A Transformative Framework For Multi-Dimensional Privacy-Preserving Data Mining: Mapping, Adaptability, And Fidelity

Mangal Sonawane[1], Lata Ragha[2]
[1]Dept of Computer Engineering
[2]Professor and HoD, Dept of Computer Engineering
[1]Terna Engineering College, Nerul, Navi-Mumbai
[2]Fr. C. Rodrigues Institute of Technology, Vashi, Navi-Mumbai

Abstract- *In response to the escalating challenge of preserving privacy amidst the ever-expanding accumulation of personal data in business applications, this project introduces a pioneering framework for multi-dimensional privacy-preserving data mining. Diverging from conventional methodologies, it takes a transformative approach by eliminating the reliance on problem-specific algorithms. Instead, the project proposes a groundbreaking strategy: mapping the original dataset into a novel, anonymized counterpart that intricately replicates the nuances of the data, including inter-dimensional correlations.Unlike prior methods, the emphasis of this methodology lies in its adaptability and accuracy. By eschewing problem-specific algorithms, it offers a versatile solution applicable across diverse business contexts. The innovative framework not only addresses immediate concerns related to data privacy but positions itself as a catalyst for future advancements in the field.The adaptability of the methodology transcends the limitations of conventional approaches, providing a scalable solution across diverse business contexts. Moreover, its commitment to accuracy ensures that the transformed dataset retains a high fidelity to the original, preserving individual privacy and nuanced interconnections within the data.This project stands at the nexus of technological advancement and ethical considerations, offering a forward-looking perspective on privacy preservation in the digital age. By pushing the boundaries of conventional data mining techniques, it seeks to establish a new paradigm that harmonizes the imperatives of data-driven business applications with the paramount importance of protecting individuals' privacy rights in an increasingly interconnected and data-rich environment.*

Keywords *privacy, data mining, privacy-preserving data mining,K-anonymity.*

I. INTRODUCTION

In the dynamic landscape of data mining, Privacy Preserving Data Mining (PPDM) represents a multidimensional strategy crafted to counter the escalating concerns related to the confidentiality and security of sensitive information within extensive databases. The evolution of the data mining field and the increasing sophistication of algorithms heighten the risks associated with privacy breaches, necessitating a proactive and comprehensive approach to safeguarding individual privacy. The core tenet of the PPDM approach lies in reconciling the extraction of valuable insights through data mining with the imperative to protect the privacy of individuals whose information resides within these databases. One key facet of this approach involves the development and implementation of advanced encryption and anonymization techniques. By encrypting sensitive data or anonymizing it in a manner that preserves its utility for analysis while mitigating the risk of identification, PPDM endeavors to strike a delicate balance.

Cryptographic methods, such as homomorphic encryption, enable computations on encrypted data without decryption, ensuring that sensitive information remains confidential throughout the analysis process. Conversely, anonymization techniques focus on obscuring personally identifiable information while retaining the overall statistical patterns and trends in the data. Furthermore, the PPDM approach incorporates access control mechanisms and secure data-sharing protocols. Access control mechanisms define and regulate who can access specific data within the database, minimizing the risk of unauthorized users obtaining sensitive information. Secure data-sharing protocols facilitate the controlled sharing of information between different entities, ensuring that only relevant and non-sensitive insights are disclosed.

In addition to technological safeguards, the PPDM approach recognizes the significance of ethical guidelines and legal frameworks. It seeks alignment with existing privacy regulations and norms, fostering a responsible and transparent use of data mining techniques. The integration of legal and ethical considerations into the technical fabric of privacy-preserving strategies establishes a holistic approach beyond mere technological solutions. Essentially, the PPDM approach represents a dynamic and adaptive framework that evolves alongside advancements in data mining techniques. It addresses the immediate challenges posed by extracting insights from massive databases and anticipates and mitigates potential risks to individual privacy. This approach aims to pave the way for a responsible and privacy-conscious era in data mining through a synergistic combination of encryption, anonymization, access controls, and ethical considerations.

II. LITERATURE SURVEY

Hewage et al. analyze the strengths and disadvantages of present input privacy-preserving data mining (PPDM) and privacy-preserving data stream mining (PPDSM) techniques in resolving the trade-off between data mining precision and privacy. A systematic literature review based on 104 primary studies from reputable databases categorizes PPDM methods into perturbation, non-perturbation, secure multi-party computation, and combinations. PPDSM techniques are classified into perturbation, non-perturbation, and others. Challenges unique to PPDSM, such as high volume and speed, are noted. While numerous studies acknowledge the accuracy-privacy trade-off, solutions, particularly in PPDSM, still need exploration. Most PPDM methods find classification applications, clustering applications, and association rule mining. [1]

Data analysis, validation, and publication rely on the safe and secure transfer of private information, making privacy-preserving data mining (PPDM) necessary. The data mining industry has developed several algorithms specifically to protect user privacy. This article provides a comprehensive overview of algorithms for privacy-preserving data mining, discusses the many datasets utilized in the study, and evaluates the methods concerning several criteria. The strengths and weaknesses of each study's findings are noted in the survey. Future PPDM studies can use this poll as a reference when deciding on research methods. [2]

By meticulously organizing the literature into subcategories, Yousra et al. provide a bird's-eye view of a new perspective and systematically analyze a long list of publications. The basic concepts of existing privacy-preserving data mining technologies are introduced, together with their benefits and drawbacks. It is possible to classify the many existing privacy-preserving data mining approaches using terms like "distortion," "association rule," "hide association rule," "taxonomy," "clustering," "associative classification," "outsourced data mining," "distributed," and "k-anonymity," each highlighting the respective method's benefits and drawbacks. The evolution, research obstacles, future trends, gaps, and shortcomings are all exposed by this thorough examination. Additional substantial changes are required for more effective privacy protection and preservation. [3]

Safeguarding healthcare data privacy is crucial for ensuring accurate records and fostering confidence among data custodians for mining purposes. While association rule mining is widely applied in healthcare, the focus has predominantly been on positive associations, neglecting the negative implications of certain diagnostic techniques. Negative association rules, particularly in bridging diseases and drugs, can offer valuable insights, especially concerning physicians and social organizations. The challenge lies in conducting healthcare data mining that protects patient identity, given the sensitivity of the information. This study explores metaheuristic-based data sanitization, utilizing the Tabu-genetic algorithm to optimize the selection of item sets for sanitization, minimizing changes made to the primary data source. The suggested privacy-preserving data mining approach outperforms state-of-the-art algorithms on benchmark healthcare datasets in terms of "Hiding Failure," "Artificial Rule Generation," and "Lost Rules," as shown by the experiments. [4]

