iJRASET
International Journal For Research in Applied Science and Engineering Technology

INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY

Volume: 7 **Issue:** IX **Month of publication:** September 2019

DOI: http://doi.org/10.22214/ijraset.2019.9027

www.ijraset.com

Call: 08813907089 | E-mail ID: ijraset@gmail.com

International Journal for Research in Applied Science & Engineering Technology (IJRASET)
ISSN: 2321-9653; IC Value: 45.98; SJ Impact Factor: 7.177
Volume 7 Issue IX, Sep 2019- Available at www.ijraset.com

Privacy Preservation in Data Centric Environment using K-Anonymity

Shrinkhala Shinghai[1], Somesh Dewangan[2], Rahul Mishra[3]
[1]*M.Tech Scholar,* [2, 3]*Assistant Professor, RSR-RCET, Bhilai, Chhattisgarh, India*

Abstract: ***Due to the expansion in healthcare information systems, the availability of therapeutic reports has benefitted human administration organizations to inquire about work. In numerous cases, these are creating concerns while sharing helpful records. Protection methods for an unstructured helpful substance highlight on acknowledgment and ejection of individual identifiers from the substance, which may be missing for shielding security and data utility. Considering sensitive social protections information, protection security could be a noteworthy concern, when patients' restorative administration's data is utilized for investigation purposes. In this article, we have compared two methods K-anonymity and the inbuilt simulator ARX tool to ensure who can provide higher privacy on medical databases in data-intensive environments. The outcomes declare that the proposed approach has superior execution than those of the related works concerning variables such as information protection with k-anonymity.***

Keyword: Data Centric Environment, K-Anonymity, Healthcare data, Privacy Preservation

I. INTRODUCTION

In healthcare association, the medical communities still depend on paper-based and handwritten records for analysis purpose. The digitalized data shared by social association is generally not compressed; thus, allowing the probability of sharing this data among various social elements. Privacy is the legal right of each individual. Privacy preservation signifies that one is liberated from all impedance and can regulate the level of closeness. According to the AICPA (American Institute of Certified Public Accountants) and CICA (Canadian Institute of Charted Accountants) privacy is defined as, “It may be termed as the right and obligation of individuals as well as organizations with due respect to the collection, use, retention, and disclosure of someone’s data"[12]. Driven by consistent preferences or regulations that require definite data to be mined and appropriated, there is an excitement for the exchange and generation of data among distinctive parties. The privacy preservation is used in many fields like cloud environment [12] and web based services [4], etc.

The advancement in cloud Computing allows the healthcare suppliers to intensify their administrations with the utilization of cloud services like SaaS (Software as a Service) and DaaS (Database as a Service) model. Due to enhancement in innovation, healthcare administrator authorizes to develop the appropriation of PHR (Personal Health Records), EMR (Electronic Medical Records) and EHR (Electronic Health Records).

Electronic Medical Record/Electronic Health Record (EMR/EHR) systems are used to consolidate and store different type of patient data and their records (Dean, Lam, Natoli, Butler, Aguilar, &Nordyke, 2010; Lau, Mowat, Kelsh, Legg, Engel-Nitz, &Watson, 2011; Makoul, Curry, & Tang, 2001).

Privacy protecting information examines the systems and tries to reveal sensitive individual data while executing the anonymized information suitable for investigation. As of late, there has been a lot of work focusing on the impression of prosperity record data at the individual record level, moreover at the level of companion examination and recognizing worldly patterns of medicines and patient arrangements (Perer, & Sun, 2012; Aigner, & Miksch, 2006). In both cases, the privacy of information is in danger as a result of the mind-boggling biological system of the human services industry, including both trusted and untrusted clients (Fung, Wang, Chen, &Yu, 2010). Here we are comparing the privacy between an in-built tool and applied K-anonymity algorithm giving privacy on medical databases in data-intrinsic environments.

II. COVERAGE PROBLEM

According to Dalenius (1977), privacy protection means not allowing an adversary to obtain any person-specific sensitive information of a targeted individual even though he has some background knowledge from external sources. In PPDP, the attack models are classified into two categories based on their attack principles: First, linking attacks and second, background knowledge. The main focus of the adversary is to gain information about the victim with the help of previously known knowledge.

