iJRASET
International Journal For Research in Applied Science and Engineering Technology

INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY

Volume: 3 **Issue:** II **Month of publication:** February 2015

DOI:

www.ijraset.com

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

ISSN: 2321-9653
International Journal for Research in Applied Science & Engineering Technology (IJRASET)

Privacy Preserving in Data Mining

Pranoti S. Talokar[1], Renuka D. Mane[2], Darshana K. Patil[3]
Computer Science Department, H.VPM's COET Amravati

Abstract*—Data mining is an increasingly important technology for extracting useful knowledge hidden in large collection of data. It is today well observe that database represent important role in many application and for this reason its security becomes compulsory. To keep the data very confidentially is relevant because of value and need .In recent years, many organization have data collector and collect the data in digital form. .The problem with data mining output is that it also reveals some information, which are considered to be private and personal. Easy access to such personal data poses a threat to individual privacy. There has been growing concern about the chance of misusing personal information behind the scene without the knowledge of actual data owner.Privacy is becoming an increasingly important issue in many data mining applications in distributed environment. Privacy preserving data mining technique gives new direction to solve this problem. In this paper we have given a brief discussion on different privacy preservation techniques and their advantages and disadvantages.

***Keywords*— Data Mining, Privacy, Cryptography, Data perturbation, k-Anonymity.**

I. INTRODUCTION

Data mining is a powerful tool that can investigate and extract previously unknown patterns from large amounts of data. The process of data mining requires a large amount of data to be collected into a central site. In modern days organizations are extremely dependent on data mining in results to provide better service, achieving greater profit, and better decision-making. For these purposes organizations collect huge amount of data. Huge volumes of Data collected in this manner also include sensitive data about individuals. It is obvious that if a data mining algorithm is run against the union of different databases, the extracted knowledge not only consists of discovered patterns and correlations that are hidden in the data but it also reveals the information which is considered to be private. Privacy is an important issue in many data mining applications like which are deals with health care, security, financial and other types of sensitive data. Privacy preserving data mining technique gives new direction to solve this problem. PPDM is a popular search area that "develops technique for modifying original data in some way, so that the private data and knowledge remain private even after mining process."

PPDM can be considered in two aspects:

Protecting sensitive data values, e.g., names, social security numbers of some people etc.

Protecting confidential knowledge in data, e.g., hiding confidential knowledge and not affecting the non- confidential knowledge and data utilities.

There are three different techniques of privacy preserving in data mining.

Data Perturbation

K-anonymity

Cryptography

II. DIFFERENT TECHNIQUES OF PRIVACY PRESERVING DATA MINING

A huge number of methods for privacy preserving data mining have been proposed .

A. Data Perturbation

Data perturbation is a data security technique that adds 'noise' to databases allowing individual record confidentiality. In Data Perturbation technique Organizations store large amounts of data, and most may be considered confidential. Thus, security and protection of the data is a concern. This concern applies not just to those who are trying to access the data illegally, but to those who should have legitimate access to the data. Our interest in this area relates to restricting access of confidential database attributes to legitimate organizational users(i.e., data protection).Data perturbation techniques are statistically based methods that seek to protect confidential data by adding random noise to data, there by protecting the original data. Note that these techniques are not encryption techniques, where the, data is first modified, then (typically) transmitted, and then received,'decrypted' back to the original data. Techniques that seek to accomplish masking of individual confidential data elements while maintaining underlying aggregate relationships of the database are called **data perturbation techniques**. These techniques modify actual data values to 'hide' specific confidential individual record information.

In Data Perturbation section we will describe the various computation techniques which we are using for data. Two techniques

are use in Data Perturbation

1) *Additive Perturbation***:** Data perturbation is one of the most popular models is used in privacy preserving data mining. It is specially convenient for applications where the data owners need to export/publish the privacy of sensitive data.

Definition : In this noise is added to the data records.
For example ,
1 noise 'v' is added to each tupple 't' in the perturbed data set, which equals (t+v),is similar but not equal to the original one .
2 Z = X+Y
X is the original value, Y is random noise and Z is the perturbed value Data Z and the parameters of Y are published.