Web-scale data mining tools and systems include things like online search, recommender systems, crowdsourced platforms, and analytics programs; in light of recent data breaches and new legislation like GDPR, preserving users' privacy is more important than ever and has been brought back into the spotlight. The first part of this tutorial provides a historical context for differential privacy definitions and approaches by discussing the major privacy breaches over the past two decades and the lessons learned from them. Then, zero in on how companies like Apple, Google, and Microsoft have put privacy-protecting data mining techniques to use in real-world contexts by showcasing real-world case studies like their differential privacy deployments for iOS/macOS, RAPPOR, LinkedIn salary data, and Windows telemetry. In closing, Kenthapadi et al. draw on professional experiences to identify some outstanding issues and concerns facing the data mining and machine learning community. [5]

Knowledge-based applications now absolutely require a high level of private space. Integrating people's right to privacy in a meaningful way is crucial for successful data mining. The healthcare, pharmaceutical, research, and security service provider industries, to mention a few, all have a vested interest in privacy-preserving data mining. For comprehensive answers, there are compromises between data security and privacy. Shah et al. conduct a comprehensive literature review of PPDM methods, classify those methods, and then provide tentative implications for when each category of method should be employed. [6]

This article surveys data mining methods appropriate for massive datasets used in PPDM. This is the introduction backdrop for the next complete description of the most prevalent PPDM approaches. Data collection, publication, dissemination, and output are just some of the many detailed data lifecycle steps discussed in detail in this PPDM process. Metrics for gauging the privacy, quality, and complexity of the suggested PPDM approaches are then analyzed to complete the evaluation of these methods. After that, Mendes et al. look at the aforementioned PPDM methods from the perspective of their use in various real-world contexts and the reasoning for the selection of those methods. Finally, some unanswered questions and potential avenues for further study are outlined. [7]

"Data mining" refers to gleaning useful information from massive data stores. Data mining and security both benefit greatly from privacy-preserving data mining. Data set complementation is the current privacy preservation solution, although it is ineffective if all datasets are exposed due to the generic nature of the dataset reconstruction algorithm. The suggested method safeguards personal information by first transforming real-world sample datasets into a collection of fictitious ones and then using cryptographic privacy protection for sensitive values. RSA is used as the method of encryption in this case. This approach protects users' anonymity while also boosting precision. The Nave Bayesian classification technique is used in this paper, along with a new privacy-preserving approach. [8]

Protecting individuals' anonymity requires that datasets have as much of their original quality as feasible. Information loss after privacy has been protected is difficult to define and quantify. Methods have been created to evaluate the generalizability, suppressiveness, and randomization of a dataset's information quality. Different metrics focus on the data or the results of data mining operations like Clustering and Classification. A variety of information metrics, as well as their applications and limits, are discussed in this survey. [9]

The need for privacy in knowledge-based applications has grown significantly. Integrating people's right to privacy in a meaningful way is crucial for successful data mining. Industries as diverse as healthcare, pharmaceuticals, academia, and security services all stand to benefit from privacy-preserving data mining. Perturbation, Secure Sum Computations, and Cryptographic-based techniques comprise the primary Privacy Preserving Data Mining (PPDM) methods classifications. For comprehensive answers, there are compromises between data security and privacy. The paper's authors conduct a comprehensive literature review of PPDM methods, classify those methods, and then provide tentative implications for when each category of method should be employed. [10]

III. METHODOLOGY

A multi-dimensional data mining platform ensuring user anonymity has been proposed and implemented. A novel and adaptable data mining method safeguarding individual privacy has been created, which involves transferring the old dataset to a new, anonymized one without the need for additional problem-specific algorithms. The structure and correlations between various dimensions in the anonymized data remain virtually unchanged from the original dataset. During this process, information is compressed into several groups of uniform size, each with its own set of statistics tracking various data types. These categories reliably store statistics related to averages and relationships between different dimensions, rendering individual records indistinguishable within a set. The degree of indiscernibility is defined as the least size, k, of any given group, where a higher degree implies a more discrete system. Condensing multiple records into a single statistical group entity results in increased information loss.

A condensation method is provided, constructing restricted clusters in the dataset and generating pseudo-data based on their statistical properties. Condensing cluster statistics produces pseudo-data, hence the term "condensation." Cluster sizes are determined to maintain k-anonymity, defining the limitations of the clusters. This approach, compared to the perturbation model, effectively preserves privacy and offers various advantages. Importantly, this method utilizes pseudo-data instead of alterations to the original data, enhancing privacy preservation. Data mining methods can be directly applied to pseudo-data without rewriting. The challenge lies in adapting data mining algorithms to effectively handle incomplete or partially certain data resulting from generalizations or suppressions, and it accommodates real-time data changes.

Condensation, as a technique, aims to preserve the confidentiality of personal information without compromising the ability to mine data for actionable insights and patterns. In the context of privacy-preserving data mining, businesses and applications collect extensive personal data, but individuals often hesitate to share sensitive information to ensure its privacy.

Traditional privacy-preserving data mining approaches commonly employ perturbation techniques, which slightly alter data to hide individual details while allowing general analysis. However, these methods treat each data item or dimension independently, overlooking linkages between distinct parts of the data. In contrast, the condensation strategy provides an alternative for maintaining privacy during data mining without developing new algorithms for each task. Instead, it transforms source information into a masked copy, preserving linkages, correlations, and original traits and patterns in the anonymized data, thus protecting individuals' privacy while enabling valuable insights from anonymized data.

