A Hybrid Approach to Association Rule Hiding through BDA (Blocking, Distortion and Anonymization) Technique

Ambrish Kumar Sharma[1], Swati Namdev[2]

Research Scholar, Oriental University, Indore (M.P.)[1]
Supervisor, Oriental University, Indore (M.P.)[2]

Abstract: In this research paper we are explained the Association Rule Mining technique of data mining, which is used in many fields mostly like medical, marketing, retailing and analyzed the human's behavior. Many strategies are used for obtaining the frequent itemset. Here the main problem of Association Rule Mining is disclosing the personal information of the particular. To rectify this issue many techniques are proposed but all have some limitations to overcome these limitations we proposed new method called BDA technique.

Keywords: BDA (Blocking, Distortion and Anonymization Technique), Association Rule Hiding, Data Mining

I. INTRODUCTION

As the traditional definition of data mining says "Data Mining" is entails discovering hidden insights from massive datasets collected from different data sources.

This process is helpful in business, medical, research for decision making and other type of information need. Data Mining is a step-by-step process which includes: -

- Extraction and transformation
- Store and manage the data
- Present analyzed data in user understandable form like graph

Association Rule Mining

In short, we can say that Association rule mining is just a way to find interesting pattern in data. It finds two things in the data set, characteristics which occur together and characteristics which are correlated.

Support: - Support shows the items features which occurs together, it contains the fractions of transaction that contain both X and Y.

$$\sigma (X+Y) / \text{total items}$$

Confidence(C): - It is basically a ratio of the no. of occurrence of data record that contains all the items in consequent and the antecedent to the number of records that contains all the items in the antecedent[6].

$$\text{Confidence}(X \Rightarrow Y) = \text{Supp}(XUY)/\text{Supp}(X)$$

Apriori

Apriori is used in mining frequent item sets. This algorithm is used basically in a database that contains huge number of transactions[7,8]. This algorithm has useful in different field like medical, healthcare, pharmacy etc. It is helpful to analyze the medical condition of patient and effects of different medicines to the patient health. The name Apriori is just as it leverages prior knowledge of frequent itemset properties. To improve the performance of the generation of frequent item set we can use the important property called Apriori property which is helpful in reducing the search space[9].

Association Rule Hiding Technique- Privacy Preserving Data Mining (PPDM) helps protect personal or sensitive information that might be exposed through regular data mining. One way to do this is by using the Association Rule Hiding (ARH) technique. This method removes sensitive rules from the database, so important private data is not revealed to data miners.

When we talk about frequent patterns, associations, and correlations, we mean items or values that show up repeatedly in a dataset.

E.g.: Diseases+ Sensitivity = Frequent Itemset

In short, frequent mining shows which items appear together in transaction or relation. Frequent mining is generation of association rules from transactional dataset. If there are two items N & N' purchased frequently then it's good to put them together.

Anonymization based PPDM

Data Anonymization is one of the most effective techniques to preserve the sensitive information about an individual to along with the consideration of data reusability factor. This technique is used to protect sensitive data by drop or modified the personal information or the data by which individual identity of the user can be hidden. This technique is done by various ways like dropping value, generalization, encryption etc.

II. BACKGROUND AND RELATED WORK

D. Dhinakaran et al. [1] A secure privacy preservation technique was developed to protect the sensitive healthcare data of patients and facilitate secure transmission to clinicians for expert consultation. The proposed model ensures both data confidentiality and diagnostic support by integrating advanced cryptographic and machine learning techniques. Healthcare data were sourced from benchmark datasets and encrypted using a novel FHE-HECC approach, which combines Fully Homomorphic Encryption with Hyperelliptic Curve Cryptography. To enhance encryption efficiency, an optimal-key was generated using the Improved Dragonfly Glowworm Optimization (IDGO) algorithm, effectively reducing memory usage and processing time.
For decryption, the same FHE-HECC scheme was applied, ensuring end-to-end security. Encrypted data were then processed for disease detection using a Weighted Deep Neural Network (W-DNN), where weights were optimized to maximize accuracy, precision, and Negative Predictive Value, while minimizing the False Positive Rate.

Sreedhar, C., Kallam et al [2], The research paper by Sreedhar et al. (2025) presents a literature review that highlights the growing concern over the security and privacy of healthcare big data, especially in the context of increasing digitalization and data breaches. The authors discuss existing encryption and authentication techniques but note that many are insufficient in handling the scalability, complexity, and real-time access demands of medical data systems. Prior studies have explored various machine learning and cryptographic methods; however, they often lack efficiency or compromise on speed and accuracy. To bridge these gaps, the paper introduces the CRHSM (Clustering-based Robust Healthcare Security Mechanism) integrated with Recursive Feature Elimination (RFE) for feature selection, offering an advanced hybrid approach. This paper sets the foundation for proposing a system that improves secure storage and retrieval of healthcare data while maintaining performance and data integrity in large-scale applications.

Tariq Emad Ali et al [3], has highlighted the rapid adoption of Internet of Things (IoT) devices in the healthcare sector has raised significant concerns regarding data security and patient privacy. While numerous techniques have been explored to enhance security, many existing approaches suffer from limitations that hinder their effectiveness in real-world applications. Recent research has focused on addressing these challenges through the integration of emerging technologies such as big data, blockchain technology, machine learning (ML), deep learning (DL), edge computing, and software-defined networking (SDN) within the Healthcare IoT (HIoT) ecosystem. These studies have provided a comprehensive evaluation of secure data access frameworks, particularly emphasizing the role of cryptographic platforms in safeguarding sensitive health information. The literature further highlights both the opportunities and challenges associated with HIoT, underscoring its transformative potential in improving patient care and clinical outcomes.

Hangyu Xie et al [4], Differential privacy (DP) has emerged as a pivotal technique in safeguarding data privacy within machine learning, particularly in sensitive fields like healthcare. It maintains data privacy by making sure that any single data entry does not significantly alter the overall analytical results. In medical applications, the adoption of DP has addressed challenges related to the confidentiality of electronic health records, diagnostic data, and teaching datasets. Central to DP are the privacy budget (ε) and noise injection mechanisms, primarily

Laplacian and Gaussian perturbations, which introduce statistical noise to preserve privacy without severely compromising data utility.

Sai Dikshit Pasham et al [5], As data volume and complexity grow, frameworks that support secure and privacy-aware data sharing in distributed computing environments are increasingly essential. Core methods such as encryption, homomorphic encryption, differential privacy, secure multi-party, and federated learning offer robust mechanisms for protecting sensitive data while enabling analytical utility. These techniques are especially impactful in high-risk domains like healthcare, finance, and smart cities.
Emerging computing paradigms like edge and cloud computing facilitate decentralized data processing, enhancing privacy control. However, challenges such as computational overhead, regulatory compliance, data heterogeneity, and scalability persist. With advancements in AI and the looming impact of quantum computing, there is a pressing need for adaptive cryptographic solutions and privacy-aware AI models. Future developments in blockchain, post-quantum cryptography, and AI-driven privacy tools are expected to significantly transform the landscape. A collaborative approach involving researchers, technologists, and policymakers is vital to build effective, privacy-centric big data ecosystems.

III. PROBLEM DEFINITION

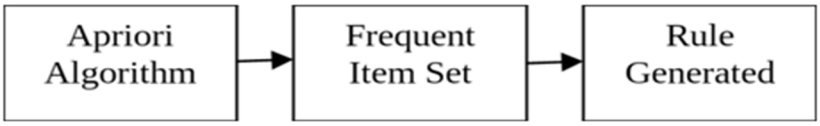
Accuracy and privacy are two different things when we achieve one lead opposite effect on other. We try to review a good number of PPDM technique. At last, we come to the conclusion that there is a no single method in PPDM algorithm that fulfills the all criteria of privacy like complexity, cost, utility but these algorithms are better on some specific criteria.
The main problem in anonymization technique they judge all the sensitive attributes at same stage or a level and apply the generalization on all the attribute. This created some issues like -

- Utility of the data
- Data loss
- Privacy issue
- Only work for centralized data

Apply these methods to minimize association rule on horizontal and vertical data set. They use Apriori algorithm which is comparatively slow on huge data set. Some limitations in association rule hiding techniques are-

- Work for distributed system only.
- Time taken by horizontal partitioned data set is twice than the vertically partitioned dataset.
- Data transfer rate in horizontal partitioned dataset is much slower than vertically partitioned dataset.

Above two points show that the association hiding technique perform better in vertically partitioned data, but in vertically partitioned data we lost integrity and Privacy.



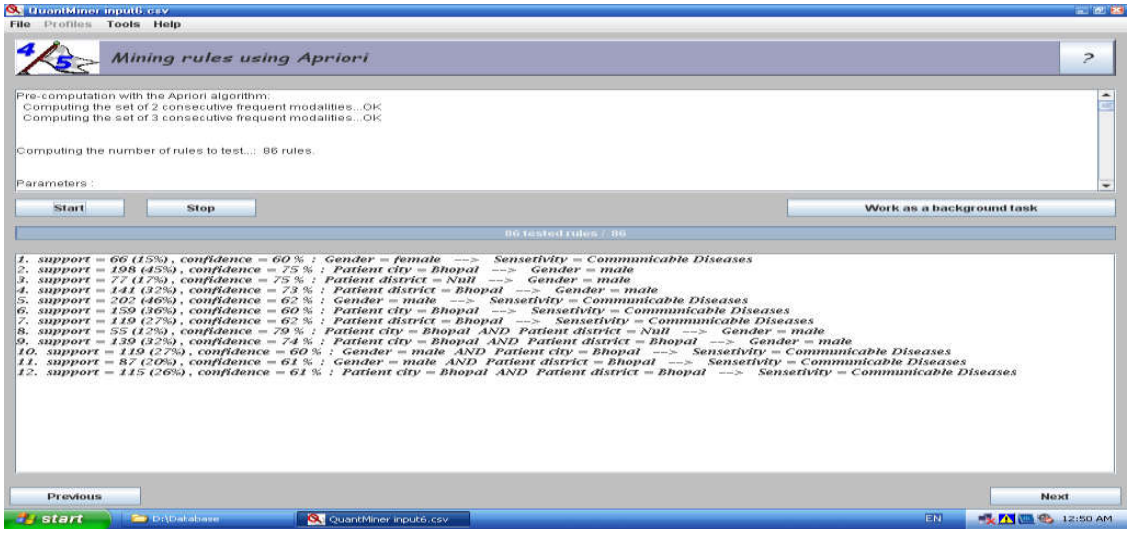


Figure 1: Represents the frequent item set using Apiori algorithm

IV. PROPOSED APPROACH

There are two different terms data noise and data distortion. It is mandatory to protect the sensitive data. Data Perturbation method is used for preserving the sensitive data. This technique is mainly Perform by adding some garbage value or by adding some unknown value [12]. The major problem with data perturbation technique to retrieve the original dataset. To control this problem new method is developed in [10] author develops a new distribution-based algorithm Vidya[13] & Rizvi[11] develop strategy for Privacy Preservation.

Blocking based technique: These Method is used to hide some sensitive data when data is shared or distribute. The individual information contains sensitive association Rule. Before share this information in public or for some research purpose sensitive information should be hide. For this reason, first it must to identify the sensitive information or rule. Second it must replace the sensitive data by some unknown value (?). Authors in [14] hide the actual values; they replace '1' by '0' or '0' by '1' or with any garbage value in a particular transaction.

Data Distortion Technique: In this method, the values of 1s are replaced with 0s, and vice versa. Rule hiding using the data distortion approach can be achieved in two ways first, by reducing the confidence of the rule, and second, by lowering its support.

Items	Confidence
Disease=> Sensitivity	4/4 =100%
Pin code, Disease=>Sensitivity	3/3=100%
Disease, Sensitivity=>Pin code	3/4=75%

Table 1: Item-sets showing the confidence %

Here we apply data distortion technique for hiding the following set of Association rules, to hide the rule, Disease=> Sensitivity process starts with the original dataset, followed by the application of Data Distortion or Data Blocking techniques to obscure sensitive association rules. For an added layer of privacy, Anonymization techniques are also employed. Data Distortion is typically divided into two categories, each targeting the concealment of sensitive rules.

Category-1: Increase support of LHS

- Look for transaction which support

Disease= Sensitivity =1

- Update the values of Disease from 1 to 0 where Disease=Sensitivity =1

Consider an example of medical dataset

Transaction	Items
N_1	Pin code, Disease, Sensitivity
N_2	Pin code, Disease, Sensitivity, City
N_3	Pin code, Disease, Sensitivity
N_4	Disease, Sensitivity
N_5	Pin code, State
N_6	Pin code, Disease

Table 2: Transactional data showing list of six transactions

In table 3 Item Occur in the Transaction Denoted by: 1
Item does not occur: 0
R_1-sensetivity, R_2-Diseases, R_3-Pincode, R_4,-City R_5-State

Transaction	Items	R_1, R_2, R_3, R_4, R_5	Size
N_1	$R_1, R_2, R_3,$	11100	3
N_2	R_1, R_2, R_3, R_4	11110	4
N_3	$R_1, R_2, R_3,$	11100	3
N_4	R_1, R_2	11000	2
N_5	R_1, R_5	10001	2
N_6	R_1, R_3	10100	2

Table 4: List of four transaction records

Category-1: Now hide LHS Rule in database D
Here we are trying to hide Disease=> Sensitivity
Step-1: Now Look for record which support Disease=Sensitivity=1
Step-2: Update the table, use value 0 for the Item disease in
all the transaction where Disease=> Sensitivity Transactions which support Disease=Sensitivity=1
Now change the value of disease from 1to 0 & also set the container.

Transaction	Items	I_1, I_2, I_3, I_4, I_5	Size
N_1	$I_1, I_2, I_3,$	10100	3
N_2	I_1, I_2, I_3, I_4	10110	4
N_3	$I_1, I_2, I_3,$	10100	3
N_4	I_1, I_2	10000	1
N_5	I_1, I_5	10001	2
N_6	I_1, I_3	10100	2

Table 5: Updated table after hiding item-2

Category-2: Now RHS Rule will be hidden for the rule Disease=> Sensitivity
To hide the rule, now change the value from 1 to 0 on RHS.
For hiding the rule, we follow two steps:
Step-1: Search for transaction support
Light Disease= Sensitivity=1
Step-2: Now change the values from 1 to 0 on RHS and also set the updated container size.

Transaction	Items	R_1, R_2, R_3, R_4, R_5	Size
N_1	$R_1, R_2, R_3,$	01100	3
N_2	R_1, R_2, R_3, R_4	01110	4
N_3	$R_1, R_2, R_3,$	01100	3
N_4	R_1, R_2	01000	1
N_5	R_1, R_5	10001	2
N_6	R_1, R_3	10100	2

Table 6: Updated Table after Hiding Item-5

Now the confidence for the rule, $\{R_5\} => \{R_1, R_2\}$, is reduced from 100% to 0%.

ii) Data blocking technique: Data blocking technique is used to maximize (Catogary-1) or minimize (Catogary-2) the support of the items by replacing 0's or 1's value by unknowns "?", so that it becomes hard for an opponent to understand the hidden value of "?"[20]

Transaction	Items	R_1, R_2, R_3, R_4, R_5	Size
N_1	$R_1, R_2, R_3,$	1?100	2
N_2	R_1, R_2, R_3, R_4	10110	3
N_3	$R_1, R_2, R_3,$	?0100	1
N_4	-	??000	?
N_5	R_1, R_5	10001	2
N_6	R_1, R_3	10100	2

Table 7: Category-1 from the data distortion table (Maximize Support)

Transaction	Items	R_1, R_2, R_3, R_4, R_5	Size
N_1	$R_1, R_2, R_3,$	?1100	2
N_2	R_1, R_2, R_3, R_4	01110	3
N_3	$R_1, R_2, R_3,$	01100	2
N_4	R_2	??000	?
N_5	R_1, R_5	10001	2
N_6	R_1, R_3	10100	2

Table 8: Category-2 from the data distortion table (Minimize Support)

Implement either Category-1 or Category-2 but not both as it will leads to opposite operations performed so far

Anonymization approaches are used to attain anonymity. These approaches include suppression, pseudonymisation, generalization and Anonymization. Generalization approaches replace specific quasi-identifier values with less specific values [16].

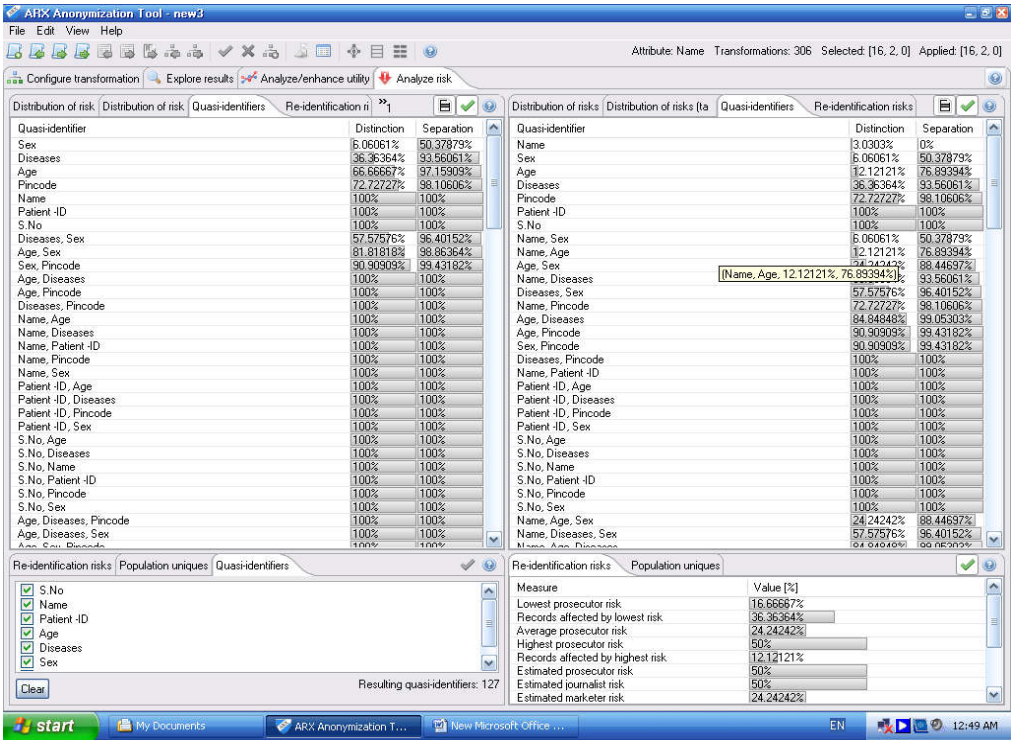


Figure 2: Represents the anonymization quasi-identifier in input and output data

V. RESULTS AND ANALYSIS

Here we use medical dataset which is available online which consist thousands of records and 103 sensitive disease it is basically structured dataset

- Firstly, we apply pre-processing to deal zero and null values.
- After Prepossessing we apply Apriori algorithm for frequent item-set and got 86 rules in which we take 12 rules those have higher support level.
- Now we use blocking technique to hide sensitive data like city

Then we apply Anonymization technique(in which we use generalization technique)

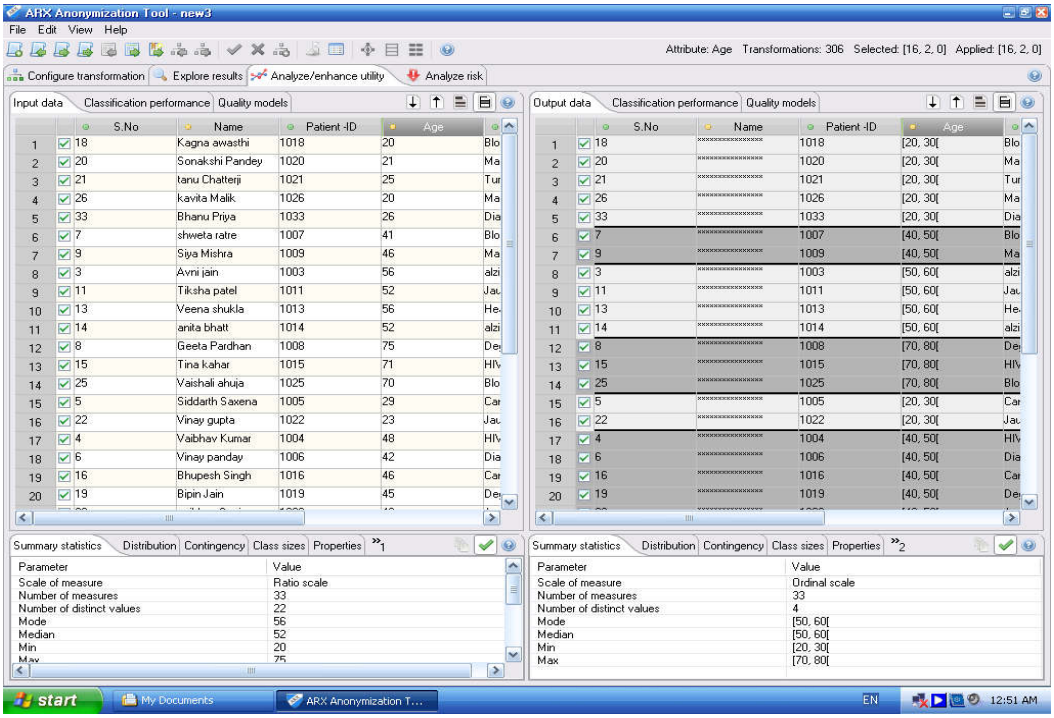


Figure 3: Input data set

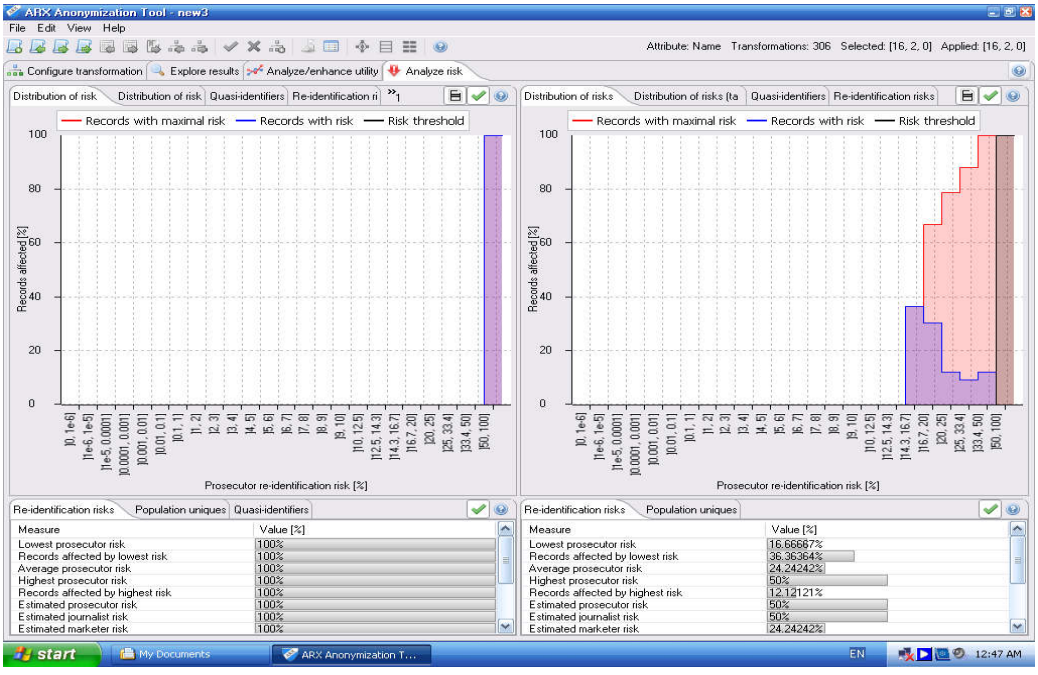


Figure 4: Risk Analysis result data set is on left hand which represents lowest risk level

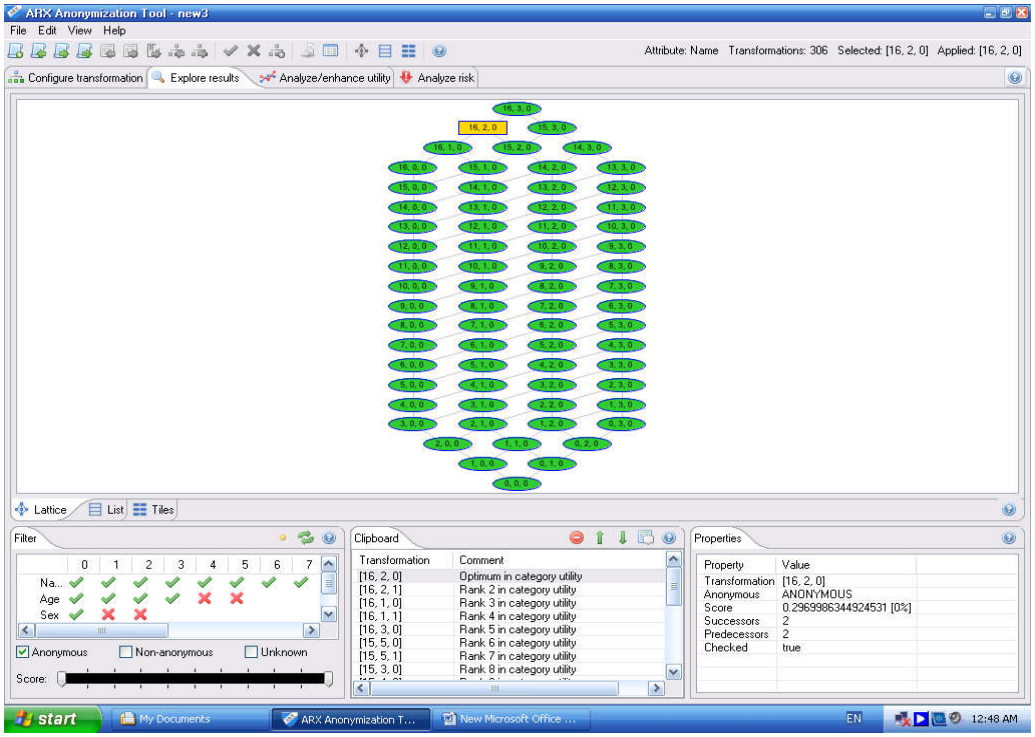


Figure 5: Result explorer

Result explorer represents the secure data in oval shape and sensitive data in rectangular which we secure by using with distortion technique. By using this hybrid BDA algorithm, we can achieve better result which depicts minimum information loss and better privacy.

VI. CONCLUSION

This paper addresses the challenges associated with privacy in data mining and proposes a novel technique for privacy preservation. It explores the process of association rule generation and introduces a hybrid approach for hiding sensitive data, referred to as the BDA technique [17], which sequentially applies Data Distortion, Data Blocking and Anonymization. Experimental results demonstrate that all association rules containing sensitive items have been effectively hidden. The proposed algorithm has been implemented and validated through numerical examples. For future work, more efficient association rule hiding methods may be developed to potentially outperform the BDA technique in terms of performance and effectiveness.

VII. REFERENCES

1. Dhinakaran, D., Srinivasan, L., Gopalakrishnan, S. and Anish, T.P., 2025. An efficient data mining technique and privacy preservation model for healthcare data using improved darts game optimizer-based weighted deep neural network and hybrid encryption. Biomedical Signal Processing and Control, 100, p.107168.
2. Sreedhar, C., Kallam, S., Ghantasala, G. P., Anthoniraj, S., Kumarganesh, S., Sagayam, K. M., Pandey, B. K., and Pandey, D."Enhancing healthcare data security using RFE and CRHSM for big data," Computers in Biology and Medicine, vol. 190, p. 110063, 2025.
3. T. E. Ali, F. I. Ali, P. Dakić, and A. D. Zoltan, "Trends, prospects, challenges, and security in the healthcare internet of things," Computing, vol. 107, no. 1, pp. 28, Jan. 2025.
4. H. Xie, Y. Zhang, Z. Zhongwen, and H. Zhou, "Privacy-preserving medical data collaborative modeling: A differential privacy enhanced federated learning framework," J. Knowl. Learn. Sci. Technol., vol. 3, no. 4, pp. 340–350, Dec. 2024.
5. S. D. Pasham, "Optimizing blockchain scalability: A distributed computing perspective," The Metascience, vol. 1, no. 1, pp. 185–214, Dec. 2023
6. Patel Shreya, Aniket Patel" Heuristic Based Approach for Privacy Preserving in Data Mining" International Journal of Scientific Research in Science, Engineering and Technology (www.ijsrset.com) Volume 4 Issue 8 ,2018.
7. Masooda Modak, Rizwana Shaikh, "Privacy Preserving Distributed Association Rule Hinding Using Concept Hirerachy", Science Direct Elsevier, Procedia computer Science(2016).

8. Krishna Kumar Tripathi, “Discrimination presvention with classification and Privacy Presevation in Data Mining”, Science Direct Elsevier, Procedia computer Science(2016).
9. Yamini M. Babnekar, Dr. Sheetal S. Dhande” Implementing the Privacy Preservation Data Mining Based on Association Sharing Technique ”, IJIRCCE Vol. 4, Issue 5, May (2016)
10. Marathe Shashank S, Manjusha Yeola, “Generating and Hiding Sensitive Association Rules,” International journal of Advance Foundation and Research in Computer (IJAFRC), Volume 2, Issue 5, May - 2015.
11. J. Han, “Mining knowledge at multiple concept levels”, Proceeding of In ACM International Conference on Information and Knowledge Management (CIKM’95), Maryland, USA, pp. 19 - 24 , November 1995.
12. J. Han and M. Kamber, “Data Mining: Concepts and Techniques”, Morgan Kaufmann Publishers, August 2000.
13. J. Han, J. Pei and Y. Yin, “Mining frequent patterns without candidate generation”, Proceeding of 2000 ACM-SIGMOD International Conference Management of Data (SIGMOD’00), pp. 1-12, 2000.
14. R. Agrawal and A. Srikant, " Privacy-preserving data mining”, in proceedings of SIGMOD00, pp. 439-450.
15. Evfimievski, A.Srikant, R.Agrawal, and Gehrke , "Privacy preserving mining of association rules", KDD02, pp. 217-228.
16. T. Jahan, G.Narsimha and C.V Guru Rao, “Data Perturbation and Features Selection in Preserving Privacy” in proceedings IEEE 2012.
17. J. Vaidya and C. Clifton, “Privacy preserving association rule mining in vertically partitioned data”, in The Eighth ACM SIGKDD International conference on Knowledge Discovery and Data Mining, Alberta, CA, July 2002, IEEE 2002.
18. S. Lohiya and L. Ragha, “Privacy Preserving in Data Mining Using Hybrid Approach”, in proceedings of 2012 Fourth International Conference on Computational Intelligence and Communication Networks, IEEE 2012.
19. A.Parmar, U. P. Rao, D. R. Patel, “Blocking based approach for classification Rule hiding to Preserve the Privacy in Database” , in proceedings of International Symposium on Computer Science and Society, IEEE 2011.
20. Apoorva Joshi,, Pratima Gautam” An optimized algorithm for association rule hiding technique using Hybrid Approach” International Journal of Computer Sciences and Engineering, Vol.-7, Issue-1, Jan 2019