International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 9 Issue 06, June-2020

Survey on Privacy Preserving Data Mining Techniques

Shilpa Rathod
Information Technology Department
V.V.P. Engineering College, Rajkot

Dr. Darshana Patel
H.O.D. Information Technology Department
V.V.P. Engineering College, Rajkot

***Abstract*:- Privacy preservation in Data Mining has become more prominent and popular because of its property of maintaining privacy of sensitive data for analysis purposes. In this decade, enormous volume of data is created by many sectors especially healthcare, and it is vital to analyze and extract the right information out of it. For instance, the integration of patient's medical records and health test data helps to identify the relation between atypical test result and disease. Incorporating association rule mining on this data aids in creating new information which contributes in disease prevention. During association rule mining procedure, it is crucial to maintain the privacy and security of data, the business's vital information should not be leaked. In this paper, we provide an effective solution of privacy preservation along with association rule mining. Our paper is focused on healthcare datasets; however, it can be extended and implemented in various areas**

Index Terms- Data Mining, Privacy Preserving, Association Rule Mining, Cryptography

I. INTRODUCTION

(A)

Data Mining is a set of method that applies to large and complex databases. This is to eliminate the randomness and discover the hidden pattern. It has attracted a great deal of attention in the information industry and in society. Data mining includes the utilization of refined data analysis tools to find previously unknown, valid patterns and relationships in huge data sets. These tools can incorporate statistical models, machine learning techniques, and mathematical algorithms, such as neural networks or decision trees. Thus, data mining incorporates analysis and prediction. In recent data mining projects, various major data mining techniques have been developed and used, including association, classification, clustering, prediction, sequential patterns, and regression.

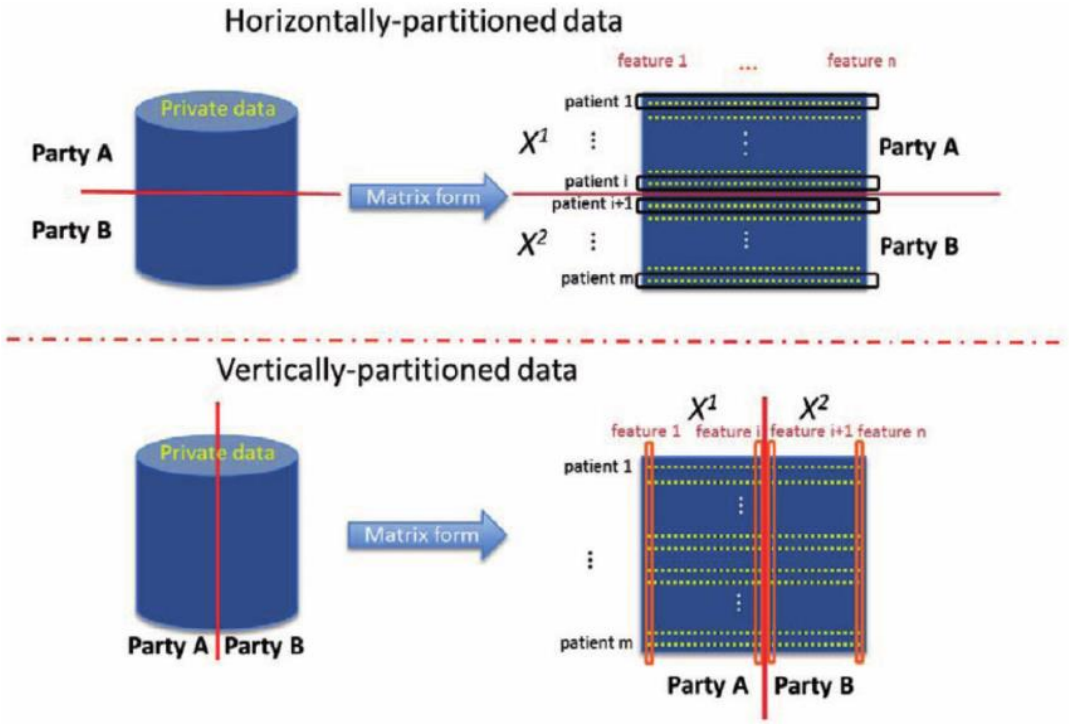
(B) Data Partition Model

Partitioning a data set is splitting the data into two, sometimes three smaller data sets. These are called Training, Validation and Test. This technique is best practice when creating a predictive model but is only possible when working with enough data. Test data sets are less common due the volume of data required.

If a predictive model is created to fit a specific data set, it is possible to create a highly predictive model. To ensure that this model will predict new data well, it should be tested on a different sample of data to see how accurate it is. Data partitioning is used to split the original data set before the model is created so that there is 'new' data available to assess the model.In distributed environment, two major types of data partition model among all participants that is, Horizontal Partitioned and Vertical Partitioned data model are given.

Here, we discuss some of the existing approaches towards privacy persevering distributed association rule mining of both kinds ofdatapartitioning. In the horizontally partitioned data, all the participants have the same schema, but each participant contains the records of different entities.

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 9 Issue 06, June-2020



Scenario- horizontal partition: Researchers at the global level, many private hospitals with their local HER system data want to undertake the association rule mining. Each local EHR system contains the data about the patient's ID, sex, age, disease, treatment, duration etc. These hospitals are interested in a collaborative research for finding the relation among the sex, age, diagnosis and disease by the global dataset. As a result of the collaborative association rule mining, all the hospitals learn the rule.

Scenario – vertical partition: In Each participant has deferent schema and it stores the data of the same set of entities. Privacy preserving association rule mining in vertically partition data discussed in.As shown in the figure above,Healthexaminationdataandoutpatientdatacanbeintegratedbasedoncommon IDfordiscovering the correlations between abnormal test results and some disease. Medical researchers are interested in computing the association rule mining on the integrated data of the health examination record and outpatient record. Association rules discovered from this collaboration help to discover some correlations between some disease sand patient's attributes. However, sharing of patient's information violates the privacy. Hence, Privacy preserving association rule mining in vertically partitioned healthcare data has received the substantial attention of medical researchers.

(C) Privacy Preserving Data Mining

Privacy has become crucial in knowledge based applications. Proper integration of individual privacy is essential for data mining operations. This privacy based data mining is important for sectors like Healthcare, Pharmaceuticals, Research, and Security Service Providers, to name a few. The main categorization of Privacy Preserving Data Mining (PPDM) techniques falls into Perturbation, Secure Sum Computations and Cryptographic based techniques.

With modern world getting digitized, there is an increase in electronic data. It is important to analyze socio-economic trends of the individuals of the society. Privacy concern is important when data disclosure is taken into account. Privacy can be defined as prevention of

Unwanted disclosure of information when data mining is performed on aggregate results. Privacy must be addressed at all the levels while mining is carried out[3][1]

Before data mining tasks are carried out, several methods must be applied to protect the privacy of individuals. Privacy preserving data mining is the branch which includes the studies of privacy concern when mining is applied. Various methods like data hiding, masking, suppression, aggregation, perturbation, anonymization, SMC are studied in literature Based on the location of computation carried out for mining results, PPDM techniques can be classified as described in Figure 1. The mining can be entrusted to a trusted third party who collects all sensitive data. Another scenario is when the individual parties privatize their data before mining process is carried out. The classification thus can be broadly categorized as: Central/Commodity Server and Distributed. The mplementation of various techniques related to Fuzzy and Neural Networks is still rudimentary and is discussed in brief here.

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 9 Issue 06, June-2020

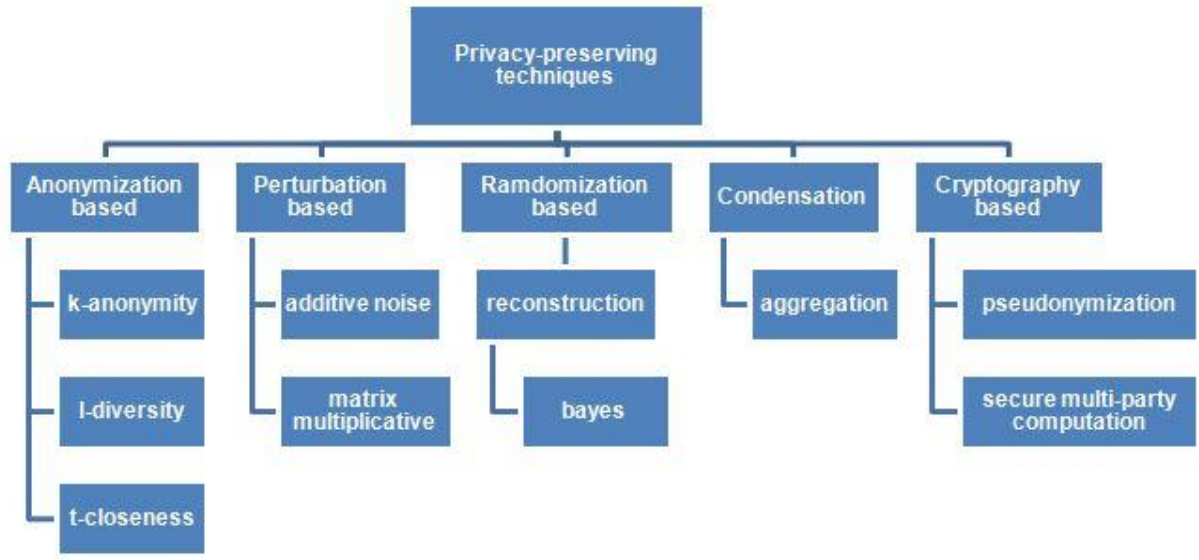


Figure : Taxanomy of privacy-preserving techniques.

LITERATURE REVIEW

Various methods

A)Anonymization based PPDM

The basic form of the data in a table consists of following four types of attributes:
(i) Explicit Identifiers is a set of attributes containing information that identifies a record owner explicitly such as name, SS number etc. (ii) Quasi Identifiers is a set of attributes that could potentially identify a record owner when combined with publicly available data. (iii) Sensitive Attributes is a set of attributes that contains sensitive person specific information such as disease, salary etc. (iv) Non-Sensitive Attributes is a set of attributes that creates no problem if revealed even to untrustworthy parties.

Anonymization refers to an approach where identity or/and sensitive data about record owners are to be hidden. It even assumes that sensitive data should be retained for analysis. It's obvious that explicit identifiers should be removed but still there is a danger of privacy intrusion when quasi identifiers are linked to publicly available data. Such attacks are called as linking attacks. For example attributes such as DOB, Sex, Race, and Zip are available in public records such as voter list.

Such records are available in medical records also, when linked, can be used to infer the identity of the corresponding individual with high probability as shown in fig.3.

Sensitive data in medical record is disease or even medication prescribed. The quasi-identifiers like DOB, Sex, Race, Zip etc. are available in medical records and also in voter list that is publicly available. The explicit identifiers like Name, SS number etc. have been removed from the medical records. Still, identity of individual can be predicted with higher probability. Sweeney proposed k-anonymity model using generalization and suppression to achieve k-anonymity i.e. any individual is distinguishable from at least k-1 other ones with respect to quasi-identifier attribute in the anonymized dataset. In other words, we can outline a table as kanonymous if the Q1 values of each raw are equivalent to those of at least k- 1 other rows. Replacing a value with less specific but semantically consistent value is called as generalization and suppression involves blocking the values. Releasing such data for mining reduces the risk of identification when combined with publically available data. But, at the same time, accuracy of the applications on the transformed data is reduced. A number of algorithms have been proposed to implement k-anonymity using generalization and suppression in recent years. Although the anonymization method ensures that the transformed data is true but suffers heavy information loss. Moreover it is not immune to homogeneity attack and background knowledge attack practically . Limitations of the k-anonymity model stem from the two conventions. First, it may be very tough for the owner of a database to decide which of the attributes are available or which are not available in external tables. The second limitation is that the kanonymity model adopts a certain method of attack, while in real situations; there is no reason why the attacker should not try other methods. However, as a research direction, kanonymity in combination with other privacy preserving methods needs to be investigated for detecting and even blocking k-anonymity violations.[12][20]**[29]**

B) Perturbation Based PPDM

Perturbation being used in statistical disclosure control as it has an intrinsic property of simplicity, efficiency and ability to reserve statistical information. In perturbation the original values are changed with some synthetic data values so that the statistical information computed from the perturbed data does not differ from the statistical information computed from the original data to a larger extent. The perturbed data records do not agree to real-world record holders, so the attacker cannot

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 9 Issue 06, June-2020

perform the thoughtful linkages or recover sensitive knowledge from the available data. Perturbation can be done by using additive noise or data swapping or synthetic data generation. In the perturbation approach any distribution based data mining algorithm works under an implicit assumption to treat each dimension independently. Relevant information for data mining algorithms such as classification remains hidden in inter-attribute correlations. This is because the perturbation approach treats different attributes independently. Hence the distribution based data mining algorithms have an intrinsic disadvantage of loss of hidden information available in multidimensional records. Another branch of privacy preserving data mining that manages the disadvantages of perturbation approach is cryptographic techniques.[12][28][29]

C) Randomized Response Based PPDM

Randomized response is statistical technique introduced by Warner which is used to solve a survey problem. In Randomized response technique, the data is twisted in a way that the central place cannot have chances better than a predefined threshold, whether the data contains correct information or incorrect information. The information received by each single user is twisted and if the number of users is large, the aggregate information of these users can be calculates with good quantity of accuracy. This is important for decision-tree classification. It is based on combined values of a dataset, somewhat individual data items. The data collection process in randomization method is carried out using two steps [29]. During first step, the data providers randomize their data and transfer the randomized data to the data receiver. In second step, the data receiver rebuilds the original distribution of the data by using a distribution reconstruction algorithm.

Randomization Response Model Randomization method is relatively very simple and does not require knowledge of the distribution of other records in the data. Hence, the randomization method can be implemented at data collection time. A trusted server does not require to contain the entire original records in the anonymization process [6]. The weakness of a randomization response based PPDM technique is that it treats all the records equal irrespective of their local density. These indicate to a problem where the outlier records become more subject to oppositional attacks as compared to records in more compressed regions in the data . One key to this is to be uselessly adding noise to all the records in the data. But, it reduces the utility of the data for mining purposes as the reconstructed distribution may not yield results in conformity of the purpose of data mining.[6][28][12]

D) Condensation approach based PPDM

Condensation approach compresses and packs the raw input data into multiple groups or clusters. Each group or cluster has constraint which is defined for it in terms of its size. This size is referred as the level of that privacypreserving approach. Greater is the level, the greater will be the amount of privacy. This size is chosen in a way so as to preserve k-anonymity. After condensing data into clusters, statistics of data in each group is analyzed and maintained separately for each cluster. This statistics from each cluster is used further to generate pseudo data for corresponding clusters. In the process of data mining, data owner publish this pseudo data instead of original data. Various data mining tasks use this pseudo data as input. In this way actual data remains hidden from other parties [7]. This technique is referred as condensation because of its approach of using condensed statistics of the clusters in order to generate pseudo-data. In this approach, a greater amount of information is lost because of the condensation of a larger number of records into a single statistical group entity. Following figure shows steps in followed in the condensation approach.[29]

E) Cryptography Based PPDM

Consider the situation where multiple medical institutions wish to conduct a joint research for some mutual benefits without sharing unnecessary information. In this scenario, research regarding symptoms, diagnosis and medication based on various parameters is to be conducted and at the same time privacy of the individuals is to be protected. Such scenarios are referred to as distributed computing scenarios [4].The parties involved in mining of such tasks can be untrusted parties, competitors; therefore protecting privacy becomes a major concern. Cryptographic techniques are ideally meant for such scenarios where multiple parties collaborate to compute results or share non sensitive mining results and thereby avoiding disclosure of sensitive information. Cryptographic techniques find its utility in such scenarios because of two reasons: First, it offers a well defined model for privacy that includes methods for proving and quantifying it. Second a vast set of cryptographic algorithms and constructs to implement privacy preserving data mining algorithms are available in this domain. The data may be distributed among different collaborators vertically or horizontally. All these methods are almost based on a special encryption protocol known as Secure Multiparty Computation (SMC) technology. Moreover, the data mining results may breach the privacy of individual records. There exist a good number of solutions in case of semi-honest models but in case of malicious models very less studies have been made. [4][5][12]

REVIEW PAPERS

This section describes about study of existing work/methods in specific area such As web log mining, prepossessing methods and observation from experiments performed of different Research Papers.

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 9 Issue 06, June-2020

3.1 Privacy Preserving Distributed Association Rule Mining Approach on Vertically Partitioned Healthcare Data
Publication: ICDS
Year:2018

Problem Identified:
In data mining we need to use some data to be integrated for better results and partitioned but they consist of some sensitive information so we need to preserve the privacy of such data

Proposed System:
- This paper investigate how the medical research can be improved by the collaborative association rule mining on vertically partitioned healthcare data. Privacy of patients must be preserved during this collaboration..
- The privacy preserving distributed data mining (PPDDM) problem can be solved using the cryptography techniques.
- The system model consist of three systems(i.e Participant A (EHR with Medical Examination data of patients), Participant B (EHR with Outpatient Medical Records) and Key Generation Center (KGC)

Results:
The results compute the communication cost and computation cost For different sized datasets

3.2 An Efficient Approach for Privacy Preserving in Data Mining
Publication: IEEE
Year : 2014

Problem Identified:
Many techniques are present for privacy preserving in data mining but they have some shortcomings like information loss and data utility. This research work is mainly focus on combined method of randomization and k-anonymity techniques to preserve the privacy, increase data utility and decrease information loss.

Proposed System
The proposed approach uses the combined techniques of randomization and k-anonymization Mainly proposed approach is dived into two algorithms. In algorithm I randomization is performed on dataset using attribute transitional probability matrix and in algorithm II k-anonymity is performed on randomized dataset which is result of algorithm I.

Results:

By using Randomization technique attacker cannot identifY a pattern of data. K-anonymity method has shortcoming of homogeneity and background attack. In the proposed method we combined K-anonymity with randomization. It makes difficult for the attacker to identify Y background and homogeneity attack. Apart from that it protects private data with better accuracy and gives no loss of information which increases data utility. Data can also be reconstructed by our proposed approach.

This had applied data perturbation on the different datasets. All the datasets are downloaded from archive.ics.uci.edu (Machine Learning, UCI).

3.3 Cryptanalysis of a Privacy-Preserving Aggregation Protocol
Publication: IEEE
Year: 2017

Problem Identified:
Privacy-preserving aggregation protocols allow an untrusted aggregator to evaluate certain statistics over a population of individuals without learning each individual's privately owned data. In this note, we show that a recent protocol for computing an aggregate sum due to Jung, Li, and Wan (IEEE Transactions on Dependable and Secure Computing, 2015) is universally breakable, that is, anyone is able to recovereachindividual'sprivatedatafromthecorrespondingciphertext. We also describe an alternate collusion attack against their companion product protocol..

Proposed System:
We show that their sum protocol can be trivially broken to obtain the plaintext value from the ciphertext. We also point out a collusion attack.

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 9 Issue 06, June-2020

Result:
Compared with existing proactive methods, PRMRAP considers not only horizontal resizing but also vertical resizing of VMs which makes it even quicker and much more cost-effective.

3.4 Privacy-preserving association rule mining for horizontally partitioned healthcare data: a case study on the heart diseases
Publication: Sådhanå (2018) 43:127
Problem Identified:
Initially, investigate how the accuracy of medical research can be increased by the collaborative association rule mining and at the same time, privacy issues need to be focused. We propose an efficient approach (PPDARM for medical research) on horizontally partitioned data for higher accurate association results and preserving the privacy of patients. We analyze the proposed scheme with heart disease dataset available publically on UCI repository. The analysis of results shows that the accuracy of association among the diseases and symptoms increases by the collaborative mining and privacy of patients is preserved.

Proposed System:
Four collaborative participants or EHR systems are used to evaluate the proposed approach. Therefore, the original dataset is randomly divided into four partitions. The results of each EHR system and the collaboration of four EHR systems are shown in table 2. Experiments are conducted using NetBeans on a machine with an intel core i3 CPU, 4 GB RAM, and 2.0 GHz speed. Result of any single EHR system is compared to the result of the collaboration in terms of accuracy

Results:
The result shows that prediction of heart disease using any single EHR system has lower accuracy (confidence) in some EHR systems compared with the result of collaborative EHR systems. The analysis of experimental result based on heart disease dataset shows that the female gender indicates lower chances of coronary heart disease.

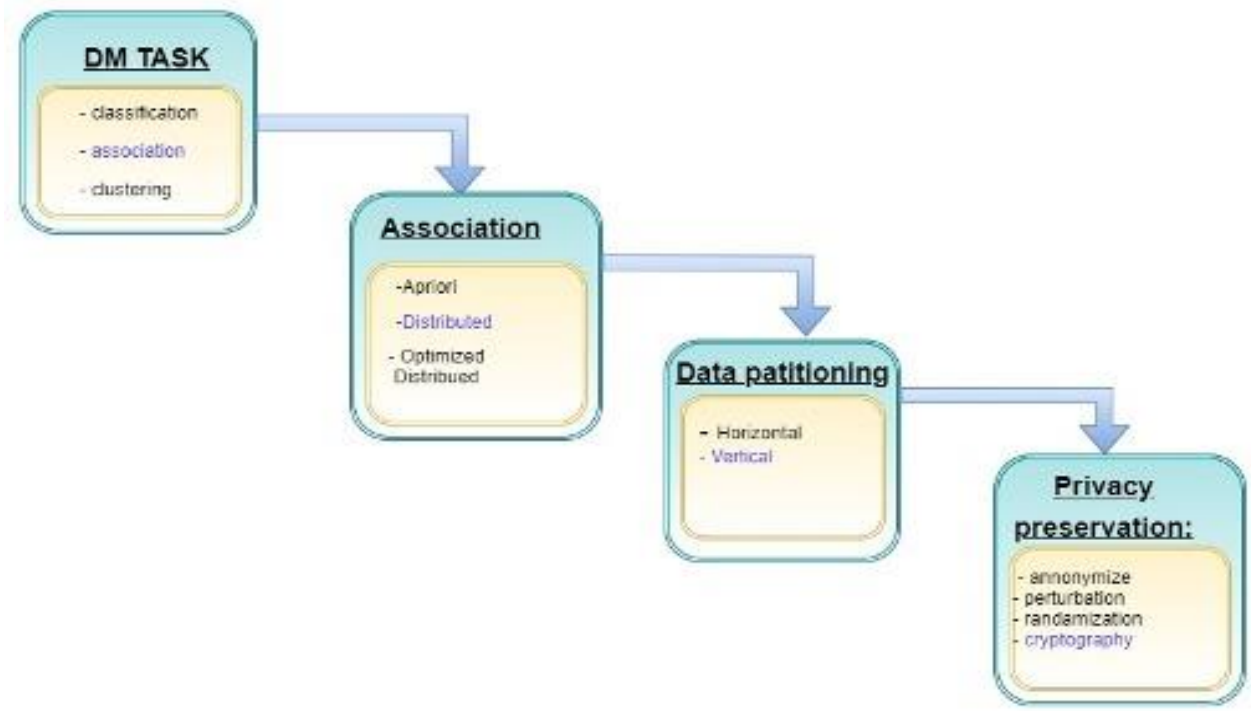
3.5 Privacy Preserving Health Data Mining
Publication: IJCST Vol. 6, Issue 4,
Year: 2015
Problem Identified:
Releasing of medical data without influencing individual privacy, we need to anonymize the document .Here in our work we concentrate on the clustering of data. Therefore the major areas which related to our work are information extraction, clustering and privacy.

Proposed systems:
In the proposed system we are anonymzing the medical data without affecting the quality of medical data. The data can be used for the data mining purpose; here we are considering the clustering of data. The data is anonymized using LKC privacy technique.

Results:
Stable service quality can be ensured and proposed schedule-based mechanism has demonstrated to be a better choice when workload variation can be predicted.

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 9 Issue 06, June-2020

PROPOSED FLOW



CONCLUSION AND FUTURE WORK

The importance of data mining in Healthcare for improving the Medical research Privacy issues during the collaborative data mining for medical research have been taken . To solve this, an efficient approach for privacy preserving association rule mining on vertically partition healthcare data applies. The theoretical analysis of proposed algorithm is done by fundamental study and literature review.

Until now we have studied the existing system and its implementation. Now , we are going to apply cryptographic algorithm and compare our new approach ECC with RSA on to the association rule vertical partitioning data to reduce cost of the model while working with more than 2 participants for collaboration.

REFERENCES

[1] B. Fung, K. Wang, R. Chen, P. Yu, “Privacy-Preserving Data Publishing: A Survey Of Recent Developments”, ACM Computing Surveys, June 2010.
[2] J. Han, M. Kamber, “Data Mining: Concepts and Techniques”, Morgan Kaufmann Publishers Inc., 2005
[3] Alpa Shah, Ravi Gulati” Privacy Preserving Data Mining: Techniques, Classification and Implications - A Survey”, International Journal of Computer Applications (0975 – 8887) Volume 137 – No.12, March 2016
[4] Benny Pinkas, "Cryptographic Techniques for Privacy preserving data mining", SIGKDD Explorations, Vol. 4, Issue 2, 12-19, 2002.
[5] S. Taneja, S. Khanna, S. Tiwalia, “A Review on Privacy Preserving Data mining: Techniques and research challenges”, International Journal of Computer Science and Information Technologies, vol. 5, 2014.
[6] Ahmed K. Elmagarmid, Amit P. Sheth “PrivacyPreserving Data Mining Models and algorithm” advances in database systems 2008.
[7] Agrawal, R., Srikant, R., et al., 1994. Fast algorithms for mining association rules, in: Proceeding of 20th international conference on very large data bases, VLDB, pp. 487–499.
[8] Domadiya, N., Rao, U.P., 2018. Privacy-preserving association rule mining for horizontally partitioned healthcare data: a case study on the heart diseases. S¯adhan¯a 43, 127.
[9] ElGamal, T., 1985. A public key cryptosystem and a signature scheme based on discrete logarithms. IEEE transactions on information theory 31, 469–472.
[10] Ge, X., Yan, L., Zhu, J., Shi, W., 2010. Privacy preserving distributed association rule mining based on the secret sharing technique, in: Proceedings of 2nd International Conference on Software Engineering and Data Mining (SEDM), IEEE. pp. 345–350.
[11] Murut, Chris , “Privacy-Preserving Distributed Mining of Associative Rules on Horizontally Partitioned data”, IEEE transactions on knowledge and data engineering, 2-13,2004.
[12] Hina Vaghashia, Amit Ganatra,”A Survey: Privacy Preservation Techniques in Data Mining”, International Journal of Computer Applications (0975 – 8887) Volume 119 – No.4, June 2015
[13] Datta, A., Joye, M., 2016. Cryptanalysis of a privacy-preserving aggregation protocol. IEEE Transactions on Dependable and Secure Computing 82, 23–30.
[14] Leelavathy, J., Selvabrundha, S., 2017. A novel approach to classify users based on keystroke behavior. Cluster Computing , 1–9.
[15] Kantarcioglu, M., Clifton, C., 2004. Privacy preserving distributed mining of association rules on horizontally partitioned data. IEEE Transactions on Knowledge & Data Engineering , 1026–1037.
[16] Cleveland heart disease data details. [Online] vailable:http://archive.ics.uci.edu/ml/machine-learning-databases/heart-disease/heart-disease.names [accessed: 28-May-2016]
[17] K. Saranya, K. Premalatha, S. Rajasekar, A survey on privacy preserving data mining, in International Conference on Electronics & Communication System (IEEE, 2015)
[18] Ordonez C 2006 Association rule discovery with the train and test approach for heart disease prediction. IEEE Trans. Inf. Technol. Biomed. 10(2): 334–343

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 9 Issue 06, June-2020

[19] Shin A M, Lee I H, Lee G H, Park H J, Park H S, Yoon K I, Lee J J and Kim Y N 2010 Diagnostic analysis of patients with essential hypertension using association rule mining. Healthc. Inform. Res. 16(2): 77–81
[20] Australian Bureau of Statistics 2016 Causes of death in australia. [Online] Available: http://abs.gov.au/ausstats/abs. nsf/Products [accessed: 12-May-2016]
[21] Leading-causes-of-death, australia. [Online] Available: http://www.aihw.gov.au/deaths/leading-causes-of-death [accessed: 12-May-2016]
[22] Gulbinat W The role of WHO as an intergovernmental organization in the coordination of telematics in health care. [Online] Available: https://www.hon.ch/Library/papers/gul binat.html [accessed: 19-April-2016]
[23] Heart disease dataset. [Online] Available: http://archive.ics. uci.edu/ml/machine-learningdatabases/heart-disease/cleve. mod [accessed: 28-May-2016]
[24] Somy.M.S, 2Gayatri.K.S, 3 Ashwini.B "Privacy Preserving Health Data Mining" IJCST Vol. 6, ISSue 4, oCT - DeC 2015
[25] NIKUNJ DOMADIYA* and UDAI PRATAP RAO "Privacy-preserving association rule mining for horizontally partitioned healthcare data": a case study on the heart diseases Sådhanå (2018) 43:127 Indian Academy of Sciences https://doi.org/10.1007/s12046-018-0916-9Sadhana(0123456789().,-volV)
[26] Amit Datta and Marc Joye "Cryptanalysis of a Privacy-Preserving Aggregation Protocol"IEEE Transactions on Dependable and Secure Computing 14(6):693–694, 2017.
[27] Manish Shannal Atul Chaudhar/ Manish Mathuria3 Shalini Chaudhar/ Santosh Kumar5 "An Efficient Approach for Privacy Preserving in Data Mining" 978-1-4799-3140-8/14/$31.00 ©2014 IEEE
[28] Sweeney L, "Achieving k-Anonymity privacy protection uses generalization and suppression" International journal of Uncertainty, Fuzziness and Knowledge based systems, 10(5), 571-588, 2002.
[29] Gayatri Nayak, Swagatika Devi, "A survey on Privacy Preserving Data Mining: Approaches and Techniques", ternational Journal of Engineering Science and Technology, Vol. 3 No. 3, 2127-2133, 2011.