

Enhancing data publishing privacy: split-and-mould, an algorithm for equivalent specification

Supriya G Purohit, Veeragangadhara Swamy

Department of Computer Science and Engineering, Rao Bahadur Y. Mahabaleswarappa Engineering College,
Ballari Affiliated to Visvesvaraya Technological University, Belagavi, India

Article Info

Article history:

Received Oct 15, 2023

Revised Nov 18, 2023

Accepted Dec 16, 2023

Keywords:

Big data

Data privacy

Data publish

Privacy preservation

Security

ABSTRACT

Data sharing and publication have been popular in recent years due to the abundance of options. Evaluating and extracting data from sizable valuable databases i.e., data mining has various challenges which include issues with security, privacy, and data integrity. Anonymized data is used in the majority of privacy preserving data publication approaches, depending on a few utilitarian measures. However, applications that have particular needs for the data they utilize might not be able to use the anonymized data. Practical data anonymization must work to accomplish two opposing objectives: to maintain the data's usefulness and to satisfy a specific privacy need. The utility loss when data is anonymized is frequently measured using generic utility metrics, such as the specific values generalized in a specific ontology. As a need for an application, we suggest equivalent specification, a technique that enables a data user to characterize some properties of the anonymized data. We also introduce the "split-and-mould" algorithm, a heuristic anonymization algorithm that applies a generalization method to the user-provided parameters. Our preliminary results indicate that the specification format and procedure can improve significantly the utility of the anonymized data for data mining that develop predictive models, like decision trees (DTs) and Naïve Bayes.

This is an open access article under the [CC BY-SA](#) license.



Corresponding Author:

Supriya G Purohit

Department of Computer Science and Engineering, Rao Bahadur Y. Mahabaleswarappa Engineering College, Ballari Affiliated to Visvesvaraya Technological University

Belagavi, Karnataka, India

Email: sup1purohit@gmail.com

1. INTRODUCTION

In the era of big data and information sharing, maintaining a delicate balance between information utility and individual privacy has become an increasingly pressing concern. With the proliferation of sensitive information being collected and disseminated across various domains, protecting the privacy of individuals is paramount. The demand for privacy-preserving methods in data publication has encouraged to the creation of novel algorithms [1] that protect and safe-guard sensitive data while permitting data to be used for a range of analytical and research applications. In corporate intelligence, policy-making, and scientific studies, data release is essential. It is essential to provide insightful trends, patterns, and insights that may guide innovation and decision-making. But the disclosure of unprocessed, raw data may be a serious threat to people's privacy, particularly when those data points can be connected together to zero down on identity of a particular people. The difficulty of re-identification is increased in this day and age due to developments in data analytics and machine learning. Several methods have been put forward in the field of privacy-preserving data publication to solve this urgent issue. One such method is "equivalent

specification,” in which sensitive features are hidden but data is altered to retain its analytical value [2]. By using these techniques, it is expected to strike a compromise that will allow data to be shared with various stakeholders like analysts, researchers, and other interested parties while protecting the privacy of the persons whose sensitive and personal information it contains. Even though efficient, there are still numerous obstacles in the privacy-preserving data publication strategies [3]. Current algorithms cannot always provide sufficient privacy protection, and in some situations, they might significantly reduce the usefulness of the data, which would reduce the efficacy of data-driven research. In order to attain the ideal equilibrium between privacy and functionality is nevertheless a difficult and developing task.

We provide the “split-and-mould” technique in this context, providing an improved method for similar specification. By efficiently dividing the data into privacy-sensitive and utility-rich components, and then shaping them in a manner that ideally maintains privacy without reducing data usefulness, this algorithm aims to solve the shortcomings of previous approaches. “split-and-mould” offers a more reliable and adaptable method of achieving comparable specifications, which is an innovative solution to the enduring issue of data publication privacy. The “split-and-mould” algorithm’s main elements, methodology, and possible ramifications for the larger fields of data privacy and publication are described in this work. The algorithm, its experimental validation, and a comparison with current approaches will all be covered thoroughly in the next sections, which also attempts to show how good the algorithm is in protecting data publication privacy without sacrificing data usefulness. People working in any technological field use their own terms to define the characteristics and technologies with the specification. Similarly, several project-related terms have been discussed in this section. i) Data mining: data mining, often referred to as data discovery, is examining information from various outlook and refining it to meaningful and useful knowledge that may be utilized to increase revenue, cut expenses, or both. Data analysis is done using a variety of analytical techniques, including data mining tools. Using the tool, users may categorize the data, list the connections made, and examine it from various angles. Finding correlations or patterns between hundreds of variables in large relational databases is a technique known as data mining. ii) Database publishing: data publishing is a branch of automated media creation that uses specialised techniques to produce paginated pages from source data stored in traditional databases. Mail order catalogues, direct marketing, report generating, pricing lists, and telephone directories are common examples. The database content can be text and images, but it can also include metadata about formatting and additional rules that may apply to the document generating process. Database publishing may be implemented as a component into broader workflows where publications are developed, authorised, amended, and distributed. iii) Anonymization: anonymization is the elimination of information that might lead to the identification of an individual, either just on the basis of the deleted information or when coupled with additional information.

