Non-homogenous Slicing Anonymization with Subsequent Data utility Analysis for Privacy Preservation Data mining

Rajul Jain, Pranjali Singh

Abstract—With propulsion in the amount of data processed and released every day, privacy and security have become an indispensable factor in the data sphere. But data privacy and data utility seem to be in a constant tug-of-war with each other, with one factor having to compromise for the other. But if either utility or privacy is deprioritized beyond a certain point then the data might be rendered as either useless or vulnerable to severe privacy breaches. For this reason, it is essential to publish data in such a way that individual privacy remains intact, and the data is still useful for knowledge discovery, which is the main agenda behind Privacy-Preserving Data Mining (PPDM). This paper proposes a refinement of an existing PPDM technique known as slicing anonymization. Slicing has been previously proven to be an efficient technique for preserving the high quality of data while achieving high data privacy in publishing. In this paper, we target higher data utility and more secure data publishing using the concepts of probabilistic non-homogenous suppression and attribute correlation. We also validate the results by comparing the pre-defined data quality metrics of the most used classification algorithms before and after applying this technique on the candidate dataset obtained from the Madhya Pradesh State Election Commission (MPSEC). The closeness of the results proves that our proposed algorithm maintains high data quality and ensures strong privacy preservation at the same time.

Index Terms— Data mining, Data utility, non-homogenous slicing anonymization, Privacy preserving data mining, Slicing

I. INTRODUCTION

In the current data-driven world- in order to realize the full potential of data which is an extremely useful resource-companies and individuals often unknowingly exchange and publish information that is sensitive. This data is open to some serious privacy and security breaches. As of now, there exists a plethora of privately outsourced or publicly published data exposing people's finances, interests, background, health, and demographics. For example, cancer researchers often seek patient's diagnostic data to analyze it and consequently, research on a potential cure. However, this data is sensitive and can be used to carry out deleterious activities. This is where Privacy-Preserving Data Publishing (PPDP) comes into play. Privacy-Preserving Data Publishing (PPDP) algorithms are not tailor-made for specific mining tasks and more often than not, the resultant anonymized data loses its usefulness.

PPDM techniques make sure that data is published in such a way that there is no risk on individual privacy, and, at the same time, ensuring that the data isn't distorted beyond certain pre-defined metrics and its utility/ effectiveness remains intact as much as possible. Some of these techniques are applied in the Data Collection phase, which transforms the data by adding random noise while keeping the data distribution intact, so that, at the time of mining, individual values of sensitive attributes aren't exposed but the statistical distribution of the data can be reconstructed. Other techniques sanitize data before publishing it. These sanitization/redaction techniques are carried out differently for different data mining tasks. For example, if the cancer diagnostic data needs to be fed into a Support Vector Machine, the corresponding PPDM algorithm will aim at achieving anonymization while incurring a minimal loss of accuracy of the resulting classifier.

II. RELATED WORK

A considerable amount of work is done towards defining the perfect trade-off between retaining the utility of data while protecting private information. There are various ways datasets are mined depending upon the work that needs to be accomplished. Employing classification algorithms is one such way, which we intend to do in this paper. However, as also mentioned in reference (Sweeney, 2002) in order to prevent a potential breach of privacy, the accuracy of the algorithms needs to be compromised to a certain extent.

For privacy preserving, a number of algorithms have been developed. In reference (Sweeney, 2002), the author has introduced the k-anonymity model which aims at making a tuple indistinguishable from at least k-1 tuples. This model formed the underlying basis of many systems guaranteeing privacy protection. However, it fails at preventing attacks due to record and linkage of attributes.

The k-anonymity model was further refined in reference (J. Li, Wong, Fu, & Pei, 2006) by basing it on clustering. The authors experimentally proved that the resultant algorithm increased the scalability and significantly reduced the distortions when data is k-anonymized by defining a distance

Manuscript revised on September 7, 2019 and published on September 21, 2019

Rajul Jain*, Department of Computer Science and Engineering, Maulana Azad National Institute of Technology, Bhopal, India*
Pranjali Singh*, Department of Computer Science and Engineering, Maulana Azad National Institute of Technology, Bhopal, India*

doi: 10.32622/ijrat.78201919

metric to measure the generalization distances between tuples in the resultant data.

