Received: 27 March 2021 | Revised: 15 October 2021 | Accepted: 3 December 2021

DOI: 10.1002/cpe.6808

RESEARCH ARTICLE

WILEY

A new utility-aware anonymization model for privacy preserving data publishing

Yavuz Canbay[1] | Seref Sagiroglu[2] | Yilmaz Vural[3]

[1]Department of Computer Engineering, Sutcu Imam University, Kahramanmaras, Turkey

[2]Department of Computer Engineering, Gazi University, Ankara, Turkey

[3]Department of Computer Science, University of California, Santa Barbara, California, USA

Correspondence
Yavuz Canbay, Department of Computer Engineering, Sutcu Imam University, Kahramanmaras, Turkey.
Email: yavuzcanbay@ksu.edu.tr

Abstract

Most of data in various forms contain sensitive information about individuals and so publishing such data might violate privacy. Privacy preserving data publishing (PPDP) is an essential for publishing useful data while preserving privacy. Anonymization, which is a utility based privacy preserving approach, helps hiding the identities of data subjects and also provides data utility. Since data utility is effective on the accuracy of analysis model, new anonymization algorithms to improve data utility are always required. Mondrian is one of the near-optimal anonymization models that presents high data utility and is frequently used for PPDP. However, the upper bound problem of Mondrian causes a decrease in potential data utility. This article focuses on this problem and proposes a new utility-aware anonymization model (u-Mondrian). Experimental results have shown that u-Mondrian presents an acceptable solution to the upper bound problem, increases total data utility and presents higher data utility than Mondrian for different partitioning strategies and datasets.

KEYWORDS

anonymization, privacy preserving data publishing, utility-aware model

1 | INTRODUCTION

Digitalization in society gives the opportunity of dealing with huge amount of data about the physical world, machines, people, and so forth. Today, many institutions and companies (data curators) collect and store huge amount of data from individuals (clients, patients, users, etc.). Main purposes can be listed as achieving their tasks, improving services, extracting behavior patterns, detecting anomalies, making plans, creating policies, developing decision-making mechanisms and so on. Due to some mutual benefits or regulations, it is inevitable to publish data to public or researchers for analysis.[1–3] However, privacy concern is one of the most important issues to be addressed in data publishing.[4]

It is well known that privacy is a serious issue and is protected by law. Preserving the privacy of individuals can be defined as the right of an individual in real and cyber space in where an individual identifies his or her boundary which varies from culture to culture, country to country, or even individual to individual.[5]

In the literature, some works define data privacy as "informational self-determination" in Reference 6 and "the appropriate use of responders' information and the ability to decide what information of a responder goes where" in Reference 7. Data privacy is a major need and a requirement for privacy preserving data publishing (PPDP) and privacy preserving data mining (PPDM).[8,9] In addition, some regulations such as General Data Protection Regulation (GDPR)[10] and Personal Data Protection Law (PDPL)[11] also require data privacy.

While PPDP allows data curators or holders to share anonymized datasets, PPDM enables performing data mining operations without violating the privacy of individuals.[12,13] In PPDP, any attack on the direct release of raw data may compromise the privacy of individuals. Therefore, some approaches that eliminate privacy violations are required.

Cryptography and anonymization are two main approaches used to minimize privacy violations and preserve data privacy.[5] While the data are encrypted with algorithmically strong keys in cryptography, the identities of individuals are hidden in anonymization. If a dataset will be published to public, in order to preserve the privacy of data subjects, anonymization is the suitable solution. But, if some personal data will be shared among a community in a private manner, cryptographic solutions can be utilized.[14]

A number of privacy preserving models have been developed against privacy disclosure attacks such as record linkage, attribute linkage, probability, and table linkage. Some well-known and fundamental privacy preserving models for data publishing are briefly explained below:

- k-anonymity presents a solution for record linkage attack and ensures that a record is similar to at least $k-1$ other records over quasi-identifier attributes within the published dataset.[15]
- l-diversity provides the diversity of sensitive information in each equivalence class and it bounds the probability of successful inference.[16]
- t-closeness presents a solution for attribute linkage and probability attacks, and provides a balance between the distribution of sensitive attributes in any equivalence class and the distribution of sensitive attributes in the entire dataset.[17]
- δ-presence, which is a solution for table linkage attack, proposes to bound the probability of inferring the presence of any target record within a specified range.[18]

Differential privacy, which was proposed by Dwork,[19] is another important approach for data privacy. It provides that the result of any analysis will be almost the same if an individual participates the dataset or not. It aims to limit the effects of background information based disclosure attacks, and is successfully applied in many areas and different types of data.[20–24] Unlike the models such as k-anonymity, l-diversity, t-closeness, and δ-presence, differential privacy is not originally proposed for PPDP. It mainly allows data curators to give noisy answers to some statistical queries on private data and it is used in privacy preserving data analysis, intensively. Since our article applies k-anonymity, differential privacy is out of scope of this article.

A comprehensive list of the privacy preserving models and privacy attacks were given in Reference 25. Among the privacy preserving models, k-anonymity has lower conceptual or implementation complexity than the others and this advantage leads us to prefer k-anonymity in this work. In addition, while many studies in this field accept that each record belongs to a unique individual, some recent works accept that a dataset may include more than one record of any individual.[26–28] Due to the nature of datasets used in this article, we accepted that each record belongs to a unique individual.

The computational complexity of k-anonymity is an important issue to find an optimal solution. Previous works prove that optimal solutions for k-anonymity have an exponential increment in the solution space. As reported in the literature, k-anonymity is an NP-Hard problem and it requires near-optimal solutions.[29–33]

Mondrian[34] is a near-optimal solution for k-anonymity and a frequently used anonymization model with high utility. It supports multidimensional anonymization, and employs k-dimensional tree (KD-tree)[35–39] and multidimensional generalization.[14] Beside these advantages, Mondrian has a disadvantage which is known as the upper bound problem[40] that needs to be addressed.

In this article, we focused on the upper bound problem of Mondrian to find an acceptable solution and provided an extended version of the model presented in Reference 41. Hence, we proposed a utility-aware model applying an outlier-oriented concept to Mondrian for strict and relaxed partitioning, and obtained comprehensive results for different datasets. Detailed information about the proposed model is given in Section 4. In addition, the proposed model provides a solution for the anonymity requirements presented in some regulations such as GDPR and PDPL.

The contributions of this article are listed below:

- The upper bound problem of Mondrian has been solved,
- Data utility of Mondrian has been improved,
- A new utility-aware anonymization model for PPDP has been proposed and,
- The relation between outlier management and data utility has been represented with decision rules.

This article is organized as follows. Section 2 introduces the upper bound problem of Mondrian. In Section 3, related works and some background information are presented. The proposed u-Mondrian anonymization model is introduced and explained in Section 4. Experimental studies and detailed results are given in Section 5. Finally, the article is concluded in Section 6.

2 | THE UPPER BOUND PROBLEM OF MONDRIAN

Mondrian is a multidimensional anonymization model that employs KD-tree and multidimensional generalization. It firstly partitions data space into some smaller partitions by employing KD-tree and then anonymizes each of these partitions using multidimensional generalization.

Mondrian uses strict and relaxed partitioning strategies to partition data space. Strict partitioning enables to obtain non-overlapping multidimensional regions while relaxed partitioning allows to obtain potentially overlapping regions. In Reference 34, some detailed information for strict and relaxed partitioning are given.

Mondrian has lower and upper bounds that define minimum and maximum sizes of equivalence classes for strict and relaxed partitioning strategies. The lower bounds are accepted as k for the both of partitioning strategies, and the upper bounds for strict and relaxed partitioning are accepted as $2d(k-1)+t$ and $2k-1$, respectively. Note that d, k, and t represent dimension, anonymity level, and maximum number of copies of any distinct point in a set, respectively. A detailed information about these terms and partitioning strategies are explained in Reference 34. While the main purpose of these bounds is limiting the sizes of equivalence classes, they also limit potential data utility. In the literature, this problem is known as the upper bound problem.[40]

In order to describe the upper bound problem more precisely, we provided some definitions and graphical illustrations below.

Mapper function: Let R be a data space, V be the set of all data points in R, and $\{R_1, R_2, \dots, R_n\}$ be the sub-spaces of R; a mapper function m maps each data point in V into its projection in any sub-space R_i in $\{R_1, R_2, \dots, R_n\}$.

An illustration of mapper function is presented in Figure 1. The points in the space R are mapped into their projections in the sub-spaces of R_1 and R_2, as shown on the right of the figure.

Partition: Let R_i be any sub-space in $\{R_1, R_2, \dots, R_n\}$; the set of data points in R_i is called as a partition P_i and $P = \{P_1, P_2, \dots, P_n\}$.

Bound: Let P_i be a partition and $P_i \in \{P_1, P_2, \dots, P_n\}$; total number of data points in P_i is called as the bound b_i of P_i.

In Figure 2A, an illustration for the bound is presented. The bounds of P_1 and P_2 partitions are obtained as 3 and 2, respectively.

Upper bound: Let b_i be the bound of P_i, $B = \{b_1, b_2, \dots, b_n\}$ be the set of bounds of $\{P_1, P_2, \dots, P_n\}$ and $P = \{P_1, P_2, \dots, P_n\}$; the maximum b_i in B is accepted as the upper bound up_b of P.

In Figure 2B, an illustration for the upper bound is shown. It can be seen that the upper bound for this partitioning is obtained as 3, which is observed in P_1.

The upper bound problem: Let R be a data space, P be the set of partitions, $P = \{P_1, P_2, \dots, P_n\}$, up_b be the upper bound of P, and k be the parameter of k-anonymity; if $k < up\_b$, then it is called as the upper bound problem.

While the situation of $k = up\_b$ is ideal to obtain high data utility, the upper bound problem occurs when $k < up\_b$. This problem causes a decrease in potential data utility and hence some solutions considering this problem are required.

In the following, illustrations for the upper bound problem and a simple solution are presented. Figure 3A indicates an example of 2-dimensional dataset. If a mapper function is applied to this dataset, a possible partitioning for $k = 3$ can be obtained as shown in Figure 3B. Note that P_1 and P_2 represent the partitions, and the lower and upper bounds for these partitions are 3 and 5, respectively. However, P_1 and P_2 include 5 and 4 data

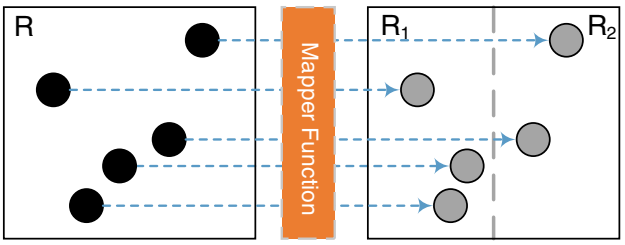

FIGURE 1 An illustration of mapper function

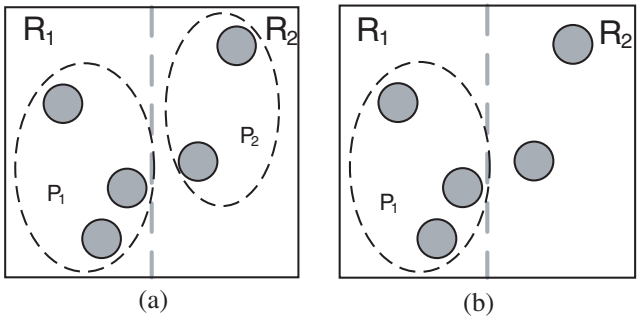

FIGURE 2 (A) An illustration for bound, (B) an illustration for upper bound

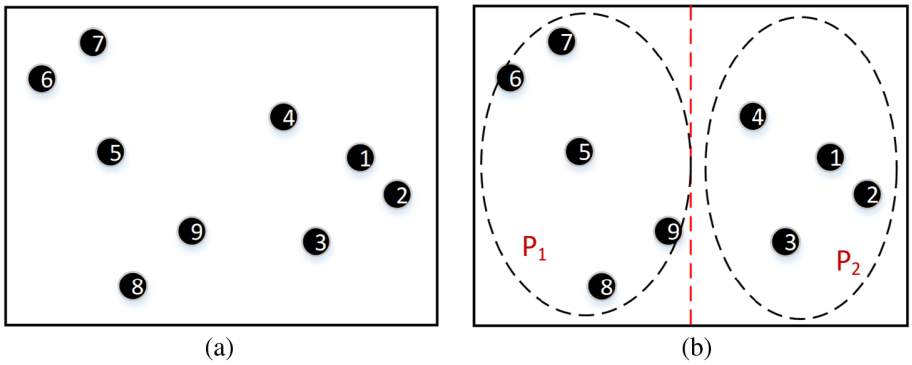

FIGURE 3 (A) 2D dataset, (B) partitioning of dataset ($k = 3$)

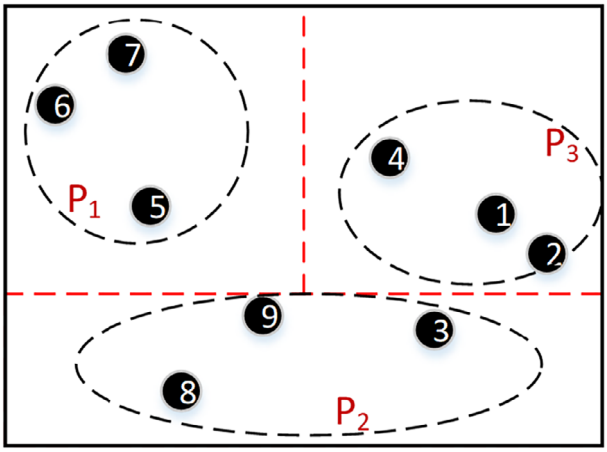

FIGURE 4 Utility-aware partitioning of dataset ($k = 3$)

points, respectively. In this case, it is accepted that any two-points in P_1 and any one-point in P_2 do not present any utility and they also cause a decrease in total data utility. Finally, the upper bound problem occurs.

If the bounds of P_1 and P_2 are examined deeply, it can be obviously seen that these bounds can be optimized to gain more utility. A possible solution for the upper bound problem occurred in Figure 3B is indicated in Figure 4. Other solutions might be applied for illustration, but for the sake of simplicity, the scenario given in Figure 4 was selected to show a utility-aware partitioning that enhances total data utility.

By using discernibility metric (DM), which is detailed in Section 3.3, data utilities of the approaches presented in Figures 3B and 4 can be measured as $DM_1 = 5^2 + 4^2 = 41$ and $DM_2 = 3^2 + 3^2 + 3^2 = 27$, respectively. Because of $DM_1 > DM_2$, it can be clearly seen that utility-aware partitioning presents higher utility than the other one. According to the results, it can be clearly observed that a solution considering the upper bound problem presents higher data utility than classical solutions.

As can be seen from the above explanations and illustrations, the upper bound problem is an important issue for data utility that needs to be addressed.

3 | RELATED WORKS AND PRELIMINARIES

In this section, a brief for related works, and some preliminary information about outliers, outlier management and data utility in anonymization are given.

3.1 | Related works

In the literature, there exist a number of studies enhancing data utility of Mondrian in a problem-specific manner and proposing some variations of Mondrian. A comparison of these works is given in Table 1.

TABLE 1 Comparisons for Mondrian based studies in the literature

Model	Scope	Enhancement target	Enhancement approach	Partitioning strategy	Guarantee k	Consider data distribution	Consider upper bound problem
Liu's model[42]	PPDP	Attributes	Attribute weighting	N/A	No	No	No
Enhanced Mondrian[43]	PPDP	Attributes	Missing value handling	N/A	No	No	No
Nergiz's model[44]	PPDP	Attributes	Data relocation	N/A	No	No	No
InfoGain Mondrian[45]	PPDM	Splitting approach	Entropy	N/A	No	Yes	No
LSD Mondrian[45]	PPDM	Splitting approach	Least square deviance	N/A	No	No	No
Selection Mondrian[45]	PPDM	Splitting approach	Imprecision	N/A	No	No	No
Canbay's model[46]	PPDP	Outlier	Density based splitting	Strict	No	Yes	No
Canbay's model[41]	PPDP	Outlier	Density and nearest neighbor based partitioning	Strict	Yes	Yes	No
Tang's model[40]	PPDP	Upper bound	Flexible splitting	Relaxed	No	No	Yes
u-Mondrian (present model)	PPDP	Upper bound	Density and nearest neighbor based partitioning	Strict and relaxed	Yes	Yes	Yes

From Table 1, it can be understood that only Tang et al.[40] directly studied the upper bound problem of Mondrian and developed a solution. They used Adult dataset to test the proposed model and DM metric to measure data utility. In the experimental studies, they compared their model with Mondrian and Bottom-Up algorithms, and finally achieved acceptable results.

If we compare u-Mondrian, proposed in this article, with the model of Tang et al.,[40] some comments can be given as below:

- The model of Tang et al. is only applicable for relaxed partitioning, but u-Mondrian can be applied on both strict and relaxed partitioning,
- The model of Tang et al. focuses on just improving the size of each equivalence class and does not consider the distribution of data, but u-Mondrian applies a density and nearest neighborhood based hybrid approach to improve the sizes of equivalence classes and,
- The upper bound of u-Mondrian is k, which is smaller than the upper bound of the model of Tang et al. (if the size of the last remaining partition is bigger than k, we ignore it).

In addition, Canbay et al.[41] focused on outliers to enhance total data utility and proposed a new utility optimization function to measure the effects of outliers in dataset. Since our article extends the anonymization model proposed in Reference 41, some significant differences between the study of Canbay et al.[41] and the present study can be given as in Table 2.

3.2 | Outliers and outlier management in anonymization

In anonymization, outliers are defined as the data group that decreases total data utility and the data group except outliers is accepted as normal dataset.[47–53] In addition to these general definitions, some problem-specific definitions are also presented in the literature. The definitions available in the literature are given below:

- The data group whose value is bigger than a threshold value.[54]
- The data group whose distance to the others is bigger than a threshold value.[51–53]
- The data group which is completely suppressed.[47–49]
- The data group that violates anonymity standard.[55]
- The data group that violates the privacy criteria.[56]
- The data group whose density is lower than the others.[46]
- The data group which is the complement of k closest data points.[41]

TABLE 2 Comparisons for the study of Canbay et al.[41] and the present study

Canbay et al.[41]	Present study
Introduces an outlier-oriented version of Mondrian and proposes UO function to measure data utility in outlier concept	Proposes an extended version of the anonymization model introduced by Canbay et al.[41]
Directly focuses on outliers	Focuses on the upper bound problem and proposes a solution by applying an outlier concept
Supports only strict partitioning	Supports both strict and relaxed partitioning
Uses iteration number for stopping criteria	Uses decision rules for stopping criteria
Evaluates the results for only the number of outliers	Evaluates the results by considering both the sizes of equivalence classes and proximity of dataset
Uses UO function to measure data utility	Uses DM, AECS, and GCP metrics to measure data utility for detailed results
Uses only Adult dataset for experiment	Uses both Adult and Diabetes datasets for detailed experiments
Presents only a simple block diagram and general steps of the proposed model	Provides detailed theoretical and mathematical information about the proposed model
Presents a brief information about the outliers and outlier determination process	Provides detailed information about the upper bound problem and outlier determination process
Presents a simple graphic for result	Presents detailed graphics for comprehensive results
Does not provide any graphical illustrations	Provides graphical examples to explain the related situations (e.g., upper bound problem, outlier detection, etc.)

Outliers are considered to reduce total data utility and they should be managed in the anonymization process. The approaches for outlier management existing in literature are presented below:

(i) Detecting outliers before anonymization and removing them from dataset.[54]
(ii) Detecting outliers after anonymization and removing global outliers from dataset and reusing local outliers to gain information.[51–53]
(iii) Detecting outliers after anonymization, reusing the convenient ones to gain information and then removing the final remaining.[47–49]
(iv) Detecting outliers after anonymization and modifying them.[55]
(v) Detecting outliers after anonymization and removing them from dataset.[56]
(vi) Detecting outliers after anonymization and reusing all of them to gain information.[46]
(vii) Detecting outliers before anonymization and reusing them to gain information.[41]

While the approaches (i) and (v) aim only to preserve data privacy, the approaches (ii), (iii), (iv), (vi), and (vii) focus on gaining utility from outliers while preserving data privacy. In addition, the approach (iv) modifies original values of the outliers which this case negatively affects the truthfulness of data. Finally, we adopted the approach of (vii) in this study.

3.3 | Data utility in anonymization

Data utility is defined as the similarity between anonymized data and original data.[14] In the literature, it is accepted that an original dataset presents higher data utility than the anonymized version of the same dataset. Because, anonymization transforms original data into a new form and this transformation causes some information loss. Higher data utility presents more accurate analysis model.

In the context of anonymization, a formal definition for data utility can be given as below.

Data utility: Let V be an original dataset, V_A be an anonymous version of V, *Similarity*( ) is a function that measures the similarity between two datasets, for $u \geq 0$ and $u = Similarity\,(V, V_A)$; u presents the utility of V_A.

Information metrics existing in the literature can be used as *Similarity*( ) function. In this article, we utilized DM,[32] average equivalence class size (AECS),[34] and generalized certainty penalty (GCP)[57] metrics which are directly related to Mondrian and its extensions. Let T be a dataset, EC be the set of equivalence classes, qid be quasi-identifier, k be the parameter of k-anonymity, G be any equivalence class, NCP be the metric of normalized certainty penalty, N be the number of records, and d be the dimensionality. DM, AECS, and GCP metrics are calculated according to Equations (1)–(3), respectively.

$$\mathrm{DM}(T) = \sum_{qid_i} \left| T\left[qid_i\right] \right|^2, \tag{1}$$

$$\text{AECS}(T) = \frac{|T|}{|EC| * k}, \tag{2}$$

$$\text{GCP}(T) = \frac{\sum_{G \in T} |T| * NCP(G)}{d * N}. \tag{3}$$

DM, AECS, and GCP metrics measure the size of equivalence classes, the size of average equivalence classes, and the perimeter of each equivalence class, respectively. Although DM and AECS are considered sufficient to evaluate the data utility in a number of studies, GCP is also used to demonstrate the main contribution of this article.

4 | THE PROPOSED UTILITY-AWARE MODEL: U-MONDRIAN

In order to explain the proposed anonymization model, the definitions for outlier and normal data, decision rules and outlier determination approach are first presented and the proposed model is then introduced in the following subsections.

4.1 | Defining outlier and normal data for u-Mondrian

u-Mondrian is an outlier-oriented anonymization model that enhances total data utility by gaining information from outliers. We adopted and revised the definitions for outlier and normal data from Reference 41 and presented below:

Normal data: Let P_i be a partition in $P = \{P_1, P_2, \dots, P_n\}$, up_b be the upper bound of P and for $k \leq |P_i| \leq up\_b$; the group of k closest data points (a subset of data points with the size of k that are closer to each other than any other subsets) in P_i is accepted as normal data.

Outlier data: The data group except normal data in P_i is accepted as outlier data.

Remember that Mondrian accepts the upper bound up_b as $2k - 1$ for relaxed partitioning and $2d(k-1) + t$ for strict partitioning.

4.2 | Creating decision rules

To evaluate the relation between outlier management and data utility, three decision rules are created by using DM, AECS, and GCP metrics. These decision rules are used to determine if outlier management increases total data utility or not. It is accepted that if the results of these rules equal to 1, it means that outlier management increases total data utility, but 0 means no increment. These rules are used as stopping criteria for the iteration of outlier determination phase in the proposed model. Let V be a dataset, ES be the number of equivalence classes, U be the value of information metric without outlier management, U' be the value of information metric with outlier management, S be the maximum iteration number of outlier detection phase, N_i be the normal dataset in the ith iteration, and O_S be the outlier dataset in the Sth iteration.

The decision rules created in this work are presented below.

Data utilities in terms of DM metric are calculated as

$$U = DM(V), \text{ without outlier management,}$$

$$U' = \sum_{i=1}^{S} DM(N_i) + DM(O_S), \text{ with outlier management.}$$

Decision rule A is created as in Equation (4):

$$A = \begin{cases} 1, & \text{if } U' < U, \\ 0, & \text{otherwise.} \end{cases} \tag{4}$$

Data utilities in terms of AECS metric are calculated as

$$U = \frac{|V|}{ES(V)}, \text{ without outlier management,}$$

$$U' = \frac{|V|}{\sum_{i=1}^{S} ES(N_i) + ES(O_S)}, \text{ with outlier management.}$$

Decision rule B is created as in Equation (5):

$$B = \begin{cases} 1, & \text{if } U' < U, \\ 0, & \text{otherwise.} \end{cases} \tag{5}$$

Data utilities in terms of GCP metric are calculated as

$$U = GCP(V), \quad \text{without outlier management,}$$
$$U' = \sum_{i=1}^{S} GCP(N_i) + GCP(O_S), \quad \text{with outlier management.}$$

Decision rule C is finally created as in Equation (6)

$$C = \begin{cases} 1, & \text{if } U' < U, \\ 0, & \text{otherwise.} \end{cases} \tag{6}$$

It should be emphasized that for each of these decision rules:

- "1" represents an increment in the data utility.
- "0" represents no increment or remaining the same in the data utility.

For example, if the output of decision rule A is 1, it means that an increase occurs in total data utility.

4.3 | Determination of outlier and normal data

Considering the definitions for outlier and normal data groups as presented in Section 4.1, an approach that finds k closest data points in a dataset is required. Due to finding k closest data points requires a combinatorial searching space, a near-optimal hybrid approach presented in Reference 41 was borrowed. Note that determination of outlier and normal data groups is performed for each partition P_i in $\{P_1, P_2, \ldots, P_n\}$. Some definitions for the hybrid approach are given below.

Density based reference point: Let P_i be a partition in $P = \{P_1, P_2, \ldots, P_n\}$, r be any point in P_i, up_b be the upper bound of P and *density_scoring*( ) be a density scoring function. The data point in P_i having maximum *density_scoring*( ) is accepted as a density based reference point, which is given in Equation (7)

$$ref = \begin{cases} r \text{ with } max\left(density\_scoring(P_i)\right), & \text{if } up\_b > k, \\ \text{None}, & \text{otherwise.} \end{cases} \tag{7}$$

$k-1$ nearest neighbors: Let *ref* be a density based reference point, $p \in P_i$, and $dist(ref, p)$ be a distance function. $k-1$ data points in P_i with $min(dist(ref, p))$ is accepted as $k-1$ nearest neighbors (NN^{k-1}) of *ref*, which is given in Equation (8)

$$NN^{k-1} = \begin{cases} \text{for all } p \in \left(P_i - \{ref\}\right) : \min(dist(ref, p)), & \text{if } up\_b > k, \\ \text{None}, & \text{otherwise.} \end{cases} \tag{8}$$

By using these definitions, the main workflow of the hybrid approach is explained below.

It is well-known that a dense region includes closer points than the other regions. Hence, a density-based approach enables to obtain closer data points in a dataset. In the proposed hybrid approach, local outlier factor (LOF)[58] is used as *density_scoring*( ) function to score each data point based on their densities. LOF algorithm produces density scores inversely proportional to the densities of each data point. It means that the data point with a lower LOF score has a higher density than the other points. Hence, the point with maximum density is already the point with minimum LOF score. After scoring the densities of all points, a point is determined as a reference point *ref*, which has the maximum density or in another words minimum LOF score as presented in Equation (7).

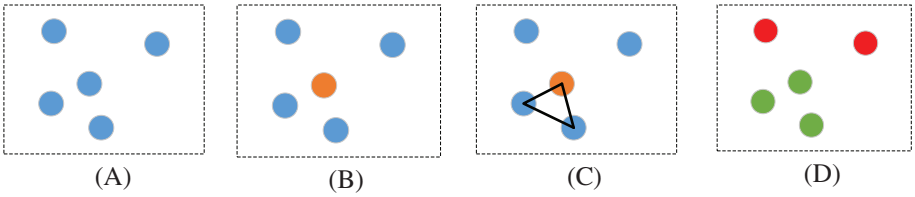

FIGURE 5 (A) A candidate partition of 2D dataset, (B) determination of a reference point (orange colored point), (C) determination of $k-1$ nearest neighbor of the reference point, (D) determination of normal and outlier data groups (represented with green and red colors, respectively)

After the determination of the reference point *ref*, $k-1$ nearest neighbors of *ref* are found by using the function of $dist(ref, p)$ which measures the distance between *ref* and *p*. $k-1$ nearest neighbors (NN^{k-1}) of *ref* are then searched and k closest data points (including *ref* and NN^{k-1}) are obtained.

Then, for each partition P_i in $\{P_1, P_2, \dots, P_n\}$, normal and outlier data groups are determined from Equations (9) and (10):

$$Normal_i = \left\{ \{ref_i\} \cup NN_i^{k-1} \right\}, \tag{9}$$

$$Outlier_i = \{P_i - Normal_i\}. \tag{10}$$

Finally, *NormalSet* and *OutlierSet* for *P* are determined from Equations (11) and (12):

$$NormalSet = \bigcup_{i=1}^{n} Normal_i, \tag{11}$$

$$OutlierSet = \bigcup_{i=1}^{n} Outlier_i. \tag{12}$$

In order to explain the hybrid approach in detail, an illustration is presented in Figure 5, for $k = 3$. In Figure 5A, a candidate partition of a dataset is given. A reference point indicated with orange color is determined by using LOF in Figure 5B. $k-1$ nearest neighbors of the reference point are found in Figure 5C. Finally, normal and outlier data groups are determined in Figure 5D and represented with green and red colors, respectively. Note that this illustration is constructed based on only one partition, but in the proposed model this approach is applied to all partitions.

4.4 | u-Mondrian anonymization model

Since our model is based on Mondrian, in order to give a clear understanding, the approach in Mondrian is first explained and u-Mondrian is then introduced. Mondrian consists of two phases, which are KD-tree phase and Generalization phase. In KD-tree phase, the data are partitioned into smaller partitions under some constraints and these partitions are then generalized in Generalization phase.

Let V be a D-dimensional dataset. A KD-tree is first applied over V and some partitions $\{P_1, P_2, \dots, P_m\}$ are then obtained. Each partition P_i in $\{P_1, P_2, \dots, P_m\}$ is generalized with function of $gen()$ and finally anonymized dataset V^G is achieved. The algorithm for Mondrian is given in Algorithm 1.

Algorithm 1. The algorithm of Mondrian model

Input: dataset $V = \{v_1, v_2, \dots, v_N\}$; $v_N \in R^D$, k parameter of k-anonymity

Initialize: $data = V$

1. Apply KD-tree

Iterate:

- Choose dimension; $dim = \{d \in D,\ \ d = max_{range}(data, D)\}$
- Calculate frequencies on *dim*; $f_{dim} = \{\{f_1, f_2, \dots, f_M\} = calc_f(data, dim)\}$
- Find the median of f_{dim}; $\mu = \{m \in data_{dim}, m = med(f_{dim})\}$
- Split data; $lhs = \{l \in data,\ \ l_{dim} \le \mu\}$
  $rhs = \{l \in data,\ \ l_{dim} > \mu\}$
- Repeat for; $data = lhs$

$data = rhs$, until the stopping criteria is met

Return:

- Send the set of partitions $\{P_1, P_2, \ldots, P_m\}$ to Phase 2

2. Generalize each P_i in $\{P_1, P_2, \ldots, P_m\}$

Iterate:

- Generalize; $P_i^G = gen(P_i)$
- Repeat for each P_i

Return:

- Return the set of anonymized partitions $V^G = \{P_1^G, P_2^G, \ldots, P_m^G\}$

Output: anonymized dataset V^G

The proposed u-Mondrian model is an anonymization model that increase total data utility by applying an outlier-oriented concept to Mondrian. In order to apply this concept to Mondrian, we added a new phase, which is called as "outlier determination phase," between KD-tree phase and Generalization phase. The algorithm of u-Mondrian is introduced in Algorithm 2.

Let V be a D dimensional dataset. KD-tree is first applied over V and some partitions $\{P_1, P_2, \ldots, P_m\}$ are then obtained. For each P_i in $\{P_1, P_2, \ldots, P_m\}$, outlier and normal data groups are determined and stored in *OutlierSet* and *NormalSet*, respectively. According to the evaluation, either *NormalSet* is generalized or *OutlierSet* is accepted as the new dataset to be examined for the next iteration. This iteration continues until the result of the decision rule equals to 1.

Algorithm 2. The algorithm of u-Mondrian model

Input: dataset $V = \{v_1, v_2, \ldots, v_N\}$; $v_N \in R^D$, k parameter of k-anonymity

Initialize: $data = V$

1. Apply KD-tree

Iterate:

- Choose dimension; $dim = \{d \in D, \quad d = max_{range}(data, D)\}$
- Calculate frequencies on dim; $f_{dim} = \{\{f_1, f_2, \ldots, f_M\} = calc_f(dim, data)\}$
- Find median of f_{dim}; $\mu = \{m \in data_{dim}, m = med(f_{dim})\}$
- Split data; $lhs = \{l \in data, \quad l_{dim} \leq \mu\}$
  $rhs = \{l \in data, \quad l_{dim} > \mu\}$
- Repeat for; $data = lhs$
  $data = rhs$, until the stopping criteria is met

Return:

- Send the set of partitions $\{P_1, P_2, \ldots, P_n\}$ to Phase 2

2. Determining outliers in $\{P_1, P_2, \ldots, P_n\}$

Iterate:

- Score each point $p \in P_i$; $scores = \{\{s_1, s_2, \ldots, s_R\} = density\_scoring(P_i)\}$
- Determine ref. point $ref_i \in P_i$; $ref_i = r$ *with* $max(scores)$
- Find NN^{k-1} of ref_i, $x \in P_i$; $NN_i^{k-1} = \left\{x : (k-1) = argmin\|x - ref_i\|^2\right\}$
- Determine normal data group; $Normal_i = \left\{ref_i \cup NN_i^{k-1}\right\}$
- Determine outlier data group; $Outlier_i = \{P_i - Normal_i\}$
- Store $Normal_i$ in *NormalSet*
- Store $Outlier_i$ in *OutlierSet*
- Repeat for each P_i

Evaluate:

- If $DecisionRule = 1$; $data = OutlierSet$ and go to Phase 1
  else go to Phase 3

3. Generalize each partition NS_i in *NormalSet*

Iterate:

- Generalize; $NS_i^G = gen\left(NS_i\right)$
- Repeat for each NS_i

Return:

- Return the set of anonymized partitions $V^G = \left\{NS_1^G, NS_2^G, \ldots, NS_n^G\right\}$

Output: anonymized dataset V^G

The workflow of u-Mondrian is presented in Figure 6. This figure is constructed according to the assumptions given below:

1. 3-Anonymity is applied.
2. Each square box represents a data point.
3. All colored boxes except black represent normal data groups.
4. Black colored boxes represent outlier data groups.
5. Each color except black represents the proximity among data points, it means that the same-colored data points are closer to each other than the others.
6. Each group including a number of boxes with same color represents partition.
7. A sharp transformation from one color to another demonstrates high level generalization (e.g., in Figure 6, orange color in Partition 1 transform to black).
8. A soft transformation from one color to another demonstrates low level generalization (e.g., in Figure 6, the first blue partition in *NormalSet* transforms to the light blue partition in EqClass 1).

The steps of u-Mondrian illustrated in Figure 6 are given below:

1. Chose a dimension on the original dataset.
2. Calculate the frequencies of the data points on this dimension.
3. Find the median of the frequencies.
4. Split the dataset into two subsets (LHS and RHS) based on the value of median.
5. Check if allowable cutting exits.
6. If yes go to Step 1.
7. Obtain partitions satisfying lower and upper bounds.
8. Detect normal data group and outlier data group for each partition.
9. Separate dataset as normal and outliers, and store in *NormalSet* and *OutlierSet*, respectively.
10. Check if the value of decision rule is 1.
11. If yes, store *NormalSet*, accept *OutlierSet* as the new dataset and go to Step 1.
12. If no, generalize each partition in *NormalSet* and obtain equivalence classes.
13. Collect and combine all equivalence classes, and then obtain anonymous dataset.

The main difference between Mondrian and u-Mondrian is the outlier detection phase. This phase aims to improve data utility by applying an outlier-oriented concept to Mondrian. In addition, u-Mondrian iterates outlier detection and KD-tree phases to manage outliers.

Outliers are managed according to an outlier management approach adapted and modified from Reference 49. In our modified approach, outlier management is performed by partitioning of outlier dataset into two new subsets iteratively and gaining information from all subsets without removing any of them. Partitioning of outlier dataset iteratively contributes to improve data utility by applying a better separation. For example, Figure 7 shows how u-Mondrian enhances data utility. A sample 2D dataset is presented in Figure 7A. KD-tree is applied to this dataset and two partitions are obtained as presented in Figure 7B. Density based reference point determination is performed in Figure 7C. $k-1$ nearest neighbors of these reference points are found in Figure 7D. In each partition, k closest data points (including reference point and its $k-1$ nearest neighbors) are found and accepted as normal data group (blue colored points), then stored in *NormalSet*; the others are accepted as outlier data group (red colored points) and stored in *OutlierSet* in Figure 7E. In the next iteration, *OutlierSet* is accepted as the new dataset to be examined and the steps presented above are then repeated until the output of the decision rule is 1. Finally, the resulting *NormalSet* is obtained as shown in Figure 7F. After the generalization of the partitions in *NormalSet*, anonymized dataset is obtained.

Partitioning process of Mondrian is only composed from KD-tree phase which can be illustrated as in Figure 7B. But, u-Mondrian employs a utility-aware partitioning strategy which is presented in Figure 7B–F. Hence, data utility comparison for Mondrian and u-Mondrian can be performed by considering all sub-figures in Figure 7. By using DM metric, the data utility of Mondrian can be calculated as $DM_{\text{Mondrian}} = 5^2 + 4^2 = 41$;

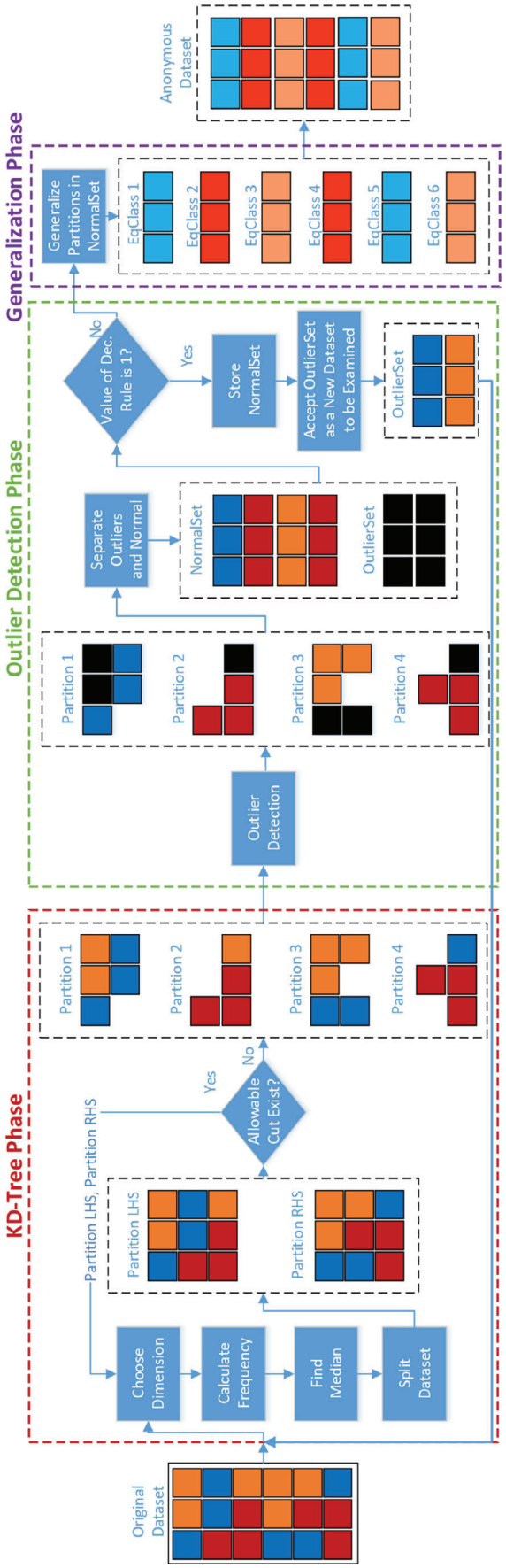

FIGURE 6 The workflow of u-Mondrian

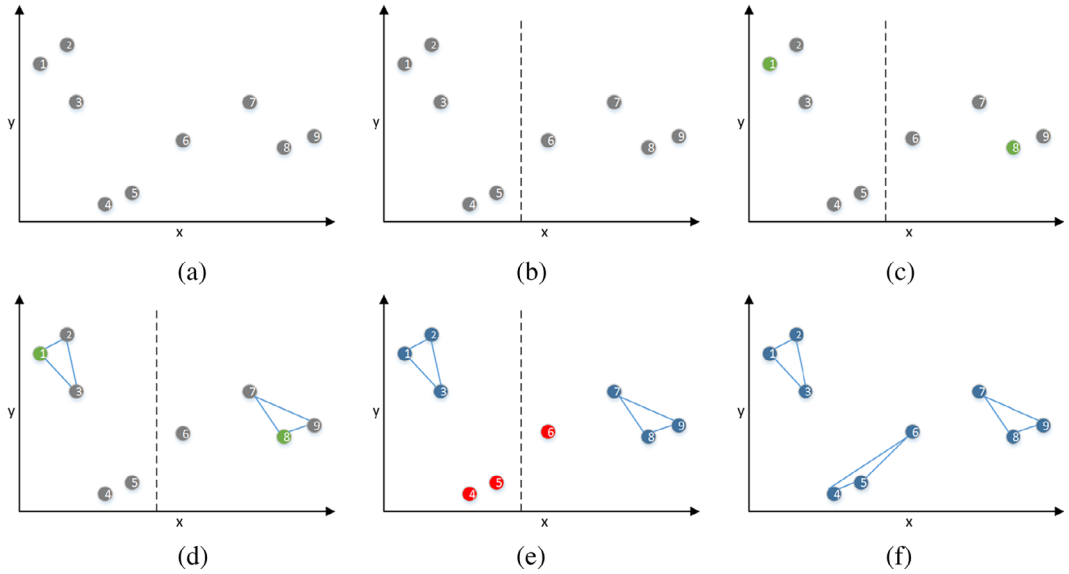

FIGURE 7 (A) 2D dataset, (B) a possible KD-tree partitioning, (C) determination of reference points (green colored points) in each partition, (D) determination of $k-1$ nearest neighbors of the reference points, (E) composing normal data group (blue colored points) and outliers (red colored points) in each partition, (F) the resulting *NormalSet* ($k=3$)

and the data utility of u-Mondrian can be calculated as $DM_{u\text{-}Mondrian} = 3^2 + 3^2 + 3^2 = 27$. Because of $DM_{u\_Mondrian} < DM_{Mondrian}$, it can be clearly seen that u-Mondrian presents higher data utility than Mondrian.

Finally, the computational complexity of u-Mondrian is also presented below. Let M be the iteration number, N be the number of elements in dataset and k be the parameter of k-anonymity. Computational complexity of u-Mondrian is obtained as $O\left(M * \left(N \log N + \frac{N}{k} N^2\right)\right)$. It is known that Mondrian has the complexity of $O(N \log N)$. Although u-Mondrian presents higher complexity than Mondrian, it can be ignored due to the main purpose of this study is only to enhance data utility capability of Mondrian.

5 | EXPERIMENTAL STUDIES AND RESULTS

In the experimental studies, Adult dataset[59] is used to test the proposed model due to being a defacto-standard and frequently used dataset in anonymization studies. Adult dataset includes 48,842 records and 18,680 of them are unfortunately incomplete. By removing the incomplete records, 30,162 records were employed in the experiments. Since the proposed model supports only numerical values, the quasi-identifiers are determined as age, final_weight, capital_gain, capital_loss, and hours_per_week throughout this study.

In addition, in order to verify the proposed model with another dataset, Diabetes dataset[59] is also employed. It includes 101,766 records and 55 features with categorical and numerical values. Since we consider only numerical values, the numerical features of mean_age, num_lab_procedures, num_procedures, num_medications, number_outpatient, number_emergency, number_impatient, and number_diagnosis are determined as quasi-identifiers for the related experiments.

In the experiments, the proposed u-Mondrian model is tested and then compared to Mondrian for different k values, metrics, partitioning strategies and datasets. Since we use decision rules introduced in previous sections, DM, AECS, and GCP metrics are employed to measure the utility of the models. In addition, the number of equivalence classes (ES) is also measured to evaluate the utility-aware partitioning strategy. Since k values are generally chosen as small values, we preferred to use the values as presented in Table 3.

The results of u-Mondrian and Mondrian for strict and relaxed partitioning, and in addition for Adult and Diabetes datasets, are presented in Table 3. For each k value, data utilities in terms of ES, DM, AECS, and GCP are observed. It can be clearly seen from the results that u-Mondrian produces more ES and also presents less DM, AECS, and GCP values than Mondrian for both datasets and partitioning strategies. (In Table 3, S represents strategy and M represents model.)

Figures 8 and 9 represent the results of ES, DM, AECS, and GCP of u-Mondrian and Mondrian for Adult and Diabetes datasets, respectively. Note that while DM, AECS, and GCP are inversely proportional to data utility, ES is vice versa. The overall results show that the proposed u-Mondrian model presents higher data utility than Mondrian.

Table 4 represents data utility enhancement ratios of u-Mondrian against Mondrian for both partitioning strategies and datasets. As it can be clearly seen that the results validate the success of the proposed u-Mondrian model, demonstrate increments drastically in data utility and provide an acceptable solution for PPDP with higher data utility.

TABLE 3 The overall results of u-Mondrian and Mondrian

			Adult dataset				Diabetes dataset			
S.	**M.**	***k***	**ES**	**DM**	**AECS**	**GCP**	**ES**	**DM**	**AECS**	**GCP**
Strict	Mondrian	5	4553	207,996	6.62	3952.62	12,791	1,068,636	7.95	48,485.08
		10	2222	425,890	13.57	6141.81	6683	1,752,930	15.22	64,090.00
		20	1105	856,562	27.29	8582.58	3434	3,254,622	29.63	83,088.25
		30	749	1,267,990	40.26	10,367.28	2314	4,778,192	43.97	95,088.58
		40	550	1,725,228	54.84	11,586.60	1725	6,393,690	58.99	104,847.97
		50	438	2,153,108	68.86	12,664.65	1369	8,037,076	74.33	113,145.87
		60	376	2,514,582	80.21	13,721.26	1148	9,518,942	88.64	119,656.23
		70	316	3,010,522	95.44	14,689.80	986	10,113,834	103.21	125,731.62
	u-Mondrian	5	6029	150,725	5.11	2636.11	20,353	502,625	5.00	29,901.53
		10	3012	301,200	10.00	4173.57	10,176	1,015,500	10.00	44,138.30
		20	1507	602,800	20.33	6096.85	5088	2,030,400	20.02	59,553.38
		30	1005	904,500	32.40	7732.23	3392	3,046,500	30.05	69,425.36
		40	754	1,206,400	40.10	8885.01	2544	4,067,200	40.07	76,710.43
		50	602	1,505,000	52.00	9247.82	2035	5,082,500	50.26	83,005.60
		60	502	1,807,200	61.31	10,384.18	1694	6,105,600	60.00	87,044.38
		70	430	2,107,000	76.20	10,951.68	1451	7,109,900	70.00	91,561.05
Relaxed	Mondrian	5	4096	223,054	7.36	6266.90	16,384	634,830	6.21	44,251.69
		10	2048	444,618	14.72	9561.73	8192	1,266,198	12.42	63,888.38
		20	1024	888,678	29.45	13,938.86	4096	2,528,934	24.84	85,786.45
		30	512	1,776,890	58.91	19,792.52	2048	5,057,234	49.69	111,330.67
		40	512	1,776,890	58.91	19,792.52	2048	5,057,234	49.69	111,330.67
		50	512	1,776,890	58.91	19,792.52	1414	8,163,834	71.97	128,781.89
		60	256	3,553,734	117.82	27,497.32	1024	10,113,834	99.38	140,764.48
		70	256	3,553,734	117.82	27,497.32	1024	10,113,834	99.38	140,764.48
	u-Mondrian	5	6026	150,650	5.05	4575.39	20,334	508,350	5.01	39,393.76
		10	3008	300,800	10.04	7276.47	10,176	1,016,600	10.04	59,528.66
		20	1504	601,600	20.08	10,413.83	5088	2,022,400	20.01	77,304.74
		30	992	892,800	30.15	10,984.57	3392	3,052,800	30.02	91,045.88
		40	752	1,203,200	40.10	14,076.83	2544	4,044,800	40.02	98,504.34
		50	602	1,505,000	52.00	16,620.57	2035	5,075,000	50.25	111,447.68
		60	496	1,785,600	61.31	14,657.89	1696	6,105,600	60.00	113,718.31
		70	428	2,097,200	76.20	16,776.98	1448	7,095,200	70.00	117,812.94

Considering the improvements given in Table 4 and the results in Figures 8 and 9, overall evaluations for Mondrian and u-Mondrian is presented below under different subtitles as:

- Overall evaluation for the results of strict and relaxed partitioning strategies;
  - The values of DM, AECS, and ES for Mondrian with strict partitioning are rather different from the values of Mondrian with relaxed partitioning. This means that Mondrian with different partitioning strategies is not stable. But, DM, AECS, and ES values of u-Mondrian for both partitioning strategies are almost the same which means that u-Mondrian presents a stable solution for different partitioning strategies.

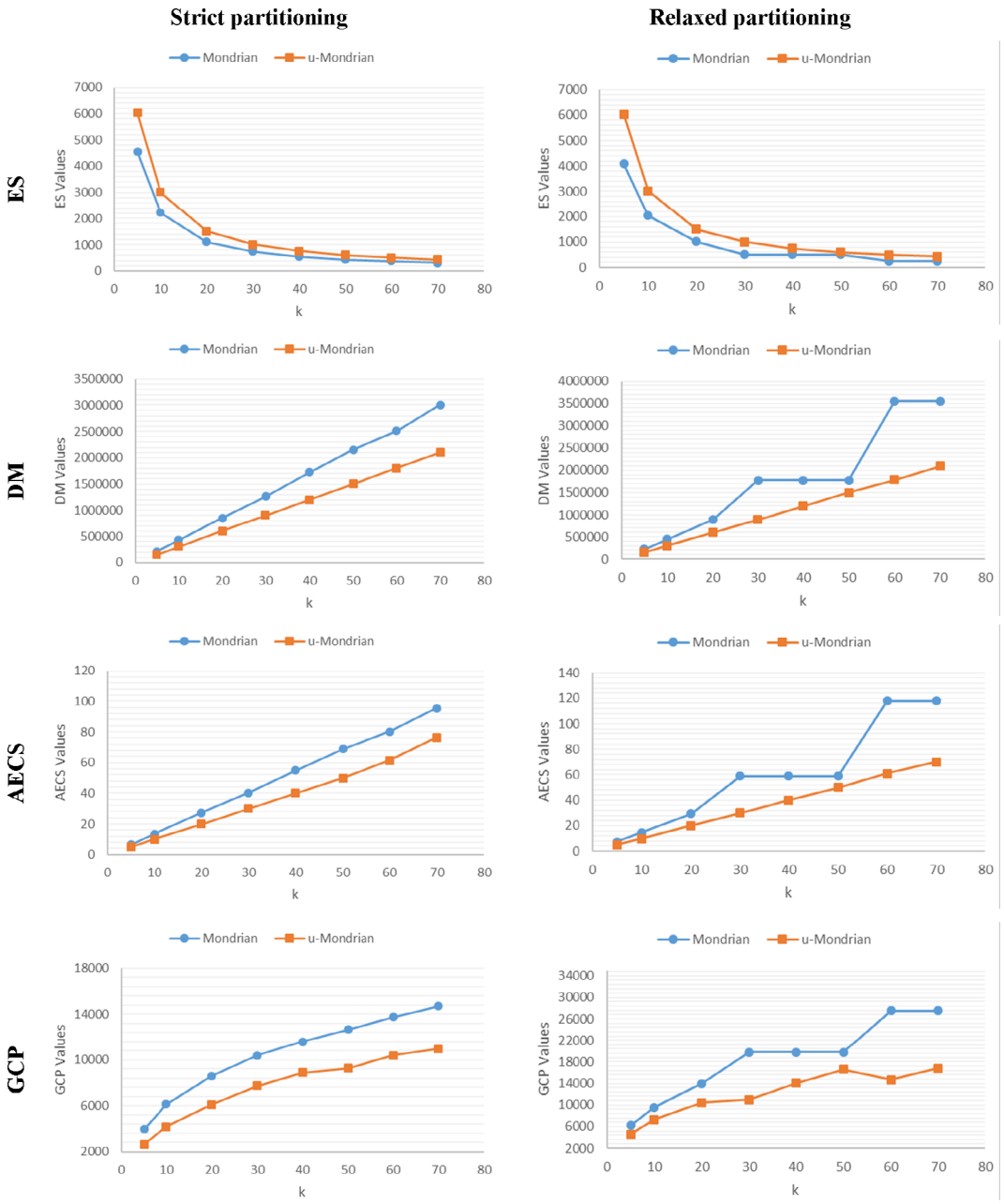

FIGURE 8 The results of u-Mondrian and Mondrian models for Adult dataset

  - Unlike the stability of DM, AECS, and ES values of u-Mondrian, GCP values represent a meaningful change for u-Mondrian with strict and relaxed partitioning.
  - If we observe GCP values for Mondrian, it can be seen that Mondrian is in the same trend with u-Mondrian. Therefore, it can be understood that this trend originates from the differences in partitioning strategies.
- Overall evaluation for the results of Mondrian and u-Mondrian models;
  - u-Mondrian generates more equivalence classes than Mondrian.
  - u-Mondrian presents lower DM, AECS, and GCP values than Mondrian.
  - The enhancement in the values of GCP metric especially proves that u-Mondrian facilitates to obtain higher data utility than Mondrian by grouping closer data points.
  - u-Mondrian presents higher data utility than Mondrian.
- Overall evaluation for u-Mondrian;
  - The proposed model offers a solution to the upper bound problem of Mondrian.

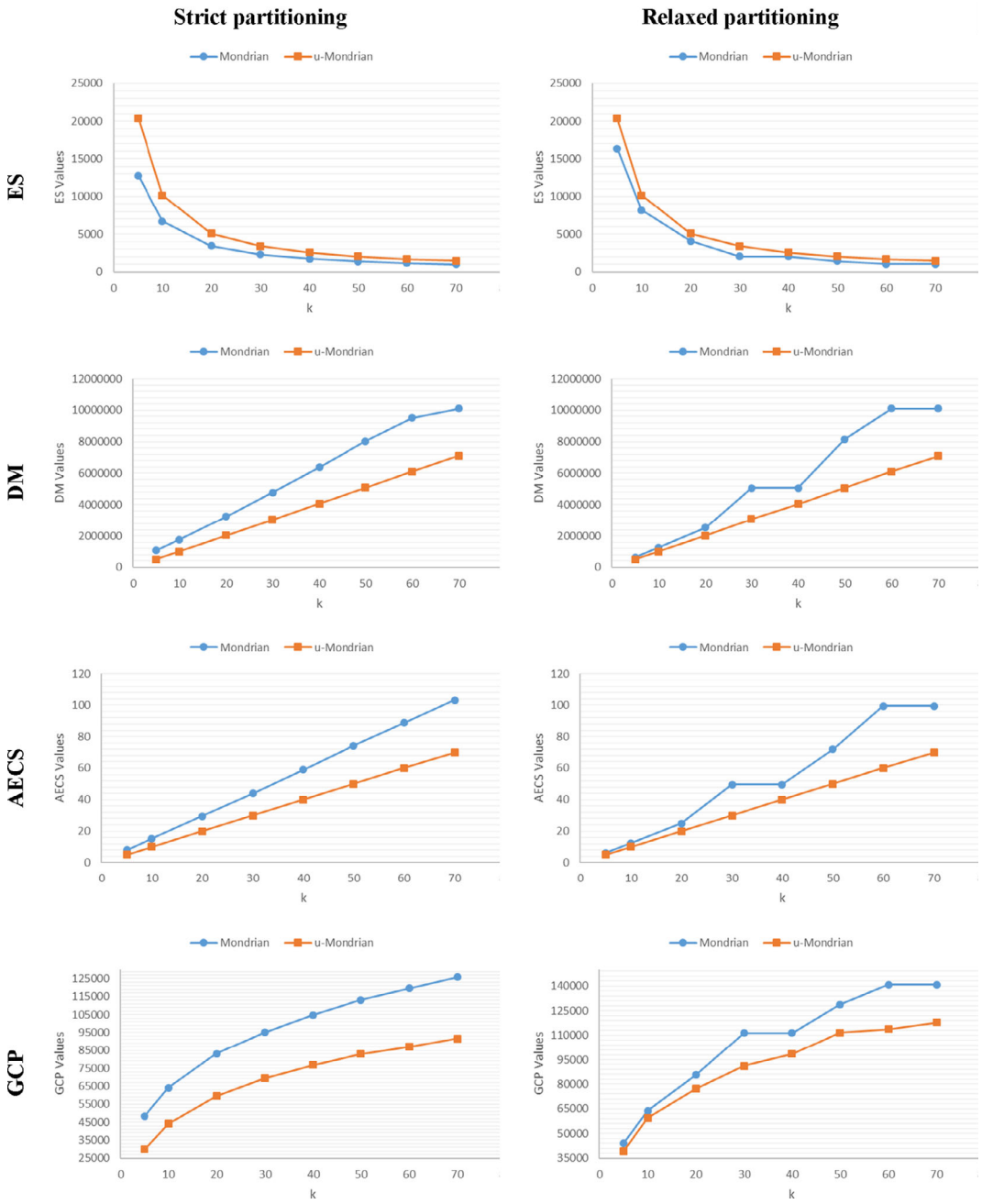

FIGURE 9 The results of u-Mondrian and Mondrian models for Diabetes dataset

- This solution enhances the upper bounds of each equivalence class by applying an outlier concept.
- u-Mondrian is a reliable solution that was proved by using two different datasets.
- Each equivalence class contains exactly *k* members (we except the bound of the last partition if it could not be partitioned any more).
- u-Mondrian enhances data utility for different datasets and this proves that u-Mondrian is a reliable solution.
- Considering the enhancement ratios of the proposed models presented in Table 4, u-Mondrian might be confidently used for PPDP with higher data utility.
- u-Mondrian presents 30.10%, 26.87%, and 33.30% better data utility than Mondrian in terms of DM, AECS, and GCP metrics, respectively, for strict partitioning model on Adult dataset.
- u-Mondrian presents 49.75%, 48.82%, and 46.69% better data utility than Mondrian in terms of DM, AECS, and GCP metrics, respectively, for relaxed partitioning model on Adult dataset.
- u-Mondrian presents 52.96%, 37.14%, and 38.32% better data utility than Mondrian in terms of DM, AECS, and GCP metrics, respectively, for strict partitioning model on Diabetes dataset.

TABLE 4 Improvements achieved with the proposed u-Mondrian model

Dataset	Partitioning strategy	*k*	ES (%)	DM (%)	AECS (%)	GCP (%)
Adult	Strict	5	32.41	27.53	22.86	33.30
		10	35.55	29.27	25.74	32.04
		20	36.38	29.62	25.52	28.96
		30	34.17	28.66	19.54	25.41
		40	37.09	30.07	26.87	23.31
		50	37.44	30.10	24.48	26.97
		60	35.63	26.98	23.57	15.70
		70	36.07	30.01	20.16	25.44
	Relaxed	5	47.11	32.46	31.42	26.99
		10	46.87	32.34	31.82	23.90
		20	46.87	32.30	31.82	25.28
		30	93.75	49.75	48.82	44.50
		40	46.87	32.28	31.82	28.87
		50	17.57	15.30	12.01	16.02
		60	93.75	49.75	48.09	46.69
		70	67.18	40.98	36.20	38.98
Diabetes	Strict	5	59.11	52.96	37.14	38.32
		10	52.26	42.06	34.22	31.13
		20	48.16	37.61	32.42	28.32
		30	45.67	36.50	31.65	28.17
		40	47.47	36.38	32.07	26.83
		50	48.64	36.76	32.37	26.63
		60	46.68	36.31	32.31	29.18
		70	47.16	29.70	32.17	27.17
	Relaxed	5	24.10	19.84	19.33	10.27
		10	24.21	19.71	19.16	6.82
		20	24.21	20.02	19.44	9.88
		30	65.62	39.63	39.60	18.22
		40	24.21	20.01	19.46	11.52
		50	43.91	37.83	30.19	13.46
		60	65.62	39.63	39.62	19.21
		70	41.40	29.84	29.56	16.30

– u-Mondrian presents 39.63%, 39.62%, and 19.21% better data utility than Mondrian in terms of DM, AECS, and GCP metrics, respectively, for relaxed partitioning model on Diabetes dataset.

6 | CONCLUSION

In this article, a new utility-aware anonymization model called u-Mondrian was designed, developed, compared, and presented successfully. u-Mondrian presents a solution for the upper bound problem of Mondrian by applying an outlier-oriented concept, increases total data utility by managing outliers, and finally presents better results than Mondrian. To reveal the relation between outlier management and data utility, decision rules based on information metrics were also proposed in this study for the first time.

The proposed u-Mondrian model is tested and then compared to Mondrian for different *k* values, metrics, partitioning strategies, and datasets. The results have shown that u-Mondrian achieves better data utilities in term of DM, AECS, and GCP metrics than Mondrian for both partitioning strategies and datasets.

It is expected that the proposed utility-aware anonymization model might help researchers to improve anonymization with higher data utility and preserve data privacy. In the future, we are planning to utilize different outlier determination approaches and develop new anonymization models by employing a metric tree to partition data space.

DATA AVAILABILITY STATEMENT

The data that support the findings of this study are openly available in UCI Machine Learning Repository at http://archive.ics.uci.edu/ml.[59]

ORCID


Yavuz Canbay https://orcid.org/0000-0003-2316-7893

REFERENCES

1. Hasan AT, Jiang Q. A general framework for privacy preserving sequential data publishing. International Conference on Advanced Information Networking and Applications Workshops; 2017; Taipei, Taiwan.
2. Almasi MM, Siddiqui TR, Mohammed N, Hemmati H. The risk-utility tradeoff for data privacy models. International Conference on New Technologies, Mobility and Security; 2016; Larnaca, Cyprus.
3. Chen X, Huang V. Privacy preserving data publishing for recommender system. IEEE Annual Computer Software and Applications Conference Workshops; 2012; Izmir, Turkey.
4. Wang R, Zhu Y, Chang C-C, Peng Q. Privacy-preserving high-dimensional data publishing for classification. *Comput Secur*. 2020;93:101785.
5. Fang W, Wen XZ, Zheng Y, Zhou M. A survey of big data security and privacy preserving. *IETE Tech Rev*. 2017;34(5):544-560.
6. Chibba M, Cavoukian A. Privacy, consumer trust and big data: privacy by design and the 3 C's. ITU Kaleidoscope: Trust in the Information Society; 2015; Barcelona, Spain.
7. Jain P, Gyanchandani M, Khare N. Big data privacy: a technological perspective and review. *J Big Data*. 2016;3(1):25.
8. Nayahi JJV, Kavitha V. Privacy and utility preserving data clustering for data anonymization and distribution on Hadoop. *Future Gener Comput Syst*. 2017;74:393-408.
9. Tang Q, Wu Y, Liao S, Wang X. Utility-based k-anonymization. International Conference on Networked Computing and Advanced Information Management; 2010; Seoul, South Korea.
10. General Data Protection Regulation. Accessed March 12, 2020. https://eur-lex.europa.eu/eli/reg/2016/679/oj
11. Personal Data Protection Law. Accessed March 12, 2020. https://www.kvkk.gov.tr/Icerik/6649/Personal-Data-Protection-Law
12. Eom CS-H, Lee CC, Lee W, Leung CK. Effective privacy preserving data publishing by vectorization. *Inf Sci*. 2020;527:311-328.
13. Majeed A, Lee S. Anonymization techniques for privacy preserving data publishing: a comprehensive survey. *IEEE Access*. 2020;9:8512-8545.
14. Fung BC, Wang K, Fu AW, Philip SY. *Introduction to Privacy-Preserving Data Publishing: Concepts and Techniques*. CRC Press; 2010.
15. Sweeney L. K-anonymity: a model for protecting privacy. *Int J Uncertain Fuzziness Knowl Based Syst*. 2002;10(5):557-570.
16. Machanavajjhala A, Gehrke J, Kifer D, Venkitasubramaniam M. L-diversity: privacy beyond k-anonymity. International Conference on Data Engineering; 2006; Atlanta, USA.
17. Li N, Li T, Venkatasubramanian S. t-Closeness: privacy beyond k-anonymity and l-diversity. IEEE International Conference on Data Engineering; 2007; Istanbul, Turkey.
18. Nergiz ME, Atzori M, Clifton C. Hiding the presence of individuals from shared databases. ACM SIGMOD International Conference on Management of Data; 2007; Beijing, China.
19. Dwork C. Differential privacy. International Colloquium on Automata, Languages, and Programming; 2006:1-12; Springer.
20. Zhao J, Chen Y, Zhang W. Differential privacy preservation in deep learning: challenges, opportunities and solutions. *IEEE Access*. 2019;7:48901-48911.
21. Huang X, Guan J, Zhang B, Qi S, Wang X, Liao Q. Differentially private convolutional neural networks with adaptive gradient descent. IEEE Fourth International Conference on Data Science in Cyberspace (DSC); 2019.
22. Ziller A, Usynin D, Braren R, Makowski M, Rueckert D, Kaissis G. Medical imaging deep learning with differential privacy. *Sci Rep*. 2021;11(1):13524.
23. Fioretto F, Van Hentenryck P, Zhu K. Differential privacy of hierarchical census data: an optimization approach. *Artif Intell*. 2021;296:103475.
24. Xu C, Ren J, Zhang D, Zhang Y, Qin Z, Ren K. GANobfuscator: mitigating information leakage under GAN via differential privacy. *IEEE Trans Inf Forensics Secur*. 2019;14(9):2358-2371.
25. Abdelhameed SA, Moussa SM, Khalifa ME. Privacy-preserving tabular data publishing: a comprehensive evaluation from web to cloud. *Comput Secur*. 2017;72:74-95.
26. Gong Q, Luo J, Yang M, Ni W, Li X-B. Anonymizing 1: M microdata with high utility. *Knowl Based Syst*. 2017;115:15-26.
27. Tao Y, Tong Y, Tan S, Tang S, Yang D. Protecting the publishing identity in multiple tuples. IFIP Annual Conference on Data and Applications Security and Privacy; 2008:205-218; Springer.
28. Poulis G, Loukides G, Gkoulalas-Divanis A, Skiadopoulos S. Anonymizing data with relational and transaction attributes. Joint European Conference on Machine Learning and Knowledge Discovery in Databases; 2013:353-369; Springer.
29. Meyerson A, Williams R. On the complexity of optimal k-anonymity. ACM SIGMOD-SIGACT-SIGART Symposium on Principles of Database Systems; 2004; Paris, France.
30. Aggarwal G, Feder T, Kenthapadi K, et al. Approximation algorithms for k-anonymity. *J Priv*. 2005;20051120001:1-18.
31. Aggarwal G, Feder T, Kenthapadi K, et al. Anonymizing tables. International Conference on Database Theory; 2005:246-258; Springer.
32. Bayardo RJ, Agrawal R. Data privacy through optimal k-anonymization. International Conference on Data Engineering; 2005; Tokyo, Japan.
33. Aggarwal C. On k-anonymity and the curse of dimensionality. International Conference on Very Large Data Bases; 2005; Trondheim, Norway.

34. LeFevre K, DeWitt D, Ramakrishnan R. Mondrian multidimensional k-anonymity. International Conference on Data Engineering; 2006; Atlanta, USA.
35. Kumar N, Zhang L, Nayar S. What is a good nearest neighbors algorithm for finding similar patches in images? European Conference on Computer Vision; 2008; Marseille, France.
36. Dolatshah M, Hadian A, Minaei-Bidgoli B. Ball*-tree: efficient spatial indexing for constrained nearest-neighbor search in metric spaces; 2015. arXiv preprint arXiv:151100628.
37. Witten IH, Frank E, Hall MA, Pal CJ. *Data Mining: Practical Machine Learning Tools and Techniques*. Morgan Kaufmann; 2016.
38. Wang J, Wang N, Jia Y, et al. Trinary-projection trees for approximate nearest neighbor search. *IEEE Trans Pattern Anal Mach Intell*. 2014;36(2):388-403.
39. Bentley JL. Multidimensional binary search trees used for associative searching. *Commun ACM*. 1975;18(9):509-517.
40. Tang Q, Wu Y, Wang X. New algorithm with lower upper size bound for k-anonymity. International Conference on Communication Systems, Networks and Applications; 2010; Hong Kong, China.
41. Canbay Y, Sagiroglu S, Vural Y. A Mondrian-based utility optimization model for anonymization. International Conference on Computer Science and Engineering (UBMK); 2019.
42. Liu KC, Kuo CW, Liao WC, Wang PC. Optimized data de-identification using multidimensional k-anonymity. IEEE International Conference on Trust, Security and Privacy in Computing and Communication, IEEE International Conference on Big Data Science and Engineering; 2018; New York, USA.
43. Gong Q, Yang M, Chen Z, Luo J. Utility enhanced anonymization for incomplete microdata. International Conference on Computer Supported Cooperative Work in Design; 2016; Nanchang, China.
44. Nergiz ME, Gök MZ. Hybrid k-anonymity. *Comput Secur*. 2014;44:51-63.
45. LeFevre K, DeWitt D, Ramakrishnan R. Workload-aware anonymization. Proceedings of the 12th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. 2006:277-286.
46. Canbay Y, Vural Y, Sagiroglu S. OAN: outlier record-oriented utility-based privacy preserving model. *J Fac Eng Archit Gazi Univ*. 2020;35(1):355-368.
47. Vural Y. *P-Gain: Privacy Preserving Utility-Based Data Publishing Model*. Ph.D. thesis. Hacettepe University; 2017.
48. Vural Y, Aydos M. A new approach to utility-based privacy preserving in data publishing. IEEE International Conference on Computer and Information Technology; 2017; Helsinki, Finland.
49. Vural Y, Aydos M. ρ-Gain: utility based data publishing model. *J Fac Eng Archit Gazi Univ*. 2018;33(4):1355-1368.
50. Lee H, Kim S, Kim JW, Chung YD. Utility-preserving anonymization for health data publishing. *BMC Med Inform Decis Mak*. 2017;17:104.
51. Ramana KV, Kumari V, Raju K. Impact of outliers on anonymized categorical data. Advances in Digital Image Processing and Information Technology; 2011; Berlin, Germany.
52. Wang HW, Liu R. Hiding outliers into crowd: privacy-preserving data publishing with outliers. *Data Knowl Eng*. 2015;100:94-115.
53. Wang HW, Liu R. Hiding distinguished ones into crowd: privacy-preserving publishing data with outliers. International Conference on Extending Database Technology: Advances in Database Technology; 2009; Saint Petersburg, Russia.
54. Majeed A. Attribute-centric anonymization scheme for improving user privacy and utility of publishing e-health data. *J King Saud Univ Comput Inf Sci*. 2019;31(4):426-435.
55. Nergiz ME, Gök MZ, Özkanlı U. Preservation of utility through hybrid k-anonymization. International Conference on Trust, Privacy and Security in Digital Business; 2013:97-111; Springer.
56. Prasser F, Kohlmayer F, Lautenschlaeger R, Kuhn KA. ARX—a comprehensive tool for anonymizing biomedical data. AMIA Annual Symposium Proceedings; 2014:984-993; American Medical Informatics Association.
57. Ghinita G, Karras P, Kalnis P, Mamoulis N. Fast data anonymization with low information loss. Proceedings of the 33rd International Conference on VLDB Endowment; 2007:758-769.
58. Breunig MM, Kriegel HP, Ng RT, Sander J. LOF: identifying density-based local outliers. Proceedings of the 2000 ACM SIGMOD International Conference on Management of Data; 2000:93-104; ACM.
59. Dheeru D, Taniskidou EK. UCI Machine Learning Repository. Accessed March 25, 2019. http://archive.ics.uci.edu/ml

How to cite this article: Canbay Y, Sagiroglu S, Vural Y. A new utility-aware anonymization model for privacy preserving data publishing. *Concurrency Computat Pract Exper*. 2022;e6808. doi: 10.1002/cpe.6808