Progress in Artificial Intelligence (2024) 13:31–50
https://doi.org/10.1007/s13748-024-00315-2

REGULAR PAPER

Explainable machine learning models with privacy

Aso Bozorgpanah[1] · Vicenç Torra[1]

Received: 7 June 2022 / Accepted: 12 February 2024 / Published online: 28 March 2024

Abstract

The importance of explainable machine learning models is increasing because users want to understand the reasons behind decisions in data-driven models. Interpretability and explainability emerge from this need to design comprehensible systems. This paper focuses on privacy-preserving explainable machine learning. We study two data masking techniques: maximum distance to average vector (MDAV) and additive noise. The former is for achieving k-anonymity, and the second uses Laplacian noise to avoid record leakage and provide a level of differential privacy. We are interested in the process of developing data-driven models that, at the same time, make explainable decisions and are privacy-preserving. That is, we want to avoid the decision-making process leading to disclosure. To that end, we propose building models from anonymized data. More particularly, data that are k-anonymous or that have been anonymized add an appropriate level of noise to satisfy some differential privacy requirements. In this paper, we study how explainability has been affected by these data protection procedures. We use TreeSHAP as our technique for explainability. The experiments show that we can keep up to a certain degree both accuracy and explainability. So, our results show that some trade-off between privacy and explainability is possible for data protection using k-anonymity and noise addition.

Keywords Machine learning · Data privacy · Microaggregation · k-anonymity · Noise addition · Local differential privacy · Irregularity · Explainability · eXplainable artificial intelligence

1 Introduction

Machine learning (ML) models have great potential for enhancing products, processes, and research. The role of predictions made by machine learning algorithms in our lives is increasing. Data-driven models built using machine learning prove their usefulness. Nevertheless, ML algorithms usually do not explain their predictions, which is a barrier to the adoption of machine learning. Solutions for this challenge fall within the so-called eXplainable Artificial Intelligence (XAI). Explainability is widely acknowledged as a crucial feature for the practical deployment of AI models. XAI provides ways to understand why a ML model yields a predicted output for a certain input.

Unfortunately, some data-driven models are difficult to understand. A black-box predictor is a machine learning obscure model. That is, a model with internal structures that are difficult to interpret by humans. Following Marcinkevičs and Vogt [1], we distinguish between interpretability and explainability. Interpretability focuses on designing models that are themselves understandable to a human [2], whereas explainability involves providing rationales for decisions. In other words, explainability makes clear and understandable explanations for the decisions. Post hoc explainability [3, 4] deals with methods to explain decisions from black-box models after predictions are made.

One of the post hoc explainability methods is SHAP (SHapley Additive exPlanations) [5], which is among the most popular techniques in the category of local model-agnostic methods. These methods can be applied to various types of models, including black-box models.

The main aim of SHAP is to explain individual predictions, and it is based on game theoretically optimal Shapley Values. Nevertheless, like other model-agnostic feature attribution methods [6, 7], they are computationally expensive.

This work was partially supported by the Wallenberg AI, Autonomous Systems and Software Program https://wasp-sweden.org/ (WASP) funded by the Knut and Alice Wallenberg Foundation.

✉ Aso Bozorgpanah
asob@cs.umu.se

Vicenç Torra
vtorra@cs.umu.se

[1] Department of Computing Science, Umeå University, MIT Building, 90187 Umeå, Sweden

1.1 Motivation

Although understanding why a model makes a prediction is important, XAI principles are not enough [8]. Other important principles need also to be carefully addressed for artificial intelligence and machine learning deployment in the real world, privacy is one of them. Data and machine learning models should not disclose sensitive information. Moreover, data-driven models should consider the privacy requirements in the whole life cycle. That is, they need to be built using privacy by design. Otherwise, traces of the data used in the training can be found in the model. This is one of the principles considered in the paper.

Masking methods are the tools that allow us to protect data when this needs to be shared with third parties. This is also useful for building privacy-preserving models, to prevent these traces from appearing in the model. Masking methods allow us to implement k-anonymity and also to provide privacy for reidentification.

The effects of masking methods on XAI and, more concretely, on explainability are unknown. In [9], it is claimed that there is a conflict between privacy and explainability and that both are incompatible. The goal of this paper is to assess to what extent we can keep information useful for explainability using tools for building privacy-preserving models.

1.2 Related works

The combination of explainability and privacy technologies has recently attracted the interest of the machine learning community. A particular area of research focuses on creating privacy-preserving strategies that let AI models give explanations without endangering individual user data. Differential privacy methods, for example, have been used with XAI systems to make sure that explanations are produced without accidentally revealing private data [10–12]. Furthermore, improvements in federated learning make it possible to train collaborative models across decentralized data sources while improving explainability [13–15].

More concretely, Patel et al. [10] study the construction of differentially private local approximation mechanisms. This is to provide differentially private explanations. Their conclusion is that in sparse regions the performance of their approach is poorer. A different approach is the one introduced by Nori et al. [12]. It is about building a differentially private model what is interpretable. The authors propose an approach to build Explainable boosting machines (EBMs) using Gaussian differential privacy (GDP). The model is based on very shallow trees, and to maintain explainability, each tree in the ensemble is constrained to utilize only one feature at a time. This work is similar to the one by Kwatra et al. [16] in which decision trees are built in a federated learning framework. The work uses k-anonymity instead of differential privacy as the protection mechanism.

Our approach is different. We consider the effects of masking in SHAP. That is, our scenario is about sharing protected data, and building both machine learning models and explanations from the protected dataset. Then, the goal is to understand whether this protection alters the explanations built from both the data and the model. This paper complements our results [11] that also considers other masking procedures based on dimensionality reduction (e.g., the ones based on principal components and non-negative factor matrix factorization).

There are other related works about privacy and explainability with a loose connection with our scenario, for example, the work by Bogdanova et al. [17]. Their goal is to build explanations (SHAP) in a collaborative way in a distributed environment. That is, in a federated learning type of scenario. The limitation of their approach is the same as most approaches based on Shapley values. That is, a large number of features make the approach unfeasible. A similar work is described in [18] discussing alternative interpretable models (decision trees and linguistic fuzzy models). Nevertheless, the paper does not provide experimental results. A related result is provided by Wang [19], which concerns the computation of Shapley values in vertically partitioned federated learning. The connection of these works with ours is small, because they are focused on federated learning, while we consider a centralized approach.

Grant and Wischik [9] claim from a law perspective that both explainability and privacy cannot come together. We show that this is indeed possible to a certain extent, and for this, we evaluate how masking methods can affect the results of SHAP. Our analysis is rooted in previous research that shows to what extent masking affects machine learning models.

From the perspective of explainability, model-agnostic methods explain individual predictions. The two most common methods are LIME and SHAP. LIME is based on a linear approximation around an instance. SHAP is based on the Shapley value of a game built from the model. In this paper, we use SHAP because it has better properties (e.g., local accuracy). More particularly, we use TreeSHAP, as proposed by Lundberg et al. [20], which is a tree-based model approach that proved to be very fast in comparison with KernelSHAP. The approach was latter improved in [21] and [22]. They consider features' correlation in SHAP. There is nonzero estimation for a feature when that feature is correlated with another feature that has an influence on the prediction. Likewise, hiding the underlying biases [4] is an effective improvement in the SHAP approach.

With respect to data privacy, there are quite a few masking methods for data protection [23]. Microaggregation is one of the methods that provide k-anonymity. Among the existing methods for microaggregation, MDAV is one of the effective ones, also for big data sets [24]. There are implementations for both numerical and categorical features [25]. For numerical data, there is a method [26] based on principal component analysis to obtain significant savings in running time and memory, without significant degradation of information utility. Rebollo-Monedero et al. [27] proposed a variation of MDAV that improves data utility. More recently, Parra-Arnau and colleagues [28] suggest another version of microaggregation on numerical microdata. They show that disclosure risk is guaranteed via differential privacy through record-level perturbation. That is, their approach for microaggregation provides not only higher utility but also higher privacy than classical microaggregation.

Randomized response [29] and the Laplace distribution mechanism [30] are alternative data protection methods available in the literature. While microaggregation provides k-anonymity, these methods can be used to provide local differential privacy. Randomized response is used for categorical data and the Laplace distribution mechanism for numerical data. See, e.g., Palia et al. [31] about the use of differential privacy for machine learning.

The two main privacy models in the data privacy literature are k-anonymity and differential privacy. They are two alternative definitions of privacy. The first is to avoid privacy from reidentification, and the second is to ensure that any information computed on the records in a database does not let us infer the presence (or absence) of a particular record in the database. k-Anonymity is usually implemented using microaggregation or generalization. We use MDAV as one of the most significant methods for microaggregation. Differential privacy for numerical data is typically implemented using Laplacian noise addition. We also evaluate this approach in this paper.

1.3 Contributions

The main contributions of our work are to study the effect of data privacy techniques for databases (masking methods) [32] to metrics devised for explainability and, more particularly, to SHAP [33].

To do so, we develop a series of experiments comparing the effects of data masking procedures on the explainability of models according to SHAP on three different data sets. In order to do this analysis, we apply two masking methods (microaggregation and Laplacian noise addition). Applying these methods to data sets, we obtain protected versions of the data sets. In this way, we are able to build and compare machine learning models as well as SHAP results on both the original data and the protected one. This is to analyze to what extent masking has an effect on explainability. The results show that some properties are preserved in the protected data. That is, that data protection permits to keep, up to some extent, the qualities of the data and its usefulness for explainability.

Moreover, we observe that the results of MDAV are qualitatively similar to the ones obtained with the original data. A good selection of the distortion level in microaggregation using MDAV permits keeping a good balance between disclosure risk, model accuracy, and explainability. In contrast, noise addition does not seem so effective in preserving explainability.

1.4 Structure of the paper

The structure of the paper is as follows: In Sect. 2, we discuss SHAP, the method to explain individual predictions. Furthermore, we introduce the data protection methods we use. They are MDAV (maximum distance to average vector) and additive noise. Then, our methodology to analyze the effect of masking on explainability is discussed in Sect. 3. Our experiments are described in Sect. 4. Then, evaluation is given in Sect. 5. The paper finishes with some conclusions and directions for future work.

2 Preliminaries

This section gives an overview of some concepts that we need in the rest of the paper. We begin describing the interpretability and explainability as understood in machine learning, as well as pointing out their differences. Then, we describe SHAP, a method for explainability. The section concludes with a review of two data protection mechanisms.

2.1 Interpretability and explainability

EXplainable Artificial Intelligence (XAI) advocates for tools that help understand automated decisions. Some principles are usually required for XAI: fairness, accountability, transparency, robustness and safety, ethics, and privacy [34]. They are depicted in Fig. 1. FATE (fairness, accountability, transparency, and ethics) [7] provides a basic framework for understanding these values. In this work, we focus on privacy and transparency.

As deep learning and other highly accurate black-box models develop, the demand for transparency, interpretability, and explainability grows. Within XAI, interpretability and explainability are understood differently [8].

Interpretability is defined as the ability for a machine learning model to be understood by its users [8, 34]. So, it comes from the design of the model itself. By contrast, explainability is an interface between humans and an automated decision-maker (e.g., a ML model). In this case, an accurate proxy of the decision-maker that is comprehensible to humans [8, 34] can provide explainability. SHAP is an example of such an approach [4].

