdescription attribute was generalized to level 5.



Anonymized Dataset

Show 10 entries

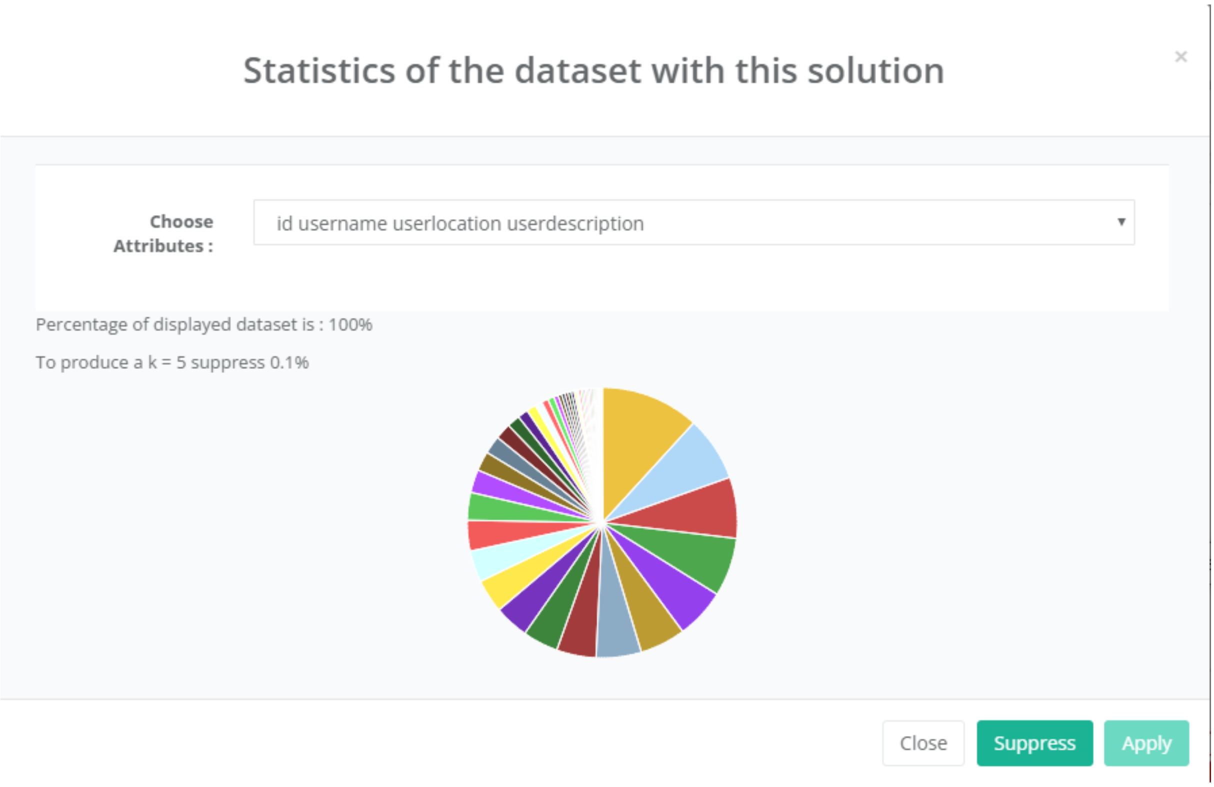
id	username	userlocation	userdescription	usercreated	userfollowers	userfriends	userfavourites	user
1.33773E18-1.46324E18	******	******	A*****	08/04/2009 17:52	405	1692	3247	FALS
1.33773E18-1.46324E18	******	******	M*****	21/09/2009 15:27	834	666	178	FALS
1.33773E18-1.46324E18	******	******	h*****	25/06/2020 23:30	10	88	155	FALS
1.33773E18-1.46324E18	******	******	H*****	10/09/2008 11:28	49165	3933	21853	VERD
1.33773E18-1.46324E18	******	******	C*****	23/04/2020 17:58	152	580	1473	FALS
1.33773E18-1.46324E18	******	******	G*****	26/01/2020 21:43	105	108	106	FALS
1.33773E18-1.46324E18	******	******	E*****	10/06/2013 17:49	2731	5001	69344	FALS
1.33773E18-1.46324E18	******	******	I*****	25/03/2019 04:14	21924	593	7815	VERD
1.33773E18-1.46324E18	******	******	D*****	30/10/2009 17:53	887	1515	9639	FALS
1.33773E18-1.46324E18	******	******	P*****	12/11/2012 04:18	671	2368	20469	FALS

Showing 1 to 10 of 11,020 entries

Figure 96 - Anonymized Dataset - Unsafe Solution

In Figure 96 it is possible to verify that the userdescription attribute is not totally generalized, showing the first letter of each value, as in the second level of the hierarchy in Figure 88.

