Comparison of machine learning models applied on anonymized data with different techniques

Judith Sáinz-Pardo Díaz
Instituto de Física de Cantabria (IFCA), CSIC-UC
Avda. los Castros s/n. 39005 - Santander (Spain)
Email: sainzpardo@ifca.unican.es

Álvaro López García
Instituto de Física de Cantabria (IFCA), CSIC-UC
Avda. los Castros s/n. 39005 - Santander (Spain)
Email: aloga@ifca.unican.es

2023 IEEE International Conference on Cyber Security and Resilience (CSR) | 979-8-3503-1170-9/23/$31.00 ©2023 IEEE | DOI: 10.1109/CSR57506.2023.10224917

***Abstract*—Anonymization techniques based on obfuscating the quasi-identifiers by means of value generalization hierarchies are widely used to achieve preset levels of privacy. To prevent different types of attacks against database privacy it is necessary to apply several anonymization techniques beyond the classical *k-anonymity* or *ℓ-diversity*. However, the application of these methods is directly connected to a reduction of their utility in prediction and decision making tasks. In this work we study four classical machine learning methods currently used for classification purposes in order to analyze the results as a function of the anonymization techniques applied and the parameters selected for each of them. The performance of these models is studied when varying the value of k for *k-anonymity* and additional tools such as *ℓ-diversity*, *t-closeness* and *δ-disclosure privacy* are also deployed on the well-known *adult dataset*.**

I. Introduction

Digitization makes the amount of data being generated on a daily basis increasingly larger and more valuable for use and processing. The easier sharing of information with anyone else anywhere in the world has meant that the research and development of secure protocols for such releasing is proceeding apace. The continuous generation of individuals' personal data makes it essential to develop privacy preserving techniques and to include them in data science pipelines.

The use of anonymization techniques is key in the publication, processing and analysis of sensitive data. This is a particularly critical issue in a data science process where we are dealing with data containing information about individuals that allow to identify them.

To this aim, different privacy models have been proposed with the overall idea of providing data privacy and anonymity guarantees. When handling a database that is to be anonymized, one must clearly identify certain types of attributes or information in it according to their nature: *identifiers* (variables which uniquely characterize an individual, e.g. ID number or name and surname), *quasi-identifiers* (variables accessible to the attacker that, through their combination, make it possible to identify an individual, e.g. gender, age or address) and *sensitive attributes* (confidential information about an individual that should not be associated with him/her, e.g. salary, clinical history, etc) [1]. Therefore, during an anonymization process, identifiers must be eliminated, and a taxonomy tree or scheme of hierarchies can be established for each quasi-identifier, allowing them to be generalized or even deleted [2]. All database records that are identical with respect to the quasi-identifiers form an *equivalence class*. In this context, some classical anonymization techniques that focus both on acting on the quasi-identifiers (such us *k-anonymity*) but also on the sensitive attributes and their distribution in the different equivalence classes of the database (such us *ℓ-diversity*, *t-closeness* and *δ-disclosure privacy*) can be applied.

The relevance of anonymization is self-evident in many areas where sensitive data are available and when machine learning (ML) models are potential tools to be applied in order to carry out an inference process with them. However, it should be noted that a too strict level of anonymity may compromise their usefulness for processing and inference. This can occur either because a large number of records had to be eliminated to reach the desired level of anonymization, or because the hierarchies applied on the quasi-identifiers dilute their initial information. It is necessary to achieve a balance between the level of data anonymization and the utility of the data for analysis. In this work we will use a classical dataset in the field of anonymization, together with some pre-established hierarchies on its quasi-identifiers, and we will anonymize it using four different methods. On these data, anonymized with different levels and different techniques, an inference process will be carried out using four classical machine learning models, from the classical *k-Nearest Neighbors (kNN)* to three ensemble methods based on the use of decision trees, *Random Forest (RF)*, *Adaptive Boosting (AB)* and *Gradient Tree Boosting (GB)*.

The remainder of this paper is structured as follows: in Section II the related work in the area is presented. In Section III the four anonymization techniques under study are exposed. Section IV presents the data used, the machine learning models employed, and the results obtained in each case according to different metrics. Finally, Section V draws the conclusions obtained from the study performed and Section VI presents the code availability.