A. Condensation Approach

- Condensation methods are designed to restrict the exposure of the data mining process to solely the most pertinent information. To maintain the primary patterns and characteristics while diminishing the risk of divulging sensitive information, it is customary to select a representative subset of the data or to summarize the data.
- Data generalization is a prevalent tool within the condensation method. Specifics are exchanged for a broader, less detailed category during generalization. For instance, an age range might substitute the actual age of an individual, or a more extensive geographical region could supplant the precise location.
- Another applicable method is data suppression, which involves removing sensitive information or data attributes that could be exploited to single out particular individuals. This could involve concealing unusual details or eliminating outliers.
- To heighten the challenge for potential attackers attempting to discern specific patterns or sensitive information, data may undergo "randomization," entailing the introduction of noise or randomness into the data. Randomization methods can be beneficial for both numerical and categorical information.
- For confidentiality purposes, data values may undergo "perturbation," involving intentional and measured alterations. The introduction of noise to the numerical values is a potential method.
- Aggregated data involves grouping data into more generalized categories. For instance, aggregated data may portray total sales by region or month rather than individual sales records.
- The representation of information within a hierarchical framework can aid in safeguarding personal details. Such structures enable users to access data at varying levels of detail based on their access permissions and requirements.
- Addressing the challenge of utility-precision is crucial in the condensation technique. Striking a balance between data privacy and accessibility is a significant hurdle. The application of more privacy measures may result in a reduction in the data's utility. Hence, data miners and organizations must thoughtfully evaluate the trade-off between these two factors.
- Privacy-preserving algorithms, in certain instances, execute data analysis while concealing sensitive data. Techniques like differential privacy can be employed to ensure that mining results do not disclose individual-level information.

The condensation approach is integral to a comprehensive framework that navigates the balance between safeguarding individual privacy and conducting analytical examinations to extract insights from data. The characteristics of the data, confidentiality requirements, and project objectives collectively influence the selection of data mining methods. The condensation method empowers data analysts and researchers to handle anonymized data with the same analytical treatment as the original data, preserving the confidentiality of essential information and enabling valuable data analysis and mining processes.

B. Input and Output of Condensation Approach

The following describes the inputs and outputs of the "Condensation Approach to Privacy Preserving Data Mining.":

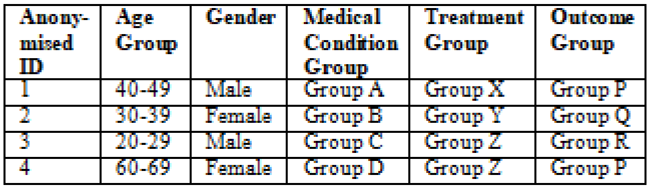
Input:A multi-dimensional dataset, encompassing private information, undergoes processing through the Condensation Method. The dataset incorporates various features or dimensions, including age, gender, medical problems, treatments, and outcomes. Each record in the dataset serves as a representation of an individual, and the identification of individuals is feasible based on the amalgamation of attributes within each record.

Table 1. Example Dataset (Before Anonymization):

Patient ID	Age	Gender	Medical Condition	Treatment	Outcome
1	45	Male	Diabetes	Insulin	Improved
2	32	Female	Asthma	Inhaler	Stable
3	28	Male	Hypertension	Medication	Worsened
4	62	Female	Arthritis	Medication	Improved

Output: The Condensation Method involves the transformation of the original dataset to generate a novel, anonymous, and private dataset. The resultant dataset maintains identical dimensions to the input dataset, yet features modified values and records. These modifications serve to safeguard the identities and sensitive information of individuals. The primary objective is to obscure the discernibility of patients' identities, all the while preserving the data's utility for data mining applications.

Table 2. Example Dataset (After Anonymization):



Anony-mised ID	Age Group	Gender	Medical Condition Group	Treatment Group	Outcome Group
1	40-49	Male	Group A	Group X	Group P
2	30-39	Female	Group B	Group Y	Group Q
3	20-29	Male	Group C	Group Z	Group R
4	60-69	Female	Group D	Group Z	Group P

In the anonymized dataset, original patient IDs have been substituted with anonymized IDs. Age values have been organized into age groups (e.g., 40-49), while gender information remains unaltered. Medical conditions are categorized (e.g., Group A, Group B, etc.), treatment information is grouped (e.g., Group X, Group Y, etc.), and outcomes are also categorized (e.g., Group P, Group Q, etc.).

Explanation:

- The initial entry in the anonymized dataset, identified by Anonymized ID 1, signifies a cohort of patients aged 40-49, of male gender, with medical conditions categorized under Group A. This cohort received treatment from Group X and reported outcomes within Group P.
- The second entry, Anonymized ID 2, characterizes a cohort of patients aged 30-39, of female gender, with medical conditions falling under Group B. This cohort received treatment from Group Y and reported outcomes within Group Q.
- The third entry, Anonymized ID 3, depicts a cohort of patients aged 20-29, of male gender, with medical conditions categorized under Group C. This cohort received treatment from Group Z and reported outcomes within Group R.
- The fourth entry, Anonymized ID 4, portrays a cohort of patients aged 60-69, of female gender, with medical conditions falling under Group D. This cohort received treatment from Group Z and reported outcomes within Group P.

Applying the Condensation Approach has resulted in the transformation of the dataset to safeguard the privacy of individual patients. Consequently, each record now signifies a cohort of patients sharing similar characteristics, rendering it challenging to discern specific individuals. The anonymized dataset thus generated proves suitable for conducting data mining tasks without compromising patient privacy, all the while preserving crucial correlations and patterns inherent in the data.

Enhancements:

1. Clustering-based Anonymization
2. Homomorphic-based Anonymization
3. T-Closeness Privacy Model-based Anonymization
4. Condensation-based Anonymization

The enhancement integrates Clustering-based Anonymization, Homomorphic-based Anonymization, T-Closeness Privacy Model-based Anonymization, and Condensation-based Anonymization, collectively referred to as CHTC. This unified approach is designed to improve data privacy and utility in anonymization.

C. Overview of the proposed concept:

- Clustering-based Anonymization: Initially, the dataset undergoes partitioning into clusters based on similarity metrics or clustering algorithms. Each cluster signifies a group of records sharing similar characteristics. Anonymizing data within these clusters preserves individuals' privacy by concealing individual identities within groups.
- Homomorphic-based Anonymization: For each cluster, encryption of sensitive attributes occurs using homomorphic encryption techniques, enabling computations on encrypted data without decryption. This step ensures that data processors or analysts can process the data without compromising individual-level information.
- T-Closeness Privacy Model-based Anonymization: To achieve a higher level of privacy protection, the T-Closeness Privacy Model is independently applied to each cluster. This model ensures that the distribution of sensitive attribute values within each cluster is indistinguishable from the distribution in the entire dataset, thereby preventing attribute disclosure attacks.
- Condensation-based Anonymization: In the final step, condensation techniques are employed to reduce the dimensionality of the data, aggregating or summarizing similar records within each cluster. This process helps mitigate re-identification risks, as fewer identifying attributes are retained in the final anonymized dataset.