The statistics for the unsafe solutions can be verified in Figure 97.



Statistics of the dataset with this solution

Choose Attributes : id username userlocation userdescription

Percentage of displayed dataset is : 100%

To produce a k = 5 suppress 0.1%

Figure 97 - Statistics - Unsafe Solution

The dataset is entirely anonymized but needs to suppress 0.1% of the dataset. The tool should allow the suppression of an unsafe solution, to turn it into a safe solution, but this button does not work, so the unsafe solutions could not be corrected. Each one of the different colours of the circle is the percentages of different solutions.

9.7. Conclusions using Pfizer Vaccine Tweets Dataset

This dataset was bigger than the one used previously. The Amnesia tool takes almost 5 minutes to anonymize the dataset and more than 3 minutes to display the solution graph. The time that takes to display the solution graph is always the same every time this page is left to visualize the anonymized dataset. Compared with ARX Anonymization Tool, Amnesia took longer because the anonymization result was displayed immediately in the first tool. The Amnesia tool also showed different results for the dataset anonymization while ARX just retrieved one solution. This is because Amnesia retrieves a solution for each level of each attribute while ARX only retrieves a solution for the quasi-identifier attributes. As it is possible to verify in Amnesia results, the quasi-identifier attributes were always anonymized for the same hierarchy level in the safe solutions, so the quasi-identifier attributes only have one safe solution. The ARX Data Anonymization Tool only retrieves this solution. The variation in the Amnesia, in the safe solution, is for the id attribute, a sensitive attribute, which is unchanged in the anonymization result.

The solution in Amnesia is very confused, contrary to what happens in ARX. In ARX the solution is clearer, and it is simpler to access the result of the anonymization because it only has the results for the quasi-identifier attributes. The fact that Amnesia retrieves the solutions that violate the k-anonymity process makes it very difficult to choose a solution.

Both tools allow the download of the anonymized datasets in CSV formats.

10. Conclusions and Future Work

The volume of data used online has increased exponentially, which could bring high risks for the individuals. Personal data is increasingly used in marketing campaigns, forecasting future trends, helping in scientific and medical research and many other examples. One of the main problems is because data can contain sensitive information regarding each individual and the disclosure of such data can cause huge damages in life, like identity deft, money extraction and many others. Also, an individual could want to create a dataset with personal information and publish the data and could not do that with the original information. For this, it is important to understand the anonymization process. Some regulations protect individuals against data disclosure. In Europe, the European General Data Protection Regulation (EU GDPR) is the regulation created to protect individuals and their data.

In this thesis, we studied the use of the anonymization tools in two datasets. An assessment was done using OSSpal, to verify which tool could be the best to perform the anonymization process in personal information. Some anonymization algorithms like k-anonymity and l-diversity as well as the anonymization techniques, like generalization, suppression, anatomization, permutation, and perturbation were briefly explained since they were not the thesis main objective.

One of the contributions of this thesis is the assessment of the data anonymization tools using the OSSpal methodology. This methodology is used to evaluate free open-source tools to help users and organizations to choose the best ones. For this assessment, some categories are used. The categories used were functionality, operational software characteristics, support and services, documentation, software technology attributes, community and adaption, and development process. With this evaluation, ARX Data Anonymization Tool has an evaluation of 3.1 points, Amnesia has 2.45 points and UTS Anonymization Toolbox has 1.6 points, in a punctuation where the maximum value is 5. Therefore, it was possible to conclude that ARX Data Anonymization Tool is the best tool to be used in this situation, followed by Amnesia and, the tool with the worst result is UTD Anonymization Toolbox.