II. Related work

A. Privacy preservation

When managing data including sensitive attributes it is important to focus on data privacy in order to avoid possible security breaches, especially when these may involve a threat to an individual's privacy. To this end, privacy preserving

technologies are in the foreground [3]. Specifically, privacy-preserving technologies encompass a wide range of techniques that can be grouped into different subsets (see Figure 1). On the one hand, those that focus on cryptographic methods, such as Homomorphic Encryption (HE) or Secure Multi-Party Computation (SMPC), which make it possible to process and analyze the information without the need to decrypt it [3], thus preserving its integrity and privacy. On the other hand, techniques such as differential privacy are also emerging. The implementation of differential privacy is proposed in [4], and the main idea is to achieve that the absence of a single record does not impact the overall dataset characteristics. Several methods can be used in order to add such an statistical noise, e.g. Laplace and Exponential mechanisms among others [5].

Other key privacy preserving technologies are in the focus of this work, namely anonymity techniques. Specifically, *k-anonymity* was introduced in 1998 [6], and seeks to make an individual in a dataset indistinguishable from at least $k-1$ other individuals [3]. This approach is very simple but it is also widely used even nowadays ([7], [8], [9]). However, it is susceptible to numerous attacks [10], so other techniques must also be taken into account, such as ℓ*-diversity* or *t-closeness* among many others, focusing on the sensitive attributes and their distribution in the database [10], [11]. Details about these techniques and the attacks they aim to prevent will be detailed in Section III.

Finally, another group which could be included is *privacy preserving machine/deep learning*. If we are focusing on data analysis by means of ML or DL models, data decentralization techniques, such as federated learning, split learning or gossip learning allow collaboration between different clients to apply this kind of models to their data without the need for them to share it with each other or with a central server [12].

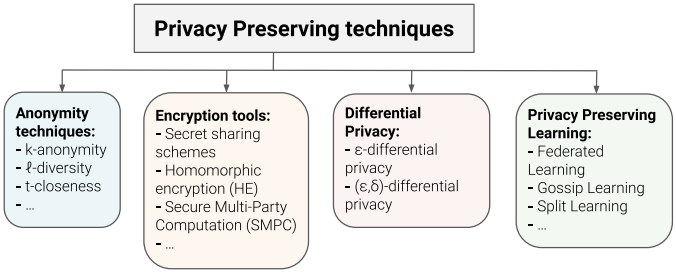

Fig. 1. Privacy Preserving techniques. Inspired in Figure 2.1 of [3].

B. Data utility

Here, the largest challenge that we find when applying anonymization techniques to preserve the privacy of a database lies in achieving a balance between the privacy of the data, and its usefulness for further analysis. There is no point in obtaining data with a level of anonymity that makes them insusceptible to any attack (which is an optimal scenario from a privacy point of view), if they are of no practical use at all. It is therefore essential to analyze the usefulness of the data after anonymization in order to select the level that best achieves this trade-off. To this end, in [13] different *k-anonymization* algorithms (Mondrian, OLA, TGD, CB) are studied together with the performance of four machine learning models by varying the value of k, concluding that with an increasingly strong *k-anonymity* constraint, the classification performance generally degrades. Also, in this work the authors note that "for very large k of up to 100 the performance losses remain within acceptable limits". On the other hand, and following this same line, in [14] different artificial intelligence models (i.e. neural networks, logistic regression, decision trees, Bayesian classifier) are considered together with three datasets, in order to analyze the performance of the different models before and after applying *k-anonymity* to the data with a value of $k=2$. The authors highlight that, according to their results, certain machine learning algorithms are more suited to be used with privacy-preserving data mining techniques than others. Finally, in [15] several examples applying both *k-anonymity* and *differential privacy* are presented in order to mitigate inference attacks.

Among our contributions we highlight the study of additional anonymization techniques to *k-anonymity*, which focus on sensitive attributes. On the one hand, we first study the influence of different values of k on the four analyzed machine learning models using two metrics: accuracy and AUC. On the other hand, we consider three other anonymity techniques and analyze their effect on the perfomance of the different models under study. For this purpose, we use a classical dataset and a set of predefined hierarchies to carry out the anonymization using the *ARX Software* [16].

