ANALYSIS OF PRIVACY PROTECTION AWARENESS WHEN MINING AND SHARING FEMALE STUDENTS DATA THROUGH ONLINE

[1]R.S.Maragathavalli, [2]K. Mythili, [3]M. Gayathri

[1]Research Scholar,[2]Assistant Professor,[3]Assistant Professor

[1,2] CSA Department, [3]CSE Department,

Sri Chandrasekharendra Saraswathi Viswa Mahavidhyalaya University, Kancheepuram, Tamil Nadu, India

***ABSTRACT*— In today's digital environment, the education system in tamilnadu also updated to the digital innovative. Schools are increasingly adopting digital teaching. Technology and education are a great combination with a right reason and vision. Technology improves education to a great extent and it has now become a need for revolutionizing education for the better. With technology, educators, students and parents have a variety of learning tools at their fingertips. The online data processes, such as Exam details, students welfare scheme details, teachers personal details, promotion, transfer details, result statistics, students census information's like gender, community, environment details etc.. are done through the mails, google sites, google docs, google forms, forums, blogs, facebook etc.. The search engines and social networks, collect data about their users and their users' online activities. Mining and sharing of this data has been a key driver of innovation and improvement in the quality of these services, but has also raised major user privacy concerns. This paper is aimed to help educational institutes to aware when mine and share student data through online.**

I. INTRODUCTION

Recent advances in technology and communications have dramatically changed the landscape of education in India. Today's classrooms increasingly utilized adapted learning materials, virtual classrooms, smart classes for interacting with other students and teachers, and a wealth of other interactive technologies that help foster and enhance the learning process. Online forums help teachers share lesson plans; social media help students team up across classrooms; also emails, sites, blogs and private contract websites are used to manage the school administration works through online.

Adoption of digital technologies like online activities has established their potential to transform the educational process, but they have also called attention to possible challenges. In particular, the information sharing, web hosting, and telecommunication innovations that have enabled these new education technologies raise questions about how best to protect student privacy during use. This document will address a number of these questions, and present some requirements and best practices to consider, when evaluating the use of online educational services.

This thesis discusses privacy and security considerations relating to computer software and web based tools provided by a third party to the colleges and schools that students access the Internet and education records.

Privacy Challenges

A district administration like DDCE or CEO or DEO office may decide to use an online system to allow students to log in and access class materials. In order to create student accounts, the Office or School will likely need to give the provider the students' names and contact information from the students' education records, which are protected by Educational System. Online educational services increasingly collect a large amount data as part of their operations, often referred to as "metadata." Metadata refer to information that provides meaning and context to other data being collected; the provider collects metadata about student activity, including time spent online, desktop access, success rates, and keystroke information. If the provider de-identifies these metadata by removing all direct and indirect identifying information about the individual students (including school and most geographic information), the provider can then use this information to develop new personalized learning products and services (unless the district's agreement with the provider precludes this use).

Aim of the Study

The primary goal of this study is to find the privacy protection awareness of the students while sharing and mining data through online environment.

Statement Of the Problem

Students Privacy particularly the female student's privacy awareness is very importance when keeping the personal records through online. We aware with the online environment and protect the data's from the intruder and virus programs.

Scope of the Study

Women's data security is more important while surfing, chatting, sharing through internet or online. Communication through internet is more risky now a days, it have lot of loop holes to stolen one's personal data's and gives lot of damage to their

functional activities and livings. Particularly female student's data's must handle with safe and secure level at education institutions. But there is no enough knowledge to the students, employers and parents to protect the female student's data's. In this study, we discuss the level of awareness and protection skills of the students, when mining and sharing female student's data through online.

Objectives

[1] To test the awareness level about viruses and malware's depends upon the age group.
[2] To test awareness level about network tracking depends upon the age group
[3] To test the awareness level of protecting personal details depends upon the gender.
[4] To know the awareness level of handling privacy setting tools depends upon the gender.
[5] To know the knowledge about the private browsing depends upon the gender and course who studying.
[6] To know the awareness about spam mails.
[7] To know the awareness about women's privacy protection through online.

Hypotheses Of The Study

The following hypotheses were formulated for the present study.

1. There is significant Different between Age of the Student and Awareness about Virus and malware's.
2. There is significant Different between Age of the Student and Awareness about tracking the network activities.
3. There is significant Different between Age of the Student and Awareness about protecting personal details from unknown users.

Location Of The Study

The investigator has chosen colleges and schools in Vellore district as the area of study. Vellore district is the third most populous district of Tamil Nadu. It is located with neighbour states of Andhra Pradesh and Karnataka. Vellore district has the blend of rich heritage and culture of the ancient Dravidian civilization.

II. LITERATURE REVIEW

Rakesh Agrawal and Ramakrishnan Srikanth says that, the primary task in data mining is the development of models about aggregated data, can we develop accurate models without access to precise information in individual data records. We consider the concrete case of building a decision-tree classifier from training data in which the values of individual records have been perturbed. The resulting data records look very different from the original records and the original distribution. A Recent survey of web users classified 17% of respondents as privacy fundamentalists who will not provide data to a website even if privacy protection measures are in place. However the concern of 56% of respondents constituting the pragmatic majority were significantly reduced by the presence of privacy protection measures. The Remaining 27% were marginally concerned and generally willing to provide data to web sites, although they often expressed a mild general concern about privacy.

Amongst several existing algorithm, the Privacy Preserving Data Mining (PPDM) renders excellent results related to inner perception of privacy preservation and data mining. Truly, the privacy must protect all the three mining aspects including association rules, classification, and clustering (Sachan et al. 2013). The problems faced in data mining are widely deliberated in many communities such as the database, the statistical disclosure control and the cryptography community (Nayak and Devi 2011).

Vaidya et al. (2008) developed an approach for vertically partitioned mining data. This technique could modify and extend a variety of data mining applications as decision trees. More efficient solutions are needed to find tight upper bound on the complexity.

Kantarcıoglu and Vaidya (2003) emphasized the use of secure logarithm and summation, where the distributed naive Bayes classifier are securely determined. The experimental results strongly supported the concept of few useful protected protocols that facilitated the secure deployment of different types of distributed data mining algorithms.

The classification of privacy preserving methods and standard algorithms for each class is reviewed by Sathiyapriya and Sadasivam (2013), where the merits and limitations of different methods are exemplified. The optimal sanitization is found to be NP-Hard in the presence of privacy and accuracy trade-off.

III. METHODOLOGY

Method Of Data Collection

The researcher collected the needed data through the use of questionnaire and its administration in the selected faculties. The administration of the questionnaire was carried out by the researcher. A total of 200 copies of the questionnaire were distributed to elicit responses from the students and retrieved on the spot by the researcher.

Method of Data Analysis

Responses from the questionnaire were analyzed using the descriptive statistics of frequency counts and percentage, and inferential statistics of Chi-square(x2). Descriptive statistics of frequency counts and percentages were used in analyzing demographic variables and research questions while the inferential statistics of Chi-square(x2) was also used to test the stated hypotheses at 0.05 significant level .

Tools Used For Data Collection

In this research 25 questions were formed to find out the privacy protection awareness level of the students while browsing and sharing data through online.

First part is the General Information format that is used as background variables knowing about the student's age, course type.

1 to 12th questions are about to know their (students) facilities to use internet access and the basic knowledge about the privacy protection.

13 to 25th question is about to know their advanced skills about privacy protection while surfing, working with and sharing files through online. These questions are constructed using lickert scale which says agree and disagree in 4 point scale.

Samples of 200 students are taken for the present study. These 200 Students were taken from 9 Government and private Engineering, Arts Colleges and 3 schools in Vellore district, Tamilnadu.

Variables of the Study

Variables are defined as the conditions or characteristics or situations that the experimenter manipulates, controls or observes. The variables selected in the present study by the investigator are as follows.

1. Independent Variable

In this study, the investigator considered **Awareness of internet utilization** (that mean, Awareness about Privacy Settings, Sharing Files, Private Browsing, Spam Mails, Clearing History, Webcam Usage and etc..) are the independent variable.

2. Dependent Variable

The dependent variable is subject to change. It is the measure or indicator of the effect of changes caused by independent variables. In the present study, **Privacy protection Level (score)** has been treated as a dependent variable.

IV. DATA ANALYSIS AND INTERPRETATION

Descriptive Analysis

The descriptive analysis of the data involves percentiles, mean and standard deviation. Mean is one of computing measures of central tendency whereas standard deviation is one of the measures of variability.

TABLE 1Distribution of sample (Age wise)

Age	Frequency	Percent	Valid Percent	Cumulative Percent
Up to 17	29	14.5	14.5	14.5
17-20	39	19.5	19.5	34.0
21 – 25	132	66.0	66.0	100.0
Total	200	100.0	100.0	

Relational Analysis

Correlation is the relationship between two or more paired variables or two or more sets of data. The degree of relationship is measured and represented by the coefficient of correlation. Correlation analysis explains qualitative relationship between two variables. The Chi-square ($\chi 2$) test is also used to compare a sample variance to a theoretical population variance. It is applied to examine whether two attributes are associated or not. Multiple regression analysis also employed to explain quantitative relationship among more than two variables in the form of equations.

Factor analysis

Factor analysis is a statistical technique used to study the inter-relationships among the variables in an effort to reduce the large number of variables into a few dimensions called factors that summarize the available data. Its aims at grouping the original input variables into factors in which the input variables are underlying.

Table 2 Identifying factors

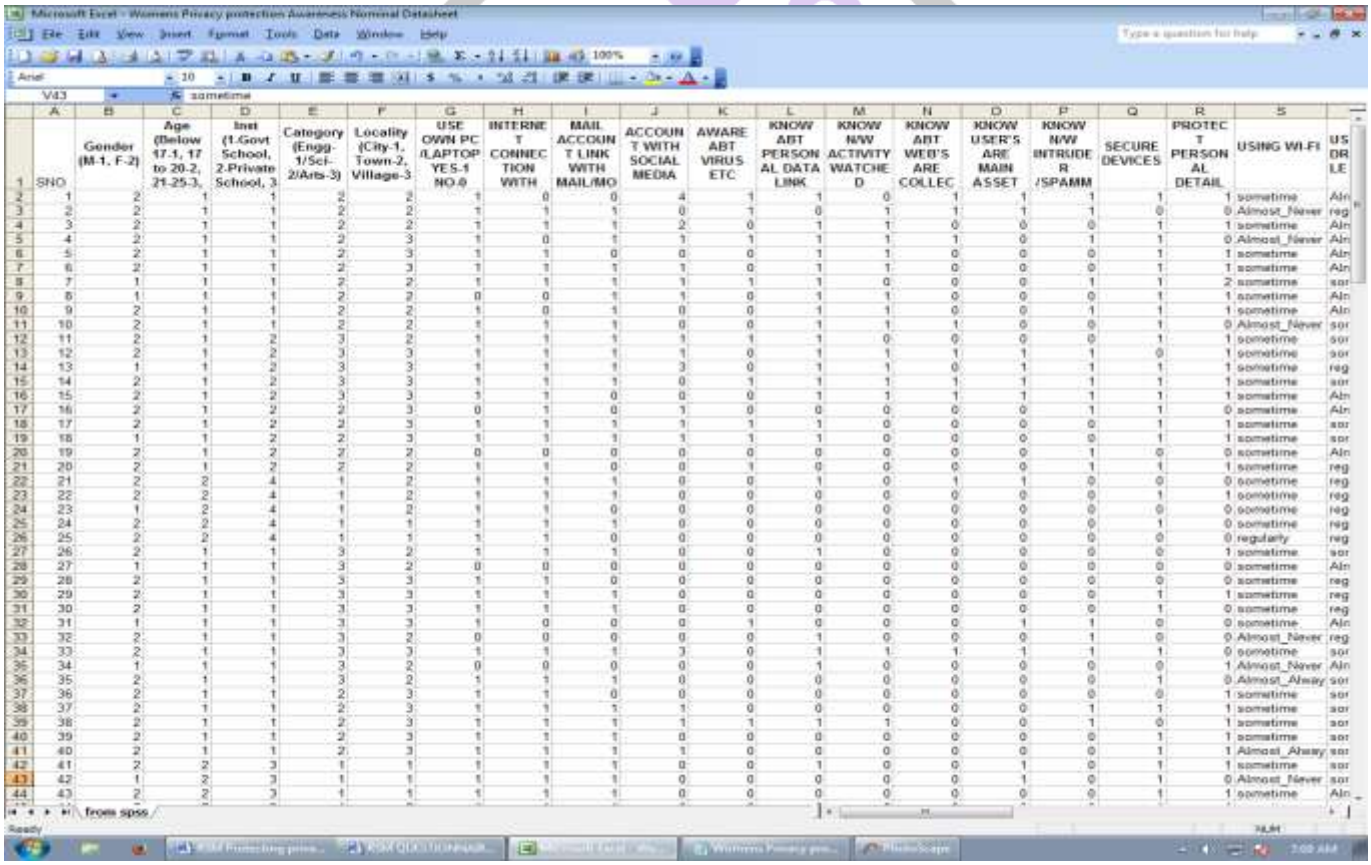
Communalities		
	Initial	Extraction
Own Computer	1.000	.648
High Speed Net Connection	1.000	.664
Email Account	1.000	.734
No of Social Media Accounts	1.000	.603
Awareness about Virus, Malwares, Spywares,Spammers	1.000	.580
Awareness about Personal Data Sharing	1.000	.586
Awareness about Tracking	1.000	.567
Aware about Internet Traking and Metadata	1.000	.774

Aware about Metadata investment	1.000	.783
Indruder Awareness	1.000	.691
Password Security Awareness	1.000	.362
Chat and Sharing Awareness	1.000	.675
Use of WiFi Connection	1.000	.708
Use of Online Storage Drives	1.000	.602
Use of Google Docs	1.000	.675
Clearing History Datas Frequently	1.000	.600
Use of private browsing	1.000	.594
Sharing Datas with group chat	1.000	.735
Facebook privacy settings and tools	1.000	.449
Facebook activity log clearing	1.000	.669
Awareness about Spam mails	1.000	.659
Webcam Awareness	1.000	.567
GPS Tracking Awareness	1.000	.547
Virtual Keyboard use in Net Banking	1.000	.665
Awareness about Teamviewer and RDC	1.000	.690

V. RESULT AND DISCUSSIONS

Analysis Of Data Collection Using Weka Tool

A total of 200 records were taken for the data mining analysis.



SNO	Gender (M-1, F-2)	Age (Below 17-1, 17 to 20-2, 21-25-3,	Inst (1-Govt School, 2-Private School, 3	Category (Engg-1/Sci-2/Arts-3)	Locality (City-1, Town-2, Village-3	USE OWN PC /LAPTOP YES-1 NO-0	INTERNET CONNECTION WITH	MAIL ACCOUNT LINK WITH MAIL/MO	ACCOUNT WITH SOCIAL MEDIA	AWARE ABT VIRUS ETC	KNOW ABT PERSONAL DATA LINK	KNOW N/W ACTIVITY WATCHED	KNOW ABT WEB'S ARE COLLEC	KNOW USER'S ARE MAIN ASSET	KNOW N/W INTRUDER /SPAMM	SECURE DEVICES	PROTECT PERSONAL DETAIL	USING WI-FI	US DR LE
1	2	1	1	2	2	1	0	0	4	1	1	0	1	1	1	1	1	sometime	An
2	2	1	1	2	2	1	1	1	0	1	0	1	1	1	1	0	0	Almost_Never	reg
3	2	1	1	2	2	1	1	1	2	0	1	1	0	0	0	1	1	sometime	An
4	2	1	1	2	3	1	0	1	1	1	1	1	1	0	1	1	0	Almost_Never	An
5	2	1	1	2	3	1	1	0	0	0	1	1	0	0	0	1	1	sometime	An
6	2	1	1	2	3	1	1	1	1	0	1	1	0	0	0	1	1	sometime	An
7	1	1	1	2	2	1	1	1	1	1	1	0	0	0	1	1	2	sometime	sor
8	1	1	1	2	2	0	0	1	1	0	1	1	0	0	0	1	1	sometime	An
9	2	1	1	2	2	1	0	1	0	0	1	1	0	0	1	1	1	sometime	An
10	2	1	1	2	2	1	1	1	0	0	1	1	1	0	0	1	0	Almost_Never	sor
11	2	1	2	3	2	1	1	1	1	1	1	0	0	0	0	1	1	sometime	sor
12	2	1	2	3	3	1	1	1	1	0	1	1	1	1	1	0	1	sometime	sor
13	1	1	2	3	3	1	1	1	3	0	1	1	0	1	1	1	1	sometime	reg
14	2	1	2	3	3	1	1	1	0	1	1	1	1	1	1	1	1	sometime	sor
15	2	1	2	3	3	1	1	0	0	0	1	1	1	1	1	1	1	sometime	An
16	2	1	2	2	3	0	1	0	1	0	0	0	0	0	1	1	0	sometime	An
17	2	1	2	2	3	1	1	1	1	1	1	0	0	0	0	1	1	sometime	sor
18	1	1	2	2	3	1	1	1	1	1	1	0	0	0	0	1	1	sometime	sor
19	2	1	2	2	2	0	0	0	0	0	0	0	0	0	0	1	0	sometime	An
20	2	1	2	2	2	1	1	0	0	1	0	0	0	0	1	1	1	sometime	reg
21	2	2	4	1	2	1	1	1	0	0	1	0	1	1	0	0	0	sometime	reg
22	2	2	4	1	2	1	1	1	0	0	0	0	0	0	0	1	1	sometime	reg
23	1	2	4	1	2	1	1	0	0	0	0	0	0	0	0	0	0	sometime	reg
24	2	2	4	1	1	1	1	1	0	0	0	0	0	0	0	1	0	sometime	reg
25	2	2	4	1	1	1	1	0	0	0	0	0	0	0	0	0	0	regularly	reg
26	2	1	1	3	2	1	1	1	0	0	1	0	0	0	0	0	1	sometime	sor
27	1	1	1	3	2	0	0	0	0	0	0	0	0	0	0	0	0	sometime	An
28	2	1	1	3	3	1	1	0	0	0	0	0	0	0	0	0	0	sometime	reg
29	2	1	1	3	3	1	1	1	0	0	0	0	0	0	0	1	1	sometime	reg
30	2	1	1	3	3	1	1	1	0	0	0	0	0	0	0	1	0	sometime	reg
31	1	1	1	3	3	1	0	0	0	1	0	0	0	1	1	0	0	sometime	An
32	2	1	1	3	2	0	0	0	0	0	1	0	0	0	1	0	0	Almost_Never	reg
33	2	1	1	3	3	1	1	1	1	0	1	1	1	1	1	1	0	sometime	sor
34	1	1	1	3	2	0	0	0	0	0	1	0	0	0	0	0	1	Almost_Never	An
35	2	1	1	3	2	1	1	1	0	0	0	0	0	0	0	1	0	Almost_Alway	sor
36	2	1	1	2	3	1	1	0	0	0	0	0	0	0	0	0	1	sometime	sor
37	2	1	1	2	3	1	1	1	1	0	0	0	0	0	1	1	1	sometime	sor
38	2	1	1	2	3	1	1	1	1	1	1	0	0	0	1	0	1	sometime	sor
39	2	1	1	2	3	1	1	1	0	0	0	0	0	0	0	1	1	sometime	sor
40	2	1	1	2	3	1	1	1	1	0	0	0	0	0	0	1	1	Almost_Alway	sor
41	2	2	3	1	1	1	1	1	0	0	1	0	0	1	0	1	1	sometime	sor
42	1	2	3	1	1	1	1	1	0	0	0	0	0	1	0	1	0	Almost_Never	sor
43	2	2	3	1	1	1	1	1	0	0	0	0	0	0	0	1	1	sometime	An

Attribute Selection

Attribute selection searches through all possible combinations of attributes in the data and finds which subset of attributes works best for prediction. Attribute selection methods contain two parts: an attribute evaluator and a search method. The evaluator such as correlation-based, wrapper, information gain and chi-squared are determines what method is used to assign a worth to each subset of attributes. The search method such as best-first, forward selection, random, exhaustive, genetic algorithm and ranking are determines what style of search is performed.

In this study, ChiSquaredAttributeEval is used as evaluator and Ranker as search method. These methods extract 29 attributes from 30 attributes of this study. So we select the attributes that mostly related to this study that have the chi squared value more than 30. The selected attributes as follows.

Chi-Squred Value	Attribute Name
86.965	Awareness_abt_Privacy_setting_tool
59.0799	Awareness_abt_Sharing_files_with_groups
60.9646	Awareness_abt_private_browsing
73.9207	Awareness_abt_Spam_mails
62.9586	Awareness_abt_Clear_History
30.8112	Awareness_abt_usage_of_webcame
36.3478	Awareness_abt_remote_desktop

After selecting this attributes, we load the test and train dataset into weka. The aim of our work is to Detecting Privacy protection awareness when mining and sharing Female Student's Data through Online. Normally, many test methods involve the classification of large scale data. But too many tests could complicate the main analysis process and lead to the difficulty in obtaining the end results, particularly in the case where many tests are performed. This kind of difficulty could be resolved with the aid of machine learning could be used directly to obtain the end result with the aid of several Data Mining Algorithms which perform the role as classifiers.

After the Detailed Analysis of Women's Privacy protection Awareness Nominal Data Table using Weka, the following conclusion table was derived.

Table 3 Relation between Nominal class attribute and WEKA classifiers

S.No	Nominal Class Attribute	Predicted By Weka Classifiers		
		Naïve Bayes	***SMO***	***J48 Tree***
1	Awareness_abt_Privacy_setting_tool	75.5%	96.5%	93.5%
2	Awareness_abt_Sharing_files_with_groups	67.5%	94%	91.5%
3	Awareness_abt_private_browsing	61.5%	93.5%	95%
4	Awareness_abt_Spam_mails	73%	93.5%	88%
5	Awareness_abt_Clear_History	67%	95.5%	91%
6	Awareness_abt_usage_of_webcam	68%	90%	89.5%
7	Awareness_abt_remote_desktop	69%	97.5%	93%
8	Awareness Scale (Awareness Level of Womens Privacy protection through online)	99%	99.5%	100%
	Average	72.56%	95%	92.68%

Noted the above table we are clearly concluded that, SMO (Support Vector Machine) Classifier gives high prediction level (95%) with more accuracy.

Analysis Result

1. 54.5% of students at age level of 21-25 have awareness about viruses. 59% of students at the age level of 17-20 have more awareness about viruses. 41.4% of students at the age level of 16-17 have less awareness about virus and malwares.
2. 48.5% of students have more awareness about network activities tracking at age level of 21-25.
3. 62.1% of students have more awareness about protecting personal details at age level of 21-25.
4. 44% of Students at all age groups (88/200) have awareness about handling privacy setting tools while using social medias and cloud databases.
5. 59 % Students at all age groups (112/200) have awareness about sharing files with social media groups like chat groups or forums.
6. 65.5% of Students at all age groups (131/200) have awareness and use of private browsing on the internet.
7. 56.5% of Students at all age groups (113/200) have awareness about spam mails.
8. 57% of user have knowledge about Clearing History and cookies. History collect users every day browser activities frequently, based on this, they send spam mails to us, track our hobbies etc etc.
9. 48.5% of students have awareness about usage of Web camera. Introducer use unprotected web camera's like a remote desktop.
10. 32.5% of students have less awareness about women's privacy protection. 19.5% of Students have Average level of Women's privacy protection. 29% Have moderate level of Awareness. 19% of students have high level of awareness about women's privacy protection while sharing data through online.

VI. CONCLUSIONS

In the results acquired, it may be concluded that awareness about the female students privacy protection is 67.5% (included Average, Moderate and High). 32.5% of participants doesn't have enough awareness to privacy protection and skill to effective handling of internet. We proposed for future work with Analysis of Privacy protection awareness when mining and sharing Female Student's Data through Online, the researcher take larger sample covered wide geographical area and large age range with most popular questions that are used for day to day life of the students.

REFERENCES

[1] Shearin Sybil, Henry Lieberman (2001), "Intelligent Profiling by Example", ACM Conference on Intelligent User Interfaces, Santa Fe, NM, January 2001.

[2] Ke Wang and Philip S.Yu (2008) " Privacy-Preserving Data Publishing: A Survey of Recent Developments", Simon Fraser University, Burnaby & University of Illinois, Chicago.

[3] Rakesh Agrawal and Ramakrishnan Srikanth(2012), "Fast Algorithms for Mining

[4] Association Rules", IBM, Almaden Research Center, CA.

[5] Kamalika Das(2009), " Privacy Preserving Distributed Data Mining based on Multi objective Optimization and Algorithmic Game Theory", University of Maryland.

[6] M. Sachan, D. Contractor, T. A. Faruquie, and L.V.Subramaniam, "Using Content and Interactions for Discovering Communities in Social Networks", in Proc. of the 21st Int. Conf. on World Wide Web, 2012

[7] Gayatri Nayak and Swagatika Devi (2011), "A Survey On Privacy Preserving Data Mining: Approaches And Techniques", International Journal of Engineering Science and Technology.

[8] Li Xiong and Subramanyam Chitti, Ling Liu, "k Nearest Neighbor Classification acrossMultiple Private Databases", Emory & Georgia institute of technology.

[9] Srikanth Kandula, Sankalp Singh, and Dheeraj Sanghi. Argus " A Distributed Network Intrusion Detection System", In Proceedings of USENIX SANE, 2002.

[10] Chris Clifton and Murat Kantarcioglu and Jaideep Vaidya (2002)," Defining Privacy for Data Mining", Proceedings of the National Science Foundation Workshop on Next Generation Data Mining, pp.274-281

[11] K. Sathiyapriya, Dr. G. Sudha Sadasivam,, "A Survey on Privacy Preserving Association Rule," International Journal of Data Mining & Knowledge Management Process (IJDKP), vol. 3, no. 2, pp. 119-131, 2013.