D. *CHTC Approach*

The CHTC approach adeptly amalgamates the strengths of four anonymization techniques to offer a resilient and all-encompassing privacy protection solution. The articulated concept adeptly harmonizes privacy preservation and data utility through the application of clustering, homomorphic encryption, T-closeness privacy, and condensation, rendering it well-suited for diverse data sharing and analysis scenarios.

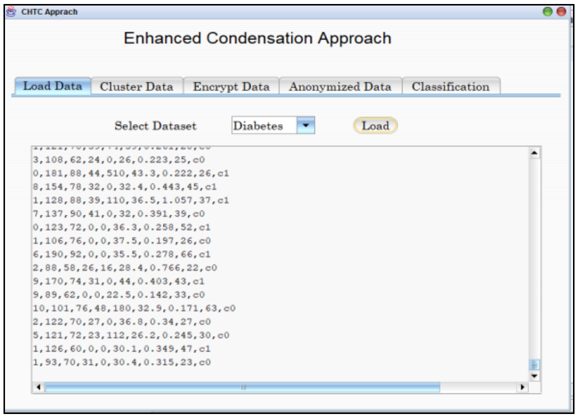
- Novelty: The uniqueness of the CHTC approach is evident in its amalgamation of multiple privacy-enhancing techniques, ensuring a harmonious equilibrium between data privacy preservation and utility during the anonymization process. The key facets that distinguish this approach are as follows:
- Comprehensive Approach: The CHTC approach harmonizes four discrete anonymization techniques – Clustering-based Anonymization, Homomorphic-based Anonymization, T-Closeness Privacy Model-based Anonymization, and Condensation-based Anonymization. This integration is groundbreaking as it addresses diverse facets of data privacy, presenting a holistic solution rather than relying on a singular technique.
- Clustering-based Anonymization: While clustering-based anonymization is not entirely groundbreaking, its incorporation as the initial step in the process enhances value by grouping akin records. This aids in privacy preservation by concealing individual identities within clusters, thereby complicating the linkage of specific data points to individuals.
- Homomorphic-based Anonymization: The use of homomorphic encryption to safeguard sensitive attributes within clusters is a novel augmentation. Homomorphic encryption allows computations on encrypted data, fortifying security and averting data processors from accessing individual-level information.
- T-Closeness Privacy Model: The application of the T-Closeness Privacy Model to each cluster independently constitutes an innovative stride. T-closeness ensures the similarity of the distribution of sensitive attribute values within each cluster to the overall dataset distribution, thwarting attribute disclosure attacks and ensuring heightened privacy protection.
- Condensation-based Anonymization: The employment of condensation techniques in the concluding step is a novel and invaluable addition. By diminishing the dimensionality of the data and summarizing akin records within clusters, the risk of re-identification is minimized, thereby ensuring more robust privacy protection.
- Balancing Privacy and Utility: The CHTC approach underscores the importance of balancing data privacy preservation and data utility. While certain anonymization techniques may be highly effective in ensuring privacy, they can lead to a substantial loss of data utility. CHTC endeavors to strike an optimal balance between these two aspects, rendering it suitable for diverse data-sharing and analysis scenarios.
- Integration of Techniques: While individual techniques may have been employed independently in previous research, the distinctive feature of CHTC lies in the amalgamation of clustering, homomorphic encryption, T-Closeness, and condensation into a singular approach. The synergistic fusion of these techniques fortifies overall privacy protection and utility in the anonymized dataset.

The novelty inherent in the CHTC approach resides in the integration of various state-of-the-art anonymization techniques, facilitating comprehensive and balanced privacy preservation without compromising data utility. This holistic methodology offers a robust and innovative solution for privacy-conscious data sharing and analysis.

IV. RESULT AND DISCUSSION

The results of the proposed system are presented in this section

Step 1: Load Dataset

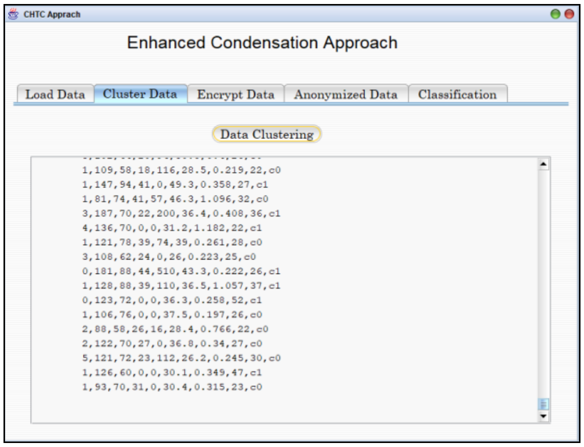


```
3,108,62,24,0,26,0.223,25,c0
0,181,88,44,510,43.3,0.222,26,c1
8,154,78,32,0,32.4,0.443,45,c1
1,128,88,39,110,36.5,1.057,37,c1
7,137,90,41,0,32,0.391,39,c0
0,123,72,0,0,36.3,0.258,52,c1
1,106,76,0,0,37.5,0.197,26,c0
6,190,92,0,0,35.5,0.278,66,c1
2,88,58,26,16,28.4,0.766,22,c0
9,170,74,31,0,44,0.403,43,c1
9,89,62,0,0,22.5,0.142,33,c0
10,101,76,48,180,32.9,0.171,63,c0
2,122,70,27,0,36.8,0.34,27,c0
5,121,72,23,112,26.2,0.245,30,c0
1,126,60,0,0,30.1,0.349,47,c1
1,93,70,31,0,30.4,0.315,23,c0
```

Fig. 1. Load dataset

In Figure 1, the initial step entails loading the dataset and activating the refined condensation approach. This methodology refines condensation, summarizing data while retaining crucial information. The enhanced variant incorporates optimizations, potentially integrating advanced algorithms or techniques to enhance the condensation outcome. The outcome is a more efficient and streamlined representation of the dataset, establishing the foundation for subsequent steps in data analysis or mining.