2) *Random substitution:* Defination: In this random substitution of values is done in order to perturb the records. For example, if possible values {m, n, p} of the attribute Name take the substitution rule as {m → p, n → m, p → n}, and thus the value of m, n, p substituted according to this rule. Randomized data perturbation technique allows Systemic transformation of original data and the the modified data is then submitted as a result of client's query.Using this approach we can achieve confidentiality at client as well as data owner sites. In randomization perturbation approach the Privacy of data can be protected by perturbing the Sensitive data.

For example-

We treat the original values (x1,x2,...,xn) from a column to be randomly drawn from a random variable X, which has some kind of distribution. The randomization processs changes the original data by adding random noises R to the original data values,and generates a perturbed data column Y, Y = X + R.The resulting record (x1+r1, x2+r2,...,xn+rn) and the distribution of R are published. Adding of Random Noise is used in this Technique.

a) Advantages:
i) It is very simple technique.
ii) Different attributes are treated independently.

b) Disadvantages:
i) Does not reconstruct the original vale rather than only distortion.
ii) The perturbation approach does not provide a clear understanding of the level of indistinguishability of different records.

B. k-anonymity

Anonymization means identifying information is removed from the original data to protect personal or private information. There are many ways to perform data anonymization basically this method uses k-anonymization approach .If each row in the table cannot be distinguished from at least other k-1 rows by only looking a set of attributes, then this table is K-anonymized on these attributes.Many organizations are increasingly publishing micro data – tables that contain sensitive information about Individuals.However, if individuals can be uniquely identified in the micro data then their private information (such as their medical condition) would be disclosed , and this is unacceptable. To avoid the identification of records in micro data, uniquely identifying information like names and social security numbers can be removed from the table. But, it still does not ensure the privacy of individuals in the data. Because we can re-identify that data by linking. Below is a demonstration of how such data can be re-identified.

Example 1. The National Association of Health Data Organizations (NAHDO) reported that 37 states in the USA have legislative mandates to collect hospital level data and that 17 states have started collecting ambulatory care data from hospitals, physicians offices, clinics, and so forth. The leftmost circle in Figure 1 contains a subset of the fields of information, or attributes, that NAHDO recommends these states collect; these attributes include the patient's ZIP code, birth date, gender, and ethnicity. The leftmost circle in Figure 1 contains a subset of the fields of information, or attributes, that NAHDO recommends these states collect; these attributes include the patient's ZIP code, birth date, gender, and ethnicity. In Massachusetts, the Group Insurance Commission (GIC) is responsible for purchasing health insurance for state employees. GIC collected patient- specific data with nearly one hundred attributes per encounter along the lines of those shown in the leftmost circle of Figure 1 for approximately 135,000 state employees and their families. Because the data were believed to be anonymous, GIC gave a copy

of the data to researchers and sold a copy to industry . For twenty dollars I purchased the voter registration list for Cambridge Massachusetts and received the information on two diskettes.

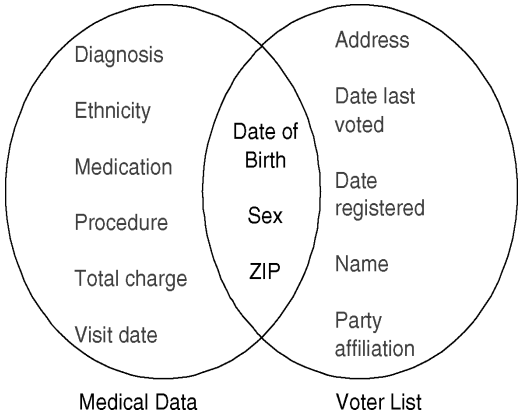


Fig.1 Linking to re-identify record owner[2,3]

The rightmost circle in Figure 1 shows that these data included the name, address, ZIP code, birth date, and gender of each voter. This information can be linked using ZIP code, birth date and gender to the medical information, there by linking diagnosis, procedures, and medications to particularly named individuals. For example, William Weld was governor of Massachusetts at that time and his medical records were in the GIC data. Governor Weld lived in Cambridge Massachusetts. According to the Cambridge Voter list, six people had his particular birth date; only three of them were men; and, he was the only one in his 5-digit ZIP code.

To counter linking attacks using quasi- identifiers, Samarati and Sweeney proposed a definition of privacy called k-anonymity[7, 8]. A table satisfies k- anonymity if every record in the table is indistinguishable from at least k − 1 other records with respect to every set of quasi-identifier attributes; such a table is called a k- anonymous table. Hence, for every combination of values of the quasi-identifiers in the k-anonymous table, there are at least k records that share those values. This ensures that individuals cannot be uniquely identified by linking attacks.

C. Methods For Achieving K-Anonymity

There are two methods for Achieving K-anonymity
Suppression
Generalization

1) *Suppression:* In Suppression technique sensitive data value are removed or suppressed before published. Suppression is used to protect an individual privacy from intruders attempt to accurately predict a suppressed value. Information loss is an important issue in suppression by minimizing the number of values suppressed[9][10]. In this method certain values of the attributes are replaced by asterisk '*'.All or some values of column may be replace by '*'.
2) *Generalization:* Aggregation is also known as generalization or global recording. It is used for protecting an individual privacy in a released data set by perturbing the original data set before its releasing. Aggregation change k number of records of a data by representative records. The value of an attribute in such a representative record is generally derived by taking the average of all values, for the attributes, belonging to the records that are replaced. Another method of aggregation or generalization is transformation of attribute values. For ex- an exact birth date can be changed by the year of birth. Such a generalization makes an attribute value less informatics. In this method individuals values of attributes are replace by the broader category[9][10].

An Example, medical records from a fictitious hospital located in upstate New York. Note that the table contains no uniquely identifying attributes like name, social security number, etc. In this example, we divide the attributes into two groups: the sensitive attributes (consisting only of medical condition) and the non-sensitive attributes (zip code, age, and nationality). An attribute is marked sensitive if an adversary must not be allowed to discover the value of that attribute for any individual in the dataset. Attributes not marked sensitive are non-sensitive. Furthermore, let the collection of attributes {zip code, age, nationality}

be the quasi-identifier for this dataset. (here "*" denotes a suppressed value so, for example, "zip code = 1485*" means that the zip code is in the range [14850−14859]and "age=3*" means the age is in the range [30 − 39]

	Non-Sensitive			Sensitive
	Zip Code	Age	Nationality	Condition
1	13053	28	Russian	Heart Disease
2	13068	29	American	Heart Disease
3	13068	21	Japanese	Viral Infection
4	13053	23	American	Viral Infection
5	14853	50	Indian	Cancer
6	14853	55	Russian	Heart Disease
7	14850	47	American	Viral Infection
8	14850	49	American	Viral Infection
9	13053	31	American	Cancer
10	13053	37	Indian	Cancer
11	13068	36	Japanese	Cancer
12	13068	35	American	Cancer

Original Table

	Non-Sensitive			Sensitive
	Zip Code	Age	Nationality	Condition
1	130**	< 30	*	Heart Disease
2	130**	< 30	*	Heart Disease
3	130**	< 30	*	Viral Infection
4	130**	< 30	*	Viral Infection
5	1485*	≥ 40	*	Cancer
6	1485*	≥ 40	*	Heart Disease
7	1485*	≥ 40	*	Viral Infection
8	1485*	≥ 40	*	Viral Infection
9	130**	3*	*	Cancer
10	130**	3*	*	Cancer
11	130**	3*	*	Cancer
12	130**	3*	*	Cancer

anonymous Table

a) Advantages.

i) By replacing actual value with more general value it become very difficult to find or guess actual data.

ii) K-anonymous techniques is very fact and efficient as compared to previous techniques.

iii) By replacing actual value with _*' unauthorized user get confused and it creates more possible combination related to original dataset t.

b) Disadvantage

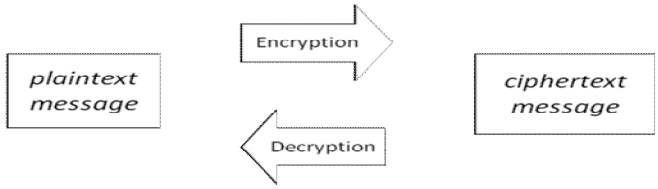
i) The main problems with generalization are it fails on high-dimensional data due to the curse of dimensionality it causes too much information loss due to the uniform distribution assumption

ii) The database with the tuple data does not be maintained confidentially.