III. Anonymization techniques deployed

In this study four classical anonymity techniques that focus on obfuscating quasi-identifiers by applying value generalization hierarchies (VGH) to them will be applied. Specifically, generalization consists of replacing the original values by other consistent but less specific ones, in this case by means of taxonomy tree or hierarchies. Maximal generalization or value suppression consists of replacing it with an special character such as '*'.

Thus, we provide the dataset with a hierarchy scheme through which we can generalize the various quasi-identifiers to a more complete domain, as will be explained in the next section. Different examples of hierarchy trees can be found in Figure 1 of [13] and in Figure 2 of [17] among others.

Specifically, we will start with the application of one of the most classic and widely used techniques, due to its easy interpretability and implementation, that is *k-anonymtity*. A database is said to be *k-anonymous* if and only if there are at least k records in each equivalence class of the database. If this condition is verified the probability of that record corresponding to an individual is $1/k$. In the same way, another of the techniques applied, ℓ*-diversity*, is also widely used for preserving privacy of sensitive attributes of a database (especially to prevent homogeneity attacks), due again to its easy interpretability. We say that ℓ*-diversity* is verified if there are at least ℓ different values for the sensitive attribute in

each equivalence class of the database [10]. Finally, two other techniques to be employed that focus on the distribution of sensitive attributes in the different equivalence classes of the database are *t-closeness* and *δ-disclosure privacy*. A brief description of this last two methods is given below (see [1]).

- *t-closeness:* is verified if the distribution of the values of the sensitive attribute in each equivalence class are no more than a distance t apart from the distribution of the sensitive attribute in the whole database. In order to measure the distance between the distributions, for numerical attributes the Earth Mover's distance (EMD) with the ordered distance is applied, while for categorical attributes the equal distance is used [11].
- *δ-disclosure privacy:* is satisfied if $\left|\log\left(\frac{p(EC,s)}{p(DB,s)}\right)\right| < \delta$ is fulfilled for every equivalence class (EC) and every value s of the sensitive attribute, being $p(EC,s)$ its distribution in an equivalence class and $p(DB,s)$ in the whole database [18].

It is essential to study several anonymity techniques together, since each one of them can prevent a different type of attack on the database. Specifically, the above four techniques can be used respectively to prevent different kind of attacks, as summarized in Table I for the methods presented previously.

TABLE I
MAIN ATTACKS THAT PREVENT EACH ANONYMITY TECHNIQUE UNDER STUDY. EXTRACTED FROM TABLE 2 OF [1].

Technique	***Main attacks prevented***
k-anonymity	Linkage and re-identification attacks
ℓ-diversity	Homogeneity and background attacks
t-closeness	Similarity and skewness attacks
δ-disclosure privacy	Skewness and inference attacks

IV. EXPERIMENTAL RESULTS AND ANALYSIS

The data used in this analysis are presented below along with the hierarchies established for each quasi-identifier. In addition, a brief description of the machine learning models analyzed through this study is presented. The performance of these models will be analyzed in two cases: varying the value of k fixed for *k-anonymity*, and once anonymized with $k = 5$, applying the other three techniques focused on the distribution of quasi-identifiers: *ℓ-diversity*, *t-closeness* and *δ-disclosure privacy*.

A. Dataset and hierarchies

In order to carry out this comparative study, a classic dataset regarding privacy and anonymity techniques will be used: the *adult dataset* (available in [19]), which is an extraction of information from the 1994 U.S. Census database. A sub-sample containing 32561 records will be used, and the objective is to predict whether or not the income of an individual exceeds 50000$ by year, based on census data.

Specifically in this study six quasi-identifiers (which would act as the features when training a machine learning model) will be considered, namely: *age*, *sex*, *education*, *relationship*, *occupation* and *native-country*, together with *salary class* (which can take as values $> 50K$ or $\leq 50K$) acting as the sensitive attribute and the label of the data. The following hierarchies have been applied to each quasi-identifier (extracted from those presented in the *ARX Software* GitHub repository https://github.com/arx-deidentifier/arx):

- *age:* five levels of hierarchies are applied, starting with 5-year intervals: $[15, 20)$, $[20, 25)$,..., $[75, 80)$, and finally the case where the age is greater than 80. Further levels are defined with 10, 20, 40 and 80 years intervals.
- *sex:* no hierarchies are applied.
- *education:* two levels of hierarchies are applied for each of the 16 possible values. In the first level there are five values: *Primary School*, *High School*, *Undergraduate*, *Professional Education* and *Graduate*. The second level only takes three possible alternatives: *Primary education*, *Secondary education* and *Higher education*.
- *relationship:* no hierarchy is applied.
- *occupation:* a hierarchy level is included that encompasses in three categories the different options, namely: *technical*, *non-technical* and *other*.
- *native country:* a hierarchy level is introduced that generalizes the country by continent. Thus, the possible options according to the countries present in the original database are: *Africa*, *Asia*, *Europe*, *North America*, *South America* and *unknown* (if the value of the native country field was ?).

As already mentioned, the anonymization process will be carried out using the *ARX Software* with the hierarchies described above for the quasi-identifiers and with *salary class* as the sensitive attribute.

B. Machine learning models and evaluation metrics

Through this study we present a classical machine learning classification problem. That is, given a set of inputs, $\mathbf{X} \in \mathrm{D}^{nxm}$ with n the number of records and m the number of features, and the corresponding labels (outputs) $\mathbf{y}$, with $y_i \in \{1, \ldots, n_C\}\ \forall i \in \{1, \ldots, n\}$, being n_C the number of different classes, our objective is to estimate a function $\hat{f}$ which approximates $\mathbf{y} = \hat{f}(\mathbf{x})$ [20]. Specifically, we are going to apply the following four supervised machine learning models, with the quasi-identifiers as features and the sensitive attribute as label:

- ***k-Nearest Neighbors (kNN):*** This non-parametric method find the k nearest point in the training split to the test input. Thus, it computes the posterior probability that an element belongs to certain class. Its main drawback is that its performance is poor in cases of high dimensional inputs [20]. While large values of k reduce the effect of noise in classification, a value $k = 1$ actually induces a Voronoi tessellation of the point $\mathbf{X}_i, i \in \{1, \ldots, n\}$.
- ***Random Forest (RF):*** The main motivation for introducing this method comes from trying to reduce the variance of an estimator by aggregating several of them. In particular, N decision trees can be trained and their

ensemble be computed. In order to avoid obtaining highly correlated predictors, the Random Forest method learns trees based on randomly chosen subsets of the input data and of the features [20]. Specifically, a new training subset is constructed using bootstrapping, then decision trees are trained on it using a subset of the predictor variables. This process is repeated to finally obtain a unified prediction.

- ***Adaptive Boosting (AB):*** Boosting is a ML approach based on the idea of creating a highly accurate prediction rule by combining many relatively weak rules [21]. Thus, with adaptive boosting we first train a classifier (by giving equal weight to all data), calculate the error associated with it, and compute a new distribution to weight the data based on whether or not it was correctly classified. This process is repeated M times and finally a weighted average of the classifiers is performed. The details of this method can be found in pseudocode form in Algorithm 16.2 of [20].
- ***Gradient Tree Boosting (GB):*** This ensemble classifier trains several individual predictors sequentially. As in the case of Adaptive Boosting, again the fundamental idea lies in combining weak predictors (decision trees) to create a robust one. Typically, with this method the first predictor learns to predict the data mean, then the second one explains the errors of the first one, the third one explains the errors of the second one, and so on. A detailed formulation can be found in [22].

The Python library *scikit-learn* (version 1.2.0) has been used to train and test the different models. In addition, the *GridSeachCV* function is used to select the optimal parameters for each of them in order to perform a 5-fold cross-validated grid-search. Then, the model is retrained with the optimum parameters calculated with the cross-validation. Specifically, in each model the hyper-parameter grid selected is presented bellow together with some other fixed parameters:

- ***kNN.*** Number of neighbors: [3, 4, ... , 50]. Metric to calculate the distance: *minkowski*.
- ***RF.*** Maximum depth of the tree: [2, 3, ... , 9]. Number of trees: 100. Criterion: gini impurity.
- ***AB.*** Number of estimators: [50, 100, 150]. Learning rate: [0.01, 0.1, 0.5, 1].
- ***GB.*** Number of estimators: [50, 100, 150]. Learning rate: [0.01, 0.1, 0.5, 1]. Maximum depth of the individual estimators: [2, 4, 6, 8, 10].

Regarding the error metrics used to evaluate the performance of each of the four machine learning methods presented above, since we are dealing with a classification problem, the accuracy and the area under the ROC curve will be analyzed (AUC). Note that although the anonymization process is carried out on the complete database (in order to simulate the real case study), a stratified random train-test split (75%-25%) will be performed when training the models.

