iJRASET
International Journal For Research in Applied Science and Engineering Technology

INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY

Volume: 5 **Issue:** XI **Month of publication:** November 2017

DOI:

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:6.887
Volume 5 Issue XI November 2017- Available at www.ijraset.com

Data Renovation Algorithm for Protecting Sensitive Categorical Data

G. S. Karthick [1], M.Sridhar[2]
[1, 2]*Department of Computer Science, Bharathiar University*

Abstract: ***In the fast growing technological world, improvements and automation of various kinds of fields leads to increasing in growth of data. Such collection of data can be used frequently by economists, statisticians, scientists, forecasters, communication engineers, business and health care predictions in order to take any decision based on the historical factors. At one side data mining becoming the forefront of many fields, on the other side privacy risk factors also tremendously huge. Data given to the users like researchers or to any other third parties may contain sensitive data about an individual that must be protected from breach. In recent periods, many techniques were proposed and developed by the researchers to preserve the privacy in data. This proposed Data Renovation Algorithm (DRA) particularly paying attention on protecting sensitive categorical attributes in a dataset. It alters the original sensitive categorical attribute values and produce modified dataset. Both the original and modified datasets are applied to data mining techniques individually and the results produced for both datasets are equal.***
Keywords: Data mining, Privacy, Sensitive data, Masking techniques, Perturbation, J48 Classification.

I. INTRODUCTION

Data mining is a finding of solutions for the problems by analyzing data that readily available in dataset. Data mining can be seen as a intimidation to privacy because of the prevalent propagation of electronic data managed by corporations [1]. There may be a chance of using the sensitive or confidential data in illegal manner and this introduced the privacy preserving in data mining to safeguard the sensitive data from illegal usage.The sensitive data item can be any one of numerical, categorical or both. Many algorithms and techniques were developed in order to protect the sensitive data by modifying or renovating it. These modifications on sensitive data can be done by Data Masking technique. Data Masking is the process of replacing the existing sensitive information in the datasets with information that looks true, and then becomes meaningless to others who wish to misuse it [2].Many of the recent researches focused on data perturbation techniques mainly of how to protect the sensitive numerical attributes from unauthorized access. The sensitive numerical data protected by adding random noise to it and also encrypting techniques has been applied to protect sensitive categorical data. This proposed algorithm deals with hiding the sensitive categorical data by modifying original value with duplicate value by random consonants and vowels replacement.

II. REVIEW OF LITERATURE

Privacy preserving data mining technique is a broad research area in data mining and statistical databases in which mining algorithms are analysed and altered to acquire data privacy [3].Privacy preserving data mining techniques is classified into five major categories: Anonymization, perturbation, randomization, condensation and cryptography [4].

Anonymization is a process of removing the identifier attributes and changing the values of quasi identifier with less values, that migrates data appear similar and thus it is too difficult to identify the value of sensitive attribute. In perturbation technique original values are modified with the artificial values and the result obtained from the modified data does not differ from result obtained from original data. The modification can be achieved by adding noise, swapping data, multiplicative perturbation and projection perturbation [5]. The authors of [6-7] proposed an additive perturbation method (AP) that adds noise to the original data for ensuring the privacy and also revealed a good approximation of the original data. Kim et. al stated that perturbation is a technique that multiplies each data element by a random number has a Gaussian distribution reduced with a less average and variance [8].

III. METHODOLOGY

Consider a patient's health data, it may include various sensitive data items that can be either numerical or categorical or both. This provides importance to privacy preserving data mining aspects on opening access to sensitive data to third parties. The aim of privacy preserving data mining (PPDM) algorithms is to ex-tract relevant knowledge from large amounts of data while protecting at the sam time sensitive information [9].

A. Dataset

Dataset is a collection of heterogeneous records 'N', each record is made-up of 'M' number of attributes. Table 1 shows the original dataset of the patients. This table includes sensitive numerical attributes and sensitive categorical attributes which should not be released for open access.

TABLE 1
PATIENT'S ORIGINAL DATASET

Patient Id	Name	Age	Gender	Marital Status	Disease
1001	John	23	M	Unmarried	SwineFlu
1002	Kamalraj	27	M	Married	Dengu
1003	Latha	34	F	Divorce	Malaria
1004	Prabhu	30	M	Married	Aids
1005	Manoj	20	M	Unmarried	SwineFlu
1006	Sheela	18	F	Unmarried	Dengu
1007	Santhiya	25	F	Married	Malaria
1008	Kishore	31	M	Divorce	Aids

B. Modified Dataset

In a given dataset, sensitive data items must be modified or hidden to provide security. Modified dataset is a dataset in which sensitive attributes are identified and its values are altered by using any masking technique. Table 2 shows the modified dataset of the patients.

TABLE 2
PATIENT'S MODIFIED DATASET

Patient Id	Age	Gender	Marital Status	Disease
##	20-30	M	Unmarried	Qouieyos
##	20-30	M	Married	Mumky
##	30-40	F	Divorce	Bumptty
##	30-40	M	Married	Dups
##	20-30	M	Unmarried	Qouieyos
##	10-20	F	Unmarried	Mumky
##	20-30	F	Married	Bumptty
#s#	30-40	M	Divorce	Dups

C. Masking Techniques

Preservation of sensitive data from disclosure is an important issue in data mining process. Masking technique can be defined as a process of preserving the sensitive data from breach. Masking techniques basically applied on two different data types are as follows.

1) *Continuous data type:* These are the numerical data type on which basic arithmetic operations performed to produce a meaningful result.
2) *Categorical data type:* These are the non-numerical data type on which basic arithmetic operations cannot be applied. Further it can be classified into two categories of categorical data type.

a) *Nominal data type:* It can be two or more categories, but does not have any specificorder or quantitative values. (e.g. gender, city)

b) Ordinal data type: It also can be of two or more categories, but have an order. (e.g. measuring of happiness, discomfort, educational status)

D. Types of Masking Techniques

Masking techniques are alienated into two types shown in Fig 1

1) Perturbative: Perturbation is a mechanism of changing the sensitive attribute values and the datasets are unfair before establishing. The original dataset may alter and the new unique combinations of data items will be included in the perturbed dataset. In perturbation method statistics computed on the perturbed dataset do not differ from the statistics obtained on the original dataset [10]. It is well suited for masking the numerical attributes than the categorical attributes. Additive noise, resampling, rounding are the few Perturbative masking techniques.

2) Non-Perturbative: Non-Perturbation technique does not alter the values of sensitive attribute; instead values are covered up or detached. Top coding, local suppression, global recoding are the few non-Perturbative masking techniques.

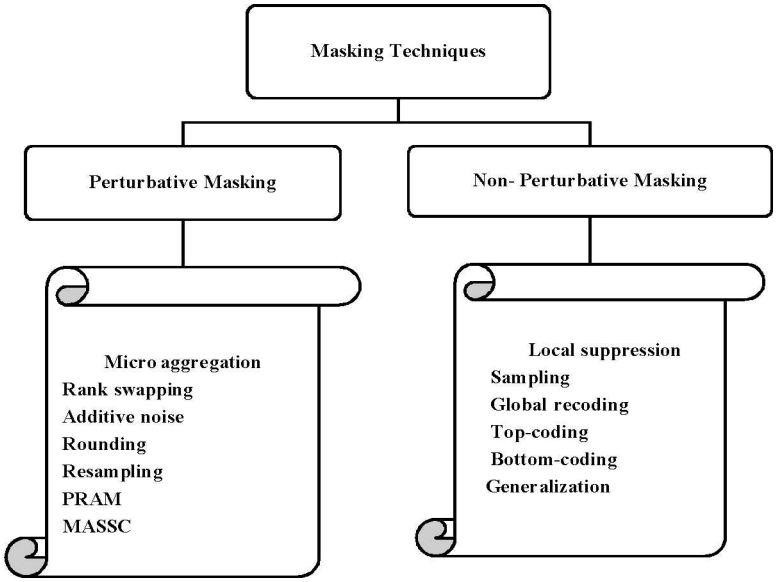


Fig1 Categorization of Masking Techniques

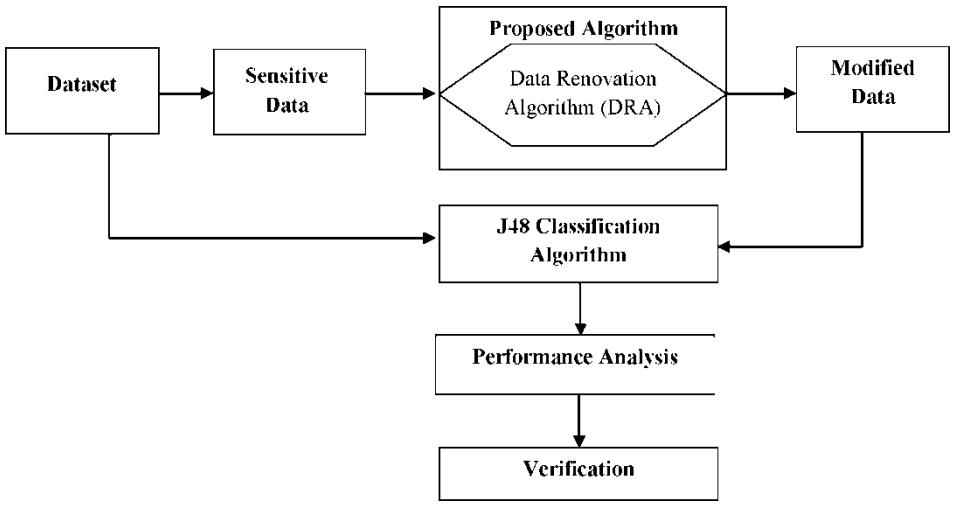
IV. OBJECTIVE OF THE PROBLEM

As privacy preserving data mining plays a major role in protecting sensitive attributes in a datasets. The main intention of this research work is to ensure privacy for sensitive categorical attributes by using a new perturbative masking technique. In this sensitive categorical attributes are selected and applied into proposed data renovation algorithm which produces the modified dataset. Then data mining techniques such as classification, clustering is applied on both the original dataset and modified dataset. The result produced by both the datasets are compared and analyzed to verify the accuracy and performance.

A. Proposed Solution

1) Identifying and selecting of sensitive categorical attributes Modification of categorical attributes using Data Renovati
2) Algorithm(DRA)
3) Application of Data Mining Technique- J48 Classification
4) Original dataset Modified dataset
5) Performance analysis Verifying the quality of proposed algorithm

ISSN: 2321-9653; IC Value: 45.98; SJ Impact Factor:6.887
Volume 5 Issue XI November 2017- Available at www.ijraset.com



B. Proposed Algorithm - Data Renovation Algorithm (DRA)

The basic principle of this new Perturbative masking algorithm is to modify the sensitive categorical data; the Data Renovation Algorithm (DRA) is given below on Table 4.1

TABLE 3
DATA RENOVATION ALGORITHM (DRA)

Begin
Consider a Dataset D includes R records, D= {r_{lt} r2, r3.., r_n}. Every record in R contains a set of attributes R={A1,A2, Am} where A_i €R and R_i €D
Predict the private or sensitive or confidential categorical attribute A_c
For each Confidential Categorical attribute A_c, convert the String A_c into a Character array $A_c[i]$
Obtain the length of Character array $A_c[i]$
For each element in $A_c[i]$ € A_c
Identify whether $A_c[i]$ is Consonant or Vowel
If $A_c[i]$ is Consonant then
Randomly replace $A_c[i]$ with C[i] € C// C is a set of Consonants
Else if $A_c[i]$ is Vowel then
Randomly replace $A_c[i]$ with V[i] € V // V is a set of Vowels
Next
Return Masked Char
End For each
Next
Return Masked Str
End

V. EXPERIMENTAL RESULTS

The dataset used here is the hospital dataset which contains patient's personal information with their diseases details. It is a synthetic dataset, which includes 500 instances. There are totally 11 attribute in which 5 attributes are numerical and 6 attributes are categorical attributes. In this research paper, we have performed classification using J48 decision tree algorithm on original hospital dataset and modified hospital dataset in weka tool. Weka tool provide inbuilt algorithms for J48. Then the result produced by both the datasets are compared and analyzed to predict whether the modified dataset generated by proposed algorithm affects the result or not.

A. J48 Classification Algorithm

J48 classifier is a simple C4.5 decision tree for classification and it creates a binary tree. With this technique, a tree is constructed to model the classification process. Once the tree is built, it is applied to each tuple in the dataset and results in classification for that tuple [11] [12].

TABLE 4
J48 CLASSIFICATION ALGORITHM

Algorithm [11] J48:
INPUT: D //Training data OUTPUT:T
//Decision tree
DTBUILD (*D)
{
T=p;
T= Create root node and label with splitting attribute; T= Add arc to root node for each split predicate and label;
For each arc do
D= Database created by applying splitting predicate to D;
If stopping point reached for this path, then T'= create leaf node and label with appropriate class;
Else
T'= DTBUILD(D);
T= add T' to arc;
}

B. Performance Analysis

The result produced by both the datasets are compared and analyzed to predict whether the modified dataset generated by proposed algorithm attains the accuracy same as original dataset. The performance is measured in terms of following factors.

1) *Classification Accuracy:* The following Table 5 and Fig 3 shows the accuracy rates produced by J48 algorithm for both original and modified datasets. Classification accuracy of modified is same as the accuracy provided by original dataset classification. Therefore, proposed algorithm does not affect the result.

TABLE 5
CLASSIFICATION ACCURACY RATE

Datasets	Correctly Classified Instances (%)	Incorrectly Classified Instances (%)	Kappa Statistics	Time Take (seconds)
Original	86.6534	13.3466	0.8404	0.11
Modified	86.6534	13.3466	0.8404	0.03

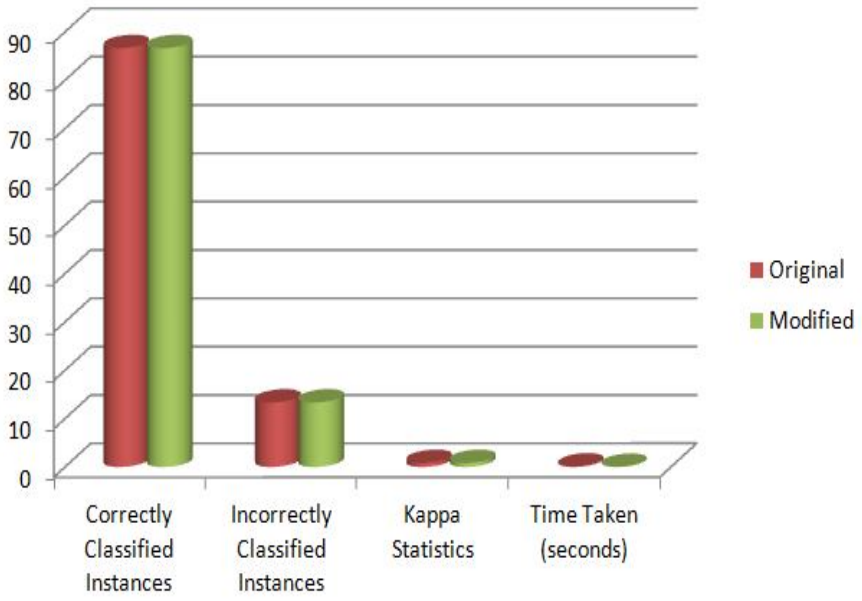


Fig 3 Classification Accuracy of datasets

2) *Error Rate:* Error rate is measured in terms of mean absolute error, root mean square error, relative absolute error and root relative squared error. The error rate produced for original and modified datasets are same. Thus, the proposed algorithm does not affect the result. Table 6 and Fig 4 shows the Classification error rates of both datasets.

TABLE 6
CLASSIFICATION ERROR RATE

Datasets	Mean Absolute Error	Root Mean Square Error	Relative Absolute Error (%)	Root Relative Squared Error (%)
Original	0.0387	0.1391	16.1399	40.1804
Modified	0.0387	0.1391	16.1399	40.1804

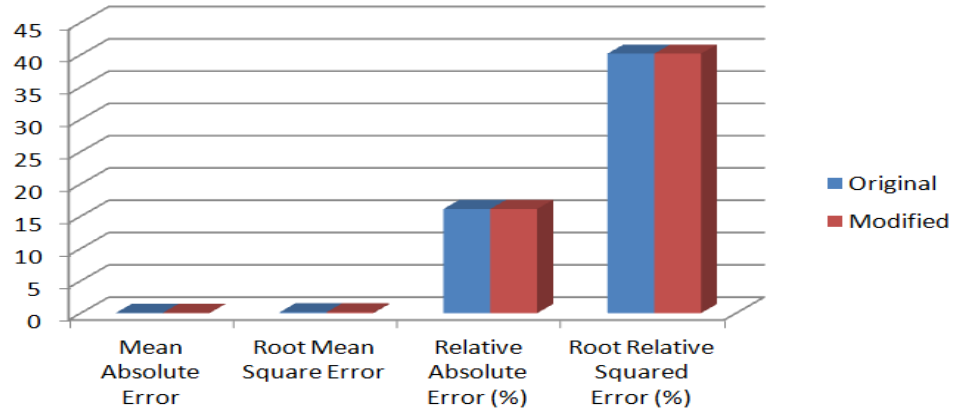


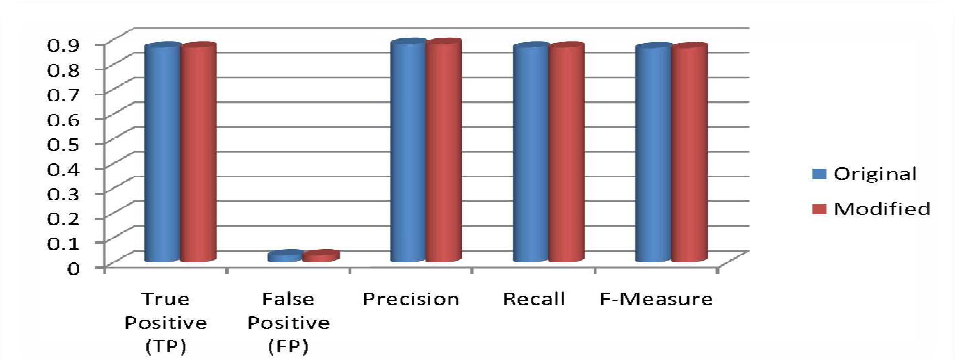
Fig 4 Classification Error Rate

3) *Detailed Accuracy:* Class accuracy can be calculated with weighted average values of TP Rate, FP Rate, Precision, Recall and F-Measure. The weighted average produced for both the datasets are same. Thus, the proposed algorithm does not affect the result. Table 7 and Fig 5 shows the Classification weighted average of both datasets.

TABLE 7
CLASSIFICATION WEIGHTED AVERAGE

Datasets	True Positive	False Positive	Precision	Recall	F-Measure
Original	0.867	0.027	0.881	0.867	0.865
Modified	0.867	0.027	0.881	0.867	0.865

Fig 5 Classification Weighted Average

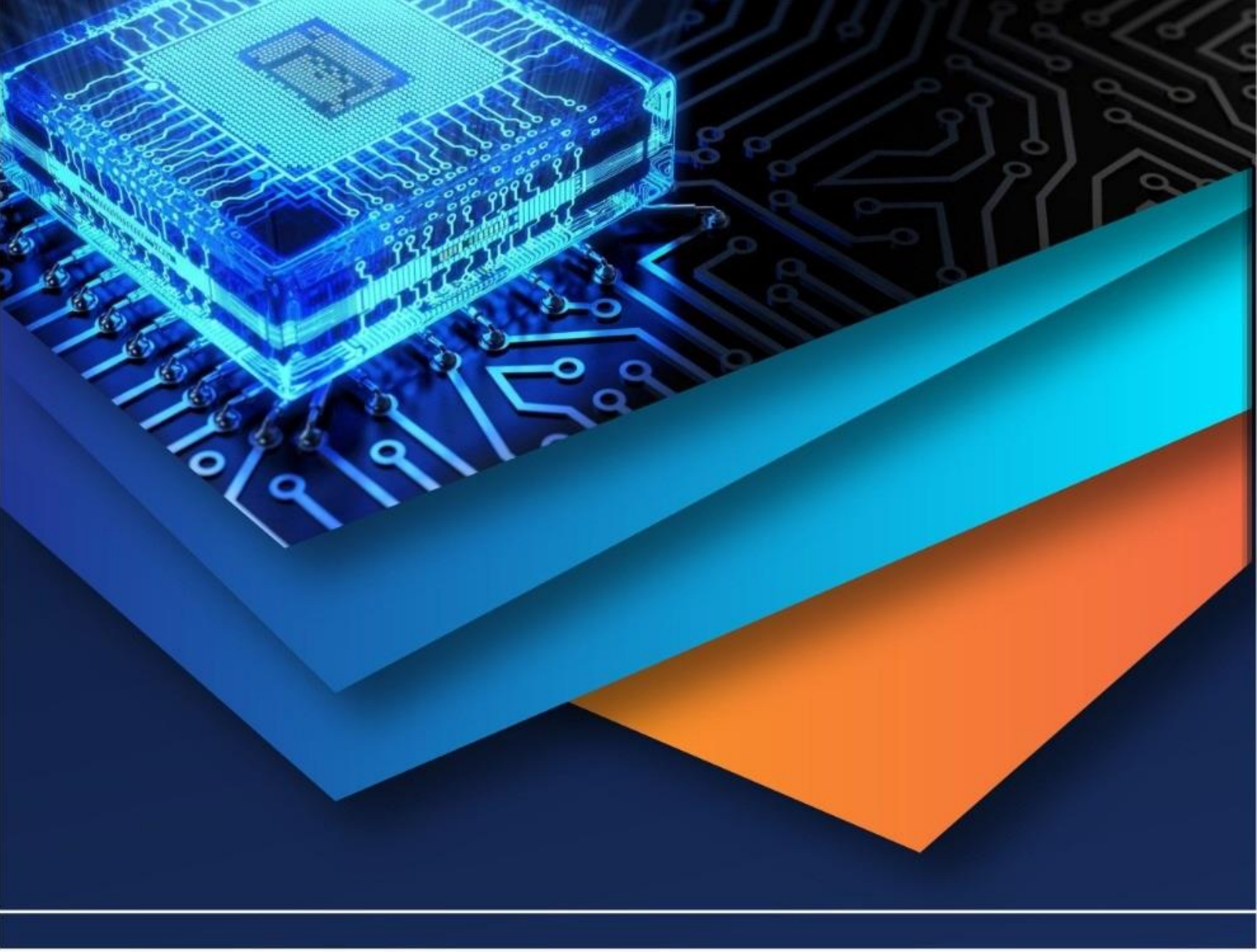


VI. CONSLUSION

In this research work, sensitive categorical data has been modified has been protected by using the proposed Data Renovation Algorithm (DRA). Then, the original and modified dataset has been applied to the data mining J48 classification algorithm in order to find out whether the modified dataset affects the actual results of data mining. From the factors of performance analysis, we found that the proposed algorithm does not affect the actual data mining techniques. Therefore, it is best for sensitive categorical data masking. In future, new masking algorithm is to be proposed to protect both sensitive numerical and sensitive data simultaneously.

REFERENCES

[1] C.C.Aggarwal; P.S.Yu.: A general Survey of Privacy-Preserving Data Mining Models and Algorithms, Sprringer,2008.
[2] Data Masking: What You Need to Know, A Net 2000 Ltd. White Paper
[3] Evfimievski, A., R. Srikant, R. Agrawal and J. Gehrke. “Privacy preserving mining of association rules”. Proceedings of the 8th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, July 23-25, ACM Press, Edmonton, AB., Canada, pp. 1-12,2002.
[4] M. B. Malik, M. A. Ghazi, and R. Ali, ``Privacy preserving data mining techniques: Current scenario and future prospects," in Proc. 3rd Int. Conf. Comput.Commun. Technol. (ICCCT), Nov. 2012, pp. 26-32
[5] Charu C. Aggarwal, Philip S. Yu,” A General Survey of Privacy-Preserving Data Mining Models and Algorithms”, in springer ISBN 978-0- 387-70991-8, e-ISBN 978-0-387-70992-5,DOI 10.1007/978-0-387-70992-5.
[6] Agrawal, R., &Srikant, R, “ Privacy preserving data mining.”, Paper presented at the 2000 ACM SIGMOD Conference on Management of Data, pp. 439-450,2000.
[7] Huang, Z., Du, W., & Chen, ” Deriving private information from randomized data”,In Proceedings of the 2005 ACM SIGMOD international conference on management of data, pp. 37-48).
[8] Kim, J. J., & Winkler, W. E,” Multiplicative noise for masking continuous data”,WashingtonD.C.: Statistical Research Division, U.S.Bureau of the Census,2003.
[9] A Survey of Quantification of Privacy Preserving Data Mining Algorithms Elisa Bertino, Dan Lin, and Wei Jiang.
[10] V.Ciriani, S.De Capitan di Vimercati, S. Forest, and P. Samurai “Micro data Protection” Springer US, Advances in Information Security (2007).
[11] Margaret H. Danham,S. Sridhar, " Data mining, Introductory and Advanced Topics", Person education, lsted., 2006.
[12] Aman Kumar Sharma, SuruchiSahni, "A Comparative Study of Classification Algorithms for Spam Email Data Analysis", IJCSE, Vol. 3, No. 5, 2011, pp. 18901895.



10.22214/IJRASET

45.98

IMPACT FACTOR: 7.129

IMPACT FACTOR: 7.429

INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY