1 2 9 0

UNIVERSIDADE D
COIMBRA

Security and Privacy

Anonymization of Datasets with Privacy, Utility and Risk Analysis

Francisco Catarino Mendes - 2019222823
Leonardo Oliveira Pereira - 2020239125
Department of Informatics Engineering
University of Coimbra

Introduction

The main objective of this assignment is to perform a detailed analysis of the anonymization process of a dataset, with the help of the ARX tool. The work is organized into 5 steps, in which will be performed a detailed analysis of the choices made.

To begin with, a search for the right dataset is going to be made, from which a dataset will be selected. Then, one will sanitize and characterize that dataset for the anonymization process, as well as conduct a detailed analysis, selection, and configuration of the appropriate anonymization/privacy models that will be applied to the dataset.

From there, an analysis and an optimization are to be performed to the utility and privacy levels of the selected anonymization/privacy models, together with an analysis of the risk of re-identification of the selected anonymization/privacy models.

1 - Selection, Importing, and Characterization of the Dataset

In this phase, the dataset being studied is going to be selected, imported, and characterized. After an extensive search, and with the help of the Professor, a great dataset was found. Titled ”Employability Classification of Over 70,000 Job Applicants”, it contains a comprehensive collection of information regarding job applicants and their respective employability scores. It was arranged to assist organizations and recruiters in evaluating the suitability of candidates for various employment opportunities.

The dataset was gathered from diverse sources, including job portals, career fairs, and online applications, over a specified period. The information was collected in a standardized manner, ensuring consistency across various data points. The dataset encompasses a wide range of industries, positions, and qualifications. It also comprises structured data, organized into multiple columns or features, that will be classified next.

When anonymizing a dataset, the goal is to protect individual privacy by ensuring that the individuals cannot be identified while still retaining the utility of the dataset for analysis. The main goal of anonymizing this dataset is to enable analysis of factors that influence employability without compromising the privacy of the job applicants.

1.1 - Classification of the attributes

The attributes (columns) of the dataset go as follows:

a) **Column #0 -** unique id of the person;

b) **Age -** age of the applicant, either older than 35 years old or younger than 35 years old;

c) **EdLevel -** the education level of the applicant (Undergraduate, Master, Phd, etc);

d) **Acessibility -** this attribute raises some questions, as its definition was not specified by the makers of the dataset;

e) **Employment -** whether the applicant is currently employed at the time of the application;

f) **Gender -** the gender of the applicant;

g) **MentalHealth -** this attribute raises some questions, as its definition was not specified by the makers of the dataset;

h) **MainBranch -** whether the applicant is a professional developer or not;

i) **YearsCode -** number of years the applicant has of coding experience;

j) **YearsCodePro -** number of years the applicant has of coding experience, but in a professional context;

k) **Country -** the country of the applicant;

l) **PreviousSalary -** the applicant's previous job salary;

m) **HaveWorkedWith -** a description of the code languages the applicant has worked with/has experience of;

n) **ComputerSkills -** the actual number of the code languages the applicant has worked with/has experience of;

o) **Employed -** if the applicant was successful or not (hired or not);

From the list above, it is clear that two of the attributes are not very clear for interpretation. Therefore, it was considered that "Accessibility" is whether the job applicant requires special accommodations for accessibility due to a disability or any physical condition, and "MentalHealth" is whether the applicant has reported any mental health conditions that could affect their work.

Now, regarding the classification of these attributes, Table 1 shows the results obtained:

Classification of the attributes				
Attribute	**Identifying**	**Quasi-Identifying**	**Sensitive**	**Insensitive**
Column #0	X			
Age		X		
Accessibility			X	
EdLevel		X		
Employment				X
Gender		X		
MentalHealth			X	
MainBranch		X		
YearsCode		X		
YearsCodePro		X		
Country		X		
PreviousSalary			X	
HaveWorkedWith		X		
ComputerSkills		X		
Employed				X

Table 1: Classification of attributes

Identifying - the only identifying attribute chosen was the ID number present in the dataset, represented by the column ”Column #0”, as it is the only attribute to uniquely identify an individual without being considered as sensitive. It was pondered if the ”Country” attribute should go in here, as it can probably identify people from unusual countries in this database, but in the end, it was chosen not to.

Quasi-Identifying - the majority of the attributes come in this section. Things like ”Age”, ”EdLevel”, ”Gender”, ”MainBranch”, ”YearsCode”, ”YearsCodePro”, ”Country”, ”HaveWorkedWith”, and ”ComputerSkills” while not explicitly identifying someone, can be combined with data from other public sources to de-anonymize (re-identify) the owner of a record. As said above, ”Country” is a tricky one, therefore it was decided to be a quasi-identifier. ”ComputerSkills” and ”HaveWorkedWith” are other attributes that should be mentioned. These are very important quasi-identifiers, mainly the latter, as it is very specific from person to person and can be powerful in linkage attacks. Furthermore, ”MainBranch” could be considered insensitive, but it was thought to be useful to classify it here due to the context of the dataset.

Sensitive - here are the individual-specific private/sensitive attributes that should not be publicly disclosed, those being ”Accessibility”, ”MentalHealth”, ”PreviousSalary”, and ”HaveWorkedWith”. Concerning the first two, they characterize health conditions, which should be always private. The ”PreviousSalary” attribute is also sensitive, as it is very specific from person to person, and can be a target of discrimination. Finally, the ”HaveWorkedWith” attribute pops up as sensitive, due to it being somewhat unique, and due to the fact that it contains the skills of a person, which is something that can be said to be private, but its position is also arguable, as this information can be considered to have no privacy risks.

Insensitive - ending the classification is the insensitive category, in which only two attributes were colocated. Both the ”Employment” and the ”Employed” features do not fall into the previous three categories, because neither the employment status nor the results of the application are that relevant in terms of identification and privacy, despite that an argument could be made for ”Employment” to be a quasi-identifier.

1.2 - Analysis of the privacy risks

As for the privacy risks of the dataset, it contains multiple quasi-identifiers such as ”Age”, ”Gender”, ”Country”, ”YearsCode”, and ”EdLevel”. Alone, these might not identify an individual, but combined, they could potentially lead to identification, especially in smaller or more homogeneous populations. Then, attributes like ”PreviousSalary” can also be quite unique, especially when combined with ”YearsCodePro” and ”Country”.

Furthermore, given the nature of the data, it’s possible that some of the information, like ”YearsCode”, ”YearsCodePro”, or ”PreviousSalary”, could be linked with data from other sources such as professional networks or company databases to re-identify individuals. Another problem is if any of the applicants have a unique combination of skills or come from countries with fewer professionals in the dataset, as this increases the risk of linkability.

Regarding sensitive attributes, they should not be forgotten in this analysis. For example, data corresponding to ”MentalHealth” and ”Accessibility” is sensitive. Disclosure could lead to discrimination or stigmatization. ”PreviousSalary” must also be treated carefully, as it is financially sensitive.

Lastly, looking at the distribution of the attributes, it is clear that the coding model to be used is mainly towards a generalization, as this is a case of dealing with quasi-identifiers that can indirectly lead to the identification of individuals when combined with other data. By generalizing detailed information into broader categories, it becomes harder to pinpoint an individual. It also allows for the data to retain its usefulness for analysis at a higher level.

2 - Coding Model and Privacy requirements

As it was just said, the coding model used in this dataset is biased to mostly a generalization. Figure 1 shows its application in ARX:

Figure 1: Coding Model

In the construction of the hierarchies, for a categorical attribute, a specific value can be replaced with a general value according to a given taxonomy. For a numerical attribute, exact values can be replaced with an interval that covers those.

All the Quasi-Identifiers were put into hierarchies, except ”Age” and ”HaveWorkedWith, since the first is already generalized (>35, >35) and the second is represented by the ”ComputerSkills” attribute. Tables 2 to 9 show all these hierarchies:

Level-0	Level-1
Master	Higher Education
PhD	Higher Education
NoHigherEd	Lower Education
Other	Lower Education
Undergraduate	Lower Education

Table 2: EdLevel Hierarchy

Level-0	Level-1
Man	{Man, NonBinary,...
NonBinary	{Man, NonBinary,...
Woman	{Man, NonBinary,...

Table 3: Gender Hierarchy

Level-0	Level-1
Dev	{Dev, NotDev}
NotDev	{Dev, NotDev}

Table 4: MainBranch Hierarchy

Level-0	Level-1	Level-2	Level-3	Level-4
0	[0, 5[	[0, 10[	[0, 20[	[0, 40[
1	[0, 5[	[0, 10[	[0, 20[	[0, 40[
2	[0, 5[	[0, 10[	[0, 20[	[0, 40[
3	[0, 5[	[0, 10[	[0, 20[	[0, 40[
4	[0, 5[	[0, 10[	[0, 20[	[0, 40[
5	[5, 10[	[0, 10[	[0, 20[	[0, 40[
6	[5, 10[	[0, 10[	[0, 20[	[0, 40[
7	[5, 10[	[0, 10[	[0, 20[	[0, 40[
8	[5, 10[	[0, 10[	[0, 20[	[0, 40[
9	[5, 10[	[0, 10[	[0, 20[	[0, 40[
10	[10, 15[	[10, 20[	[0, 20[	[0, 40[
11	[10, 15[	[10, 20[	[0, 20[	[0, 40[
12	[10, 15[	[10, 20[	[0, 20[	[0, 40[
13	[10, 15[	[10, 20[	[0, 20[	[0, 40[
14	[10, 15[	[10, 20[	[0, 20[	[0, 40[
15	[15, 20[	[10, 20[	[0, 20[	[0, 40[
16	[15, 20[	[10, 20[	[0, 20[	[0, 40[

Table 5: YearsCode Hierarchy

Level-0	Level-1	Level-2	Level-3
0	[0, 10[	[0, 20[	[0, 40[
1	[0, 10[	[0, 20[	[0, 40[
2	[0, 10[	[0, 20[	[0, 40[
3	[0, 10[	[0, 20[	[0, 40[
4	[0, 10[	[0, 20[	[0, 40[
5	[0, 10[	[0, 20[	[0, 40[
6	[0, 10[	[0, 20[	[0, 40[
7	[0, 10[	[0, 20[	[0, 40[
8	[0, 10[	[0, 20[	[0, 40[
9	[0, 10[	[0, 20[	[0, 40[
10	[10, 20[	[0, 20[	[0, 40[
11	[10, 20[	[0, 20[	[0, 40[
12	[10, 20[	[0, 20[	[0, 40[
13	[10, 20[	[0, 20[	[0, 40[
14	[10, 20[	[0, 20[	[0, 40[
15	[10, 20[	[0, 20[	[0, 40[
16	[10, 20[	[0, 20[	[0, 40[

Table 6: YearsCodePro Hierarchy

Level-0	Level-1
Slovakia	Europe
Slovenia	Europe
Spain	Europe
Sweden	Europe
Switzerland	Europe
Turkey	Europe
Ukraine	Europe
United Kingdom ...	Europe
Australia	Oceania
Fiji	Oceania
New Zealand	Oceania
Argentina	America
Barbados	America
Belize	America
Bolivia	America
Brazil	America
Canada	America

Table 7: Country Hierarchy

Level-0	Level-1	Level-2	Level-3	Level-4
0	[0, 15[	[0, 30[	[0, 60[	[0, 108[
1	[0, 15[	[0, 30[	[0, 60[	[0, 108[
2	[0, 15[	[0, 30[	[0, 60[	[0, 108[
3	[0, 15[	[0, 30[	[0, 60[	[0, 108[
4	[0, 15[	[0, 30[	[0, 60[	[0, 108[
5	[0, 15[	[0, 30[	[0, 60[	[0, 108[
6	[0, 15[	[0, 30[	[0, 60[	[0, 108[
7	[0, 15[	[0, 30[	[0, 60[	[0, 108[
8	[0, 15[	[0, 30[	[0, 60[	[0, 108[
9	[0, 15[	[0, 30[	[0, 60[	[0, 108[
10	[0, 15[	[0, 30[	[0, 60[	[0, 108[
11	[0, 15[	[0, 30[	[0, 60[	[0, 108[
12	[0, 15[	[0, 30[	[0, 60[	[0, 108[
13	[0, 15[	[0, 30[	[0, 60[	[0, 108[
14	[0, 15[	[0, 30[	[0, 60[	[0, 108[
15	[15, 30[	[0, 30[	[0, 60[	[0, 108[
16	[15, 30[	[0, 30[	[0, 60[	[0, 108[

Table 8: ComputerSkills Hierarchy

Level-0	Level-1
	Unknown
APL	Program Langua...
APL;ASP.NET;My...	Program Langua...
APL;ASP.NET;jQu...	Program Langua...
APL;Angular;AWS...	Program Langua...
APL;Ansible;Flow;...	Program Langua...
APL;Ansible;npm;...	Program Langua...
APL;Assembly;An...	Program Langua...
APL;Assembly;Ba...	Program Langua...
APL;Assembly;Ba...	Program Langua...
APL;Assembly;Ba...	Program Langua...
APL;Assembly;Ba...	Program Langua...
APL;Assembly;Ba...	Program Langua...
APL;Assembly;Ba...	Program Langua...
APL;Assembly;Ba...	Program Langua...
APL;Assembly;Ba...	Program Langua...
APL;Assembly;Ba...	Program Langua...

Table 9: HaveWorkedWith Hierarchy

a) **EdLevel Hierarchy** - generalization to (Higher Education, Lower Education) in order to not lose that much utility and data;

b) **Gender Hierarchy** - not much to say, the gender is generalized to its 3 possible types;

c) **MainBranch Hierarchy** - again, there is not much one can do with 2 variations of a non-numerical attribute;

d) **YearsCode Hierarchy** - generalization set to groups of 5 years, in order to lose the lowest amount of utility while enhancing privacy;

e) **YearsCodePro Hierarchy** - here, the generalization is set to groups of 10, as the numbers tend to be lower;

f) **Country Hierarchy** - all the countries are generalized to their respective continent, enhancing privacy, crucially in those where the amount of individuals is very low;

g) **ComputerSkills Hierarchy** - the number of skills is put into groups of 15 in level 1, 30 in level 2, 60 in level 3, and the maximum at level 4.

h) **HaveWorkedWith Hierarchy** - the technical skills are dispersed in 2 basic groups, one that just refers to ”programming languages” and the other for those who do not possess skills, ”unknown”.

Moving on to the attributed weights in ARX, these are shown in Figure 2 below:

a) **Age: 0,8 -** a significant factor that can narrow down an individual’s identity, especially in combination with other attributes;

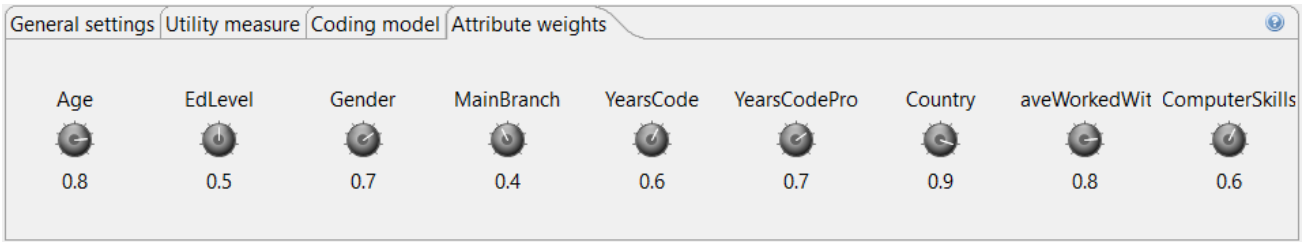

Figure 2: Attributed Weights

b) **EdLevel: 0,5 -** can provide some clues about an individual's background but may not be as specific as other attributes;

c) **Gender: 0,7 -** can narrow down an individual's identity, especially in combination with other attributes;

d) **MainBranch: 0,4 -** indicates an individual's primary field of employment, which could provide some clues about their background but may not be as specific as other attributes;

e) **YearsCode: 0,6 -** the number of years an individual has been coding can provide some insight into their experience level;

f) **YearsCodePro: 0,7 -** similar to YearsCode, YearsCodePro provides information about coding experience, but it is slightly more important due to the years characterizing a professional status;

g) **HaveWorkedWith: 0,8 -** a significant factor, as the name of the languages and technologies can be very specific;

h) **ComputerSkills: 0,6 -** provide some insight into an individual's proficiency in programming languages, but is not as relevant as "HaveWorkedWith".

To end this section, the suppression limit must also be mentioned, which is the maximal number of records that can be removed from the input dataset. ARX tells us that the recommended value for this parameter is "100%", so that is exactly what Figure 3 presents:

Figure 3: Suppression Limit

3 - Privacy Models Application

At this point in the anonymization, the selected privacy models, used for privacy assurance and privacy measurement, are applied to the dataset. The privacy models chosen are listed below:

K-Anonymity - aims at protecting datasets from re-identification in the prosecutor model. If one record in the table has some value QID, at least k-1 other records also have the value QID. Each group of indistinguishable records forms a so-called equivalence class. For this dataset, which includes quasi-identifiers like ”Age”, ”Gender”, and ”Country”, applying k-anonymity means that any attempt to identify an individual based on these attributes would yield at least k individuals with the same characteristics, thereby reducing the risk of re-identification. Its applicability is high, especially since the dataset includes quasi-identifiers that could be combined to identify individuals. In the analysis made, the value of k is set to 5, because this ensures that no individual can be singled out uniquely within the dataset, and provides a good balance between privacy and utility.

L-Diversity - the idea is that sensitive attributes must be “diverse” within each quasi-identifier equivalence class. Each equivalence class has at least L well-represented values. Since there exist sensitive attributes like ”MentalHealth” and ”Accessibility”, applying l-diversity would ensure that each combination of quasi-identifiers has a diverse set of values for these sensitive attributes, making it harder to infer any individual’s sensitive information. Its applicability is high for protecting sensitive attributes against attribute disclosure. Since the sensitive attributes only can be of two forms (yes or no), it makes sense that the value of l is set to 2, and similarly to the value of k, it maintains utility and privacy at the same time.

T-Closeness - distribution of the sensitive values in each equivalence class must be “close” to the corresponding distribution in the original table. Again, this model is particularly useful for attributes like ”MentalHealth” and ”Accessibility”, where one wants to prevent an attacker from deducing an individual’s attribute value based on the group they belong to. The applicability is lower than l-diversity but can be considered at least moderate. The value of t selected was 0,6, so that the distribution of a sensitive attribute in any group is almost identical to the overall distribution. This can somewhat minimize the risk of inferring an individual’s sensitive attributes based on their group while maintaining a high value of utility.

β-Likeness - this model restricts the relative maximal distance between distributions of sensitive attribute values, also considering positive and negative information gain. It recognizes that not all values of a sensitive attribute are equally likely and that some values may be more revealing than others. It sets a constraint on the conditional probability of the sensitive attribute within each group defined by the quasi-identifiers. It limits how much more probable it is to observe a particular sensitive value in any group compared to the whole dataset. A parameter β is also introduced to control the allowed deviation in probability. This privacy model was chosen mainly because it offers a probabilistic approach to protecting sensitive attributes. The dataset contains attributes like ”PreviousSalary” or ”HaveWorkedWith,” which could potentially be guessed by looking at the distribution of values, and β-Likeness can reduce this risk. The value of β chosen was 5, which gives a fair amount of privacy and utility to the anonymization.

,

3.1 - K-anonymity + L-diversity

In this combination, k-anonymity and l-diversity come together. Using the value specified above for k, the values 2 for ”MentalHealth” and ”Accessibility”, and the value 1000 for ”PreviousSalary” due to the range of values being of great number, the privacy models were applied to the dataset. In Figures 4 to 6, one can see the results obtained:

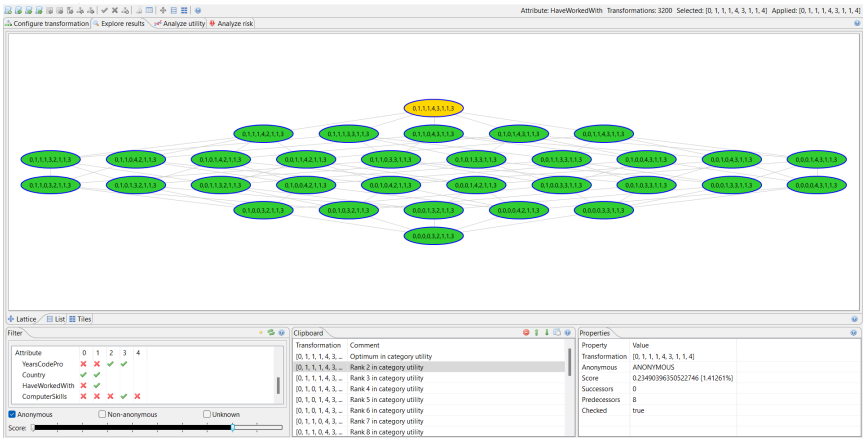

Figure 4: K-anonymity + L-diversity (1)

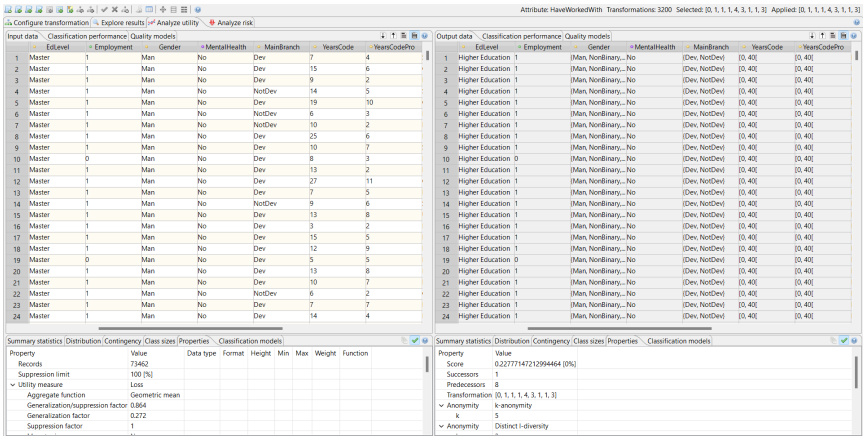

Figure 5: K-anonymity + L-diversity (2)

Figure 6: K-anonymity + L-diversity (3)

It must be said that these first images correspond to the optimal transformation that ARX recommends us to use. But then, in the ”explore results” section, in the filter, level 4 of ”ComputerSkills” was checked, which led to a better transformation to be created, supposedly. But when comparing the risks in the attack model section of the new one to the optimal one, the truth is that there are no differences between the two. Figures 7 to 9 show the information related to this new transformation:

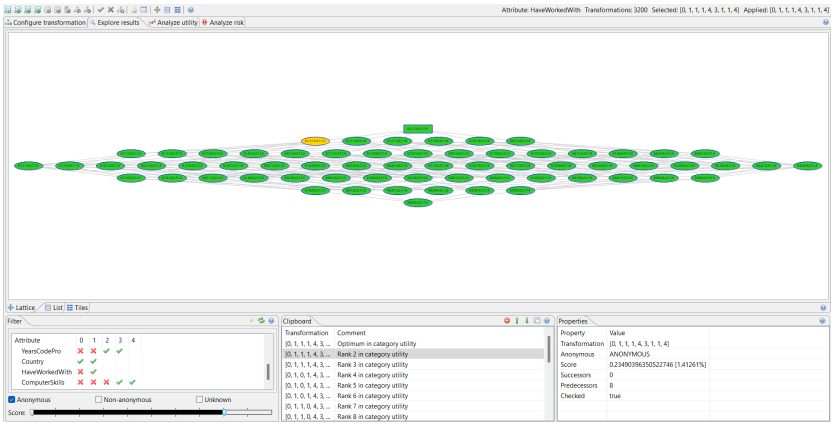

Figure 7: New Transformation (1)

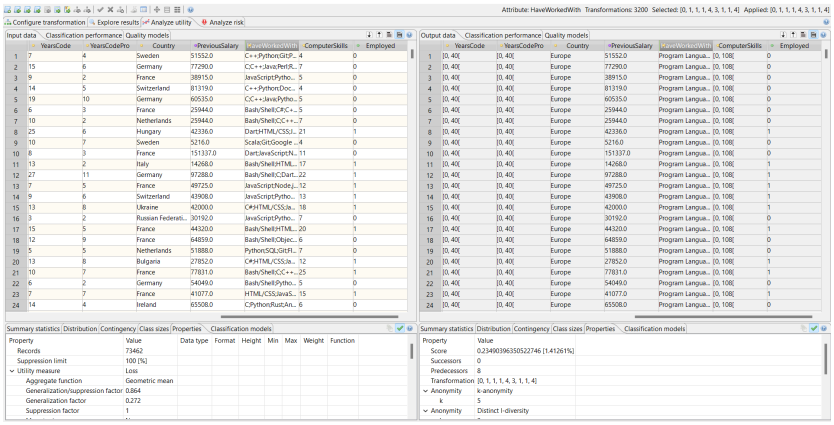

Figure 8: New Transformation (2)

Figure 9: New Transformation (3)

When analyzing the best possible results, one can see that the exact score in the properties pane which assesses the utility of the dataset post-transformation is about 0.234, which suggests a moderate level of data utility relative to the maximum possible score. Also, the 'Anonymous' property is set to 'ANONYMOUS', confirming that this particular transformation meets the criteria set by the selected privacy models. Basically, the utility is a bit low, but this means that privacy is high.

3.2 - K-anonymity + T-closeness

Regarding the combination of k-anonymity and t-closeness, the values applied to k and t are the ones specified before, where t is 0,6 for the three sensitive attributes. Figures 10 to 12 represent the outcome of the application of these privacy models:

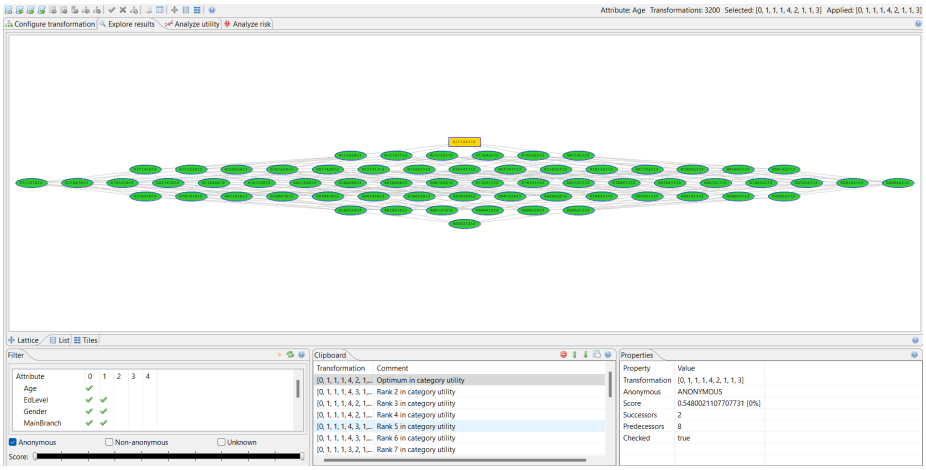

Figure 10: K-anonymity + T-closeness (1)

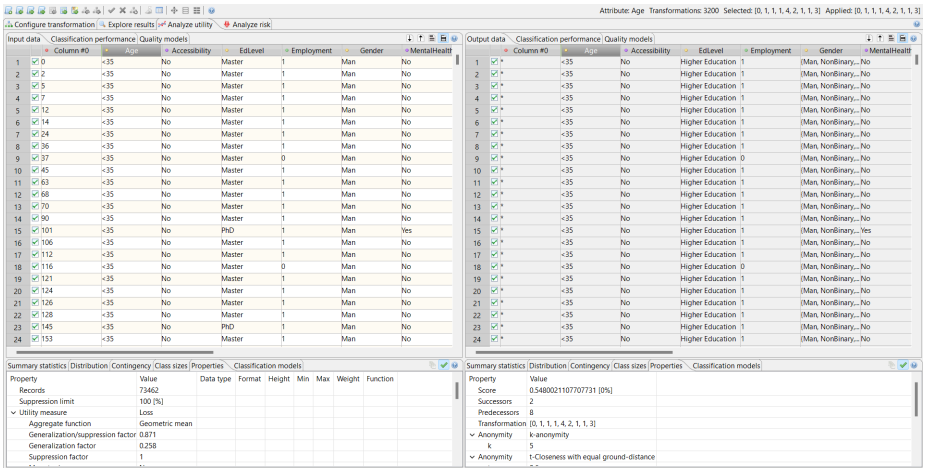

Figure 11: K-anonymity + T-closeness (2)

Figure 12: K-anonymity + T-closeness (3)

Again, better transformations could be created, but would end up matching the optimal transformation in terms of risk, like in the previous combination.

In this situation, some things should be said. Concerning the score in the properties pane which assesses the utility of the dataset post-transformation, it is about 0,5408. This score seems to have almost the same amount of privacy and utility, which is great, suggesting a good balance between data utility and privacy. Also, the properties panel indicates that the transformation is considered ”ANONYMOUS,” meeting the criteria for k-anonymity and t-closeness. Finally, the absence of any re-identification risk (0%) indicates that the current state of the dataset is considered to have a very low likelihood of re-identification under the applied privacy models.

3.3 - K-anonymity + L-diversity + β-Likeness

For the final combination, three privacy models are chosen, those being k-anonymity, l-diversity, and a new one, β-Likeness. The values of k, l, and β are the same, the difference is that l is applied to ”MentalHealth” and ”Accessibility”, and β is applied to ”PreviousSalary”.

Datasets often have skewed distributions of sensitive attributes, which l-diversity does not always handle well. β-Likeness helps to mitigate the risk by ensuring that the likelihood of inferring such commonly known sensitive values doesn’t increase significantly within any group of the dataset.

Figures 13 to 15 show the results of the application of these privacy models to the dataset:

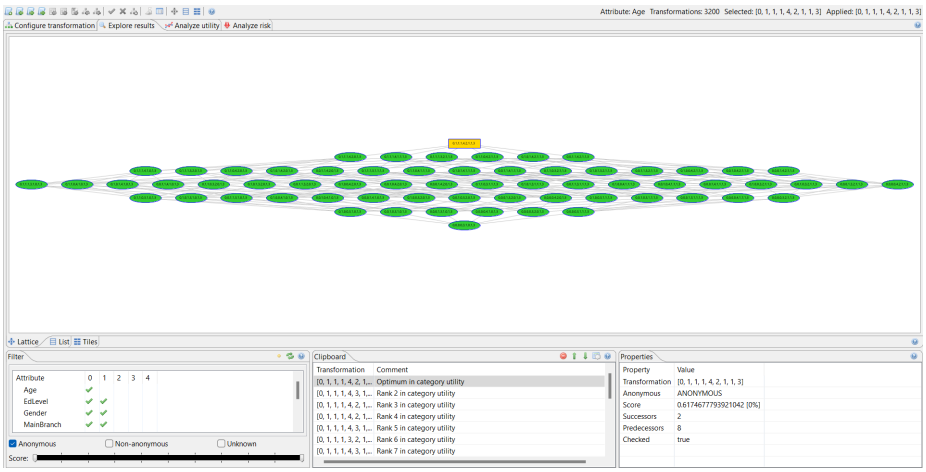

Figure 13: K-anonymity + L-diversity + β-Likeness (1)

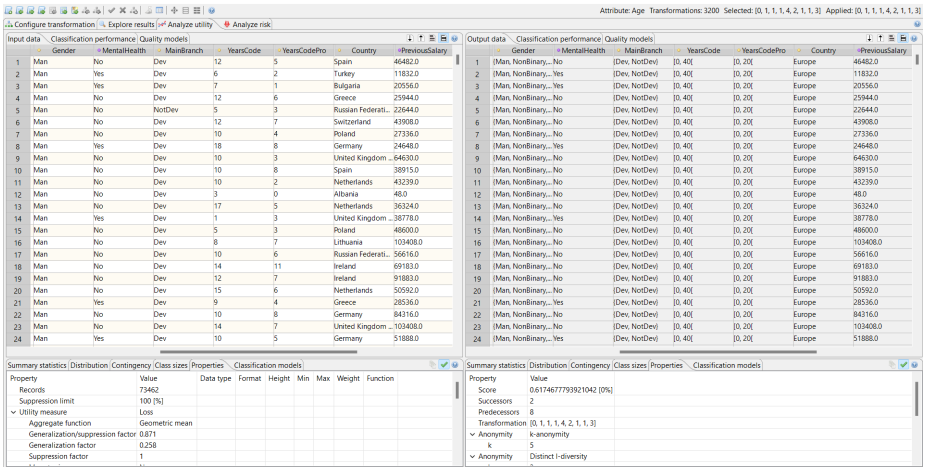

Figure 14: K-anonymity + L-diversity + β-Likeness (2)

Figure 15: K-anonymity + L-diversity + β-Likeness (3)

One more time, it should be mentioned that better transformations could be originated. Still, this scenario would end up being equal to the optimal transformation in terms of risk, like in the previous combinations.

The use of k-anonymity ensures that each record is indistinguishable from at least (k-1) other records. L-diversity enriches this by ensuring that sensitive attributes within these groups have at least 'l' diverse values. β-Likeness adds a probabilistic measure to ensure that the distribution of sensitive attributes within any group is not disproportionately different from the distribution of the attribute in the entire dataset. Analyzing the results, the utility score appears to be approximately 0,617, which indicates a more-than-moderate level of data utility after applying the transformations. This means that this combination of privacy models is the one that gives the anonymization of the dataset more utility, but not to the point of ignoring the privacy component. And again, the dataset at this transformation level is considered "ANONYMOUS," which means it has successfully met the anonymity criteria defined by the selected combination of privacy models.

4 - Analysis of Utility, Privacy and Risk Assessment

The final part of this assignment is the analysis of the utility, privacy, and risk factors. Starting with the utility part, it was mostly shown in Figures 5, 11, and 14. From what was reported in the previous sections, k-anonymity + l-diversity reached the highest score in terms of privacy, k-anonymity + t-closeness was the most balanced of the three, and k-anonymity + l-diversity + β-likeness got the highest score when it comes to utility. This suggests that adding β-likeness to a combination and dividing the sensitive attributes between different privacy models can probably contribute to improving the utility of the anonymization of the dataset.

As for the re-identification risk, it is the likelihood that de-identified data can be re-identified and linked back to an individual or group of individuals. It arises when there is still enough information in the de-identified dataset to allow someone to re-identify individuals.

To calculate the re-identification risk, it is required to determine QIDs Distinction and Separation. Distinction is the degree to which variables make records distinct, while Separation is the degree to which combinations of variables separate the records. In ARX, in the "analyze risk section", there is a tab for these values of the quasi-identifiers. The following Figures 16, 17, and 18 represent these values for the 3 combinations of privacy models made in section 3 of this work:

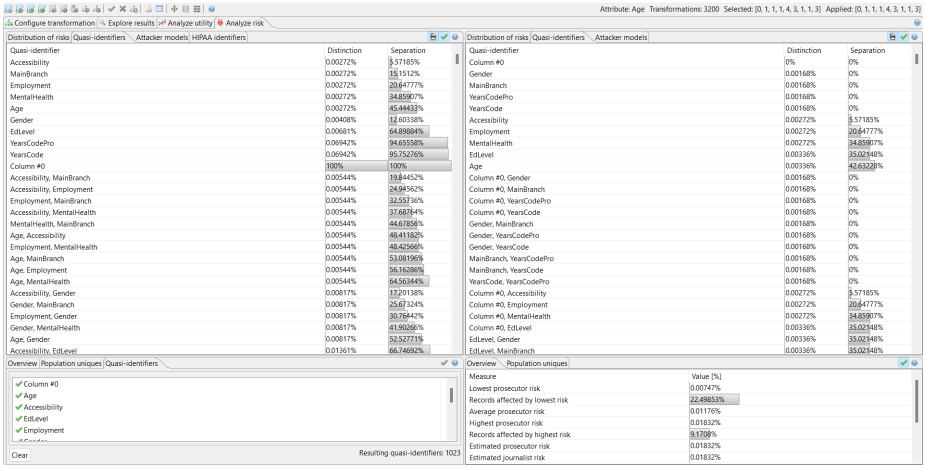

Figure 16: K-anonymity + L-diversity + β-Likeness QId's

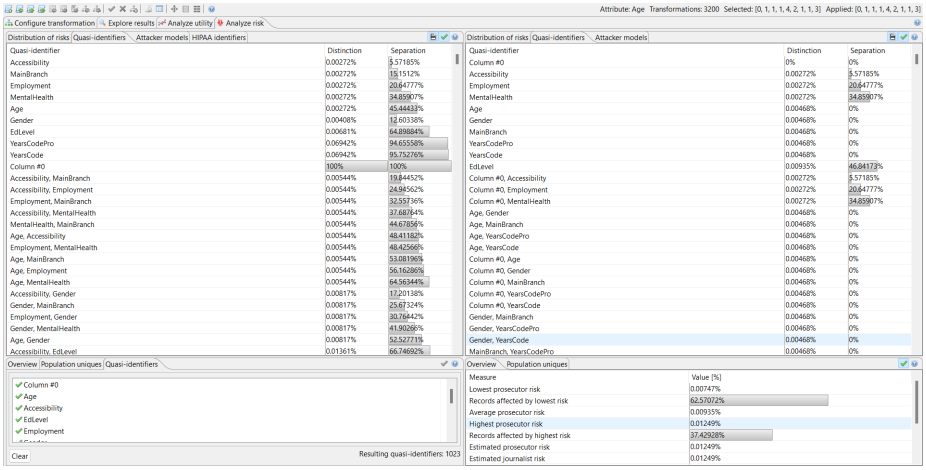

Figure 17: K-anonymity + L-diversity + β-Likeness QId's

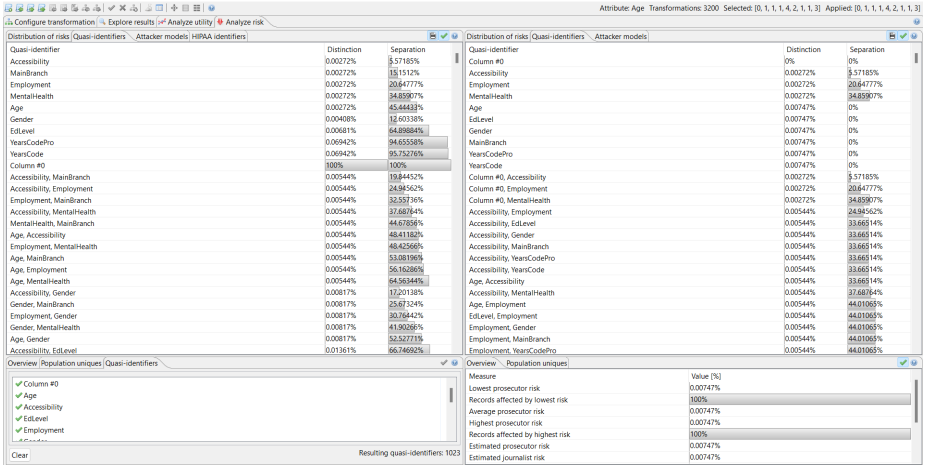

Figure 18: K-anonymity + L-diversity + β-Likeness QId's

From this ARX analysis, one can conclude that the best QId's following these metrics are "YearsCode", "YearsCodePro", "Column 0", "EdLevel", and "Age".

Moving on to risk, firstly, the different attack models should be considered. There are three types of models to consider:

Prosecutor - the attacker only has 1 individual as a target and knows that he is present in the dataset;

Journalist - the re-identification of any individual will benefit the attacker, which means every record is a target;

Marketer - in this model, the attack is considered successful only if a great number of people is re-identified. The more, the merrier.

From Figures 6, 12, and 15, one can gather the information presented in Table 10 below:

Attack Models – Risk Analysis							
Privacy Models	Prosecutor			Journalist			Marketer
	Records at Risk	Highest Risk	Success Rate	Records at Risk	Highest Risk	Success Rate	Success Rate
K + L	0	0,01832	0,01176	0	0,01832	0,01176	0,01176
K + T	0	0,01249	0,00935	0	0,01249	0,00935	0,00935
K + L + β	0	0,00747	0,00747	0	0,00747	0,00747	0,00747

Table 10: Attacker Models - Risk Analysis

Records at Risk - this represents the number of records in the dataset that are considered at risk of re-identification. For all combinations, this value is 0, which indicates that no records are directly at risk given the anonymization techniques applied.

Highest Risk - this value assesses the highest probability of re-identifying an individual within the dataset. For K + L, it is 0,01832; for K + T, it is 0,01249; and for K + L + β, it is 0,00747. These numbers indicate that the K + L + β combination has reduced the highest risk the most effectively compared to the other models.

Success Rate - this rate likely refers to the probability of successfully re-identifying an individual based on the attacker's background knowledge and capabilities. Similar to the highest risk, the values decrease from K + L to K + L + β, showing an improvement in privacy protection.

The fact that all models show a success rate for each attack model means that while risk is reduced, it is not eliminated. However, the success rate is consistent across all attack models for each privacy model. Overall, the combination of K + L + β offers the strongest protection for this dataset.

Finally, there is also the tab for "distribution of risks". Figures 19, 20, and 21 show those for the 3 respective combinations of privacy models made previously:

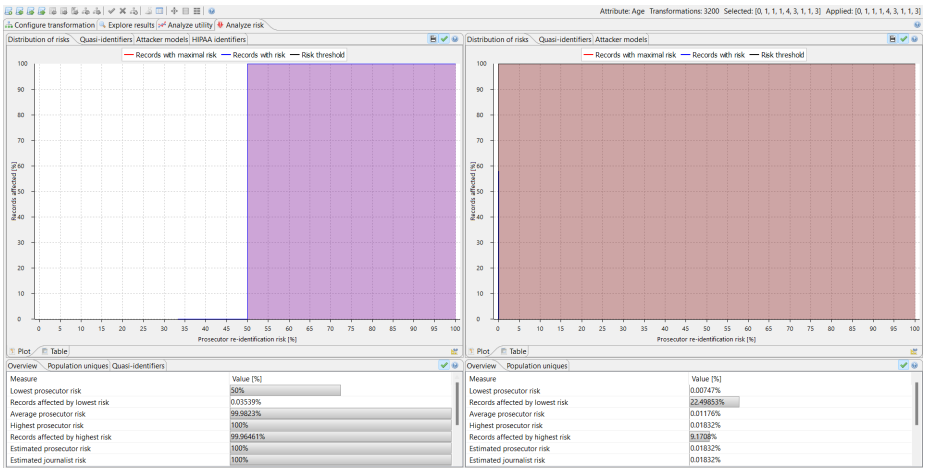

Figure 19: K-anonymity + L-diversity + β-Likeness D.O.R.

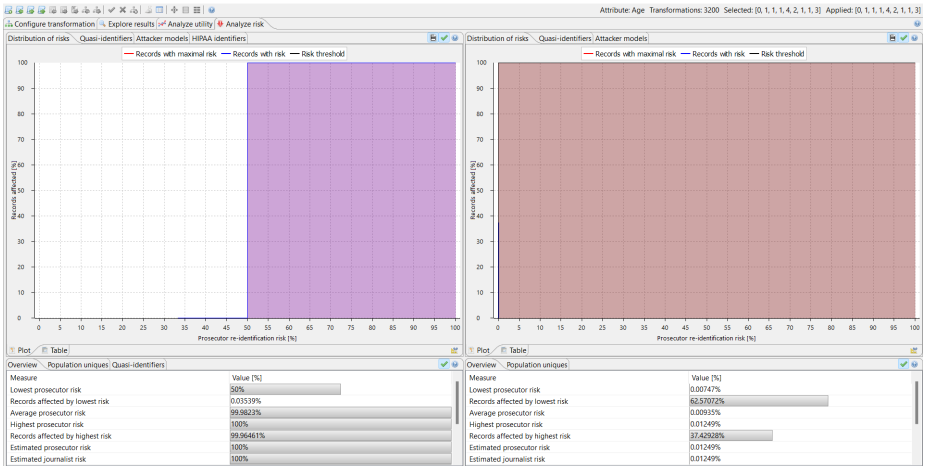

Figure 20: K-anonymity + L-diversity + β-Likeness D.O.R.

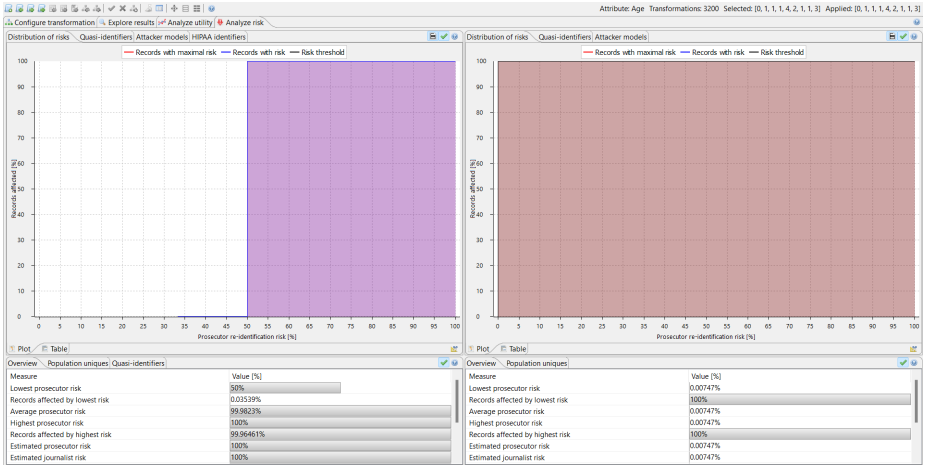

Figure 21: K-anonymity + L-diversity + β-Likeness D.O.R.

Conclusion

For this work, the process of anonymizing the dataset has been done with the objective of balancing privacy protection with data utility. Through a systematic approach involving dataset selection, characterization, and application of privacy models, significant insights and actions have been taken to mitigate privacy risks while preserving the dataset's utility.

The initial phase involved the classification of attributes into identifying, quasi-identifying, sensitive, and insensitive categories. This classification provided a comprehensive understanding of the dataset's composition and potential privacy vulnerabilities.

Also, hierarchies were constructed to generalize detailed information into broader categories, enhancing anonymity while retaining the utility of the dataset for analysis. Attribute weighting further refined the anonymization process by assigning weights to attributes based on their sensitivity and contribution to re-identification risks. This ensured that appropriate emphasis was placed on protecting sensitive information while balancing the need for accurate analysis.

Therefore, the application of privacy models such as K-Anonymity, L-Diversity, T-Closeness, and β-Likeness demonstrated a holistic approach to anonymization. Each model was selected based on its suitability to protect specific aspects of privacy within the dataset.

When checking the results obtained, it becomes clear why the anonymization process is so important, since it significantly reduces the risk of attacks by journalists, prosecutors, and marketers.

In conclusion, the anonymization process undertaken in this study exemplifies a robust framework for privacy-preserving data analysis. By integrating advanced privacy models with thoughtful parameter selection and rigorous analysis, the confidentiality of sensitive information has been secured.