There are both global and local model-agnostic explanation methods. The latter is about explaining particular decisions (an output for a particular input), and the former is about the behavior of the system in general. In this paper, we use SHAP, one of the most used local model-agnostic explanation methods, that has been used extensively in the literature. It is model-agnostic because it does not try to inspect the system, but only relies on the output of the system.

2.1.1 SHapley Additive exPlanations (SHAP)

SHAP [33] is a method to explain individual predictions. It is based on the Shapley value of game theory. KernelSHAP and TreeSHAP are two different implementations of SHAP.

SHAP values explain the output of a function f as a sum of the effects ϕ_i of each feature i, computed from a conditional expectation. To compute SHAP values, Lundberg et al. [35] define $f_x(S) = E[f(x) \mid x_S]$ where S is a set of input features, and $E[f(x) \mid x_S]$ is the expected value of the function conditioned on a subset S of the input features. SHAP value takes these conditional expectations as the game (i.e., the set function in game theory) and then applies the classic Shapley value [36] to this game to obtain a value for each feature i. That is, from the set function $f_x(S)$, SHAP computes ϕ_i as follows (see, e.g., [37, 38] for details):

$$\phi_i(f, X) = \sum_{S \subseteq N \setminus \{i\}} \frac{|S|!\,(M - |S| - 1)!}{M!} \left[f_x(S \cup \{i\}) - f_x(S)\right] \tag{1}$$

KernelSHAP is an implementation of SHAP that is model-agnostic and, thus, can be used for any ML model f. It estimates for an instance x the contributions of each feature value of x to the prediction. All possible subsets are considered to compute the game and, then, the Shapley value, so its run time is exponential in the number of features. In contrast, TreeSHAP [35] is faster, but it is specific for tree-based ML models. It estimates SHAP values of tree ensembles and takes advantage of these trees to provide Shapley values.

2.2 Data privacy

Due to both GDPR regulations and XAI principles (see Fig. 1), AI systems need to be private by design. This means that privacy needs to be taken into account during the whole cycle. This requirement is essential when data-driven models are built using personal data. There are mainly two approaches to this. One consists of using sanitized data (using masking methods) and the other is building a privacy-preserving model (using e.g. differential privacy in the process of building the data).

When we have a researcher who wants to consider, develop, and test different machine learning methods, the former method is preferable. The study of effects of data protection mechanisms on data-driven models has been studied by several researchers (see [23] and references therein). Two methods that have been extensively used for data protection are microaggregation (to provide k-anonymity) and additive noise (to provide local differential privacy). In short, they modify the data in a way that is still useful but is privacy-preserving. We describe below these two methods. We have selected these two approaches because they are representative of methods providing either k-anonymity or differential privacy, the two most significant privacy models in the literature.

The implications of data protection mechanisms on model-agnostic explanation methods are unknown. When data are modified using, e.g., microaggregation, data-driven models are affected; however, the effect of the change is known. Small perturbations usually do not affect the accuracy of the models, and a good trade-off can be found between privacy and accuracy. For example, we have recently proven [39] that we can integrate protected databases in a way that data-driven models have good accuracy while reidentification attacks are unsuccessful. In contrast, the effects of data protection on SHAP are not known.

In Table 1, two data privacy methods are shown in a simple tabular dataset with three instances. In the MDAV method, the "Price" and "Area" attributes within each cluster have been replaced with the average values for those attributes within their respective clusters. This is a simplified illustration for demonstration purposes, and actual microaggregation involves more sophisticated clustering algorithms and calculations.

Also, Table 1 Laplacian noise has been added to the "Price" attribute to provide local differential privacy. In real used datasets, the noise parameters and distribution are chosen based on privacy requirements. It is noteworthy that the irregularity values tend to be closer to zero, the better performance. As indicated in Table 1, the irregularity observed in the maximum distance to average vector (MDAV) method is less than the noise addition method.

2.2.1 Microaggregation

Microaggregation [25, 40–42] is one of the most efficient approaches for data protection in relation to the trade-off risk-utility. It consists of building small clusters with the original data and then replacing each of the data with a cluster center that represents the whole cluster. Microaggregation permits us to implement k-anonymity (one of the privacy models) and

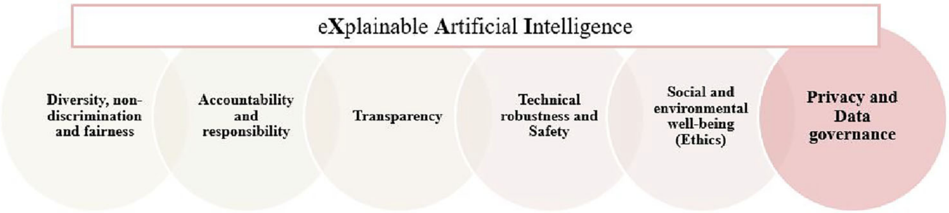

eXplainable Artificial Intelligence
Diversity, non-discrimination and fairness
Accountability and responsibility
Transparency
Technical robustness and Safety
Social and environmental well-being (Ethics)
Privacy and Data governance

Fig. 1 Six principles of eXplainable Artificial Intelligence

Table 1 Data privacy methods' samples and *Irregularity* values on simple datasets

House	Price	Area	Bedrooms	Bathrooms	Irregularity
Original dataset					
1	300000	1800	3	2	0
2	500000	2500	4	3	0
3	200000	1500	2	1	0
Protected dataset with MDAV					
1	333333	1767	3	2	– 108059.24
2	444444	2167	4	3	– 109149.815
3	222222	1667	2	1	– 99680.037
Added Laplacian Noise to "Price" Attribute					
1	288512	1800	3	2	– 199343.212
2	510230	2500	4	3	– 141550.215
3	220125	1500	2	1	– 101635.867

can be applied to any kind of data (from standard SQL with numerical and categorical data to complex data as graphs).

Let X be a data set where $x \in X$ are records described in terms of features $V_1, \ldots, V_n$. Then, microaggregation partitions the records in X into a set of clusters, where each cluster should have at least k elements. This is to ensure a pre-established degree of privacy. Then, each record in X is replaced by the representative of the cluster. E.g., the center of the cluster is defined as the mean of the records assigned to the cluster. In this way, we get a protected data set X' where there are groups of k indistinguishable records. Here, k is a parameter of the process and represents the privacy level. The larger the k, the larger the privacy. Naturally, the larger the k, the larger the perturbation. For a given k, the number of clusters c will be $|x|/c \geq k$.

For a given c, microaggregation is formally defined in terms of the minimization of the total Sum of Square Error. This is given in Eq. 2, where x is an individual record in the file (i.e., $x \in X$), C_i is the i-th cluster, and $\bar{x_i}$ is the centroid of C_i.

$$\mathrm{SSE} = \sum_{i=1}^{c} \sum_{x \in C_i} x - \bar{x_i} \tag{2}$$

This expression is the objective function. Then, we have the constraints $|C_i| \geq k$ for all $i = \{1, \ldots, C\}$ to ensure all clusters have at least k records.

Naturally, the lower SSE, the higher the within-group homogeneity [42], and the better the protection. When data is numerical, it is usual to use the Euclidean distance in the expression above. Nevertheless, other distances are also possible.

There are different heuristic solutions to the optimization problem above. In this work, we use maximum distance to average vector (MDAV). There are public implementations of this method (see, e.g., implementations in R and in μ-Argus), and it permits dealing with both numerical and categorical features. The algorithm is effective with the size of the files we use in our experiments.

2.2.2 Noise addition for local differential privacy

In our experiments, we also consider noise addition as a tool for data protection. Noise addition provides local differential privacy.

Dwork and her colleagues [30] introduced differential privacy as a framework to ensure that computation is safe from a privacy perspective. Informally, differential privacy has as its goal that the removal or addition of a single database item does not affect the outcome of any analysis (i.e., a query). This is to provide strong privacy guarantees. The privacy level is established by a parameter ϵ, maximum privacy guarantee is with $\varepsilon = 0$. Formally, the mechanism K constructed from query q satisfies differential privacy if for all pair of databases D_1 and D_2 differing in one record the following holds:

$$P_r[K_q(D_1) \in S] \leq e^{\epsilon} \times P_r[K_q(D_2) \in S]$$

The standard approach for preserving differential privacy in numerical data is adding noise with the Laplacian distribution to the query output [32]. Formally, given $q(X)$, the output of query q on the database X, differential privacy returns $K_q(X) = q(X) + r$ where r is a random noise drawn from $L(0, b)$ where $b = \Delta(q)/\varepsilon$. $Delta(q)$ stands for the sensitivity of the query for all pairs of datasets $D_1, D_2 \in \mathcal{D}$ that differ in one record.

Local differential privacy is a variation for databases, where the goal is to protect the records of the database. Then, each record is protected independently, and noise is added to features so that the values of a record cannot be distinguished from the ones of other records. Each variable has a range of possible variables. This provides a sensitivity (i.e., Δ).

3 Proposed method

The methodology for our experiments[1] consists of two main steps, as shown in Fig. 2. First, data were masked (using two data privacy techniques). Second, we applied the explainable machine learning algorithms on the protected data to create machine learning models and explain the results of their predictions. The protected explainable machine learning models were compared with the ones obtained from data sets without masking. The goal of this comparison is to show the effect of privacy methods on XAI models. Three metrics were used to assess protected explainable machine learning models' functionality. They are *Irregularity*, *Utility*, and *Rank correlation*. Utility and rank correlation are metrics used in the literature, while irregularity is a new metric to show how SHAP-related information changes in the process.

3.1 Irregularity

We define *Irregularity* to assess the change on the information associated to SHAP. It is calculated by Eq. 4. It comprises (sums) two parts. The first corresponds to features' $Distortion\ Rate(DR)$. It measures in what extent the order of importance of the features changes. Given n features, this is computed by Eq. 3.

$$DR\left(X, X'\right) = \frac{\sum_{i=1}^{n} m_i}{\mathrm{n}} \tag{3}$$

where $m_i = 1$ if and only if the order of the ith feature changes with masking.

The second part takes into account the SHAP value coefficients and their variation. SHAP values are modified when the masking methods are applied. So, $\phi_i(f, X)$ and $\phi_i(f, X')$ are, in general, somehow different. The expression accounts for this variation.

$$\text{Irregularity}\left(X, X'\right) = \frac{\sum_{i=1}^{n} m_i}{\mathrm{n}} + \left(\sum_{i=1}^{n} \frac{\phi_i(f, \mathbf{X})}{\phi_i(f, \mathbf{X}')}\right) \tag{4}$$

The Shapley value $\phi_i(f, \mathbf{X}')$ corresponds to Equation 1, and n corresponds to the number of features.

4 Empirical work

We have applied to each data set the following data protection mechanisms: MDAV microaggregation and Laplacian noise addition. Each mechanism has different parameterizations. This results in several protected versions of the same original dataset. Then, explainable machine learning models are trained on protected and non-protected data sets, separately. More concretely, we use TreeSHAP to compute SHAP values for all data sets. As explained in 2.1.1, TreeSHAP builds a tree-based ML model from which the SHAP value is computed. Finally, we analyze the results.

4.1 Datasets

The analysis has been conducted on three multivariate data sets. Their properties are listed in Table 2. These real data sets are "Cervical cancer (risk factors)", "Breast cancer Coimbra", and "USA House", which were previously used in the literature on data privacy [43] and explainability [44]. These datasets have been normalized substracting the mean and dividing by the standard deviation.

4.2 Parameterizations

Datasets have been protected by microaggregation (MDAV algorithm) and Laplacian noise. For MDAV microaggregation, 15 masked datasets have been obtained using the parameter $k = \{1, \ldots, 15\}$. Microaggregation was applied considering all variables at the same time, and, thus, producing k-anonymous files with increasing protection levels. Laplacian noise addition was applied by parameters $b = [1, 5]$. For all these original and protected files, TreeSHAP models were built.

[1] The code related to this paper is published in https://github.com/bozorgpanah/The-Explainable-Machine-Learning-Model-withPrivacy.

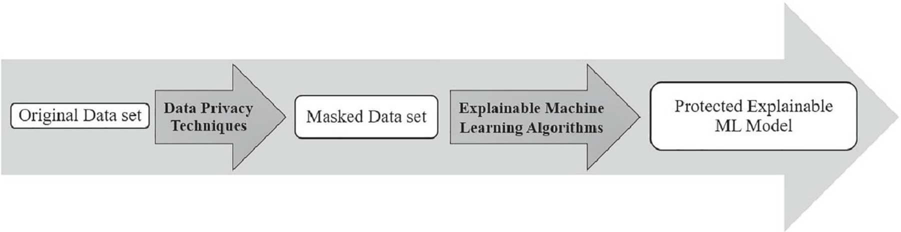

Original Data set
Data Privacy Techniques
Masked Data set
Explainable Machine Learning Algorithms
Protected Explainable ML Model

Fig. 2 The process of training a protected explainable machine learning model

Table 2 Datasets' properties

Name	Attribute characteristics	Instances	Attributes
Cervical cancer (risk factors)[a]	Integer, Real	858	36
Breast cancer Coimbra	Integer	116	10
USA House[b]	Integer	5,000	5

[a]Used by Molnar[44] to analyze explainable methods
[b]Used by Rodriguez-Hoyoz et al. [43] to analyze MDAV behavior
The data sets are available from the UCI Machine Learning repository

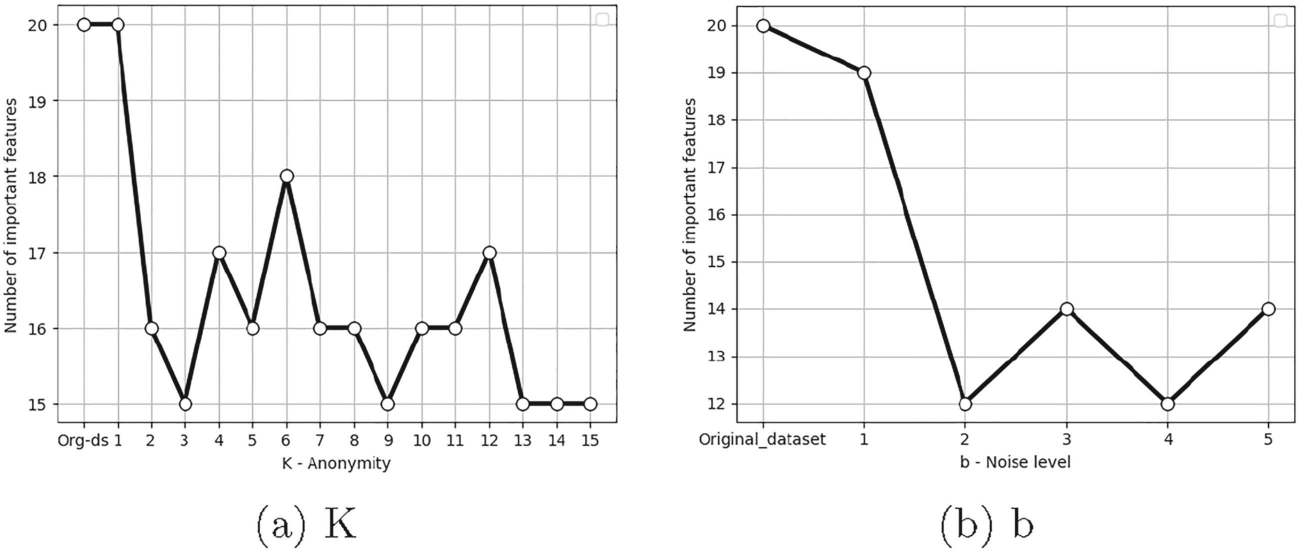


Fig. 3 Selected more important features by SHAP on masked the Cervical cancer (risk factors) datasets by MDAV and noise addition methods

4.3 Cervical cancer risk factors

We have analyzed the results of TreeSHAP for original and protected files. Figure 3 plots[1] the top 20 features (according to SHAP) for models built from the original dataset and a selection of protected ones. Each plot indicates how important features are in terms of their impact on the output. The bar chart shows the average impact of each feature on the final prediction. Features are ranked in decreasing order of SHAP values, taking into account their contribution to the four target classes. Colors in the bars represent target classes: 'Citology', 'Hinselmann', 'Schiller', and 'Biopsy' are colored by blue, purple, red, and green, respectively.

More particularly, Fig. 4a corresponds to prominent features for the original dataset. Then, results corresponding to data protected with MDAV and k equal to 3, 10, and 13 are shown in Fig. 4 b, d, and f, respectively. Results for masked data using additive noise are depicted in Fig. 4c and e with $b = 2$ and $b = 3$. We can compare the 20 most important features for the model on the original data and the models on the protected data. We can see that among the 20 most important features for the original data we keep 15, 16, and 15 among the 20 most important ones when data are protected using MDAV. While after applying noise addition, for $b = 2$ and $b = 3$, 13 and 10 influence features are raised up. We can also observe that the structure of the plots corresponding to MDAV is more similar to the original one, than when using noise addition.

[1] In linear plots, "Org-ds" is an abbreviation of "Original dataset".

Moreover, mean SHAP values on the x-axis change significantly for additive noise. As the ordering of important features and the average SHAP values are different for different k and b, we study them further in Sect. 5. Molnar [44] averages the absolute Shapley values per feature across the data to show its global importance.

To better compare the effect of data protection on the variables selected as important, Fig. 3 displays the number of variables with respect to k and b values. Figure 3 (left) shows the case of MDAV for different values of k, the value fluctuates slightly in the range [14, 18]. For k = 6 the number of features reach the peak with 18 features. In contrast, for noise addition (right), values are in the range [12, 14] for values of b between 2 and 5.

To quantify how individual target variables independently affect the model output, some SHAP summary plots are presented in Figs. 5 and 6. They show the distribution of Shapley values for each feature in relation to the impact on the prediction of the four target classes 'Hinselmann', 'Schiller', 'Citology', and 'Biopsy'. Each point corresponds to an instance, the position on the x-axis corresponds to its Shapley value, and the color to the feature value from low to high. Redder points mean higher values for the feature, whereas bluer points mean a lower value for the attribute. In the y-axis direction, overlapping points provide a sense of the SHAP value distribution per feature. Figure 5 depicts four subplots (one for each target class), which correspond to Figs. 4a, and 6 corresponds to Fig. 4d

According to Fig. 5, *DX:Cancer*, the most influencing feature has the top rank for the four target classes, and a high Shapley value of this feature means its influence on "Cervical cancer" risk's probability. Moreover, a low number of years on $hormonal\ contraceptives$ reduce the predicted cancer risk; in contrast, a large value of years increases the risk, as depicted in Fig. 5a, b, and d. Figure 5 c shows that the prediction is not affected by *hormonal contraceptives (years)* in 'Citology' target mode, while this feature does impact the other three target classes.

In Fig. 6, for masking data with MDAV and $k = 10$, *hormonal contraceptives (years)* emerges as the most relevant feature for all four target classes. Protection with $k = 10$ widens the distribution and SHAP values are further changed. Masking with different values of k leads to different results.

The impact of input variables on the prediction of model's output for column 'Hinselmann' target class can be observed in Fig. 7 for both non-protected and protected data set. It qualitatively describes the overall relationship between risk factors' explanatory variables in "Cervical cancer (risk factors)".

Figure 7a shows the SHAP summary plots that characterize the overall impacts of 20 top-ranking risk factors features on the original data set with Shapley value = [−0.2, 0.4].

In subplot Fig. 7a (original data), higher *DX:Cancer* values are associated an increased risk of "Cervical cancer (risk factors)". Figure 7b, d, and f illustrates what we have explained, that after masking by MDAV results are mostly similar to the original dataset's plot, even if the SHAP values vary slightly. In any case, they remain in [−0.2, 0.4].

In Fig. 7b, we can see that 15 out of the 20 most important risk factors on the original data set are also relevant after applying MDAV with k = 3. Similarly, Fig. 7d and f gives 16 and 15, respectively. Although the density of the points is different (because of the masking process), the outcome is not affected drastically. So, these results seem to show that protection using MDAV result in quite acceptable results for some values for parameter k.

According to our experiments and Fig. 3, masking "Cervical cancer (risk factors)" data set with $k = 6$ seems to be a good alternative as we get a model similar with respect to explainability to the one of the original data set.

In contrast, more negative SHAP values and variations on points density in Fig. 7c and e indicate that applying noise addition on "Cervical cancer (risk factors)" data set decreases accuracy in prediction (see also Fig. 14). In other words, we may misidentify the important relationships between "Cervical cancer" risk and explanatory variables. As Fig. 7c and e show, the priority features and the SHAP value distribution per all features are changed. In this case, we cannot extract the same explanation after protection by adding noise. Thus, it is not a reliable method to preserve "Cervical cancer (risk factors)" data set.

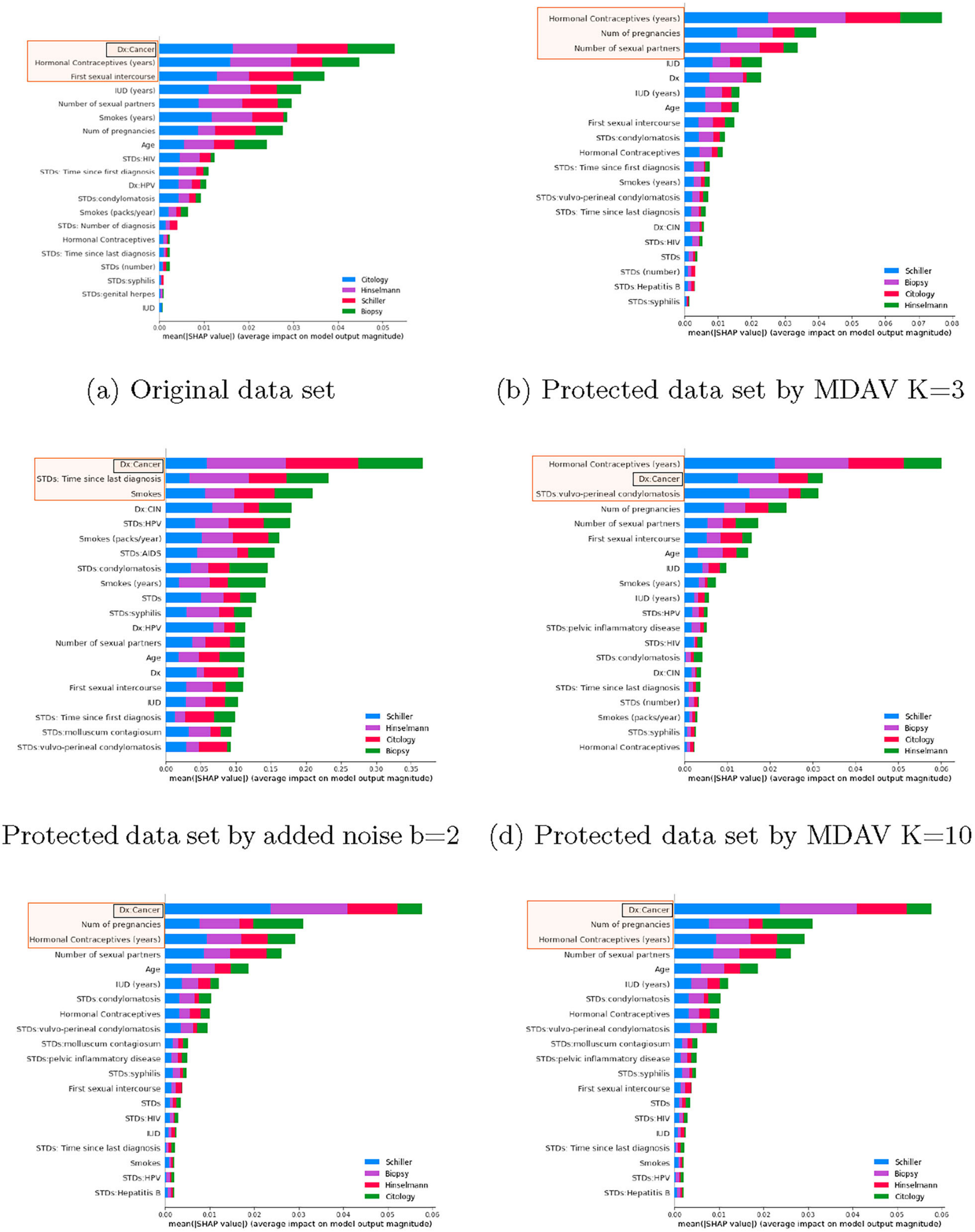
4.4 Breast cancer Coimbra

Figure 8 reports SHAP feature importance plots on the original "Breast cancer Coimbra" dataset and protected ones by MDAV and additive noise.

It can be seen that the relative influence of features has changed. Among the four most relevant features of the original data set, *Glucose*, *Resistin*, *Homa*, and *Adiponectin*, there are three out of four retrieved features by MDAV with $k = 3$, and one for each $k = 10$ and $k = 13$. Further, two out of four relative influences of features are displayed after applying noise addition with different $b = 2$ and $b = 3$.

Hence, for the "Breast cancer Coimbra" data set, the MDAV method seems to provide better-protected data in terms of explainability. It is also relevant to show that the mean SHAP values have also changed significantly for noise addition (see Fig. 8c, e).

Figure 9 displays the relationship between the feature value and its impact on the prediction. It depicts the original data on subfigure a and masked data sets with various k in MDAV; $k = 3, 10, 13$ through subfigures (b, d, f). Figure 10 represents the distortion rate with respect to the parameters of MDAV and noise addition.



(a) Original data set

(b) Protected data set by MDAV K=3

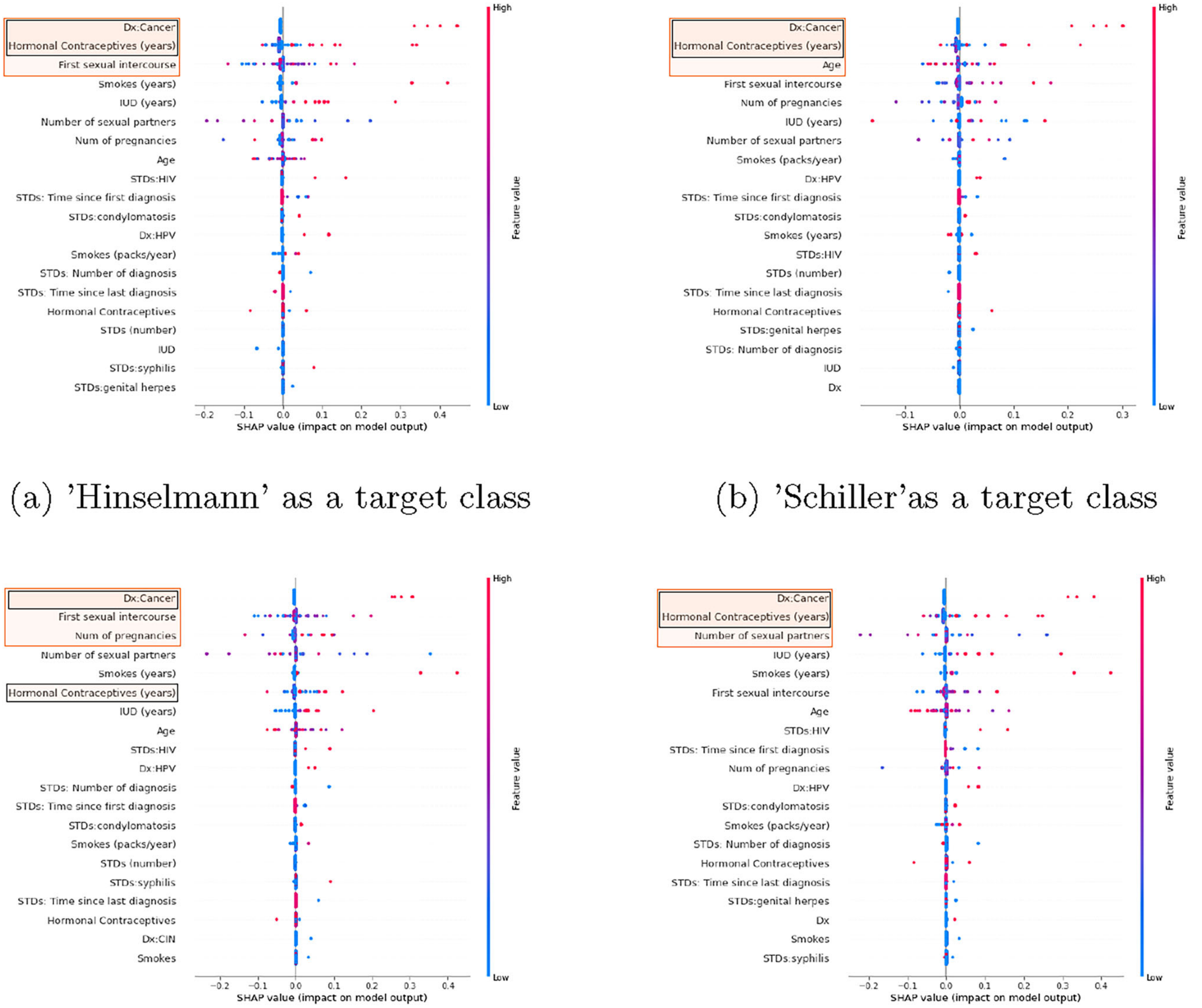
(c) Protected data set by added noise b=2

(d) Protected data set by MDAV K=10

(e) Protected data set by added noise b=3

(f) Protected data set by MDAV K=13

Fig. 4 SHAP feature importance plots for protected and non-protected “Cervical cancer (risk factors)” dataset



(a) 'Hinselmann' as a target class

(b) 'Schiller'as a target class

(c) 'Citology' as a target class

(d) 'Biopsy' as a target class

Fig. 5 SHAP summary plots for original Cervical cancer (risk factors) dataset to four targets classes

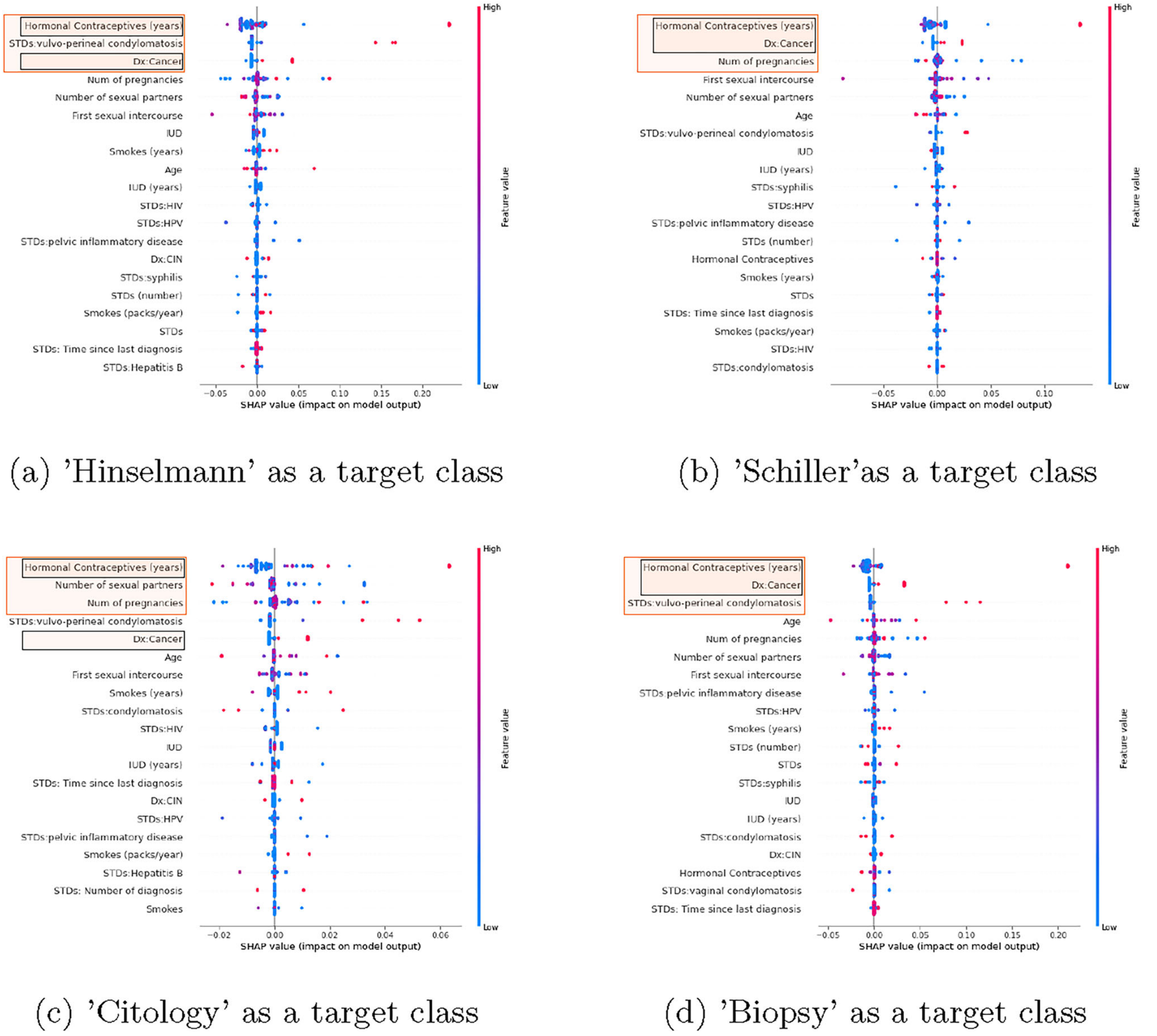
The results of additive noise are given in Fig. 9 (subfigures c, e) for $b = 2, 3$. We see that for $b = 2$ the three most important features have dropped significantly in importance and that the one in the last position in the original data set is for $b = 2$ the most relevant one. If we compare noise addition with MDAV methods on the Breast cancer Coimbra dataset, the results after MDAV are more similar to the original dataset plot in comparison with noise addition results. These changes are more clear when we take into account the x-axis values of plots (c) and (e).

The *Distortion Rate* (see Expression 3) is used to estimate features' relative influence after masking. Figure 10 summarizes the distortion rate with respect to different k and b values for microaggregation and noise addition, respectively. Figure 10a shows that the distortion rates are in the range [0.55, 1] for k-anonymity, while it is [0.8, 1] for noise addition. The less distortion the better, because less distortion shows less disorder after masking. The distortion rate is a sub-parameter of irregularity, that is discussed in Sect. 3.1.

We can underline that Fig. 10 shows that values of $k = 6$ and 12 lead to low values of distortion rate, lower than for noise addition. So, these results show that we can tune MDAV parameters to get acceptable explainability results.

4.5 USA House

Figure 11 depicts that *Avg. Area Income* has a high relative influence in the model, and this feature remains the most



(a) 'Hinselmann' as a target class

(b) 'Schiller'as a target class

(c) 'Citology' as a target class

(d) 'Biopsy' as a target class

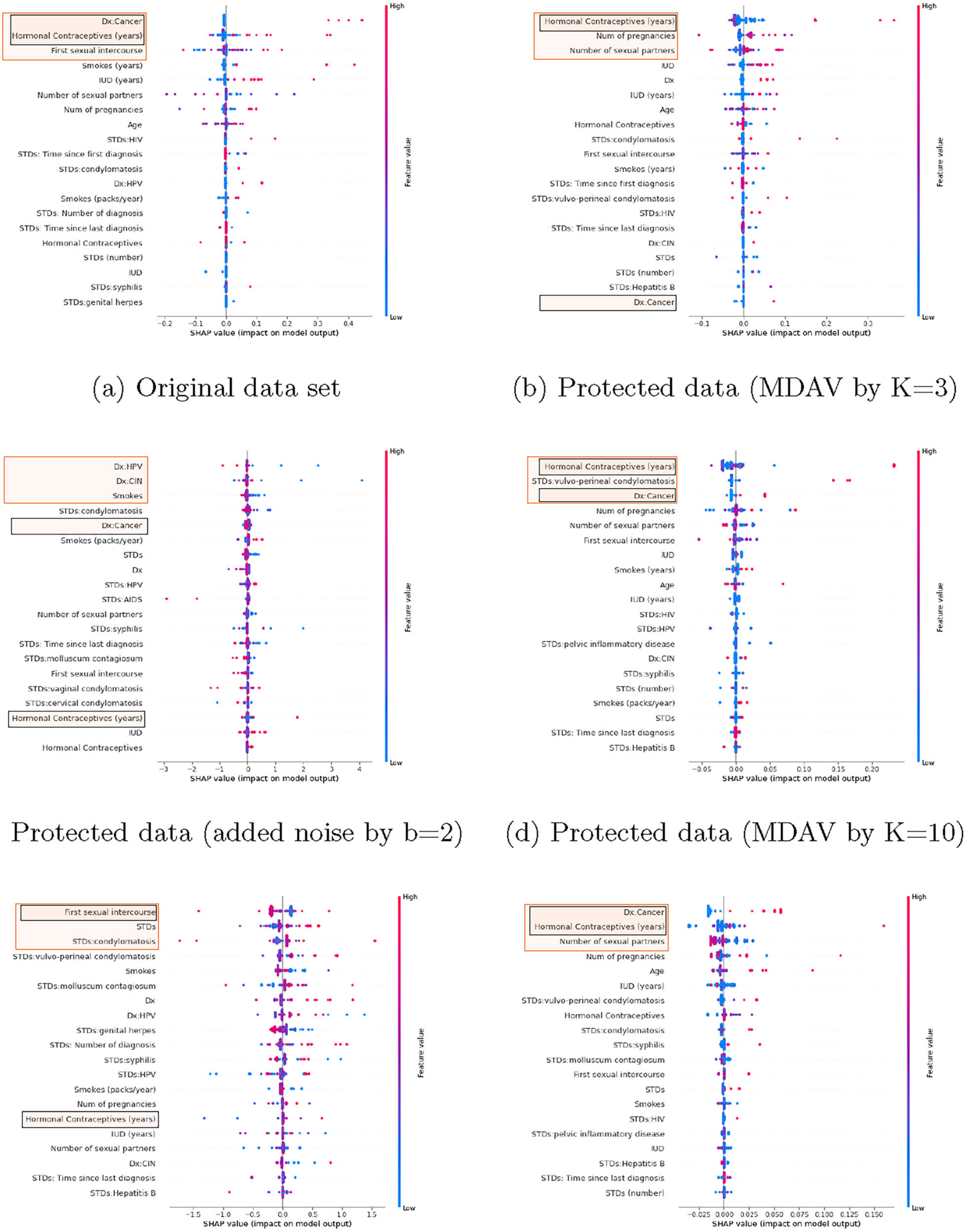
Fig. 6 SHAP summary plots for masked "Cervical cancer (risk factors)" dataset to four targets classes with k-value = 10

important one even after masking. Furthermore, the feature with more relative influence has a larger SHAP value. The amount of *Avg. Area Income* has a crucial role in the explanation of house sales prediction.

The *Avg. Area House′s Age* and *Area Population* are the next two impact features after *Avg. Area Income*. Figure 11b and d shows symmetry with the same scale we see in a, whereas in c the Shapley value increased 2/3 times. It means the sequence of features is the only efficient parameter to analyze different protected data. Hence, as b and d depict for MDAV and $k = 3, 6$, the ordering of features is kept as for the original data while for additive noise, disorder emerges between top features.

In particular, according to the data from the original model, old houses with fewer rooms and bedrooms are sold earlier. By contrast, c just explains that a house with fewer rooms and bedrooms will be sold easier. Therefore, the results show that the explanation after noise addition is not complete as the one obtained after MDAV.

In Fig. 12, feature values are on the x-axis and the corresponding SHAP value is on the y-axis. This is for a particular feature, and it is used to visualize the impact of changing another feature. In particular, we select *Avg. Area House′s Age* when *Avg. Area Income* increases from 55000 to 85000. In these subfigures, the red points represent higher values of *Avg. Area Income*, and the blue points represent lower ones. Since the red points' density in all over the images are more



(a) Original data set

(b) Protected data (MDAV by K=3)

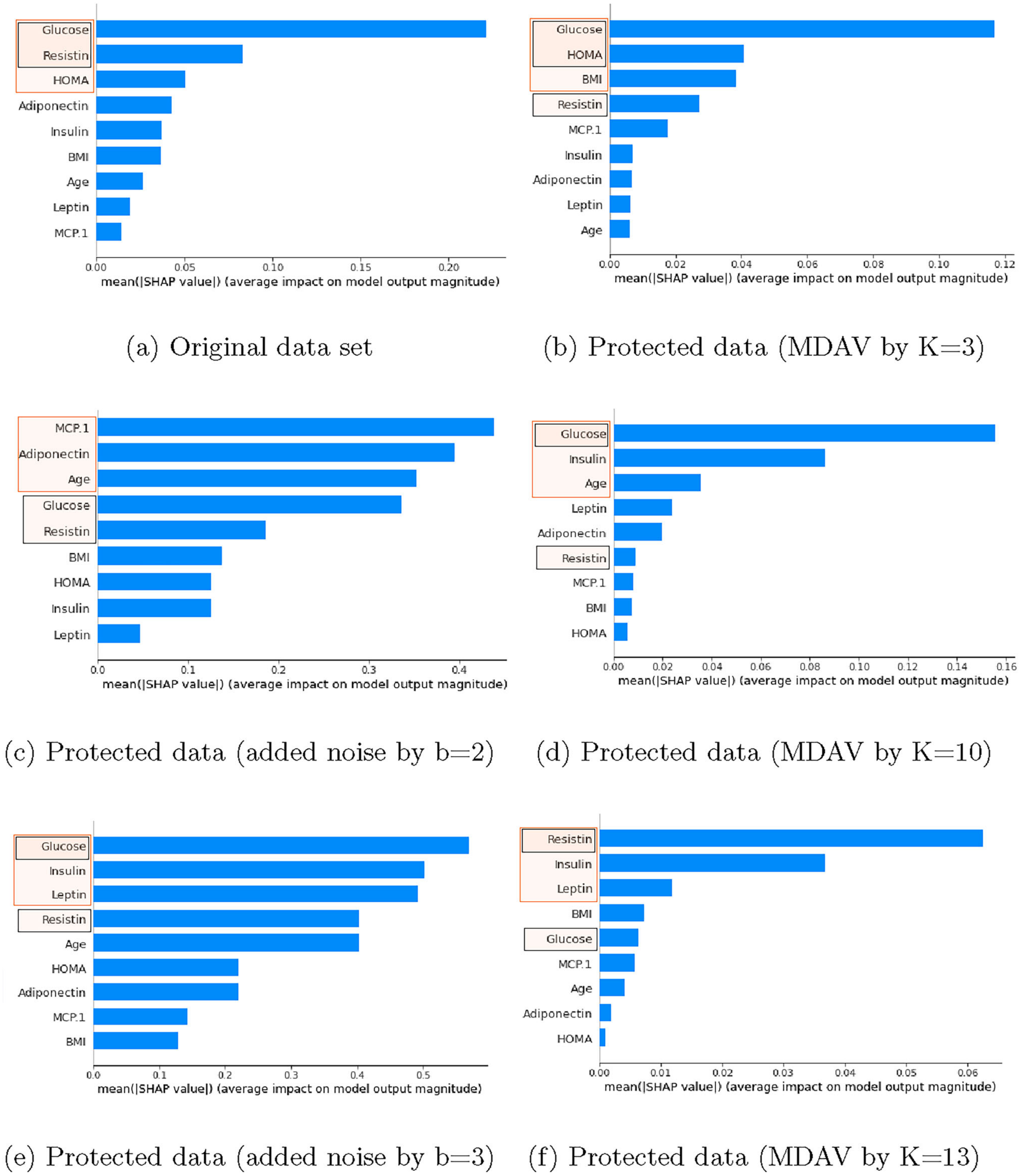
(c) Protected data (added noise by b=2)

(d) Protected data (MDAV by K=10)

(e) Protected data (added noise by b=3)

(f) Protected data (MDAV by K=13)

Fig. 7 SHAP summary plots for original “Cervical cancer (risk factors)” dataset on Hinselmann target class



(a) Original data set

(b) Protected data (MDAV by K=3)

(c) Protected data (added noise by b=2)

(d) Protected data (MDAV by K=10)

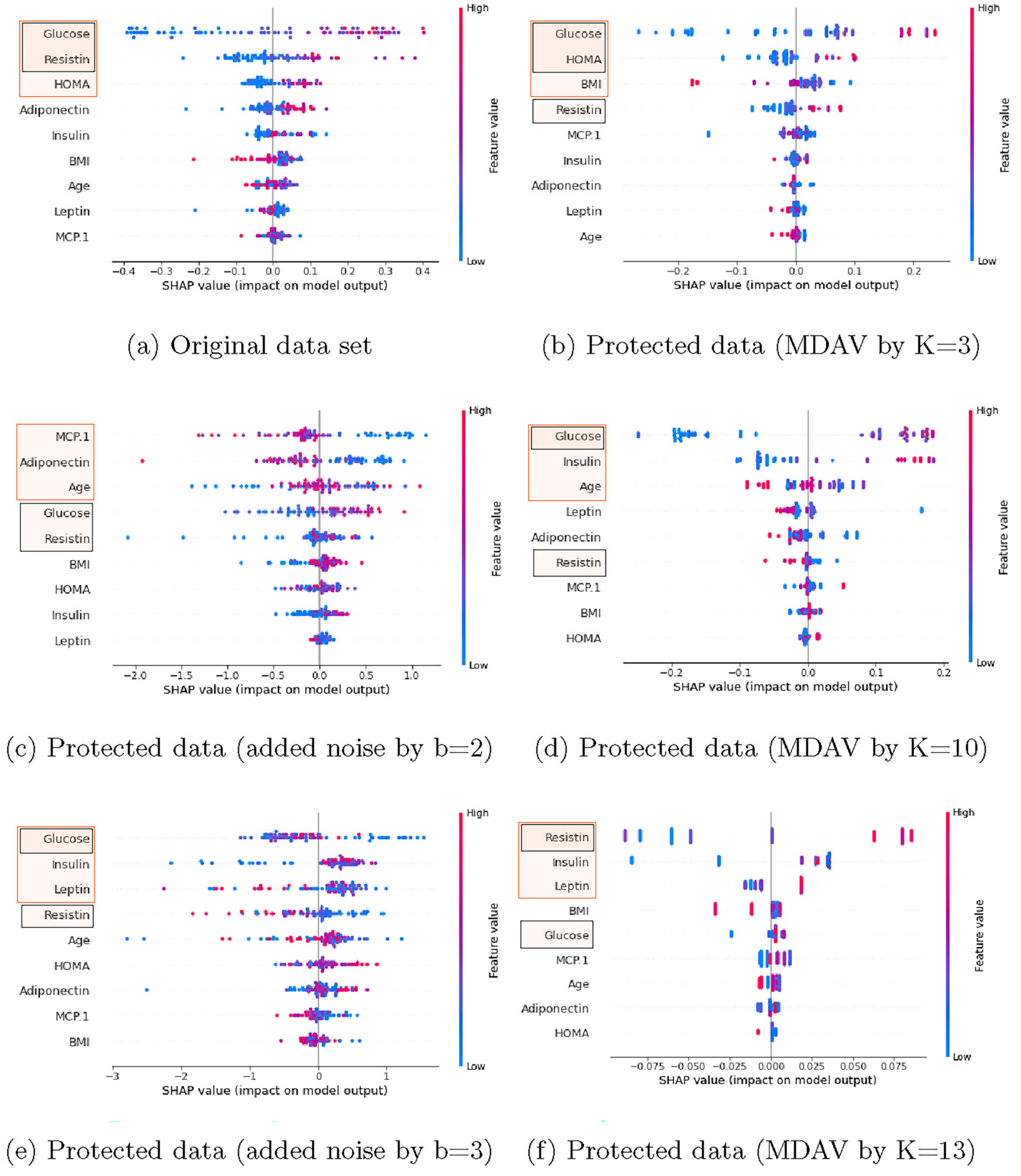
(e) Protected data (added noise by b=3)

(f) Protected data (MDAV by K=13)

Fig. 8 SHAP feature importance plots on “Breast cancer Coimbra” dataset in original and protected data

than the blue ones, we can state that the *Avg. Area Income* feature has a positive impact on selling houses. Also, when *Avg. Area House*’*s Age* is less than 6 years and *Avg. Area Income* is under 70000, SHAP values are less than zero, which suggests that new houses with less *Avg. Area Income* feature has a low chance of being sold.

By contrast, Fig. 12 explains when *Avg. Area House*$'$*s Age* is more than 6 years and *Avg. Area Income* more than



(a) Original data set

(b) Protected data (MDAV by K=3)

(c) Protected data (added noise by b=2)

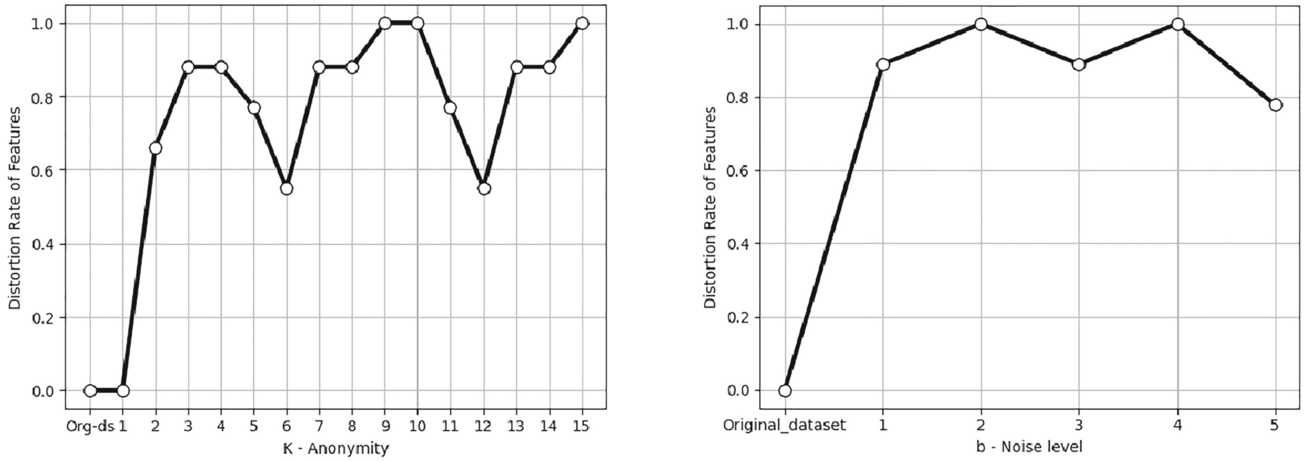
(d) Protected data (MDAV by K=10)

(e) Protected data (added noise by b=3)

(f) Protected data (MDAV by K=13)

Fig. 9 SHAP summary plots on "Breast cancer Coimbra" dataset in original and protected data

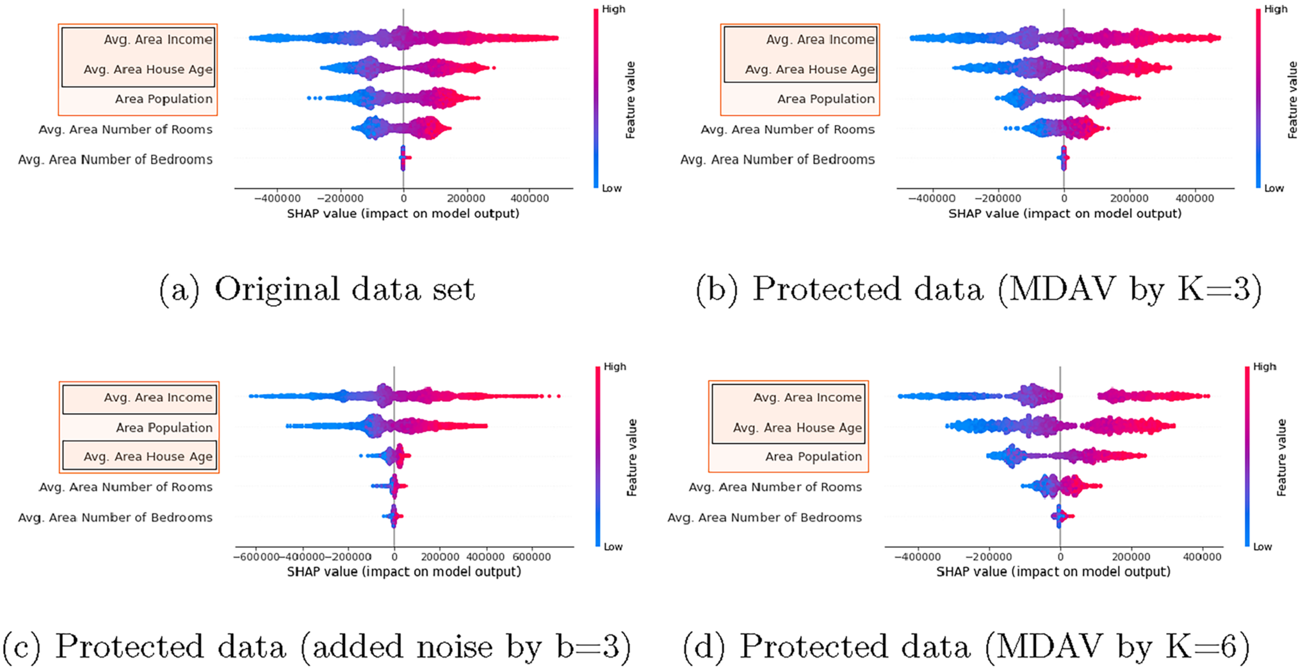
70000, overall SHAP values are positive, which means that increasing the *Avg. Area House's Age* and *Avg. Area Income* values, increases the probability of the house being sold. This explanation is observed in a, b, and d, but c shows a different analysis. Houses newer than 8 years have less chance to be sold, and some houses in the range [8, 17] will be sold earlier. In contrast, the range of the house's age in the original data is [6, 8.5]. Generally, for noise addition, the Shapley value is



(a) K

(b) b

Fig. 10 Distortion rate of features after applying data privacy methods on "Breast cancer Coimbra" dataset



(a) Original data set

(b) Protected data (MDAV by K=3)

(c) Protected data (added noise by b=3)

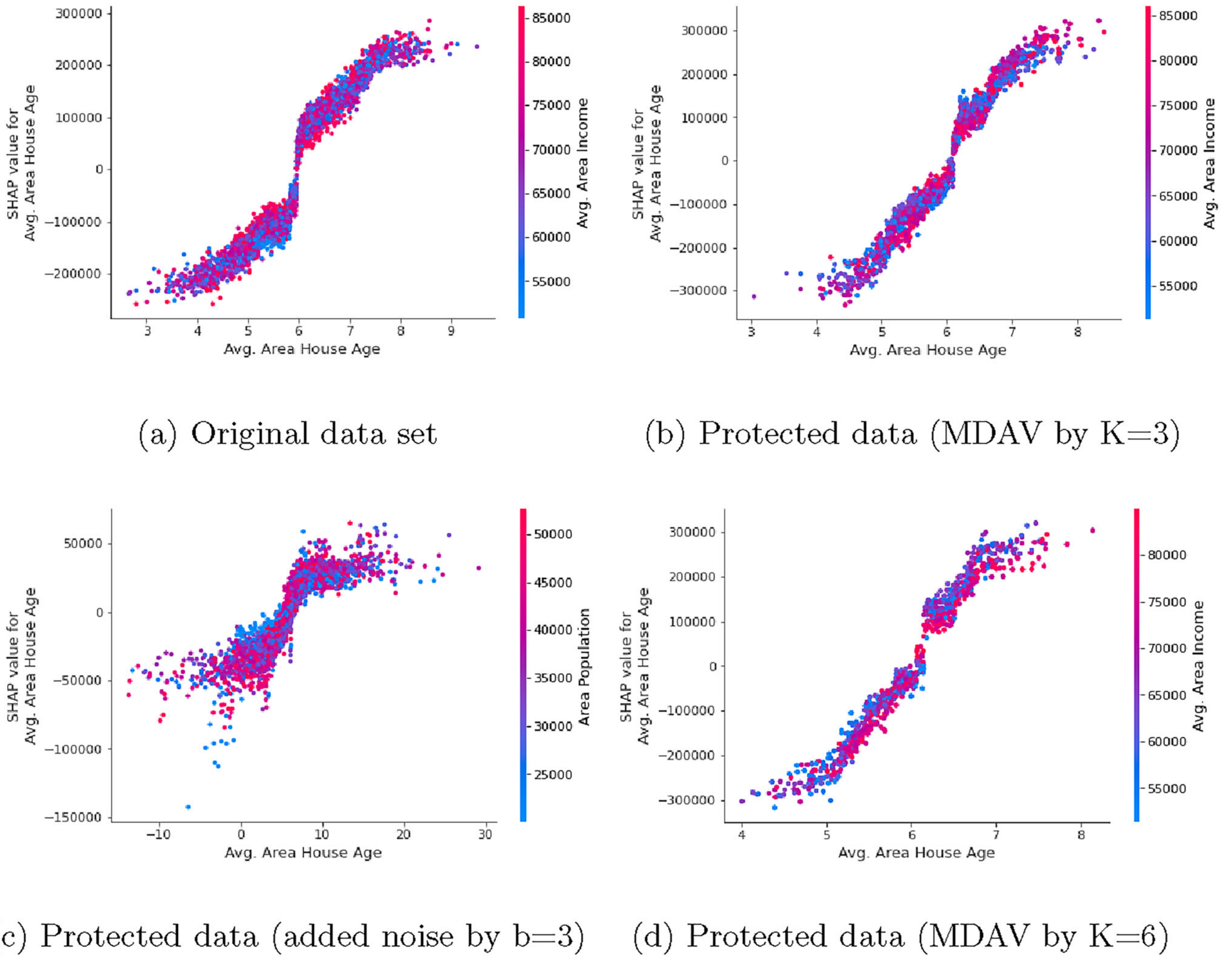
(d) Protected data (MDAV by K=6)

Fig. 11 SHAP Summary plots for the original and masked "USA House" data set

changed dramatically. Then, the explanation and prediction are affected significantly.

In addition, increasing the protection decreases the density. This is because when the value of k for k-anonymity increases, the number of repeated (indistinguishable) records also increases. This growth is valid as long as the main content is not changed. As a general fact, and from what we see also in the other results (see Sects. 4.3 and 4.4), the value for k is better to be within [6, 12]. Also, the k values (range) depend on the data set's size. In this range, the data diversity decreases, but dataset's content remains with a correct explanation.

From the perspective of explainability, all diagrams in Fig. 12a, b, and d have some linearity. Figures show that houses with both *Avg. Area House′s Age* larger than 6 years old and *Avg. Area Income* more than 70000, have a positive impact when selling. Figure 12c is not linear and the results are different when compared to the case of non-masked data. Furthermore, *Avg. Area House Age* is changed to $[-15, 25]$ for $b = 3$, when negative years are not valid.



(a) Original data set

(b) Protected data (MDAV by K=3)

(c) Protected data (added noise by b=3)

(d) Protected data (MDAV by K=6)

Fig. 12 SHAP dependency analysis for the original and masked "USA House" data set

5 Experimental evaluation

To analyze the models, we have used three metrics *Irregularity*, *Utility*, and *Rank correlation*. They are presented in the following sections.

Irregularity: Figure 13 displays irregularity, using Eq. 4, with respect to k and b. In general, irregularity for MDAV for various k-values is about 1.25 on "USA House", and "Breast cancer Coimbra" data sets. The irregularity fluctuates for "Cervical cancer (risk factors)". "Cervical cancer (risk factors)" has 36 features in four target classes, and SHAP is appropriate for multi-class data sets. Nevertheless, after masking, the irregularity grows significantly. As Figs. 3, 4, 5, and 7 show, the results about explainability after masking by MDAV have not changed significantly, which seems consistent with these results.

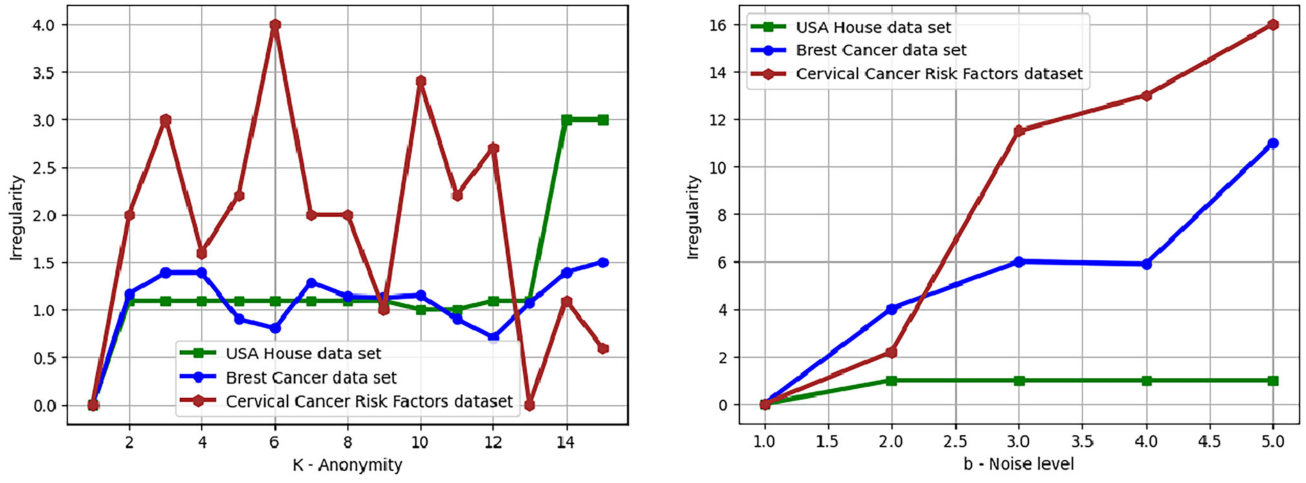
We achieved that the protected models (by MDAV) are explainable as understandable as models trained on real data sets (without protected data), so both $explainability$ and $privacy$ can be compatible in an ML models.

Figure 13b shows that irregularity for "Cervical cancer (risk factors)" and "Breast cancer Coimbra" increases continuously when b grows from zero to 16 and 11, respectively. Hence, MDAV behaves better in this analysis.

Utility: After masking datasets by microaggregation and local differential privacy with various privacy levels k and b, we built machine learning models. For this purpose, we considered decision tree learning. Then, we assessed the accuracy of the models built. This was applied to the datasets considered in the experiments' Sects. 4.3, 4.4, and 4.5.

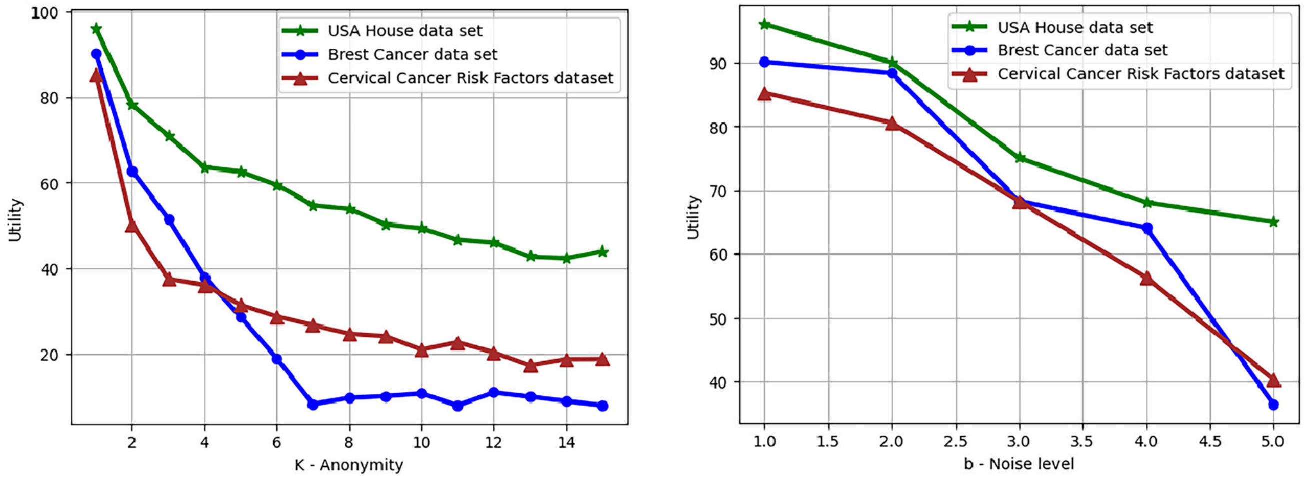
Figure 14a displays the utility for different k-anonymity in [1, 15], as the value of k increases, the accuracy declines slightly for "Cervical Cancer Risk Factors" data set, its progress starts about 85.21 and continues to reach 20 smoothly.

Most changes are done by $k = [1, 9]$ then for $k > 9$ the utility value is not modified anymore, and it remains about 20. Similarly, "USA House" dataset for $k = [1, 12]$ decrements the utility from 96.01 to 43.99 then it is mostly stable for $k > 13$. Also, there is the same linear decrease for "Breast



(a) K (b) b

Fig. 13 Irregularity after masking three data sets



(a) K (b) b

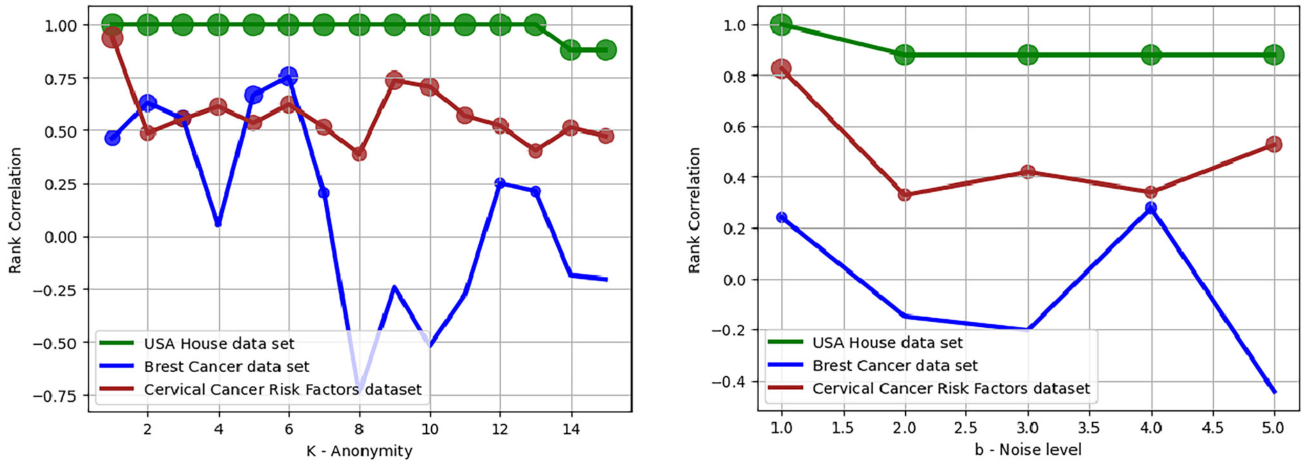
Fig. 14 Utility values for various K-anonymity and **b** as additive noise level

cancer Coimbra” datasets $k = [1, 7]$. In general, the k value should be selected within a range that provides high accuracy. In the case of additive noise, utility sharply declines. Figure 14b shows that for $b = [1, 5]$ efficiency decreases significantly for “Breast cancer Coimbra” and “Cervical cancer (risk factors)” data sets. This behavior is not so acute on “USA House”. Observe that there is a considerable change from 96.01 to 65 for $b = [1, 5]$, respectively.