C. Results and analysis

First, we will analyze the scenario in which the only anonymization technique applied is *k-anonymity* for different values of k, and with a record suppression limit of 100% (i.e., there is no limit on the number of records that can be deleted to reach the anonymity condition). Specifically, the following values of k will be taken: $k \in \{2, 5, 10, 25, 50, 75, 100\}$, although the analysis has been performed on more values of k, just the most significant ones are shown below.

For this purpose, once the anonymization process with the hierarchies exposed in Subsection IV-A has been performed using the *ARX Software*, we test the different machine learning models. Table II shows the results obtained both for the accuracy and for the AUC compared also to those obtained by applying the models on the raw data.

TABLE II
ACCURACY AND AUC OBTAINED FOR EACH MACHINE LEARNING MODEL WHEN VARYING THE VALUE OF k FOR k-ANONYMITY.

	kNN		***RF***		***AB***		***GB***	
k	*Acc.*	*AUC*	*Acc.*	*AUC*	*Acc.*	*AUC*	*Acc.*	*AUC*
Raw	0.8056	0.6779	**0.8334**	**0.7351**	**0.8350**	0.7377	**0.8352**	**0.7437**
2	0.8176	0.7143	0.8210	0.7215	0.8266	0.7372	0.8249	0.7346
5	**0.8199**	0.7234	0.8208	0.7209	0.8197	0.7342	0.8221	0.7258
10	0.8099	0.6932	0.8139	0.6913	0.8146	0.7171	0.8168	0.7157
25	0.8170	0.7194	0.8163	0.6987	0.8164	0.7167	0.8177	0.7139
50	0.8124	**0.7268**	0.8114	0.7285	0.8051	**0.7399**	0.8114	0.7285
75	0.8129	0.7064	0.8129	0.7064	0.8091	0.7195	0.8129	0.7030
100	0.8107	0.7254	0.8083	0.6981	0.8047	0.7117	0.8103	0.7274

Overall, starting first to analyze the results in terms of accuracy, we can observe that, as expected, a higher value of k goes together with a reduction in accuracy in most of the cases. This is clearly seen in the three ensemble methods in which for both accuracy and AUC the best value is achieved with the raw data, and the worst with $k = 100$ except for AUC with GB, which is obtained when $k = 75$. However, with *kNN* the minimum for the accuracy and the AUC is reached when using the raw data. This peculiarity can be attributed to the characteristics of the data, the dimensionality of the problem and the hierarchies applied. With respect to the AUC again, it can be observed that there is a lot of variability in the results, which start clearly decreasing for the first values of k and the ensemble methods. It is especially striking that the optimal value for the AUC with *AB* method is achieved when $k = 50$ (although it is close to that obtained by using the raw data).

Besides, we are also interested in analyzing for each of the previously exposed values of k, how close the resulting database (Ω_k) is to being optimal for the stated level of anonymization. For this purpose the *average equivalence class size metric*, which measures how well the equivalence classes are created to fit the best case [17], is studied. Specifically, the optimal value for this metric would be one, indicating that all equivalence classes are of length k. Since in our example record suppression has been allowed with a limit of 100%, instead of analyzing the number of equivalence classes as a function of the number of initial records (as exposed in the classic definition of this technique [17]), the number of records

resulting from anonymization in each case will be considered. This is shown in Equation 1, in which *ECs* is the set of equivalence classes.

$$C^*_{AVG_k}(\Omega_k) = \frac{|\Omega_k|}{k|ECs|} \tag{1}$$

As intuitively expected, the lowest value for this metric is reached for the highest value of k analyzed ($k = 100$), obtaining in particular the following values: $C^*_{AVG_k} = 7.88, 10.94, 10.75, 7.68, 6.90, 5.13, 4.17 \ \forall k \in \{2, 5, 10, 25, 50, 75, 100\}$ respectively.

Next, once a value of $k = 5$ for *k-anonymity* has been fixed, the other three techniques described in Section III will be applied. In particular, a value $\ell = 2$, $t = 0.7$, and $\delta = 1.5$ have been fixed in each case. Note that $\ell = 2$ is the only possible value other than one, and for the other two cases, the values have been chosen after testing different values in order to achieve a balance with the number of deleted records and the utility. In addition, the values of δ has been chosen in order to be the most restrictive scenario of the five under study (this will be discussed further below in terms of the other parameters fulfilled). The results as a function of the accuracy are shown in Table III, and the ROC curves and the AUC obtained with each ML model analyzed and with each level of anonymity are displayed in Figure 2.

TABLE III
ACCURACY OBTAINED WITH EACH MACHINE LEARNING MODEL ACCORDING TO THE ANONYMIZATION TECHNIQUE APPLIED.

ML model	*Raw*	$k = 5$	$k = 5$, $\ell = 2$	$k = 5$, $t = 0.7$	$k = 5$, $\delta = 1.5$
kNN	0.8056	0.8199	0.7894	0.8164	0.8061
RF	0.8334	0.8208	**0.8025**	**0.8218**	0.8070
AB	0.8350	0.8197	0.8017	**0.8218**	0.8085
GB	**0.8352**	**0.8221**	**0.8025**	0.8210	**0.8098**

Here again the case of *kNN* stands out for its counter-intuitiveness: both accuracy and AUC obtained by using the raw data are the second worst values of the five scenarios considered. However, with the ensemble methods the optimum is reached when using the models with the raw data (both for the accuracy and the AUC), i.e. without applying any anonymization technique. In the same line, both with *kNN* and the ensemble models the worst results in terms of accuracy are obtained when the classification is performed with the data resulting after applying *ℓ-diversity* with $\ell = 2$. This is really intuitive from the meaning of this property: for the same set of quasi-identifiers, the label will take in all cases two different values.

Similarly, the worst performance in terms of AUC clearly occurs with all four models in the case where *δ-disclosure* is applied with $\delta = 1.5$. Again this is really intuitive if we analyze the parameters that are satisfied for each technique in that scenario, because although *ℓ-diversity* is only verified for $\ell = 1$, the value of t for *t-closeness* is the most stringent of all the cases analyzed, being $t = 0.47$. This has been verified

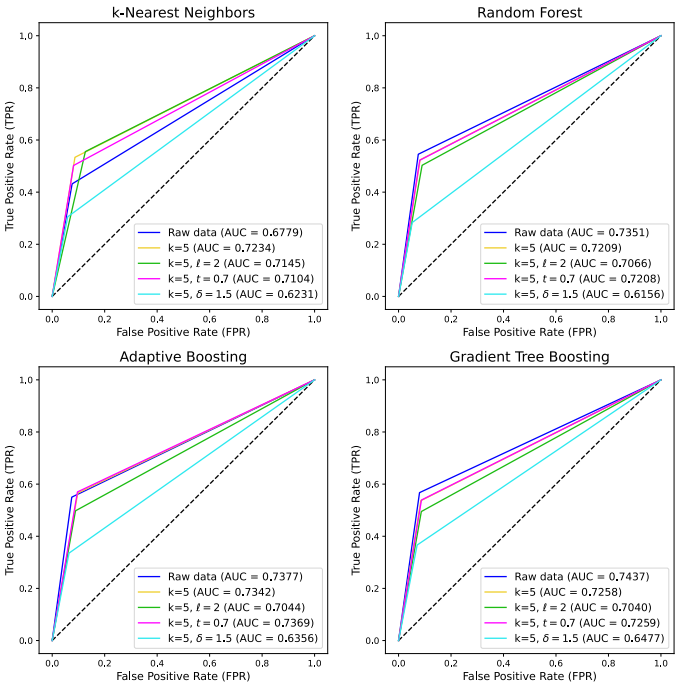

Fig. 2. ROC curves and AUC obtained with each machine learning model and anonymity level for the *adult dataset*.

by means of the Python library *pyCANON* [1] (version 1.0.0.), and we can also observe that although in the case in which the value of δ has been set to 1.5, this property is verified for $\delta = 1.16$ (more restrictive). In the case where we set $\ell = 2$, we obtain values $t = 0.64$ and $\delta = 4.88$, while in the case where we set $t = 0.7$, $\ell = 1$ and $\delta = 4.10$

As expected, the three ensemble methods perform quite similarly in all scenarios, which is plainly illustrated in Figure 2. In particular, the best value for the AUC is obtained with *GB* and the raw data, and in the worst scenario ($\delta = 1.5$) the best value is also obtained with this method for both accuracy (0.8098) and AUC (0.6477). Note that when $\delta = 1.5$, although the prediction does not seem to be bad in terms of accuracy it is actually poor in view of the AUC with the four models. This reflects the need to test different error metrics in order to select an optimal anonymization technique and ML model.

In order to carry out a classification task, it should be noted that the optimum scenario would be for all the records that constitute an equivalence class to have the same label (i.e. the same sensitive attribute). However, this is in contrast to the three anonymization techniques analyzed that focus on sensitive attributes, since it would make homogeneity attacks feasible, among others. Let us therefore look at the *classification metric (CM)* obtained in each of the four cases analyzed, defined as shown in Equation 2 (see [23]):

$$CM = \frac{1}{N}\sum_{i=1}^{N} penalty(r_i), \tag{2}$$

where $penalty(r_i) = 1$ if the row r_i has been deleted or if its associated label (i.e. SA) takes a value other than the

majority value in the equivalence class to which it belongs, and 0 otherwise, $\forall i \in \{1, \ldots, N\}$ with N the number of records in the original database.

The results obtained for each of the four cases and the CM metric are as follows: 0.2569 ($k = 5$), 0.3089 ($k = 5$, $\ell = 2$), 0.2575 ($k = 5$, $t = 0.7$), 0.4589 ($k = 5$, $\delta = 1.5$). These values contrast with those obtained in Table III for the accuracy, where the results for $\delta = 1.5$ are better than those for $t = 0.7$ in 3 of the 4 cases, while they agree with those obtained for the AUC (see Figure 2). Note that for the AUC the worst results are obtained when $\delta = 1.5$, as would be expected based on the values calculated for *CM*.

V. Conclusions and future directions

In this paper the performance of four classical machine learning models in a classification task has been analyzed after subjecting the *adult dataset* to different levels of anonymity. In particular, the scaling of the accuracy and the AUC has been analyzed when increasing the value of k for *k-anonymity*, as well as the optimally of the anonymization by using the $C^*_{AVG_k}$ metric exposed in Equation 1. As a result, it is noteworthy that in the case of *kNN* model the best results in terms of both metrics are not obtained either with the raw data or with the lowest value of k. As for the ensemble methods, in all cases the best performance concerning accuracy are obtained when training with the raw data, as expected intuitively.

Furthermore, these same models have been analyzed when applying *ℓ-diversity*, *t-closeness* and *δ-disclosure privacy* in addition to *5-anonymity*, with $\ell = 2$, $t = 0.7$ and $\delta = 1.5$ respectively. The AUC and the accuracy obtained in each case are studied, as well as the *classification metric (CM)* defined in Equation 2. In general terms in this case the best results are obtained when using the ensemble models, and it is observed that in agreement with *CM* the worst results in terms of AUC are reached with $\delta = 1.5$ in addition to $k = 5$.

As future lines to further extend this work, we are interested in the extrapolation to other datasets, anonymization techniques and their associated parameters. In addition, the inclusion of *differential privacy (DP)* during the inference process using machine learning models is also an attractive field of study.

VI. Code availability

For completeness and reproducibility of the work, the anonymized data together with the code carried out is available in the following GitHub repository: https://github.com/IFCA/anonymity-ml.

Acknowledgment

The authors would like to thank the funding through the European Union - NextGenerationEU (Regulation EU 2020/2094), through CSIC's Global Health Platform (PTI+ Salud Global) and the support from the project AI4EOSC "Artificial Intelligence for the European Open Science Cloud" that has received funding from the European Union's Horizon Europe research and innovation programme under grant agreement number 101058593.

References

[1] J. Sáinz-Pardo Díaz and Á. López García, "A python library to check the level of anonymity of a dataset," *Scientific Data*, vol. 9, no. 1, pp. 1–12, 2022.

[2] J. Domingo-Ferrer, D. Sánchez, and J. Soria-Comas, "Database anonymization: privacy models, data utility, and microaggregation-based inter-model connections," *Synthesis Lectures on Information Security, Privacy, & Trust*, vol. 8, no. 1, pp. 1–136, 2016.

[3] K. Kim and H. C. Tanuwidjaja, *Privacy-preserving Deep Learning: A Comprehensive Survey*. Springer, 2021.

[4] C. Dwork and A. Roth, "The algorithmic foundations of differential privacy," *Foundations and Trends® in Theoretical Computer Science*, vol. 9, no. 3–4, pp. 211–407, 2014.

[5] T. Zhu, G. Li, W. Zhou, and S. Y. Philip, *Differential privacy and applications*. Springer, 2017.

[6] P. Samarati and L. Sweeney, "Protecting privacy when disclosing information: k-anonymity and its enforcement through generalization and suppression," 1998.

[7] K. Arava and S. Lingamgunta, "Adaptive k-anonymity approach for privacy preserving in cloud," *Arabian Journal for Science and Engineering*, vol. 45, no. 4, pp. 2425–2432, 2020.

[8] N. Li and S. K. Das, "Applications of k-anonymity and l-diversity in publishing online social networks," in *Security and Privacy in Social Networks*. Springer, 2013, pp. 153–179.

[9] H. Takabi, J. B. D. Joshi, and H. A. Karimi, "A collaborative k-anonymity approach for location privacy in location-based services," in *2009 5th International Conference on Collaborative Computing: Networking, Applications and Worksharing*, 2009, pp. 1–9.

[10] A. Machanavajjhala, D. Kifer, J. Gehrke, and M. Venkitasubramaniam, "l-diversity: Privacy beyond k-anonymity," *ACM Transactions on Knowledge Discovery from Data (TKDD)*, vol. 1, no. 1, pp. 3–es, 2007.

[11] N. Li, T. Li, and S. Venkatasubramanian, "t-closeness: Privacy beyond k-anonymity and l-diversity," in *2007 IEEE 23rd international conference on data engineering*. IEEE, 2006, pp. 106–115.

[12] J. Sáinz-Pardo Díaz and Á. López García, "Study of the performance and scalability of federated learning for medical imaging with intermittent clients," *Neurocomputing*, vol. 518, pp. 142–154, 2023.

[13] D. Slijepčević, M. Henzl, L. D. Klausner, T. Dam, P. Kieseberg, and M. Zeppelzauer, "k-anonymity in practice: How generalisation and suppression affect machine learning classifiers," *Computers & Security*, vol. 111, p. 102488, 2021.

[14] H. Wimmer and L. Powell, "A comparison of the effects of k-anonymity on machine learning algorithms," in *Proceedings of the Conference for Information Systems Applied Research ISSN*, vol. 2167, 2014, p. 1508.

[15] A. Goldsteen, G. Ezov, R. Shmelkin, M. Moffie, and A. Farkash, "Anonymizing machine learning models," in *Data Privacy Management, Cryptocurrencies and Blockchain Technology: ESORICS 2021 International Workshops, DPM 2021 and CBT 2021, Darmstadt, Germany, October 8, 2021, Revised Selected Papers*. Springer, 2022, pp. 121–136.

[16] F. Prasser and F. Kohlmayer, "Putting statistical disclosure control into practice: The arx data anonymization tool," in *Medical data privacy handbook*. Springer, 2015, pp. 111–148.

[17] V. Ayala-Rivera, P. McDonagh, T. Cerqueus, L. Murphy *et al.*, "A systematic comparison and evaluation of k-anonymization algorithms for practitioners," *Transactions on data privacy*, vol. 7, no. 3, pp. 337–370, 2014.

[18] J. Brickell and V. Shmatikov, "The cost of privacy: Destruction of data-mining utility in anonymized data publishing," in *Proceedings of the 14th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, ser. KDD '08. New York, NY, USA: Association for Computing Machinery, 2008, p. 70–78.

[19] D. Dua and C. Graff, "UCI machine learning repository," http://archive.ics.uci.edu/ml, 2017.

[20] K. P. Murphy, *Machine learning: a probabilistic perspective*. MIT press, 2012.

[21] R. E. Schapire, "Explaining AdaBoost," in *Empirical Inference*. Springer Berlin Heidelberg, Jan. 2013, pp. 37–52.

[22] J. H. Friedman, "Stochastic gradient boosting," *Computational statistics & data analysis*, vol. 38, no. 4, pp. 367–378, 2002.

[23] V. S. Iyengar, "Transforming data to satisfy privacy constraints," in *Proceedings of the Eighth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, ser. KDD '02. New York, NY, USA: Association for Computing Machinery, 2002, p. 279–288.