FACULDADE DE CIÊNCIAS E TECNOLOGIA
UNIVERSIDADE D COIMBRA

departamento de engenharia informática
1995 – 2020

Project Report

Assignment 1: Anonymisation of Datasets with Privacy, Utility, and Risk Analysis – ARX Project

Curricular Unit:

Security and Privacy (SP)

Master's Degree in Data Science and Engineering

Mariana Lopes Paulino, 2020190448

Rui Alexandre Coelho Tapadinhas, 2018283200

2023/2024

Index

1. Introduction

1.1 Dataset

The chosen dataset contains information about the 2023 list of world billionaires. It has the most various details about them, from the rank they are in to their date of birth and the company that made them wealthy enough to be considered for this list.
This dataset was collected from Kaggle, an online platform that has millions of different datasets, and then it was cleaned and explored to the final form we are using for this project. Our cleaned dataset consists of 2640 entries and 35 attributes that categorize it in the most detailed way possible.

The range of these attributes is very wide. With this fact we decided to clean the dataset and take out the columns that were just filling even more space.
The whole dataset is composed by:

- rank
- finalWorth
- category
- personName
- age
- country
- city
- source
- industries
- countryOfCitizenship
- organization
- selfMade
- status
- gender
- birthDate
- lastName
- firstName
- title
- date
- state
- residenceStateRegion
- birthYear
- birthMonth
- birthDay
- cpi_country
- cpi_change_country
- gdp_country
- gross_tertiary_education_enrollment
- gross_primary_education_enrollment_country
- life_expectancy_country
- tax_revenue_country_country
- total_tax_rate_country
- population_country
- latitude_country
- longitude_country

Upon a more in-depth examination of these attributes, it becomes evident that there is a significant amount of redundant information distributed across various columns. For instance, the billionaire's name is unnecessarily split into first and last names, which does not add any value to the dataset and our analysis. Consequently, we removed all columns that did not provide meaningful insights.

1.2 Cleaned Dataset

After the cleaning process, the cleaned dataset ended up with only 12 columns, being those essential to discover the identity of the billionaire.

- rank
- finalWorth
- category
- personName
- age
- city
- source
- countryOfCitizenship
- selfMade
- status
- gender
- birthDate

In the following table a list of the attributes is presented, and contains a little description of every column that is still considered for this assignment along with its data type.

Attribute	Description	Type
rank	The rank between #1 and #2640 the person is based on their worth in billions, being number 1 the richest.	int
finalWorthInMillions	The final worth of the billionaire in millions of US Dollars.	int
companyCategory	The category of the company associated with the wealth of the billionaire (ex. Technology, Automotive, Food & Beverage, …).	string
personName	The first and last name of the billionaire.	string
age	The age of the person, as of 2023.	int
city	The city where the company is based.	string
source	The name of the company that the billionaire is linked to.	string
countryOfCitizenship	The country of citizenship of the billionaire	string
selfMade	Value that indicates if the billionaire was self-made or not.	boolean
inherited	This column contains the information needed to determine if the wealth was inherited or not or if it is Unknown.	string
gender	The gender of the person (Female or Male).	string
birthDate	The birthdate of the billionaire in the format Year/Month/Day.	datetime series

Figure 1- Description of the Cleaned Dataset

The dataset used in this project is a mirror image of the list of the world's richest people. This means that although we are very likely to know at least one name on it, it does not mean that the person would like to have their net worth all over the internet. So our final goal in this project is to anonymise this dataset and do a risk analysis on it after implementing some privacy models.

After a short analysis to our clean and final version of the dataset we could still check that some information was missing, and some entries had incomplete information and empty columns. So, after a thoughtful analysis, the conclusion we arrived was that if we took them off and dropped them, a lot of information would be missing. Therefore, we made the decision to keep those entries in order to not compromise the information available.

2. Attribute Classification

The first step necessary for this assignment was the categorization of the attributes in categories, they could be categorized in four different types (*Insensitive*, *Sensitive*, *Quasi-Identifier* or *Identifier*).

In the Insensitive group, there should only be information that is not relevant or can identify the billionaire in our project by any means nor association with other attributes.
Sensitive attributes have information that belongs to an individual, and he doesn't want to make it public. Having the possibility of that information being harmful or some private detail the person doesn't want to share with others and make it public.
Quasi-Identifiers are the columns that contain pieces of information that alone cannot identify a person by itself but when combined with more Quasi Identifiers can determine the identity of a person.
Identifiers are the type of data that we want to protect from attackers, like names or Social Security numbers, private details that are unique and identify a single individual.

The presence of these categories in the chosen dataset is clear. After looking at the columns we have available we can see that there are columns that identify the individual, some that can be combined and arranged to identify the person. And, some columns are just information that the person doesn’t really want to be public.

Attribute	Category
rank	Sensitive Attribute
finalWorthInMillions	Sensitive Attribute
companyCategory	Quasi Identifier
personName	Identifier
age	Quasi Identifier
city	Quasi Identifier
source	Quasi Identifier
countryOfCitizenship	Quasi Identifier
selfMade	Sensitive Attribute
inherited	Sensitive Attribute
gender	Quasi Identifier
birthDate	Quasi Identifier

Figure 2 - Categories applied to the attributes

2.1 Justification of the Categories Chosen

In order to better justify the reason that lead us to the categories as shown above in the table we are going to present them one by one following the reason.

- **rank** - ***Sensitive Attribute*** - The rank is considered to be a sensitive attribute because the individual that it is connected to may not want this information to be public,

- **finalWorthInMillions** - ***Sensitive Attribute*** - As said above for the rank attribute, the individual this information may concern may not want this information to be made public,

- **companyCategory** - ***Quasi Identifier*** - The company category is classified as a Quasi Identifier because this information linked with any other that is also classified as a Quasi Identifier can lead to the discovery of the identity of the person,

- **personName** - ***Identifier*** - The name of the individual is an identifier because even alone it identifies one person only, this information is going to be taken out of the assignment since the objective is to protect data,

- **age** - ***Quasi Identifier*** - The age when looked up with other Quasi Identifiers can lead to the discovery of the name of the billionaire, so it is considered a Quasi Identifier,

- **city** - ***Quasi Identifier*** - The city when looked up with other Quasi Identifiers can lead to the discovery of the name of the billionaire, so it is considered a Quasi Identifier,

- **Source - Quasi Identifier -** The source of income, or more in depth the name of the company the billionaire belongs to. When looked up with other Quasi Identifiers, it can lead to the discovery of the name of the billionaire, so it is considered a Quasi Identifier,

- **countryOfCitizenship** - ***Quasi Identifier*** - The country of citizenship when looked up with other Quasi Identifiers can lead to the discovery of the name of the billionaire, so it is considered a Quasi Identifier,

- **selfMade** - ***Sensitive Attribute*** - The self-made attribute is considered as only sensitive information because it only says if the person is an entrepreneur and worked for his/her fortune or not. This information is something that the person does not really want to see public, but it does not really help to disclosure the identity of the billionaire,

- **inherited** - ***Sensitive Attribute*** - The inherited attribute only tells the attackers if the person inherited their money or not, this does not quite help to reveal the identity of

the person. However, it is an information the person does not really want to share with others,

- **gender - *Quasi Identifier* -** The gender of the billionaire is not a big help to discover the identity of the person. The number of billionaires present is mostly classified as Male, but when searching for a Female it may be different. Besides this fact, also if we make the right combination of Quasi Identifiers, we can discover more than the identity of one individual.

- **birthDate** - ***Quasi Identifier*** - The birthdate is considered a Quasi Identifier as well because when combined with other Quasi Identifiers can help to disclosure the identity of the billionaire.

The division of the attributes into categories is one of the steps that really influences if the anonymisation of the data will be made correctly. Our main difficulty was to separate the Sensitive Information of the Quasi Identifiers. But when applied to the chosen dataset, we know for sure that the names it contains are public, and it contains information about public figures that whether they like it or not, the information can help to identify their names.

To sum up the work made in this section, the separation of the attributes into classes was a really important factor for the assignment. The choice between Sensitive and Quasi Identifier was also made because the information that these attributes carry is considered as Quasi Identifier. If the dataset was not about famous people, these attributes could be only considered as Sensitive information.

2.2 Distinction and Separation Analysis

For a better analysis and a cleaner one also, we decided to take a further look into just the Quasi Identifiers alone, and the combinations of two of the Quasi Identifiers. Here we can see the values for the Distinction and Separation.

Quasi-identifier	Distinction	Separation
gender	0.07576%	22.27976%
companyCategory	0.68182%	91.29804%
countryOfCitizenship	2.91667%	87.48768%
age	3.0303%	97.93408%
city	28.10606%	99.11376%
source	34.31818%	99.03157%
birthDate	78.06818%	99.86709%
companyCategory, gender	1.32576%	93.16118%
countryOfCitizenship, gender	4.43182%	90.19429%
age, gender	5.5303%	98.3757%
companyCategory, countryOfCitizenship	16.9697%	98.36959%
companyCategory, age	29.92424%	99.80677%
city, gender	32.87879%	99.29418%
city, countryOfCitizenship	34.88636%	99.36429%
age, countryOfCitizenship	37.19697%	99.67817%
source, gender	39.35606%	99.23614%
companyCategory, source	39.65909%	99.1427%
companyCategory, city	53.56061%	99.83476%
source, countryOfCitizenship	53.93939%	99.81389%
city, source	76.28788%	99.95304%
age, birthDate	78.10606%	99.88913%
age, city	79.96212%	99.97459%
gender, birthDate	80.41667%	99.92114%
age, source	81.85606%	99.97537%
countryOfCitizenship, birthDate	90.75758%	99.97086%
companyCategory, birthDate	92.31061%	99.98694%
city, birthDate	97.12121%	99.9948%
source, birthDate	98.14394%	99.99833%

Figure 3 - Distinction and Separation Values for the Quasi Identifiers and their Combinations

After a closer analysis into the table above, all the Quasi Identifier attributes have a very high Separation level but the gender. It is the only attribute who doesn't have a lot of different values, it only has two possible values, so, the separation can be very high. The Distinction value indicates that the attribute it refers to can be used to identify a group of individuals with similar characteristics. In other words, it can not identify a single person but can identify a group of people when the value is high. When put together the values for Distinction and Separation come up to close to 100% meaning that these are good Quasi Identifiers since these attributes can be a risk to the privacy.

3. Privacy Risks

The chosen dataset contains information about the richest people on the planet, and some information is sensitive. Even knowing if someone is present or not in the dataset is sensitive. Knowing if a person is in the dataset means that they have a very high worth or their assets, as companies or investments, are worth a lot.

Any leak of information would be a privacy violation if the data leaked was not public yet, and would be an asset for an attacker to better choose his target in identity theft, intelligence selling, stalking or blackmailing.

3.1 Risk Analysis

In the ARX software, we can analyse three types of attacks:

- **Prosecutor attacker model:**
  - Has a specific target and assumes that the target's data is contained in the dataset. So it only needs to try to identify which record belongs to the target by combining the dataset Quasi-Identifiers to the information the attacker already has from other sources.
- **Journalist attacker model:**
  - Also has a specific target, but the attacker doesn't know if the target's data is in the dataset. This attacker uses the information he already knows from another sources and joins them with the dataset and tries to find sensitive information with the records that matches the previous known data.
- **Marketer attacker model:**
  - The attacker's goal is to re-identify / de-anonymise a large portion of records of the dataset. This type of attack can occur using statistical analysis, machine learning and external data sources in order to find the identities of the records in the anonymised dataset.

Figure 4. Risk according to each attacker model

According to the analysis done by the ARX software, we can see that our cleaned dataset has a very high risk for any type of attack. Being all the values for the "Records at risk", "Highest risk" and "Success rate" of each attack type 100% or very close to that.

So, by looking at these values, we can reach the conclusion that the cleaned dataset needs to be protected by anonymising the dataset.
Since our goal is to reduce the risk, we have established a maximum risk rate for each type of attack to be 10% or less to consider our anonymisation models successful while keeping the record suppression to a maximum of 25%.

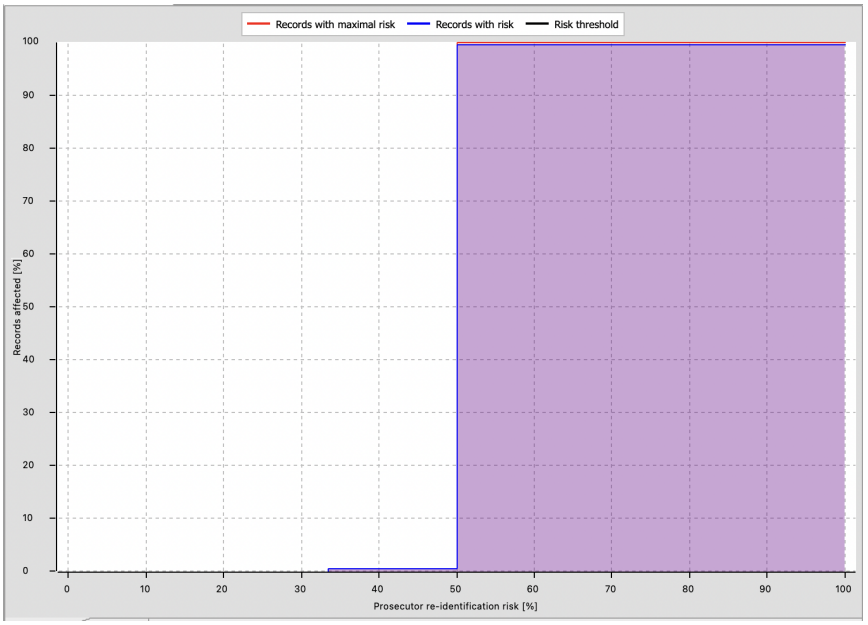

Figure 5. Risk distribution on the cleaned dataset

In this figure, we can see the distribution of the risk according to the percentage of the records. And we can notice that 1% of the records have between 33% and 50% of risk, and the remaining 99% of the records have a risk of more than 50%.

4. Hierarchies used for *Quasi Identifiers*

In order to apply the privacy models, we need to hierarchize all the Quasi Identifiers, to reach our Generalization goal.

4.1 Company Category, City, Source, Country of Citizenship, Gender, Birth Date

For this attribute, the chosen hierarchy was a priority hierarchy by frequency, more specifically a prioritization by frequency from highest to lowest, with 10 levels, where in each level the categories with the least occurrences are anonymised.

4.2 Age

In the age column, we also applied the priority hierarchy with 10 levels. In each level the ages the most occur in the dataset are anonymised until it reaches to level 10 where all the entries are anonymised and take the value (*). In this column there is another detail we would like to highlight, being it the fact that there is age 0, which was imputed for the people that didn't have data about their age.

As specified in the subsections above, the hierarchisations used were only the priority type because it was the one type that gave us the best results for all of our models.

5. Attribute Weights

As we studied our dataset, we could quickly understand that the attributes have different weights when trying to discover the billionaire.
Although the dataset has information that is public and can be accessed by anyone, we are treating it the way we would like our data to be treated. For example, the age of the person isn't quite important as the source of income to disclosure the identity of the billionaire.

So, in the figure below, we display the weights we consider to be more reasonable for this analysis. The attributes that have a weight equal or bigger than 0.5 are the ones that help the most when trying to discover the identities.

The city where the company is based, or the birthdate or even the country of citizenship of the billionaire can be a little more difficult to discover so, those attributes should way less in this analysis.