After this and the second contribution mentioned in Chapter 1, two datasets were used to assess the tools. The first dataset is smaller than the second one. The first dataset was inserted in ARX Data Anonymization Tool, in Amnesia and UTD Anonymization Toolbox. The second dataset was only performed in ARX Data Anonymization Tool and Amnesia because the UTD Anonymization Toolbox requires the user to insert the string values manually in a file, which is not a simple task to perform manually and that could cause errors. All the tools performed the anonymization immediately, but the anonymization process was not simple in all of them. In Amnesia it is simple to follow all the steps to anonymize the data, but the tool stops various times along the process, which turns the anonymization process into something slow. In the results obtained, the solution graph retrieved by the tool is very confusing even for a dataset with only 7 attributes. If the dataset had 20 attributes, it could be almost impossible to understand the solution graph. The solution graph retrieves safe and unsafe solutions. The unsafe solutions are the ones that violate k-anonymity definition. The tool informs the user that it is possible to turn unsafe solutions into safe, by just clicking in a Suppression button. Nevertheless, this operation always retrieves an error, so it is not possible to verify the result of a safe solution created from an unsafe solution. In UTD Anonymization Toolbox, as it was mentioned, the process is very manual and can origin some errors in the anonymization process. It does not have a user interface which makes the process more difficult and, as all the logs are retrieved in the command line, it is also not simple to understand everything that is retrieved. Also, the documentation of the tool is not entirely satisfactory, not helping the user going through the data anonymization process. ARX Data Anonymization Tool was simple to use and has more algorithms that could be used to anonymize the data than the other tools. The results are simpler to understand because the solutions retrieved are just the ones for the quasi-identifier attributes and not for the quasi-identifier, sensitive and identifier attributes as in Amnesia. The process is not so simpler to follow in Amnesia but every time that a step is missing the tool displays a pop-up warning what is missing.

The second dataset was bigger and had some symbols in the text of the original dataset. These symbols needed to be removed because Amnesia does not allow the upload of the dataset with these symbols. This dataset was only uploaded to Amnesia and ARX Data Anonymization Tool. In Amnesia, it took almost 5 minutes to anonymize the dataset and almost 3 minutes to display the solution graph, so the anonymization of a bigger dataset is a long process in this tool. In ARX, the anonymization results are displayed instantaneously. Amnesia, once again, retrieves a confusing solution graph, with just 3 safe solutions, while ARX retrieves just one solution. This difference is because in ARX the results are only displayed for the quasi-identifier attributes, and it is possible to verify that in Amnesia, the quasi-identifier attributes were always anonymized for the same hierarchy level (in the safe solutions), so the quasi-identifier attributes only have one safe solution, and this is the same number of solutions that the ARX Data Anonymization Tool retrieves. Once again, it is not possible to know the result of a safe solution from an unsafe solution because the tool retrieves an error when clicking on the suppression button.

Concluding, the ARX Data Anonymization Tool retrieves results that are simpler to visualize and understand and it is the faster tool used. Amnesia is also a good tool but has some errors, and it is slower, and the solution graph is very confused. UTD is not a tool that we recommend being used since the anonymization process needs to be very manual and it is the most difficult one to use to obtain results.

The third contribution for this thesis is the best tools to use according to dataset characteristics. If the dataset that is going to be used is big, the tool chosen should be ARX Data Anonymization, because this is the one that supports big datasets and do not take longer time to anonymize the data. Amnesia supports datasets up to 4MB, but the anonymization process is slow. UTD Anonymization Toolbox should not be used with bigger datasets because the data needs to be inserted manually and this can lead the user to errors.

For the fourth contribution for this thesis, we had the main weaknesses and difficulties in practical use of each tool, which some of them were already explained but, in short, ARX Data Anonymization does not mentioned in detail the steps that need to be performed to anonymize a dataset. This tool also shows some algorithms or privacy models to link to the sensitive attributes where the user could choose between different options for the same algorithm, but it does not allow to pick another privacy if it is not already linked to the attribute. In Amnesia, the tool does not upload big datasets and the dataset could not contain symbols, otherwise the tool retrieves an error. The solution graph with the possible anonymization solutions has a lot of values, which make difficult to choose a solution and to visualize them. The solution graph has safe solutions and unsafe solutions. In the tool documentation is represented that the unsafe solutions can be transformed into safe solutions by clicking in a button to suppress the unsafe solutions and turn this one into a safe solution, but the button always retrieves an error. Besides this, the tool is slow processing big datasets and returning the results for these datasets. Finally, UTD Anonymization Toolbox is a very manual tool where it is easy to introduce human errors. This tool does not have a graphical interface and the values need to be inserted manually in some files to be read by the tool, which make it a difficult and slow process.

As future work, we intend to study ARX Data Anonymization Tool with the other anonymization algorithms to verify the differences in the results. ARX also has a variety of information about the quality of the information that can be analysed in the future, and it completes the present work. Another issue that can be addressed in the future is the comparison of the information loss between the results from each tool using specific metrics that helps with these values. All these studies will help in the decision of the best tool to use, but with the present analyses we are able to say that the most complete one is ARX Data Anonymization Tool.

References

Aggarwal, G., Feder, T., Kenthapadi, K., Motwani, R., Panigrahy, R., Thomas, D., & Zhu, A. (2005). Approximation Algorithms for k-Anonymity. *Journal of Privacy Technology*.

Almokbily, R. S., & Rauf, A. (2018). Anatomization through generalization (AG): A hybrid privacy-preserving approach to prevent membership, identity and semantic similarity disclosure attacks. In *2018 International Conference on Computing, Mathematics and Engineering Technologies (iCoMET)* (pp. 1-7).

Amnesia. (n.d.). Retrieved from https://amnesia.openaire.eu/index.html

Amnesia. (2020, July 7). *Amnesia.* Retrieved from https://amnesia.openaire.eu/Scenarios/AmnesiaKMAnonymityTutorial.pdf

Amnesia. (2020). *Documentation*. Retrieved from Amnesia: https://amnesia.openaire.eu/about-documentation.html#

Arora, D. K., Bansal, D., & Sofat, S. (2014). Comparative Analysis of Anonymization Techniques. In *International Journal of Electronic and Electrical Engineering* (pp. 773-778). International Research Publication House.

ARX - Data Anonymization Tool. (2021). *Privacy models*. Retrieved from https://arx.deidentifier.org/overview/privacy-criteria/

ARX. (n.d.). *ARX - Data Anonymization Tool*. Retrieved from ARX - Data Anonymization Tool: https://arx.deidentifier.org/

Brickell, J., & Shmatikov, V. (2008). The Cost of Privacy: Destruction of Data-Mining Utility in Anonymized Data Publishing. In *Proceedings of the 14th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* (pp. 70-78). Nevada, USA: Association for Computing Machinery.

Brito, F. T., & Machado, J. (2017). Preservação de Privacidade de Dados: Fundamentos, Técnicas e Aplicações. In *Jornadas de Atualização em Informática - Chapter: 3* (p. 40). Brasil: Sociedade Brasileira de Computação - SBC.

Cantiello, P., Mastroianni, M., & Rak , M. (2021). A Conceptual Model for the General Data Protection Regulation. In B. M. O. Gervasi, *Computational Science and Its Applications – ICCSA 2021* (pp. 60-77). Cham: Springer International Publishing.

Dados.gov. (2021). *Sobre o dados.gov*. Retrieved from Dados.gov: https://dados.gov.pt/pt/docs/about_dadosgov/

data.world. (2017). *CDC Nutrition, PhysicalActivity, and Obesity by State*. Retrieved from data.world: https://data.world/basilhayek/cdc-nutrition-physical-activity-and-obesity-by-state

Fedesoriano. (2021, September 10). *Heart Failure Prediction Dataset*. Retrieved from Kaggle: https://www.kaggle.com/fedesoriano/heart-failure-prediction

Fung, B. C., Wang, K., Chen, R., & Yu, P. S. (2010). Privacy-Preserving Data Publishing: A Survey of Recent Developments. *ACM Comput. Surv. Volume 42*.

Ghinita, G., Karras, P., Kalnis, P., & Mamoulis, N. (2007). Fast Data Anonymization with Low Information Loss. In *Proceedings of the 33rd International Conference on Very Large Data Bases* (pp. 758–769). Vienna: VLDB Endowment.

Gunawan, D., & Mambo, M. (2018). Set-Valued Data Anonymization Maintaining Data Utility and Data Property. In *Proceedings of the 12th International Conference on Ubiquitous Information Management and Communication.* Malaysia: Association for Computing Machinery.

He, Y., & Naughton, J. (2009). Anonymization of Set-Valued Data via Top-down, Local Generalization. *Proc. VLDB Endow.*, 934-945.

Israni, K., Chopra, S., & Jewani, K. (2017). PRIVACY PRESERVING USING ANONYMIZATION AND PERTURBATION IN CLASSIFICATION. *Indian Journal of Scientific Research*, 164.

Kaggle. (2021, January 21). *Pfizer Vaccine Tweets*. Retrieved from Kaggle: https://www.kaggle.com/gpreda/pfizer-vaccine-tweets

Li, D., Xianmang, H., Cao, L., & Chen, H. (2015). Permutation anonymization. *Journal of Intelligent Information Systems*.

Li, N., Li, T., & Venkatasubramanian, S. (2007). t-Closeness: Privacy Beyond k-Anonymity and l-Diversity. *IEEE 23rd International Conference on Data Engineering (ICDE)*, 106-115.

Machanavajjhala, A., Kifer, D., Gehrke, J., & Venkitasubramaniam, M. (2007). L-Diversity: Privacy beyond k-Anonymity. *ACM Trans. Knowl. Discov. Data*, 3-es.

Meindl, B., Kowarik, A., & Templ., M. (2018, March 22). *Using the interactive GUI - sdcApp*. Retrieved from sdcMicro5.1.1: http://sdctools.github.io/sdcMicro/articles/sdcMicro.html#introduction-and-main-features

Murthy, S., Bakar, A. A., Rahim, F. A., & Ramli, R. (2019). A Comparative Study of Data Anonymization Techniques. *2019 IEEE 5th Intl Conference on Big Data Security on Cloud (BigDataSecurity), IEEE Intl Conference on High Performance and Smart Computing, (HPSC) and IEEE Intl Conference on Intelligent Data and Security (IDS)*, (pp. 306-309). Washington, DC, USA.

Nergiz, M. E., Atzori, M., & Clifton, C. (2007). Hiding the presence of individuals from shared databases. *Proceedings of the ACM SIGMOD International Conference on Management of Data*, 665-676.

Pereira, A. K., Sousa, A. P., Santos, J. R., & Bernardino, J. (2018). Open Source Data Mining Tools Evaluation using OSSpal Methodology. *Proceedings of the 13th International Conference on Software Technologies - ICSOFT*, 672-678.

Prasser, F., Eicher, J., Spengler, H., Bild, R., & Kuhn, K. (2020). Flexible data anonymization using ARX-Current status and challenges ahead. *Software: Practice and Experience*.

Preda, G. (2021, 11 23). *Pfizer Vaccine Tweets*. Retrieved from Kaggle.com: https://www.kaggle.com/gpreda/pfizer-vaccine-tweets

Rao, P. S., & Satyanarayana, S. (2018). Privacy preserving data publishing based on sensitivity in context of Big Data using Hive. *Journal of Big Data*, Article number: 20.

Ren, W., Wang, L., Choo, K.-K. R., & Xhafa, F. (2019). *Security and Privacy for Big Data, Cloud Computing and Applications.* London, United Kingdom: The Institution of Engineering and Technology.

Samarati, P., & Sweeney, L. (1998). *Protecting privacy when disclosing information: k-anonymity and its enforcement through generalization and suppression.* technical report, SRI International.

Sartor, N. (2019, February 4). *Top 5 Free Data Anonymization Tools*. Retrieved from aricloak: https://aircloak.com/top-5-free-data-anonymization-tools/

Sharma, S., Choudhary, N., & Jain, K. (2019). A Study on Models and Techniques of Anonymization in Data Publishing. *International journal of scientific research in science, engineering and technology*, 84-90.

Terrovitis, M., Mamoulis, N., & Kalnis, P. (2008). Privacy-Preserving Anonymization of Set-Valued Data. *VLDB Endowment*.

Union, E. (2016). Regulation (EU) 2016/679 Of the European Parliament and of the Council. *Office Journal of the European Union*.

UT Dallas Data Security and Privacy Lab. (n.d.). *Anonymization ToolBox*. Retrieved from UTD Anonymization Toolbox: http://cs.utdallas.edu/dspl/cgi-bin/toolbox/index.php?go=home

Wikipedia. (2021, June 1). *Kaggle*. Retrieved from Wikipedia - The Free Encyclopedia: https://en.wikipedia.org/wiki/Kaggle

Annexes

Annex A

In this annexe are the hierarchies created for section 6.2 from Chapter 6.