In the Table 1, we can see the decision made for the diseases in the ordered list of attributes {address, gender, age, diseases}. Suppose Raghvendra wants to check his disease as young person in Whitefield using data provided by an organization by training a decision tree (DT). Raghvendra might arrive at the conclusion as depicted in the Figure 1 by applying the DT approach to learn as compared with the conclusion depicted in the Table 1, if he used the original data table. The approach using the DT shows that a young person from whitefield would have probability of brain tumor disease is more than probable of getting Kidney damaged.

Table 1. Data table

Name	Age	Gender	Address	Disease	Name	Age	Gender	Address	Disease
Anoop	Middle	Male	K R Pura	Brain tumor	Mala	Old	Female	K R Pura	Kidney damage
Anurag	Middle	Male	Whitefield	Kidney damage	Seema	Old	Female	Whitefield	Kidney damage
Manoj	Middle	Male	Hoskote	Brain tumor	Harshitha	Old	Female	Hoskote	Brain tumor
Collin	Middle	Male	Hoodi	Kidney damage	Prema	Old	Female	Hoodi	Kidney damage
Bhagya	Middle	Female	K R Pura	Kidney damage	Santhosh	Young	Male	K R Pura	Brain tumor
Shanthamma	Middle	Female	Whitefield	Brain tumor	Raghvendra	Young	Male	Whitefield	Brain tumor
Rudramma	Middle	Female	Hoskote	Brain tumor	Soma	Young	Male	Hoskote	Kidney damage
Manjula	Middle	Female	Hoodi	Kidney damage	Bheem	Young	Male	Hoodi	Kidney damage
Manjunath	Old	Male	K R Pura	Kidney damage	Sushma	Young	Female	K R Pura	Brain tumor
Rakesh	Old	Male	Whitefield	Brain tumor	Anusha	Young	Female	Whitefield	Kidney damage
Veeresh	Old	Male	Hoskote	Brain tumor	Thriveni	young	Female	Hoskote	Kidney damage
Shiva	Old	Male	Hoodi	Kidney damage	Vanaja	young	Female	Hoodi	Brain tumor

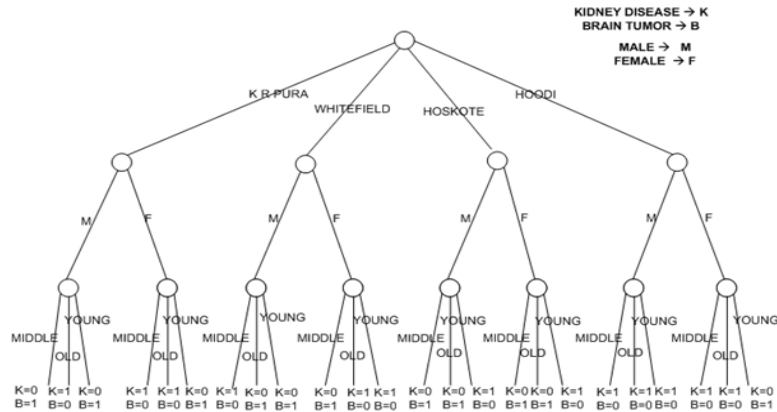


Figure 1. DT learned from Table 1

In the Table 2, again we made the decision for the diseases in the ordered list of attributes {age, gender, diseases}. Raghvendra might create the DT shown in Figure 2 utilizing the same learning technique. However, with a probability of 1.0, which is the DT makes the blatantly erroneous prediction that a young Whitefield man would have a probability of having both diseases equally 0.5 (2 for brain tumor and 2 for kidney damage).

Table 2. Transformed data table according to the ordered list of attributes {AGE, GENDER, ADDRESS}

Name	Age	Gender	Address	Disease
Seema	Old	Female	Whitefield#hoodi#K R Pura#hoskote	Kidney damage
Prema	Old	Female	Whitefield#hoodi#K R Pura#hoskote	Kidney damage
Mala	Old	Female	Whitefield#hoodi#K R Pura#hoskote	Kidney damage
Harshitha	Old	Female	Whitefield#hoodi#K R Pura#hoskote	Brain tumor
Sushma	Young	Female	K R Pura#hoskote	Brain tumor
Thriveni	Young	Female	K R Pura#hoskote	Kidney damage
Anusha	Young	Female	Whitefield#hoodi	Kidney damage
Vanaja	Young	Female	Whitefield#hoodi	Brain tumor
Bhagya	Middle	Female	K R Pura#hoskote	Kidney damage
Rudramma	Middle	Female	K R Pura#hoskote	Brain tumor
Shanthamma	Middle	Female	Whitefield#hoodi	Brain tumor
Manjula	Middle	Female	Whitefield#hoodi	Kidney damage
Manjunath	Old	Male	K R Pura#hoskote	Kidney damage
Veeresh	Old	Male	K R Pura#hoskote	Brain tumor
Rakesh	Old	Male	Whitefield#hoodi	Brain tumor
Shiva	Old	Male	Whitefield#hoodi	Kidney damage
Santhosh	Young	Male	K R Pura# hoskote	Brain tumor
Soma	Young	Male	K R Pura# hoskote	Kidney damage
Raghvendra	Young	Male	Whitefield#hoodi	Brain tumor
Bheem	Young	Male	Whitefield#hoodi	Kidney damage
Anoop	Middle	Male	K R Pura#hoskote# whitefield#hoodi	Brain tumor
Manoj	Middle	Male	K R Pura#hoskote# whitefield#hoodi	Brain tumor
Anurag	Middle	Male	K R Pura#hoskote# whitefield#hoodi	Kidney damage
Collin	Middle	Male	K R Pura#hoskote# whitefield#hoodi	Kidney damage

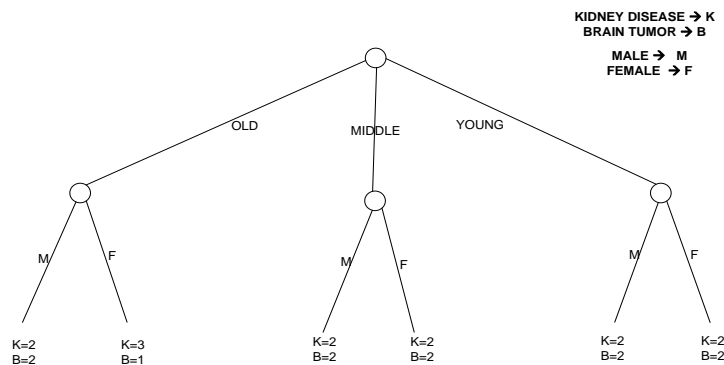


Figure 2. DT learned from Table 2 after the data transformation

2. LITERATURE SURVEY

Numerous sources and techniques are used to collect and handle data, which raises privacy concerns [1]. The privacy of a person has so far been protected through the techniques like randomization, k-anonymization, ℓ -diversity, t closeness, cryptography, and several such methods. However, each method has its own drawbacks, such as information loss, privacy violations, and low data utility. K anonymization strategy is one of the most popular anonymization-based approaches out of all of them. This method, however, faces the problem of information loss. Therefore, data mining is a difficult work for data miners. In contrast to the conventional technique, the research focuses on reducing information loss utilising the 2-level k anonymization approach while still preserving privacy. This strategy's primary goal is to reduce data loss while maintaining privacy.

For storage, computation, and data utilization in the cloud environment, a significant amount of data and applications must be shared with numerous parties and stakeholders. The cloud platform is run by a third party; thus, owners cannot have complete faith in this setting. When effectively transferring data between parties, it might be difficult to protect privacy. Applying the k-anonymization, differential privacy, and machine learning techniques, this study [2] suggests a unique method which divides data into sensitive and non-sensitive parts, adds noise into sensitive data, and classification tasks are performed. For a variety of uses, it enables several owners to share their data in a cloud environment. The model outlines the communication protocol for the several untrusted parties that are involved in processing the data of the owners.

To address the key issues with privacy protection in the big data context, Wu *et al.* [3] researched medical data from the data collection, data transport, and data sharing. The work proposed the MNSSp3 (medical big data privacy protection platform based on Internet of things), which aims to provide an efficient medical data sharing solution with the privacy protection algorithms for various data types and support for data analytics. In order to offer users mining techniques, the platform focuses on the transmission and sharing security of medical big data and realizes the separation of data and users to secure the security of medical data. Additionally, the site gives users the option to add privacy algorithms on their own. Three case studies were presented to demonstrate the platform's functionality after authors examined its needs and design elements. The results of the trials clearly demonstrate the usefulness and viability of the suggested platform.

With the fast advancement of computer, networking, and database technologies, vast amounts of digital data are now being collected and integrated. This data must be shared/published among numerous parties for analysis or to cater to compulsory disclosure by the law of the land. When these data contain personally sensitive information, an individual's privacy becomes a major problem. Many privacy preserving data publishing (PPDP) methods are being developed by researchers to address these privacy problems [4]-[7]. Some popular strategies are k-anonymity [8]-[11], ℓ -diversity [12], [13], t-closeness [14], m-invariance [15], personalized privacy [16], slicing [17], and others. Aggarwal [18], [19] demonstrate that k-anonymity loses a significant amount of information because of the homogeneity and background attacks. ℓ -diversity technique is more useful for data than k-anonymity, although it is vulnerable to skewness attack. Because quasi-identifiers are already disclosed, this strategy cannot prevent membership disclosure [20]. The sensitive attributes distribution in the individual buckets must approximate the spread and placing of the attributes within the original table, according to t-closeness [14]. The usability of the information in addition to the relationship among four times daily (QID) and SAs are both negatively impacted by this situation. A dynamic data republishing technique called m-invariance [15] permits record modification and deletion as well as a variety of data releases. However, privacy is not assured by the m-invariance when the span of attribute existence is disturbed, i.e., if a record continues beyond its original shelf life, for instance.

Personalized privacy [16] eliminates these restrictions, but it suffers due to the "play safe dilemma," where the record owner may choose to be careful by checking the guarding node to "Any illness," that help them keep in a more secure private zone. However, for several data mining initiatives, taking the safe route would lead to inaccurate findings. Since slicing [17] divides the contents of the table horizontally as well as vertically, it is a more effective tactic for preserving privacy and avoiding data loss. Vertical partitioning is achieved by classifying traits according to their connection. To calculate the correlation amongst the attributes, a mean-square contingency coefficient is used and they cluster the attributes using the computationally expensive k-medoid technique, which divides the attributes into columns. The 'Mondrian' approach [11] is used to divide tuples into buckets in order to achieve horizontal partitioning, even though it is not the best technique for tuple partitioning.

The works [21], [22] proposed an advancement in k-anonymity or ℓ -diversity by presenting a systematized clustering approach to k-anonymize the attributes. The sensitive characteristics distributed in every individual bucket must approximate with the attribute's distribution within the original table, according to t-closeness [14]. In order to avoid corruption attacks [23], an independent ℓ -diversity idea was put out in [24]. For greater data utility and preservation, it combines perturbations and generalization. It was stated in [25], [26]

how to protect unprocessed textual medical data by sanitizing sensitive association criteria by modifying the confidence and backing of linked attributes. Big data and its analytics are a rapidly growing field which is transforming all sections of database and its management like for example, big data related to traffic, disease outbreaks and detection, smart grids, product preferences and purchase patterns, and other areas. Fan and Jin [27] addressed privacy preservation for large data analytics in order to protect differential privacy for individual data sources. The authors present a general approach for producing analytical results from a sampling database. Zaman *et al.* [28] a strategy for improved classification precision, with a non-interactive technique to meet e-differential privacy. Weng *et al.* [29] proposes protecting the privacy of outsourced multimedia content. The paradigm used by the authors to provide privacy is based on robust hash techniques and a partial method of encryption.

For organizations that distribute microdata for informal examination, privacy is a big problem. Most of the PPDPs i.e., privacy-preserving methods anonymize data in accordance with either specific privacy considerations or a more general utility standard. As a consequence, when data is anonymized, both the record owner's privacy concerns as well as analyst's requirements for data efficacy are taken into consideration, which compromises the accuracy of many data mining tasks. As a result, the authors of the study [4] propose a novel approach that takes into consideration the privacy requirements of data miners (analysts) and record owners in the format of sensitivity flags and application specific requirements.

The stipulation-based anonymization algorithm [4] is divided into two sections. The initial portion of the article revised the greedy personalized-generalization algorithm [5] to accommodate the record owner's "play safe" request. The SA_LIST is searched for the value of each tuple's sensitive property called disease. If it is discovered, just the sensitivity flag is taken into account for anonymization. In the second phase, the quasi-IDs generalization and the generalization of SA was completed.

Individuals' sensitive information is frequently contained in detailed person particular data. Individual privacy must be protected when such information is shared. PPDP approaches and tools are provided for releasing relevant information while maintaining data privacy. The intricacy of its depiction, as well as the needs of the modern industry, have prompted a great deal of study in this area. The author provides a brief assessment of several strategies for PPDP in this work. Authors have also highlighted current research on anonymization and addressed various threats that may occur during the anonymization process.

The majority of currently used PPDP methods are unable to handle a wide variety of features with varied degrees of sensitivity. To cope with several diverse (that is, numeric and qualitative) sensitive characteristics, they thus suggest a novel method. Instead of applying an identical amount of confidentiality and privacy to all data records without taking into account the record owner's specific needs, this method takes into account personal sensitivity rating flags [6] and applies privacy-preserving techniques to those records that need them. Additionally, the connection between the characteristics is established, and by overgeneralizing other traits, highly related features are generalized using as little as possible. As a consequence, by retaining the most details from the original data, our method will provide enough privacy. The recommended method [6] exceeds earlier research by means of theoretical assessment and mathematical analysis when put through enough testing.

For firms that publish or exchange personal data for ad hoc research, privacy is a big issue. As a remedy for this PPDP, several anonymization techniques, to generalize and bucketize, are being researched. Recent research has shown that generalization significantly reduces information, particularly for data with multiple dimensions.

Contrarily, bucketing does not restrict membership disclosure. The creative method used Ashoka and Poornima [7] illustrates how well an attribute can be used to categorize data and shows how characteristics are related to one another. It permits the utilization of information gained about the characteristics with regard to susceptible attributes. Ashoka and Poornima [4] demonstrate how this strategy maintains a better degree of data usefulness while being less difficult than prior ones.

3. METHOD

Figure 3 is the proposed architecture which has modules mentioned are anonymized data collection and data transformation using split-and-mould mechanism:

- Anonymized data collection: data is in the format of ordered attributes. We will specify the order of attributes.
- Data transformation: data transformation can be done based on the split-and-mould mechanism. And finally, the transformed data pushed to the database.

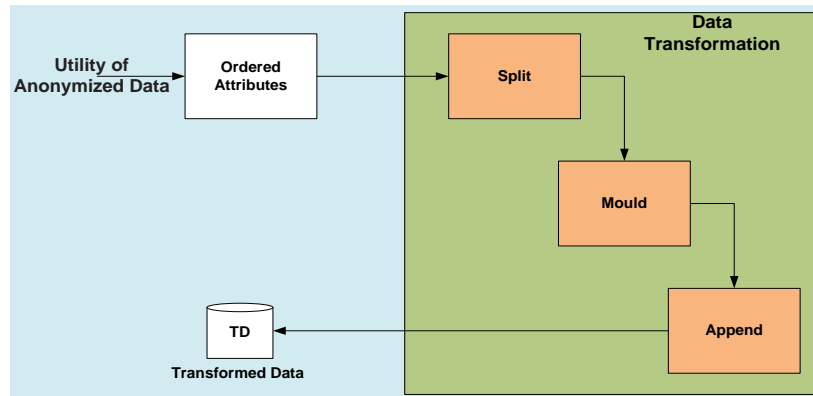


Figure 3. System architecture

The “split-and-mould” algorithm is a method designed for enhancing data publishing privacy, particularly in scenarios where the data contains sensitive information that needs to be protected while preserving its utility for analysis and research purposes. Here’s an explanation of the methodology behind the “split-and-mould” algorithm:

i. Data input:

The algorithm takes as input a data table (often denoted as DT) that contains sensitive data. It also requires a utility specification (S) which defines the desired attributes or features that should be preserved for data analysis. An ℓ -diversity requirement (r) is specified to ensure that the data retains a certain level of diversity among its records.

ii. Initialization:

- Initialize a list (often denoted as L) to store subsets of the data. Initially, L contains the entire data table DT.
- Set a quasi-identifier (T. qi) to a default value, which is often represented as $\langle *, \dots, * \rangle$ to mask sensitive attributes.

iii. Attribute ordering:

The attributes of the quasi-identifier T are ordered according to the utility specification S. This reordering helps identify which attributes are more important for analysis and need to be preserved. The ordering helps in reducing efforts of privacy on redundant attribute. The utility specification perfects the attribute-based privacy to be provided to selected sensitive attributes. The reordering supports to analyse the attributes to be preserved for the privacy.

iv. Iterative processing:

For each attribute (often denoted as A_i) in the order determined by the utility specification:

For each subset E in the list L:

- Split ($L' = \text{split}(E, i)$): the subset E is divided into multiple subgroups based on the values of attribute A_i . This partitioning is done to segment the data for further processing.
- Mould ($L' = \text{mould}(L', i, r)$): each of these subgroups is then transformed in a way that guarantees ℓ -diversity with respect to attribute A_i , as specified by the requirement r. This ensures that sensitive information remains sufficiently obscured within each subgroup.
- Append: the transformed subgroups (L') are appended to a new list (L'').

v. Update list L:

The list L is updated with the contents of L'' to include the transformed and split subgroups. The updated list meets the ℓ -diversity requirement and utility specification. A balance between data utility and privacy protection is achieved, which makes data publishing safe and yet useful for analytics, especially where sensitive attributes are a part of the input data.

vi. Output:

Once all the attributes have been processed in this manner, and all subsets within the list L meet the ℓ -diversity requirement and utility specification, the algorithm returns the final data in list L. The “split-and-mould” algorithm’s methodology is aimed at achieving a balance between data utility and privacy protection. It does so by segmenting the data into subsets based on important attributes, applying privacy-preserving transformations, and ensuring that the resulting data complies with ℓ -diversity requirements. This makes it a valuable tool for data publishing in scenarios where privacy is a concern, such as healthcare, finance, and research.

3.1. Algorithms

The Algorithms 1 for Split-and-mould will take the Data Table DT, utility specification S and an ℓ -diversity specification as input. The attributes are ordered according to the ℓ -diversity requirement and a new list of attributes is appended resulting in a modified data table satisfying the ℓ -diversity requirement specification. The Algorithm 2 for moulding takes a collection of equivalence classes denoted as L, a specific attribute position i, and an ℓ -diversity requirement r as input. The attributes are modified according to the requirements specified. A revised set of ECs (equivalent classes) meeting the ℓ -diversity requirement is generated.

Algorithm 1. Algorithm for split-and-mould

Input: a data table DT, a utility specification S, and an ℓ -diversity requirement r.

Output: a modified data table that satisfies the ℓ -diversity requirement.

