International Journal of Advanced Technology in Engineering and Science
Vol. No.3, Special Issue No. 01, November 2015
www.ijates.com
ijates
ISSN 2348 - 7550

PRESERVATION PRIVACY TECHNIQUES AND CRYPTOGRAPHY RSA IMPLEMENTED IN RAPIDMINER TOOL USING NAIVE BAYESIAN CLASSIFICATION

N. Ambika Devi

Assistant Professor, Department of Computer Science and Information Technology

Nadar Saraswathi College of Arts and Science, Theni (India)

ABSTRACT

Data mining is the process of extracting or mining knowledge from large databases. Data mining software is one of a number of analytical tools for analyzing data. It allows users to analyze data from many different dimensions or angles, categorize it, and summarize the relationships identified. Privacy-preserving data mining plays an important role in the areas of data mining and security. In the area of privacy preserving data mining, the data mining algorithms are analyzed for their impact on data privacy. The goal of privacy preserving data mining is to develop algorithms to modify the original data set so that the privacy of confidential information remains preserved and as such, no confidential information could be revealed as a result of applying data mining tasks. The existing privacy preservation technique is the data set complementation approach and it fails if all data sets are leaked as the data set reconstruction algorithm is generic. The proposed method provides privacy preservation by converting the original sample data sets in to a group of unreal data sets and then applying cryptographic privacy protection to sensitive values. The cryptographic technique implemented is RSA. This method provides privacy preservation with improvement in accuracy. This work covers the application of new privacy preserving approach with the Naïve Bayesian classification algorithm and all of those techniques are implemented in Rapid Miner tool.

Keywords: Anonymization, Rapid Miner Tool, Privacy, Data Mining, Cryptography

I. INTRODUCTION

1.1 Data Mining

Data mining (the analysis step of the "Knowledge Discovery in Databases" process, or KDD), a field at the intersection of computer science and statistics, is the process that attempts to discover patterns in large data sets. It utilizes methods at the intersection of artificial intelligence, machine learning, statistics, and database systems The overall goal of the data mining process is to extract information from a data set and transform it into an understandable structure for further use aside from the raw analysis step, it involves database and data management aspects, data preprocessing, model and inference considerations, interestingness metrics, complexity considerations, post-processing of discovered structures, visualization, and online updating. Generally, data mining (sometimes called data or knowledge discovery) is the process of analyzing data from

different perspectives and summarizing it into useful information - information that can be used to increase revenue, cuts costs, or both.

1.2 Data

Data are any facts, numbers, or text that can be processed by a computer. Today, organizations are accumulating vast and growing amounts of data in different formats and different databases. This includes:

- operational or transactional data such as, sales, cost, inventory, payroll, and accounting
- nonoperational data, such as industry sales, forecast data, and macro-economic data
- meta data - data about the data itself, such as logical database design or data dictionary definitions

1.3 Information

The patterns, associations, or relationships among all this data can provide information. For example, analysis of retail point of sale transaction data can yield information on which products are selling and when.

1.4 Knowledge

Information can be converted into knowledge about historical patterns and future trends. For example, summary information on retail supermarket sales can be analyzed in light of promotional efforts to provide knowledge of consumer buying behavior. Thus, a manufacturer or retailer could determine which items are most susceptible to promotional efforts.

1.5 Data Mining Techniques

There are several major data mining techniques have been developed and used in data mining projects recently including association, classification, clustering, prediction and sequential patterns. We will briefly examine classification data mining technique with example to have a good overview of them.

1.6 Classification

Classification is a classic data mining technique based on machine learning. Basically classification is used to classify each item in a set of data into one of predefined set of classes or groups. Classification method makes use of mathematical techniques such as SEMMA methodology, linear programming, neural network and statistics. In classification, we make the software that can learn how to classify the data items into groups. For example, we can apply classification in application that “given all past records of employees who left the company, predict which current employees are probably to leave in the future.” In this case, we divide the employee’s records into two groups that are “leave” and “stay”. And then we can ask our data mining software to classify the employees into each group.

1.7 Naive Bayesian Classification

Naive Bayesian Classification is based on the Bayesian theorem. It is particularly suited when the dimensionality of the inputs is high. Parameter estimation for naive Bayes models uses the method of maximum Likelihood. In spite over-simplified assumptions, it often performs better in many complex real world situations. Its advantage is requires a small amount of training data to estimate the parameters.

Theory:

Derivation:

D : Set of tuples

- Each Tuple is an 'n' dimensional attribute vector
- X : (x1,x2,x3,…. xn)

Let there be 'm' Classes: C1, C2, C3…Cm

Naïve Bayes classifier predicts X belongs to Class Ciiff

- P (Ci/X) > P(Cj/X) for 1<= j <= m , j <>i

Maximum Posteriori Hypothesis

- P(Ci/X) = P(X/Ci) P(Ci) / P(X)
- Maximize P(X/Ci) P(Ci) as P(X) is constant

With many attributes, it is computationally expensive to evaluate P(X/Ci).

Naïve Assumption of "class conditional independence"

n

$P(X/Ci) = \pi P(xk/Ci)$

k-1

P(X/Ci) = P(x1/Ci) * P(x2/Ci) *…* P(xn/ Ci)

II. RELATED WORK

The goal of privacy preserving data is to develop methods without increasing the risk of misuse of the data used to generate those methods. The topic of privacy preserving data has been extensively studied by the data community in recent years. A number of effective methods for privacy preserving data mining have been proposed. Most methods use some form of transformation on the original data in order to perform the privacy preservation. The transformed dataset is made available for mining and must meet privacy requirements without losing the benefit of mining. Privacy control deals with various kinds of methods that can provide privacy against sensitive attributes.

III. METHODOLOGY

3.1 Privacy Preservation in Data Mining (PPDM)

Data mining is the process of analyzing data from various perspectives and acquiring the useful information. Data mining is a method of extracting useful information through mapping the proper relation among huge array of data objects. Privacy control is a novel concept of data mining due to safety reasons. Nowadays, the data through the internet and other social media are plenty. Hence the privacy preservation deserves the serious attention. Privacy Preservation in Data Mining (PPDM) is a novel technique in data mining, where mining algorithms are incorporated. The significance of PPDM varies from different perspective because while publishing the data, the individual's identity and other details should not get disclosed. As well the information loss due to privacy preservation highly affects the data utility. PPDM, balance the trade-off between utility and privacy preservation by using various anonymization techniques.

3.2 Taxonomy

In general, the personal identifications will be removed before publishing the data for mining purpose. Privacy preservation is a serious issue and it can be gained through different techniques. Figure 1 describes the taxonomy of Privacy preservation in data mining. The two main approaches of Privacy preservations are Anonymization and Cryptography.

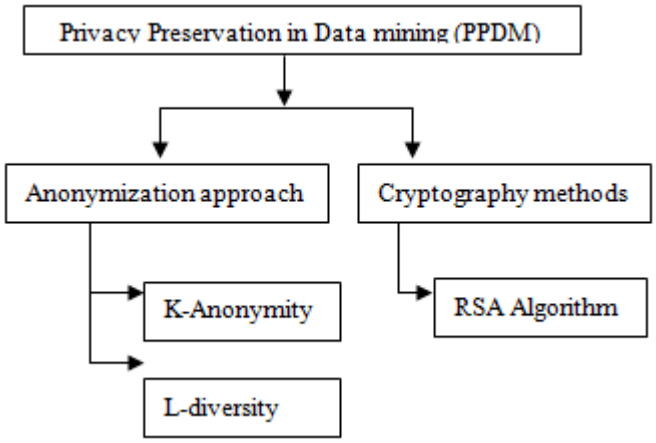


Figure 1Taxonomy

3.3 Privacy

Privacy can be defined as the technique to builds the concept of 'personal space', which means free from interference by other people or a committee of peoples. Data privacy means at the time of sharing or publishing sensitive information (personal information) it should be granted that only relevant person has an access on it and no one else is able to misuse it. Personally identifiable information (PII) is a subset of confidential information that can put your identity at risk if it is lost or stolen.

3.4 Privacy Important

When publishing micro data, there are three types of privacy disclosure threats. The first type is membership disclosure. When the data set to be published is selected from a large population and the selection criteria are sensitive (e.g., only work class are selected), one needs to prevent adversaries from learning whether one's record is included in the published data set. The second type is identity disclosure, which occurs when an individual is linked to a particular record in the released table. In some of the situations, one wants to protect against identity disclosure when the adversary is uncertain of membership. In this case, protection against membership disclosure helps protect against identity disclosure. The third type is attribute disclosure, which occurs when new information about some individuals is revealed, i.e., the released data make it possible to infer the attributes of an individual more accurately than it would be possible before the release.

3.5 Anonymization Techniques

Anonymity is derived from the Greek word, anonymia (ἀνωνυμία) birthed to anonymity which means "without an identity (name)" or "namelessness". Micro data contains records each of which contains information about an individual entity, such as a person, a household, or an organization. Several micro data anonymization techniques have been proposed. The most popular techniques k-anonymity and L-diversity. Anonymization

reduces the risk of identity disclosure whereas the data remains still realistic. Micro data contains information about an individual, a household or an enterprise. Each such dataset will be having i) Personal identification like Name, Address or Social Security Number (SSN) which uniquely identifies an individual ii) Sensitive Attributes (SAs) like salary and disease iii) The values of Quasi Identifiers (QI) such as Gender, Age, Zip code will leads to identity disclosure when taken together. In these techniques, k-anonymity prevents the identification of individual records in the data and l-diversity prevents the association of an individual record with the sensitive value attribute.

3.5.1K-anonymity

Definition **1:** k-anonymity as the property that each record is indistinguishable with at least k-1 other records with respect to the quasi-identifier. In other words, k-anonymity requires that each QI group contains at least k records. The protection k-anonymity provides is simple and easy to understand. If a table satisfies K-anonymity for some value k, then anyone who knows only the quasi-identifier values of one individual cannot identify the record corresponding to that individual with confidence greater than 1/k. K-anonymity model for multiple sensitive attributes mentioned that there are three kinds of information disclosure.

1) Identity Disclosure: When an individual is linked to a particular record in the published data called as identity disclosure.

2) Attribute Disclosure: When sensitive information regarding individual is disclosed called as attribute disclosure.

3) Membership Disclosure: When information regarding individual's information belongs from data set is present or not is disclosed is said to be membership disclosure.

3.5.2 ℓ-diversity

Definition 2:A QI group is said to have ℓ-diversity if there are at least ℓ "well-represented" values for the sensitive attribute. A table is said to have ℓ-diversity if every QI group of the table has ℓ-diversity.

3.5.3Generalization

A common approach is generalization, which replaces quasi-identifier values with values that are less specific but semantically consistent, a QI group is also called an "anonymity group" or an "equivalence class". The original Table 1 shows an example micro data table and its anonymized versions using various anonymization techniques. The three QI attributes are {Age, Gender, Education no} and the sensitive attribute SA is Work class. A generalized table that satisfies 4- anonymity is shown in Table 2.

3.5.4 Suppression

Suppression refers to removing a certain attribute value and replacing occurrences of the value with a special value "*", indicating that any value can be placed instead.

An Original Micro data Table and Its Anonymized Versions Using Various Anonymization Techniques

Age	Gender	Education no	Work class

17	Male	13	State-gov
25	Female	9	Self-emp-not inc
30	Female	7	State-gov
45	Male	13	Private
56	Female	17	Private
60	Male	12	state-gov
72	Male	12	Self-emp-not inc
90	Female	8	Private

Table 1 Micro data

Age	Gender	Education no	Work class
[17-90]	*	13	State-gov
[17-90]	*	9	Self-emp-not inc
[17-90]	*	7	State-gov
[17-90]	*	9	private

Table 2 Generalization & Suppression

3.5.5 Multi set Generalization

The multi set of exact values provides more information about the distribution of values in each attribute than the generalized interval. Therefore, using multi sets of exact values preserves more information than generalization. For example, Table 2 is a generalized table, and Table 3 is the result of using multi sets of exact values rather than generalized values. For the Age attribute of the first bucket, we use the multi set of exact values {22, 22, 33, 52} rather than the generalized interval [22-52]. Then the multi set of age ratio is {22:2, 33:1, 52:1}.

Age	Gender	Education no
22:2,33:1,52:1	M:1,F:3	13:1,9:2
22:2,33:1,52:1	M:1,F:3	13:1,9:2
22:2,33:1,52:1	M:1,F:3	13:1,9:2
22:2,33:1,52:1	M:1,F:3	13:1,9:2

Table 3 Multi set Generalization

3.5.6 Bucketization

The Bucketization method first partitions tuples in the table into buckets and then separates the quasi-identifiers with the sensitive attribute by randomly permuting the sensitive attribute values in each bucket. A bucketized table that satisfies 2-diversity is shown in Table 4.

Age	Gender	Zip code	Work class

22	M	47906	State-gov
22	F	47906	private
33	F	47905	private
52	F	47905	Self-emp-not inc
54	M	47302	private
60	M	47302	State-gov
60	M	47304	State-gov
64	F	47304	private

Table 4 Bucketization Table

3.5.7 Slicing

Slicing can be effectively used for preventing attribute disclosure, based on the privacy requirement of 'l-diversity. Slicing first partitions attributes into columns. Each column contains a subset of attributes. This vertically partitions the table. The sliced table shown in Table 5 contains four columns, where each column contains exactly one attribute. For example, the sliced table in Table 6 contains two columns: the first column contains {Age; Gender} and the second column contains {Work class, Zip code}.Slicing also partition tuples into buckets. Each bucket contains a subset of tuples. This horizontally partitions the table. For example, both sliced tables in Table 6 contain two buckets, each containing four tuples. Within each bucket, values in each column are randomly permutated to break the linking between different columns.

For example, in the first bucket of the sliced table shown in Table 6, the values {(22; M); (22; F); (33; F); (52; F)} are randomly permutated and the values {(47906; State-gov),(47906; private),(47905; private), (47905; Self-emp-not inc)} are randomly permutated so that the linking between the two columns within one bucket is hidden. Given a micro data table T, a slicing of T is given by an attribute partition and a tuple partition.

Age	**Gender**	**Zip code**	**Work class**
22	F	47906	private
22	M	47905	private
33	F	47906	State-gov
52	F	47905	Self-emp-not inc
54	M	47302	State-gov
60	F	47304	Private
60	M	47302	State-gov
64	M	47304	Private

Table 5 one attribute per column slicing

(Age, Gender)	**(zip code, work class)**

(22,M) (22,F) (33,F) (52,F)	(47905,private) (47906 State-gov) (47905; Self-emp-not inc) (47906; private)
(54,M) (60,M) (60,M) (64,F)	(47304,private) (47302,private) (47302, State-gov) (47304, State-gov)

Table 6 sliced table.

3.6 Cryptography Methods

Cryptography is the technique which focuses mainly on securing the information from the third parties. Information security has various aspects like data confidentiality, authentication and data integrity. Cryptographic methods like symmetric-key cryptography, public-key cryptography, cryptanalysis and cryptosystems are widely used privacy preservation methods.

3.6.1 RSA algorithm

The RSA Algorithm uses two keys. d and e, which work in pairs, for decryption and encryption respectively. A plain text message P is encrypted to cipher text by,

C=Pe mod n

The plain text I recovered by:

P=cd mod n

Because of symmetry in modular arithmetic, encryption and decryption are mutual inverses and commutative. Therefore,

P=cd mod n = (Pe)d mod n= (Pd)e mod n.

Thus one can apply the encrypting transformation first and then the decrypting one, or the decrypting transformation first followed by the encrypting one. For example, the sensitive attribute work class is encrypted and decrypted in Table 7 & Table 8.

Age	Gender	Education no	Work class
17	Male	13	????????
25	Female	9	??????????
30	Female	7	???????????
45	Male	13	?????????
56	Female	17	?????????
60	Male	12	???????????
72	Male	12	??????????
90	Female	8	???????????

Table 7 Encryption Data

17	Male	13	State-gov
25	Female	9	Self-emp-not inc
30	Female	7	State-gov
45	Male	13	Private
56	Female	17	Private
60	Male	12	state-gov
72	Male	12	Self-emp-not inc
90	Female	8	Private

Table 8 Decryption Data

3.7 System architecture

The System architecture of the Anonymization and Cryptography data are,

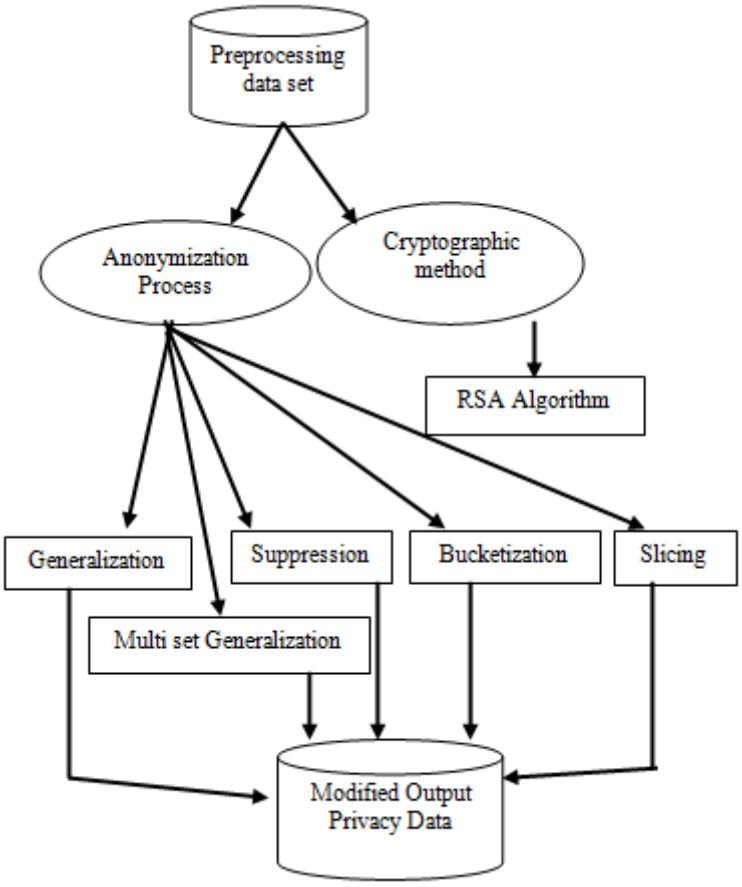


Figure2 System Architecture of the Anonymization

IV. EXPERIMENTAL DATA

In this paper, I used the Adult data set from the UC Irvine machine learning repository. In this data set we have to use 14 attributes. They are Name, Age, work class, Education, Education no, Marital status, Occupation, Relationship, Race, Gender, Hours per week, Native, Salary, Zip code. There are 600 valid tuples in total.

V. BENEFITS OF ANOYMIZATION

Generalized Data, in order to perform data analysis or data mining tasks on the generalized table, the data analyst has to make the uniform distribution assumption that every value in a generalized interval/set is equally possible, as no other distribution assumption can be justified. This significantly reduces the data utility of the generalized data.

We observe that this multi set-based generalization is equivalent to a trivial slicing scheme where each column contains exactly one attribute, because both approaches preserve the exact values in each attribute but break the association between them within one bucket.

While bucketization has better data utility than generalization, it has several limitations. First, bucketization does not prevent membership disclosure. Because bucketization publishes the QI values in their original forms, an adversary can find out whether an individual has a record in the published data or not. The effectiveness of slicing in membership disclosure protection. For this purpose, we count the number of fake tuples in the sliced data. We also compare the number of matching buckets for original tuples and that for fake tuples. Our experiment results show that bucketization does not prevent membership disclosure as almost every tuple is uniquely identifiable in the bucketized data.

Slicing is its ability to handle high-dimensional data. By partitioning attributes into columns, slicing reduces the dimensionality of the data. Each column of the table can be viewed as a sub-table with a lower dimensionality. Slicing is also different from the approach of publishing multiple independent sub-tables in that these sub-tables are linked by the buckets in slicing. It preserves better data utility than generalization. It preserves more attribute correlations with the Sensitive Attributes than bucketization. It can also handle high-dimensional data and data without a clear separation of Quasi Identifiers and Sensitive Attributes. Slicing can be effectively used for preventing attribute disclosure, based on the privacy requirement of ‘l-diversity.

VI. RAPID MINER 6.0

Data mining refers to extracting or “mining” knowledge from large amount of data. There are also many Data mining tools such as SAS, WEKA, Mine set and Rapid Miner. We will focus on Rapid Miner in this topic. In this paper, we do the classification in Rapid miner, and introduce how to uses Rapid miner to do these actions. At the same time we will use the tool to analyze the adult data sets.

Rapid Miner (Formerly YALE) is the worldwide leading open source data mining solution due to the combination of its leading edge technologies and its functional range. Application of Rapid miner covers a wide range of real world data mining tasks.

Use Rapid miner and explore many data. Simplify the construction of experiments and the evaluation of different approaches. It is used for business and industrial applications as well as for research, education, training, rapid prototyping, and application development and supports all steps of the data mining process including results visualization, validation and optimization.

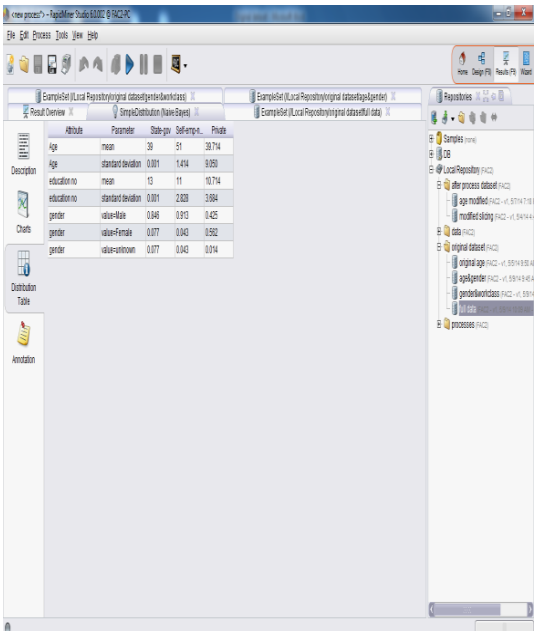
Rapid Miner has a very large set of operators, which makes it very suitable for comparing different machine learning/statistical methods. It is also very good for model building and validation. However, the learning curve

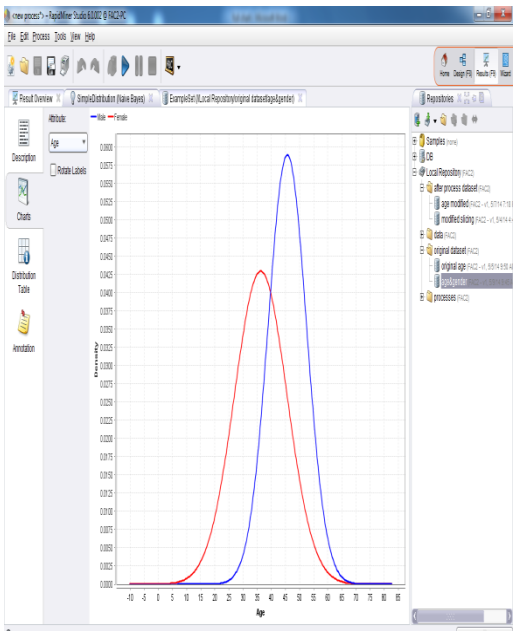
for the software is rather steep. Rapid Miner have the biggest and most active user communities. It is quickly implement (and integrate) new and emerging machine learning algorithms into their systems.

VII. RESULTS AND DISCUSSION

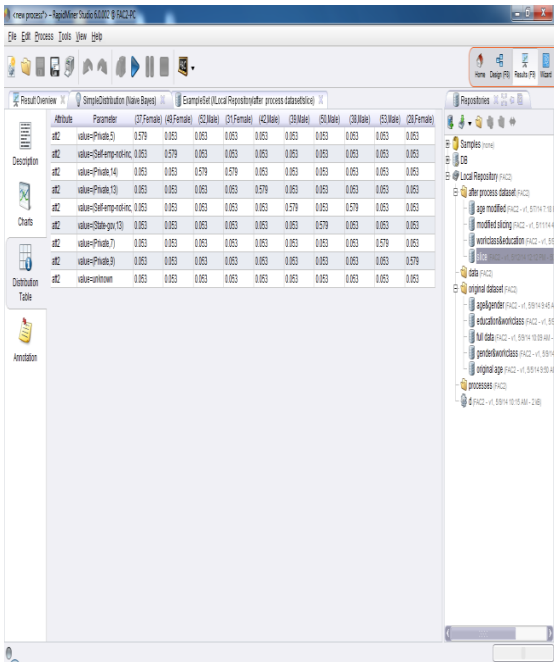
This section presents the first result is the original Adult data set for four attributes. Then the second result is applying six techniques. And then they are applied in Rapid Miner tool to show the accuracy is given below,

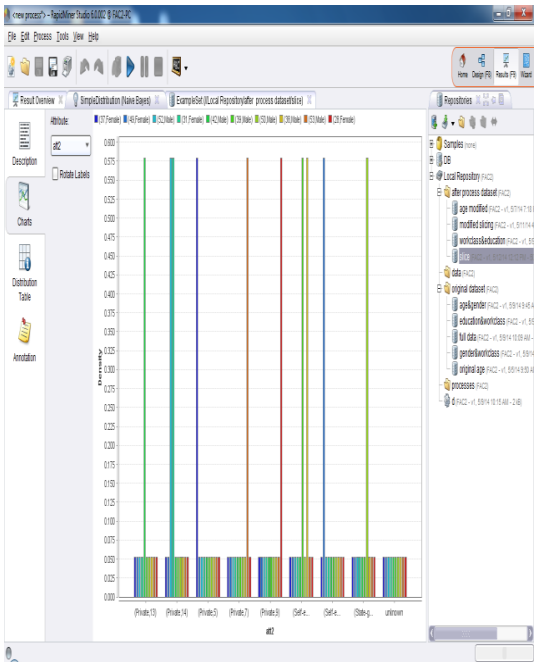
Original Adult Data set



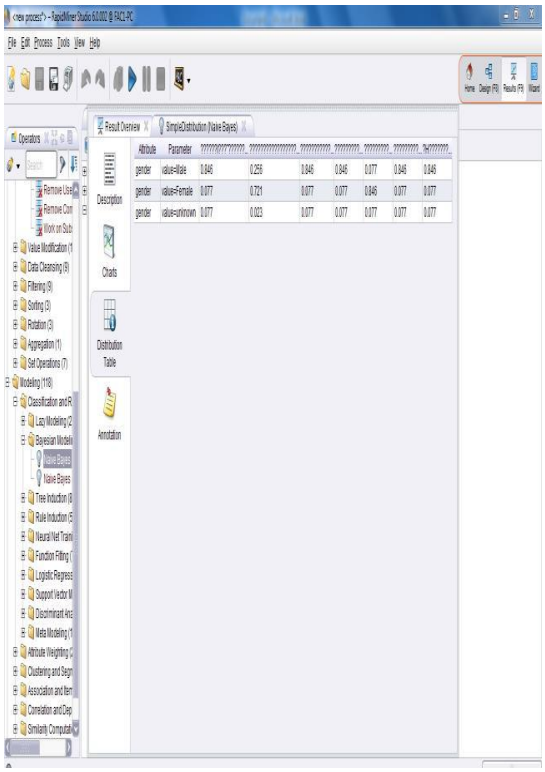


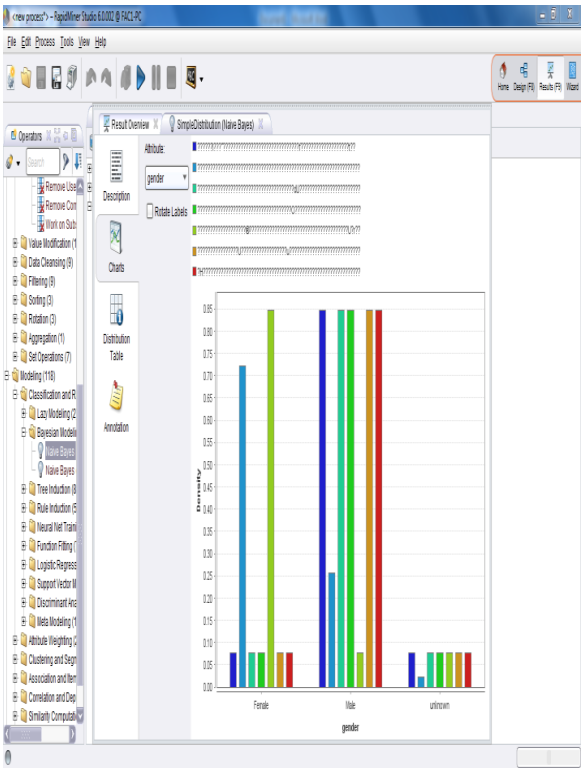
Modify Privacy Adult Data set (Anonymization)





Modify Privacy Adult Data set (Cryptography)





VIII. CONCLUSION

In this paper, finally conclude that every anonymization techniques have their own significance. Generalization causes too much of information loss, suppression & multi set generalization are better utility than generalization. Bucketization fails in privacy preservation due to identity disclosure. Slicing performs better than generalization, bucketization and many other anonymization methods. Slicing provides high dimensional data by partitioning highly correlated attributes into columns and further breaks the association of uncorrelated attributes. Thus slicing in combination with correlation analysis has the high data utility and ensures privacy in PPDM.

The proposed method provides privacy preservation by converting the original sample data sets in to a group of unreal data sets and then applying anonymization techniques and cryptographic privacy protection to sensitive values. The cryptographic technique implemented is RSA. All the above techniques are implemented in Rapid Miner tool. These methods provide privacy preservation with improvement in accuracy. This work covers the application of new privacy preserving approach with the Naïve Bayesian classification algorithm to report the statistical view.

IX. FUTURE ENHANCEMENT

In Future works, to analyze the P-sensitive, M-Invariance, T-Closeness and other anonymization techniques and those to apply different data mining open source tools using various datasets in neural networks, clustering models and association rules.

ijates
ISSN 2348 - 7550

REFERENCES

[1]. L. Sweeny," K-Anonymity: A Model for protecting privacy", International journal of Uncertainty, fuzziness, Knowledge based systems, Vol.no.10, Issue.07, 2002.

[2]. Y.Lindell and B.Pinkas," Privacy Preserving data mining", Journal of Cryptology, Vol.no.15, no.3,pp 177-206, 2002.

[3]. J.Zhang,D.K.Kang,A.Silvescu and V.Honavar," Learning accurate and concise naïve bayes classifiers from attribute value taxonomies and data", Knowl.Inf.Syst, Vol.no.09, no.2, pp 157-179, 2002.

[4]. D.Newman,S.Hettich,C.Blake and C.Merz," UCI repository of Machine learning databases", 1998.

[5]. I.H.witten and E.Frank," Data mining:Practical machine learning tools and techniques", 2 nd ed.Morgan Kaufmann, 2005.

[6]. Machanavajjhala, A.Gehrke, J.Kifer, D.Venkatasubramaniam,"L-diversity:Privacy beyond K-Anoymity",in:ICDE, 2006.

[7]. R.Leela,P.Revathi,"Enhancing the Utility of Generalization for privacy preserving Re publication of Dynamic datasets",International Journal of Computer applications, Vol.no.13, no.6, Jan 2002.

[8]. G.Nayak, "A survey on Privacy Preserving data mining: approaches and Techniques", International Journal of Engineering Science and Technology, Vol.no.03, no.3, pp 2117-2133, March 2002.

[9]. R.Agrawal,R.Srikant,Privacy Preserving Data mining",ACM SIGMOD Record,Newyork, Vol.no.29, no.2, pp 439-450, 2000.

[10]. P.Wang,"A survey on Privacy Preserving data mining", International Journal of Digital Content Technology and its Applications, Vol.no.04,no.9,Dec 2002.