D. Cryptography

The cryptographic approach for the privacy preserving data mining is assume that the data is stored at several private parties and they accept the describe the result of specific data mining operation. The parties use a cryptographic protocol for encrypting and decrypting the messages. That is they use encrypted messages to do some operation efficient. They blindly run their algorithm These mining process could be occurred in between two untrusted parties, or even between competitors . We describe here results of a body of cryptographic research that shows how separate parties can jointly compute any function of their inputs, without revealing any other information. As we argued above, these results achieve maximal privacy that hides all information except for the designated output of the function. This body of research attempts to model the world in a way which is both realistic and general. While there are some aspects of the "real world" that are not modeled by this research, the privacy guarantees and the generality of the results are quite remarkable. Oblivious transfer is a basic protocol that is the main building block of secure computation. It might seem strange at first, but its role in secure computation should become clear later.

The common definition of privacy in the cryptographic community limits the information that is leaked by the distributed computation to be the information that can be learned from the designated output of the computation. run in order to compute the function does not leak any "unnecessary" information. In simple words,Cryptography is the practice and study of hiding information. In many cases, multiple parties may require to share private data. They want to share information without leakage at their end.



For example, different branches in an educational institute wish to share their sensitive sales data to co-ordinate themselves without leaking privacy. This requires secure and cryptographic protocols for sharing the information across the different

parties. Cryptography , in the presence of a intruder extends from the traditional tasks of encryption and authentication .In an ideal situation, in addition to the original parties there is also a third party called “trusted party”. All parties send their inputs to the trusted party, who then computes the function and sends the appropriate results to the other parties. The protocol that is run in order to compute the function does not leak any unnecessary information. Sometimes there are limited leaks of information that are not dangerous. This process requires high level of trust.

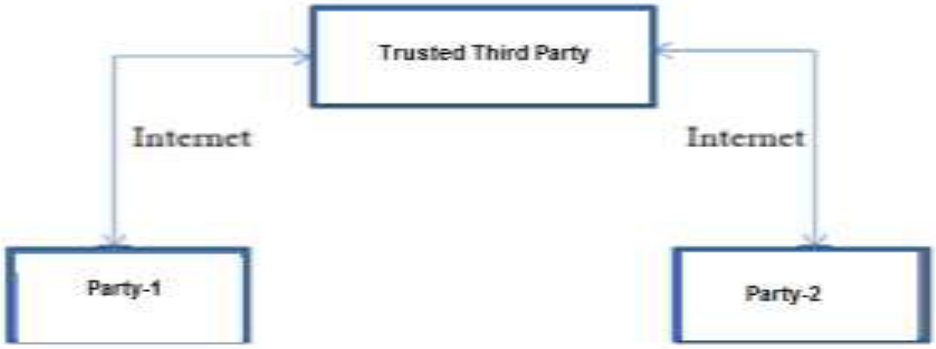


FIG 4: System using Semi Trusted Third Party

Some of the advantages and disadvantage of this method are as follows.

1) Advantages:

a. Cryptography offers a well-defined model for privacy for proving and quantifying it.
b. There exit a vast range of cryptographic algorithms.

2) Disadvantages

a. It is difficult to scale when more than a few parties are involved.
b. It does not guarantee that the disclosure of the final data mining result may not violate the privacy of individual records

By using following techniques we can overcome this problem.

E. Techniques of Cryptography

1) *Symmetric Cryptography*: Symmetric cryptography includes methods of encryption that are best suited for processing large streams of data. It is distinguished the use of a single key for encrypting and decrypting messages by the sender and receiver.This type of cryptography is categorized by the use of stream or block ciphers. Stream ciphers operate by encrypting single bits or bytes of information (or plaintext) at a time and implements a feedback mechanism to constantly change the key. Alternatively, block ciphers encrypts data into individual fixed group of bits (a common size is 128 bits) using the same key. An advantage of symmetric cryptography is that its methods are inexpensive for creating and processing encrypted data. The disadvantage of this example of cryptography is that is both the sender and receiver of the message have to agree on the key. If the key is discovered, the encrypted information becomes compromised.

2) *Asymmetric Cryptography*: Asymmetric cryptography (also called public key cryptography) encryption methods are best used for key exchange and user authentication. This type of cryptography is commonly used in digital signatures. It is distinguished by the use of a private and public key that are created with one-way functions using multiplication and exponentiation. One key is public and published in a public directory while the private key is only known by the receiver of the message.

3) *Elliptic Curve Cryptography.*: Elliptic curve cryptography is a standard method used by NIST, NSI and IEEE for

government and financial institution use. It is based on public key encryption and used in mobile and wireless environments. Public keys are created by utilizing the following algebraic equation - **y^2=x^3 + 3 + Ax + B** where the x and y points on a curve are used to calculate a public key. The private key is a random number. The appeal of elliptic curve cryptography is that it offers security with smaller key sizes which result in faster computations, lower power consumption, memory and bandwidth use. Very few attributes are required compared to traditional cryptographic approaches.Implementation is a complex task.

III.CONCLUSION

The ever increasing ability to identify and collect large amounts of data, analyzing the data using data mining process and decision on the results gives prospective benefits to organizations .But, such repositories also contains private and sensitive information and releasing the personal information can cause significant damage to data owner. Hence there is increased need to discover and distribute the databases, without compromising the privacy of the individual's data.In this paper, We have discussed different privacy preservation techniques such as Method of Perturbation, Method of Cryptography and k-anonymity and their advantages and disadvantages.

IV.ACKNOWLEDGMENT

We take this opportunity to express our gratitude and indebtedness to our guide Prof. P.D.Kaware as well as H.O.D, Prof. A. B**.** Raut, Computer Science and Engineering Department, they are a constant source of guidance and inspiration in preparing this work. Their constant help and encouragement helped us to complete our paper. We are grateful to Principal Dr. A.B. Marathe**,** for his encouragement and support. We are also thankful to all the staff members of Computer Science and Engineering department, whose suggestions helped us to complete the paper work and those who have directly and indirectly helped for completion of the paper.

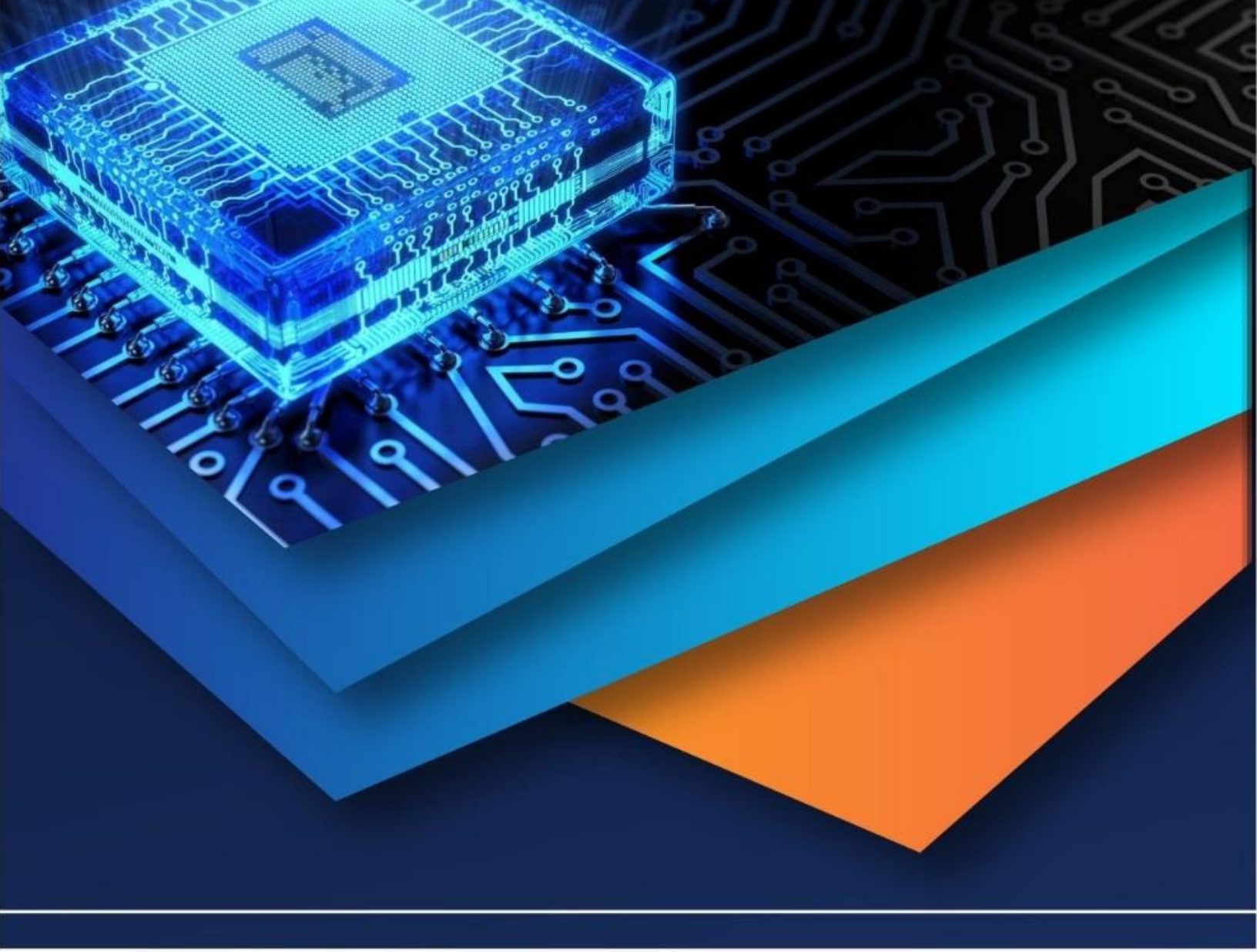
REFERENCES

[1]Benjamin C. M.,Fung, Ke Wang, Rui Chen, Philip S. Yu. 2010. Privacy-Preserving Data Publishing: A Survey of Recent Developments. ACM Computing Surveys, Vol. 42, No. 4, Article14.
[2] Samarati P (2001). "Protecting respondent's identities in microdata release". IEEE Transactions on Knowledge and Data Engineering, 13(6):1010–1027.
[3] Sweeney L (2002). "Achieving k-anonymity privacy protection using generalization and suppression". International Journal on Uncertainty, Fuzziness and Knowledge-based Systems, 10(5):571–588.
[4] R. Agrawal and R. Srikant, "Privacy-preserving data mining", In ACM SIGMOD, pages 439–450, May 2000
[5] A. Machanavajjhala, J. Gehrke, D. Kifer, and M. Venkitasubra- maniam. ℓ-diversity: Privacy beyond k-anonymity. Available at http://www.cs.cornell.edu/
[6]Lindell Y. , Pinkas B., "Privacy Preserving Datamining*", International Journal of Cryptology, Citesheer, 2002.
[7]P. Samarati. Protecting respondents' identities in microdata release. In *IEEE Transactions on Knowledge and Data Engineering*, 2001.
[8] L. Sweeney. k-anonymity: a model for protecting privacy. *International Journal on Uncertainty, Fuzziness and Knowledge-based Systems*, 10(5):557–570, 2002.
[9].Ashish C. Patel , Udai Pratap Rao, Dhiren R. Patel, "*Privacy Preserving Association Rules in Unsecured Distributed Environment Using Cryptography*" IEEE Department of Computer Engineering, Sardr Vallabhbhai National Institute of Technology, Surat, Gujarat, India-395007
[10] Anand Sharma and vibha ojha ""*Privacy preserving Data Mining by" cryptography"* in Springer-LNCS-CICS- Vol:89, *"Recent Trends in Network Security and Applications" .pp.*576- 581.2010.
[11]. Anand Sharma,Vibha Ojha "*Implementation of Cryptography for Privacy Preserving Data Mining"* 1 CSE Deptt., MITS, Lakshmangarh, Sikar, Rajasthan ,CSE Deptt., IITM, Gwalior, Madhya Pradesh

[12] Aniket Patel1, HirvaDivecha2 ,Samir Patel*" A Study of Data Perturbation Techniques For Privacy Preserving Data Mining"*Assistant Professor Assistant Professor,Department of Computer Engineering Department of Information Technology U V Patel College of Engineering Sigma Institute Of Engineering, Kherva-Mehsana, India Baroda,Gujarat

[13] Rick L. Wilson and Peter A. Rosen, Oklahoma State University,USA" "Protecting Data through Perturbation' Techniques: The Impact on Knowledge Discovery in Databases"

[14]"*A Survey of Perturbation Technique For Privacy-Preserving of Data"* Lokesh Patel1, Prof. Ravindra Gupta2
1M. Tech, 2Ass. Professor, SSSIST, Sehore Volume 3, Issue 6, June 2013)
[15]"*Protecting Data through Perturbation Techniques: The Impact on Knowledge Discovery in Databases"* Rick L. Wilson and Peter A. Rosen, Oklahoma State University, USA, Apr-June 2003



10.22214/IJRASET

45.98

IMPACT FACTOR: 7.129

IMPACT FACTOR: 7.429

INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY