



FACULTY OF SCIENCES OF THE
UNIVERSITY OF PORTO

MASTER'S DEGREE IN INFORMATION SECURITY

Privacy Enhancing Technologies

Anonymization of Datasets with Privacy, Utility and
Risk Analysis

David Capela Rui Fernandes

December, 2021

Abstract

This report describes the first assignment of the Privacy Enhancing Technologies course within the context of the Master's Degree in Information Security at the Faculty of Sciences of the University of Porto.

This assignment aims at anonymizing a dataset containing personal information of individuals while maintaining a reasonable level of data utility. To do so, several privacy models were applied, followed by a careful evaluation of the re-identification risk and utility metrics as to ensure that an acceptable trade-off between privacy protection and utility was achieved considering the requirements we defined.

Keywords: Data Anonymization, Privacy Preserving Data Publishing, Privacy Models, Privacy-Utility Trade-off

Contents

Acronyms	v
1 Introduction	1
2 Selection, Importing & Goal of Dataset	2
2.1 Privacy & Utility Requirements	2
2.2 Data Preprocessing	2
3 Characterization of the Dataset & Coding Models	4
3.1 Distribution of Data for Each Attribute	6
3.2 Classification of Attributes	11
3.3 Risk Analysis	14
4 Privacy Models	15
4.1 Model 1: k -Anonymity & ℓ -Diversity	15
4.1.1 Results	17
4.1.2 Risk Analysis & Data Utility	18
4.2 Model 2: k -anonymity and ℓ -diversity with attribute weights	19
4.2.1 Results	20
4.2.2 Risk Analysis & Data Utility	21
4.3 Model 3: t -Closeness and k -Anonymity	22
4.3.1 Results	22
4.3.2 Risk Analysis & Data Utility	23
4.4 Comparative Analysis	24
4.4.1 The Privacy-Utility Trade-off	26
5 Conclusion	27
A Data Generation using Faker	29
B Generalization of the City attribute	30

List of Figures

1	Importing data into ARX	5
2	Excerpt of the dataset	5
3	Summary statistics for the Name attribute	6
4	Summary statistics for the SSN attribute	6
5	Summary statistics for the Gender attribute	6
6	Distribution of values for the Gender attribute	7
7	Summary statistics for the City attribute	7
8	Summary statistics for the Date of Birth attribute	7
9	Summary statistics for the Race/Ethnicity attribute	8
10	Distribution of values for the Race/Ethnicity attribute	8
11	Summary statistics for the Parental Level of Education attribute	8
12	Distribution of values for the Parental Level of Education attribute	9
13	Summary statistics for the Test Preparation Course attribute	9
14	Distribution of values for the Test Preparation Course attribute	9
15	Summary statistics for the Average Score attribute	10
16	Distribution of values for the Average Score attribute	10
17	Identification of QIDs through ARX (i)	12
18	Identification of QIDs through ARX (ii)	13
19	Risk of re-identification on each attacker model	14
20	Gender transformation levels	16
21	City transformation levels	16
22	Generalization hierarchy for the Birth Date attribute	16
23	Excerpt of the dataset anonymized with Model 1	17
24	Risk associated with each attacker model considering Model 1	18
25	Supression limit for Model 2	19
26	Coding model for Model 2	19
27	Attribute weights for Model 2	19
28	Excerpt of the dataset anonymized with Model 2	20
29	Risk for Model 2	21
30	Anonymized dataset for Model 3	22
31	Risk analysis for Model 3	23
32	Trade-off between <i>perfect utility</i> and <i>perfect privacy</i> [1]	26

List of Tables

1	Characterization of the original dataset	3
2	Characterization of the original dataset after introducing synthetic data . . .	3
3	Possible values and type for each attribute	4
4	Transformations applied to each of the QID	16
5	General results after anonymizing with Model 1	17
6	General results after anonymizing with Model 2	20
7	General results after anonymizing with Model 3	22
8	Comparative analysis of the data utility	24
9	Comparative analysis of the re-identification risk	24

Listings

1	Python script used to generate random synthetic data	29
2	Python script used to replace each city with the corresponding State	30

Acronyms

PII Personally Identifiable Information

PPDDM Privacy Preserving Distributed Data Mining

PPDM Privacy Preserving Data Mining

PPDP Privacy Preserving Data Publishing

QID Quasi-Identifier

1 Introduction

Often times, one wants to release data to the public for a variety of reasons – *e.g.* to compute some statistics or to perform other data analysis tasks. However, prior to anonymization, data usually contains sensitive information about individuals and sharing it could potentially violate their privacy. Privacy Preserving Data Publishing (PPDP) addresses this issue by providing mechanisms to ensure privacy protection while maintaining the utility of data.

In this assignment, we perform a detailed analysis of the anonymization process of a dataset. To do so, we first start by choosing the dataset to anonymize, define the privacy and utility requirements, apply different privacy models and, finally, compare the privacy protection and the data utility offered by each of them. This being said, is clear that the challenge does not only lie in anonymizing data itself, but also in preserving an acceptable level of information utility.

Throughout this report, we explain the reasoning behind the choices made on each of the steps of this process.

Document Structure

The structure of this report is as follows:

- Section 2, **Selection, Importing & Goal of Dataset**, specifies the goal of the release of the anonymized dataset and defines both privacy and utility requirements for the anonymization process.
- Section 3, **Characterization of the Dataset & Coding Models**, provides a characterization of the original dataset by classifying attributes and analyzing the re-identification risk.
- In Section 4, **Privacy Models**, we apply different privacy models to the target dataset and analyse the privacy protection and data utility offered by each of them.
- Section 5, **Conclusion**, concludes the work and suggests future improvements.

2 Selection, Importing & Goal of Dataset

In the following section, we describe the dataset, define the privacy and utility requirements for the anonymization process and specify the goal with which the anonymized dataset would be released.

For this assignment, we generated our own dataset based on a fictional dataset available at <https://www.kaggle.com/spscientist/students-performance-in-exams>, consisting of the scores obtained in tests by high school students from the United States, as well as some social and economic aspects.

2.1 Privacy & Utility Requirements

With the use of anonymization operations, it is inevitable that some information will be lost. However, even after the anonymization process, the resulting data should preserve an acceptable level of utility. This being said, it should still be possible to answer questions along the lines of the following:

- How does the parental level of education affect one’s school performance?
- How does one’s gender relate to their school performance?

On the other hand, when it comes to privacy, the following requirements must be met:

- Under no circumstances should the re-identification risk be greater than 50%.
- No record in the anonymized dataset should be unique.

It is important to clarify that in this context *anonymity* refers to the indistinguishability from other records with respect to QID.

2.2 Data Preprocessing

In the original dataset, the following attributes are present (see Table 1):

- **Gender**
- **Ethnicity:** This attribute was already subject to anonymization in the original dataset. As a matter of fact, the values for this attribute are grouped into **Group A**, **B**, **C**, **D** and **E**.
- **Parental level of education**
- **Lunch:** Whether the student had a standard or a free/reduced lunch.

- **Test preparation course:** Whether or not the student completed a test preparation course.
- **Math, Reading & Writing score:** These attributes will be combined in a single attribute, **Average Score**, since we are only interested in the overall school performance and not the performance in specific subjects. The values of the resulting attribute will be in the range $[0, 100]$ and will be calculated according to

$$\text{Average Score} = \frac{\text{Math Score} + \text{Reading Score} + \text{Writing Score}}{3}$$

and rounded to the nearest integer.

Number of Records	1000
Number of Attributes	8
Attributes	Gender, Race/Ethnicity, Parental Level of Education, Lunch, Test Preparation Course, Math Score, Reading Score, Writing Score

Table 1: Characterization of the original dataset

Considering the fact that this dataset doesn’t include any personally identifiable information (PII), we added some identifying attributes and other sensitive information – *i.e.* name, social security number, birth date (in the `yyyy-MM-dd` format) and city – as to make it more suitable for the anonymization process. This was done using **Faker**, a Python library that allows the generation random data (see Listing 1 and Table 2).

Number of Records	1000
Number of Attributes	9
Attributes	Name, SSN, Birth Date, City, Gender, Race/Ethnicity, Parental Level of Education, Test Preparation Course, Average Score

Table 2: Characterization of the original dataset after introducing synthetic data

3 Characterization of the Dataset & Coding Models

In this section, we provide an overview of the original dataset and classify attributes into *identifying*, *quasi-identifying*, *sensitive* and *insensitive*. Finally, we analyze the distribution of data the privacy risks.

As previously stated in Section 2, this dataset contains the scores obtained in tests by high school students. Essentially, it has 1,000 records, each with 9 attributes, none of which have null values. A more in-depth view of the dataset is presented in Table 3, and an excerpt is presented in Figure 2.

Attribute	Possible Values	Type
Name	US Names (first and last name)	String
SSN	9 digit numbers in the format AAA-GG-SSSS	String
Gender	Male/Female	String
City	US Cities	String
Birth Date	Values in the format yyyy-MM-dd ranging from 2002-01-01 to 2005-01-01	DateTime
Race/Ethnicity	Group A, B, C, D or E	String
Parental Level of Education	Some High School, High School, Some College, Associate's Degree, Bachelor's Degree, Master's Degree,	String
Test Preparation Course	Completed, None	String
Average Score	[0, 100]	Integer

Table 3: Possible values and type for each attribute

Next, we import the dataset into ARX (Fig. 1), an open source data anonymization tool developed in Java that supports several privacy models such as k -anonymity, ℓ -diversity t -closeness and differential privacy, as well as utility metrics for analyzing data utility and re-identification risks.

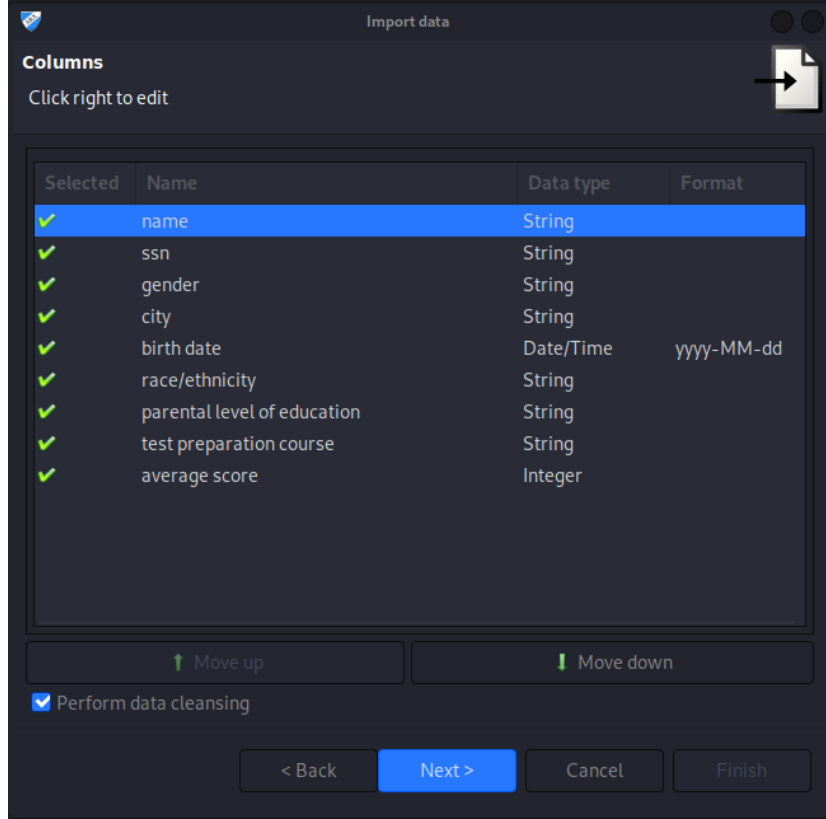


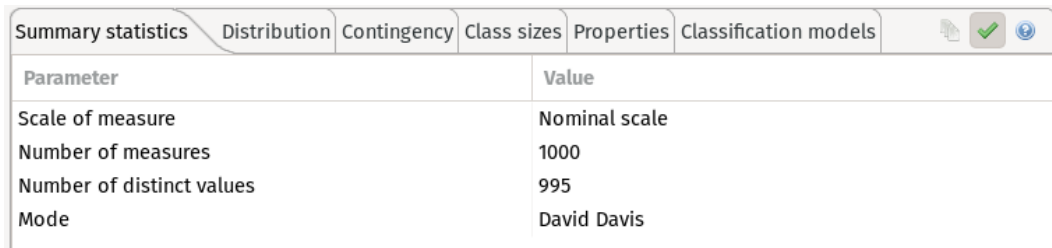
Figure 1: Importing data into ARX

Input data									
	name	ssn	gender	city	birth date	race/ethnicity	parental level of education	test preparation course	average
1	✓ Allison Hill	501-85-4342	female	Olympia	2004-07-13	group B	bachelor's degree	none	168
2	✓ Noah Rhodes	094-83-8915	female	Union	2002-07-01	group C	some college	completed	188
3	✓ Angie Hende...	397-24-6079	female	Strongsville	2003-04-24	group B	master's degree	none	216
4	✓ Daniel Wagner	391-48-3057	male	Saint Cloud	2004-12-31	group A	associate's degree	none	118
5	✓ Cristian Santos	457-06-4314	male	Sunset	2003-05-04	group C	some college	none	179
6	✓ Connie Lawr...	451-60-4316	female	Grove City	2002-02-19	group B	associate's degree	none	180
7	✓ Abigail Shaffer	232-35-9245	female	North Las Ve...	2003-02-26	group B	some college	completed	213
8	✓ Gina Moore	064-20-1585	male	Racine	2003-10-05	group B	some college	none	96
9	✓ Gabrielle Davis	088-86-5633	male	Ferguson	2002-05-15	group D	high school	completed	150
10	✓ Ryan Munoz	828-70-6783	female	Nixa	2004-02-12	group B	high school	none	114
11	✓ Monica Herr...	775-76-3728	male	Kokomo	2003-01-05	group C	associate's degree	none	129
12	✓ Jamie Arnold	567-08-6343	male	Ashwaubenon	2003-08-01	group D	associate's degree	none	106
13	✓ Lisa Hensley	839-67-6859	female	Dracut	2003-12-04	group B	high school	none	170
14	✓ Michele Willi...	549-88-7780	male	Kuna	2003-04-16	group A	some college	completed	173
15	✓ Dylan Miller	587-31-7804	female	Olney	2002-12-02	group A	master's degree	none	122

Figure 2: Excerpt of the dataset

3.1 Distribution of Data for Each Attribute

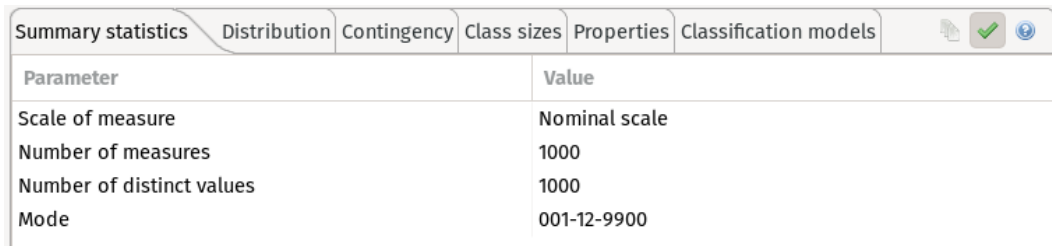
Given the nature of the data, it is crucial to ensure that it follows a reasonable distribution. To do so, we analyzed the summary statistics and distribution for each attribute using ARX, as presented in Figures 3 to 16.



The screenshot shows the ARX software interface with the 'Summary statistics' tab selected. The table displays the following data:

Parameter	Value
Scale of measure	Nominal scale
Number of measures	1000
Number of distinct values	995
Mode	David Davis

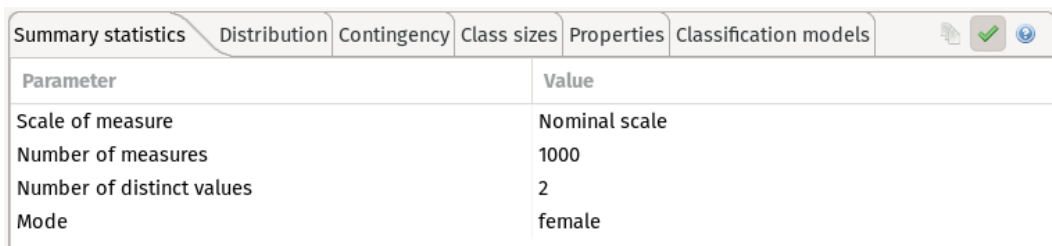
Figure 3: Summary statistics for the **Name** attribute



The screenshot shows the ARX software interface with the 'Summary statistics' tab selected. The table displays the following data:

Parameter	Value
Scale of measure	Nominal scale
Number of measures	1000
Number of distinct values	1000
Mode	001-12-9900

Figure 4: Summary statistics for the **SSN** attribute



The screenshot shows the ARX software interface with the 'Summary statistics' tab selected. The table displays the following data:

Parameter	Value
Scale of measure	Nominal scale
Number of measures	1000
Number of distinct values	2
Mode	female

Figure 5: Summary statistics for the **Gender** attribute



Figure 6: Distribution of values for the **Gender** attribute

Summary statistics	Distribution	Contingency	Class sizes	Properties	Classification models
Parameter		Value			
Scale of measure		Nominal scale			
Number of measures		1000			
Number of distinct values		816			
Mode		Huntington Beach			

Figure 7: Summary statistics for the **City** attribute

Summary statistics	Distribution	Contingency	Class sizes	Properties	Classification models
Parameter		Value			
Scale of measure		Interval scale			
Number of measures		1000			
Number of distinct values		659			
Mode		2002-08-20			
Median		2003-07-13			
Min		2002-01-01			
Max		2004-12-31			
Arithmetic mean		2003-07-08			
Sample variance		2169w ² , 16d ² , 49h ² , 2108m ² , 2258s ² , 456576ms ²			
Population variance		2167w ² , 7d ² , 454h ² , 1670m ² , 3307s ² , 382784ms ²			
Standard deviation		46w, 4d, 46m, 34s, 91ms			
Range		156w, 3d, 0ms			
Kurtosis		Not available			

Figure 8: Summary statistics for the **Date of Birth** attribute

Summary statistics		Distribution	Contingency	Class sizes	Properties	Classification models			
Parameter		Value							
Scale of measure		Nominal scale							
Number of measures		1000							
Number of distinct values		5							
Mode		group C							

Figure 9: Summary statistics for the Race/Ethnicity attribute

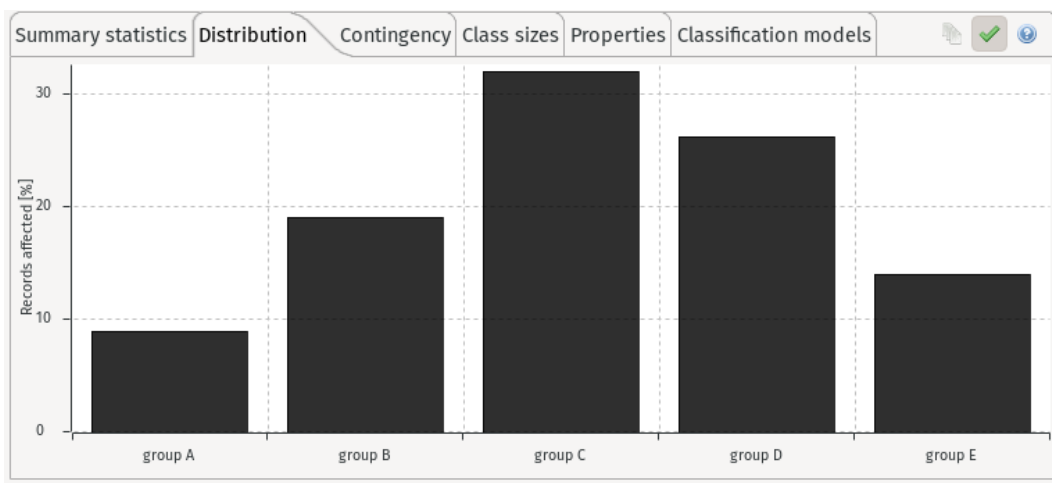


Figure 10: Distribution of values for the Race/Ethnicity attribute

Summary statistics		Distribution	Contingency	Class sizes	Properties	Classification models			
Parameter		Value							
Scale of measure		Nominal scale							
Number of measures		1000							
Number of distinct values		6							
Mode		some college							

Figure 11: Summary statistics for the Parental Level of Education attribute

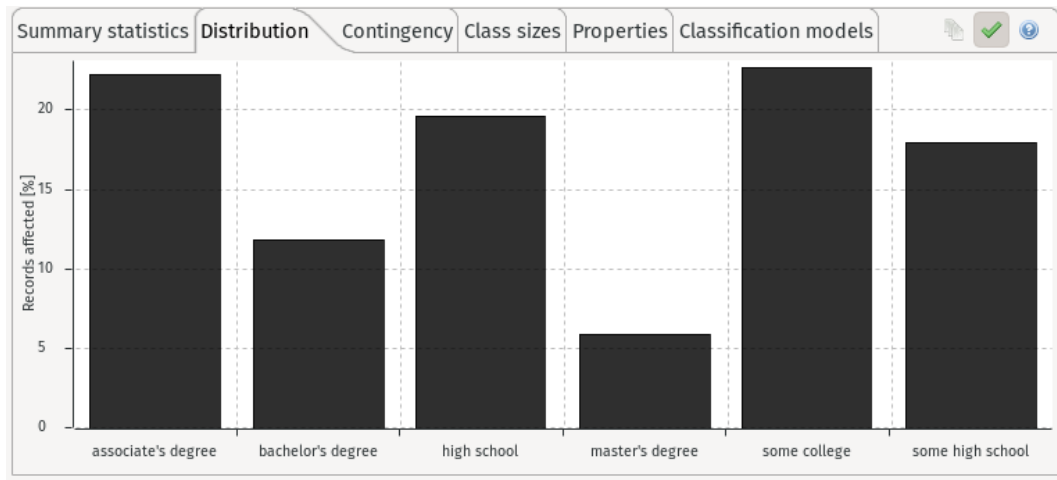


Figure 12: Distribution of values for the Parental Level of Education attribute

Parameter	Value
Scale of measure	Nominal scale
Number of measures	1000
Number of distinct values	2
Mode	none

Figure 13: Summary statistics for the Test Preparation Course attribute

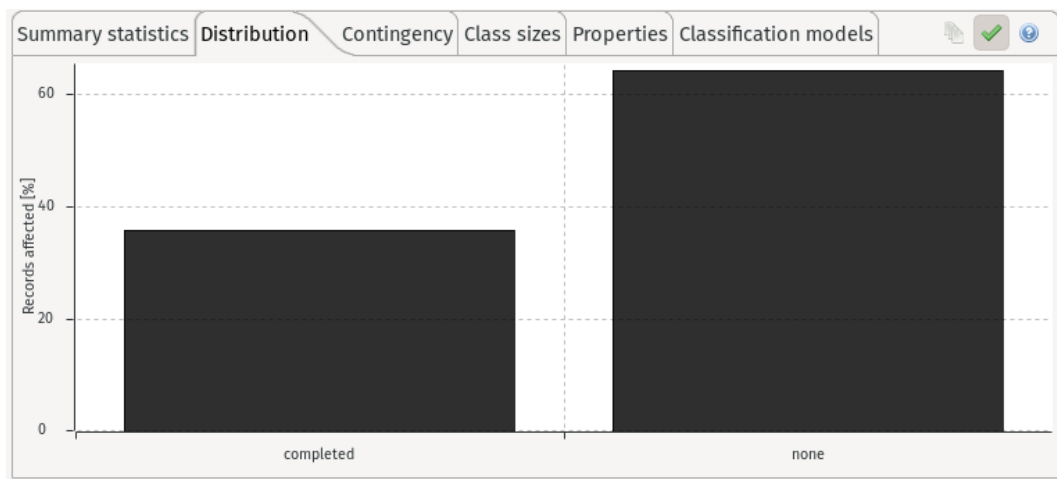


Figure 14: Distribution of values for the Test Preparation Course attribute

Summary statistics		Distribution	Contingency	Class sizes	Properties	Classification models
Parameter		Value				
Scale of measure		Ratio scale				
Number of measures		1000				
Number of distinct values		73				
Mode		68				
Median		68				
Min		9				
Max		100				
Arithmetic mean		67				
Sample variance		203				
Population variance		203				
Standard deviation		14				
Range		91				
Kurtosis		0				
Geometric mean		66				

Figure 15: Summary statistics for the Average Score attribute

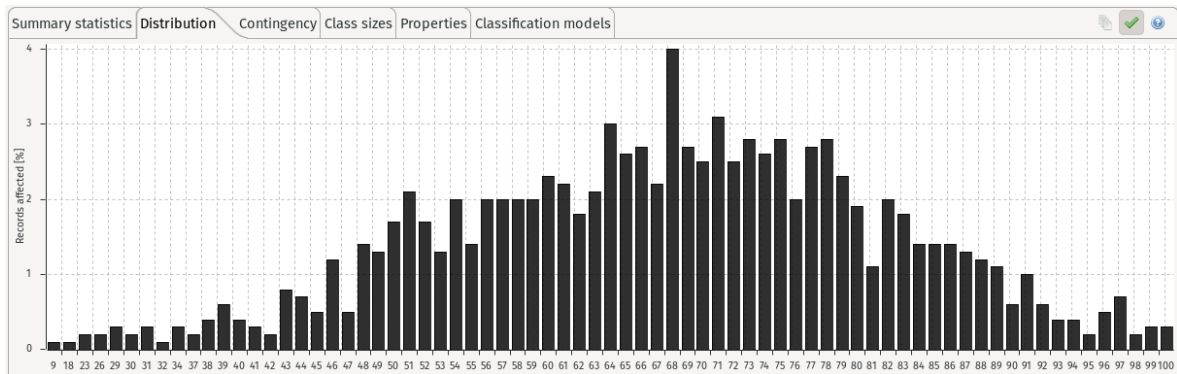


Figure 16: Distribution of values for the Average Score attribute

3.2 Classification of Attributes

The attributes can be classified as follows:

- **Identifying Attributes:** Name, SSN. As any real world scenario, more than one student have the same name. Identifying Attributes are associated with a high risk of re-identification and, as such, they will be removed before disclosing the dataset.
- **Quasi-Identifiers:** Birth Date, City, Gender, Race/Ethnicity. These attributes do not explicitly identify a record owner but can be combined with data from public sources in an attempt to de-anonymize the owner of a record.
- **Insensitive Attributes:** Parental Level of Education, Test Preparation. These attributes do not pose any privacy risks and, as such, they will be kept unmodified.
- **Sensitive Attributes:** Average Score. As any sensitive attribute, its disclosure may be undesirable from the record owner standpoint and, as such, it will be kept unmodified but will be subject to restrictions imposed by the privacy model – *e.g.* l -diversity, t -closeness, etc.

We were particularly careful when differentiating quasi-identifiers from sensitive attributes because a wrong classification may result in a lower privacy protection – and hence a higher re-identification risk – than what was expected. As a matter of fact, incorrectly classifying an attribute as a quasi-identifier will result in a greater loss of information due to the anonymization operations that will be applied to each QID. On the other hand, misclassifying an attribute as a quasi-identifier when it is, in fact, a sensitive attribute can have an undesirable effect on the privacy protection in the sense that an attacker is then able perform an attribute linkage attack.

The classification of the attributes can be checked in ARX (Figs. 17 and 18) – high values of distinction¹ and separation² indicate probable QIDs.

¹Distinction measures the degree to which variables make records distinct.

²Separation measures the degree to which combinations of variables separate the records.

Distribution of risks	Quasi-identifiers	Attacker models	HIPAA identifiers	
Quasi-identifier				
		Distinction	Separation	
test preparation course		0.2%	46.01321%	
gender		0.2%	49.98519%	
race/ethnicity		0.5%	76.67407%	
parental level of education		0.6%	81.25906%	
birth date		65.9%	99.91211%	
city		81.6%	99.95756%	
gender, test preparation course		0.4%	73.01782%	
race/ethnicity, test preparation course		1%	87.34234%	
gender, race/ethnicity		1%	88.27127%	
parental level of education, test preparation course		1.2%	89.81542%	
gender, parental level of education		1.2%	90.65425%	
race/ethnicity, parental level of education		3%	95.61662%	
birth date, test preparation course		80.5%	99.95455%	
gender, birth date		80.8%	99.95756%	
city, test preparation course		88.8%	99.97578%	
birth date, race/ethnicity		89.9%	99.97858%	
gender, city		90.5%	99.97998%	
birth date, parental level of education		91.3%	99.98158%	
city, race/ethnicity		95.2%	99.98999%	
city, parental level of education		95.6%	99.99119%	
city, birth date		100%	100%	
gender, race/ethnicity, test preparation course		2%	93.63964%	
gender, parental level of education, test preparation course		2.4%	94.92913%	
race/ethnicity, parental level of education, test preparation course		5.9%	97.6034%	
gender, race/ethnicity, parental level of education		6%	97.79459%	
gender, birth date, test preparation course		89.7%	99.97818%	
gender, city, test preparation course		94.9%	99.98959%	
gender, birth date, parental level of education		95%	99.98979%	
birth date, parental level of education, test preparation course		95.3%	99.99039%	
birth date, race/ethnicity, test preparation course		95.6%	99.99099%	
gender, birth date, race/ethnicity		96%	99.99199%	
city, race/ethnicity, test preparation course		97.2%	99.99439%	
birth date, race/ethnicity, parental level of education		97.7%	99.9954%	
gender, city, race/ethnicity		97.7%	99.9954%	

Figure 17: Identification of QIDs through ARX (i)

gender, city, race/ethnicity	97.7%	99.9954%
city, parental level of education, test preparation course	97.9%	99.9958%
gender, city, parental level of education	97.9%	99.9958%
city, race/ethnicity, parental level of education	98.4%	99.9968%
city, birth date, parental level of education	100%	100%
city, birth date, race/ethnicity	100%	100%
city, birth date, test preparation course	100%	100%
gender, city, birth date	100%	100%
gender, race/ethnicity, parental level of education, test preparation course	11.5%	98.8036%
gender, birth date, parental level of education, test preparation course	97.4%	99.99479%
gender, birth date, race/ethnicity, test preparation course	98.1%	99.9962%
gender, birth date, race/ethnicity, parental level of education	98.7%	99.9974%
gender, city, race/ethnicity, parental level of education	99%	99.998%
birth date, race/ethnicity, parental level of education, test preparation course	99.1%	99.9982%
gender, city, race/ethnicity, test preparation course	99.1%	99.9982%
city, race/ethnicity, parental level of education, test preparation course	99.3%	99.9986%
gender, city, parental level of education, test preparation course	99.3%	99.9986%
city, birth date, parental level of education, test preparation course	100%	100%
city, birth date, race/ethnicity, parental level of education	100%	100%
city, birth date, race/ethnicity, test preparation course	100%	100%
gender, city, birth date, parental level of education	100%	100%
gender, city, birth date, race/ethnicity	100%	100%
gender, city, birth date, test preparation course	100%	100%
gender, birth date, race/ethnicity, parental level of education, test preparation course	99.5%	99.999%
gender, city, race/ethnicity, parental level of education, test preparation course	99.7%	99.9994%
city, birth date, race/ethnicity, parental level of education, test preparation course	100%	100%
gender, city, birth date, parental level of education, test preparation course	100%	100%
gender, city, birth date, race/ethnicity, parental level of education	100%	100%
gender, city, birth date, race/ethnicity, test preparation course	100%	100%
gender, city, birth date, race/ethnicity, parental level of education, test preparation course	100%	100%

Figure 18: Identification of QIDs through ARX (ii)

3.3 Risk Analysis

Figure 19 shows an overview of several measures for re-identification risks considering three different attacker models – the prosecutor attacker model, the journalist attacker model, and the marketer attacker model.

In the prosecutor scenario, an attacker targets a specific individual and knows whether or not they are in the dataset; in the journalist scenario, an adversary selects a target at random because the re-identification of any record achieves the purpose; and finally, in the marketer scenario, an attacker targets as many individuals as possible, which means that an attack is considered successful if a considerable portion of the records can be re-identified.

As we can see, regardless of the attacker model we consider, the values for the re-identification risk are very close to 100%. This is highlighted by the fact that the birth date uniquely identifies 65.9% of the records; furthermore, the combination of the birth date and the ethnicity is enough to uniquely identify 89.9% of the records.

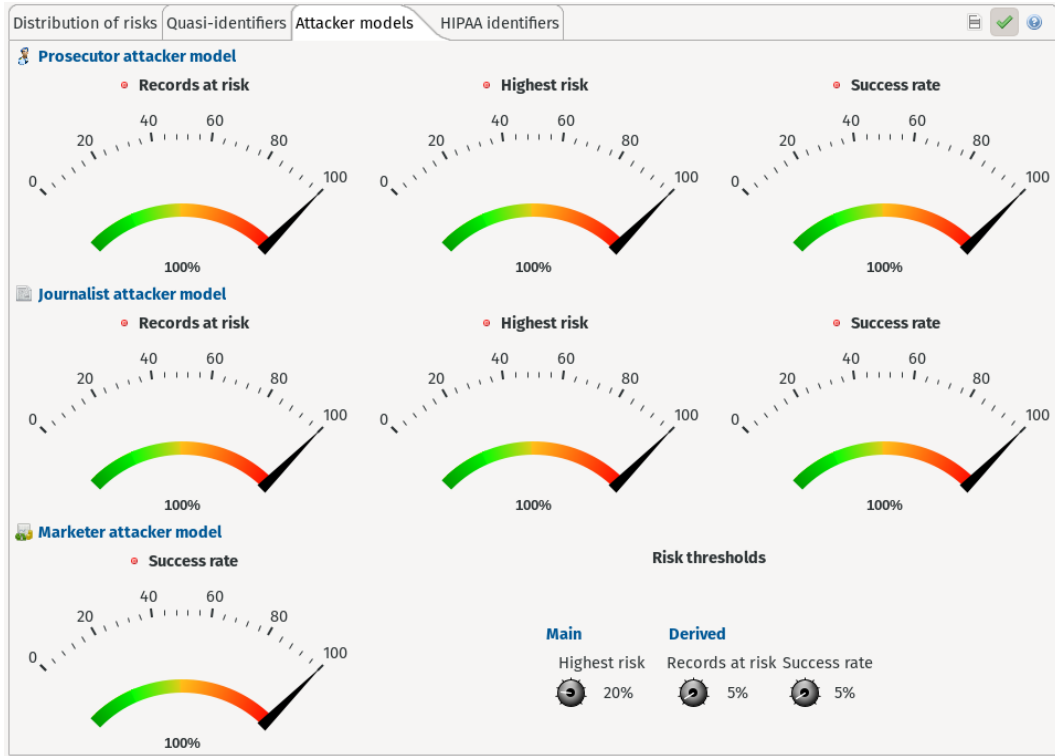


Figure 19: Risk of re-identification on each attacker model

In order to reduce the re-identification risk, we applied different privacy models to the data, which we will describe in the following sections.

4 Privacy Models

In an anonymization process, the first thing to do is removing PII before disclosing the dataset. PII includes but is not limited to:

- Name
- Telephone number
- Personal identification numbers such as social security number, passport number, driver’s license number, credit card number, etc.
- Email address
- Biometric data

However, simply removing PII from the dataset is not sufficient because an adversary can use background knowledge and cross-correlation with other databases to re-identify individual records. A famous example of such kind of attack includes the de-anonymization of a Massachusetts hospital discharge database, carried out by Latanya Sweeney, who joined it with a public voter database and then used the combination of both to determine the values of medical attributes for each person who appeared in both databases [2].

4.1 Model 1: k -Anonymity & ℓ -Diversity

k -Anonymity requires that each record in the dataset must be indistinguishable from at least $k - 1$ other records with respect to every quasi-identifier [3]. In other words, these k records form an *equivalence class* such that the minimum equivalence class size is k . As a result, the probability of linking a record owner to its record is at most $1/k$.

However, it is important to note k -anonymity alone does not ensure privacy as it assumes that each record represents a distinct record owner. However, if more than one record refers to the same individual, a set of k records will represent fewer than k individuals, which causes a record owner to be under-protected because the probability of re-identification will be greater than $1/k$ [4]. Even though this is not a problem in this dataset as each record belongs to a different individual, it is something worth keeping in mind since it allows for homogeneity attacks as well as background knowledge attacks [5].

On the other hand, ℓ -diversity addresses the limitations of k -anonymity by requiring that each equivalence class to have at least ℓ distinct records. This encompasses the idea that a sensitive attribute must be "diverse" within each equivalence class.

In the first privacy model, we applied k -anonymity, with $k = 3$ and distinct ℓ -diversity, with $\ell = 2$. In this case, we applied a character masking so the gender of the given individual stays anonymous (Fig. 20), replaced each city with the corresponding State³ (Fig. 21), replaced the date of birth with the most generalized date interval in the hierarchy defined in Figure 22 and, finally, we kept ethnicity attribute unmodified due to the fact that it was already generalized in the initial dataset.

This is summarized in Table 4:

Quasi-identifier	Transformation	Levels
Gender	Generalization – Character masking	0–1
City	Generalization – Class grouping	0–1
Birth Date	Generalization – Date intervals	0–3
Race/Ethnicity	None	0

Table 4: Transformations applied to each of the QID

Level-0	Level-1
female	*
male	*

Figure 20: Gender transformation levels

LEVEL 0 : LEVEL 1
"El Dorado": "Arkansas",
"McAlester": "Oklahoma",
"Levittown": "New York",
"South Whittier": "California",
"Huntington": "New York",
"McMinnville": "Oregon",

Figure 21: City transformation levels

[2002-01-01, 2002-12-31[[2002-01-01, 2002-12-31[[2002-01-01, 2003-12-30[[2002-01-01, 2003-12-30[[2002-01-01, 2005-12-27[[2002-01-01, 2005-12-27[
[2002-12-31, 2003-12-30[[2002-12-31, 2003-12-30[
[2003-12-30, 2004-12-28[[2003-12-30, 2004-12-28[
[2004-12-28, 2005-12-27[[2004-12-28, 2005-12-27[[2003-12-30, 2005-12-27[[2003-12-30, 2005-12-27[

Figure 22: Generalization hierarchy for the Birth Date attribute

³This transformation was made using a Python script (see Listing 2) prior to importing the dataset into ARX.

4.1.1 Results

Applying the transformation specified in Table 4 with levels $[1, 1, 3, 0]$, we obtained the following results:

		img	en	gender	city	birth date	race/ethnicity	parental level of education	post preparation course	average score
1	✓	*	*	*	Alabama	[2002-01-01, 2004-12-31[group D	bachelor's degree	none	67
2	✓	*	*	*	Alabama	[2002-01-01, 2004-12-31[group D	some high school	completed	78
3	✓	*	*	*	Alabama	[2002-01-01, 2004-12-31[group D	some high school	none	51
4	✓	*	*	*	Arizona	[2002-01-01, 2004-12-31[group A	high school	none	52
5	✓	*	*	*	Arizona	[2002-01-01, 2004-12-31[group A	high school	none	51
6	✓	*	*	*	Arizona	[2002-01-01, 2004-12-31[group A	some college	none	59
7	✓	*	*	*	Arizona	[2002-01-01, 2004-12-31[group A	associate's degree	none	59
8	✓	*	*	*	Arizona	[2002-01-01, 2004-12-31[group A	bachelor's degree	completed	55
9	✓	*	*	*	Arizona	[2002-01-01, 2004-12-31[group C	associate's degree	completed	46
10	✓	*	*	*	Arizona	[2002-01-01, 2004-12-31[group C	high school	completed	76
11	✓	*	*	*	Arizona	[2002-01-01, 2004-12-31[group C	some college	none	59
12	✓	*	*	*	Arizona	[2002-01-01, 2004-12-31[group C	associate's degree	completed	77
13	✓	*	*	*	Arizona	[2002-01-01, 2004-12-31[group C	high school	none	68
14	✓	*	*	*	Arizona	[2002-01-01, 2004-12-31[group C	high school	none	51
15	✓	*	*	*	Arizona	[2002-01-01, 2004-12-31[group C	some college	completed	70
16	✓	*	*	*	Arizona	[2002-01-01, 2004-12-31[group D	some college	completed	72
17	✓	*	*	*	Arizona	[2002-01-01, 2004-12-31[group D	high school	none	46
18	✓	*	*	*	Arizona	[2002-01-01, 2004-12-31[group D	some college	completed	74
19	✓	*	*	*	Arizona	[2002-01-01, 2004-12-31[group E	associate's degree	completed	80
20	✓	*	*	*	Arizona	[2002-01-01, 2004-12-31[group E	some college	completed	89
21	✓	*	*	*	Arizona	[2002-01-01, 2004-12-31[group E	associate's degree	completed	81
22	✓	*	*	*	Arkansas	[2002-01-01, 2004-12-31[group B	some college	completed	91
23	✓	*	*	*	Arkansas	[2002-01-01, 2004-12-31[group B	high school	none	71
24	✓	*	*	*	Arkansas	[2002-01-01, 2004-12-31[group B	high school	none	23
25	✓	*	*	*	Arkansas	[2002-01-01, 2004-12-31[group B	high school	none	60

Figure 23: Excerpt of the dataset anonymized with Model 1

Model 1: k -Anonymity & ℓ -Diversity ($k = 3, \ell = 2$)	
Number of suppressed records	113
Minimal Class Size	3
Average Class Size	8
Maximal Class Size	42
Number of Classes	109

Table 5: General results after anonymizing with Model 1

4.1.2 Risk Analysis & Data Utility

Despite the considerably lower re-identification risk when compared to the original dataset (see Fig. 24), the results we obtained using this privacy model were not considered satisfactory, given the generalization level of many of the attributes, as well as due to a significant loss in quality of data we considered as priority.

Even after the application of the privacy model, it is still possible to answer the first question we set as requirements in Section 2.1 – *How does the parental level of education affect one’s school performance?*. As a matter of fact, since we considered the **Parental Level of Education** as an insensitive attribute, it won’t be subject to any anonymization operation and, as a result, all the information that relates it to the sensitive attribute **Average Score** will be preserved. However, we note that since the **Gender** attribute was suppressed, all the information associated with it will be lost and, as such, the second question set as an utility requirement – *How does one’s gender relate to their school performance?* – becomes impossible to answer based on the anonymized version of the dataset.

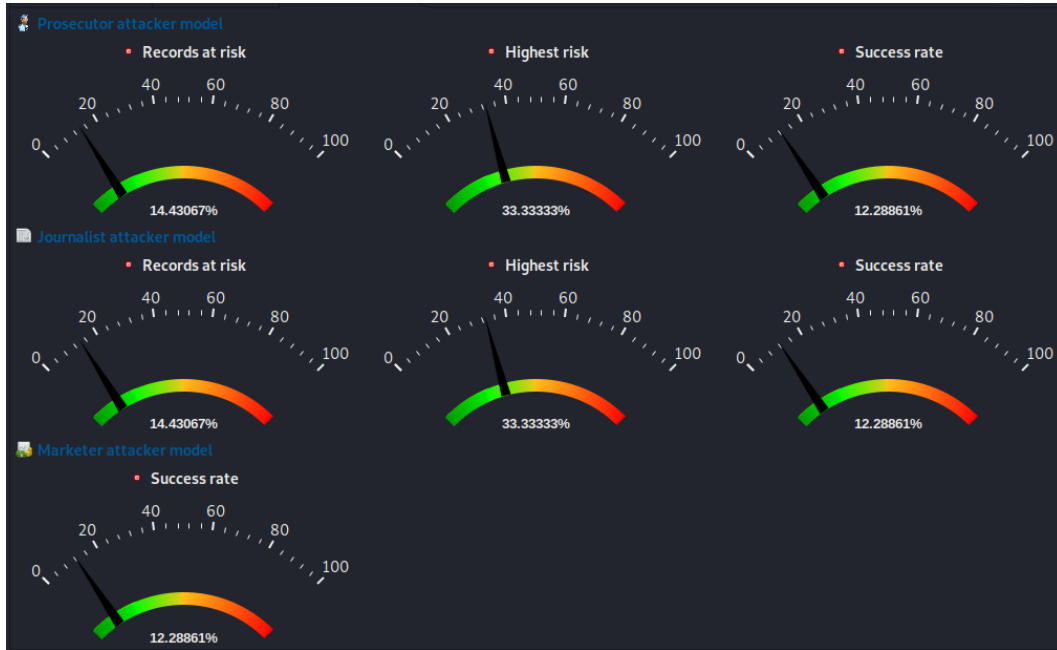


Figure 24: Risk associated with each attacker model considering Model 1

To address this limitation, we experimented with different attribute weights for mitigating information loss on QIDs. Thus, we consider another privacy model, described in Section 4.2, in which preserve the **Gender** attribute in the anonymized version of the dataset and apply a higher level of generalization to the **City** and **Birth Date** attributes.

4.2 Model 2: k -anonymity and ℓ -diversity with attribute weights

Based on limitations we identified in the previous model, we now consider a new privacy model in which we will again apply k -anonymity and ℓ -diversity, with $k = 3$ and $\ell = 2$. However, in order to be able to answer to the second question we set as an utility requirement – *How does one’s gender relate to their school performance?* – we made the following alterations to the previous privacy model:

- Set the supression limit to 50% (see Fig 25). This limits the maximum numbers records ARX can supress.
- In the coding model, adjust the trade-off between suppression and generalization (see Fig 26). In this case we chose to have slightly more generalization than suppression.
- Increase the attribute weight for the **Gender** (see Fig. 27). This leads to less information loss on this attribute, which, in turn, makes it possible to answer to the second question set as an utility requirement – *How does one’s gender relate to their school performance?*

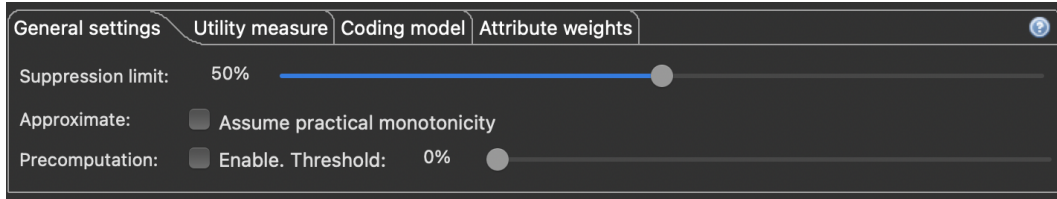


Figure 25: Supression limit for Model 2

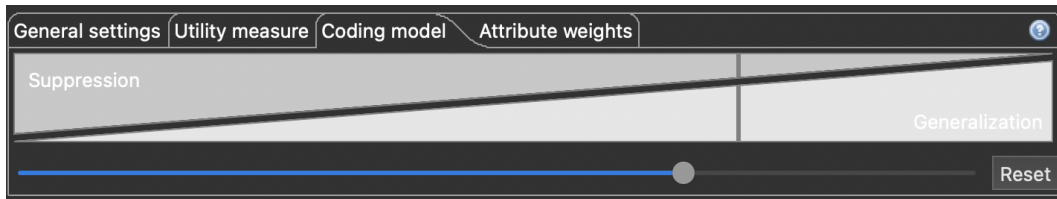


Figure 26: Coding model for Model 2

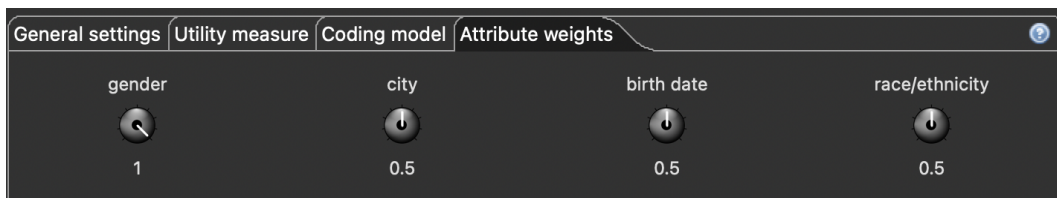


Figure 27: Attribute weights for Model 2

4.2.1 Results

Using the same attribute transformations specified in Table 4, but now considering the levels $[0, 1, 2, 0]$ and the attribute weights specified in Figure 27, the anonymized dataset is as follows:

	gender	city	birth date	race/ethnicity	parental level of education	post preparation course	average score
41	female	California	[2002-01-01, 2003-12-30[group D	high school	completed	65
42	female	California	[2002-01-01, 2003-12-30[group D	some college	completed	78
43	female	California	[2002-01-01, 2003-12-30[group E	bachelor's degree	none	40
44	female	California	[2002-01-01, 2003-12-30[group E	bachelor's degree	completed	80
45	female	California	[2002-01-01, 2003-12-30[group E	associate's degree	none	87
46	female	California	[2002-01-01, 2003-12-30[group E	associate's degree	none	70
47	female	California	[2002-01-01, 2003-12-30[group E	some college	completed	82
48	female	California	[2003-12-30, 2004-12-31[group B	master's degree	completed	70
49	female	California	[2003-12-30, 2004-12-31[group B	high school	completed	76
50	female	California	[2003-12-30, 2004-12-31[group B	associate's degree	completed	85
51	female	California	[2003-12-30, 2004-12-31[group B	bachelor's degree	none	79
52	female	California	[2003-12-30, 2004-12-31[group C	master's degree	completed	86
53	female	California	[2003-12-30, 2004-12-31[group C	bachelor's degree	completed	66
54	female	California	[2003-12-30, 2004-12-31[group C	bachelor's degree	none	90
55	female	California	[2003-12-30, 2004-12-31[group C	high school	none	65
56	female	California	[2003-12-30, 2004-12-31[group C	associate's degree	none	89
57	female	California	[2003-12-30, 2004-12-31[group C	associate's degree	completed	69
58	female	California	[2003-12-30, 2004-12-31[group C	bachelor's degree	completed	73
59	female	California	[2003-12-30, 2004-12-31[group C	high school	none	63
60	female	California	[2003-12-30, 2004-12-31[group C	high school	none	78
61	female	California	[2003-12-30, 2004-12-31[group C	associate's degree	none	56
62	female	California	[2003-12-30, 2004-12-31[group C	associate's degree	none	50
63	female	California	[2003-12-30, 2004-12-31[group D	master's degree	completed	93
64	female	California	[2003-12-30, 2004-12-31[group D	high school	completed	95
65	female	California	[2003-12-30, 2004-12-31[group D	master's degree	none	60
66	female	California	[2003-12-30, 2004-12-31[group D	some college	none	84
67	female	California	[2003-12-30, 2004-12-31[group D	master's degree	none	88

Figure 28: Excerpt of the dataset anonymized with Model 2

Model 2: k -Anonymity & ℓ -Diversity ($k = 3, \ell = 2$, with attribute weights)	
Number of suppressed records	454
Minimal Class Size	3
Average Class Size	4.92
Maximal Class Size	16
Number of Classes	111

Table 6: General results after anonymizing with Model 2

4.2.2 Risk Analysis & Data Utility

Considering the limitations of the previous privacy model, **Model 2** takes into account the fact that we must keep information about the **Gender** attribute. As a result of the increase in data utility, the number of records at risk increased, as expected.

Besides, the average class size decreased from 8 to 4.92, about half the size of the previous model, which may have a significant impact on the privacy protection offered, since each record is now, on average, indistinguishable from other 3.92 instead of the previous 7. In addition, 454 records, which equates to about 45% of the records in the dataset, were suppressed, numbers that we considered completely unacceptable. This is even more relevant considering that, in an ideal scenario, the dataset should still retain enough utility to so that it can be used for other tasks than the initial.

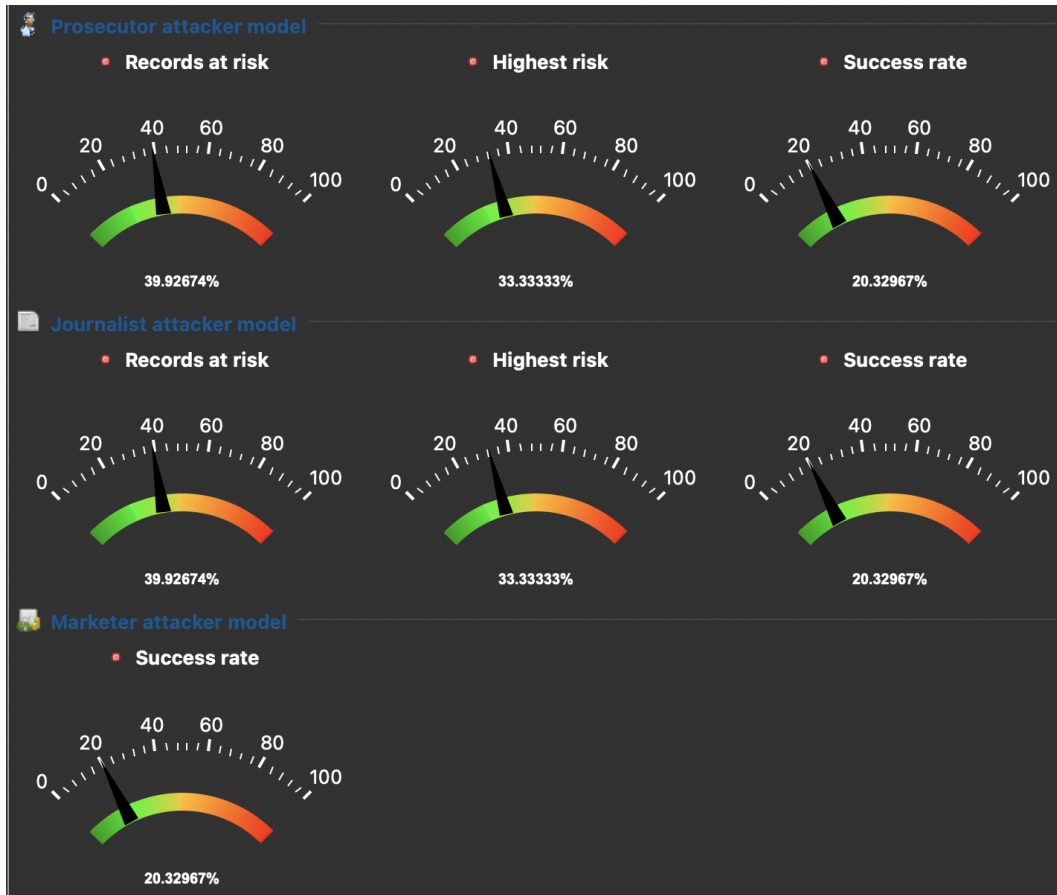


Figure 29: Risk for Model 2

4.3 Model 3: t -Closeness and k -Anonymity

When we applied the previous privacy model, Model 2, too much information was lost. In order to avoid this issue, we now consider another model, in which we consider k -anonymity, with $k = 3$ and t -closeness, with a threshold of 1. With t -closeness, we try to achieve a closer distribution of sensitive values – in this case, the **Average Score** – in each equivalence classes to the initial dataset.

4.3.1 Results

Once again, after applying the transformation specified in Table 4, this time with levels $[0, 1, 3, 0]$, we obtained the following results:

		img	gender	city	birth date	race/ethnicity	parental level of education	test preparation course	average scc
130	✓	*	female	Georgia	[2002-01-01...	group C	some college	completed	62
131	✓	*	female	Georgia	[2002-01-01...	group C	associate's degree	completed	79
132	✓	*	female	Georgia	[2002-01-01...	group C	some college	none	51
133	✓	*	female	Georgia	[2002-01-01...	group C	associate's degree	completed	71
134	✓	*	female	Georgia	[2002-01-01...	group C	some high school	completed	90
135	✓	*	female	Georgia	[2002-01-01...	group C	associate's degree	none	67
136	✓	*	female	Georgia	[2002-01-01...	group D	some high school	completed	99
137	✓	*	female	Georgia	[2002-01-01...	group D	some high school	none	80
138	✓	*	female	Georgia	[2002-01-01...	group D	some college	none	65
139	✓	*	female	Illinois	[2002-01-01...	group B	associate's degree	none	48
140	✓	*	female	Illinois	[2002-01-01...	group B	associate's degree	none	63
141	✓	*	female	Illinois	[2002-01-01...	group B	some high school	completed	59
142	✓	*	female	Illinois	[2002-01-01...	group B	high school	none	87
143	✓	*	female	Illinois	[2002-01-01...	group B	some college	completed	60
144	✓	*	female	Illinois	[2002-01-01...	group B	some college	none	84
145	✓	*	female	Illinois	[2002-01-01...	group C	some college	none	58
146	✓	*	female	Illinois	[2002-01-01...	group C	some college	none	71
147	✓	*	female	Illinois	[2002-01-01...	group C	some college	none	85
148	✓	*	female	Illinois	[2002-01-01...	group C	some college	none	63
149	✓	*	female	Illinois	[2002-01-01...	group C	high school	none	44
150	✓	*	female	Illinois	[2002-01-01...	group C	associate's degree	completed	64
151	✓	*	female	Illinois	[2002-01-01...	group C	bachelor's degree	none	71
152	✓	*	female	Illinois	[2002-01-01...	group D	some college	none	72
153	✓	*	female	Illinois	[2002-01-01...	group D	some high school	none	58
154	✓	*	female	Illinois	[2002-01-01...	group D	some college	none	82
155	✓	*	female	Illinois	[2002-01-01...	group D	some high school	none	91
156	✓	*	female	Illinois	[2002-01-01...	group D	some college	none	71

Figure 30: Anonymized dataset for Model 3

Model 3: k -Anonymity & t -Closeness ($k = 3, t = 1$)	
Number of suppressed records	235
Minimal Class Size	3
Average Class Size	5.71
Maximal Class Size	27
Number of Classes	134

Table 7: General results after anonymizing with Model 3

4.3.2 Risk Analysis & Data Utility

Similar to **Model 2**, the resulting dataset is able to answer the two questions set as utility requirements since it does not suppress the **Gender** attribute nor the **Parental Level of Education**. Although the result might seem close to the previous model, they differ both in the privacy protection they offer and in the data utility they retain.

Comparing the data utility metrics to the other models we can observe an increase in the number of equivalence classes and, most importantly, only 23.5% of the records are suppressed. On the other hand, we also note that the number of records at risk, as well as the success rate for each attacked decreased.

In short, the re-identification risk of this model is still not optimal; however, it performs slightly better than **Model 2** in the sense that it retains enough information while guaranteeing an acceptable level of privacy.

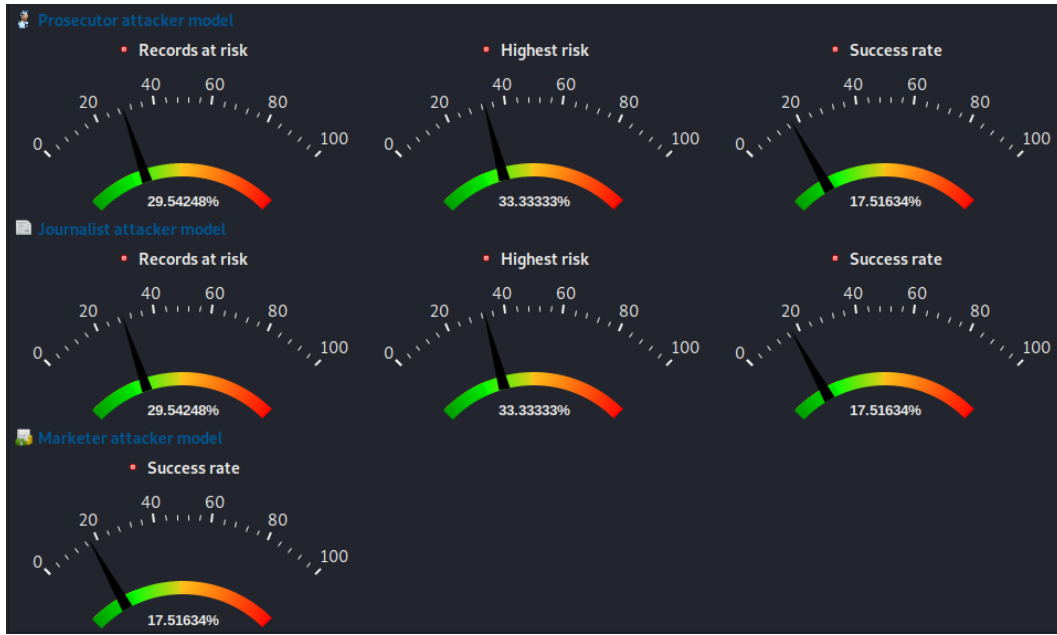


Figure 31: Risk analysis for **Model 3**

4.4 Comparative Analysis

After applying the different privacy models described previously, it is clear that the process of anonymizing a dataset should be an iterative one. As a matter of fact, we had to consistently change parameters in the privacy model in order to achieve an acceptable anonymized dataset that could not only answer the questions intended, but also have a low risk of re-identification. This process is not straight forward or intuitive due to all objectives considered.

Considering the values for the re-identification risk and data utility presented previously and summarized in Tables 8 and 9, we can see that the first model we considered, **Model 1**, despite maintaining a considerable level of privacy, it does not meet the utility requirements. As such, we considered a new privacy model, **Model 2**, that, even though it meets the privacy requirements, too much information is lost in the process. This lead us to consider a third privacy model, **Model 3**, that has both an acceptable level of privacy protection and data utility.

Utility Metric	Privacy Model		
	1	2	3
Number of suppressed records	113	454	235
Minimal Class Size	3	3	3
Average Class Size	8	4.92	5.71
Maximal Class Size	42	16	27
Number of Classes	109	111	134

Table 8: Comparative analysis of the data utility

		Privacy Model		
		1	2	3
Prosecut or attacker model	Records at risk	14.43%	39.93%	29.54%
	Highest risk	33.33%	33.33%	33.33%
	Success rate	12.29%	20.33%	17.51%
Journali st attacker model	Records at risk	14.43%	39.93%	29.54%
	Highest risk	33.33%	33.33%	33.33%
	Success rate	12.29%	20.33%	17.51%
Marketer attacker model	Sucess rate	12.29%	20.33%	17.52%

Table 9: Comparative analysis of the re-identification risk

This being said, the following trends were identified:

- Since we considered k -anonymity, with $k = 3$, the highest risk is $1/k = 1/3 \approx 33.33\%$ regardless of other considerations. Values of k greater than this would result in more privacy protection; however, the loss of information would be too excessive.
- Constraints on the **Gender** attribute lead to a considerable number of suppressed records. For example, in **Model 2**, about 45% of the records were suppressed.
- The constraint on the **Gender** attribute caused the average equivalence class size of the anonymized dataset to reduce, which, in turn, hinders the privacy protection offered.

Overall, considering this pattern, we can safely say that a greater level of utility usually implies a lower privacy protection, and vice versa. This problem is addressed in Section 4.4.1.

It is important to note that this models could be further improved. However, as various works have shown, finding the optimal anonymization is a NP-hard problem [6] and, taking these values into account, we consider that **Model 3** performs reasonably well.

Finally, we also want to point out that we expect that in a larger dataset, the re-identification risk would have been smaller given the fact that as the number of records increases, more records will have the same value for the birth date (*i.e.* this attribute would have a smaller value of distinction) and, as such, an attacker who knows the value of this attribute for a given record would not have enough information to carry out a record linkage attack. This is particularly relevant considering that, as stated in Section 3.3, the birth date uniquely identifies 65.9% of the records.

4.4.1 The Privacy-Utility Trade-off

In an ideal situation, we'd like to have a dataset with both maximal utility and maximal privacy. However, as illustrated in Figure 32, such scenario is impossible in the sense that maximum privacy means insufficient information utility and, on the other hand, maximum utility usually means little to no privacy protection. Indeed, privacy and utility are competing goals.

In fact, in a real scenario lies in between these two, where the privacy protection level is acceptable and the data also retains its utility. In this case, we want to make sure that the loss of information is minimal as to guarantee that it can still be useful for data analysis. This is particularly relevant in the fields of Privacy Preserving Data Mining (PPDM), whose goal is to extract knowledge from large amounts of data and provide accurate results while preventing sensitive information from disclosure; and Privacy Preserving Distributed Data Mining (PPDDM), which allows multiple parties to perform collaborative data mining without sharing any piece of data other than the final result.

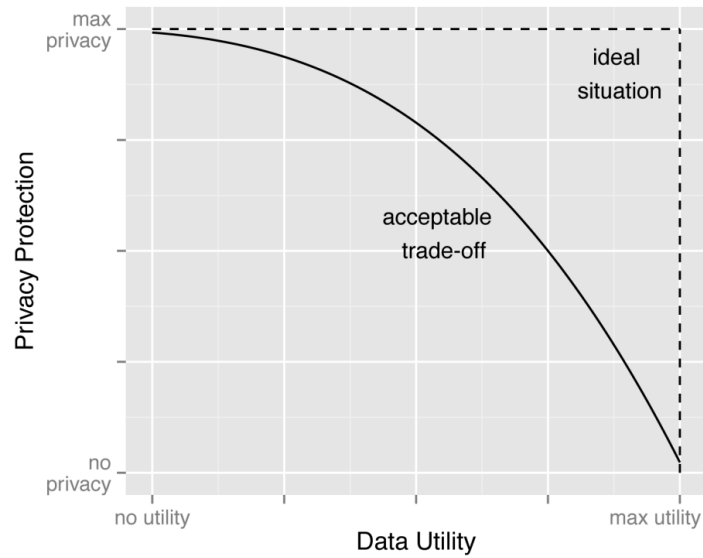


Figure 32: Trade-off between *perfect utility* and *perfect privacy* [1]

5 Conclusion

In this assignment, we have applied different privacy models and analyzed different anonymization techniques in an attempt to reduce the re-identification risk while maintaining an acceptable data utility level. In doing so, we were able to better understand the differences between privacy models and how it affects trade-off between privacy protection and data utility. In addition, we were able to conclude that the best privacy model to apply to a given dataset is highly dependent on the goal with which we release it.

As future work, we would like to try other privacy models such as *LKC*-privacy and (ϵ, δ) -differential privacy, as well as other utility metrics, such as information loss and discernability. This would allow us to have a clear understanding of the advantages and weaknesses of each privacy models. We would also like to experiment with other datasets to see how these results would generalise to high-dimensional datasets.

References

- [1] K. Emam and L. Arbuckle, *Anonymizing Health Data: Case Studies and Methods to Get You Started*. O'Reilly Media, 2013.
- [2] L. Sweeney, “Weaving Technology and Policy Together to Maintain Confidentiality,” *Journal of Law, Medicine & Ethics*, vol. 25, no. 2-3, p. 98–110, 2021.
- [3] —, “ k -Anonymity: A Model for Protecting Privacy,” *Int. J. Uncertain. Fuzziness Knowl. Based Syst.*, vol. 10, pp. 557–570, 2002.
- [4] B. C. Fung, K. Wang, A. W.-C. Fu, and P. S. Yu, *Introduction to Privacy-Preserving Data Publishing: Concepts and Techniques*, 1st ed. Chapman & Hall/CRC, 2010.
- [5] A. Machanavajjhala, D. Kifer, J. Gehrke, and M. Venkitasubramaniam, “ ℓ -Diversity: Privacy Beyond k -anonymity,” *22nd International Conference on Data Engineering (ICDE’06)*, pp. 24–24, 2006.
- [6] B. C. M. Fung, K. Wang, R. Chen, and P. S. Yu, “Privacy-Preserving Data Publishing: A Survey of Recent Developments,” *ACM Comput. Surv.*, vol. 42, no. 4, Jun. 2010.

A Data Generation using Faker

```
#!/usr/bin/env python3

import datetime
import pandas as pd
from faker import Faker
from city_state_dic import city_to_state_dict
from random import randrange

fake = Faker()
Faker.seed(42)

# Read the CSV file
df = pd.read_csv('StudentsPerformance.csv')
# Create name, SSN, birth date & city columns
df['name'] = [fake.name() for _ in range(1000)]
df['ssn'] = [fake.ssn() for _ in range(1000)]
df['birth date'] = [fake.date_between_dates(
    date_start=datetime.date(2002, 1, 1),
    date_end=datetime.date(2005, 1, 1)) for _ in range(1000)]
df['city'] = [list(city_to_state_dict.keys())[randrange(2361)]
              for _ in range(1000)]

# Create the average score column and drop the other scores
df['average score'] = ((df['math score'] +
                        df['reading score'] +
                        df['writing score']) / 3).astype(int)

df = df.drop('math score', axis=1)
df = df.drop('reading score', axis=1)
df = df.drop('writing score', axis=1)
df = df.drop('lunch', axis=1)

# Save the CSV file
df.to_csv('StudentsPerformanceUS.csv', encoding='utf-8',
          index=False)
```

Listing 1: Python script used to generate random synthetic data

B Generalization of the City attribute

```
#!/usr/bin/env python3
import pandas as pd

# city_to_state_dict is a Map with city:state
from city_state_dic import city_to_state_dict

# Read the CSV file
df = pd.read_csv('StudentsPerformanceUS.csv')

# Change each row to the corresponding state
for index, row in df.iterrows():
    df.city[index] = city_to_state_dict[df.city[index]]

# Save the CSV file
df.to_csv('StudentsPerformanceUS-CSG.csv', encoding='utf-8',
          index=False)
```

Listing 2: Python script used to replace each city with the corresponding State