For the SSN_Hierarchy, the file created as the following content:

```
range
name SSN_Hierarchy
type int
height 2

1.2012012E8,8.9645123E8        has        1.2012012E8,3.2012012E8
3.2012012E8,5.2012012E8                   5.2012012E8,7.2012012E8
7.2012012E8,8.9645123E8
```

The file for Sex_Hierarchy has the content below:

```
distinct
name Sex_Hierarchy
type string
height 2

Sex has Female Male
```

For the Job_Hierarchy, the file created had the following content:

```
distinct
name Job_Hierarchy
type string
height 2

Job has Teacher Scientist Engineer Accountant
```

Finally, for the Disease_Hierarchy, the hierarchy file was created with the following information:

```
distinct
name Disease_Hierarchy
type string
height 6

Hype* has Hypertense
Aneu* has Aneurysm
Diab* has Diabetes
Chol* has Cholesterol
Canc* has Cancer

Hyp** has Hype*
Ane** has Aneu*
Dia** has Diab*
Cho** has Chol*
Can** has Canc*

Hy*** has Hyp**
An*** has Ane**
Di*** has Dia**
Ch*** has Cho**
Ca*** has Can**

H**** has Hy***
A**** has An***
D**** has Di***
C**** has Ch*** Ca***

Disease has H**** A**** D**** C****
```

Annex B

This is the content of the DATA file of section 6.3 from chapter 6.

```
1,Joana,26,Female,Teacher,273649999,Cancer
2,João,27,Male,Scientist,537957333,Hypertense
3,Carolina,22,Female,Engineer,658658224,Diabetes
4,Ana,21,Female,Accountant,125896345,Aneurysm
5,Paulo,10,Male,Teacher,546895665,Cholesterol
6,Maria,56,Female,Scientist,756985321,Cancer
7,Pedro,45,Male,Engineer,120365120,Hypertense
8,Ricardo,76,Male,Accountant,125639856,Diabetes
9,Francisca,34,Female,Teacher,125630235,Aneurysm
10,Maria,36,Female,Scientist,896451230,Cholesterol
11,Teresa,44,Female,Engineer,890246879,Cancer
12,Vera,23,Female,Accountant,215634895,Hypertense
13,Rui,19,Male,Teacher,123456789,Diabetes
14,Tiago,67,Male,Scientist,120365201,Aneurysm
15,Henrique,43,Male,Engineer,890560235,Cholesterol
16,Carlos,89,Male,Accountant,452782982,Cancer
17,Rita,34,Female,Teacher,325658985,Hypertense
18,Catarina,17,Female,Scientist,415756324,Diabetes
19,Micaela,21,Female,Engineer,120120120,Aneurysm
20,Micael,42,Male,Accountant,321654987,Cholesterol
```