Considering the impact of k on utility, we would select a value for k or b that leads to a user acceptable risk level and at the same time leads to good accuracy. So, we train protected XAI models, which they provide a fair privacy-utility tradeoff and it is the goal for user purposes.

Rank correlation: In order to compare the results of SHAP, we consider rank correlation. Figure 15 displays rank correlation between features for k-anonymity in (a) and additive noise in (b).

Figure 15 shows a high correlation between features in the “USA House” data set compared with “Cervical cancer risk factors”, and “Breast cancer Coimbra” data sets. The large points illustrate high correlation and low correlations are displayed by smaller points. In Fig. 15a, rank correlation fluctuates for different k-values widely for “Breast cancer Coimbra” and slightly for “Cervical cancer risk factors”. In Fig. 15b, the highest rank correlation is for the “USA House” and the lowest one for “Breast cancer Coimbra”. The results show clearly for the three datasets that rank correlation for MDAV is higher than for noise addition. For the three data sets, Fig. 15b “Breast cancer Coimbra” shows that, for several values of b, there is no high correlation after masking.



(a) K

(b) b

Fig. 15 Rank correlation between features on different k-anonymity and *b* as additive noise level

6 Conclusion

In this paper, we studied data masking's impact on explainability. We have compared data privacy methods considering SHAP for explainable machine learning models. The evaluation was for decision trees, thus comparing the models built from the original and the masked data sets. Two masking methods are considered MDAV and noise addition. The former provides k-anonymity and the second differential privacy. Three different data sets were considered. Experimental results were evaluated using three metrics: *Irregularity*, *Utility*, and *Rank correlation*.

We found that the MDAV method is the best-performing one. Our observations indicate that the explainable models derived after masking maintain a significant alignment with the explanation models originating from the original dataset. This alignment becomes particularly pronounced for select values of k across various datasets. After protecting data by MDAV, the data had fewer irregularities, which preserve the utility (accuracy) of prediction. Also, the correlation ranges between 0.5 and 1.0. On the opposite, when data were masked using noise addition, irregularities increased after masking data, which led to utility reductions, and at the same time correlation had a larger range.

To sum up, we consider that explainable machine learning models can be considered along with privacy if data privacy methods preserve the three considered metrics. That is important features ranking (Irregularity), correlation among instances, and utility.

Future research includes doing research on other XAI requirements within a privacy-preserving framework to assess to what extent these tools apply in privacy-by-design machine learning.

Funding Open access funding provided by Umea University.

References

1. Marcinkevičs, R., Vogt, J.E.: Interpretable and explainable machine learning: a methods-centric overview with concrete examples. Data Mining and Knowledge Discovery, Wiley Interdisciplinary Reviews, p. e1493 (2023)
2. Carvalho, D.V., Pereira, E.M., Cardoso, J.S.: Machine learning interpretability: a survey on methods and metrics. Electronics **8**(8), 832 (2019)
3. Vale, D., El-Sharif, A., Ali, M.: Explainable artificial intelligence (XAI) post-hoc explainability methods: risks and limitations in non-discrimination law. AI Ethics **2**(4), 815–826 (2022)
4. Slack, D., Hilgard, S., Jia, E., Singh, S., Lakkaraju, H.: Fooling lime and shap: adversarial attacks on post hoc explanation methods. In: Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society, pp. 180–186 (2020)
5. Roth, A.E. (ed.): The Shapley Value: Essays in Honor of Lloyd S. Cambridge University Press, Shapley (1988)
6. Ribeiro, M.T., Singh, S., Guestrin, C.: "Why should I trust you?" Explaining the predictions of any classifier. In: Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, pp. 1135–1144 (2016)

7. Štrumbelj, E., Kononenko, I.: Explaining prediction models and individual predictions with feature contributions. Knowl. Inf. Syst. **41**(3), 647–665 (2014)
8. Arrieta, A.B., Díaz-Rodríguez, N., Del Ser, J., Bennetot, A., Tabik, S., Barbado, A., García, S., Gil-López, S., Molina, D., Benjamins, R., Chatila, R., Herrera, F.: Explainable Artificial Intelligence (XAI): concepts, taxonomies, opportunities and challenges toward responsible AI. Inf. Fusion **58**, 82–115 (2020)
9. Grant, T.D., Wischik, D.J.: Show us the data: privacy, explainability, and why the law can't have both. Geo. Wash. L. Rev. **88**, 1350 (2020)
10. Patel, N., Shokri, R., Zick, Y.: Model explanations with differential privacy. In: Proceedings of the 2022 ACM Conference on Fairness, Accountability, and Transparency, pp. 1895–1904 (2022)
11. Bozorgpanah, A., Torra, V., Aliahmadipour, L.: Privacy and explainability: the effects of data protection on Shapley values. Technologies **10**(6), 125 (2022)
12. Nori, H., Caruana, R., Bu, Z., Shen, J.H., Kulkarni, J.: Accuracy, interpretability, and differential privacy via explainable boosting. In: International Conference on Machine Learning, pp. 8227–8237. PMLR (2021)
13. Renda, A., Ducange, P., Marcelloni, F., Sabella, D., Filippou, M.C., Nardini, G., Stea, G., Virdis, A., Micheli, D., Rapone, D., Baltar, L.G.: Federated learning of explainable AI models in 6G systems: towards secure and automated vehicle networking. Information **13**(8), 395 (2022)
14. Huong, T.T., Bac, T.P., Ha, K.N., Hoang, N.V., Hoang, N.X., Hung, N.T., Tran, K.P.: Federated learning-based explainable anomaly detection for industrial control systems. IEEE Access **10**, 53854–53872 (2022)
15. Bárcena, J.L.C., Ducange, P., Ercolani, A., Marcelloni, F., Renda, A.: An approach to federated learning of explainable fuzzy regression models. In: 2022 IEEE International Conference on Fuzzy Systems (FUZZ-IEEE), pp. 1–8. IEEE (2022)
16. Kwatra, S., Torra, V.: A k-Anonymised federated learning framework with decision trees. In: Proceedings of the DPM/CBT at ESORICS 2021, pp. 106–120 (2021)
17. Bogdanova, A., Imakura, A., Sakurai, T.: DC-SHAP method for consistent explainability in privacy-preserving distributed machine learning. Hum. Centric Intell. Syst. **3**(3), 197–210 (2023)
18. Bárcena, J.L.C., Daole, M., Ducange, P., Marcelloni, F., Renda, A., Ruffini, F., Schiavo, A.: Fed-XAI: Federated Learning of Explainable Artificial Intelligence Models. In: 3rd Italian Workshop on Explainable Artificial Intelligence (XAI. it 2022) (2022)
19. Wang, G.: Interpret federated learning with shapley values. arXiv preprint arXiv:1905.04519 (2019)
20. Lundberg, S.M., Lee, S.I.: A unified approach to interpreting model predictions. Adv. Neural Inf. Process. Syst. 30 (2017)
21. Janzing, D., Minorics, L., Blöbaum, P.: Feature relevance quantification in explainable AI: A causal problem. In: International Conference on Artificial Intelligence and Statistics (ICML), pp. 2907–2916. PMLR (2020)
22. Sundararajan, M., Najmi, A.: The many Shapley values for model explanation. In: International Conference on Machine Learning, pp. 9269–9278. PMLR (2020)
23. Torra, V.: Guide to Data Privacy: Models, Technologies Solutions. Springer, Berlin (2022)
24. Domingo-Ferrer, J., Martínez-Ballesté, A., Mateo-Sanz, J.M., Sebé, F.: Efficient multivariate data-oriented microaggregation. VLDB J. **15**(4), 355–369 (2006)
25. Domingo-Ferrer, J., Torra, V.: Ordinal, continuous and heterogeneous k-anonymity through microaggregation. Data Min. Knowl. Disc. **11**(2), 195–212 (2005)
26. Monedero, D.R., Mezher, A.M., Colomé, X.C., Forné, J., Soriano, M.: Efficient k-anonymous microaggregation of multivariate numerical data via principal component analysis. Inf. Sci. **503**, 417–443 (2019)
27. Rebollo-Monedero, D., Forné, J., Soriano, M., Allepuz, J.P.: p-Probabilistic k-anonymous microaggregation for the anonymization of surveys with uncertain participation. Inf. Sci. **382**, 388–414 (2017)
28. Parra-Arnau, J., Domingo-Ferrer, J., Soria-Comas, J.: Differentially private data publishing via cross-moment microaggregation. Inf. Fusion **53**, 269–288 (2020)
29. Pastore, A., Gastpar, M.C.: Locally differentially-private randomized response for discrete distribution learning. J. Mach. Learn. Res. **22**, 1–56 (2021)

30. Dwork, C., Kenthapadi, K., McSherry, F., Mironov, I., Naor, M.: Our data, ourselves: privacy via distributed noise generation. In: Annual international conference on the theory and applications of cryptographic techniques, pp. 486–503. Springer, Berlin (2006)
31. Palia, A., Tandon, R.: Optimizing noise level for perturbing geolocation data. In: Future of Information and Communication Conference, pp. 63–73. Springer, Cham (2018)
32. Torra, V.: Data Privacy: Foundations, New Developments and the Big Data Challenge. Springer, Berlin (2017)
33. Lundberg, S.M., Erion, G., Chen, H., DeGrave, A., Prutkin, J.M., Nair, B., Katz, R., Himmelfarb, J., Bansal, N., Lee, S.I.: Explainable AI for trees: from local explanations to global understanding. arXiv preprint arXiv:1905.04610 (2019)
34. Guidotti, R., Monreale, A., Ruggieri, S., Turini, F., Giannotti, F., Pedreschi, D.: A survey of methods for explaining black box models. ACM Comput. Surv. (CSUR) **51**(5), 1–42 (2018)
35. Lundberg, S.M., Erion, G.G., Lee, S.I.: Consistent individualized feature attribution for tree ensembles. arXiv preprint arXiv:1802.03888 (2018)
36. Shapley, L.S.: A value for n-person games. Annals of Mathematics Studies (Contributions to the Theory of Games, ed. HW Kuhn and AW Tucker), pp. 307–331 (1953)
37. Myerson, R.B.: Game Theory: Analysis of Conflict. Harvard University Press, Harvard (1997)
38. Torra, V., Narukawa, Y.: Modeling Decisions: Information Fusion and Aggregation Operators. Springer, Berlin (2007)
39. Jiang, L., Torra, V.: Data protection and multi-database data-driven models. Future Internet **15**(3), 93 (2023)
40. Mortazavi, R., Jalili, S.: Fast data-oriented microaggregation algorithm for large numerical datasets. Knowl. Based Syst. **67**, 195–205 (2014)
41. Torra, V.: Microaggregation for categorical variables: a median based approach. In: International Workshop on Privacy in Statistical Databases, pp. 162–174. Springer, Berlin (2004)
42. Nin, J., Herranz, J., Torra, V.: How to group attributes in multivariate microaggregation. Int. J. Uncertain. Fuzziness Knowl. Based Syst. **16**(supp01), 121–138 (2008)
43. Rodriguez-Hoyos, A., Estrada-Jiménez, J., Rebollo-Monedero, D., Mezher, A.M., Parra-Arnau, J., Forne, J.: The fast maximum distance to average vector (F-MDAV): an algorithm for k-anonymous microaggregation in big data. Eng. Appl. Artif. Intell. **90**, 103531 (2020)
44. Molnar, C.: Interpretable machine learning. Lulu.com (2020)