Journal of Information Engineering and Applications www.iiste.org
ISSN 2224-5782 (print) ISSN 2225-0506 (online) DOI: 10.7176/JIEA
Vol.9, No.2, 2019

# Feature Based Data Anonymization for High Dimensional Data

Esther Gachanga*
School of Computer Science & Technology, P.O Box 62000- 00100, Nairobi, Kenya

Michael Kimwele
School of Computer Science & Technology,P.O Box 62000- 00100, Nairobi, Kenya

Lawrence Nderu
School of Computer Science & Technology, P.O Box 62000- 00100, Nairobi, Kenya

**Abstract**
Information surges and advances in machine learning tools have enable the collection and storage of large amounts of data. These data are highly dimensional. Individuals are deeply concerned about the consequences of sharing and publishing these data as it may contain their personal information and may compromise their privacy. Anonymization techniques have been used widely to protect sensitive information in published datasets. However, the anonymization of high dimensional data while balancing between privacy and utility is a challenge. In this paper we use feature selection with information gain and ranking to demonstrate that the challenge of high dimensionality in data can be addressed by anonymizing attributes with more irrelevant features. We conduct experiments with real life datasets and build classifiers with the anonymized datasets. Our results show that by combining feature selection with slicing and reducing the amount of data distortion for features with high relevance in a dataset, the utility of anonymized dataset can be enhanced.
**Keywords:** High Dimension, Privacy, Anonymization, Feature Selection, Classifier, Utility
**DOI**: 10.7176/JIEA/9-2-03
**Publication date**: April 30th 2019

## 1. Introduction

Advances in information technology have enabled the collection and storage of large amounts of data. These data has high dimensionality and data may be shared via publications for research and scientific purposes (Wagner & Eckhoff, 2018). Privacy is one of the major concerns, when sharing personal data, because without appropriate protection, personal information is vulnerable to misuse (Fung, Wang, & Yu, 2005). Micro-data contain different types of information which include; personally identifying information (PII) such as national identification number (ID), social security number (SSN) personal identification number (PIN) etc, quasi identifying information (QID) such as sex, date of birth, postal address, marital status etc and sensitive information such as disease, financial information etc (Samarati & Sweeney, 1998). To secure privacy in complex research environments, a broad spectrum of measures have been /must be implemented, including legal, contractual as well as technical measures (Fabian Prasser & Kohlmayer, 2015) and (W. Li et al., 2018). Privacy enhancing technologies (PET) together with data anonymization techniques are some of the technical measures implemented to enhance the privacy of data. In this paper we consider research efforts aimed at anonymizing quasi identifiers and sensitive attributes with the intention of ensuring that PII, QID and sensitive attributes are protected. We introduce feature based data anonymization and show that relevant feature selection can be used to reduce the amount of data distortion in a given dataset while preserving privacy and quality. The rest of this paper is organized as follows; section 2 presents the literature review while section 3 presents our approach. In section 4 we conduct experiments and discuss the results of the experiments in section 5 and finally, section 6 concludes the paper.

## 2. Related works

To protect the privacy of individuals in data various privacy enhancing technologies have been proposed (PET) (Abdou Hussien, Darwish, Hefny, & Hussien, 2015)

A study by (Samarati & Sweeney, 1998) introduced the k- anonymity model. K-anonymity ensures that there are at least $k$ people with the same quasi-identifier such that the risk of identity disclosure is reduced to $1/k$(Sweeney, 2002b). A table satisfies k-anonymity if every record in the table is indistinguishable from at least k−1 other records with respect to every set of quasi-identifier attributes; such a table is called a k-anonymous table. The enforcement of k-anonymity ensures that individuals cannot be uniquely identified by linking attack (Machanavajjhala, Kifer, Gehrke, & Venkitasubramaniam, 2007). Machanavajjhala et al. (2007) further observed that due to lack of diversity in the sensitive attribute, k-anonymity can create groups that leak sensitive information and hence privacy breach.

To overcome some of the shortcomings of the k-anonymity model, (Machanavajjhala et al., 2007) proposed the $l$-diversity model. The ℓ-Diversity Principle states that *"An equivalence class is said to have L-diversity if*

*there are at least L "well-represented" values for the sensitive attribute. A table is said to have L-diversity if every equivalence class of the table has L-diversity"* (Machanavajjhala et al., 2007). The degree of privacy protection is determined by the number of distinct sensitive values in each QI group. According to (Xiao & Tao, 2006b) the ℓ -diversity model guarantees stronger privacy than k-anonymity model. The *l-diversity* model requires that the values of attributes in a group have enough degree of diversity in the sensitive attribute and this effectively prevents the elimination attack. However *l*- diversity *has* some defects in handling numerical sensitive attributes as It's designed for static data with one sensitive attribute (Han, Yu, & Yu, 2008).

The work of (Ninghui, Tiancheng, & Venkatasubramanian, 2007)presented the *t*-closeness principle. The t-closeness principle requires that the distribution of a sensitive attribute in any equivalence class is close to the distribution of the attribute in the overall table as much as is possible. Therefore the distance between the two distributions should be no more than a threshold ***t*** (Ninghui et al., 2007). *T*- closeness requires that the sampling distribution of the confidential attribute within each of the k-anonymous groups be similar to the sampling distribution over the whole data set (Soria-comas & Domingo-ferrer, 2013).

Enforcing *t*-closeness destroys the correlation between quasi identifier attributes and sensitive attributes (Ahmed Ali Mubark, Emad Elabd, 2016).

(Dwork, 2006) introduced Differential Privacy (DP). DP is a mathematical framework that is widely accepted for protecting data privacy and captures ones increased risk to privacy by participating in a database. In DP, a randomized function *A randomized function K gives* $\epsilon$-differential privacy *if for all datasets D*1 *and D*2 *differing on at most one element, and all* $S \subseteq Range\ (K)$,

$$\Pr[K(D1) \in S] \leq \exp(\epsilon)\, x \Pr[K(D2) \in S \qquad (1)$$

The k-anonymity, ℓ-diversity and t-closeness are also commonly referred to as Syntactic privacy models and mainly utilize anonymization methods to enhance privacy. Syntactic models address the trade-off between privacy and utility by requiring the anonymized data set to follow a specific pattern that is known to limit the risk of disclosure privacy (Fabian Prasser & Kohlmayer, 2015).

**2.1 Data Anonymization**

In Cox (1980) and Dalenius (1986) Anonymization refers to the PPDP approach that seeks to hide the sensitive data and/or the identity of record owners, assuming that sensitive data must be retained for other purposes such as scientific research, data analysis e.t.c. Anonymous technology is a safe and effective method of privacy preservation. A study by (Samarati & Sweeney, 1998) conducted the pioneering studies on k-anonymity. This study utilized the concept of data anonymization. data anonymization approaches assume that data publisher has a table T that includes four subsets of attributes namely: 1) Personally Identifying Information (PII) which contains information that explicitly identifies a record owner and are removed from the released dataset for example name, social security number, passport number and cell-phone number; 2) quasi-identifiers (QID) containing information that could potentially identify a record owner and typically transformed in the released dataset such as date-of-birth, gender and ZIP code. 3) Sensitive attributes (S) containing sensitive information about data owner such as financial or medical information and/which should be protected, and 4) non-sensitive attributes, which refers to any information that does not fall into the previous three categories and can be published as it is, when needed e.g. one's favorite food (Canbay & Sever, 2015).

Anonymity as method of privacy protection has been utilized to effectively balance the relationship between the efficiency and the security of the data. The basic idea of anonymization is that from a transformed table, the attacker cannot easily analyze the sensitive attribute of a tuple, and therefore cannot identify a specific individual's sensitive information. This effectively addresses the issue of re-identification in published datasets Wang et.al (2016).

Anonymity principles focus on a universal approach that exerts the same amount of privacy preservation for all individuals. These principles do not take the personalized privacy preservation requirement into consideration (Liu, Xie, & Wang, 2015). Data sets are anonymized to satisfy certain privacy requirements such as *k*-anonymity, or *l*-diversitybefore they are shared with data users. Data anonymization preserves privacy by eliminating the link between people and sensitive information and therefore preventing their identifiability from the dataset (Lee, Kim, Kim, & Chung, 2017) .

The main challenge with any anonymization process is to balance between the utility and the privacy of the data. Hiding data reduces the utility of the data, while disclosing the data reduces privacy. During the anonymization process there is a tradeoff between privacy and information loss (Bhaladhare & Jinwala, 2012). Different anonymization levels leads to different amounts of information loss (Gal, Tucker, Gangopadhyay, & Chen, 2014). Various methods have been used to implement data anonymization processes.

**2.1.1 Generalization**

Generalization is a widely used anonymization approach, which replaces quasi-identifier values with values that

are less- specific but semantically consistent. As a result, more records will have the same set of quasi-identifier values. An equivalence class of an anonymized table is defined to be a set of records that have the same values for the quasi-identifiers (Sunitha, Venkata Subba Reddy, & Vijayakumar, 2012). The number of equivalence classes $C$ in anonymized data set is determined by the specified value of $k$ in a $k$- anonymity model (Chiu & Tsai, 2007). This is given by the following formulae;

$$C = M/k \qquad [2]$$

Where given $C$ is the equivalence classes, $M$ is the number of records in a table and $k$ is the value of $k$ specified in $k$ - anonymity. generalization preserves privacy, however a considerable amount of information in the micro-data is lost, which severely compromises the accuracy of data analysis (Xiao & Tao, 2006b). Generalization does not work well on high dimensional data due to the curse of dimensionality (Mohanapriya, 2013).

**2.2.2 Suppression**

Suppression means to remove data from a table so that the data is not released. Sensitive information and all other information that may allow inference to sensitive information is not released (Sweeney, 2002). Suppression can be applied at the tuple level, where, a tuple can be suppressed only in its entirety. Suppression is used to moderate the generalization process when a limited number of outliers would force a great amount of generalization.

**2.2.3 Anatomy**

Anatomy was proposed to overcome the defects of generalization. According to (Xiao & Tao, 2006a) Anatomy releases all the quasi-identifier and sensitive values directly in two separate tables. This is combined with a grouping mechanism; so as to captures a large amount of correlation in the micro-data and to protect privacy. Anatomy releases a quasi-identifying table (QIT) and a sensitive table (ST). The construction of the table is done by partitioning the tuples of the micro-data into several QI groups based on a certain strategy. The grouping is done for each tuple in the table using the QIT table (Xiao & Tao, 2006a).

**2.2.4 Data Swapping**

This technique transforms a database by exchanging values of sensitive variables among individual records. The swapping/ exchange of records is done in such a way so as to maintain lower-order frequency counts or marginals (Fienberg & Mcintyre, 2005).

**2.2.5 Perturbation**

In (Smith, 2006)perturbation is the most natural way to anonymize numerical data. For an attribute whose value is $x$ the data publisher, publishes $\tilde{x} = x + r$ where $r$ is a random value drawn from a bias free distribution. The larger the perturbation, the more blurred the value and thus the more protected the data. Perturbation works by changing the values of data items in a given dataset (Smith, 2006).

**2.2.6 Slicing**

According to (T. Li, Li, Zhang, & Molloy, 2012) slicing partitions a data set horizontally and vertically. Horizontal partitioning is done by grouping tuples into buckets. Values in each column are randomly permutated (or sorted) within each bucket, so as to break the linking between different columns. Vertical partitioning is done by grouping attributes into columns based on the correlations between them. Each column contains a subset of attributes that are highly correlated. The basic idea with slicing is to preserve the association within each column and to break the association across columns and therefore reduce the dimensionality of the data. By grouping highly correlated attributes together Slicing preserves utility, and preserves the correlations between such attributes. Slicing protects privacy as it breaks the associations between the uncorrelated attributes, that are infrequent and thus identifying (T. Li et al., 2012).

**2.3 Feature selection**

A study by (H. Liu, 1998) defined feature selection as the study of algorithms selecting an optimal subset of the input feature set. Optimality is normally dependent on the evaluation criteria or the application's needs. (Hall, 2000) describe the task of feature selection as a search problem, with each state in the search space specifying a subset of the possible features. According to (Qinbao Song, Jingjie Ni, & Guangtao Wang, 2013) feature selection involves identifying a subset of the most useful features that produces compatible results as the original entire set of features. The work of (Sharma & Bala, 2014) assert that the central assumption when using a feature extraction technique is that data contains many redundant or irrelevant features.

Redundant features not only affect the performance of classification algorithm but also require an additional computational cost (Miao & Niu, 2016). Thus, feature subset selection should be able to identify and remove as much of the irrelevant and redundant information as much as possible. Feature ranking algorithms assign a weight to each feature of the data set and rank the relevance of features according to their weights (Hall, 2000).

**2.3.1 Information Gain**

Information gain has been used as a measure of feature relevancy for filter based feature selection and evaluates

the worth of an attribute by measuring the information gain with respect to a class (Shaltout, El-Hefnawi, Rafea, Moustafa, & El-Hefnawi, 2014). Given two attributes X and Y that belong to dataset D, the Information Gain or mutual information between attributes can be calculated using conditional entropy as;

$$IG[X;Y] = H(X) - H\left(\frac{X}{Y}\right) + H(Y) - H\left(\frac{Y}{X}\right) \quad [3]$$

Where IG [X; Y] measures the degree of uncertainty about $X\boldsymbol{x}$ due to the knowledge of $y$. IG [X; Y] also measures the two way association between the attributes $\boldsymbol{x}$ and $\boldsymbol{y}$.

### 2.4 Data Utility Vs Privacy

In (Dwork & Roth, 2013), perfect privacy can be achieved by publishing nothing at all but this has no utility. Perfect utility can be achieved by publishing data as received but this has no privacy at all. In (Mivule & Turner, 2013), the more confidential data is , the more likely that the privatized data will decline in utility and therefore may become useless.

(Doka et al., 2015) argue that utility may be achieved at the expense of runtime since the anonymization process is a one-time process. Privacy preservation requires that data is protected with minimum impact on its accuracy and utility (Basso, Matsunaga, Moraes, & Antunes, 2016). (Basso et al., 2016) further present that excessive data anonymization can make the published data less useful and so it's important to me**as**ure the utility of anonymized data.

#### 2.4.1 Information Loss

According to (Ghinita, Karras, Kalnis, & Mamoulis, 2007) all privacy-preserving transformations cause a certain degree of information loss. This loss must be minimized in order to maintain high utility in the dataset and thus, the ability to extract meaningful information from the published data.

We use the term information loss to refer to loss in data accuracy caused by anonymization. The main objective of the anonymization techniques is privacy protection. However, it is important that the anonymized dataset should be as useful as possible. The utility of data is measured by the quality of the anonymized dataset (Lee et al., 2017).

### 2.5 Information Loss Metrics

A study by (Garcia-Alfaro et al., 2015) presents that several metrics have been proposed to assess the level of anonymity in a published dataset. These metrics differ in a number of ways, but they all express the risk of disclosing personal-identifiable information when releasing a given dataset (Garcia-Alfaro et al., 2015).

A survey by (Wagner & Eckhoff, 2018) explains that many authors have proposed various metrics to measure the quality of data after the anonymization process. We discuss some of these metrics.

#### 2.5.1 Generalization Height

Generalization height by (Sweeney, 2002a) is one of the earliest utility measures. Generalization height refers to the total number of generalization steps that have been performed on the original data set to anonymize it. A generalization step represents a loss of information, so the use as few generalization steps as possible maximize on utility. When the height is 0 (zero), that means no distortion (Huang, Liu, Han, & Yang, 2014). The distortion value of a whole table is the sum of all generalized values in a generalized table and is given as;

$$distortion = \sum_{ij} h_{ij} \quad [4]$$

Where $h_{ij}$ is the height of the value a generalized QI of $q_i$of the record $r_j$.

#### 2.5.2 Classification Metric

The work of (Iyengar, 2002) proposed the classification metric (CM), which was defined as the sum of the individual penalties for each row in the table normalized by the total number of rows $N$ (see equation below);

$$CM = \frac{\sum all\ rows\ penalty\ (row\ r)}{N} \quad [5]$$

A study by (Kifer & Gehrke, 2006) explains that classification metric is appropriate when one wants to train a classifier over the anonymized data.

#### 2.5.3 Suppression Ratio

Another metric used is the suppression ratio, which is the proportion of suppressed records to the total records presented by (Han, 2013) and is defined as;

$$\text{Suppression Ratio} = \frac{n_s}{n} \quad [6]$$

Where $n_s$ $\boldsymbol{n_S}$ is the number of suppressed tuples/records and $n$ is the total number of records. A lower suppression ratio implies a higher quality of data. When the suppression ratio is zero, the data quality is optimal.

#### 2.5.4 Performance of Classifiers

A classifiermust be evaluatedin-order to determine its performance. ROC curves determine the area under curve (commonly referred to as AUC) is the most popular approach used to determine the performance of a classifier (Fawcett, 2004). The performance of different classification models can be compared using a performance metric

such as accuracy (Tan, Steinbach, & Kumar, 2006), which can be defined as;

$$\text{Accuracy} = \frac{\text{Number of correct predictions}}{\text{Total number of predictions}} = \frac{f_{11} + f_{00}}{f_{11} + f_{10} + f_{01} + f_{00}} \quad [7]$$

## 3 PROPOSED APPROACH

In this section, we present a feature based anonymization model with information gain and slicing. The model first categorizes the different attributes in a dataset into quasi identifiers and sensitive attributes. Secondly it performs a feature selection by establishing the information gain with ranking for the different quasi identifiers. The quasi identifiers are ranked based on their information gain score. Third we slice the quasi identifier attributes into low information gain and high information using a user defined threshold. The low information gain quasi identifiers are anonymized via generalization and suppression and finally we evaluate the model.

### 3.1 The Model

The proposed model provides an outline of the various steps that one should follow in order to anonymize data while maintaining data quality and at the same time preserving privacy. The raw dataset is preprocessed. Preprocessing involves removing tuples with missing values. The next step involves categorizing the attributes in the dataset into the different categories as described in section 2.1. These categories are; the personally identifying information (PII), the quasi identifiers (QIDs) and the sensitive attributes (SAs). The model then establishes the information gain for the different QIDs and ranks the attributes. A feature with a high level of redundancy is ranked low. The model performs horizontal slicing as described in section 2.2.6. Slicing yields two slices namely the high information gain attributes and the low information gain attributes. The model anonymizes the low information gain attributes. Finally, the model tests the anonymized dataset for quality. To implement the model experiments were conducted. An anonymization tool built on java platform was used to conduct the experiments. The tool provided a graphical user interface for the entry of the data and a display for the anonymized data. Figure 1 presents our model;

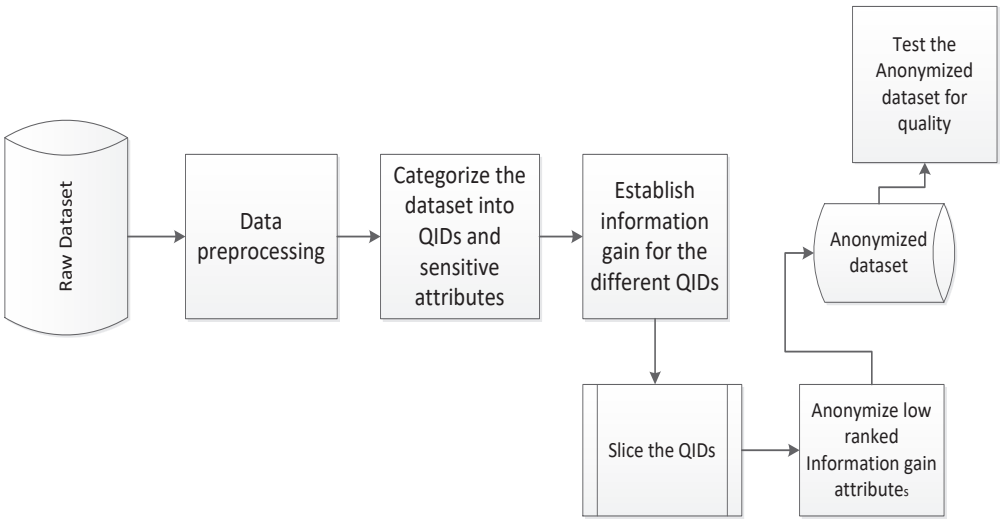


Figure 1: Feature Based Data Anonymization Model with Information Gain

The model utilizes feature selection with ranking and the slicing technique. Feature selection is useful in identifying the subset of features that are most useful. The general intuition when using a feature extraction technique is that data contains many redundant features. Redundant features affect the performance of classification algorithm and also require an additional computational cost. Slicing was used to address the challenge of high dimensionality in data.

## 4 Experiment

We adopted the adult dataset (Dheeru, Dua , Karra Taniskidou, 2017) as it is a real-world dataset and has already been utilized for benchmarking previous work on k-anonymity. Data cleaning was done by removing tuples with missing values. The dataset contains 30162 tuples after cleaning. Ten attributes were used for the experiments. The selected attributes were; Age, Work class, Education, Marital status, Occupation, Relationship, Race, Sex, Hours worked and Salary. The Dataset is conceptually organized as a table of rows (or records) and columns (or fields). Each row is termed a *tuple*. Tuples within a table are not necessarily unique. Each column is called an *attribute* and denotes a set of possible values within its domain. We begin by categorizing our attributes into quasi identifying and sensitive attribute. Occupation was used as the sensitive attribute while the other nine attributes

were used as the quasi identifier attributes. All the experiments were conducted on a java platform. We then establish the information gain for the different quasi identifiers with ranking. We vertically slice the quasi identifier attributes based on the amount of information gain. The quasi identifiers with low information gain are generalized using the approach similar to that of (Fabian Prasser & Kohlmayer, 2015) and (F. Prasser, Eicher, Bild, Spengler, & Kuhn, 2017). The purpose of this is to reduce the overall distortion of data when the data analysis task is unknown. We utilize generalization and suppression for the anonymization process. We use the *l*-diversity model for the sensitive attribute occupation.

**5 Results and Discussions**

Our first experiment aimed at establishing the information gain for the different quasi identifier attributes (see Table 1). Information Gain (IG) was used as the feature selection method since it is a filter based technique and may scale well with highly dimensional data. The attribute occupation was not used in the determination of information gain as it's a sensitive attribute and was therefore not generalized. We used a threshold of 0.02 for the information gain. We categorized features scoring less than 0.02 (work class, age, race, education and native country) as low information gain quasi identifiers while those that scored greater than 0.02 (salary, marital status and relationship) were categorized as high information gain quasi identifiers. This formed the basis for slicing our data. We sliced our data horizontally. The results for the sliced attributes are presented in table 2.

Table 1: Information Gain Score

| Attribute | Score with Ranking |
|---|---|
| Relationship | 0.394 |
| Marital Status | 0.167 |
| Salary | 0.037 |
| Work Class | 0.017 |
| Age | 0.01 |
| Race | 0.01 |
| Education | 0.006 |
| Native Country | 0.004 |

Table 2: The sliced QID Table

| High Information Gain Attribute | Low Information Gain Attribute |
|---|---|
| Relationship | Work Class |
| Marital Status | Age |
| Salary | Race |
| | Education |
| | Native Country |

The next step was to generalize the QIDs with low information gain. The purpose of this was to reduce the amount of data distortion that occurs during the anonymization process for attributes whose features were deemed to have a high level of relevance. During the anonymization process we utilized generalization and suppression. In generalization we used global transformation and a suppression rate of 1.5% for all the experiments. We used generalization hierarchies and classifier performance to measure the quality of the anonymized dataset. Intuitively a high amount of generalization implies a high amount of data distortion occurred.

We build three classifiers namely, Logistic Regression, Naïve bayes and Random forest using the anonymized dataset with k=5 and l=3 without Information Gain and k=5 and l=3 with Information Gain with our target variable as salary and presents the results in table 3 and 4.

Table 3 Classifier Performance when k=5 and l=3 without Information Gain

| Classifier | Input Data | Output/ Anonymized Data |
|---|---|---|
| Naïve Bayes | % | % |
| ≤ 50 | 86.15158 | 85.15485 |
| ≥ 50 | 86.15159 | 85.15491 |
| | | |
| Logistic Regression | | |
| ≤ 50 | 88.4457 | 86.76108 |
| ≥ 50 | 88.44572 | 86.76114 |
| | | |
| Random Forest | | |
| ≤ 50 | 86.10684 | 85.21981 |
| ≥ 50 | 86.10686 | 85.21982 |

Table 3 presents a summary of the performance of the three classifiers built on the input data and on the

anonymized dataset when the value of k=5 and the value of $l = 3$ in the *l*-diversity model with the target variable being salary.

Table 4: Classifier Performance when k=5 and l = 3 with Feature Selection

| Classifier | Input Data | output/ Anonymized Data |
|---|---|---|
| Naïve Bayes | % | % |
| ≤ 50 | 86.1553 | 85.81782 |
| ≥ 50 | 86.15531 | 85.81802 |
| | | |
| Logistic Regression | | |
| ≤ 50 | 88.44541 | 87.73816 |
| ≥ 50 | 88.44543 | 87.73837 |
| | | |
| Random Forest | | |
| ≤ 50 | 86.35558 | 84.76587 |
| ≥ 50 | 86.35559 | 84.76892 |

Table 4 presents a summary of the performance of the three classifiers built on the input data and on the anonymized dataset when the value of k=5 and the value of $l = 3$ in the *l*-diversity model with information gain.

With the Naïve bayes classifier when salary was ≤ 50, when k=5 and *l*=3 the classification accuracy for the output data was 85.15485% while that of k=5 and *l*=3 for feature selection with information gain was 85.81782%. We noted that the performance of the classifier improved by 0.66297%. For salary ≥ 50, when k=5 and *l*=3 the classification accuracy for the output data was 85.15491% while that of k=5 and *l*=3 using feature selection with information gain and slicing method was 85.81802%. We noted that the performance of the classifier improved by 0.66311%.

With the Logistic regression classifier when salary was ≤ 50, when k=5 and *l*=3 the classification accuracy for the output data was 86.76108% while that of k=5 and *l*=3 for feature selection with information gain was 87.73816%. We noted that the performance of the classifier improved by 0.97708%. For salary ≥ 50, when k=5 and *l*=3 the classification accuracy for the output data was 86.76114% while that of k=5 and *l*=3 using feature selection with information gain and slicing method was 87.73837%. We noted that the performance of the classifier improved by 0.97723%.

With the RandomForest classifier when salary was ≤ 50, when k=5 and *l*=3 the classification accuracy for the output data was 85.21981% while that of k=5 and *l*=3 for feature selection with information gain was 84.76587%. We noted that the performance of the classifier degraded by (-0.45394%). For salary ≥ 50, when k=5 and *l*=3 the classification accuracy for the output data was 85.21982% while that of k=5 and *l*=3 using feature selection with information gain and slicing method was 84.76892%. We observed that the performance of the classifier degraded by (-0.4509%).

To determine the quality of the anonymized dataset we evaluate the performance of the classifiers in table 3 and 4. This was done by comparing the results in table 3 and 4 with the results of a classifier built from the raw dataset before the anonymization process.

Table 5: % Information Loss in Classifier Accuracy for k=5

| Classifier | k=5 | k=5 with Information Gain |
|---|---|---|
| Naïve Bayes | | |
| ≤ 50 | 1.00206 | 0.34342 |
| ≥ 50 | 0.96187 | 0.34351 |
| | | |
| Logistic Regression | | |
| ≤ 50 | 1.65435 | 0.77863 |
| ≥ 50 | 1.65415 | 0.77872 |
| | | |
| Random Forest | | |
| ≤ 50 | 4.52429 | -0.89638 |
| ≥ 50 | 4.52333 | -0.89624 |

Table 5 presents the loss in the performance of the classifier when salary was the target attribute. This research compared the performance of the classifier with the input data and with output data when the value of k=5 without information gain and for k=5 with information gain. The results were presented in table 5. An analysis of the results in table 5 revealed the following;

With the Naïve bayes classifier when k=5 for salary ≤ 50, the accuracy went down by 1.00206% while with information gain the accuracy went down by 0.34342%. We observed that our approach improved the performance of the naïve bayes classifier by 0.65864%. For salary ≥ 50 the classification accuracy for k=5 reduced by 0.96187%

and when k=5 with information gain the accuracy went down by 0.34351%. We noted an improvement in the classifier performance with our approach by 0.61836%.

For the Naïve logistic regression classifier when k=5 for salary ≤ 50, the accuracy went down by 1.65435% while with information gain the accuracy went down by 0.77863%. We observed that our approach improved the performance of the naïve bayes classifier by 0.87572%. For salary ≥ 50 the classification accuracy for k=5 reduced by 1.65415% and when k=5 with information gain the accuracy went down by 0.77872%. We noted an improvement in the classifier performance with our approach by 0.87543%.

With the Random Forest classifier when k=5 for salary ≤ 50, the accuracy went down by 4.52429% while with information gain the accuracy went down by -0.89638%. We observed that our approach improved the performance of the naïve bayes classifier by 5.42067%. For salary ≥ 50 the classification accuracy for k=5 reduced by 4.52333% and when k=5 with information gain the accuracy went down by -0.89624%. We noted an improvement in the classifier performance with our approach by 5.41957%.

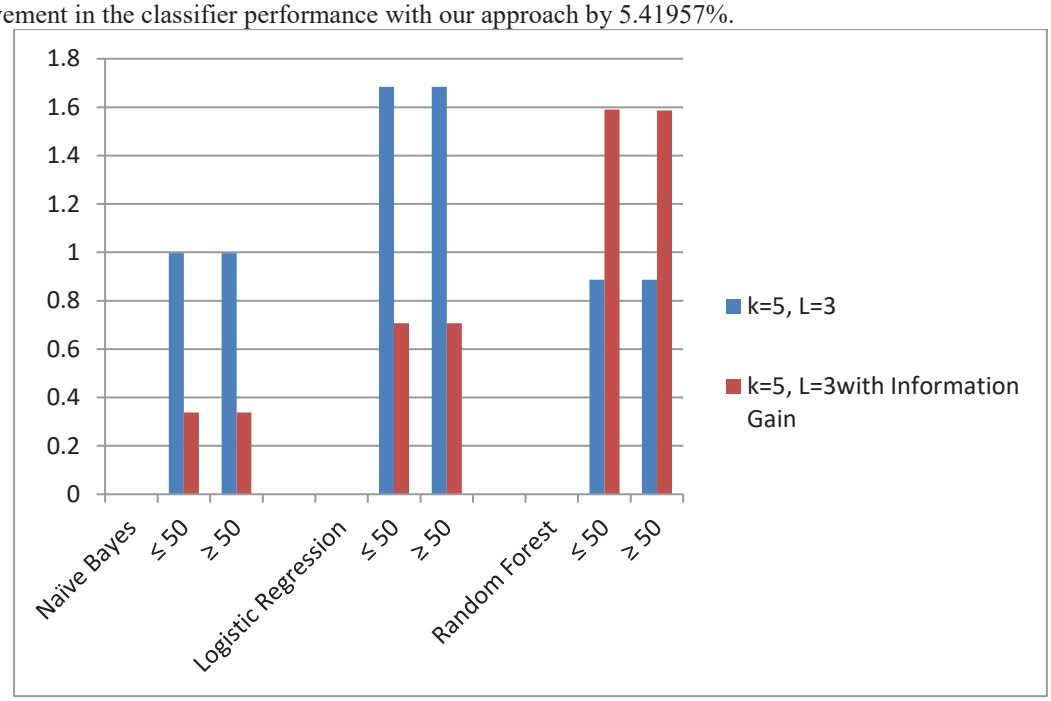


Figure 2: A comparison of loss in classifier accuracy for the three classifiers when k=5 and l=3 with and without information gain.

From figure 2 we noted that data anonymization for *k*=5 and *l*=3 with the information gain resulted in better classification accuracy with the naïve bayes and logistic regression classifiers than that for the three classifiers built from the dataset anonymized with *k*=5 and *l*=3 without the information gain.

## 6. Contribution

The contribution of this work includes the introduction of a new approach for anonymizing high dimensional data through a combination of feature selection with slicing, generalization and suppression; and an enhanced anonymization of high dimensional data with improved data utility and reduced data distortion.

## 7. Conclusions and Future Work

This paper focused on the anonymization of high dimensional data with a single sensitive attribute. Different anonymization approaches have different benefits and drawbacks. However, every approach must have the goal of preserving privacy without compromising the utility of the data much. This work proposed a privacy preservation approach that anonymizes high dimensional data while enhancing the quality of published data. For future work, we recommend that further research be undertaken in the area of privacy preservation for high dimensional data with multiple sensitive attributes.

## References

Abdou Hussien, A.-E.-E., Darwish, N. R., Hefny, H. A., & Hussien, A. A. (2015). Utility-Based Anonymization Using Generalization Boundaries to Protect Sensitive Attributes. *Journal of Information Security*, *6*(6), 179–196. https://doi.org/10.4236/jis.2015.63019

Ahmed Ali Mubark, Emad Elabd, H. A. (2016). Semantic Anonymization in publishing Categorical Sensitive Attributtes, 89–95.

Basso, T., Matsunaga, R., Moraes, R., & Antunes, N. (2016). Challenges on anonymity, privacy, and big data. In *Proceedings - 7th Latin-American Symposium on Dependable Computing, LADC 2016* (pp. 164–171). https://doi.org/10.1109/LADC.2016.34

Bhaladhare, P., & Jinwala, D. (2012). A Sensitive Attribute Based Clustering Method for k-Anonymization. *Springer-Verlag Berlin Heidelberg*, 163–170.

Canbay, P., & Sever, H. (2015). The Effect of Clustering on Data Privacy. In *2015 IEEE 14th International Conference on Machine Learning and Applications (ICMLA)* (pp. 277–282). https://doi.org/10.1109/ICMLA.2015.198

Chiu, C., & Tsai, C. (2007). A k -Anonymity Clustering Method for Effective Data, 89–99. https://doi.org/10.1007/978-3-540-73871-8

Dheeru, Dua , Karra Taniskidou, E. (2017). {UCI} Machine Learning Repository. Retrieved from http://archive.ics.uci.edu/ml

Doka, K., Xue, M., Tsoumakos, D., Karras, P., Cuzzocrea, A., & Koziris, N. (2015). Heterogeneous k - Anonymization with High Utility, 1886–1890. https://doi.org/10.1109/BigData.2015.7363963

Dwork, C. (2006). Differential Privacy. *Proceedings of the International Colloquium on Automata, Languages and Programming, Part II (ICALP)*, 1–12. https://doi.org/10.1007/11787006

Dwork, C., & Roth, A. (2013). The Algorithmic Foundations of Differential Privacy. *Foundations and Trends® in Theoretical Computer Science*, *9*(3–4), 211–407. https://doi.org/10.1561/0400000042

Fawcett, T. (2004). ROC Graphs : Notes and Practical Considerations for Researchers. *ReCALL*, *31*(HPL-2003-4), 1–38. https://doi.org/10.1.1.10.9777

Fienberg, S. E., & Mcintyre, J. (2005). Data Swapping : Variations on a Theme by Dalenius and Reiss, *21*(2), 309–323.

Fung, B. C. M., Wang, K., & Yu, P. S. (2005). Top-down specialization for information and privacy preservation. In *Proceedings - International Conference on Data Engineering* (pp. 205–216). https://doi.org/10.1109/ICDE.2005.143

Gal, T. S., Tucker, T. C., Gangopadhyay, A., & Chen, Z. (2014). A data recipient centered de-identification method to retain statistical attributes. *Journal of Biomedical Informatics*, *50*, 32–45. https://doi.org/10.1016/j.jbi.2014.01.001

Garcia-Alfaro, J., Herrera-Joancomartí, J., Lupu, E., Posegga, J., Aldini, A., Martinelli, F., & Suri, N. (2015). Data privacy management, autonomous spontaneous security, and security assurance. *Lecture Notes in Computer Science (Including Subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, *8872*, 266–276. https://doi.org/10.1007/978-3-319-17016-9

Ghinita, G., Karras, P., Kalnis, P., & Mamoulis, N. (2007). Fast data anonymization with low information loss. *Proceedings of the 33rd International Conference on Very Large Data Bases*, 758–769. Retrieved from http://dl.acm.org/citation.cfm?id=1325938%5Cnhttp://citeseerx.ist.psu.edu/viewdoc/summary?doi=10.1.1.138.3217

H. Liu, H. M. (1998). Feature Selection for Knowledge Discovery and Data Mining. *Kluwer Academic*.

Hall, M. A. (2000). Feature Selection for Discrete and Numeric Class Machine Learning 1 Introduction. *Machine Learning Proc Seventeenth International Conference on Machine Learning*, 1–16. https://doi.org/10.1.1.34.4393

Han, J. (2013). SLOMS : A Privacy Preserving Data Publishing Method for Multiple Sensitive Attributes Microdata. *Journal of Software*, *8*(12), 3096–3104. https://doi.org/10.4304/jsw.8.12.3096-3104

Han, J., Yu, H., & Yu, J. (2008). An improved l-diversity model for numerical sensitive attributes. In *3rd International Conference on Communications and Networking in China, ChinaCom 2008* (pp. 937–942). https://doi.org/10.1109/CHINACOM.2008.4685178

Huang, X., Liu, J., Han, Z., & Yang, J. (2014). A new anonymity model for privacy-preserving data publishing. *China Communications*, *11*(9), 47–59. https://doi.org/10.1109/CC.2014.6969710

Iyengar, V. S. (2002). Transforming data to satisfy privacy constraints. In *Proceedings of the eighth ACM SIGKDD international conference on Knowledge discovery and data mining - KDD '02* (p. 279). https://doi.org/10.1145/775047.775089

Kifer, D., & Gehrke, J. (2006). Injecting utility into anonymized datasets. In *Proceedings of the 2006 ACM SIGMOD international conference on Management of data - SIGMOD '06* (p. 217). https://doi.org/10.1145/1142473.1142499

Lee, H., Kim, S., Kim, J. W., & Chung, Y. D. (2017). Utility-preserving anonymization for health data publishing. *BMC Medical Informatics and Decision Making*, *17*(1), 1–12. https://doi.org/10.1186/s12911-017-0499-0

Li, T., Li, N., Zhang, J., & Molloy, I. (2012). Slicing: A new approach for privacy preserving data publishing. *IEEE Transactions on Knowledge and Data Engineering*, *24*(3), 561–574. https://doi.org/10.1109/TKDE.2010.236

Li, W., Hu, C., Song, T., Yu, J., Xing, X., & Cai, Z. (2018). Privacy-Preserving Data Collection in Context-Aware

Applications. *2018 IEEE Symposium on Privacy-Aware Computing (PAC)*, 75–85. https://doi.org/10.1109/PAC.2018.00014

Liu, X., Xie, Q., & Wang, L. (2015). A Personalized Extended (α, k)-Anonymity Model, *1*, 234–240. https://doi.org/10.1109/CBD.2015.45

Machanavajjhala, A., Kifer, D., Gehrke, J., & Venkitasubramaniam, M. (2007). *L* -diversity. *ACM Transactions on Knowledge Discovery from Data*, *1*(1), 3–es. https://doi.org/10.1145/1217299.1217302

Miao, J., & Niu, L. (2016). A Survey on Feature Selection. *Procedia Computer Science*, *91*(Itqm), 919–926. https://doi.org/10.1016/j.procs.2016.07.111

Mivule, K., & Turner, C. (2013). A Comparative Analysis of Data Privacy and Utility Parameter Adjustment , Using Machine Learning Classification as a Gauge. *Procedia - Procedia Computer Science*, *20*, 414–419. https://doi.org/10.1016/j.procs.2013.09.295

Mohanapriya, D. (2013). Slicing Technique For Privacy Preserving Data Publishing. *International Journal of Computer Trends and Technology (IJCTT)*, *4*(5), 1355–1361. https://doi.org/10.1063/1.2198933

Ninghui, L., Tiancheng, L., & Venkatasubramanian, S. (2007). t-Closeness: Privacy beyond k-anonymity and ℓ-diversity. In *Proceedings - International Conference on Data Engineering* (pp. 106–115). https://doi.org/10.1109/ICDE.2007.367856

Prasser, F., Eicher, J., Bild, R., Spengler, H., & Kuhn, K. A. (2017). A Tool for Optimizing De-identified Health Data for Use in Statistical Classification. *Proceedings - IEEE Symposium on Computer-Based Medical Systems*, *2017–June*. https://doi.org/10.1109/CBMS.2017.105

Prasser, F., & Kohlmayer, F. (2015). Medical Data Privacy Handbook. *Springer International Publishing Switzerland 2015*. https://doi.org/10.1007/978-3-319-23633-9

Qinbao Song, Jingjie Ni, & Guangtao Wang. (2013). A Fast Clustering-Based Feature Subset Selection Algorithm for High-Dimensional Data. *IEEE Trans. Knowl. Data Eng.*, *25*(1), 1–14. https://doi.org/10.1109/TKDE.2011.181

Samarati, P., & Sweeney, L. (1998). Protecting Privacy when Disclosing Information: k-Anonymity and its Enforcement Through Generalization and Suppresion. *Proc of the IEEE Symposium on Research in Security and Privacy*, 384–393. https://doi.org/http://dx.doi.org/10.1145/1150402.1150499

Shaltout, N. A. N., El-Hefnawi, M., Rafea, A., Moustafa, A., & El-Hefnawi, M. (2014). Information gain as a feature selection method for the efficient classification of Influenza-A based on viral hosts. *Proceedings of the World Congress on Engineering*, *1*, 625–631. Retrieved from http://www.scopus.com/inward/record.url?eid=2-s2.0-84907414411&partnerID=tZOtx3y1

Sharma, V. K., & Bala, A. (2014). Clustering for high dimensional data. *1st International Conference on Networks and Soft Computing, ICNSC 2014*, 365–369. https://doi.org/10.1109/CNSC.2014.6906700

Smith, S. W. (2006). Chapter 4, 1–20. https://doi.org/10.1007/978-0-387-77379-7

Soria-comas, J., & Domingo-ferrer, J. (2013). Differential Privacy via t -Closeness in Data Publishing. *Privacy, Security and Trust (PST), 2013 Eleventh Annual International Conference On*, 27–35.

Sunitha, a., Venkata Subba Reddy, K., & Vijayakumar, B. (2012). A Privacy Measure for Data Disclosure to Publish Micro Data using (N,T) - Closeness. *International Journal of Computer Applications*, *51*(6), 22–28. https://doi.org/10.5120/8047-1379

Sweeney, L. (2002a). Achieving K -Anonymity Privacy Protection Using Generalization And Suppression. *International Journal on Uncertainty, Fuzziness and Knowledge-Based Systems*, *10*(5), 1–18. https://doi.org/10.1142/S021848850200165X

Sweeney, L. (2002b). k-Anonymity: A Model for Protecting Privacy. *International Journal of Uncertainty, Fuzziness and Knowledge-Based Systems*, *10*(05), 557–570. https://doi.org/10.1142/S0218488502001648

Tan, P.-N., Steinbach, M., & Kumar, V. (2006). Classification : Basic Concepts , Decision Trees , and. *Introduction to Data Mining*, *67*(17), 145–205. https://doi.org/10.1016/0022-4405(81)90007-8

Wagner, I., & Eckhoff, D. (2018). Technical Privacy Metrics : a Systematic Survey arXiv : 1512 . 00327v2 [ cs . CR ] 20 Jun 2018. *ACM Computing Surveys*, *51*(3), 1–45.

Xiao, X., & Tao, Y. (2006a). Anatomy: Privacy and Correlation Preserving Publication. *Citeseer*, (i), 1–47. Retrieved from http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.141.1598&rep=rep1&type=pdf

Xiao, X., & Tao, Y. (2006b). Anatomy: Simple and effective privacy preservation. *Proceedings of the 32nd International Conference on Very Large Database, ACM*, 150, 139. Retrieved from http://portal.acm.org/citation.cfm?id=1164127.1164141