Annex C

Information is written in the command line. This is the complete information of what is represented in chapter 9, in section 9.3.

```
Reading data takes 0sec.s
Processing EID = 1, [[[20:80)],[[0:1]]]
        Inserted 2 (left) and 3 (right)
Processing EID = 2, [[[20.0:54.0]],[[0:1]]]
        Inserted 4 (left) and 5 (right)
Processing EID = 3, [[(54.0:80.0)],[[0:1]]]
        Inserted 6 (left) and 7 (right)
Processing EID = 4, [[[20.0:47.0]],[[0:1]]]
        Inserted 8 (left) and 9 (right)
Processing EID = 5, [[(47.0:54.0]],[[0:1]]]
        Inserted 10 (left) and 11 (right)
Processing EID = 6, [[(54.0:60.0]],[[0:1]]]
        Inserted 12 (left) and 13 (right)
Processing EID = 7, [[(60.0:80.0)],[[0:1]]]
        Inserted 14 (left) and 15 (right)
Processing EID = 8, [[[20.0:42.0]],[[0:1]]]
        Inserted 16 (left) and 17 (right)
Processing EID = 9, [[(42.0:47.0]],[[0:1]]]
        Inserted 18 (left) and 19 (right)
Processing EID = 10, [[(47.0:52.0]],[[0:1]]]
        Inserted 20 (left) and 21 (right)
Processing EID = 11, [[(52.0:54.0]],[[0:1]]]
        Removing 11 (no allowable cuts)
Processing EID = 12, [[(54.0:57.0]],[[0:1]]]
        Inserted 22 (left) and 23 (right)
Processing EID = 13, [[(57.0:60.0]],[[0:1]]]
        Inserted 24 (left) and 25 (right)
Processing EID = 14, [[(60.0:64.0]],[[0:1]]]
        Inserted 26 (left) and 27 (right)
Processing EID = 15, [[(64.0:80.0)],[[0:1]]]
        Inserted 28 (left) and 29 (right)
Processing EID = 16, [[[20.0:39.0]],[[0:1]]]
        Inserted 30 (left) and 31 (right)
Processing EID = 17, [[(39.0:42.0]],[[0:1]]]
        Inserted 32 (left) and 33 (right)
Processing EID = 18, [[(42.0:45.0]],[[0:1]]]
        Inserted 34 (left) and 35 (right)
Processing EID = 19, [[(45.0:47.0]],[[0:1]]]
        Inserted 36 (left) and 37 (right)
Processing EID = 20, [[(47.0:50.0]],[[0:1]]]
        Inserted 38 (left) and 39 (right)
Processing EID = 21, [[(50.0:52.0]],[[0:1]]]
        Removing 21 (no allowable cuts)
Processing EID = 22, [[(54.0:56.0]],[[0:1]]]
        Inserted 40 (left) and 41 (right)
Processing EID = 23, [[(56.0:57.0]],[[0:1]]]
        Removing 23 (no allowable cuts)
Processing EID = 24, [[(57.0:59.0]],[[0:1]]]
        Inserted 42 (left) and 43 (right)
Processing EID = 25, [[(59.0:60.0]],[[0:1]]]
        Removing 25 (no allowable cuts)
Processing EID = 26, [[(60.0:62.0]],[[0:1]]]
        Removing 26 (no allowable cuts)
Processing EID = 27, [[(62.0:64.0]],[[0:1]]]
        Inserted 44 (left) and 45 (right)
Processing EID = 28, [[(64.0:68.0]],[[0:1]]]
        Inserted 46 (left) and 47 (right)
Processing EID = 29, [[(68.0:80.0)],[[0:1]]]
        Inserted 48 (left) and 49 (right)
Processing EID = 30, [[[20.0:37.0]],[[0:1]]]
        Inserted 50 (left) and 51 (right)
Processing EID = 31, [[(37.0:39.0]],[[0:1]]]
        Inserted 52 (left) and 53 (right)
Processing EID = 32, [[(39.0:41.0]],[[0:1]]]
        Removing 32 (no allowable cuts)
Processing EID = 33, [[(41.0:42.0]],[[0:1]]]
        Removing 33 (no allowable cuts)
Processing EID = 34, [[(42.0:44.0]],[[0:1]]]
        Inserted 54 (left) and 55 (right)
Processing EID = 35, [[(44.0:45.0]],[[0:1]]]
        Removing 35 (no allowable cuts)
Processing EID = 36, [[(45.0:46.0]],[[0:1]]]
        Removing 36 (no allowable cuts)
Processing EID = 37, [[(46.0:47.0]],[[0:1]]]
        Removing 37 (no allowable cuts)
Processing EID = 38, [[(47.0:49.0]],[[0:1]]]
        Inserted 56 (left) and 57 (right)
Processing EID = 39, [[(49.0:50.0]],[[0:1]]]
        Removing 39 (no allowable cuts)
Processing EID = 40, [[(54.0:55.0]],[[0:1]]]
        Removing 40 (no allowable cuts)