A. *Record Linkage*

Record linkage intends to describe the records of the targeted victim in the openly issued information dependent on the quasi-identifiers of the victim. If the victim's quasi-identifiers are equivalent to the records in the published table then the opponent faces less no. of potential outcomes for the targeted record with some extra information.

To avoid the attack by record linkage, a new method is proposed by Sweeney, Samrati[15]. In this model, for each set of all quasi-identifiers having the same value in the table must have at least a k number of records. The advantage of this model is that there are other (k-1) tuples that are outlined to the same quasi-identifiers set with an attack probability 1/k.

B. *Attribute Linkage*

In this attack, the attacker gains some information about the sensitive attribute from the released table, even though the attacker cannot able to link the victim with any individual published record. According to Venkatasubramanian[14] to prevent the attribute linkage attack in L -Diversity, it’s necessary conditions are that every identity of an issued table must have at least l different values. The idea is to prevent attribute linkage by assigning different unique sensitive values. So, higher the entropy value in the published table, the lesser are the chances of probabilistic attack because the higher value of the threshold l increases its privacy and lesser is the information gain by an attacker from a published table.

III. LITERATURE REVIEW

A. Jyotir Moy Chatterjee, Raghvendra Kumar, Prasant Kumar Pattnaik and Noor Zaman", Privacy preservation in a data-intensive environment, 2018".The authors work on privacy preservation in healthcare data. According to this author, a protection procedure for an unstructured medicinal content lacks safeguarding privacy and information utility. This paper works on feature selection using Principal Component Analysis and also shares two methods K-anonymity and fuzzy system for providing privacy on medical databases in data-intensive environments.

B. Jyotir Moy Chatterjee, "Privacy Preservation in Data Centric Environment: Analysis and Segregation, 2017". In this paper, different privacy safeguarding information& strategies for improving information quality and viability are discussed. The author worked is typically centred on consolidated strategies for k-Anonymity, association rule mining, cryptographic, and information irritation systems i.e., data perturbation to protect the privacy of information and lessening data loss. This paper tries to share the privacy saving information (data) mining advancements and also the benefits and limitations of these innovations.

C. Hetaswini J., & Vimalkumar B. Vaghela, "Privacy Preserving by Anonymization Approach, 2017". In this paper, a study of the broad areas of privacy-preserving data mining and the underlying algorithms are done. The wide areas of classification include Privacy preserving data publishing; Privacy-Preserving Applications, Utility Issues, Distributed Privacy, cryptography and adversarial collaboration are analysed. A variety of data modification techniques such as randomization and k-anonymity based techniques has been studied and analysed based on their activities. A complete study is done on for distributed privacy-preserving mining, and the methods for handling horizontally and vertically partitioned data. Summary of the three main privacy models, namely, k-anonymity, l-diversity is given in this study paper. Then algorithms which are used to implement a k-anonymity model are also explained.

D. Hunka, T., Dash, S., & Pattnaik, P. K.," Web based Privacy Disclosure Threats and Control Techniques, 2016". This paper focuses on privacy threats that affect adversely on sensitive data and heads to the leakage of confidential information. So, privacy-preserving data mining techniques (PPDM) is introduced to protect the sensitive data before it gets published to the public by changing the original micro-data format and contents. The author offers an extensive study on some ramified disclosure threats to privacy and PPDM methods as a centralized solution to protect against threats.

E. Youssef Gahi, Mouhcine Guennoun, Hussein T. Mouftah, "Big Data Analytics: Security and Privacy Challenges, 2016". In this paper, the author shares the benefits of Big Data Analytics and at the same time reviews the hurdles of security and privacy in big data environments. So, the author presents some possible protection techniques and proposes some possible tracks like Rules and Legality, Encryption, MetaData and Tagged Data that enable security and privacy in a malicious big data context.

F. Kavitha, S., S. Yamini, and Raja Vadhana". An evaluation on big data generalization using K-anonymity algorithm on a cloud, 2015". This paper shares the concept of data Anonymization. It examines the top-down specialization algorithm. As big data have the obstacle which is scalability, so this paper proposes a two-phase top-down specialization technique.

G. Basu, Anirban, et al. "K-anonymity: Risks and the Reality, 2015". This paper shares the knowledge of K-anonymity which is a broad method to preserve privacy while data publishing. K-anonymity lowers the probability of re-identification of individuals

in worst-case based on quasi-identifiers to 1/k. It also assesses the probability of re-identification because of background knowledge.