In reference (Machanavajjhala, Gehrke, Kifer, & Venkitasubramaniam, n.d.) the shortcoming of the k-anonymity model in preventing linkage attacks was overcome by the proposed named l-diversity, which puts a restriction on the minimum number of distinct values of the sensitive attribute in a given equivalence class. However, this model too has its own limitation and fails in cases data is homogenous or attacked has prior background knowledge.

In reference (Friedman, Wolff, & Schuster, 2008), the authors define an extension to the definition of the k-anonymity model by describing data mining algorithms generating output adhering to the policies of the k-anonymized model. The proposed model while efficiently anonymizing data also preserves patterns in the original data while doing so.

In reference (Kisilevich, Elovici, Shapira, & Rokach, 2009), the authors proposed swapping in place of suppression in k-anonymity, thus reducing the loss of information induced by the former approach. This method also gives better performance when used with classification algorithms for predictive analysis than existing methods.

The authors in reference (T. Li, Li, Zhang, & Molloy, 2012) introduce a new approach for publishing data while preserving privacy, known as Slicing, which holds significant advantages over standard techniques like bucketization and generalization in terms of both data utility as well as privacy preservation. It is in accordance with the principles of l-diversity and can handle high-dimensional data with great accuracy, all while preventing attacks like membership disclosure. It does so by partitioning data both horizontally and vertically.

In reference (Dwork, 2006), Cynthia Dwork proposed a novel model called ϵ **-differential privacy** which provides complete privacy protection for statistical databases. Differential Privacy does not expose individual data in the released dataset, but, makes global statistical information about the data available to the public, thus maintaining its utility while ensuring privacy.

As described in reference (Yu, 2016), combining these privacy models and improving upon them in order to overcome their limitations is now the main research branch in PPDM, and, a lot of work still needs to be accomplished towards keeping data utility intact while maintaining data privacy.

Different classification algorithms are used in different scenarios and with different datasets having varied goals. The most commonly researched and used classification algorithms on sanitized data are Naive Bayes and decision tree, both of which are modified to suit classification needs and improve accuracy with sanitized data. In reference (Yang, Zhong, & Wright, 2005), the author has employed frequency mining on top of Naive Bayes to achieve higher accuracy. In reference (Agrawal & Srikant, 2000), the authors have added Gaussian noise to perturb the dataset before feeding to decision tree classifier.

III. CONCEPT AND DEFINITIONS

This work's objective is to render the privacy in data publishing while maintaining high data quality in terms of its usefulness.

1. Data privacy

While publishing any dataset, publishers come across four types of attributes, which are: Key Identifier attributes (id), which have the ability of uniquely identifying an individual tuple in the dataset; Quasi-Identifiers (Q-Id), which do not individually identify a row but when combined with other Q-ids, might lead to unique identification and consequently a privacy breach; Sensitive attributes which hold private information of a person; and other non-sensitive attributes which neither represent sensitive information, nor are they easily accessible to the attacker, and thus do not reveal individual identity.

Some of the most prevalent Privacy-preserving techniques:

1. Generalization: replacement of exact values by more general values.
   1.1 Homogenous Generalization: replacement of exact values by more general values in all the cells of the target attribute. (Usha, Shriram, & Sathishkumar, 2015)
   1.2 Non-homogenous Generalization: replacement of exact values by more general values in selective cells of the target attribute. (Usha et al., 2015)
2. Suppression: complete removal of exact values and replace with some other symbol.
   2.1 Homogenous Suppression: All the cell values of the targeted attribute are suppressed uniformly. [12]
   2.2 Non-Homogenous Suppression: Some cell values of the targeted attribute are either unsuppressed or suppressed variably. [12]
   2.3 Probabilistic non-homogenous suppression: In this paper, the method of probabilistic non-homogenous suppression is proposed. In a bivariate/ multivariate attribute, first the different values are identified for that attribute and the probability of occurrence of each value is calculated. The suppression is then applied randomly based on this probabilistic data, i.e. higher the probability of occurrence of that value in that column, higher the chances of it getting suppressed. There also exists a pre-decided or randomly generated limit factor of suppression for this dataset, i.e. the number of values that are supposed to be suppressed. For example, consider the bivariate 'Marital status' attribute from Table-1. On analysis in Fig-x, we found that, around 2% of the candidates are 'Unmarried' and the rest 98% are 'Married'. Hence if probabilistic non-homogenous suppression is applied, on say 100 tuples, with the limit factor of 60 tuples, then in all these 60 'Marital status' suppressed tuples, only about 10-12 tuples will be of 'Unmarried' status and the rest of 'Married' status. Using this technique, skewness attacks (Rohilla, 2015) on the published data can be largely reduced.

doi: 10.32622/ijrat.78201919

3. Perturbation: replacement of the original values with randomly generated values having similar statistical information.
   There exist many such algorithms which usually use a combination of these techniques to ensure privacy. For example, k-anonymity uses generalisation and suppression. Differential privacy uses the technique of perturbation for safeguarding the data.
4. Slicing: an effective anonymization technique to handle high dimensional data. It efficiently handles drawbacks of generalisation and suppression i.e. aims for high data utility. Highly correlated values are grouped together in a column, meanwhile un-correlated attributes concatenated in another by vertical partitioning, which increases the privacy of the data to be published. [13]

2. Data utility

The goal is to check the effectiveness of the resultant data. For achieving the same, standard supervised data mining algorithms and their metrics are used.

Classification is a supervised learning approach with the goal of building models which can predict the value of the class label attribute, by means of other attribute's values. Classification algorithms hold extensive applications in a multitude of sectors and often deal with sensitive data. For example, based on financial, criminal and travel data, one may want to classify passengers as a security risk.

A typical classification algorithm consists of a training phase in which the model is trained on a part of data to assess the relationship between the class label attribute and other attributes, and a testing phase in which the computed relationship is put to a test to predict the values of the class label attribute of the remaining part of the dataset using the corresponding values of other columns.

We used different standard classification algorithms on our dataset before and after applying privacy preservation techniques and compared the results obtained in both these cases.

2.1. Decision tree classification

Decision tree recursively breaks down complex data into smaller subsets while, at the same time, developing an associated decision tree incrementally. The resultant tree's leaf node represents the computed class, while the root node corresponds to the predictor. It can handle both categorical and numerical data quite accurately.

2.2. Logistic Regression

A statistical method of analyzing data, consisting of one or more independent variables to determine an outcome. The outcome is measured with a dichotomous variable (having only 2 possible values), in the case of binary logistic regression and with a non-dichotomous variable in the case of multinomial logistic regression. In this algorithm, the probabilities are computed describing the possible outcomes of a single trial using a logistic function.

2.3. K-Nearest neighbors

The KNN algorithm computes the logical distance of all the data points from the one whose value it wants to predict using standard distance functions and predicts the value of the class label attribute as the value held by a majority of the k nearest neighbors from that data point. This is lazy algorithm which spends little to no time in the training phase and directly jumps to classification, which makes it costly in terms of both computation time and space.

2.4. Naive Bayes

This classification technique is based on Bayes' Theorem and assumes that different features in the dataset are not dependent on each other and contribute independently to the probability determining the value of the class label attribute. Naive Bayes model is fairly easy to build and has great usefulness with large datasets, while at the same time performing better than many complex algorithms.

2.5. Support Vector Machine

This algorithm is extensively used in regression analysis and for classification. SVM plots each data point in an n-dimensional space and classifies the points by finding the most appropriate differentiating hyperplane.

2.5.1. Kernels in SVM

There are some datasets for which it is not possible to find a linear separating hyperplane. To handle such cases, certain mathematical functions called kernels are used to increase the dimensionality of the data points, thus making it possible to find a separating hyperplane.

There are various types of kernels available, like linear, polynomial, RBF, etc. The linear kernel, used in this paper, linearly splits the actual hypothesis into linear functions to yield a higher dimensional linear separating hyperplane.

2.6 Confusion Matrix

A confusion matrix is a table that is popularly used to assess the performance of a classifier by organizing the counts of correctly and incorrectly predicted values over the test dataset into a table. From the confusion matrix, we generate a classification report consisting of the following metrics:

- Precision (P) is the proportion of the predicted positive cases that were actually positive.
- Recall: It is the proportion of positive cases that were correctly identified.
- F1- score: The F_1 score is a weighted harmonic mean of precision and recall such that the best score is 1.0 and the worst is 0.0.
- Support: It is the number of instances of a particular class in the actual dataset.

IV. Data Analysis

The data was collected from the Madhya Pradesh State Election Commission, Bhopal (MPSEC). This data contains the information of the candidates contending from their respective districts for the position of 'Adhyaksh' and 'Parshad' for the 2016 State elections. It consisted of 74,000 tuples and 30 tuples, out of which we have identified the 10 relevant tuples for PPDP.

doi: 10.32622/ijrat.78201919

1. The data from Table-1 has been identified and categorized on the basis of attributes:
   - □ Key identifier: Auto-ID
   - □ Quasi-Identifier: Age, Gender, Marital Status, Category
   - □ Sensitive Attribute: Votes
   - □ Non-sensitive attributes: Education qualification, occupation, position, District code

2. Attribute specific analysis:

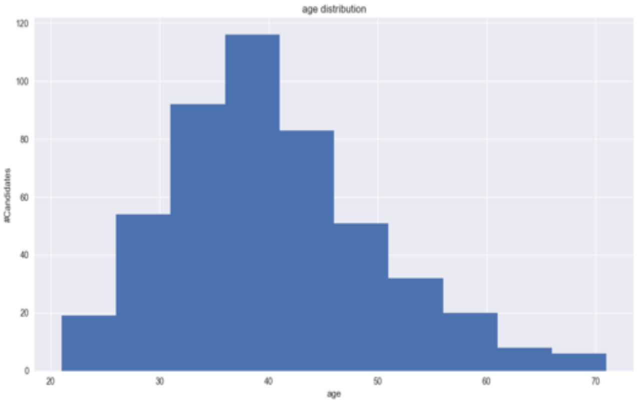


Fig. 1. Distribution of attribute age in the data set

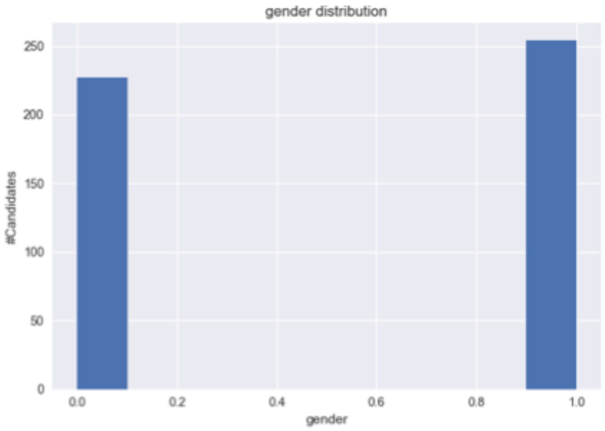


Fig. 2. Distribution of gender (0: Female; 1: Male)

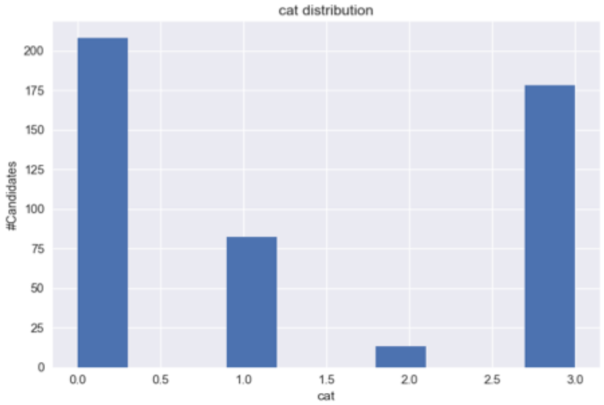


Fig. 3. Distribution of Category
(0: OBC; 1: SC; 2: ST; 3: General)

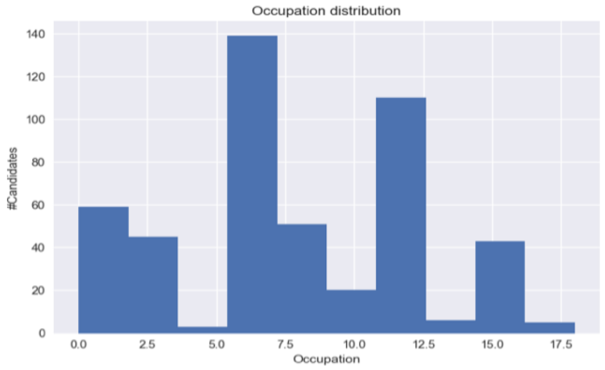


Fig .4. Distribution of Occupation (Advocate, Labour, Judge, Business, Service, Unemployed, Farmer, Pension, Housewife, Student)

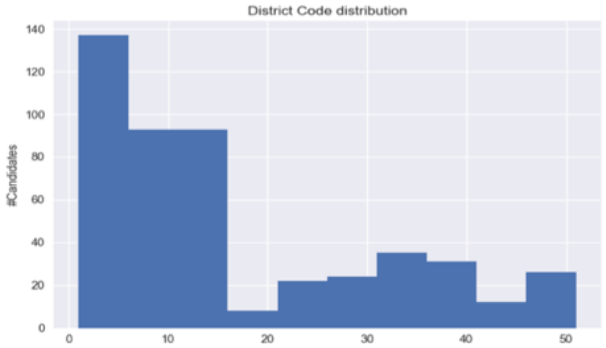


Fig .5. Distribution of District code (numeric)

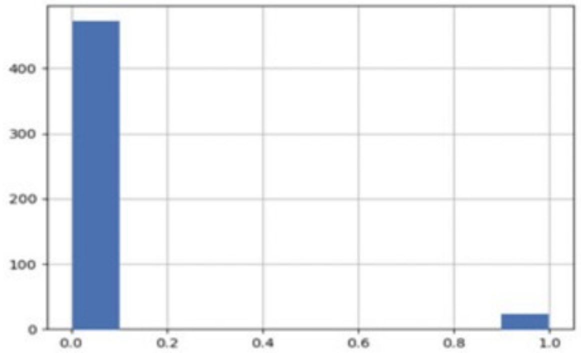


Fig. 6. Distribution of Marital Status
(0: Married; 1: Unmarried)

From the above graphs, we can conclude:

- □ More than 75% of the candidates lie between the age ranges of 20-50. In other words, less than 25% of the candidates are above the age of 50. (Fig 1)
- □ The number of male candidates are more than the number of female candidates. (Fig 2)
- □ More than 75% of candidates belong to either Other Backward Castes (OBC) or General category. This means that fewer than 25% of all the candidates belong to Scheduled Caste (SC) or Scheduled Tribe (ST). (Fig 3)
- □ Most of the candidates contesting are Farmers and Businessmen, followed by Advocates, Laborers, Servicemen, and Housewives. (Fig 4)
- □ Distribution of candidates is considerably concentrated to a few districts only, while other

doi: 10.32622/ijrat.78201919

districts see just a meagre number of contestants. (Fig 5)

- □ Number of married candidates exceed the unmarried candidates by a largely significant number. (Fig 6)

3. Correlation of other attributes with the sensitive attribute:

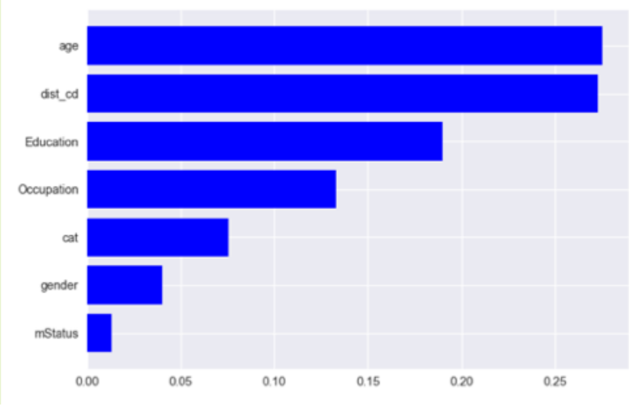
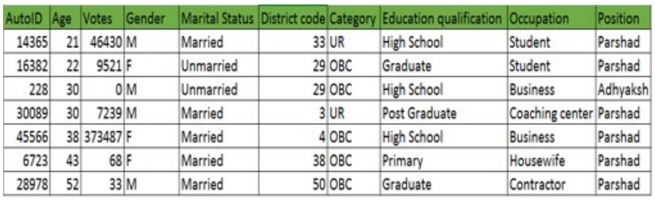


Fig. 7. Dependency of other attributes over votes

From Fig 7 we can conclude that the correlation of votes with Gender attribute and marital status attribute is the least. Thus, we can safely apply privacy preservation techniques on these attributes without degrading data utility.

Also, age and district code are most closely related to the votes attribute.

TABLE I: MPSEC Dataset Table



AutoID	Age	Votes	Gender	Marital Status	District code	Category	Education qualification	Occupation	Position
14365	21	46430	M	Married	33	UR	High School	Student	Parshad
16382	22	9521	F	Unmarried	29	OBC	Graduate	Student	Parshad
228	30	0	M	Unmarried	29	OBC	High School	Business	Adhyaksh
30089	30	7239	M	Married	3	UR	Post Graduate	Coaching center	Parshad
45566	38	373487	F	Married	4	OBC	High School	Business	Parshad
6723	43	68	F	Married	38	OBC	Primary	Housewife	Parshad
28978	52	33	M	Married	50	OBC	Graduate	Contractor	Parshad

V. Methodology

A. Vertical Partitioning

The original dataset table is divided into two tables with Key identifier as the primary key for both the tables. This reduces the amount of data to be processed, consequently the performance is increased. Dataset-1 contains quasi-identifiers, sensitive attribute along with the key identifier and Dataset-2 contains non-sensitive attributes and key identifier. Sanitization is performed on Dataset-1.

B. Algorithm

Input: Record set to be released, dependency list
Output: Anonymized record set

```
▪ Input: A dataset D [quasi-identifier attributes Q,
Sensitive values A], correlation list L, limit factor f.
▪ Output: Anonymized Dataset D*.

Begin
1. Select Data set D from a Database.
2. Select Q* as the quasi identifier having maximum
correlation on A from L.
3. Select Q', Q'',..$Q^n$ as the quasi identifier having minimum
correlation on A from L.
4. For each tuple in D, replace {A}, {Q*}, {Q'}, {Q''},... {
$Q^n$} with
{A,Q*}, Q'
5. For each tuple in D, concatenate {A}, {Q*}, {Q'},
{Q''},..{ $Q^n$ }
{A,Q*}, {Q', Q'',.. $Q^n$ }
6. Apply Homogenous Generalisation or Probabilistic
Suppression (f) on attributes in {A, Q*}, {Q', Q'',.. $Q^n$}
7. Publish final Dataset D*.
End
```

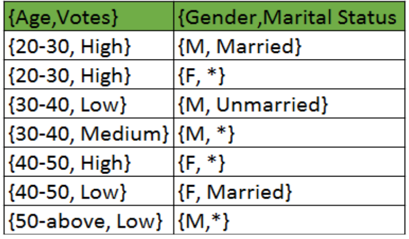
VI. Results

Table-II is obtained on application of the proposed algorithm on Dataset-1. For PPDM purpose, Dataset-1 and Dataset-2 are combined and Table-III is obtained.

In this resultant anonymized dataset, different classification algorithms are applied viz. Decision tree classifier, Naïve-Bayes, Support Vector Machine, Logistic regression and K-nearest neighbors. The anonymized dataset is fed into these classifiers for training the classifiers. After this, for testing the remaining data is used to check the accuracy percentage and data quality metrics of the testing data on these classifier algorithms, i.e. when other attributes in the given data are fed into the classifier, how accurately these algorithms predict votes. All these algorithms and metrics are implemented in python using the scikit-learn library (Pedregosa FABIANPEDREGOSA et al., 2011).

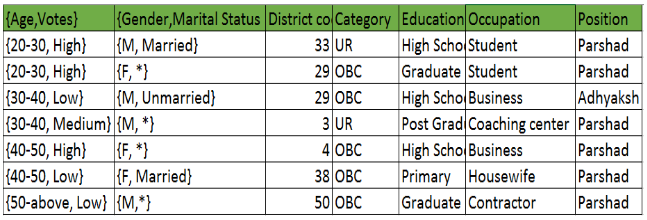
For comparative analysis, the data quality metrics of non-sanitized data for the same classification algorithm has been shown in Fig. 8. After anonymization the results obtained are shown in Fig. 9. Table IV compiles the results from the two graphs, comparing the accuracy of the above stated algorithms in both the cases. Table V, Table VI, and Table VII show the confusion matrices obtained when applied Logistic Regression, Support Vector Machine (Linear Kernel), Naïve Bayes Classifier respectively, on the resultant anonymized dataset.

TABLE II: Anonymized Table



{Age,Votes}	{Gender,Marital Status
{20-30, High}	{M, Married}
{20-30, High}	{F, *}
{30-40, Low}	{M, Unmarried}
{30-40, Medium}	{M, *}
{40-50, High}	{F, *}
{40-50, Low}	{F, Married}
{50-above, Low}	{M,*}

TABLE III: Published MPSEC Dataset Table



{Age,Votes}	{Gender,Marital Status	District co	Category	Education	Occupation	Position
{20-30, High}	{M, Married}	33	UR	High Scho	Student	Parshad
{20-30, High}	{F, *}	29	OBC	Graduate	Student	Parshad
{30-40, Low}	{M, Unmarried}	29	OBC	High Scho	Business	Adhyaksh
{30-40, Medium}	{M, *}	3	UR	Post Grad	Coaching center	Parshad
{40-50, High}	{F, *}	4	OBC	High Scho	Business	Parshad
{40-50, Low}	{F, Married}	38	OBC	Primary	Housewife	Parshad
{50-above, Low}	{M,*}	50	OBC	Graduate	Contractor	Parshad

doi: 10.32622/ijrat.78201919

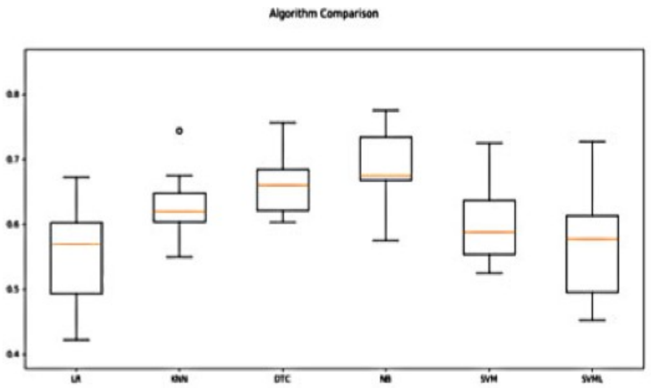


Fig. 8. Box plot depicting accuracy results of the algorithms for non-anonymized data

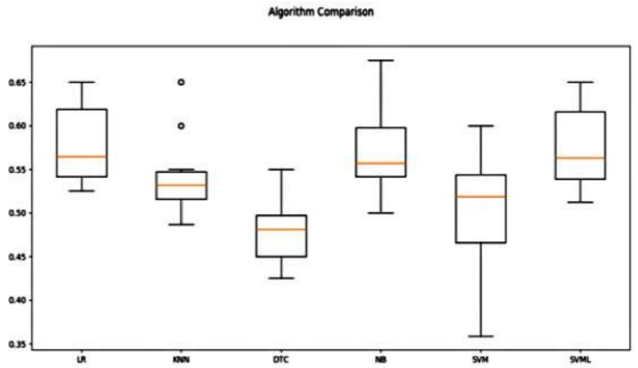


Fig. 9. Box plot depicting accuracy results of the algorithms for anonymized data

TABLE IV: Summarizing the accuracy percentages for non-sanitized and sanitized data

Model	Non-Sanitized MPSEC Data Votes Accuracy(%)	Sanitized MPSEC Data Votes Accuracy(%)
Logistic Regression	58.2	57.8
K Nearest Neighbours	62.6	54.2
Decision Tree Classifier	65.3	46.9
Naïve Bayes	68.8	57
Support Vector Machine	60.1	50.1
SVM Linear Kernel	58.7	57.5

TABLE V: Confusion matrix for sanitized data using Logistic Regression

Votes Range/ Metric	Precision	Recall	F1-score	Support
Low	0.71	0.28	0.4	18
Medium	0.73	0.67	0.7	45
High	0.48	0.68	0.56	37
Avg/total	0.64	0.6	0.59	100

TABLE VI: Confusion matrix for sanitized data using SVM Linear Kernel

Votes Range/ Metric	Precision	Recall	F1-score	Support
Low	0.75	0.17	0.27	18
Medium	0.74	0.62	0.67	45
High	0.48	0.76	0.59	37
Avg/total	0.65	0.59	0.57	100

TABLE VII: Confusion matrix for sanitized data using Naïve Bayes Classifier

Votes Range/ Metric	Precision	Recall	F1-score	Support
Low	0.45	0.28	0.34	18
Medium	0.75	0.6	0.67	45
High	0.47	0.68	0.56	37
Avg/total	0.59	0.57	0.57	100

VII. Conclusion And Future Scope

Many techniques exist for privacy preservation in data publishing, the choice of which depends on various factors like data type, the end goal of publishing, amount of data, etc. On the Madhya Pradesh State Election Commission data, we applied non-homogenous slicing for PPDM purpose, i.e. with a specific goal in mind that the resultant anonymized data can be used for multifarious purposes like problem identification and its subsequent trouble-shooting – if possible. This data can be further combined with other public data like census data for better insights. For example, to understand the pattern in the percentage of people voted from particular community in a specific district.

Apart from generalization and suppression-based anonymization, techniques like perturbation can also be used for PPDP purposes. In our case, where we have used suppression, the technique of differential privacy can also be applied to get effective results.

doi: 10.32622/ijrat.78201919

Acknowledgment

We are extremely grateful towards the Madhya Pradesh State Election Commission and Dr. Manasi Gyanchandani for providing us with this real dataset.

References

[1] Agrawal, R., & Srikant, R. (2000). Privacy-preserving data mining. *Proceedings of the 2000 ACM SIGMOD International Conference on Management of Data - SIGMOD '00*, 439–450. https://doi.org/10.1145/342009.335438

[2] Dwork, C. (2006). LNCS 4052 - Differential Privacy. *Automata, Languages and Programming*, *33*, 1–12. https://doi.org/10.1007/11787006

[3] Friedman, A., Wolff, R., & Schuster, A. (2008). Providing k-anonymity in data mining. *VLDB Journal*, *17*(4), 789–804. https://doi.org/10.1007/s00778-006-0039-5

[4] Kisilevich, S., Elovici, Y., Shapira, B., & Rokach, L. (2009). KACTUS 2: Privacy preserving in classification tasks using k-anonymity. *Lecture Notes in Computer Science (Including Subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, *5661 LNCS*, 63–81. https://doi.org/10.1007/978-3-642-10233-2_7

[5] Li, J., Wong, R. C.-W., Fu, A. W.-C., & Pei, J. (2006). *Achieving k-Anonymity by Clustering in Attribute Hierarchical Structures*. https://doi.org/10.1007/11823728_39

[6] 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

[7] Machanavajjhala, A., Gehrke, J., Kifer, D., & Venkitasubramaniam, M. (n.d.). *ℓ-Diversity: Privacy Beyond k-Anonymity*.

[8] Pedregosa FABIANPEDREGOSA, F., Michel, V., Grisel OLIVIERGRISEL, O., Blondel, M., Prettenhofer, P., Weiss, R., … Duchesnay EDOUARDDUCHESNAY, Fré. (2011). Scikit-learn: Machine Learning in Python Gaël Varoquaux Bertrand Thirion Vincent Dubourg Alexandre Passos PEDREGOSA, VAROQUAUX, GRAMFORT ET AL. Matthieu Perrot. In *Journal of Machine Learning Research* (Vol. 12). Retrieved from http://scikit-learn.sourceforge.net.

[9] Rohilla, S. (2015). Privacy Preserving Data Publishing through Slicing. *American Journal of Networks and Communications*, *4*(3), 45. https://doi.org/10.11648/j.ajnc.s.2015040301.18

[10] Sweeney, L. (2002). L. Sweeney. k-anonymity: a model for k-ANONYMITY: A MODEL FOR PROTECTING PRIVACY 1. In *International Journal on Uncertainty, Fuzziness and Knowledge-based Systems* (Vol. 10).

[11] Usha, P., Shriram, R., & Sathishkumar, S. (2015). Sensitive attribute based non-homogeneous anonymization for privacy preserving data mining. *2014 International Conference on Information Communication and Embedded Systems, ICICES 2014*. https://doi.org/10.1109/ICICES.2014.7033934

[12] Yang, Z., Zhong, S., & Wright, R. N. (2005). Privacy-Preserving Classification of Customer Data without Loss of Accuracy. *Proceedings of the 2005 SIAM International Conference on Data Mining*, 92–102. https://doi.org/10.1137/1.9781611972757.9

[13] Yu, S. (2016). Big Privacy: Challenges and Opportunities of Privacy Study in the Age of Big Data. *IEEE Access*, *4*, 2751–2763. https://doi.org/10.1109/ACCESS.2016.2577036

AUTHORS PROFILE

Rajul Jain, Department of Computer Science and Engineering, Maulana Azad National Institute of Technology, Bhopal, India

Pranjali Singh, Department of Computer Science and Engineering, Maulana Azad National Institute of Technology, Bhopal, India

doi: 10.32622/ijrat.78201919