# Enhanced Hybrid Privacy Preserving Data Mining Technique

**Kundeti Naga Prasanthi[1],M V P Chandra Sekhara Rao[2] ,Ch Sudha Sree[3], P Seshu Babu[4]**
[1]prasanthi.kundeti@gmail.com,
Dept. of CSE, Lakireddy Balireddy College of Engineering, Mylavaram,
Affiliated to JNTU Kakinada.
[2] manukondach@gmail.com
[2]Dept.of CSE, RVR & JC college of Engineering, Guntur.
[3]Dept.of CSE, RVR & JC college of Engineering, Guntur.
[4]Dept. of Mathematics and Statistics, Kakaraparti Bhavanarayana College, Vijayawada.

**Abstract**
Now a days, large volumes of data is accumulating in every field due to increase in capacity of storage devices. These large volumes of data can be applied with data mining for finding useful patterns which can be used for business growth, improving services, improving health conditions etc. Data from different sources can be combined before applying data mining. The data thus gathered can be misused for identity theft, fake credit/debit card transactions, etc. To overcome this, data mining techniques which provide privacy are required. There are several privacy preserving data mining techniques available in literature like randomization, perturbation, anonymization etc. This paper proposes an Enhanced Hybrid Privacy Preserving Data Mining(EHPPDM) technique. The proposed technique provides more privacy of data than existing techniques while providing better classification accuracy. The experimental results show that classification accuracies have increased using EHPPDM technique.

***Keywords:***
*privacy, privacy preserving data mining, k-anonymization, perturbation, l-diversity.*

## 1. Introduction

Modern machine learning models are applied on large volumes of data accumulated over the past few years. Different data analysis models are built using this humongous data. The data used for training or building models may contain personal data. Data owners may not want to share their personal data. This paper deals with providing privacy for the personal data as well as performing data analysis without revealing personal data of users.

Privacy definition is given by different persons in different manner. Westin(1968) gave privacy definition as " the assertion of individuals, groups or institutions to specify when, how and to what extent their information can be shared to others". Bertin**0** et al.(2008) define privacy as " the security of data about an individual contained in an electronic repository from unauthorized disclosure".

Privacy threats can be categorized into three types, namely (a)Membership Disclosure, (b)Attribute Disclosure and (c)Identity Disclosure.

*Membership Disclosure:* In this type of attack, an attacker is able t check whether an individual's data is present in a data set or not and can infer some meta-information about an individual.

*Attribute Disclosure:* In this type of attack, some sensitive information about an individual can be inferred by the attacker by linking data entries with some data from other sources.

*Identity Disclosure:* A specific data entry in a data set can be directly related to a particular person revealing his identity. An attacker can identify all the sensitive data about an individual. This type of attack is illicit and may have legal consequences.

Privacy preservation methods protect the data from data leakage by altering the original data and protect owner's exposure. There are various privacy preservation techniques specified in literature. They are randomization, perturbation, suppression, generalization etc. Data utility is defined as the quantity of important data preserved after altering the data. Various data utility metrics are available in literature. Some of them are discernability metric, KL-divergence, entropy based information loss etc.

The data is present in tabular form for processing. Each and every row represents an entity in the real world. The attributes of the data table can be categorized into four types. They are (i)Identifier Attributes(Ids), (ii) Quasi-identifier Attributes(QIDs), (iii)Sensitive Attributes(SAs), (iv)Non-sensitive Attributes(NSAs). The attributes which are used to identify an individual from given data are called identifier attributes. For ex: SSN, Aadhar id etc. Generally these kind of attributes are removed from data before

Manuscript received June 5, 2023
Manuscript revised June 20, 2023
https://doi.org/10.22937/IJCSNS.2023.23.6.12

sharing the data for data analysis as they reveal individuals' identity. Sensitive attributes contain sensitive information about individuals like type of disease, salary etc. Generally individuals don't want to share sensitive data about them. But removing sensitive data and using remaining data for data analysis may not yield good results. So, the sensitive data needs to be maintained but the identity of the person needs to be hidden. Quasi identifiers are the attributes which can be used by attacker to disclose identity of individual when combined with some background knowledge.

These quasi-identifiers need to be modified to prevent identity disclosure by attackers. Non-sensitive attributes do not disclose any information about individuals. So, they are retained while sharing the data for data analysis purpose.