Step 2: Cluster Data



```
1,109,58,18,116,28.5,0.219,22,c0
1,147,94,41,0,49.3,0.358,27,c1
1,81,74,41,57,46.3,1.096,32,c0
3,187,70,22,200,36.4,0.408,36,c1
4,136,70,0,0,31.2,1.182,22,c1
1,121,78,39,74,39,0.261,28,c0
3,108,62,24,0,26,0.223,25,c0
0,181,88,44,510,43.3,0.222,26,c1
1,128,88,39,110,36.5,1.057,37,c1
0,123,72,0,0,36.3,0.258,52,c1
1,106,76,0,0,37.5,0.197,26,c0
2,88,58,26,16,28.4,0.766,22,c0
2,122,70,27,0,36.8,0.34,27,c0
5,121,72,23,112,26.2,0.245,30,c0
1,126,60,0,0,30.1,0.349,47,c1
1,93,70,31,0,30.4,0.315,23,c0
```

Fig. 2. Cluster Data

Utilizing the K-Nearest Neighbors technique, a cluster of densely packed datasets is generated. The algorithm's specifics are outlined below.

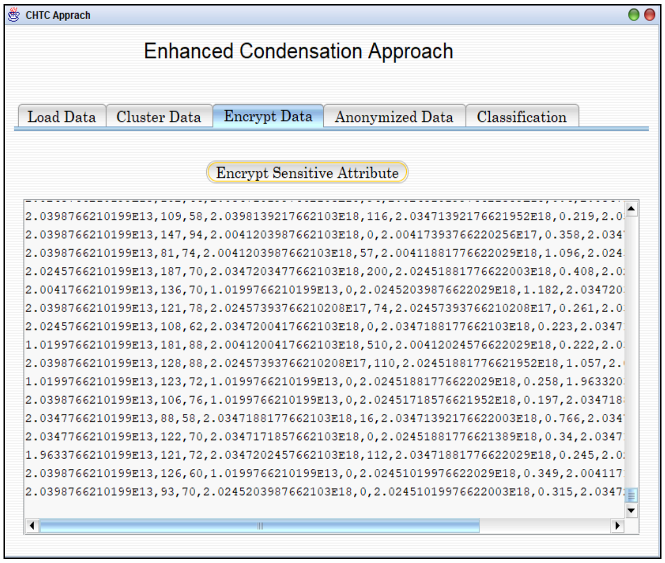
A distance measure, such as the Euclidean distance, is employed by the K-NN method to identify data points closest to a given one. The category or value of a data point is subsequently determined by the majority vote or average of its K neighbors. The K-Nearest Neighbor (KNN) algorithm finds widespread application in classification and regression problems. In predicting outcomes, the algorithm gauges the distance between the input data point and each sample in the training set, typically utilizing the Euclidean distance.Subsequently, the algorithm calculates the distances between the input data point and its K nearest neighbors. In classification scenarios, the predicted label for the input data point is based on the most frequent class label among the K neighboring points. For regression, it computes the mean or weighted mean of the target values of the K neighbors to predict the value of the input data point.

We implement a KNN model by following the below steps:

1. Data loading
2. Start with a blank slate for k.
3. Iterate from 1 to the total number of training data points to obtain the predicted class.
4. Determine the gap between the test data and each row in the dataset used for training.
5. Since Euclidean distance is the standard here, we'll utilise that. Manhattan distance, Minkowski distance, Chebyshev distance, cosine distance, etc, are other examples of distance functions or metrics. The hamming distance method can be utilised in the presence of categorical factors.
6. Using the distance data, sort the tallied distances into ascending order.
7. Collect the first k rows of the sorted array.
8. Determine the most common grouping of these rows.
9. It is expected that you will return to the projected class.

In coding, the WEKA dotnet tool is employed to generate clusters. The concise overview of WEKA is as follows. The weka.clusters.AbstractClusterer class constitutes a component of the Weka machine learning framework. Weka is utilized for data analysis as an open-source data mining, machine learning program, and modeling tool. The Abstract Clusterer class serves as the foundational class for diverse clustering algorithms within Weka.Within Weka, clustering represents unsupervised learning with the objective of grouping similar instances (data points) into clusters. The Abstract Clusterer class delineates the fundamental structure and functionality that clustering algorithms must conform to.

Step 3: Data Encryption



```
2.0398766210199E13,109,58,2.0398139217662103E18,116,2.03471392176621952E18,0.219,2.0
2.0398766210199E13,147,94,2.0041203987662103E18,0,2.00417393766220256E17,0.358,2.034
2.0398766210199E13,81,74,2.0041203987662103E18,57,2.00411881776622029E18,1.096,2.024
2.0245766210199E13,187,70,2.0347203477662103E18,200,2.02451881776622003E18,0.408,2.0
2.0041766210199E13,136,70,1.0199766210199E13,0,2.02452039876622029E18,1.182,2.034720
2.0398766210199E13,121,78,2.02457393766210208E17,74,2.02457393766210208E17,0.261,2.0
2.0245766210199E13,108,62,2.0347200417662103E18,0,2.0347188177662103E18,0.223,2.0347
1.0199766210199E13,181,88,2.0041200417662103E18,510,2.00412024576622029E18,0.222,2.0
2.0398766210199E13,128,88,2.02457393766210208E17,110,2.02451881776621952E18,1.057,2.
1.0199766210199E13,123,72,1.0199766210199E13,0,2.02451881776622029E18,0.258,1.963320
2.0398766210199E13,106,76,1.0199766210199E13,0,2.02451718576621952E18,0.197,2.034718
2.0347766210199E13,88,58,2.0347188177662103E18,16,2.03471392176622003E18,0.766,2.034
2.0347766210199E13,122,70,2.0347171857662103E18,0,2.02451881776621389E18,0.34,2.0347
1.9633766210199E13,121,72,2.0347202457662103E18,112,2.03471881776622029E18,0.245,2.0
2.0398766210199E13,126,60,1.0199766210199E13,0,2.02451019976622029E18,0.349,2.004117
2.0398766210199E13,93,70,2.0245203987662103E18,0,2.02451019976622003E18,0.315,2.0347
```

Fig. 3. Data Encryption

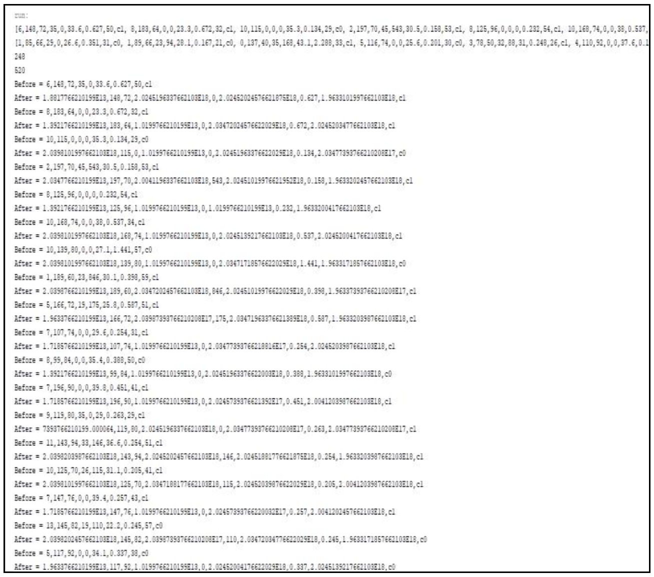
Utilizing homomorphic encryption allows the execution of computations on encrypted data without the necessity of decoding it initially. Data mining proves highly advantageous in safeguarding individual privacy, as it facilitates the analysis of data without revealing sensitive information.

- Data Encryption: Initially, the data owner encrypts their sensitive data using a homomorphic encryption scheme. This ensures that the data remains confidential even while it's being processed.
- Outsourced Computation: The encrypted data is then sent to a third party, such as a cloud service provider, for processing. The third party can perform computations on this encrypted data without decrypting it.

- Computation on Encrypted Data: Homomorphically compatible implementations of data mining algorithms carry out computations on the encrypted data.These operations could include addition, multiplication, comparison, etc.
- Result Extraction: After the computations are completed, the third party sends back the results in encrypted form.
- Decryption and Interpretation: The data owner, who possesses the decryption key, can then decrypt and interpret the results.

Using homomorphic encryption, the data owner maintains control over their sensitive information throughout the process. Despite performing computations, the third party never gains access to the raw data.

Step 4: Data Encryption

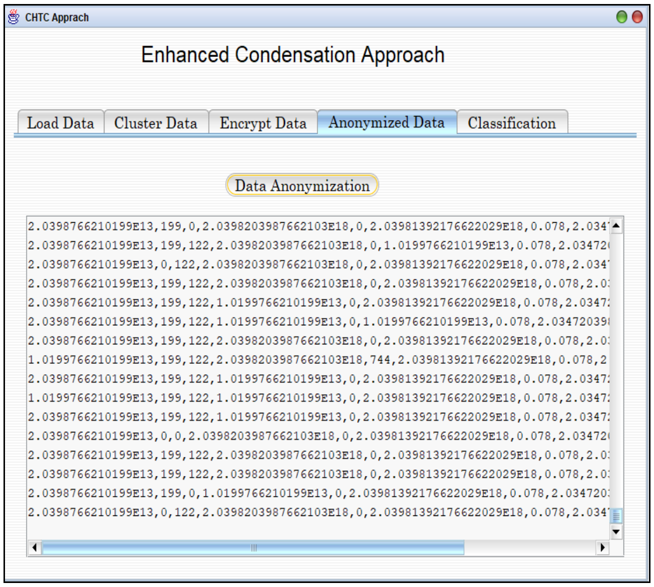


```
run:
[6,148,72,35,0,33.6,0.627,50,c1, 8,183,64,0,0,23.3,0.672,32,c1, 10,115,0,0,0,35.3,0.134,29,c0, 2,197,70,45,543,30.5,0.158,53,c1, 8,125,96,0,0,0,0.232,54,c1, 10,168,74,0,0,38,0.537,
[1,85,66,29,0,26.6,0.351,31,c0, 1,89,66,23,94,28.1,0.167,21,c0, 0,137,40,35,168,43.1,2.288,33,c1, 5,116,74,0,0,25.6,0.201,30,c0, 3,78,50,32,88,31,0.248,26,c1, 4,110,92,0,0,37.6,0.1
248
520
Before = 6,148,72,35,0,33.6,0.627,50,c1
After = 1.8817766210199E13,148,72,2.0245196337662103E18,0,2.02452024576621875E18,0.627,1.9633101997662103E18,c1
Before = 8,183,64,0,0,23.3,0.672,32,c1
After = 1.3921766210199E13,183,64,1.0199766210199E13,0,2.0347202457662202 9E18,0.672,2.0245203477662103E18,c1
Before = 10,115,0,0,0,35.3,0.134,29,c0
After = 2.0398101997662103E18,115,0,1.0199766210199E13,0,2.0245196337662202 9E18,0.134,2.0347739376621020 8E17,c0
Before = 2,197,70,45,543,30.5,0.158,53,c1
After = 2.0347766210199E13,197,70,2.0041196337662103E18,543,2.0245101997662195 2E18,0.158,1.9633202457662103E18,c1
Before = 8,125,96,0,0,0,0.232,54,c1
After = 1.3921766210199E13,125,96,1.0199766210199E13,0,1.0199766210199E13,0.232,1.9633200417662103E18,c1
Before = 10,168,74,0,0,38,0.537,34,c1
After = 2.0398101997662103E18,168,74,1.0199766210199E13,0,2.0245139217662103E18,0.537,2.0245200417662103E18,c1
Before = 10,139,80,0,0,27.1,1.441,57,c0
After = 2.0398101997662103E18,139,80,1.0199766210199E13,0,2.0347171857662202 9E18,1.441,1.9633171857662103E18,c0
Before = 1,189,60,23,846,30.1,0.398,59,c1
After = 2.0398766210199E13,189,60,2.0347202457662103E18,846,2.0245101997662202 9E18,0.398,1.9633739376621020 8E17,c1
Before = 5,166,72,19,175,25.8,0.587,51,c1
After = 1.9633766210199E13,166,72,2.0398739376621020 8E17,175,2.0347196337662138 9E18,0.587,1.9633203987662103E18,c1
Before = 7,107,74,0,0,29.6,0.254,31,c1
After = 1.7185766210199E13,107,74,1.0199766210199E13,0,2.0347739376621881 6E17,0.254,2.0245203987662103E18,c1
Before = 8,99,84,0,0,35.4,0.388,50,c0
After = 1.3921766210199E13,99,84,1.0199766210199E13,0,2.0245196337662200 3E18,0.388,1.9633101997662103E18,c0
Before = 7,196,90,0,0,39.8,0.451,41,c1
After = 1.7185766210199E13,196,90,1.0199766210199E13,0,2.0245739376621392E17,0.451,2.0041203987662103E18,c1
Before = 9,119,80,35,0,29,0.263,29,c1
After = 7393766210199.000064,119,80,2.0245196337662103E18,0,2.0347739376621020 8E17,0.263,2.0347739376621020 8E17,c1
Before = 11,143,94,33,146,36.6,0.254,51,c1
After = 2.0398203987662103E18,143,94,2.0245202457662103E18,146,2.0245188177662187 5E18,0.254,1.9633203987662103E18,c1
Before = 10,125,70,26,115,31.1,0.205,41,c1
After = 2.0398101997662103E18,125,70,2.0347188177662103E18,115,2.0245203987662202 9E18,0.205,2.0041203987662103E18,c1
Before = 7,147,76,0,0,39.4,0.257,43,c1
After = 1.7185766210199E13,147,76,1.0199766210199E13,0,2.0245739376622003 2E17,0.257,2.0041202457662103E18,c1
Before = 13,145,82,19,110,22.2,0.245,57,c0
After = 2.0398202457662103E18,145,82,2.0398739376621020 8E17,110,2.0347203477662202 9E18,0.245,1.9633171857662103E18,c0
Before = 5,117,92,0,0,34.1,0.337,38,c0
After = 1.9633766210199E13,117,92,1.0199766210199E13,0,2.0245200417662202 9E18,0.337,2.0245139217662103E18,c0
```