Procedure:

1. Initialize a list L with the original data table DT, and set the quasi-identifier T. qi to $\langle *, \dots, * \rangle$.
2. Arrange the attributes of the quasi-identifier T according to the specified utility specification S.
3. For each attribute A_i in the given order:
4. For each element E in list L:
5. Split the element E using attribute A_i , creating a new list L' .
6. Modify the elements in L' to achieve ℓ -diversity with respect to attribute A_i , following requirement r.
7. Append the modified list L' to a new list L'' .
8. Update list L with the elements in L'' .
9. Return the tuples in the final list L.

Algorithm 2. Algorithm for mould

Input: a collection of ECs (equivalence classes) denoted as L, a specific attribute position i, and an ℓ -diversity requirement r.

Output: a revised set of ECs that meet the ℓ -diversity requirement.

Procedure:

1. Continue the following steps while there exists an EC E in the set L that does not meet the ℓ -diversity requirement r.
2. Discover an alternative EC E' within the set L.
3. Merge the contents of EC E with EC E' .
4. Update the attribute position i of EC E with the union of the corresponding attribute positions in EC E' .
5. Remove EC E' from the set L.
6. Once all ECs in the set L satisfy the ℓ -diversity requirement, return the modified set L.

3.2. Machine learning models

3.2.1. Accuracy for classifier called DT

We consider the data exists as original and transformed data and do apply DT classifier algorithm from the WEKA library and compute the statistics comes under DT. And also find the accuracy achieved from data exists as original and accuracy achieved from the data which is transferred with the DT algorithm. And finally, we find utility value for the DT utility value = accuracy achieved from original dataset/accuracy achieved from the transformed dataset.

3.2.2. Accuracy for classifier called Naïve Bayes

We take the data exists as origin and transformed data and do apply Naïve Bayes classifier algorithm from the WEKA library and compute the statistics comes under Naïve Bayes. And also find the accuracy achieved from data exists as original and accuracy achieved from the transformed data with the Naïve Bayes algorithm. And finally, we find utility value for the Naïve Bayes. Utility value = accuracy achieved from original dataset/accuracy achieved from the transformed dataset

4. RESULTS AND DISCUSSION

In the experiments, we discussed on the algorithm “split-and-mould” and start comparing the output dataset with the ML models. The algorithms are all implemented in java and used with swings for GUI development. The execution is done on the PC with minimal 4 GB RAM and 3 GHz CPU. To assess the value of generated anonymized data from the “split-and-mould” algorithm, first we learn models based on knowledge obtained from the anonymized data then apply ML models to the transformed data and then check the performance. In this research we first compute the statistics and make the comparison between the machine learning algorithms such as DT and Naïve Bayes. In Figure 4 the statistics of DT classifier from the

WEKA. In Figure 5 the statistics of Naïve Bayes classifier from the WEKA. From the above results section, we can conclude that DT classifier gives more accurate results than the Naïve Bayes model where accuracy achieved in DT is 78% and accuracy achieved in Naïve Bayes is 61%.

Training results for original dataset

811 instances were correctly classified, making up 61.1111%
 87 instances that were incorrectly classified, or 38.8889%
 Kappa value: 0.2222
 The average absolute error is 0.125.
 Error in the root mean square is 0.25.
 Error percentage relative to absolute 72.9167%
 Error squared relative to the root is 91.0274%
 The total number of instances are 898.

Training results for transformed dataset

811 cases, or 61.1111%, were accurately identified.
 87 instances, or 38.888%, were misclassified.
 Value of Kappa: 0.2222
 The absolute inaccuracy is 0.125 on average.
 The root mean square error is 0.25.
 Relative error to absolute error 72.9167%
 91.0274% is the error squared in relation to the root.
 There are 898 cases in total.

Accuracy with original data set is 0.78125

Accuracy with transformed data set is 0.78125

Utility value is 1.0

Figure 4. The statistics of DT classifier from the WEKA

Training results for original dataset

811 instances were correctly classified, making up 61.1111%
 87 instances that were incorrectly classified, or 38.8889%
 Kappa value: 0.2222
 The average absolute error was 0.1257.
 Error in the root mean square is 0.2507
 Error percentage relative to absolute 73.3147%
 error squared relative to the root is 91.2846%
 The total number of instances are 898.

Training results for transformed dataset

811 cases (61.1111%) were accurately classified.
 87 cases were incorrectly classified, amounting to a 38.8889% error rate.
 The mean absolute error was 0.1257 and the root mean squared error was 0.2222 in the Kappa statistic.
 The root relative squared error is 73.3147 percent relative absolute error and 91.2846%.
 There have been 898 occurrences in total.

Accuracy with original dataset is 0.6124999999999999

Accuracy with transformed dataset is 0.6124999999999999

Utility value is 1.0

Figure 5. The statistics of Naïve Bayes classifier from the WEKA

5. CONCLUSION




In this research, “split-and-mould” mechanism helps us to get the ordered list of attributes for the preservation of particular features and their values for mining applications. The presentation on anonymised algorithm called “split-and-mould”, integrates user preferred requirements into generalization mechanism. More preferably, the proposed algorithm by generalising the attribute values that are omitted from or are of lesser importance in the datasets and the values specified in the utility specification are considered.

The results are shown in the format of specification and also, we do apply machine learning models for finding the performance in terms of accuracy with the original dataset and transformed dataset. We took Naïve Bayes and DT model for classification approaches. Further we go for adding the security to the data which are taken as sensitive data attributes, the inclusion of chaos-based encryption or block-based encryption to the sensitive attributes can be made.




REFERENCES

- [1] V. B. Vaghela, "K-anonymization approach for privacy preserving in data mining," *International Journal of Scientific and Technology Research*, vol. 9, no. 1, pp. 1794–1799, 2020.
- [2] A. K. Singh and R. Gupta, "A privacy-preserving model based on differential approach for sensitive data in cloud environment," *Multimedia Tools and Applications*, vol. 81, no. 23, pp. 33127–33150, 2022, doi: 10.1007/s11042-021-11751-w.
- [3] X. Wu, Y. Zhang, A. Wang, M. Shi, H. Wang, and L. Liu, "MNSSp3: medical big data privacy protection platform based on internet of things," *Neural Computing and Applications*, vol. 34, no. 14, pp. 11491–11505, 2022, doi: 10.1007/s00521-020-04873-z.
- [4] K. Ashoka and B. Poornima, "Stipulation-based anonymization with sensitivity flags for privacy preserving data publishing," in *Advances in Intelligent Systems and Computing*, vol. 707, 2019, pp. 445–454.
- [5] K. Ashoka and B. Poornima "A survey of latest developments in privacy preserving data publishing," 2014, doi: 10.15693/ijaist/2014.v3i12.1423.
- [6] K. Ashoka and B. Poornima, "Enhanced utility in preserving privacy for multiple heterogeneous sensitive attributes using correlation and personal sensitivity flags," *2017 International Conference on Advances in Computing, Communications and Informatics, ICACCI 2017*, vol. 2017-Janua, pp. 970–976, 2017, doi: 10.1109/ICACCI.2017.8125967.
- [7] K. Ashoka and B. Poornima, "Mutual correlation-based optimal slicing for preserving privacy in data publishing," in *Smart Innovation, Systems and Technologies*, vol. 77, 2018, pp. 593–601.
- [8] L. Sweeney, "k-anonymity: a model for protecting privacy," *International Journal of Uncertainty, Fuzziness and Knowledge-Based Systems*, vol. 10, no. 5, pp. 557–570, 2002, doi: 10.1142/S0218488502001648.
- [9] P. Samarati, "Protecting respondents' identities in microdata release," *IEEE Transactions on Knowledge and Data Engineering*, vol. 13, no. 6, pp. 1010–1027, 2001, doi: 10.1109/69.971193.
- [10] K. LeFevre, D. J. DeWitt, and R. Ramakrishnan, "Incognito: efficient full-domain K-anonymity," *Proceedings of the ACM SIGMOD International Conference on Management of Data*, pp. 49–60, 2005, doi: 10.1145/1066157.1066164.
- [11] K. LeFevre, D. J. DeWitt, and R. Ramakrishnan, "Mondrian multidimensional K-anonymity," in *Proceedings - International Conference on Data Engineering*, 2006, vol. 2006, p. 25, doi: 10.1109/ICDE.2006.101.
- [12] A. Machanavajhala, J. Gehrke, D. Kifer, and M. Venkatasubramanian, "t-Diversity: privacy beyond k-anonymity," *Proceedings - International Conference on Data Engineering*, vol. 2006, p. 24, 2006, doi: 10.1109/ICDE.2006.1.
- [13] J. Domingo-Ferrer and V. Torra, "A critique of k-anonymity and some of its enhancements," *ARES 2008 - 3rd International Conference on Availability, Security, and Reliability, Proceedings*, pp. 990–993, 2008, doi: 10.1109/ARES.2008.97.
- [14] N. Li, T. Li, S. Venkatasubramanian, "t-Closeness : privacy beyond k-anonymity and-diversity t -closeness : privacy beyond k -anonymity and -diversity," no. July, 2018.
- [15] X. Xiao and Y. Tao, "M-invariance: towards privacy preserving re-publication of dynamic datasets," *Proceedings of the ACM SIGMOD International Conference on Management of Data*, pp. 689–700, 2007, doi: 10.1145/1247480.1247556.
- [16] X. Xiao and Y. Tao, "Personalized privacy preservation," *Proceedings of the ACM SIGMOD International Conference on Management of Data*, pp. 229–240, 2006, doi: 10.1145/1142473.1142500.
- [17] T. Li, N. Li, J. Zhang, and I. Molloy, "Slicing: a new approach for privacy preserving data publishing," *IEEE Transactions on Knowledge and Data Engineering*, vol. 24, no. 3, pp. 561–574, 2012, doi: 10.1109/TKDE.2010.236.
- [18] C. C. Aggarwal, "On k-anonymity and the curse of dimensionality," *VLDB 2005 - Proceedings of 31st International Conference on Very Large Data Bases*, vol. 2, pp. 901–909, 2005.
- [19] D. Kifer and J. Gehrke, "Injecting utility into anonymized data sets," in *Proceedings of the 2006 ACM SIGMOD international conference on Management of data*, 2006.
- [20] M. E. Nergiz, M. Atzori, and C. Clifton, "Hiding the presence of individuals from shared databases," *Proceedings of the ACM SIGMOD International Conference on Management of Data*, pp. 665–676, 2007, doi: 10.1145/1247480.1247554.
- [21] M. E. Kabir, H. Wang, and E. Bertino, "Efficient systematic clustering method for k-anonymization," *Acta Informatica*, vol. 48, no. 1, pp. 51–66, Feb. 2011, doi: 10.1007/s00236-010-0131-6.
- [22] P. R. Bhaladhare and D. C. Jinwala, "Novel approaches for privacy preserving data mining in k-anonymity model," *Journal of Information Science and Engineering*, vol. 32, no. 1, pp. 63–78, 2016.
- [23] Y. Tao, X. Xiao, J. Li and D. Zhang, "On anti-corruption privacy-preserving publication," *2008 IEEE 24th International Conference on Data Engineering*, 2008.
- [24] H. Zhu, S. Tian, and K. Lü, "Privacy-preserving data publication with features of independent t-diversity," *Computer Journal*, vol. 58, no. 4, pp. 549–571, 2015, doi: 10.1093/comjnl/bxu102.
- [25] Z. Fengli and B. Yijing, "ARM-based privacy preserving for medical data publishing," in *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, vol. 9483, 2015, pp. 62–73.
- [26] D. Sánchez, M. Batet, and A. Viejo, "Utility-preserving privacy protection of textual healthcare documents," *Journal of Biomedical Informatics*, vol. 52, pp. 189–198, Dec. 2014, doi: 10.1016/j.jbi.2014.06.008.
- [27] L. Fan and H. Jin, "A practical framework for privacy-preserving data analytics," *WWW 2015 - Proceedings of the 24th International Conference on World Wide Web*, pp. 311–321, 2015, doi: 10.1145/2736277.2741122.
- [28] A. N. K. Zaman, C. Obimbo, and R. A. Dara, "A novel differential privacy approach that enhances classification accuracy," *ACM International Conference Proceeding Series*, vol. 20-22-July, pp. 79–84, 2016, doi: 10.1145/2948992.2949027.
- [29] L. Weng, L. Amsaleg, and T. Furon, "Privacy-preserving outsourced media search," *IEEE Transactions on Knowledge and Data Engineering*, vol. 28, no. 10, pp. 2738–2751, Oct. 2016, doi: 10.1109/TKDE.2016.2587258.

BIOGRAPHIES OF AUTHORS

Supriya G Purohit    working as an Assistant Professor in the Department of Computer Science and Engineering at Navodaya Institute of Technology, Raichur, Karnataka, has about 9 years of teaching experience and more than 2 years of IT experience. She received her B.E. degree in Information Science and Engineering with distinction from Visvesvaraya Technological University, Belagavi and M.E. degree in Software Engineering with Distinction from Bangalore University, Bangalore. She is a Research Scholar at, Department of Computer Science and Engineering, Rao Bahadur Y. Mahabaleswarappa Engineering College, Ballari, Karnataka. Author's areas of research include big data, ethical hacking, and cloud security. She has supervised and co-supervised more than 20 UG students. She can be contacted at email: sup1purohit@gmail.com.



Dr. Veeragangadhara Swamy    working as Professor in the Department of Computer Science and Engineering, RYMEC, Ballari, has about 25 years of teaching experience. He received his B.E. degree in CSE Engineering from University BDT Engineering College, Davangere, under Visvesvaraya Technological University, Belagavi with distinction and M.Tech. in Computer Science and Engineering from Dr. Ambedkar Institute of Technology, Bangalore, Visvesvaraya Technological University, Belagavi. He received Ph.D. degree in Computer Science and Engineering from JJT University, Rajasthan. He has published more than 30 research papers in various International Journals and more than 15 research papers in the proceedings of various International Conferences. He has received Best Faculty awards for publishing research papers at various international conferences. His areas of research include data mining, text mining, and big data. He is an active member of ISTE. Author is working as Board of Examiners Member (BOE), Visvesvaraya Technological University, Belagavi, Gulbarga Region. He is a Convener of IQAC for the Dept of Computer Science and Engineering, Rao Bahadur Y. Mahabaleswarappa Engineering College, Ballari. Author is also a National Advisory Board Member, IFERP International Conference. Author's areas of research include knowledge data discovery, data mining, and cloud security. He has supervised and co-supervised more than 20 UG students, 40 PG students and more than 15 Ph.D. students. He can be contacted at email: swamytm@gmail.com.