Processing EID = 41, [[(55.0:56.0]],[[0:1]]]
        Removing 41 (no allowable cuts)
Processing EID = 42, [[(57.0:58.0]],[[0:1]]]
        Removing 42 (no allowable cuts)
Processing EID = 43, [[(58.0:59.0]],[[0:1]]]
        Removing 43 (no allowable cuts)
Processing EID = 44, [[(62.0:63.0]],[[0:1]]]
        Removing 44 (no allowable cuts)
Processing EID = 45, [[(63.0:64.0]],[[0:1]]]
        Removing 45 (no allowable cuts)
Processing EID = 46, [[(64.0:66.0]],[[0:1]]]
        Inserted 58 (left) and 59 (right)
Processing EID = 47, [[(66.0:68.0]],[[0:1]]]
        Inserted 60 (left) and 61 (right)
Processing EID = 48, [[(68.0:71.0]],[[0:1]]]
        Inserted 62 (left) and 63 (right)
Processing EID = 49, [[(71.0:80.0)],[[0:1]]]
        Inserted 64 (left) and 65 (right)
Processing EID = 50, [[[20.0:35.0]],[[0:1]]]
        Inserted 66 (left) and 67 (right)
Processing EID = 51, [[(35.0:37.0]],[[0:1]]]
        Removing 51 (no allowable cuts)
Processing EID = 52, [[(37.0:38.0]],[[0:1]]]
        Removing 52 (no allowable cuts)
Processing EID = 53, [[(38.0:39.0]],[[0:1]]]
        Removing 53 (no allowable cuts)
Processing EID = 54, [[(42.0:43.0]],[[0:1]]]
        Removing 54 (no allowable cuts)
Processing EID = 55, [[(43.0:44.0]],[[0:1]]]
        Removing 55 (no allowable cuts)
Processing EID = 56, [[(47.0:48.0]],[[0:1]]]
        Removing 56 (no allowable cuts)
Processing EID = 57, [[(48.0:49.0]],[[0:1]]]
        Removing 57 (no allowable cuts)
Processing EID = 58, [[(64.0:65.0]],[[0:1]]]
        Removing 58 (no allowable cuts)
Processing EID = 59, [[(65.0:66.0]],[[0:1]]]
        Removing 59 (no allowable cuts)
Processing EID = 60, [[(66.0:67.0]],[[0:1]]]
        Removing 60 (no allowable cuts)
Processing EID = 61, [[(67.0:68.0]],[[0:1]]]
        Removing 61 (no allowable cuts)
Processing EID = 62, [[(68.0:69.0]],[[0:1]]]
        Removing 62 (no allowable cuts)
Processing EID = 63, [[(69.0:71.0]],[[0:1]]]
        Inserted 68 (left) and 69 (right)
Processing EID = 64, [[(71.0:74.0]],[[0:1]]]
        Removing 64 (no allowable cuts)
Processing EID = 65, [[(74.0:80.0)],[[0:1]]]
        Inserted 70 (left) and 71 (right)
Processing EID = 66, [[[20.0:34.0]],[[0:1]]]
        Inserted 72 (left) and 73 (right)
Processing EID = 67, [[(34.0:35.0]],[[0:1]]]
        Removing 67 (no allowable cuts)
Processing EID = 68, [[(69.0:70.0]],[[0:1]]]
        Removing 68 (no allowable cuts)
Processing EID = 69, [[(70.0:71.0]],[[0:1]]]
        Inserted 74 (left) and 75 (right)
Processing EID = 70, [[(74.0:76.0]],[[0:1]]]
        Inserted 76 (left) and 77 (right)
Processing EID = 71, [[(76.0:80.0)],[[0:1]]]
        Removing 71 (no allowable cuts)
Processing EID = 72, [[[20.0:32.0]],[[0:1]]]
        Inserted 78 (left) and 79 (right)
Processing EID = 73, [[(32.0:34.0]],[[0:1]]]
        Removing 73 (no allowable cuts)
Processing EID = 74,
[[(70.0:71.0]],[[0.0:0.0]]]
        Removing 74 (no allowable cuts)
Processing EID = 75,
[[(70.0:71.0]],[(0.0:1.0]]]
        Removing 75 (no allowable cuts)
Processing EID = 76, [[(74.0:75.0]],[[0:1]]]
        Removing 76 (no allowable cuts)
Processing EID = 77, [[(75.0:76.0]],[[0:1]]]
        Removing 77 (no allowable cuts)
Processing EID = 78, [[[20.0:31.0]],[[0:1]]]
        Inserted 80 (left) and 81 (right)
Processing EID = 79, [[(31.0:32.0]],[[0:1]]]
        Removing 79 (no allowable cuts)
Processing EID = 80, [[[20.0:29.0]],[[0:1]]]
        Removing 80 (no allowable cuts)
Processing EID = 81, [[(29.0:31.0]],[[0:1]]]
        Removing 81 (no allowable cuts)
Anonymization takes 0sec.s
Writing data takes 0sec.s
```