Fig. 4. Data Encryption.

In Figure 4, Step 4 focuses on Data Encryption. This critical stage involves encrypting the data into a secure and unreadable format. By applying encryption, sensitive information within the dataset is protected from unauthorized access or potential breaches. This step enhances the overall security of the data, ensuring confidentiality and safeguarding it during storage, transmission, or any other data-handling processes.

Step 5: Data Anonymization



```
2.0398766210199E13,199,0,2.0398203987662103E18,0,2.0398139217662202 9E18,0.078,2.034
2.0398766210199E13,199,122,2.0398203987662103E18,0,1.0199766210199E13,0.078,2.03472
2.0398766210199E13,0,122,2.0398203987662103E18,0,2.0398139217662202 9E18,0.078,2.034
2.0398766210199E13,199,122,2.0398203987662103E18,0,2.0398139217662202 9E18,0.078,2.0
2.0398766210199E13,199,122,1.0199766210199E13,0,2.0398139217662202 9E18,0.078,2.0347
2.0398766210199E13,199,122,1.0199766210199E13,0,1.0199766210199E13,0.078,2.0347203
2.0398766210199E13,199,122,2.0398203987662103E18,0,2.0398139217662202 9E18,0.078,2.0
1.0199766210199E13,199,122,2.0398203987662103E18,744,2.0398139217662202 9E18,0.078,2
2.0398766210199E13,199,122,1.0199766210199E13,0,2.0398139217662202 9E18,0.078,2.0347
1.0199766210199E13,199,122,1.0199766210199E13,0,2.0398139217662202 9E18,0.078,2.0347
2.0398766210199E13,199,122,1.0199766210199E13,0,2.0398139217662202 9E18,0.078,2.0347
2.0398766210199E13,0,0,2.0398203987662103E18,0,2.0398139217662202 9E18,0.078,2.03472
2.0398766210199E13,199,122,2.0398203987662103E18,0,2.0398139217662202 9E18,0.078,2.0
2.0398766210199E13,199,122,2.0398203987662103E18,0,2.0398139217662202 9E18,0.078,2.0
2.0398766210199E13,199,0,1.0199766210199E13,0,2.0398139217662202 9E18,0.078,2.034720
2.0398766210199E13,0,122,2.0398203987662103E18,0,2.0398139217662202 9E18,0.078,2.034
```

Fig. 5. Data Anonymization

Calculates the minimum, maximum, and median values for each column in a 2D array d1. It uses three lists (minlt, maxlt, and medlt) to store the calculated values for each column.

Here's a breakdown of what the code does:

- It iterates through the columns of the 2D array d1 using the outer for loop. The loop runs from 0 to d1[0].length-1, which processes each column except the last one.
- Inside the outer loop, it initialises variables min and max to extreme values: Double.MAX_VALUE and Double.MIN_VALUE. These values are chosen as initial values to ensure that any value in the column will be smaller than min and larger than max initially.
- It then enters an inner loop using the for loop with the variable j. This inner loop iterates the rough rows of the 2D array d1 for the current column i.
- Inside the inner loop, it compares each value in the current column to minimum and maximum levels. When a number is entered that is less than the existing minimum, the minimum is set to that number. If the input value exceeds the maximum allowed, the maximum is increased to that value. This way, it tracks the minimum and maximum values in the column.
- After the inner loop completes, it calculates the median med as the average of the min and max values for the current column.
- The current column's min, max, and med values are then added to three separate lists (minlt, maxlt, and medlt) to store the results for each column.

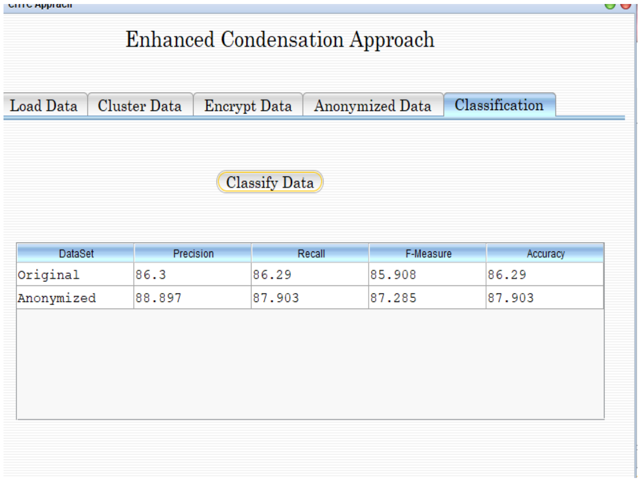
- The outer loop continues to the next column, and the process repeats.

By the end of this code, you'll have three lists: minlt containing the minimum values for each column, maxlt containing the maximum values for each column, and medlt containing the median values for each column in the 2D array d1. Following this,

- It modifies data in instances (possibly from an ARFF file). It replaces values with their corresponding min, max, or median values based on the calculated values from the previous code snippet you provided. It then saves the anonymized data to a CSV file. The code starts with a loop over at1, a list or collection of instances.
- Inside this loop, it retrieves an Instance object in1 from at1 and iterates through its attributes (excluding the last one) using a nested loop.
- It calculates the current attribute's values v1, min, max, and med. v1 is the value in the current instance, and min, max, and med are retrieved from the minlt, maxlt, and medlt lists, respectively. These lists store the minimum, maximum, and median values calculated for each attribute in the previous code snippet.
- It then checks whether v1 is less than med, greater than med, or equal to med. Depending on the comparison result, it replaces the v1 value with either min, max, or med in the current instance.
- The modified instance is added to a list lt, and its string representation is concatenated to the red string.
- After processing all attributes for the current instance, the loop continues to the next instance.The code updates some data structures dt with the anonymized data.
- It sets the text of a JTextArea component jTextArea4 with the content of the res string. This likely displays the anonymized data in a user interface.
- The code then constructs a CSV string SG containing the column names (attribute names) and the anonymized data.
- It writes the CSV string to a file named "anony.csv" by creating a File object, opening a FileOutputStream, and writing the string to the file.

This code snippet reads a list of instances, anonymizes the attribute values based on previously calculated min, max, and median values, and then saves the anonymized data to a CSV file. It's part of a data anonymization process that processes instances in an ARFF-like format and outputs the anonymized data in a CSV file.

Step 6: Classification



DataSet	Precision	Recall	F-Measure	Accuracy
Original	86.3	86.29	85.908	86.29
Anonymized	88.897	87.903	87.285	87.903

Fig. 6. Classification

Classification is done in two groups. One with an original data set and the other with anonymized data. The following are the factors which are considered for result comparisons.

For the result analysis, we have four factors.

- Precision
- Recall
- F-measure
- Accuracy

The efficacy of a classification model is commonly assessed using precision, recall, F-measure, and accuracy, such as those used in machine learning and statistics. They are instrumental in binary classification problems (where only two possible classes are positive and negative). Here's a brief explanation of each metric:

1. Precision: The accuracy of a model is quantified by how many Positive cases that are successfully anticipated comprise the total number of positive forecasts. It measures how well the model can rule out false positives.It is calculated as:

$$Precision = TP / (TP + FP) \quad (1)$$

where:

TP (True Positives) is the number of correct positive prediction.

False Positives (FP) is the number of erroneous positive predictions is the prediction.

As shown in the above figure, if we compare the precision values for the original data is 86.30, and that of the condensed approach is 88.39

2. Recall (Sensitivity or True Positive Rate):

The rate of recall indicates how many correct predictions were made out of a total number of positive examples. It provides a numerical measure of the model's sensitivity to positive cases.It is calculated as:

$$Recall = TP / (TP + FN) \qquad (2)$$

where:

TP shows the percentage of successful, positive predictions.
False negatives measure the percentage of true positives that were missed.
The recall output in our result is 86.29, while in original data, it is 87.90.

3. F-measure (F1 Score):

The F-measure balances precision and recall equally. It's a good compromise between accuracy and memory retention. This tool is invaluable when looking for a single measure that accounts for both false positives and false negatives.The formula for determining the F1 score

$$F1 = \frac{2*(Precision*Recall)}{(Precision + Recall)} \; (3)$$

Actual results values for the F-measure are 85.90 for the original data and 87.28 for the condensed approach.

4. Accuracy:

Accuracy is a measure of the overall correctness of a model's predictions. It determines what fraction of predictions were accurate (positive and negative). The formula is as follows:

$$Accuracy = \frac{(TP+TN)}{(TP+TN+FP+FN)} \qquad (4)$$

where:

TP shows the percentage of successful, positive predictions.
The True Negatives (TN) count indicates how many negative predictions were accurate.
The number of erroneous positive predictions is the False Positives (FP).
False negatives measure the percentage of true positives that were missed.
A classification model's quality and efficiency can be evaluated using these measures.
The result for accuracy for the original dataset is 86.29%, while the condensed approach is 87.90%.

V. CONCLUSION AND FUTURE SCOPE

The proposed privacy-preserving data mining method, focused on a novel condensation technique for regenerating multi-dimensional data records, presents a distinctive advantage by seamlessly integrating existing data mining algorithms without requiring constant modifications. Its compatibility with pseudo-data ensures the preservation of the original data format, eliminating the need for algorithm redesign. The incorporation of four protective steps enhances dataset security, posing challenges to both original dataset access and individual data identification. This method holds significant promise for applications in artificial intelligence and machine learning, particularly in fields like medicine, addressing the critical concern of patient data privacy. As technology evolves, further exploration of scalability and applicability across diverse domains could contribute to the ongoing advancement of privacy-preserving techniques in the data mining landscape.

REFERENCES

[1] G U. H. W. A. Hewage, R. Sinha, M. Asif Naeem, “Privacy-preserving data (stream) mining techniques and their impact on data mining accuracy: a systematic literature review”, Artificial Intelligence Review, Volume 56Issue 9, 2023.

[2] G. Sathish Kumarand K. Premalatha, “Privacy-preserving data mining - past and present”, International Journal of Business Intelligence and Data Mining, 2022

[3] Yousra Abdul Alsahib S. Aldeen,☒ Mazleena Salleh, and Mohammad Abdur Razzaque, “A comprehensive review on privacy preserving data mining” Springerplus. 2015

[4] Saad M. Darwish; Reham M. Essa; Mohamed A. Osman; Ahmed A. Ismail, “Privacy-Preserving Data Mining Framework for Negative Association Rules: An Application to Healthcare Informatics”, IEEE Access (Volume: 10), 2022

[5] Krishnaram Kenthapadi, Ilya Mironov, Abhradeep Guha Thakurta, “Privacy-preserving Data Mining in Industry”, WSDM '19: Proceedings of the Twelfth ACM International Conference on Web Search and Data Mining, 2019

[6] 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

[7] RICARDO MENDES AND JOAO P. VILELA, “Privacy-Preserving Data Mining: Methods, Metrics, and Applications”, IEEE Access 2017

[8] Alvina Anna John, S.Deepajothi, “Privacy Preservation Of Data Sets In Data Mining”, International Journal of Engineering Research & Technology (IJERT),2013

[9] S. Fletcher and M. Z. Islam, ‘‘Measuring information quality for privacy-preserving data mining,’’ Int. J. Comput. Theory Eng., vol. 7, no. 1, pp. 21–28, 2015.

[10] Shah and R. Gulati, ‘‘Privacy-preserving data mining: Techniques, classification and implications—A survey,’’ Int. J. Comput. Appl., vol. 137, no. 12, pp. 40–46, 2016