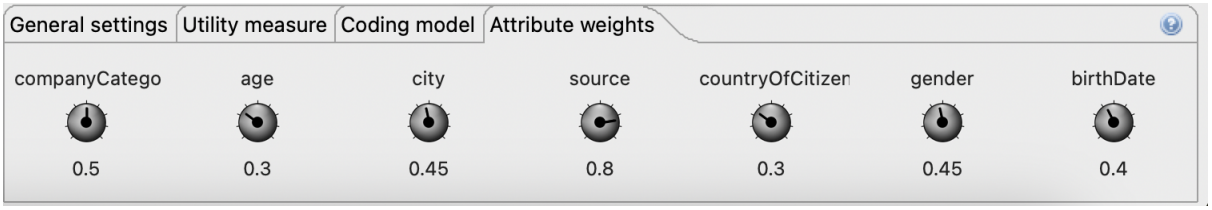

Figure 6 - Weights considered to the columns

6. Privacy Models - Definition

6.1 K-Anonymity

For a dataset to be considered k-anonymous, it needs to have k-1 records with the same Quasi-Identifiers for each record, and this condition can be achieved by 2 methods: Generalization or Suppression.

By **generalizing**, we mean reducing the specificity of an attribute, by, for example, joining 2 distinct ages in only one age group, thereby making a bigger group in that attribute of the dataset and diminishing the specificity of the attribute.

And, by **suppression**, in which a part of the information is hidden, by, for example, instead of showing a postal code of 5 digits, we only show the first 3 digits and don't disclosure the last 2 digits, just like (12345 becomes 123**), thus, increasing the size of the group of records with that same value for the postal code attribute, also diminishing the attribute specificity.

6.2 L-Diversity

L-Diversity, similarly with K-Anonymity, also ensures that each record has multiple records with the same QIDs, but also contains at least *L* different values for the sensitive attributes among the records that share the same QIDs. This technique ensures that the dataset has higher homogeneity in the sensitive attributes, making it much more difficult for an attack to infer any sensitive information about a record/person even if it has all his QID's.

An example that can illustrate this technique is when a dataset in the *finalWorth* attribute, which is classified as sensible, has 3 distinct values (1200, 1700 and 3500) inside a group of records that has the same QIDs, therefore ensuring the dataset has a 3-Diversity classification.

6.3 T-Closeness

T-Closeness, goes beyond L-Diveristy, since, besides it ensuring it has L different values in a sensitive attribute, it also ensures that the proportions of the distribution of each value of a sensitive attribute inside a group with the same QIDs, is also kept similar when comparing to the distribution of values in the whole dataset. Being *T* the value that represents the delta between the distributions in equal QIDs groups and the entire dataset.

For example, supposing the *countryOfCitizenship* has a distribution of 50% (USA), 25% (France) and 25% (Germany). In a group of 4 records with the same QIDs, there would be 2 records with the *countryOfCitizenship* of the USA, 1 with France and 1 with Germany, thus maintaining similar proportions in that attribute.

So, by keeping the distribution proportions of this attribute, the statistical analysis would be very similar between the cleaned dataset and the anonymised dataset.

7. Privacy Models - Application

7.1 First Model - k-Anonymity + l-Diversity

In our first attempt to get the best results, we chose the k-Anonymity and l-Diversity to be our models and try to achieve the best results possible.

Attribute	Privacy Model
4-Anonymity	
finalWorthInMillions	Distinct 219-Diversity
inherited	Distinct 3-Diversity
rank	Distinct 219-Diversity
selfMade	Distinct 2-Diversity

Figure 7 - Privacy Models applied in the 1st Attempt

We considered a 10-Anonimity model along with a Distinct 5-Diversity for the attribute finalWorthInMillions, a Distinct 3-Diversity for the attribute inherited, a Distinct 5-Diversity for the rank and finally a Distinct 2-Diversity for the selfMade attribute.
For the inherited attribute it was simple to discover a number that would suit this attribute the best once it only has 3 unique values. As explained for the inherited attribute we followed the same line of thought. Since the attribute only has 2 unique values the l for the l-Diversity is going to be fixed in 2. For the other two columns, their values are correlated and, they have 219 unique values so, we decided to use that as the l value for the l-Diversity applied to those columns. The value for the k-Anonymity was selected by the count of Sensitive attributes present in this dataset, in this case, 4.

After we correctly entered our data in ARX and inputted our model properly. We decided that our Suppression Limit should be less than 50% for this attempt and the coding model considered should be the one presented in figure 8.

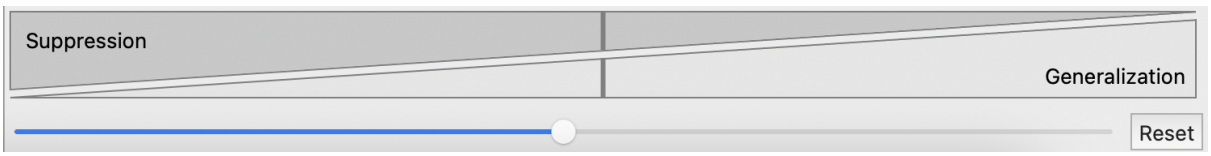

Figure 8 - Coding Model for the 1st Model (1st Iteration)

After waiting for the anonymisation, figure 9 presents the risk present after the dataset was anonymised. By this graphic, we can check that we lost all of our information and, we suppressed all the records.

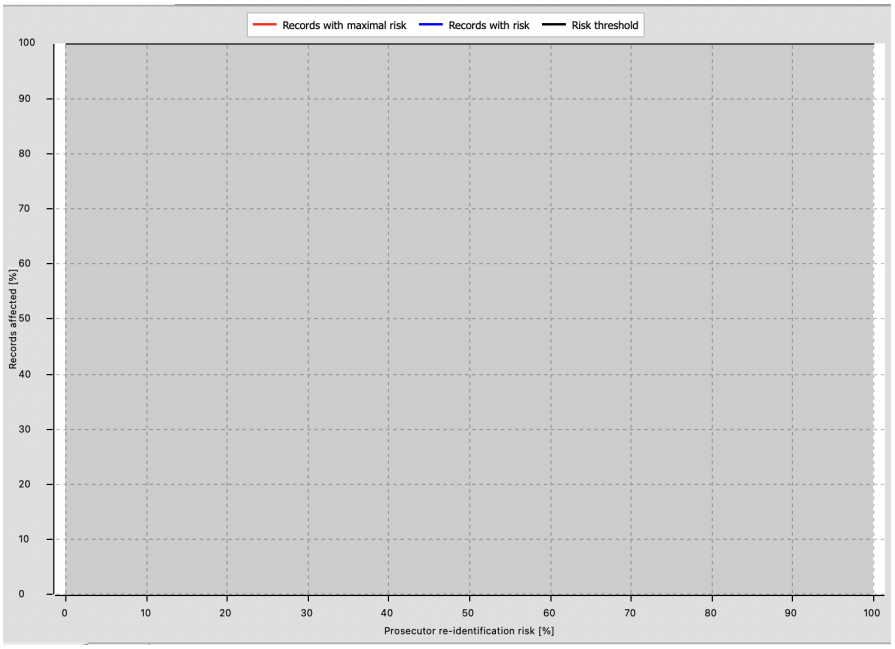

Figure 9 - Risk after the anonymisation process

After this failed iteration we decided to cut the l-Diversity of the rank and finalWorthInMillions to 5, a lot lower number that could still cover all the dataset. Because our dataset is a little small for those big numbers we decided to lower it a lot so, we could get better results.
Having Figure 10 representing the values we used for the second iteration.

Attribute	Privacy Model
4-Anonymity	
finalWorthInMillions	Distinct 5-Diversity
inherited	Distinct 3-Diversity
rank	Distinct 5-Diversity
selfMade	Distinct 2-Diversity

Figure 10 - Table that represents the values used for the 2nd iteration

After the application of this iteration we got a lot more interesting results with values and that could be considered.
In figure 11, it is presented the risk after the anonymisation process for the second iteration of this model.

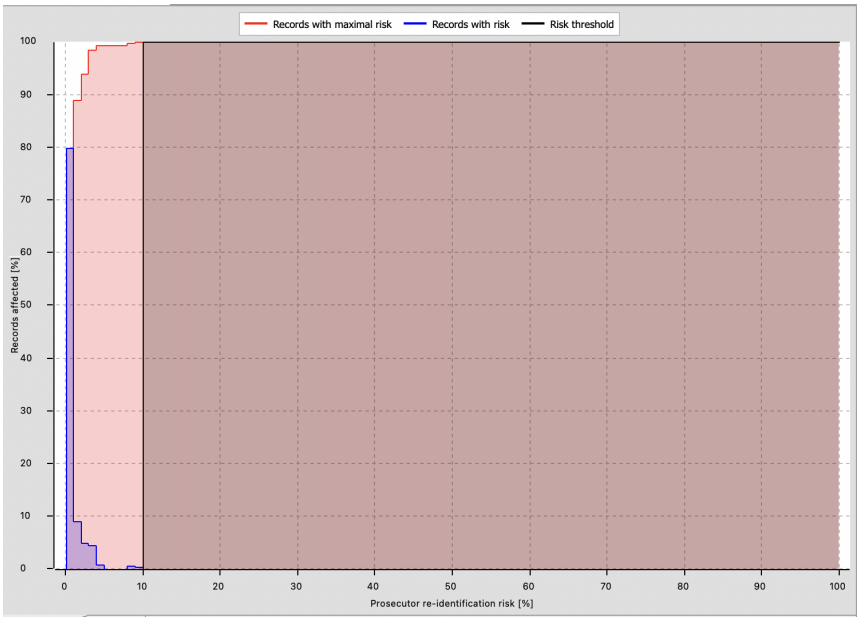

Figure 11 - Risk after anonymisation

With these values for the l-Diversity and k-Anonymity we came to the following results:

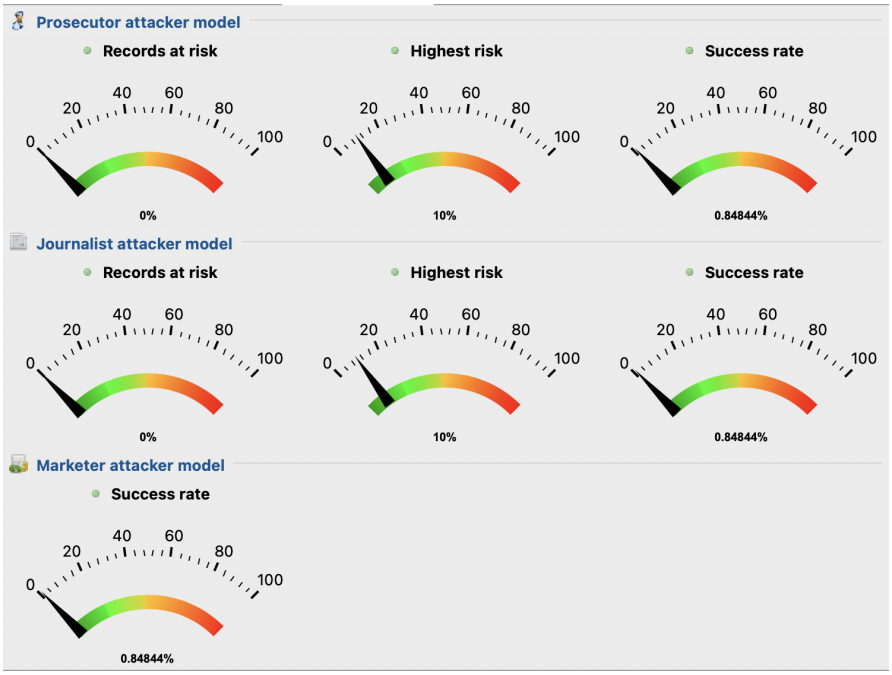

Figure 12 - Attacker Model Risks for the chosen model

Although it is not a perfect result, the result of the application of this second iteration model should be considered as a possibility for this dataset since the risk is only 10%. Despite the possibility to achieve better, the suppressed records are at only 1.78% which is very useful because the suppressed records are the records we lose when anonymising the dataset.

7.2 Second Model - l-Diversity + t-Closeness

In our second attempt to get the best results, we chose the l-Diversity and t-Closeness to be our models and try to achieve the best results possible.

For this model, we decided to apply the same number of l for the l-Diversity but making it go up to 10 because we are taking out a model, in this case, k-Anonymity. And, by adding the t-Closeness with a value of 0.25 we are trying to protect the information of the finalWorthInMillions and the rank even more.

Attribute	Privacy Model
finalWorthInMillions	Distinct 10-Diversity
finalWorthInMillions	0.25-Closeness
inherited	Distinct 3-Diversity
rank	Distinct 10-Diversity
rank	0.25-Closeness
selfMade	Distinct 2-Diversity

Figure 13 - Privacy Models applied in the 1st Attempt for the 2nd Model

After this application we get the following results:

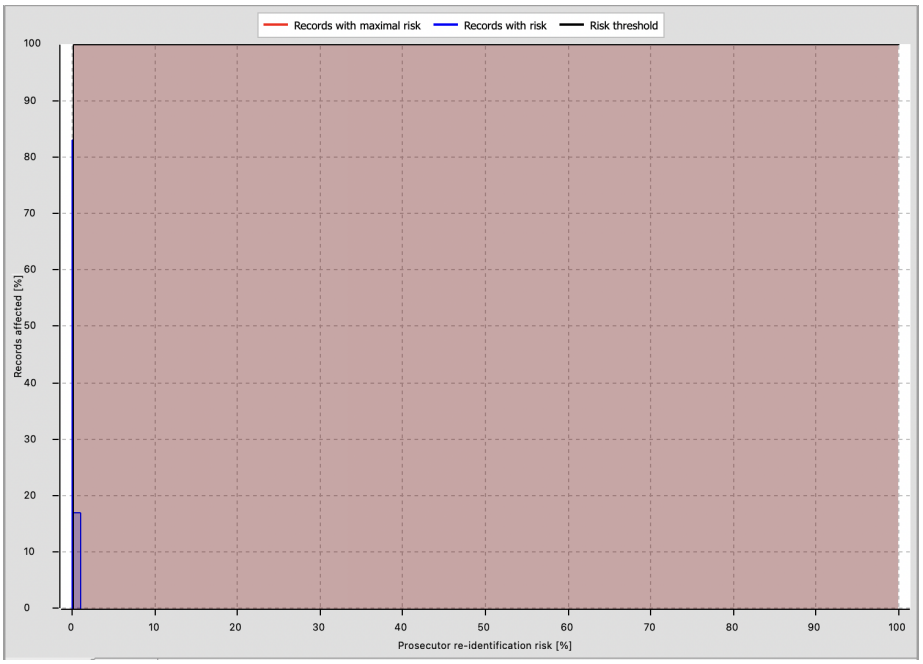

Figure 14 - Risk present in the dataset after the anonymisation process

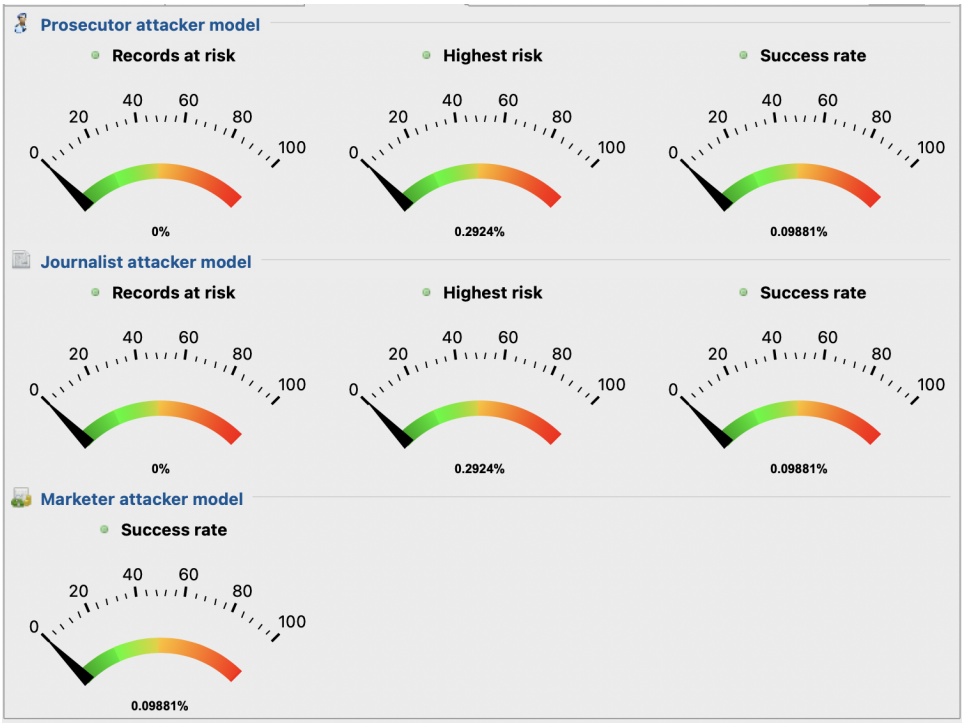

Figure 15 - Risks associated with the 2nd model applied

With this model we achieve much lower results of risk, being the highest journalist and prosecutor risk 0.29%. However, this model has 23.33% of suppressed data which is a bit higher than supposed and wanted value.

Despite the suppression level we decided that this was a good model, since the risk is very low and the lost suppressed data is still below our threshold to decide if this model is good or not.

7.3 Third Model - k-Anonymity + l-Diversity + t-Closeness

For the application of the third model we decided to combine the two models above and make some improvements so that we could achieve a lower risk like model two with the lower value of suppressed records like the first one.

The first improvement we decided to analyse was the change in the Coding Model, which before was divided in half for 50% Suppression and 50% Generalization. Now, for the third model we are going to try to have the Generalization up to 100% and see if this change makes any difference.

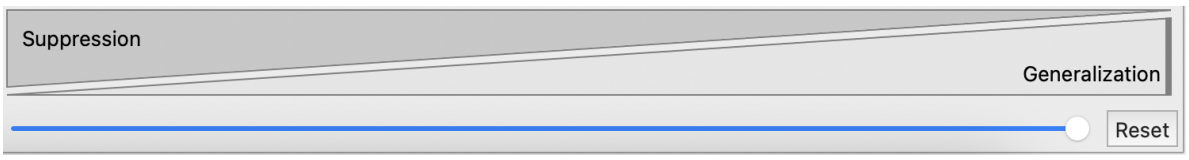

Figure 16 - Coding Model applied to the 3rd Model

Attribute	Privacy Model
100-Anonymity	
finalWorthInMillions	Distinct 10-Diversity
finalWorthInMillions	0.25-Closeness
inherited	Distinct 3-Diversity
rank	Distinct 10-Diversity
rank	0.25-Closeness
selfMade	Distinct 2-Diversity

Figure 17 - Table of the Privacy Models applied

Above, in the figure 17 that represents a table, we included all the models we have applied to the dataset, being a 100-Anonymity model, with a Distinctive 10-Diversity and 0.25-Closeness applied to both the rank attribute and the finalWorthInMillions attribute.
Also, there was a Distinct 3-Diversity applied to the inherited characteristic and a Distinct 2-Diversity to the selfMade attribute that helped us to get the best results possible in Model 1.

The explanation behind the k-Anonymity having k equalling 100 is because if the value of k were to be bigger then, the dataset would be all suppressed and 100 is the number that best minimizes risk while not suppressing all the records present in this dataset.

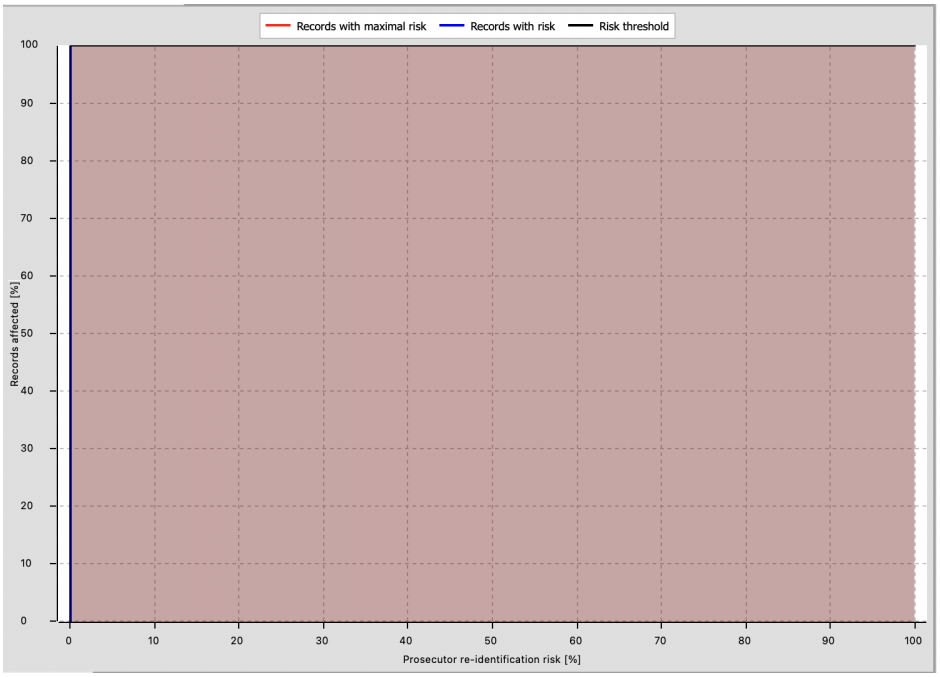

Figure 18 - Risk associated with the 3rd model

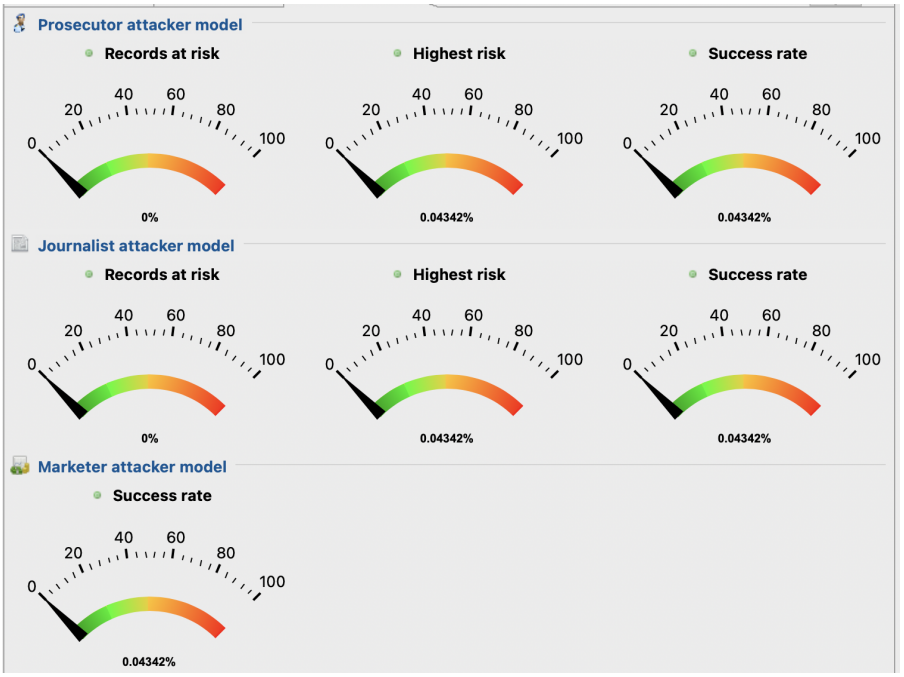

Figure 19 - Risks associated with the implemented model

In figure 18 we have a graphic that shows the risk associated with the application of this third model, and in figure 19 the risks associated with this third model as well are demonstrated.
For the third model, we were able to obtain a combination between the lowest risk possible and the lowest suppression rate possible as well. In this model, we could obtain 0.043% of highest Marketer and Prosecutor and Journalist risk. In terms of suppression, the best result we could get to was 12.76% of suppressed data.

8. Model Comparison vs Original Dataset

	Prosecutor Attacker Model			Journalist Attacker Model			Marketer Attacker Model	Suppression
	Records at Risk	Highest Risk	Success Rate	Records at Risk	Highest Risk	Success Rate	Success Rate	
Model Comparison								
k-Anonymity + l-Diversity (1)	0%	10%	0.85%	0%	10%	0.85%	0.85%	1.78%
l-Diversity + t-Closeness (2)	0%	0.29%	0.10%	0%	0.29%	0.10%	0.10%	23.33%
k-Anonymity + l-Diversity + t-Closeness (3)	0%	0.04%	0.04%	0%	0.04%	0.04%	0.04%	12.76%
Original Dataset								
-	100%	100%	99.73%	100%	100%	99.73%	99.73%	0%

Figure 20 - Comparison of the Attacker Models Risks and Suppressed Information on the Dataset with various models applied and the Original one

9. Conclusion

Throughout this project, we have analysed the performance of different data anonymisation techniques. Although some anonymise well, i.e. increase privacy, they sometimes cause relevant information to be lost for analysis.

As a result, we realised that we have to find a good balance between anonymisation and usability of the dataset. From a privacy point of view, the more anonymised the data, the better. However, the more anonymised the dataset is, the less useful the data becomes.

By applying the models mentioned above, we could experience and test three distinct privacy models and perceive the compromise between anonymisation and data suppression. So, to choose which model to apply, we need to consider the nature of the information on the dataset, and the type of analysis that is going to be done after the anonymisation process.

10. References

- https://arx.deidentifier.org/
- PowerPoints shared by the Professor in the UC Student Platform (Theoretical and Practical)