H. Zakerzadeh, Hessam, Charu C. Aggarwal, and Ken Barker "Privacy-preserving big data, 2015".This paper discusses the problem of scalability in privacy algorithms of big data. It introduces two privacy models, particularly, k-anonymity and l-diversity and introduces an algorithm based on the Map Reduce framework and can handle the scalability issue.

I. Vaibhav Lawand, Prathik Sargar, Anand Bhalerao and Pradip Jadhavn, "Analytical Approach for Privacy Preserving of Medical Data, 2015". In this paper, the author describes the data analysis processes as workflows for medical data. The authors define the workflow for privacy-aware data analysis in mental health research and promote an analytical approach for privacy-preserving of medical data to address these concerns. They share the concept of existing privacy policies for medical health records along with the security for such information. The data is classified using different attribute sets and then different privacy preservation techniques are applied to medical records like anonymization, re-identification, encryption and aggregation. For the data analysis process, the author considers the clustering and classification approach. For clustering purpose, the k-means algorithm is applied and for classification purpose SVM concept is used.

J. Freny Presswala, Amit Thakkar, Nirav Bhatt, "Survey on Anonymization in Privacy-Preserving Data Mining, 2015". This paper compiles the basics of Privacy-Preserving Data Mining with its classifications. It proposes different anonymization techniques such as k-anonymity, l-diversity and t-closeness along with attacks on them and their limitations.

K. Kupwade Patil, H., & Seshadri, R., "Big Data Security and Privacy Issues in Healthcare,2014".This paper emphasizes on big data healthcare transformation, i.e., security and patient privacy is supreme while using technologies. As healthcare clouds with big data are emerging so secure and privacy -preservation in real-time analytics will drive proactive healthcare and wellness. This paper shares some of the security and privacy issues in healthcare and anticipates a need for technological breakthroughs in computational, storage and communication abilities to meet the growing demand of securing healthcare data.

L. Bhagyashri S, Gurav YB," A Survey on Privacy-Preserving Techniques for Secure Cloud Storage, 2014" In this paper, the author focusses on cloud computing technology. The paper shares the issue of privacy and security issues in the cloud and solutions on this issue. A study on distinctive procedures and strategies is done for overpowering the issues in protection on untrusted information capacity in cloud computing like encryption-based strategies, get to control based instruments and auditability plans.

M. Zhu, Yan, and Lin Peng "Study on K-anonymity models of sharing medical information, 2007". This paper outlines the necessity for the sharing of data amidst organizations. Due to the sharing of data the chance of the linking attack may occur. So, it designs the K-anonymity model in such a way that makes use of generalization and specialization techniques to preserve the privacy of an individual. Further, it describes the l-diversity anonymity model which is an extension of the K-anonymity model.

N. Ninghui Li, Tiancheng Li and Suresh Venkatasubramanian, "t-Closeness: Privacy beyond k-Anonymity and l-Diversity,2007". The paper states that k-anonymity cannot prevent attribute disclosure, so, l-diversity was proposed. According to l- diversity each equivalence class has at least well-represented values for each sensitive attribute. So author introduces privacy concept called t-closeness, which demands that the distribution of a sensitive attribute in any equivalence class is close to the distribution of the attribute in the overall table (i.e., the distance between the two distributions should be no more than a threshold t). The author uses the Earth Mover Distance measure for t-closeness requirement.

O. Sweeney, Latanya "Achieving K-anonymity privacy protection using generalization and suppression,2002." This paper shares the concept of the K-anonymity and the MinGen algorithm. MinGen is a theoretical algorithm that combines generalization and suppression techniques to achieve K-anonymity. Moreover, it analyses MinGen to Datafly and µ-Argus. Both Datafly and µ-Argus are the working implementation of K-anonymity.

IV. METHOD AND DATASET

In this section, we will discuss the method to preserve privacy by comparing the two methods. These methods of privacy preservation using K-anonymity consist of two modules. They are:

1) Setting up of Simulator
2) Applying K –anonymity Algorithm

A. Dataset

In this work, we have worked on the Breast Cancer Wisconsin Dataset, it is having 699 instances and 10 attributes and 2 classes, to recognize malignant (cancerous) from benign (non-cancerous) specimens. The dataset includes some classification patterns or instances with an arrangement of numerical highlights or qualities.