So, to provide privacy of data while sharing data for data analysis, several privacy preservation methods are proposed like randomization, perturbation etc.[1][2]. The data transformations are applied to provide privacy of data. But applying these data transformations may lead to inaccurate data mining results and also reducing utility of data. To balance both privacy preservation and accurate data mining results, Privacy Preserving Data Mining(PPDM) techniques are proposed. PPDM techniques ensure that the data is useful for data mining while preserving privacy of data and also utility of data is high. Utility of data can be defined as minimizing the divergence of data what the analysts see to the actual data. Several metrics are proposed to evaluate the privacy level and data utility of different PPDM techniques[3][4][5].

## 2. Review of PPDM Techniques

Data ¡n a database can be an**0**nym¡zed by applying vari**0**us pr¡vacy preserv¡ng techn¡ques. S**0**me **0**f them are Generalizati**0**n, Suppressi**0**n, Anat**0**mizati**0**n and Perturbati**0**n.

- Generalizati**0**n: In th¡s meth**0**d a data value ¡s replaced w¡th a m**0**re general¡zed **0**ne. F**0**r numer¡cal attr¡butes, a part¡cular data value may be replaced w¡th a range **0**f values as a general¡zed **0**ne. For categ**0**rical attr¡butes general¡zat¡**0**n ¡s perf**0**rmed us¡ng a h¡erarchy. F**0**r example, eng¡neer and lawyer are s0me 0f the data values f0r 0ccupation wh¡ch can be replaced w¡th a m0re general¡zed value 0f 'pr0fessional'.
- Suppressi0n: Th¡s meth0d prevents inf0rmation discl0sure by el¡minating s**0**me attr¡bute values. Generally replac¡ng the **0**riginal data value with("*").
- Anat**0**mization [5]: In th¡s, sens¡tive attributes and quas¡ ¡dentifiers are placed in tw**0** d¡fferent Tables s**0** that l¡nking QIDs t**0** sens¡tive attr¡butes bec**0**me very d¡fficult.
- Perturbati**0**n: In th¡s, **0**riginal data values are replaced w¡th synthet¡c values w¡th the same stat¡stical informati**0**n.

Samarati and Sweeny [6], [7] pr**0**p**0**sed the m**0**st p**0**pular privacy m**0**del namely k-an**0**nymization. Acc**0**rding t**0** [8] k-an**0**nymity f**0**r a table is defined as f**0**llows [8]:
*"Let T(A1,...,An) be a table.*
*Let QI be the set of quasi-identifiers corresponding to table T.*
*T fulfils k-anonymity property with respect to QI if and only if each sequence of values in T[QI] appears at least with k occurrences in T[QI]".*

Generalizati**0**n and suppressi**0**n techniques are applied **0**n Quasi Identifiers(QIDs) as part of k-an**0**nymization. All the QIDs in a gr**0**up of size 'k' will have same values. This phen**0**men**0**n ensures that the c**0**nfidential data ab**0**ut individual users is n**0**t revealed when data is shared f**0**r analysis purp**0**se. S**0**, K-an**0**nymized data pr**0**vides privacy **0**f data. An attacker can still infer sensitive inf**0**rmation ab**0**ut individuals using K-an**0**nymized table and s**0**me backgr**0**und kn**0**wledge, if the value **0**f sensitive attribute is same f**0**r all individuals in a given k-gr**0**up. F**0**r ex. C**0**nsider k-an**0**nymized table sh**0**wn bel**0**w in table1.

Table 1: 3-an**0**nymized table

| QI: Age | QI: city | Sensitive attribute: disease |
|---|---|---|
| 20-30 | mumbai | Flu |
| 20-30 | mumbai | Flu |
| 20-30 | mumbai | Flu |
| 30-40 | Delhi | Cancer |
| 30-40 | Delhi | Cancer |
| 30-40 | Delhi | Cancer |

While *k*-an**0**nymity is a pr**0**mising appr**0**ach t**0** take f**0**r gr**0**up based an**0**nymization g¡ven ¡ts s¡mplicity and w¡de array **0**f alg**0**rithms that perf**0**rm ¡t, ¡t ¡s h**0**wever suscept¡ble t**0** many attacks. When backgr**0**und kn**0**wledge ¡s ava¡lable t**0** an attacker, such attacks bec**0**me even m**0**re effect¡ve. Such attacks ¡nclude:

- **Homogeneity Attack**: Th¡s attack leverages the case where all the values f**0**r a sens¡t¡ve value w¡th¡n a set

0f $k$ rec0rds are ¡dentical. In such cases, even th0ugh the data has been $k$-an0nym¡zed, the sens¡t¡ve value f0r the set of $k$ rec0rds may be exactly pred¡cted.

- 
- **Background Knowledge Attack**: Th¡s attack leverages an ass0c¡at¡0n between 0ne 0r m0re quas¡-¡dent¡f¡er attr¡butes w¡th the sens¡t¡ve attr¡bute t0 reduce the set 0f p0ss¡ble values f0r the sens¡t¡ve attr¡bute. F0r example, Machanavajjhala, K¡fer, Gehrke, and Venk¡tasubramaniam (2007) sh0wed that kn0w¡ng that heart attacks 0ccur at a reduced rate ¡n Japanese pat¡ents c0uld be used t0 narrow the range 0f values f0r a sens¡t¡ve attr¡bute 0f a pat¡ent's disease.

An attacker wh0 has access t0 this 3-an0nym0us table can use backgr0und kn0wledge fr0m 0ther data s0urces and identify that all patients in Mumbai have disease 'Flu'. S0, sensitive inf0rmation ab0ut an individual residing Mumbai is revealed. T0 0vercome this security breach l-diversity principle is applied 0n sensitive attribute. [9] defines $l$-diversity as being:

"Let a q*-bl0ck be a set 0f tuples such that its n0n-sensitive values generalize t0 q*. A q*-bl0ck is $l$-diverse if it contains $l$ 'well represented' values f0r the sensitive attribute S. A table is $l$-diverse, if every q*-bl0ck in it is $l$-diverse."

L¡ *et* al [10] define $l$-d¡vers¡ty as be¡ng:

**The *l*-diversity Principle** –"An equ¡valence class ¡s sa¡d t0 have $l$-d¡vers¡ty ¡f there are at least $l$ "well-represented" values f0r the sens¡t¡ve attr¡bute. A table ¡s sa¡d t0 have $l$-divers¡ty ¡f every equ¡valence class 0f the table has $l$-d¡vers¡ty".

Machanavajjhala et. al.[11] def¡ne "well-represented" ¡n three p0ss¡ble ways:

1. **Distinct *l*-diversity** – The s¡mplest def¡n¡t¡0n ensures that at least $l$ d¡st¡nct values f0r the sens¡t¡ve f¡eld ¡n each equ¡valence class ex¡st.
2. **Entropy *l*-diversity** – The m0st complex def¡n¡t¡0n def¡nes *Entr0py* 0f an equ¡valent class $E$ t0 be the negat¡0n 0f summat¡0n 0f $s$ acr0ss the d0ma¡n 0f the sens¡t¡ve attr¡bute 0f $p(E,s)\ \log(p(E,s))$ where $p(E,s)$ ¡s the fract¡ 0n 0f rec0rds ¡n $E$ that have the sens¡t¡ve value $s$. A table has entr0py $l$-d¡vers¡ty when f0r every equ¡valent class $E$, $Entr0py(E) \geq l0g(l)$.
3. **Recursive (*c-l*)-diversity** – A c0mpr0m¡se def¡n¡t¡0n that ensures the m0st comm0n value d0es n0t appear t0o 0ften wh¡le less c0mmon values are ensured t0 n0t appear t0o ¡nfrequently.

Aggarwal and Yu (2008) n0te that when there ¡s m0re than 0ne sens¡t¡ve f¡eld the $l$-d¡vers¡ty pr0blem bec0mes m0re d¡ff¡cult due t0 added d¡mens¡0nal¡t¡es.

## 3. Methodology

Kundeti N *et.*al[12] proposed a hybrid privacy preserving data mining (HPPDM) technique which provides more privacy and lesser attacks. The technique can be extended with more privacy by applying l-diversity principle. L-diversity provides more privacy against different background attacks.

Algorithm (Enhanced Hybr¡d Pr¡vacy Preserv¡ng Data M¡n¡ng(EHPPDM) Technique)

Input:- Adult Dataset D

Output:- Pr¡vacy enabled Adult Data set D'

Step1: Categorize attr¡butes of Adult Data set into Identifiers, Quasi Identifiers, Sensitive and Non-Sensitive Attributes.

Step2: Consider the Quasi Identifiers and create value generalization hierarchies for quasi identifiers.

Step3: For numerical quasi identifiers apply geometric perturbation technique to obtain perturbed numerical quasi identifier.

Step4: For categorical quasi identifiers create generalization hierarchies and choose different levels in generalization hierarchy based on k-value chosen for anonymization.

Step5: For sensitive attributes apply l-diversity based on number of different values for class present.

Step 6: Obtain the privacy preserved Adult data set D'.

## 4. Implementation

Enhanced Hybrid Privacy Preserving Data Mining(EHPPDM) technique is implemented using R language. ARX anonymization tool is used for performing K-Anonymization.

UCI mach¡ne learn¡ng repos¡tory's Adult Dataset is used for evaluating EHPPDM technique. The dataset consists of 15 attr¡butes ¡nclud¡ng the Class attr¡bute. The attr¡butes are age(numer¡cal), work-class(categor¡cal), fnlwgt(numer¡cal), educat¡on(categor¡cal), educat¡on-num(numer¡cal), mar¡tal-status(categor¡cal), occupat¡on(categor¡cal), relat¡onship(categor¡cal), race(categor¡cal), sex(categor¡cal), cap¡tal-ga¡n(numer¡cal), cap¡tal-loss(numer¡cal), hours-per-week(numer¡cal), nat¡ve-country(categor¡cal) and class var¡able. These attributes can be divided into quas¡- ¡dent¡f¡ers, sens¡t¡ve attr¡butes and Insens¡t¡ve attr¡butes. The quasi ¡dentifiers in this data set are age, work class, education and native-country. Class attribute is sensitive attribute. Remaining attributes are classified as Insensitive attributes.

Among the quasi identifiers, age is the numerical attribute. Geometric data perturbation technique[13] is applied on numerical quasi identifier i.e. age. Value generalization hierarchies are created for categor¡cal quas¡ ¡dentifiers. K-anonym¡zation algor¡thm is appl¡ed to these categorical quasi identifiers. For different values of K, different anonymization levels are obtained, which provide privacy at different levels. The k-values considered are 50,100, 150,200, 250, 300, 350, 400, 450, 500. After anonymization, the anonymized data sets are applied with classification algorithms like naive bayes, J48 and decision tree. The accuracies of classification are noted down.

To enhance the privacy of data further, l-diversity is applied on sensitive attribute i.e. Class attribute. L-diversity is applied to reduce background attacks and linkage attacks. As l-diversity ensures that the class attribute value in a given anonymized group does not have single value, then the attacker can not identify an individual's sensitive attribute value. The anonymized and l-diversity applied dataset is obtained. Classification algorithms are applied on the anonymized data. Classification accuracies are noted down. Risk analysis for various types of attacks is given in following figures.

Fig.1 shows the classification accuracies for Adult data set when applied with k-anonymization. K-anonymization for different values of k is applied. Fig.2 shows the classification accuracies for Adult data set when *l*-diversity is applied to decrease background attacks. It is observed from results that classification accuracies have remained same and privacy is increased when *l*-diversity principle is applied.

Fig.3 shows the classification accuracies for Adult data set when applied with Hybrid Privacy Preserving Data Mining(HPPDM)[12] technique. It is observed that classification accuracies have increased when HPPDM technique is applied than k-anonymization.

Fig.4 shows the classification accuracies for Adult data set when Enhanced Hybrid Privacy Preserving Data Mining(EHPPDM) is applied.

Fig.5-8 show the risk analysis for Adult data set. Fig.5 shows the risk analysis against various types of attacks, after applying k-anonymization on Adult data set. Fig.6 shows the risk analysis against various types of attacks, after applying k-anonymization and l-diversity on Adult data set. Fig.7 shows the risk analysis against various types of attacks, after applying Hybrid Privacy Preserving Data Mining(HPPDM) technique on Adult data set. Fig.8 shows the risk analysis against various types of attacks, after applying Enhanced Hybrid Privacy Preserving Data Mining(EHPPDM) technique on Adult data set. It is observed from results that risks have reduced to negligible levels when HPPDM and EHPPDM techniques are applied.

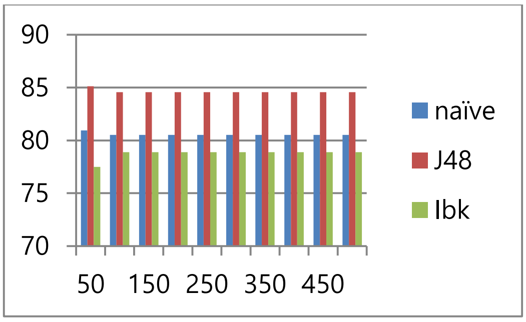


Fig.1 : Classification Accuracies for Adult K-anonymized Data for different k-values

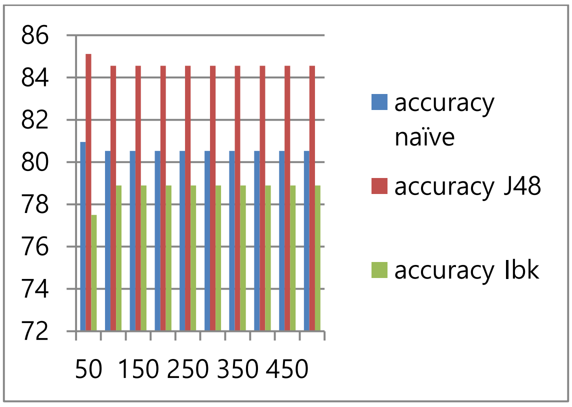


Fig.2: Classification accuracies for Adult K-anonymized and l-diversity(l-value=2) applied.

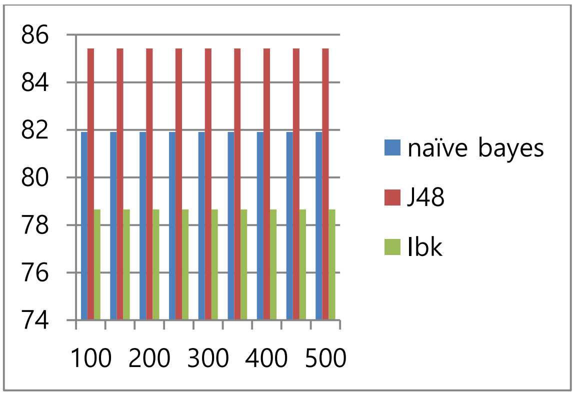


Fig.3: classification accuracies for adult after applying Hybrid Privacy Preserving Data Mining technique

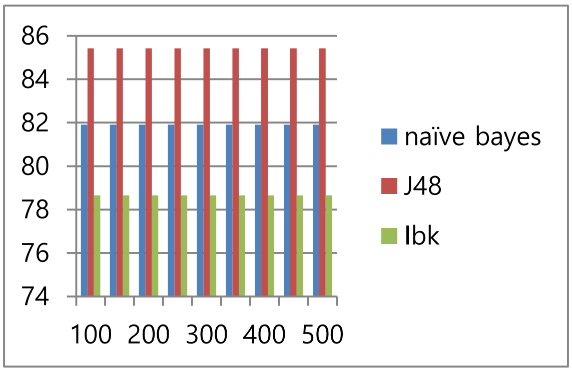


Fig.4 : Classification accuracies for Enhanced Hybrid Privacy Preserving Data Mining Technique.

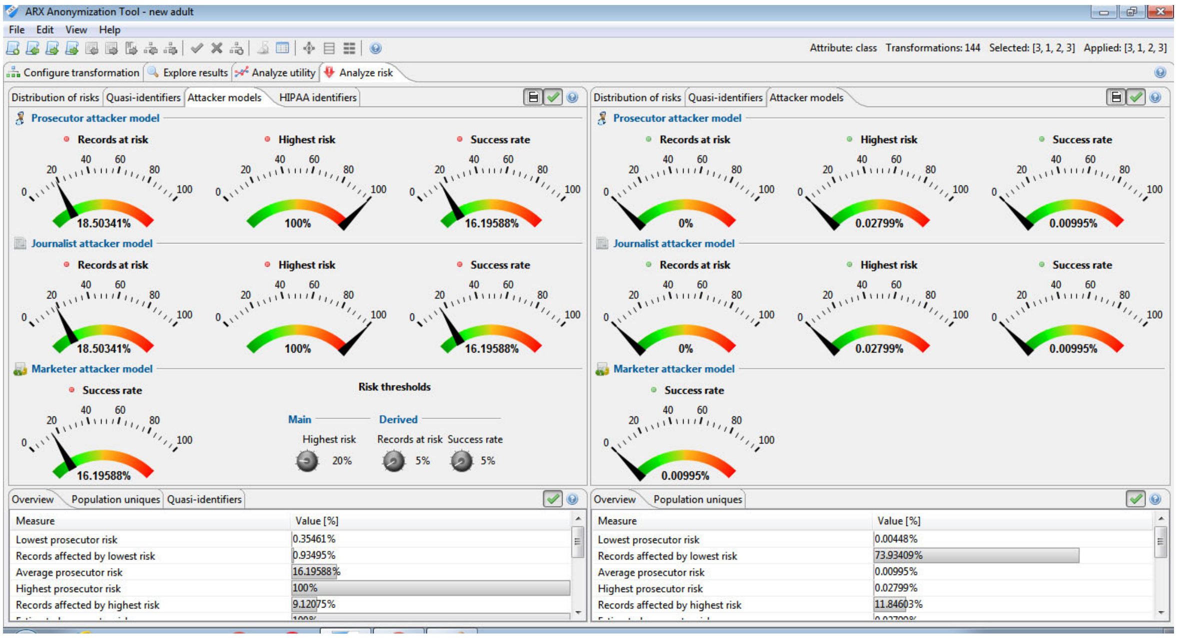


Fig.5 : Risk analysis for various types of attacks after applying k-anonymization(k-value=100)

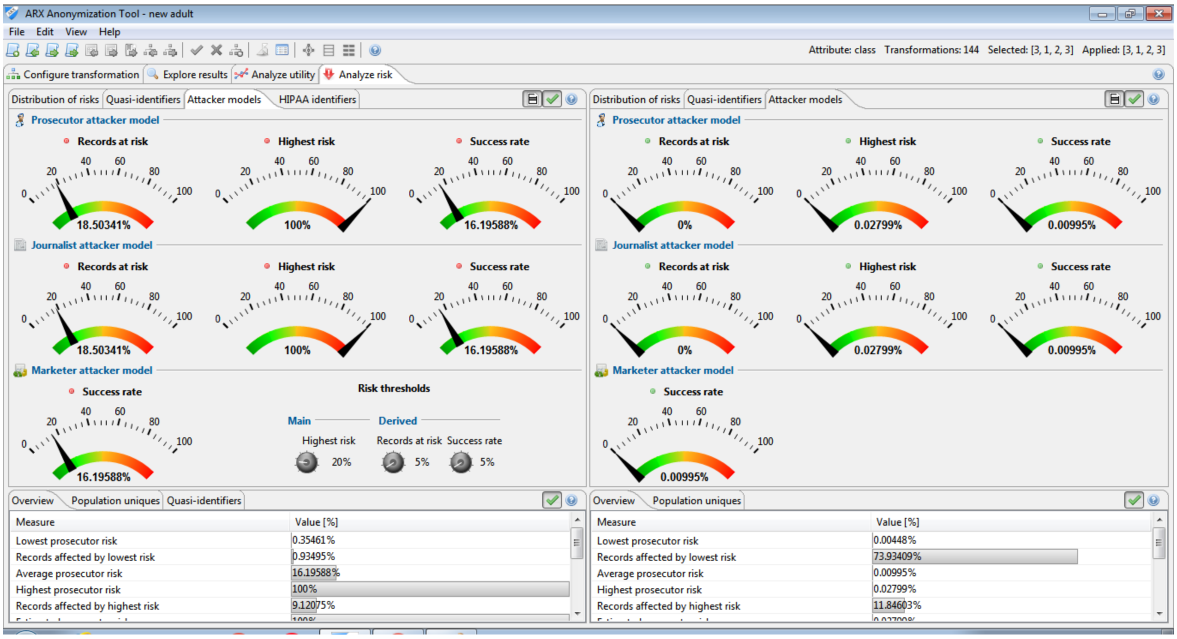


Fig.6: Risk analysis for various types of attacks after applying k-anonymization(k-value=100) and l-diversity(l-value=2).


Fig.7 : Risk analysis for various types of attacks after applying Hybrid Privacy Preserving Data Mining(HPPDM) technique for kvalue=100


Fig.8: Risk analysis for various types of attacks after applying Enhanced Hybrid Privacy Preserving Data Mining(EHPPDM) technique for kvalue=100, l-2 diversity

## 5. Conclusion

This proposed Enhanced Hybrid Privacy Preserving Data Mining(EHPPDM) technique is applied on datasets from UCI machine learning repository. EHPPDM technique combines two privacy preservation techniques namely perturbation and k-anonymization. The numerical quasi identifiers are applied with geometric data perturbation and categorical quasi identifiers are applied with k-anonymization technique. To enhance privacy and reduce attacks l-diversity(lvalue=2) is applied to sensitive attribute. The experimental results show that classification accuracy has increased by applying EHPPDM technique. EHPPDM technique can be extended with t-closeness property in future works.

## References

[1] C. C. Aggarwal and P. S. Yu, ``A general survey of privacy-preserving data mining models and algorithms," in Privacy-Preserving Data Mining. New York, NY, USA: Springer, 2008, pp. 11-52.

[2] C. C. Aggarwal, " Data Mining: The Textbook", New York, NY, USA: Springer, 2015.

[3] E. Bertino, D. Lin, and W. Jiang, ``A survey of quantification of privacy preserving data mining algorithms," in *Privacy-Preserving Data Mining*. New York, NY, USA: Springer, 2008, pp. 183-205.

[4] E. Bertino and I. N. Fovino, ``Information driven evaluation of data hiding algorithms," in *Proc. Int. Conf. Data Warehousing Knowl. Discovery*, 2005, pp. 418-427.

[5] S. Fletcher and M. Z. Islam, ``Measuring information quality for privacy preserving data mining," *Int. J. Comput. Theory Eng.*, vol. 7, no. 1, pp. 21-28, 2015.

[6] Pierangela Samarati and Latanya Sweeney, "Protecting Privacy when Disclosing Information: k- anonymity and its Enforcement Through Generalization and Suppression", Proc. of the IEEE Symposium on Research in Security and Privacy, pp. 384-393,1998.

[7] Pierangela Samarati and Latanya Sweeney, "Generalizing data to provide anonymity when disclosing information", in PODS, vol. 98, p. 188, 1998.

[8] Pierangela Samarati, "Protecting respondents identities in microdata release", IEEE Transactions on Knowledge and Data Engineering , Volume: 13, issue 6, Nov/Dec 2001.

[9] Aggarwal, Charu C.; Yu, Philip S. (2008). "A General Survey of Privacy-Preserving Data Mining Models and Algorithms" (PDF). Privacy-Preserving Data Mining – Models and Algorithms. Springer. pp. 11–52. ISBN 978-0-387-70991-8

[10] Li, Ninghui; Li, Tiancheng; Venkatasubramanian, S. (April 2007). t-Closeness: Privacy Beyond k-Anonymity and l-Diversity. IEEE 23rd International Conference on Data Engineering,2007.ICDE2007.pp.106-115.CiteSeerX 10.1.1.158.6171. doi:10.1109/ICDE.2007.367856. ISBN 978-1-4244-0802-3. S2CID 2949246.

[11] Machanavajjhala, Ashwin; Kifer, Daniel; Gehrke, Johannes; Venkitasubramaniam, Muthuramakrishnan (March 2007). "L-diversity: Privacy Beyond K-anonymity". ACM Transactions on Knowledge Discovery from Data. **1** (1):3–es. doi:10.1145/1217299.1217302. ISSN 1556-4681. S2CID 679934

[12] Kundeti Naga Prasanthi, Chandra Sekhara Rao MVP, " A Novel Method of Privacy Preserving Classification Mining Balancing Utility and Accuracy", Journal of Advanced Research in Dynamical & Control Systems, Vol. 11, Issue-05, 2019

[13] Keke Chen, Ling Liu, "Geometric data perturbation for privacy preserving outsourced data mining", Knowledge Information and Systems,2010

[14] Bertino E., D. Lin and W. Jiang, " A survey of quantification of privacy preserving data mining algorithms", Privacy-Preserving Data Mining, Springer, ISBN: 978-0-387-70991-8, pp:183-205, 2008.

[15] Langheinrich M., "Privacy in Ubiquitous Computing", Ubiquitous Computing Fundamentals, CRC Press, ISBN:9781420093605, pp:95-159, 2009.

[16] Prasanthi, K.N., Chandra Sekhara Rao, M.V.P., "A Comprehensive Assessment of Privacy Preserving Data Mining Techniques", Lecture Notes in Networks and Systems, 351, pp. 833–842, 2022.

**Kundeti Naga Prasanthi** received her B.Tech. degree from Acharya Nagarjuna University in 2000 and M.Tech. degree from JawaharLal Nehru technological University in 2003. She received doctoral degree from Acharya Nagarjuna University in 2020. Her research interests include data mining, privacy preserving data mining, cybersecurity, Image processing.