Highlights

QIDLEARNINGLIB: A *Python* Library for Quasi-Identifier Recognition and Evaluation

Sancho Amaral Simões, João P. Vilela, Miriam Seoane Santos, Pedro Henriques Abreu

- QIDLEARNINGLIB is the first library for automated QID recognition in tabular datasets.
- It integrates metrics from causality, data privacy, and data utility for flexible QID selection.
- It provides metrics to evaluate the performance of the QID selection system, against a ground-truth.
- Includes multiple optimization algorithms for QID selection based on user-defined metrics.
- Supports redundancy analysis to identify the most relevant, non-overlapping metrics.
- Provides graphical and testing tools for enhanced interpretability.

QIDLearningLib: A *Python* Library for Quasi-Identifier Recognition and Evaluation

Sancho Amaral Simões[a,*] (researcher), João P. Vilela[a,b] (researcher), Miriam Seoane Santos[a,c] (researcher) and Pedro Henriques Abreu[a] (researcher)

[a] *CISUC-DEI - FCTUC, University of Coimbra, Coimbra, Portugal*
[b] *CRACS/INESCTEC-DCS - FCUP, University of Porto, Porto, Portugal*
[c] *LIAAD, INESCTEC, FCUP, University of Porto, Porto, Portugal*

ARTICLE INFO

Keywords:
Python library
Quasi-identifiers
Data anonymization
Data Privacy
Data Utility
Causality
Optimization

ABSTRACT

Quasi-identifiers (QIDs) are attributes in a dataset that are not directly unique identifiers of the users/entities themselves but can be used, often in conjunction with other datasets or information, to identify individuals and thus present a privacy risk in data sharing and analysis. Identifying QIDs is important in developing proper strategies for anonymization and data sanitization. This paper proposes QIDLearningLib, a *Python* library that offers a set of metrics and tools to measure the qualities of QIDs and identify them in data sets. It incorporates metrics from different domains – causality, privacy, data utility, and performance – to offer a holistic assessment of the properties of attributes of a given tabular dataset. Furthermore, QIDLearningLib offers visual analysis tools to present how these metrics shift over a dataset and implements an extensible framework that employs multiple optimization algorithms such as anevolutionary algorithm, simulated annealing, and greedy search using these metrics to identify a meaningful set of QIDs.

1. Introduction

As digitization of data is on the rise, maintaining the privacy of individuals has emerged as a serious concern. Sensitive information is normally hidden within massive data sets, and thus the necessity arises to implement sound measures that will act against the risk of unauthorized access and re-identification. One of the critical elements of privacy risk assessment is the identification of QIDs —attributes that, together, can identify individuals through cross-matching external data sets. For instance, features such as age, sex, and ZIP code in isolation can appear harmless but are very revealing when taken collectively [19]. Examples in the real world have illustrated the severity of this issue, for instance, being able to re-identify individuals in ostensibly anonymized medical data sets [17]. In addition, it has been found through research that even de-identified data can be re-identified through the use of QIDs with significant risk in fields like healthcare, finance, and the social sciences. For example, [17] shows that Netflix movie ratings with outside information would reveal user names. Similarly, [1] found that 90% of a group of individuals in a mobility data set could be identified from just four spatiotemporal coordinates. These findings highlight the need for systematic methods for finding and addressing QIDs correctly.

Anonymization techniques counteract these attacks by modifying or generalizing QIDs to prevent re-identification. However, their effectiveness is dependent on a proper selection of QIDs, which is to be done *a priori*. The wrong selection of QIDs will not translate into real privacy risks, ruining the inherent presumptions of anonymization methods and providing a false sense of security. Also, a solid QID choice should preserve data utility as too frequent changes could render datasets inadequate for meaningful analysis.

To address these challenges, we introduce QIDLearningLib, a Python library designed to facilitate systematic identification and evaluation of QIDs. The library integrates statistical methods, causal inference, set theory, and privacy metrics to ascertain the threats posed by single attributes or sets thereof. It has provided metrics for the quantification of attribute distinctness, entropy, and re-identification danger [2], enabling users to make informed data anonymization and sharing decisions. Moreover, QIDLearningLib can enable machine learning algorithms to support automatic QID detection and promote the security of sensitive data. Performance metrics are offered to cross-check overlaps between predicted and ground truth QIDs, and strict evaluation of selection methodologies.

Besides its applicability, QIDLearningLib is also an educational tool. Providing tools to analyze and understand QIDs, enables researchers, developers, and educators to explore the subtleties of data privacy and develop more effective privacy-protecting techniques. Furthermore, the library facilitates adherence to regulatory frameworks such as the General Data Protection Regulation (GDPR) [20], which mandate efficient data protection practices.

The rest of this paper has the following organization. Section 3 provides background on privacy models, data utility, and causal inference methods that are relevant to QID identification. Section 4 presents the essential features of QIDLearningLib, including its metrics, optimization algorithms, and evaluation methods. Section 5 shows specific application examples of the library by describing how it might be applied in real data sets. Finally, Section 6 concludes this paper by giving an overview of some significant results and

*Corresponding author
sanchosimoes@student.dei.uc.pt (S.A. Simões); jvilela@fc.up.pt (J.P. Vilela); miriam.seoane@inesctec.pt (M.S. Santos); pha@dei.uc.pt (P.H. Abreu)
ORCID(s): 0000-0001-6013-4754 (S.A. Simões); 0000-0001-5805-1351 (J.P. Vilela); 0000-0002-5912-963X (M.S. Santos); 0000-0002-9278-8194 (P.H. Abreu)

discussing directions for future improvement on QIDLearningLib.

2. Main Goals

QIDLearningLib is developed with the following goals to address challenges in data privacy:

- **Automation:** The library aims to automate accurately and efficiently the identification of QIDs in diverse datasets, by employing statistical, causal, and optimization techniques;
- **Data Privacy/Utility:** The library offers tools to evaluate the overall privacy and utility level of a dataset. It considers the interactions between QIDs and their potential impact on data privacy and utility.
- **Evaluation:** QIDLearningLib provides a set of evaluation metrics to assess the effectiveness of QID detection algorithms. These metrics also contribute to understanding the privacy risks associated with a dataset;
- **Educational Utility:** In addition to its practical applications, QIDLearningLib serves as an educational resource. It can be used in academic settings to explore concepts related to causality, data utility, and data privacy;

In summary, QIDLearningLib aims to advance the identification, evaluation, and integration of QIDs in data privacy frameworks, providing a toolset for both practical and educational purposes.

3. Background Knowledge

The spread of data-driven technologies has increased worries regarding the privacy of individuals, especially in situations where datasets with sensitive data are shared or analyzed. One of the key challenges is how to detect and mitigate the risks associated with *QIDs* that, when merged, have the potential to re-identify individuals via linkage with outside datasets [19]. Tackling these risks demands a comprehensive understanding of privacy models, anonymization methods [4], QID detection methods, and metrics [5] for assessing both utility and privacy.

Privacy models such as k-anonymity [19], l-diversity [12], and t-closeness [3] provide formal specifications for quantifying re-identification risks. A dataset is considered to satisfy k-anonymity if each record is indistinguishable from at least $k-1$ others based on QIDs, ensuring a minimal group size to prevent trivial re-identification. However, *k-anonymity* does not protect against attribute disclosure, which is thwarted by *l-diversity* (ensuring diversity of sensitive attributes within groups) and *t-closeness* (limiting the difference between sensitive attribute distributions of the anonymized and original data). These models guide the application of anonymization techniques like *generalization*, *suppression*, and *perturbation* [13], which reduce granularity or introduce noise to hide identifiable patterns.

Several software tools have also been developed to implement these privacy models and facilitate QID detection and anonymization. *ARX* [6] is one of the most widely used open-source anonymization tools that provides a comprehensive framework for QID-based anonymization, supporting *k-anonymity*, *l-diversity*, and *t-closeness*, as well as differential privacy mechanisms. *Amnesia* [7] supports automatic QID discovery and hierarchical generalization techniques, allowing users to balance privacy and data utility through interactive anonymization workflows. The *UTD Anonymization Toolbox* [8] is a set of algorithms for QID suppression, generalization, and perturbation with a focus on large-scale anonymization techniques for big data. In addition to anonymization software, several special-purpose frameworks for QID detection have been suggested. Samarati and Sweeney's algorithm [19] provides an early, rule-based approach to discovering candidate QIDs based on uniqueness and minimality constraints. These software packages reflect the growing need for automated, scalable frameworks for QID detection in high-dimensional data.

Causal inference techniques are increasingly relevant to the problem of estimating the effect of QIDs on sensitive attributes or treatment effects.

Techniques such as covariate shift analysis [9], propensity score matching [9], and balance testing [11] identify attributes that disproportionately affect outcomes in observational studies. For instance, covariate shift metrics compute QID distribution differences between treated and control groups, while propensity score overlap metrics comparability between groups. These techniques are necessary to avoid QIDs inducing biases in causal analyses. Data utility measures, usually used to ascertain the performance of a classification system, e.g., Mean Squared Error (MSE) and accuracy, are used here to ascertain the degree to which properly anonymized datasets retain their analytical utility and the degree of predictive power QIDs possess.

MSE ascertains the precision of predictions for a target attribute, and accuracy ascertains the proportion of correct classifications based on QID groupings [16]. Range utility, distinct values utility, and completeness utility also quantify the variability, uniqueness, and quality of the data of groups, respectively. Trading these measures against privacy requirements is a commonly cited difficulty [22], as excessive anonymization can render datasets useless for analysis. Performance metrics of QID detection systems, including precision, recall, and F1-score, evaluate the extent to which algorithms retrieve true QIDs. Jaccard similarity and Dice coefficient measure overlap between predicted and true QIDs [14], while distinction and separation metrics aim for the uniqueness and discriminating power of attributes [15]. These metrics are essential to validate automatic detection systems and ensure their reliability for real-world applications.

Finally, the utility-privacy trade-off emphasizes the need for integrated frameworks that bring together statistical rigor, domain knowledge, and causal reasoning.

Current research concentrates on adaptive anonymization techniques that change with dataset characteristics [21], and machine learning-based strategies for identifying high-risk QIDs. Whereas available software like *ARX, Amnesia, UTD Anonymization Toolbox* offer solutions primarily for QID anonymization, QIDLearningLib seeks to augment these efforts by incorporating more metrics and optimization techniques for automated QID detection and assessment. Leveraging the privacy models and causal inference techniques presented, QIDLearningLib instantiates these frameworks in a set of metrics and optimization algorithms, described in the following section.

4. Features

QIDLearningLib is a framework that provides a system for quantitative assessment and QID selection with various functionalities. It supports a collection of metrics that include data privacy, data utility, and causality along with performance and QID-specific metrics. The metrics are encapsulated with their statistical properties to enable systematized assessment. The library supports metric selection by analyzing the correlation structure among metrics for various combinations of attributes, identifying redundancy and relevance in the data. Also incorporated are test scripts and visualization features to examine how metrics vary along the dataset and some optimization methods that make use of the metrics implemented to compute an appropriate set of QIDs. This section elaborates on these features in detail, including their implementation and application in the framework.

4.1. Metrics

The metrics in QIDLearningLib (tables 1 and 2) were developed for QID recognition and performance assessment of QID identifying systems against a ground-truth set of QIDs. Using the metrics from table 1 approach avoids the need for an ad hoc assessment of identification risk reduction and data utility preservation through multiple anonymization iterations, which can be computationally expensive. By evaluating QIDs in advance, these metrics facilitate a more systematic and efficient selection process, ensuring that the trade-offs between privacy and utility are analyzed before anonymization is performed. Most of the metrics from table 1 are calculated relatively to groups originated by grouping the dataset by the given QIDs ($Q_i \in Q$), hence the g in subscript, denoting generically one of these groups.

It is important to note that the metrics included in QIDLearningLib are derived from original definitions spanning causality, data utility, data privacy, performance, and QID-specific evaluations. These metrics have been adapted to incorporate the considerations associated with QIDs, aiming to provide complete insights into data characteristics and model performance. These metrics, while rooted in established statistical, causal inference, and privacy-preservation literature, were systematically adapted to address the unique challenges of QID recognition.

Traditional privacy frameworks such as *k-anonymity* [19], *l-diversity* [12], and *t-closeness* [3], originally designed to evaluate privacy risks in anonymized datasets, were redefined as dynamic metrics to assess re-identification risks associated with specific attribute combinations. For instance, rather than applying k-anonymity globally to an entire dataset, we computed it hierarchically for candidate QID sets to quantify the minimum equivalence class size generated by those attributes, thereby directly linking privacy risks to attribute combinations. Similarly, causal inference metrics such as *covariate shift* [10] and *propensity score overlap* [9], traditionally used to measure treatment-effect biases, were repurposed to evaluate distributional disparities within subgroups defined by QIDs. This adaptation enables the detection of attributes that disproportionately influence subgroup distributions—a critical signal for identifying potential QIDs. Utility metrics like *mean squared error (MSE)* and *accuracy* [27], conventionally applied to assess global model performance, were instead computed locally within QID-defined equivalence classes to ensure that predictive power is preserved at the subgroup level. Performance metrics such as Jaccard similarity and precision-recall, typically used in information retrieval, were adapted to compare predicted QID sets against ground truth labels, enabling direct evaluation of recognition accuracy. Crucially, privacy models were operationalized as evaluative metrics: for example, t-closeness, originally a privacy criterion, was transformed into a measurable divergence score between sensitive attribute distributions within QID groups and the overall dataset. This adaptation - refocusing established methodologies on attribute combinations rather than individual attributes or global properties — ensures that QID recognition balances granular privacy risks (e.g., small equivalence classes) with data utility (e.g., subgroup-level predictive validity), providing a robust framework for privacy-sensitive data analysis. Additionally, metrics such as the performance ones were developed to work with sets (the real QIDs and the predicted ones), therefore making use of several set operations. In this case, this methodology allowed us to evaluate the performance of a QID recognition system.

Given the specialized nature of these adaptations, further consultation of QIDLearningLib documentation is recommended to understand the precise implementation and interpretation of these metrics in various contexts.

Causality Metrics

Causality metrics in QIDLearningLib are designed to identify attributes that influence treatment effects, which are potential candidates for being QIDs. These metrics help in understanding the causal relationships between QIDs, sensitive attributes, and the dataset itself.

The **Covariate Shift** metric [10] quantifies the disparity in the distribution of QIDs across treated and control groups. This is calculated as:

$$\mathrm{CS}_g = \mathrm{KS}_g + D_g$$

Table 1
Overview of Designed and Implemented Metrics

Area	Metric	Formula	Description
Causality	Covariate Shift	$CS_g = KS_g + D_g$	Shift in QID distribution
	Balance Test	$T_g = \frac{\bar{X}_{T,g} - \bar{X}_{C,g}}{SE_g}$	Effect size test
	Propensity Overlap	$O_g = \frac{1}{N_T} \sum \lvert p_{T,i} - \bar{p}_C \rvert$	Degree of shared propensity score
	Causal Importance	$\sum_{q \in \text{QIDs}} \left(\sum_{i=1}^{n} \left\lvert A_{q,i} \right\rvert \right)$	Sum of causal relationships involving QIDs
		Maximize	
Data Privacy	k-Anonymity	$k_g = \lvert g \rvert$	Group size
	l-Diversity	$l_g = \lvert U(g,S) \rvert$	Unique sensitive
	t-Closeness	$t_g = D_{KL}(P_S \lvert\lvert P_{S\lvert g})$	Divergence from dataset-wide sensitive dist.
	δ-Presence	$\delta_g = \lvert P(g,S) - P(D,S) \rvert$	Sensitive attr. inference risk
	Generalization Ratio	$GR_g = \lvert P(g,S) - P(D,S) \rvert$	Distribution variation
		Minimize	
Data Utility	Mean Squared Error (MSE)	$MSE_g = \frac{1}{n_g} \sum (y_i - \hat{y}_i)^2$	Avg. squared error in target prediction
		Maximize	
	Accuracy	$Acc_g = \frac{C_g}{N_g}$	Correct predictions fraction
	Range Utility	$R_g = \max(Y_g) - \min(Y_g)$	Range of values
	Distinct Values	$D_g = \lvert U(Y_g) \rvert$	Number of unique values
	Completeness Utility	$CU_g = \frac{N_g}{T_g}$	Fraction of non-null values
	Group Entropy	$H_g = \sum_i p_i \log p_i$	Randomness within QID groups
	Information Gain	$IG_g = H(Y) - H(Y\lvert X)$	Reduction in entropy of target attribute
	Gini Index	$GI_g = 1 - \sum_{i=1}^{k} p_i^2$	Impurity of target attribute within groups
	Attribute Length Penalty	$ALP = (1-p)^2 + p^2$	Penalty based on proportion of QIDs
		Minimize	
QID-Specific	Distinction	$\frac{\lvert Q_i \rvert}{n}$	Ratio of unique QID values
	Separation	$\frac{\sum I(Q_i \neq Q_j)}{\binom{n}{2}}$	Separability of records based on QIDs
		Maximize	

Table 2
Performance Metrics

Metric	Formula	Description
Precision	$\frac{TP}{TP+FP}$	Fraction of correctly predicted positives
Recall	$\frac{TP}{TP+FN}$	Fraction of actual positives retrieved
F1 Score	$2 \cdot \frac{\text{Precision} \cdot \text{Recall}}{\text{Precision} + \text{Recall}}$	Harmonic mean of precision and recall
Jaccard Similarity	$\frac{\lvert Q \cap Q \rvert}{\lvert Q \cup Q \rvert}$	Overlap ratio between predicted and actual QIDs
Specificity	$\frac{TN}{TN+FP}$	True negative rate
False Positive Rate	$\frac{FP}{TN+FP}$	Proportion of misclassified negatives
Dice Similarity Coefficient	$\frac{2\lvert Q \cap Q \rvert}{\lvert Q \rvert + \lvert Q \rvert}$	Measures similarity between predicted and actual QIDs
F-Beta Score	$(1+\beta^2) \frac{\text{Precision} \cdot \text{Recall}}{\beta^2 \cdot \text{Precision} + \text{Recall}}$	Weighted F-score prioritizing precision/recall

where KS_g is the Kolmogorov-Smirnov statistic, defined as:

$$\text{KS}_g = \max_x \left| F_T(x) - F_C(x) \right|$$

Here, $F_T(x)$ and $F_C(x)$ are the empirical cumulative distribution functions (CDFs) for the treated and control groups, respectively. D_g represents the overall distribution difference:

$$D_g = \sum_i \left| P_g(Q_i) - P(Q_i) \right|$$

where $P_g(Q_i)$ is the distribution in group g and $P(Q_i)$ is the overall distribution. A **maximized** Covariate Shift indicates significant disparities between distributions of QIDs in treated and control groups, suggesting these attributes may act as QIDs. Conversely, a **minimized** metric implies similar distributions, reducing the likelihood that the attributes serve as QIDs.

The **Balance Test** metric [9] assesses the balance between treated and control groups based on specified QIDs. This is calculated as:

$$\text{B}_g = T_g = \frac{\bar{X}_{T,g} - \bar{X}_{C,g}}{SE_g}$$

where T_g is the t-statistic, $\bar{X}_{T,g}$ and $\bar{X}_{C,g}$ are the means of the treated and control groups, respectively, and SE_g is the standard error of the difference. A **maximized** Balance Test

metric (high T_g) indicates significant differences between treated and control groups, suggesting the QIDs have a strong impact on treatment effects. Conversely, a **minimized** metric (low T_g) implies that the distributions are similar, indicating that the attributes may not serve as effective QIDs.

The **Propensity Score Overlap** metric [9] evaluates the degree of overlap in propensity scores between treated and control groups. This is calculated as:

$$O_g = \frac{1}{N_T} \sum_{i=1}^{N_T} |p_{T,i} - \bar{p}_C|$$

where O_g is the overlap metric for group g, N_T is the number of treated instances, $p_{T,i}$ is the propensity score for treated instance i, and $\bar{p}_C$ is the mean propensity score for the control group. A **maximized** O_g indicates a lack of overlap in propensity scores between groups, suggesting limited comparability. Conversely, a **minimized** O_g implies greater overlap, enhancing comparability between treated and control samples.

Causal Importance measures the causal importance of the given QIDs in a learned causal graph. This is calculated as:

$$\mathrm{C}_q = \sum_{q \in Q} \left(\sum_{i=1}^{n} \left| A_{q,i} \right| \right)$$

where A is the adjacency matrix of the causal graph, q represents the index of a QID in the adjacency matrix, and n is the total number of attributes. The causal graph is learned using a causal discovery algorithm (e.g., PC or GES). The causal importance is the sum of the absolute values of the adjacency matrix entries related to the QIDs. A **maximized** Causal Importance indicates that the QIDs have significant causal relationships with other attributes in the dataset, suggesting they are strong candidates for being QIDs.

Data Privacy Metrics

Data privacy metrics evaluate the re-identifiability risk of QIDs, ensuring that the selected attributes do not compromise individual privacy.

The **k-Anonymity** metric [19] quantifies the indistinguishability of individuals within a dataset based on specified QIDs. This is calculated as:

$$k_g = |g|$$

where k_g is the size of group g formed by the QIDs. A low value of k_g indicates that groups formed by the QIDs are small, leading to high distinguishability of individuals within those groups. This implies that individuals can potentially be re-identified due to the lack of sufficient anonymity provided by the QIDs. Therefore, the objective is to **minimize** k_g when assessing the privacy risk associated with QIDs, as lower values indicate a greater risk of re-identification.

The **l-Diversity** metric [12] quantifies the diversity of sensitive attributes within groups defined by QIDs. This is calculated as:

$$l_g = |U(g, S)|$$

where l_g is the number of unique sensitive values in group g. A low value of l_g suggests that there are few unique values for the sensitive attribute within the groups formed by the QIDs. This lack of diversity can lead to increased risks of attribute disclosure, as it may allow an adversary to infer sensitive information about individuals in those groups. Therefore, the objective is to **minimize** l_g when assessing the privacy risk associated with QIDs and sensitive attributes, as lower values indicate a greater risk of exposing sensitive information.

The **t-Closeness** metric [3] evaluates the Kullback-Leibler (KL) divergence between the distribution of a sensitive attribute within a specific group defined by QIDs and the overall distribution of that sensitive attribute across the entire dataset. This is calculated as:

$$T(g) = D_{KL}(P(g), P(S)) = \sum_i P(g_i) \log \frac{P(g_i)}{P(S_i)}$$

where $P(g)$ is the probability distribution of the sensitive attribute S within group g, and $P(S)$ is the overall probability distribution of the sensitive attribute S across the dataset. A higher value of $T(g)$ indicates a significant divergence between the distribution of the sensitive attribute within group g and the overall distribution. This suggests that the group may contain a sensitive attribute distribution that is less similar to the general population, thus increasing the risk of re-identification or exposure of sensitive information. The objective is to **minimize** $T(g)$ when assessing privacy risk, as greater divergence implies a higher potential for privacy breaches.

The δ-**Presence** metric [19] measures the difference in the presence of sensitive attribute values between the overall dataset and the individual groups defined by QIDs. This is calculated as:

$$\delta_g = |P(g, S) - P(D, S)|$$

where $P(g, S)$ is the distribution of the sensitive attribute S within group g, and $P(D, S)$ is the overall distribution of the sensitive attribute S in the dataset D. A high value of δ_g indicates significant differences between the distributions of sensitive attributes in groups compared to the overall dataset. This suggests that individuals in a specific group may be at a higher risk of having their sensitive information disclosed. Therefore, the goal is to **minimize** δ_g to ensure that the presence of sensitive attributes is consistent across the dataset and its groups.

The **Generalization Ratio** metric [15] assesses the overall difference in the distribution of sensitive attributes between the entire dataset and its grouped subsets based on QIDs. This is calculated as:

$$GR_g = |P(g, S) - P(D, S)|$$

A high Generalization Ratio indicates that the distributions of sensitive attributes in the groups differ significantly from the overall distribution, which can lead to potential disclosure risks. It suggests that individuals in specific groups may be more easily re-identified based on the sensitive attributes they

possess, due to the lack of uniformity in their representation. Therefore, the objective is to **minimize** GR_g to ensure that the selected QIDs do not significantly alter the distribution of sensitive attributes.

Data Utility Metrics

Data utility metrics evaluate the impact of QIDs on the overall utility of the dataset, ensuring that the selected attributes do not lead to significant data loss during anonymization.

The **Mean Squared Error (MSE)** metric [27] quantifies the average squared difference between the true values and the predicted values for a target attribute, grouped by specified QIDs. This is calculated as:

$$\text{MSE}_g = \frac{1}{n_g} \sum_{i=1}^{n_g} (y_i - \hat{y}_i)^2$$

where MSE_g is the Mean Squared Error for group g, y_i are the true values corresponding to the target attribute for group g, $\hat{y}_i$ are the predicted values (mean of the target attribute) for group g, and n_g is the number of instances in group g. A **minimized** Mean Squared Error indicates that the QIDs chosen lead to more accurate predictions of the target attribute, reflecting a better representation of the underlying data patterns. Conversely, a **maximized** MSE suggests that the selected QIDs may not adequately capture the variance of the target attribute, indicating that those attributes when anonymized do not lead to a great data loss regarding the prediction of the selected target attribute. Thus, the goal is to select QIDs that maximize the MSE, to avoid high data loss while maintaining privacy.

The **Accuracy** metric [27] measures the proportion of correctly predicted instances for a given target attribute, based on the mode of the target within specified QIDs. This is calculated as:

$$\text{Acc}_g = \frac{C_g}{N_g}$$

where Accuracy_g is the accuracy for group g, C_g is the count of correctly predicted instances in group g, and N_g is the total number of instances in group g. A **maximized** Accuracy indicates that the chosen QIDs provide a good representation of the target attribute, leading to high predictive performance. This suggests that the attributes are effective for making accurate predictions, enhancing their suitability as QIDs. However, achieving very high accuracy using QIDs may lead to higher data loss during the anonymization process, as it can imply that the attributes retain too much information about the individuals. Conversely, a **minimized** accuracy value signifies that the QIDs may not be effectively capturing the relevant patterns in the data, implying that those attributes might not serve well as QIDs.

The **Range Utility** metric measures the range (difference between the maximum and minimum values) of the target attribute within groups defined by QIDs. This is calculated as:

$$\text{RU}_g = R_g = \max(Y_g) - \min(Y_g)$$

where R_g is the range utility for group g, and Y_g represents the values of the target attribute within group g. A **minimized** Range Utility indicates that the selected QIDs result in a narrower range of values in the target attribute, helping to prevent significant data loss when the dataset is anonymized. This ensures that the chosen QIDs do not excessively contribute to variability, reducing the risk of unnecessary information retention that could compromise privacy. Conversely, a **maximized** Range Utility would imply that the selected attributes allow for a wider range of values, which may preserve more information but could also increase the risk of re-identification.

The **Distinct Values Utility** metric measures the number of unique values present in the target attribute within groups defined by QIDs. This is calculated as:

$$\text{DVU}_g = D_g = |U(Y_g)|$$

where D_g is the distinct values utility for group g, and $U(Y_g)$ denotes the unique values in Y_g. A **minimized** Distinct Values Utility indicates that the chosen QIDs result in fewer unique values in the target attribute, helping to prevent significant data loss while ensuring effective anonymization. This reduction limits the potential for re-identification by decreasing the granularity of the data. Conversely, a **maximized** Distinct Values Utility would imply that the selected attributes retain a greater variety of unique values, which may preserve more information but could also increase privacy risks by making the dataset more distinguishable.

The **Completeness Utility** metric measures the proportion of instances in the target attribute that have non-null values within groups defined by QIDs. This is calculated as:

$$\text{CU}_g = CU_g = \frac{N_g}{T_g}$$

where CU_g is the completeness utility for group g, N_g is the number of non-null instances in the target attribute for group g, and T_g is the total number of instances in group g. A **minimized** Completeness Utility indicates that the chosen QIDs result in less data loss when the dataset is anonymized, ensuring that a higher proportion of valid, non-null values are retained in the target attribute. This helps maintain data quality and usability while reducing unnecessary suppression or distortion. Conversely, a **maximized** Completeness Utility would imply that anonymization leads to a significant loss of valid data, suggesting that the selected QIDs cause excessive removal or modification of information, which may hinder meaningful analysis.

The **Entropy** metric measures the randomness or disorder within groups formed by QIDs. This is calculated as:

$$\text{E}_g = \sum_{g \in G} p(g) \cdot H(g)$$

where G is the set of groups formed by the QIDs, $p(g)$ is the proportion of records in group g, and $H(g)$ is the entropy of the QIDs' distribution within group g. A **minimized** Group Entropy indicates that the groups formed by the QIDs are

more homogeneous, reducing the risk of information loss during anonymization.

The **Information Gain** metric measures how much information the QIDs provide about the target attribute. It is calculated as:

$$\mathrm{IG}_g = H(Y) - H(Y|X)$$

where $H(Y)$ is the entropy of the target attribute Y, and $H(Y|X)$ is the conditional entropy of Y given the QIDs X. A **maximized** Information Gain indicates that the QIDs provide significant information about the target attribute, enhancing their suitability as QIDs.

The **Gini Index** metric measures the impurity of the target attribute within each group formed by the QIDs. This is calculated as:

$$\mathrm{GI}_g = 1 - \sum_{i=1}^{k} p_i^2$$

where p_i is the proportion of the i-th unique value of the target attribute within a group, and k is the number of unique values. A **minimized** Gini Index indicates that the groups formed by the QIDs are purer, with more homogenous target attribute values, reducing the risk of information loss during anonymization.

The **Attribute Length Penalty** metric calculates a penalty based on the proportion of QIDs in the total set of attributes. This is calculated as:

$$\mathrm{ALP}_Q = (1-p)^2 + p^2$$

where p is the proportion of QIDs relative to the total number of attributes. A **minimized** Attribute Length Penalty indicates a better balance between QIDs and other attributes, reducing the risk of excessive information loss during anonymization.

QID-Specific Metrics

QID-specific metrics focus on the properties of QIDs themselves, providing insights into their uniqueness and separability.

The **Distinction** metric [32] measures the ratio between the number of unique values for the QIDs and the total number of records in the dataset. This is calculated as:

$$\mathrm{D}_Q = \frac{|\{Q_i\}|}{n}$$

where $|\{Q_i\}|$ is the number of unique values for the QIDs, and n is the total number of records in the dataset. A **maximized** Distinction indicates that the QIDs have a high number of unique values, making them more distinguishable and potentially increasing re-identifiability risk. Pairing this with **Completeness Utility** (*minimize*) helps to reduce data loss during anonymization.

Example: Let df be a DataFrame with the following values for attributes A and B:

A	B
1	X
2	Y
1	X
2	Y

In this case, the unique QIDs are $\{(1, X), (2, Y)\}$. Therefore, the Distinction value can be calculated as:

$$\mathrm{D} = \frac{2}{4} \times 100 = 50\%$$

The **Separation** metric [32] measures the ratio between pairs of records that have at least one differing value for their QIDs and the total number of ways that two different records can be paired. This is calculated as:

$$\mathrm{S}_Q = \frac{\sum_{i=1}^{n-1} \sum_{j=i+1}^{n} \mathbb{I}(Q_i \neq Q_j)}{\binom{n}{2}}$$

where Q_i and Q_j are the QIDs of records i and j, respectively, $\mathbb{I}(Q_i \neq Q_j)$ is an indicator function that equals 1 if Q_i and Q_j differ and 0 otherwise, and $\binom{n}{2} = \frac{n(n-1)}{2}$ is the total number of ways to choose pairs from n records. A **maximized** Separation indicates that the QIDs effectively separate records, enhancing their uniqueness and re-identifiability risk.

Example: Consider the same DataFrame df:

A	B
1	X
2	Y
1	Y
2	X

The number of differing pairs of records can be calculated, and using the total number of pairs, the Separation metric can be computed as follows:

Let: - $n = 4$ (total records), - Number of differing pairs = 6, - Total pairs = $\binom{4}{2} = 6$.

Then:

$$\mathrm{S}_Q = \frac{6}{6} \times 100 = 100\%$$

Performance Metrics

Evaluate the effectiveness of the QID recognition system by comparing results with ground truth through traditional performance metrics [30, 31] but adapted to sets (of the real and predicted QIDs). Let $\hat{Q}$ be the set of predicted QIDs, and Q be the set of actual QIDs.

Then, we have:

- $TP = |\hat{Q} \cap Q|$: Number of true positives (correctly predicted QIDs).
- $FP = |\hat{Q} - Q|$: Number of false positives (incorrectly predicted as QIDs).

- $TN = |U - (Q \cup \hat{Q})|$: Number of true negatives (correctly predicted as non-QIDs, assuming U is defined as the universal set of attributes).
- $FN = |Q - \hat{Q}|$: Number of false negatives (actual QIDs not predicted).

1. **Specificity**

$$\text{Specificity} = \frac{TN}{TN + FP}$$

2. **False Positive Rate (FPR)**

$$\text{FPR} = \frac{FP}{TN + FP}$$

3. **Precision**

$$\text{Precision} = \frac{TP}{TP + FP}$$

4. **Recall**

$$\text{Recall} = \frac{TP}{TP + FN}$$

5. **F1 Score**

$$F_1 = 2 \cdot \frac{\text{Precision} \cdot \text{Recall}}{\text{Precision} + \text{Recall}}$$

6. **F-Beta Score (Generalized F-Score)**

$$F_\beta = (1 + \beta^2) \cdot \frac{\text{Precision} \cdot \text{Recall}}{\beta^2 \cdot \text{Precision} + \text{Recall}}$$

7. **Jaccard Similarity**

$$\text{Jaccard Similarity} = \frac{|\hat{Q} \cap Q|}{|\hat{Q} \cup Q|}$$

8. **Dice Similarity Coefficient** Given parameters α and β, the Dice Similarity Coefficient is:

$$\text{Dice Similarity} = \frac{\alpha \cdot |\hat{Q} \cap Q|}{\alpha \cdot |\hat{Q}|^\beta + |Q|^\beta}$$

9. **Accuracy**

$$\text{Accuracy} = \frac{|\hat{Q} \cap Q|}{|Q|}$$

4.2. Test Scripts

QIDLearningLib includes test scripts that facilitate the evaluation of the aforementioned metrics. Users can apply these scripts to imported datasets or generate simple dummy datasets using library primitives for testing and validation.

4.3. Metric Encapsulation

The library provides an object-oriented approach, encapsulating causality, data utility, data privacy, and performance metrics. This allows users to inspect relevant statistics about the metrics and understand the data.

4.4. Metric Selection

QIDLearningLib also provides primitives to assess the metrics' relevance for a given dataset, thus, selecting them. This is based on the premise that causality, data privacy, and data utility are not independent domains when applied to QIDs. These domains exhibit interdependencies, meaning that a change in one may indirectly affect another. Consequently, when selecting an appropriate set of metrics to identify QIDs, users should prioritize metrics that enhance separability. This involves choosing metrics across different domains as well as selecting metrics within the same domain that exhibit lower correlation, thereby reducing redundancy and ensuring a more complete evaluation of the dataset.

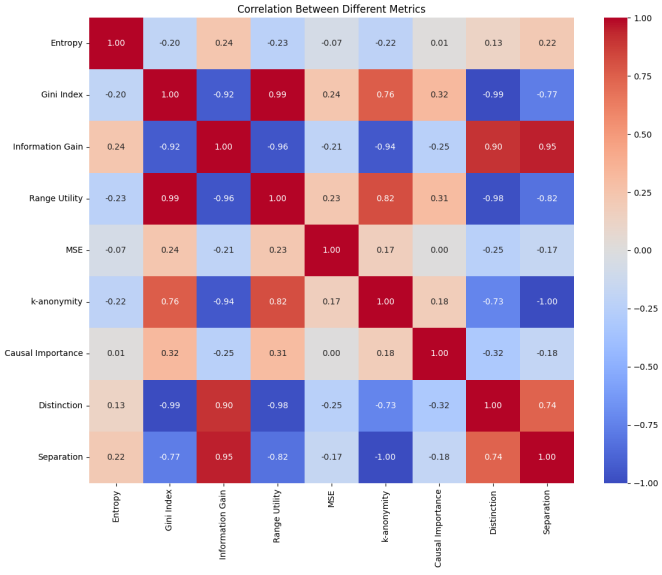


Figure 1: Metric Correlation Matrix

Figure 1 presents the correlation matrix indicating correlations among different metrics in causality, data privacy, data utility, and QID-specific domains. The metrics were calculated on a synthetically generated dataset with **10,000 samples** and **11 attributes**. For easier correlation analysis, all the metrics were calculated for all possible attribute combinations of **4 attributes** with the same attribute order across all calculations. Causal Importance is the only metric in the causality domain. Its correlations with the other metrics are very small, the largest of which is with Range Utility (0.31). This means that Causal Importance is capturing different data features, so it is a good independent feature for causality analysis. For data privacy, k-Anonymity is selected because it has a strong negative correlation with Separation (-1.00) and Distinction (-0.73), i.e., anonymity increase makes QID less distinguishable. The Gini Index, as a privacy measure, is highly related to Range Utility (0.99) and highly negatively correlated with Information Gain (-0.92), which is the established utility-privacy trade-off. Range Utility is the most indicative characteristic among the data utility measures. It has a high correlation with Gini Index (0.99) and high negative correlation with Information Gain (-0.96),

which mirrors how utility preservation and privacy protection both exhibit opposite tendencies. For QID-specific measures, Distinction and Separation are selected. The high positive correlation between them (0.74) signifies the tendency of both measures to co-occur in QID identification problems. Their high negative correlations with k-Anonymity (-0.73 and -1.00, respectively) also suggest their usefulness in measuring QID uniqueness under privacy constraints. As QID identification problems are typically interested in the two measures collectively, both are selected as complementary features. To achieve a non-overlapping and heterogeneous set of features, the ultimate selection is to use **Causal Importance** (causality), **k-Anonymity** (privacy), **Range Utility** (utility), and both **Distinction** and **Separation** (QID-specific measures). This decision ensures that all domains are addressed without becoming redundant.

4.5. Graphication

This feature enables users to visually inspect the distribution of values of a specific metric across the dataset, grouped by the chosen attributes. By generating graphs such as histograms, empirical CDF, and normal distribution's CDF and Q-Q plot (figure 2). Graphication facilitates a deeper understanding of how the metric behaves across different segments of the data. This visual analysis aids in identifying patterns, outliers, and trends, providing valuable insights into the dataset's characteristics and the impact of various factors on the metric's values.

4.6. Optimization Algorithms:

QIDLearningLib includes optimization algorithms to automate the selection of QIDs based on user-defined criteria. These algorithms evaluate candidate sets of QIDs using a fitness function that incorporates multiple privacy, causality, and data utility metrics. By using these methods, users can efficiently identify attribute combinations that balance privacy protection with data usability. Regarding the evaluation of population individuals in the evolutionary algorithm and the evaluation of neighbours in tabu search and greedy search, thread parallelization is used to speed up the performed computations, making the the optimization process more scalable and efficient.

Let $X \subset \{0,1\}^n$ denote the space of candidate solutions, where each candidate $x \in X$ represents a potential set of QIDs. The overall fitness function is defined as a linear combination of the metrics provided by the library:

$$F(x) = \sum_{i=1}^{m} w_i \cdot M_i(x),$$

where each $M_i(x)$ is a metric computed according to the definitions in QIDLearningLib and the weights w_i are selected to balance the trade-off between privacy protection and data utility. The library includes several optimization algorithms to identify the candidate x^* that maximizes $F(x)$. Their high-level formulations are as follows:

Evolutionary Algorithm: A population $P_0 \subset X$ is randomly initialized. For each generation $t = 0, 1, \ldots, T$, a new population is generated by applying selection, crossover, and mutation operators:

$$P_{t+1} = \text{Mutate}\Big(\text{Crossover}\big(\text{Select}(P_t)\big)\Big).$$

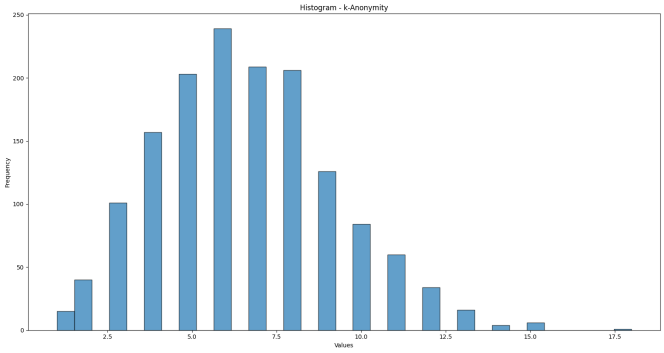
The best candidate is then given by:

$$x^* = \arg\max_{x \in P_T} F(x).$$

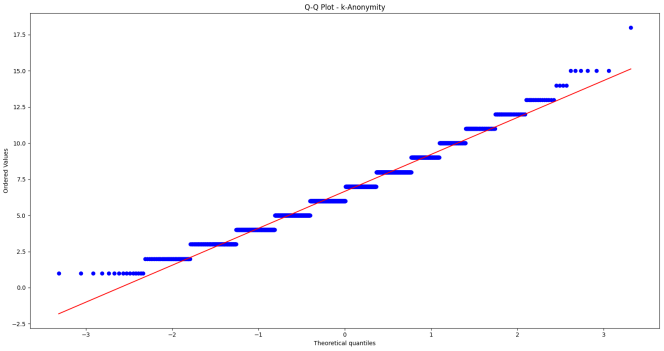
(See [26, 33] for foundational work in evolutionary algorithms.)

Tabu Search: Starting from an initial candidate $x^0 \in X$, Tabu Search iteratively selects the best neighbor $x' \in N(x^t) \setminus T$ not present in the tabu list T:

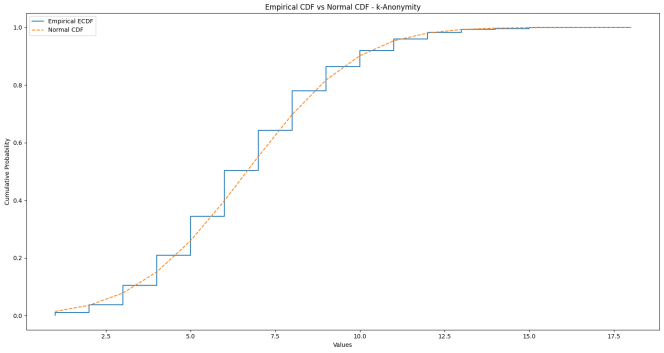
$$x^{t+1} = \arg\max_{x \in N(x^t) \setminus T} F(x).$$



(a) Histogram



(b) Q-Q Plot



(c) CDF Plot

Figure 2: Visualization of the k-Anonymity metric graphs for the Adult Income dataset.

The tabu list is dynamically updated to avoid cycles, and the process terminates when a stopping condition is met [34].

Greedy Search: Greedy Search evaluates the entire neighborhood $N(x)$ of the current candidate x and moves to the neighbor with the highest fitness:

$$x^{t+1} = \arg\max_{x' \in N(x)} F(x'),$$

terminating when no improvement is found.

Simulated Annealing: Beginning with an initial candidate x^0 and temperature T_0, Simulated Annealing iteratively selects a neighbor $x' \in N(x^t)$ and accepts it with probability

$$P(x^t \to x') = \begin{cases} 1, & \text{if } F(x') > F(x^t), \\ \exp\left(\frac{F(x') - F(x^t)}{T_t}\right), & \text{otherwise,} \end{cases}$$

while the temperature is updated as $T_{t+1} = \alpha T_t$ with $\alpha \in (0, 1)$. The algorithm terminates upon meeting a predefined criterion [35].

5. Use Cases

This section presents two demonstration use cases showing the usability and functionality of the QIDLearningLib framework in privacy-preserving data analysis. Both the use cases were intended to empirically verify whether the framework is effective in determining QID attribute combinations that conserve privacy risk while preserving utility. The examples are themselves employed to illustrate the relevance of methods conducted on synthetically produced data imitating real scenarios.

The first application (section 5.1) employs the framework on a mock medical dataset to evaluate pre-specified candidate QID sets against a set of specified metrics. The example provides a walkthrough of the metrics used to identify the fitness of different sets of attributes as QIDs. The example also shows how the framework quantifies the privacy risks for each set and identifies the optimal set of QIDs from a weighted fitness function.

The second use case (section 5.2) explores automated QID discovery in large datasets by an evolutionary algorithm (EA), which is one of the deployed optimization algorithms for the selection of QIDs in the library. The approach addresses the exponential search space of combinations of QIDs by leveraging bio-inspired optimization algorithms. The EA iteratively enhances a population of collections of candidate attributes according to the QIDLearningLib metrics. The specifics of how the EA is carried out, i.e., initialized, mutated, crossed, and selected, are described as well as the result obtained after the optimization process.

Together, these use cases introduce a thorough evaluation of the potential of the framework that points to both the practical usability feasibility and potential for QID discovery automation in privacy-protected applications.

5.1. Practical Example using a Medical Dataset

Here, we illustrate the application of QIDLearningLib on a synthetic medical dataset (table 3). The aim is to assess different combinations of attributes for their potential to serve as QIDs.

Table 3
Artificial Medical Dataset

ID	Age	Gender	Diagnosis	Treatment	Outcome
1	34	M	Hypertension	A	Improved
2	28	F	Diabetes	B	Not Improved
3	45	F	Hypertension	A	Improved
4	34	M	Diabetes	B	Improved
5	29	M	Hypertension	A	Not Improved
6	60	F	Hypertension	A	Improved
7	34	M	Diabetes	B	Improved
8	41	F	Hypertension	A	Not Improved
9	54	F	Diabetes	B	Improved
10	39	M	Hypertension	A	Not Improved

The dataset (table 3) comprises several patient-related attributes, including *Age*, *Gender*, *Diagnosis*, *Treatment*, and *Outcome* (see Table 3). For this analysis and the sake of simplicity, we consider eight candidate combinations of these attributes.

The following combinations of attributes are considered as potential QIDs:

- Combination 1: {Age, Gender}
- Combination 2: {Age, Diagnosis}
- Combination 3: {Gender, Diagnosis}
- Combination 4: {Age, Treatment}
- Combination 5: {Gender, Treatment}
- Combination 6: {Age, Gender, Treatment}
- Combination 7: {Age, Diagnosis, Treatment}
- Combination 8: {Gender, Diagnosis, Treatment}

For each combination, we compute a series of metrics that capture various aspects of the data's structure:

The **k-anonymity** metric is defined as the mean size of the groups obtained when the dataset is partitioned based on the candidate attributes. Lower values of k indicate that the attribute combination produces smaller, more specific groups, thereby enhancing the potential for re-identification. For example, when the dataset is grouped by {Age, Gender}, the mean group size is 1, which suggests that each group is highly specific.

In contrast, the **separation** metric quantifies the proportion of record pairs that differ in at least one attribute of the candidate set relative to the total number of record pairs. A higher separation value implies that the attributes effectively distinguish between records. For instance, a separation value of 0.80 for {Age, Gender} indicates that 80% of record pairs have different QID values.

The **distinction** metric is computed as the ratio of the number of unique QID combinations to the total number of records. This metric reflects the diversity of the attribute values, with higher values signaling that the attributes yield a larger variety of groupings. A distinction value of 0.75 means that 75% of the records exhibit a unique combination, which supports the discriminative capacity of the candidate QIDs.

For the **covariate shift** metric, we evaluate the difference in the distribution of the candidate attributes between treated and control groups. This metric is defined as the sum of the Kolmogorov-Smirnov statistic and the overall distribution difference:

$$\mathrm{CS}_g = \mathrm{KS}_g + D_g,$$

where

$$\mathrm{KS}_g = \max_x \left| F_T(x) - F_C(x) \right|$$

captures the maximum discrepancy between the empirical cumulative distribution functions of the treated (F_T) and control (F_C) groups, and

$$D_g = \sum_i \left| P_g(Q_i) - P(Q_i) \right|$$

measures the overall difference in the distributions. In our example, a covariate shift value of 0.60 for {Age, Gender} indicates a pronounced distributional difference, thereby enhancing the attribute combination's potential as a QID.

The **accuracy** metric assesses the predictive performance of the candidate attributes by determining the proportion of correct predictions when the mode of the target attribute is used within each group. Since a lower accuracy implies that less individual-specific information is retained (which is desirable for privacy), the fitness function employs $1 - \mathrm{Acc}$. For instance, an accuracy of 0.40 indicates that only 40% of the outcomes are correctly predicted, thus favoring a higher fitness value.

The overall fitness F for each attribute combination is then computed as a weighted linear combination of these metrics:

$$F = w_1 \cdot k + w_2 \cdot S + w_3 \cdot D + w_4 \cdot (1 - \mathrm{Acc}) + w_5 \cdot CS.$$

In our evaluation, we assign the weights as follows:

$$w_1 = 0.25, \quad w_2 = 0.25, \quad w_3 = 0.25,$$
$$w_4 = 0.15, \quad w_5 = 0.10.$$

This formulation balances the need for privacy (through low k and high $1 - \mathrm{Acc}$) with the discriminative power of the attributes (through high S, D, and CS). The numerical values for each metric, as computed on the dataset, and the corresponding fitness scores are shown in table 4.

In summary, the metric values were computed by grouping the dataset based on each candidate attribute combination, calculating the mean group size for k-anonymity, determining the proportion of distinct record pairs for separation, and evaluating the diversity of the groups for distinction. The covariate shift was derived from the combined effect of the Kolmogorov-Smirnov statistic and the overall distribution difference between treated and control groups. The accuracy metric was obtained by comparing the predicted mode of the target attribute with the actual outcomes. These metrics were then integrated into a single fitness function with appropriately chosen weights to identify the optimal QID. In this case, the combination {Age, Gender} achieves the highest overall fitness value ($F = 0.45$) and is selected for subsequent privacy-preserving analysis.

Table 4
Summary of Metrics and Fitness Values for Each Combination

Combination	k	S	D	Acc	CS	F
{Age, Gender}	1	0.80	0.75	0.40	0.60	0.45
{Age, Diagnosis}	1	0.75	0.80	0.50	0.65	0.43
{Gender, Diagnosis}	1	0.65	0.90	0.35	0.55	0.40
{Age, Treatment}	1	0.70	0.85	0.60	0.70	0.35
{Gender, Treatment}	1	0.68	0.88	0.45	0.62	0.39
{Age, Gender, Treatment }	1	0.60	0.90	0.65	0.75	0.33
{Age, Diagnosis, Treatment }	1	0.50	0.92	0.75	0.80	0.30
{Gender, Diagnosis, Treatment }	1	0.55	0.95	0.80	0.85	0.28

5.2. Application of an Evolutionary Algorithm for QID Recognition

To address the problem of QID recognition in large datasets, an evolutionary algorithm (EA) is implemented. The binary nature of the QID recognition task, along with the vast 2^D search space (where D is the number of attributes), makes evolutionary algorithms particularly suited to this challenge. Employing metrics implemented in QIDLearningLib, the EA optimizes attribute selection to identify a final set of QIDs that balance privacy, utility, and causality [23].

An EA is a bio-inspired optimization approach that iteratively refines a population of candidate solutions. Each phase below is adapted to target QID recognition with metrics designed for privacy, accuracy, and utility:

- **Initialization:** A population of binary strings representing potential QID combinations is generated, providing a diverse starting point for attribute selection [24].
- **Fitness Evaluation:** Individuals are evaluated using QIDLearningLib metrics, quantifying the suitability of each attribute set in terms of privacy, causality, and utility [25].
- **Selection:** Higher-performing solutions are selected as parents based on their fitness scores, ensuring that combinations with optimal metric performance have a greater chance of progressing [26].
- **Crossover and Mutation:** Offspring are generated by recombining attribute sets from parents, with mutations introducing variability to explore diverse attribute combinations [24].
- **Replacement and Termination:** Successive generations evolve, guided by fitness metrics from QIDLearningLib, until a predefined criterion is reached,

such as achieving a stable or maximal privacy-utility balance [23].

This EA approach effectively narrows the search for robust QID attribute combinations by optimizing privacy-utility trade-offs through QIDLearningLib metrics, providing a viable framework for secure data handling in privacy-sensitive applications.

Implementation Details

The three core components of the evolutionary algorithm were designed and implemented as follows:

1. **Representation:**
   - Each individual in the population is represented as a binary vector $\mathbf{x} = [x_1, x_2, \ldots, x_D]$, where $x_i \in \{0, 1\}$.
   - An attribute i is selected if $x_i = 1$, and not selected if $x_i = 0$.
2. **Mutation Operator:**
   - The mutation operator alters each gene x_i in the binary vector $\mathbf{x}$ with a mutation rate p_m.
   - A uniform random variable U is generated for each gene, where U is drawn from the interval $[0, 1]$. This variable determines whether the gene will be mutated or retained.
   - Mathematically, the mutated gene x_i' is defined as follows:

$$x_i' = \begin{cases} 1 - x_i & \text{if } U < p_m \\ x_i & \text{otherwise} \end{cases}$$

3. **Crossover Operator:**
   - The crossover operator combines two parent individuals $\mathbf{x}_1$ and $\mathbf{x}_2$ to produce offspring $\mathbf{x}_1'$ and $\mathbf{x}_2'$ with a crossover rate p_c.
   - For a crossover point c, the offspring are defined as:

$$\mathbf{x}_1' = \begin{cases} [x_{1,1}, \ldots, x_{1,c}, x_{2,c+1}, \ldots, x_{2,D}] & \text{if } U < p_c \\ \mathbf{x}_1 & \text{otherwise} \end{cases}$$

$$\mathbf{x}_2' = \begin{cases} [x_{2,1}, \ldots, x_{2,c}, x_{1,c+1}, \ldots, x_{1,D}] & \text{if } U < p_c \\ \mathbf{x}_2 & \text{otherwise} \end{cases}$$

4. **Fitness Function:**
   - This fitness function assesses each individual $\mathbf{x}$ based on distinction (D), separation (S), and k-anonymity (K), using the multiobjective framework in QIDLearningLib [25].
   - Given a binary vector $\mathbf{x}$ representing attribute selection, and p, the proportion of selected attributes, the function $f(\mathbf{x})$ is:

$$f(\mathbf{x}) = \frac{1}{\alpha \cdot A(\mathbf{x})} \Big(\beta_1 \cdot D(\mathbf{x}) + \beta_2 \cdot S(\mathbf{x}) - \beta_3 \cdot K(\mathbf{x}) + \beta_4 \cdot \Delta D + \beta_5 \cdot \Delta S \Big)$$

where $D(\mathbf{x})$ is the **distinction metric**, which should be maximized and quantifies unique combinations of selected QIDs to enhance identifiability. $S(\mathbf{x})$ is the **separation metric**, also maximized, and measures the differentiation across records. $K(\mathbf{x})$ is the **k-anonymity metric**, which should be minimized to evaluate group size and reduce re-identification risk. The improvements in D and S when adding an attribute are represented by $\Delta D = D(\mathbf{x}) - D(\mathbf{x}_{\text{prev}})$ and $\Delta S = S(\mathbf{x}) - S(\mathbf{x}_{\text{prev}})$, respectively. $A(\mathbf{x})$ is the **attribute length penalty**, which aims to minimize extremes and balance selection, and is defined as $A(\mathbf{x}) = (1 - p)^2 + p^2$. Finally, α is the normalization factor, and β_i are the metric weighting coefficients.

Table 5
Datasets and Predicted Quasi Identifiers (Best Individuals)

Dataset	Best Individual (x)
adult	age, workclass, fnlwgt, education, education-num, marital-status, occupation, relationship, race, sex, capital-gain, capital-loss, hours-per-week, native-country
car	buying, maint, lug_boot, eval
diabetes	time, value
heart+disease	sex, thalach, exang, oldpeak, slope, ca, thal, num
nursery	parents, has_nurs, form, housing, social, health

Table 6
Metrics for the Predicted QIDs for all Datasets

Dataset	f(x)	D(x)	S(x)	K(x)	A(x)
adult	8.99	0.99	1.00	1.00	0.88
car	90.24	0.06	0.99	16.78	0.51
diabetes	82.47	0.51	1.00	1.94	0.50
heart+disease	2.22	0.75	0.99	1.33	0.51
nursery	691.11	0.06	1.00	16.13	0.63

The EA was applied to five datasets: *adult*, *car*, *diabetes*, *heart+disease*, and *nursery*. Table 5 lists the best individuals (QID combinations) identified by the EA for each dataset. These combinations were selected based on their ability to maximize distinction and separation while minimizing k-anonymity.

Table 6 presents the fitness scores ($f(x)$) and individual metric values ($D(x)$ - distinction, $S(x)$ - separation, $K(x)$ - k-anonymity, $A(x)$) - attribute length penalty) for each dataset. The results demonstrate that the EA successfully identifies QID combinations that balance privacy and utility. For example, the *adult* dataset achieves a high distinction ($D(x) = 0.99$) and separation ($S(x) = 1.00$) with minimal k-anonymity ($K(x) = 1.00$), indicating robust QID selection. Similarly, the *nursery* dataset achieves a high fitness score ($f(x) = 691.11$), driven by strong separation and distinction metrics.

Table 7
Metrics Comparison between EA Predicted and Ground Truth QIDs

Dataset / Metric	adult	car	diabetes	heart+disease	nursery
S	0.11	0.50	0.50	0.13	0.33
FPR	0.89	0.50	0.50	0.88	0.67
P	0.47	0.50	0.50	0.13	0.33
R	0.88	0.50	0.50	0.50	0.50
F1	0.61	0.50	0.50	0.20	0.40
F2	0.74	0.50	0.50	0.31	0.45
J	0.44	0.33	0.33	0.11	0.25
D	0.61	0.50	0.50	0.20	0.40
A	0.88	0.50	0.50	0.50	0.50

Table 7 compares the EA-predicted QIDs against ground truth QIDs using performance metrics such as precision (P), recall (R), F1-score ($F1$), and Jaccard similarity (J). The results show that the EA achieves high recall ($R = 0.88$) and F1-score ($F1 = 0.61$) for the *adult* dataset, indicating strong alignment with ground truth QIDs. However, lower precision ($P = 0.47$) suggests some over-selection of attributes. The *car* and *diabetes* datasets achieve balanced precision and recall, while the *heart+disease* and *nursery* datasets exhibit lower performance, highlighting the need for further refinement in attribute selection for these datasets.

The code snippet presented in listing 1 runs a similar instance of an Evolutionary Algorithm (EA) on a dataset to identify QIDs. It compares the detected QIDs with a predefined ground truth (which is not always available) using multiple performance metrics such as specificity, recall, and accuracy.

```python
import numpy as np
import pandas as pd
from QIDLearningLib.optimizer.metric import Metric
from QIDLearningLib.optimizer.ea import
    EvolutionaryAlgorithm
from QIDLearningLib.metrics.performance import
    specificity, recall, accuracy

# Load dataset
file_path = "datasets/sample_dataset.csv"
df = pd.read_csv(file_path)

# Define ground truth QIDs
ground_truth_qids = {"Age", "Gender", "Zipcode"}  #
    Example predefined QIDs

# Define metrics
metrics = [
    Metric("Distinction", 5, Metric.distinction, maximize
        =True),
    Metric("Separation", 0.5, Metric.separation, maximize
        =True),
    Metric("k-Anonymity", -0.4, Metric.k_anonymity,
        maximize=True),
    Metric("Delta_Distinction", 0.2, Metric.
        delta_distinction, maximize=True),
    Metric("Delta_Separation", 0.2, Metric.
        delta_separation, maximize=True),
    Metric("Attribute_Length_Penalty", -1, Metric.
        attribute_length_penalty, maximize=True)
]

# Configure and execute the Evolutionary Algorithm
ea = EvolutionaryAlgorithm(df, metrics, alpha=5,
    population_size=50, generations=30,
                initial_crossover_rate=0.3,
                    initial_mutation_rate=0.2,
                elite_size=1, tournament_size=5,
                    interactive_plot=True)

best_individual, best_fitness, history = ea.run()

# Retrieve selected QIDs
selected_indices = np.where(best_individual == 1)[0]
selected_attributes = set(df.columns[selected_indices])

# Compute evaluation metrics
attributes_set = set(df.columns)  # Universal set of
    attributes
spec = specificity(selected_attributes, ground_truth_qids
    , attributes_set)
rec = recall(selected_attributes, ground_truth_qids)
acc = accuracy(selected_attributes, ground_truth_qids)

# Optionally, plot metric evolution
plot_evolution(history, metrics, ea.generations)
```

Listing 1: Python code to run the QID-selecting EA on a dataset and evaluate performance

6. Future Work and Conclusions

The growth of QIDLearningLib will also focus on maintaining progress toward resolving fundamental issues of data utility maximization and privacy. Updates in the future will also focus on adding new quantification measures of QIDs, causality-based learning, privacy risk assessment, and data utility assessment to provide an even more mathematical and systematic solution for finding QIDs. Better visualization techniques will also be pursued to make the delivered metrics more interpretable. This will consist of expanding current graphing techniques and adding new techniques to provide a better comprehension of QID selection and assessment processes. A graphical user interface (GUI) will also be developed to simplify dataset analysis and usability without compromising analytical strength. Furthermore, QIDLearningLib will be coupled with established machine learning libraries such as scikit-learn and Keras for simple integration into broader machine learning pipelines and optimization of usability in actual pipelines. Explainability features will be included to provide additional transparency into QID identification and selection processes. This will provide methods to characterize algorithmic decision-making procedures inside the library to assist users with greater insight into why the QID is found. Apart from that, scalability enhancements will be incorporated so bulk-volume datasets are handled efficiently and the library stays optimized irrespective of dataset volume variations. Dynamic metric weighting and adaptation, which see the weighting and selection change

based on dataset properties, will be incorporated to make the tool even more flexible. Because of high robustness and dependability, continuous integration ideas will be adopted, and a solid unit testing framework will be established. All these will ensure the precision and stability of the library during life cycle development.

The effectiveness of QIDLearningLib has been established by its application to simulated and actual data sets, specifically in medical data privacy applications. With the assistance of k-anonymity mechanisms, differentiation, partitioning, and covariate shift, the library helped QIDs to learn how to balance data utility and privacy threats. With this, evolutionary optimization techniques were used to find QID mixes optimal for a given fitness function based on library-provided metrics. In the future, the development of QIDLearningLib will persist in improving its methods and functionality to be able to handle data privacy issues that are present. With the introduction of state-of-the-art optimization methods, further interpretability, and integration with mainstream machine learning frameworks, the library will remain a mature and adaptive instrument for investigating and processing QIDs across datasets of all levels of complexity.

Acknowledgments

This work is partially financed through national funds by FCT - Fundação para a CiFRência e a Tecnologia, I.P., in the framework of the Project UIDB/00326/2025 and UIDP/00326/2025.

CRediT authorship contribution statement

Sancho Amaral Simões: Conceptualization; Methodology; Software; Investigation; Writing - Original Draft. **João P. Vilela:** Conceptualization; Resources; Validation; Writing - Review and Editing; Supervision. **Miriam Seoane Santos:** Visualization; Writing - Review and Editing. **Pedro Henriques Abreu:** Conceptualization; Resources; Validation; Writing - Review and Editing; Supervision.

Appendix: Installation Guide

For detailed instructions on installing and using QIDLearningLib, please refer to the documentation available on the official GitHub repository: https://github.com/smartlord7/QIDLearningLib.git.

To install the software, the repository can be cloned or downloaded as a ZIP file. The Python package of the library is located at QIDLearningLib/src/QIDLearningLib, and the documentation is available at QIDLearningLib/src/QIDLearningLib/doc/QIDLearningLib/.

In this case, these are the example commands:

```
git clone
https://github.com/smartlord7/QIDLearningLib.git
cd QIDLearningLib/src/QIDLearningLib
```

Once the repository is obtained, the user can move the package to their project directory and start using it like any other Python package.

To install `QIDLearningLib` using `pip`, execute the following command in your terminal:

```
pip install QIDLearningLib
```

Make sure you have Python and pip installed on your system. This command will download and install the `QIDLearningLib` package along with its dependencies from the Python Package Index (PyPI). It is highly recommended that documentation be consulted before using the library.

After obtaining the library, the user can integrate it into their Python project seamlessly, leveraging the functionalities provided by QIDLearningLib.

References

[1] De Montjoye, Y. A., Hidalgo, C. A., Verleysen, M., & Blondel, V. D. (2013). Unique in the crowd: The privacy bounds of human mobility. *Scientific Reports*, 3, 1–5. 10.1038/srep01376

[2] Ganta, S. R., Kasiviswanathan, S. P., & Smith, A. (2008). Composition attacks and auxiliary information in data privacy. In *Proceedings of the 14th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* (pp. 265–273). 10.1145/1401890.1401926

[3] Li, N., Li, T., & Venkatasubramanian, S. (2007). t-closeness: Privacy beyond k-anonymity and l-diversity. In *2007 IEEE 23rd International Conference on Data Engineering* (pp. 106–115).

[4] Cunha, M., Mendes, R., & Vilela, J. P. (2021). A survey of privacy-preserving mechanisms for heterogeneous data types. *Computer Science Review*, 41, 100403. 10.1016/j.cosrev.2021.100403

[5] Mendes, R., & Vilela, J. P. (2017). Privacy-preserving data mining: Methods, metrics, and applications. *IEEE Access*, 5, 10562-10582. 10.1109/ACCESS.2017.2706947

[6] Prasser, Fabian, Jannik Eicher, Hermann Spengler, and Florian Kohlmayer. "ARX—A comprehensive tool for anonymizing biomedical data." *Methods of Information in Medicine*, vol. 59, no. 03, 2020, pp. 93–108.

[7] "Amnesia - Data Anonymization." 2021. Available at: www.https://amnesia.openaire.eu/. Accessed February 2025.

[8] Data Security and Privacy Lab. (2012). *UTD Anonymization Toolbox: Privacy-preserving data publishing and analysis*. University of Texas at Dallas. Available at: https://labs.utdallas.edu/dspl/software/anonymization-toolbox/

[9] Rosenbaum, P. R., & Rubin, D. B. (1983). The central role of the propensity score in observational studies for causal effects. *Biometrika*, 70(1), 41–55.

[10] Shimodaira, H. (2000). Improving predictive inference under covariate shift by weighting the log-likelihood function. *Journal of Statistical Planning and Inference*, 90(2), 227–244.

[11] Imbens, G. W., & Rubin, D. B. (2015). *Causal Inference in Statistics, Social, and Biomedical Sciences*. Cambridge, UK: Cambridge University Press.

[12] Machanavajjhala, A., Gehrke, J., Kifer, D., & Venkitasubramaniam, M. (2006). l-diversity: Privacy beyond k-anonymity. In *22nd International Conference on Data Engineering (ICDE'06)* (pp. 24–24).

[13] Fung, B. C. M., Wang, K., Chen, R., & Yu, P. S. (2010). Privacy-preserving data publishing: A survey of recent developments. *ACM Computing Surveys*, 42(4), Article 14.

[14] Powers, D. M. W. (2020). Evaluation: From precision, recall, and F-measure to ROC, informedness, markedness, and correlation. *Journal of Machine Learning Research*, 21(1), 1–27. 10.5555/12345678

[15] LeFevre, K., DeWitt, D. J., & Ramakrishnan, R. (2006). Mondrian multidimensional k-anonymity. In *22nd International Conference on Data Engineering* (pp. 25–26).

[16] Domingo-Ferrer, J., & Torra, V. (2008). A critique of the sensitivity rules is usually employed for statistical table protection. *International Journal of Uncertainty, Fuzziness and Knowledge-Based Systems*, 16(5), 619–640.

[17] Narayanan, A., & Shmatikov, V. (2008). Robust de-anonymization of large sparse datasets. In *2008 IEEE Symposium on Security and Privacy (SP 2008)* (pp. 111–125).

[18] Pearl, J. (2001). *Causality: Models, Reasoning, and Inference* (2nd ed.). Cambridge, UK: Cambridge University Press.

[19] Sweeney, L. (2002). k-anonymity: A model for protecting privacy. *International Journal of Uncertainty, Fuzziness and Knowledge-Based Systems*, 10(5), 557–570.

[20] Voigt, P., & Von dem Bussche, A. (2017). *The EU General Data Protection Regulation (GDPR): A Practical Guide*. Cham, Switzerland: Springer International Publishing.

[21] Zhu, T., Li, G., Zhou, W., & Yu, P. S. (2017). Differentially private data publishing and analysis: A survey. *IEEE Transactions on Knowledge and Data Engineering*, 29(8), 1619–1638. https://doi.org/10.1109/TKDE.2017.2697856

[22] Soria-Comas, J., Domingo-Ferrer, J., Sánchez, D., & Martínez, S. (2014). Enhancing data utility in differential privacy via microaggregation-based k-anonymity. *The VLDB Journal*, 23(5), 771–794.

[23] Back, T. (1996). *Evolutionary Algorithms in Theory and Practice*. Oxford, UK: Oxford University Press.

[24] Eiben, A. E., & Smith, J. E. (2003). *Introduction to Evolutionary Computing*. Natural Computing Series. Springer. https://doi.org/10.1007/978-3-662-05094-1

[25] Deb, K. (2001). *Multi-Objective Optimization Using Evolutionary Algorithms*. Wiley-Interscience. Available at: http://ci.nii.ac.jp/ncid/BB00925127

[26] Goldberg, D. E. (1989). *Genetic Algorithms in Search, Optimization, and Machine Learning*. Addison-Wesley. Available at: https://doi.org/10.5555/58058

[27] Hastie, T., Tibshirani, R., & Friedman, J. (2009). *The Elements of Statistical Learning*. In Springer Series in Statistics. Springer. https://doi.org/10.1007/978-0-387-84858-7

[28] Dwork, C. (2006). Differential privacy. In *International Colloquium on Automata, Languages, and Programming (ICALP 2006)*, Lecture Notes in Computer Science, vol. 4052, pp. 1–12. Springer. https://doi.org/10.1007/11787006_1

[29] Hyndman, R. J., & Fan, Y. (1996). Sample quantiles in statistical packages. *The American Statistician*, 50(4), 361–365. https://doi.org/10.1080/00031305.1996.10473566

[30] Sebastiani, F. (2002). Machine learning in automated text categorization. *ACM Computing Surveys*, 34(1), 1–47.

[31] Levandowsky, M., & Winter, D. (1971). Distance between sets. *Nature*, 234(5323), 34–35. https://doi.org/10.1038/234034a0

[32] Golle, P. (2006). Revisiting the uniqueness of simple demographics in the US population. In *Proceedings of the Workshop on Privacy in the Electronic Society* (pp. 77–80).

[33] Holland, J. H. (1975). *Adaptation in Natural and Artificial Systems*. University of Michigan Press. Available at: https://dl.acm.org/citation.cfm?id=129194

[34] Glover, F. (1986). Future paths for integer programming and links to artificial intelligence. *Computers & Operations Research*, 13(5), 533–549. https://doi.org/10.1016/0305-0548(86)90048-1

[35] Kirkpatrick, S., Gelatt, C. D., & Vecchi, M. P. (1983). Optimization by simulated annealing. *Science*, 220(4598), 671—680. https://doi.org/10.1126/science.220.4598.671