B. Proposed Work

We have used the K-anonymity algorithm on breast cancer datasets to compare the best privacy method for sharing data. For this, we will perform on the in-built tool "ARX" and then implement the k-anonymity algorithm in JAVA. Then we will compare the results.

1) Setting up of Simulator: For our experiment, we have used the ARX simulator and input as a breast cancer dataset. ARX is comprehensive open-source software for anonymizing sensitive personal data. The "ARX tool" is having a broad range of algorithms in-built like K-anonymity, t-closeness, l-diversity, ∂-privacy, etc. A discharge gives k-anonymity insurance if the data for every individual contained in the discharge can't be recognized from in any event (k-1) people whose data likewise shows up in the discharge [1]. The basic goals that should be achieved are to discharge the utmost information which can be utilized by third-party for analysing and experimenting purposes. To ensure that the security of no individual is placed in threat because of the released information against assuming and connecting attacks are shown in Algorithm 1.

Algorithm 1: Privacy Preservation using K-anonymity Method

a) Input

i) Source database DB,

ii) Anonymized parameter k,

b) Output: A converted dataset DB'.

c) Algorithm

i) Import the dataset into the workspace.

ii) Categorize the Quasi-identifier values.

iii) Apply the necessary conditions (hierarchies) required for the simulation.

iv) Analyze the classification accuracy of the input and the anonymized dataset.

v) Set the value of k=5.

vi) Set the suppression limit to 100%.

vii) Set the QI values to 0.5 as attribute weights.

viii) Set the generalization and suppression balanced.

ix) Finally, analyze risk check for the re-identification risk of the input and the anonymized dataset.

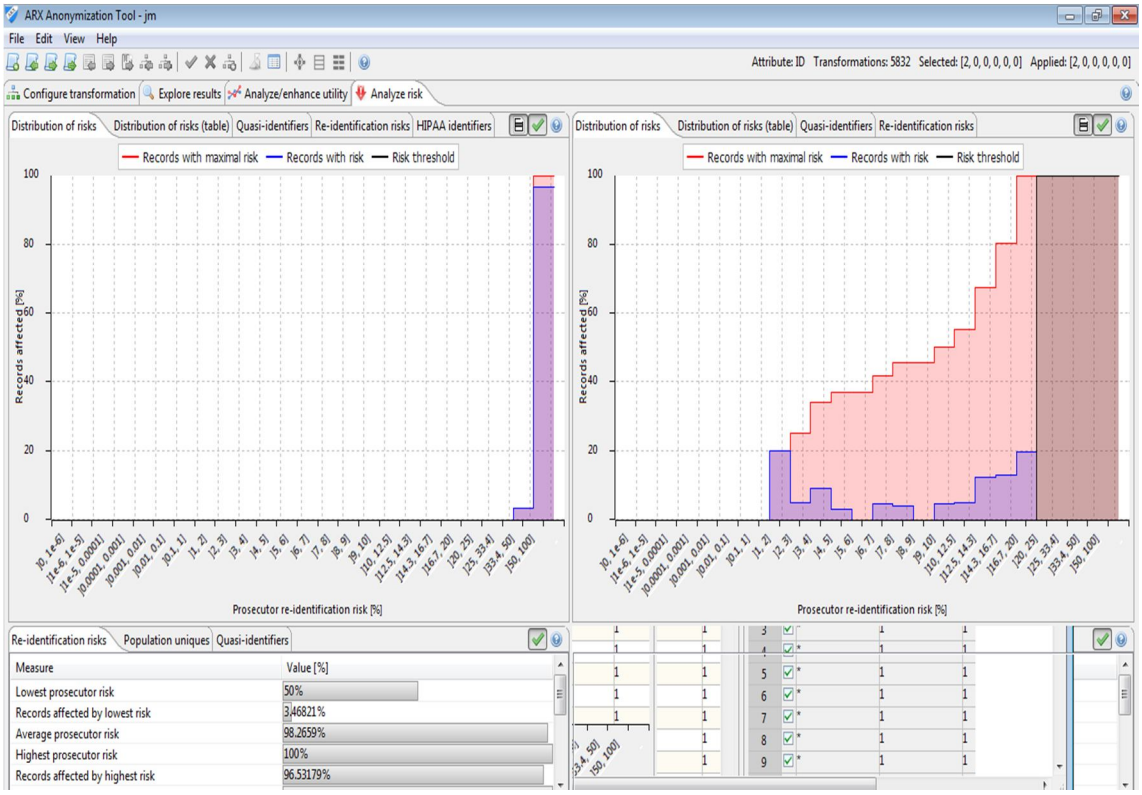


Fig 4.1: Distribution of risk/Risk Analysis

ISSN: 2321-9653; IC Value: 45.98; SJ Impact Factor: 7.177
Volume 7 Issue IX, Sep 2019- Available at www.ijraset.com

Fig 4.2: Re-identification of risk

2) *Applying K –anonymity Algorithm:* We are implementing K -anonymity algorithm using Java to check the results. Here, we will take the same input as taken in ARX, i.e., Wisconsin breast cancer data. The database is cleaned by substituting the missing qualities by zero and disposing of the excess qualities. In our situation, there is no missing value available, so no adjustment in Table 4.4. The cleaned dataset in Table 4.5 is taken as an input in the JAVA method. After obtaining the input data, we set the value of k=5 to perform K-anonymity.

The entire procedure is shown by Algorithm 2:

Algorithm 2

a) Input: A dataset T, Quasi-identifier attributes QI,

Sensitive values A, Anonymity parameter K.

b) Output: Releasing Table RT.

c) Algorithm

i) Step1: Select Data set T from a Database.

ii) Step2: Select Key attribute, Quasi-identifier attribute, and Sensitive Attribute from the given attribute list.

iii) Step3: Select the set of most sensitive values A from a list of all sensitive values that are to be preserve.

iv) Step4: For each tuple whose sensitive value belongs to set A. If t[S] ∈ A , then move all these tuples to Table.

v) Step5: Find the statistics of quasi attributes of Table, i.e., distinct value for those attributes and the total number of rows having that value.

vi) Step6: Apply generalization on quasi-identifiers of Table to make it K-Anonymized, which is an output table RT and ready to release.

Table 4.1: Cleaned data

CN	CT	UCS_h	Ma	Mi
1000000	0.5	0.1	0.1	0.1
1002900	0.5	0.4	0.5	0.1
1015400	0.3	0.1	0.1	0.1
1016300	0.6	0.8	0.1	0.1
1017000	0.4	0.1	0.3	0.1

ISSN: 2321-9653; IC Value: 45.98; SJ Impact Factor: 7.177
Volume 7 Issue IX, Sep 2019- Available at www.ijraset.com

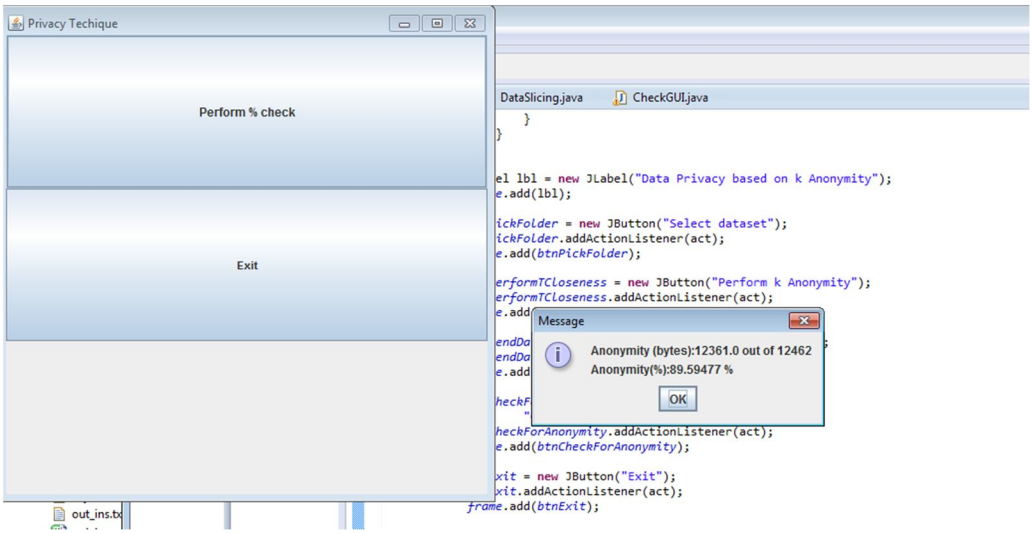


Fig 4.3: Anonymity percentage when k=5

V. RESULT & DISCUSSION

In this paper, the executed experiment is to evaluate the privacy percentage by comparing the "ARX tool" which has in-built algorithms like K-anonymity, t-closeness, l-diversity, ∂-privacy and Sweeny's K-anonymity algorithm using JAVA method.

A. The check the privacy preservation of data in the “ARX” tool, we have taken different values of K. The higher the K value, the higher the privacy preservation of data. Following is the output:

Table 5.1: ARX simulation result

Value of K	Input dataset			Anonymized output dataset		
	Records at Risk	Highest Risk	Success Rate	Records at Risk	Highest Risk	Success Rate
K=2	98.41%	100%	92.26%	84.82%	50%	39.16%
K=3	100%	100%	98.26%	59.56%	33.33%	22.02%
K=5	100%	100%	98.26%	0.00%	20%	10.30%
K=7	100%	100%	98.26%	0%	14.28%	6.81%

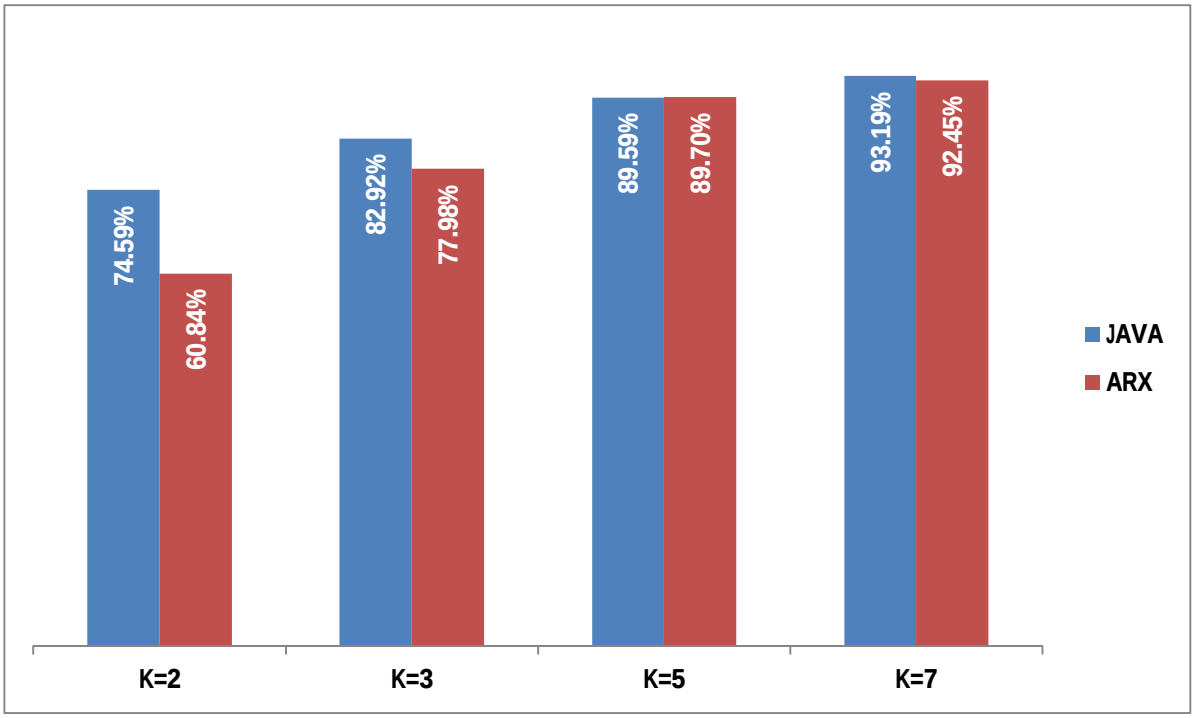
B. Now, we experiment from Java for the same k as in the ARX tool. The below table shows the result obtained by the Java process.

Table 5.2: JAVA simulation result

Value of K	Anonymity Percentage
K=2	74.59%
K=3	82.92%
K=5	89.59%
K=7	93.19%

C. After getting the result, we compare both methods in terms of the value of k and the anonymization percentage. Below is the graphical representation of the compared result.

Graphical representation of Anonymization % in Java & ARX



Hence, from the graph, we can conclude that using a k-anonymity algorithm gives better privacy over the in-built ARX tool.

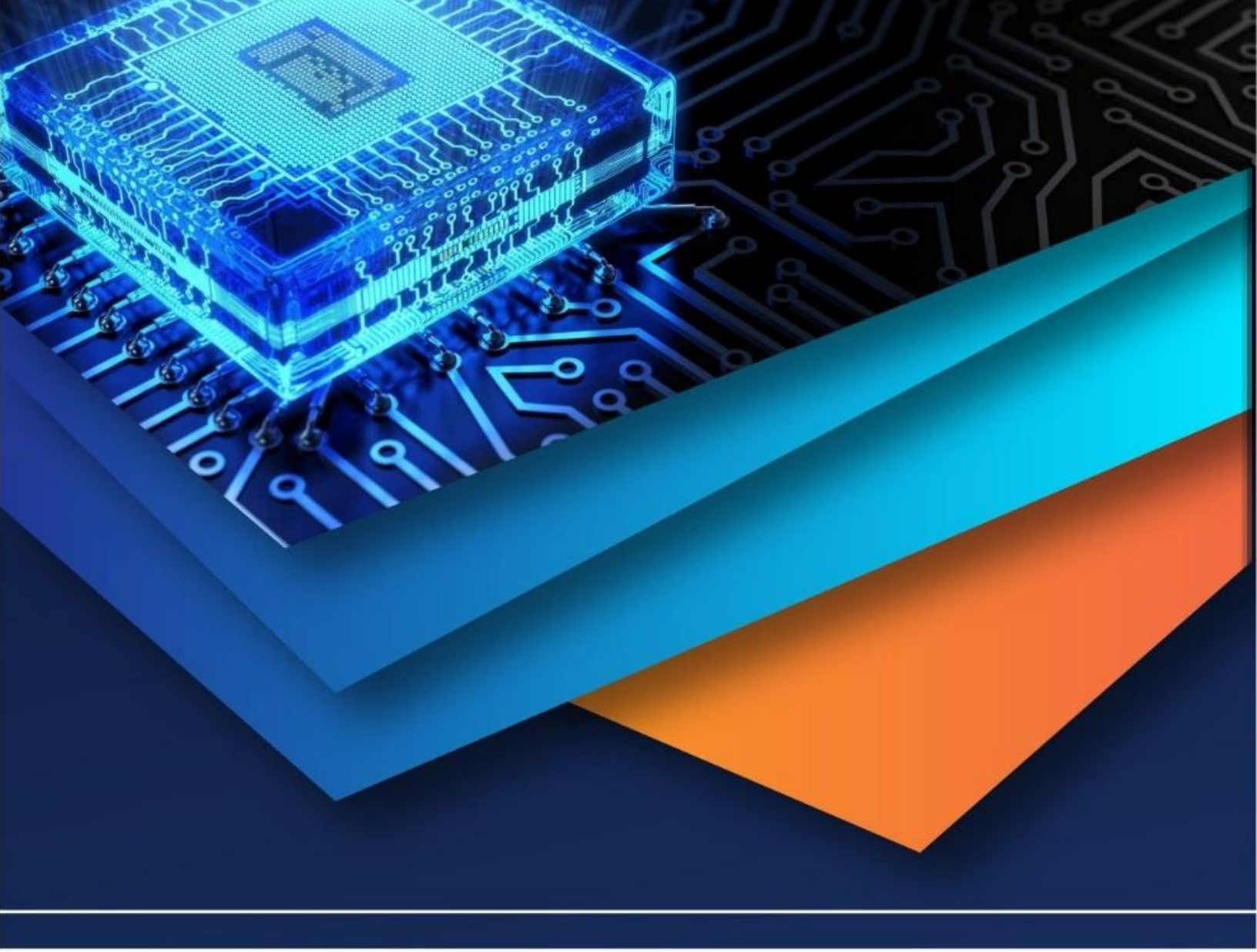
VI. CONCLUSION & FUTURE WORK

This section presents the conclusion and future scope of the present work. This paper measures the different values of k and checks for its anonymization percentage for the ARX tool and the implemented algorithm. Hence, after the result and discussion, we can conclude that using a k-anonymity algorithm gives better privacy over the in-built ARX tool. In the future, the algorithm discussed in this paper can be introduced by reducing the size of the solution space and applying improved algorithms.

REFERENCES

[1] Chatterjee, J.M., Kumar, R., Pattnaik, P.K., Solanki, V.K., & Zaman, M. (2018), “Privacy preservation in data intensive environment”, Tourism & Management Studies, 14(2), 72-79.
[2] Jyotir Moy Chatterjee (2017), “Privacy Preservation in Data Centric Environment: Analysis and Segregation”, International Journal of Engineering Research & Technology (IJERT), Vol. 6 Issue 05, May – 2017.
[3] Hetaswini J., & Vimalkumar B. Vaghela,(2017). "Privacy Preserving by Anonymization Approach”, International Journal on Recent and Innovation Trends in Computing and Communication ISSN: 2321-8169 Volume: 5 Issue: 11.
[4] Hunka, T., Dash, S., &Pattnaik, P. K. (2016). “Web based Privacy Disclosure Threats and Control Techniques”. In Design Solutions for Improving Website Quality and Effectiveness (pp. 334-341). IGI Global.
[5] Gahi, Youssef, Mouhcine Guennoun, and Hussein T. Mouftah. (2016). "Big Data Analytics: Security and privacy challenges." Computers and Communication (ISCC), IEEE Symposium on IEEE.
[6] Kavitha, S., S. Yamini, and Raja Vadhana. (2015). "An evaluation on big data generalization using kAnonymity algorithm on cloud." Intelligent Systems and Control (ISCO), IEEE 9th International Conference.
[7] Basu, Anirban, et al. (2015) "k-anonymity: Risks and the Reality" Trustcom/BigDataSE/ISPA, IEEE. Vol. 1.
[8] Zakerzadeh, Hessam, Charu C. Aggarwal, and Ken Barker. (2015). "Privacy-preserving big data publishing."Proceedings of the 27th International Conference on Scientific and Statistical Database Management. ACM.
[9] Vaibhav Lawand, Prathik Sargar, Anand Bhalerao and Pradip Jadhavn, (2015). "Analytical Approach for Privacy Preserving of Medical Data". International Journal of Engineering Research & Technology (IJERT) Vol. 4 Issue 10
[10] Freny Presswala, Amit Thakkar, Nirav Bhatt, (2015). "Survey on Anonymization in Privacy-Preserving Data Mining”. International Journal of Innovative and Emerging Research in Engineering Volume 2, Issue 2.
[11] Kupwade Patil, H., & Seshadri, R. (2014). "Big Data Security and Privacy Issues in Healthcare".IEEE International Congress on Big Data.

[12] Bhagyashri S, Gurav YB(2014). A Survey on Privacy-Preserving Techniques for Secure Cloud Storage. International Journal of Computer science and Mobile computing.
[13] Zhu, Yan, and Lin Peng. "Study on K-anonymity models of sharing medical information." Service Systems and Service Management, 2007 International Conference on. IEEE, 2007.
[14] Ninghui Li Tiancheng Li,Suresh Venkatasubramanian, (2007). “t-Closeness: Privacy Beyond k-Anonymity and l-diversity”. ICDE 2007, pp. 106–115
[15] Sweeney, Latanya. (2002). "Achieving K-anonymity privacy protection using generalization and suppression." International Journal of Uncertainty, Fuzziness and Knowledge-Based Systems 10.05: 571-588
[16] Kamakshi P, Babu AV. Preserving privacy and sharing the data in distributed environment using cryptographic technique on perturbed data. arXiv preprint arXiv:1004.4477. 2010 Apr 26.
[17] Rezgui A, Ouzzani M, Bouguettaya A, Medjahed B. Preserving privacy in web services. InProceedings of the 4th international workshop on Web information and data management 2002 Nov 8 (pp. 56-62). ACMS
[18] Dean, B.B., Lam, J., Natoli, J.L., Butler, Q., Aguilar, D., &Nordyke, R.J. (2010). Use of electronic medical records for health outcomes research: a literature review. Medical Care Research Review, 66(6), 11–38
[19] Lau, E.C., Mowat, F.S., Kelsh, M.A., Legg, J.C., Engel-Nitz, N.M., &Watson, H.N. (2011). Use of electronic medical records (EMR) for oncology outcomes research: assessing the comparability of EMR information to patient registry and health claims data. Clin. Epidemiol, 3(1), 259–272.
[20] Perer A. & Sun, J. (2012). Matrixflow: temporal network visual analytics to track symptom evolution during disease progression. In AMIA annual symposium proceedings, American Medical Informatics Association, (pp. 716-725), AMIA.



10.22214/IJRASET

45.98

IMPACT FACTOR: 7.129

IMPACT FACTOR: 7.429

INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY