Privacy Preserving Data Mining

A Project Work-II Report

Submitted in partial fulfillment of requirement of the

Degree of

BACHELOR OF TECHNOLOGY in COMPUTER SCIENCE & ENGINEERING

BY

Divyanshu Sharma, Vasudha Yadav, Bhanu Chouhan

EN16CS301090 , EN16CS301284, EN16CS301069

Under the Guidance of

Mr. Hitesh Kag

Department of Computer Science & Engineering
Faculty of Engineering
MEDI-CAPS UNIVERSITY, INDORE- 453331

MAY 2020

Privacy Preserving Data Mining

A Project Work-II Report

Submitted in partial fulfillment of requirement of the

Degree of

BACHELOR OF TECHNOLOGY in COMPUTER SCIENCE & ENGINEERING

BY

Divyanshu Sharma , Vasudha Yadav , Bhanu Chouhan>
EN16CS301090 , EN16CS301284 , EN16CS301069

Under the Guidance of
Mr. Hitesh Kag

Department of Computer Science & Engineering
Faculty of Engineering
MEDI-CAPS UNIVERSITY, INDORE- 453331

MAY 2020

Report Approval

The project work **“Privacy Preserving Data Mining”** is hereby approved as a creditable study of an engineering subject carried out and presented in a manner satisfactory to warrant its acceptance as prerequisite for the Degree for which it has been submitted.

It is to be understood that by this approval the undersigned do not endorse or approved any statement made, opinion expressed, or conclusion drawn there in; but approve the “Project Report II” only for the purpose for which it has been submitted.

Internal Examiner

Name:

Designation

Affiliation

External Examiner

Name:

Designation

Affiliation

Declaration

We hereby declare that the project entitled **“Privacy Preserving Data Mining”** submitted in partial fulfillment for the award of the degree of Bachelor of Technology in ‘Computer Science & Engineering’ completed under the supervision of **Mr. Hitesh Kag , Assistant professor,** Faculty of Engineering, Medi-Caps University Indore is an authentic work. We also like to thank our project Coordinator Mr. **Ganesh Patidar , Assistant Professor & Mr. Sachin Solanki**, **Assistant Professor** for continuously supporting us.

Further, we declare that the content of this Project work, in full or in parts, have neither been taken from any other source nor have been submitted to any other Institute or University for the award of any degree or diploma.

Divyanshu Sharma(EN16CS301090)

Vasudha Yadav (EN16CS301284)

Bhanu Chouhan (EN16CS301069)

Date: ________________________

Certificate

I, **Hitesh Kag** certify that the project entitled **"Privacy Preserving Data Mining"** submitted in partial fulfillment for the award of the degree of Bachelor of Technology by **Divyanshu Sharma, Vasudha Yadav, Bhanu Chouhan** is the record carried out by them under my guidance and that the work has not formed the basis of award of any other degree elsewhere.

____

Mr. Hitesh Kag

Assistant Professor

Computer Science & Engineering

Medi-Caps University, Indore

____

Dr. Suresh Jain

Head of the Department

Computer Science & Engineering

Medi-Caps University , Indore

Acknowledgements

I would like to express my deepest gratitude to Honorable Chancellor, **Shri R C Mittal,** who has provided me with every facility to successfully carry out this project, and my profound indebtedness to **Prof. (Dr.) Sunil K Somani,** Vice Chancellor, Medi-Caps University, whose unfailing support and enthusiasm has always boosted up my morale. I also thank **Prof. (Dr.) D K Panda,** Dean, Faculty of Engineering, Medi-Caps University, for giving me a chance to work on this project. I would also like to thank my Head of the Department **Dr. Suresh Jain** for his continuous encouragement for betterment of the project.

I express my heartfelt gratitude to my Internal Guide, Mr. Hitesh Kag, Assistant Professor, Department of Computer Science & Engineering, MediCaps University, without whose continuous help and support, this project would ever have reached to the completion.

It is their help and support, due to which we became able to complete the design and technical report.

Without their support this report would not have been possible.

Divyanshu Sharma(EN16CS301090)
Vasudha Yadav (EN16CS301284)
Bhanu Chouhan (EN16CS301069)
B.Tech. IV Year
Department of Computer Science & Engineering
Faculty of Engineering
Medi-Caps University, Indore

Abstract

In a number of business scenarios the data collaboration and analysis is essential for making critical decisions. But all the business owners are worried about end client privacy and data security. Therefore the proposed work addresses this issue and proposed to design a privacy-preserving data model. That model enables a business owner to combine their data to other business owner's data and mine the data patterns on entire data. Thus first a Cryptographic technique based on the AES algorithm is implemented and then the ciphered data is used on the server for generating the decision rules. Finally, the rules are distributed to all the target parties. And the parties can only recover the data which are contributed by self. The implementation of the proposed secure data mining model is carried out using the JAVA and WEKA technology. The experimental results demonstrate the proposed technique outperforms in terms of accuracy, memory and time complexities.

Keywords: Privacy preserving data mining, Decision tree, Decision rules, Privacy and security.

Table of Contents

List of Figures

List of Tables

Chapter-1

Introduction to Privacy Preserving data Mining

1.1 Introduction

Data mining is an essential technique of extracting the application based target patterns from raw and huge data. These techniques are usages computational algorithms for performing the required task. According to literature, data mining techniques can be supervised or unsupervised in nature. The supervised learning techniques first accept initial predefined samples or examples to learn the patterns and then the unidentified patterns are recognized on the basis of previous learning . On the other hand, the unsupervised learning algorithms are directly applicable to the data and automatically create groups of data according to the data internal similarity or differences. However, supervised learning methods are much accurate as well as efficient with respect to unsupervised learning algorithms .

In this context, data mining techniques are much helpful in various real-world applications such as prediction, classification, association, and others . But sometimes for mining and extracting patterns from data can be sensitive because of the security and privacy point of view. In order to understand these issues let us assume that, there are a number of business owners who want to conduct an analysis on the consumer behavior for a target industry. In this scenario, a number of different business owners need to combine their own part of data in a common place for utilizing the data mining services and extraction of meaningful patterns. but the end data owners' records available in the outsource data can create privacy and security issues. Therefore in this presented work, a client-controlled privacy-preserving data mining model is proposed for implementation and design**.**

This segment makes available a fundamental indication of the proposed privacy-preserving data mining technique. In the next section, the related recent development is reported in privacy-preserving data mining. Further, the proposed data mining model is demonstrated and then the presentation of the realized system is provided. Finally based on the experiments and observed facts the conclusion and future work is proposed.

1.2 Literature Review

This section provides the recent development and contributions of different researches for finding the best and efficient approach for designing the privacy-preserving data mining system.The compilation and assessment of information are expanding because of the event of processing gadgets. The examination of data is encouraging organizations and contributing helpfully to society in numerous fields. In any case, this stockpiling and stream of touchy information present genuine concerns. Techniques that permit information extraction, while saving protection, are known as security safeguarding information mining (PPDM). ***R. Mendes et al [1]*** reviews the important PPDM strategies and the measurements used to assess such methods and uses of PPDM. Besides, the present difficulties and open issues are talked about.

Protection safeguarding is basic where information mining is changed into a helpful errand. ***V. Manikandan et al [2]*** propose a security saving edge grouping that utilizations a code-based system with the edge for sharing of mystery information in the protection safeguarding instrument. The procedure empowers the data to be apportioned into various offers and dealt with freely. This technique takes less emphasis in correlation with existing strategies that not require any trust among the customers or servers. Furthermore, it gives results on the security and proficiency of the technique.

Information irritation is an appreciated information-digging practice for protection safeguarding. A significant issue in information bother is the means by which to adjust the two clashing elements – security and information utility. ***S. Upadhyay et al [3]*** proposes a Geometric Data Perturbation (GDP) utilizing information parceling in three-dimensional turns. The properties are separated into gatherings of three and each gathering of qualities is turned about various sets of tomahawks. The turning round edge is picked to such an extent that the fluctuation bolstered security metric is raised which makes the first information recreation muddled. The same number of calculations like arrangement and bunching are invariant to geometric irritation, the information utility is protected right now. The assessment shows that the technique gives great protection safeguarding results and information utility.

S. Scardapane et al [6] consider the utilization of information mining strategies in clinical settings, wherein the information to be investigated is circulated among numerous gatherings. While derivation activities could offer important clinical data, each gathering is illegal to reveal its dataset to a unified area, because of security worries of the dataset. To this end, the creator suggests an all-inclusive structure encouraging the party to execute any information mining work on depending solely on the Euclidean separation among designs, including portion strategies, unearthly grouping, etc. The issue is reworked as a decentralized network finish issue, whose arrangement doesn't require the nearness of a concentrated organizer, and full protection of the information can be guaranteed by the various systems, including irregular multiplicative updates for secure calculation. Results bolster the proposition as a proficient apparatus for grouping and characterization.

The PPDM is assuming a significant job go about as rising innovation to perform different information mining procedures on private information and to give information in a safe manner to ensure touchy information. Numerous kinds of strategy for example randomization made sure about aggregate calculations, and k-namelessness has been recommended to execute PPDM. Rajesh N. et al [4], looks into the PPDM strategy with fluffy rationale, neural system, made sure about entirety, and different encryption calculations. This will permit grabbing hold of the differing go up against the face in PPDM and furthermore help us to locate the best fitting technique.

In PPDM, anonymization based methodologies have been utilized to safeguard protection. Existing writing tends to different anonymization based methodologies for safeguarding the touchy data. In any case, the anonymization based methodologies experience the ill effects of the issue of data misfortune. To limit the loss of different anonymization based grouping approaches viz. Voracious k-part and Systematic grouping calculation have been proposed. Among them, the Systematic bunching gives lesser misfortune. Furthermore, these methods use all properties during the development of an anonymized database. Therefore, the danger of presentation of touchy private data is higher. **P. R. Bhaladhare et al [7]** propose two methodologies for limiting the divulgence chance and safeguarding protection by utilizing precise bunching. The principal procedure readies a disparate blend of semi identifiers and delicate qualities. The following method creates an identical change of a semi identifier and delicate property. They additionally assess approaches concentrating on data misfortune and execution time. They show the adequacy of the methodologies by contrasting them and the current calculations.

1.3 Objectives

Privacy-Preserving Data Mining (PPDM) is a data mining and statistical databases innovative field where data mining algorithms are analyzed for side-effects in data privacy. It is also called privacyenhanced/privacy-sensitive data mining dealing with getting valid data mining results without learning underlying data values. This reveals how many different methods and techniques can be used in a PPDM context from a technical perspective.

Data mining techniques were developed to extracts knowledge to support various domains like weather forecasting, marketing, medical diagnosis and national security. But it is still challenging to mine specific data without violating data owners ’privacy. For instance, mining patients ’private data is an ongoing problem in health care applications. As data mining become more pervasive, privacy concerns increase. Commercial issues are also linked to the privacy issue. Most organizations collect information about individuals for specific needs. Frequently different units in an organization may find it necessary to share information. In such cases, each organization/unit must ensure that individual privacy is not violated or sensitive business information revealed. To avoid these types of violations, there is a need of various data mining algorithm for privacy preserving .

Data mining needs correct input for meaningful results, but privacy concerns influence users to provide wrong information. To preserve client privacy in data mining procedures, various random perturbation of data records based techniques were proposed.

1.4 Significance/Scope

Data collection and data mining techniques are applied to several application domains. Some of these domains require handling, and often publishing sensitive personal data (e.g. medical records in health care services), which raises the concern about the disclosure of private information. Privacy-Preserving Data Mining (PPDM) techniques have been developed to allow for the extraction of knowledge from large datasets while preventing the disclosure of sensitive information. The vast majority of the PPDM techniques modify or even remove some of the original data in order to preserve privacy . This data quality degradation is known as the natural trade-off between the privacy level and the data quality, which is formally known as utility. PPDM methods are designed to guarantee a certain level of privacy while maximising the utility of the data to allow for effective data mining. Throughout this work, sanitised or transformed data will refer to the data that resulted from a privacy-preserving technique.

However, this may result in unwanted privacy violations. To protect from information leakage,

privacy preservation methods have been developed to protect owner's exposure, by modifying the original data. However, transforming the data may also reduce its utility, resulting in inaccurate or even infeasible extraction of knowledge through data mining. This is the paradigm known as Privacy-Preserving Data Mining (PPDM). PPDM methodologies are designed to guarantee a certain level of privacy, while maximising the utility of the data, such that data mining can still be performed on the transformed data efficiently. PPDM encompasses all techniques that can be used to extract knowledge from data while preserving privacy.

Chapter 2

Present Investigation

2.1 Experimental Set-up

The proposed privacy-preserving data model is demonstrated in figure 2.1. According to the diagram the system contains a client and a server application. The client interface enabled to accept the dataset and encrypt the data in the client end. In order to keep security and privacy, the data owner encrypt the data using a private key encryption algorithm. Therefore here the AES encryption algorithm is implemented. The key reason for implementing the AES algorithm is that this algorithm is efficient and less time and memory consuming. Additionally, the algorithm provides the flexibility to use the variable size of cryptographic key. Therefore the user can use any password to encrypt the data. Figure 2.3 shows the cryptographic process used for encrypting the user side.

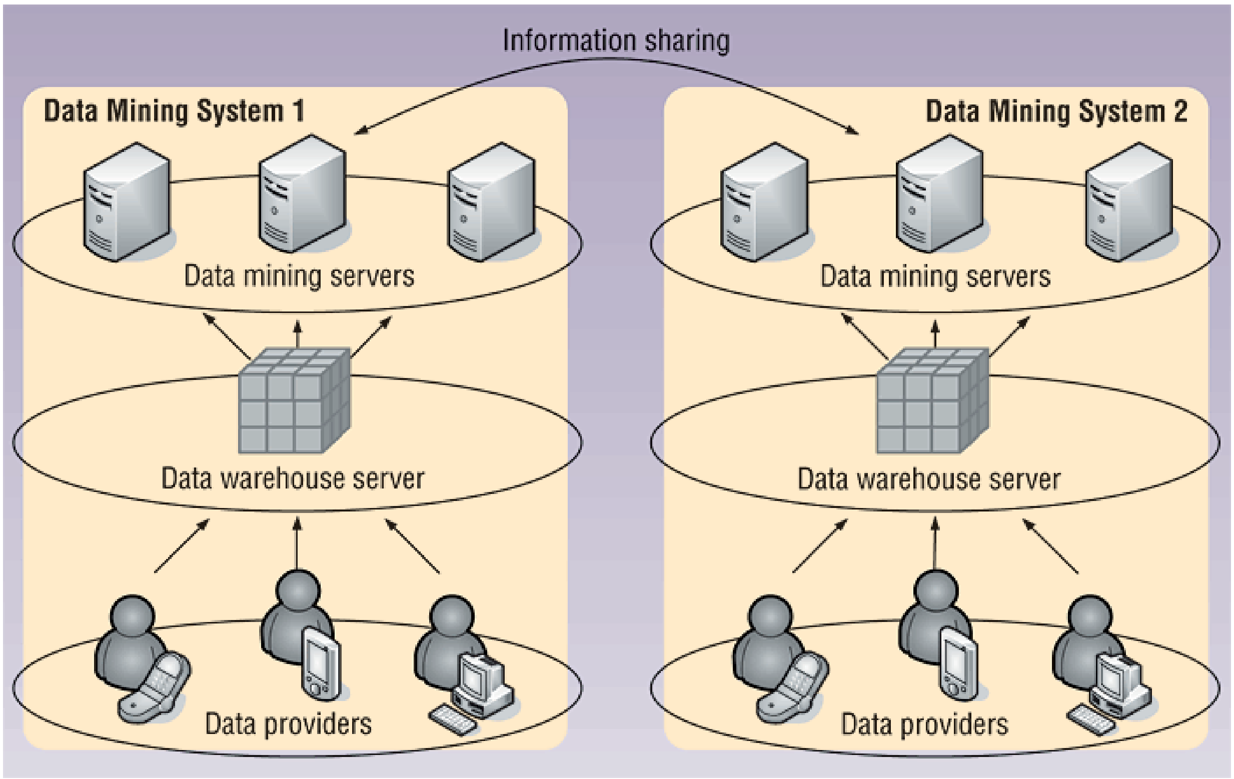

Figure 2.1

According to the given figure the user who agreed to combine and data mine with the other business parties. Produce the input part of data to the system. Additionally, a password is also provided by the user. The given password is produced to the MD5 hash generation algorithm to generate the 128-bit password. Further, the input data and the 128-bit password is produced to the AES encryption algorithm. That algorithm processes the data and generates the ciphertext. After encryption of data, the data is sent to the server for the data mining process.

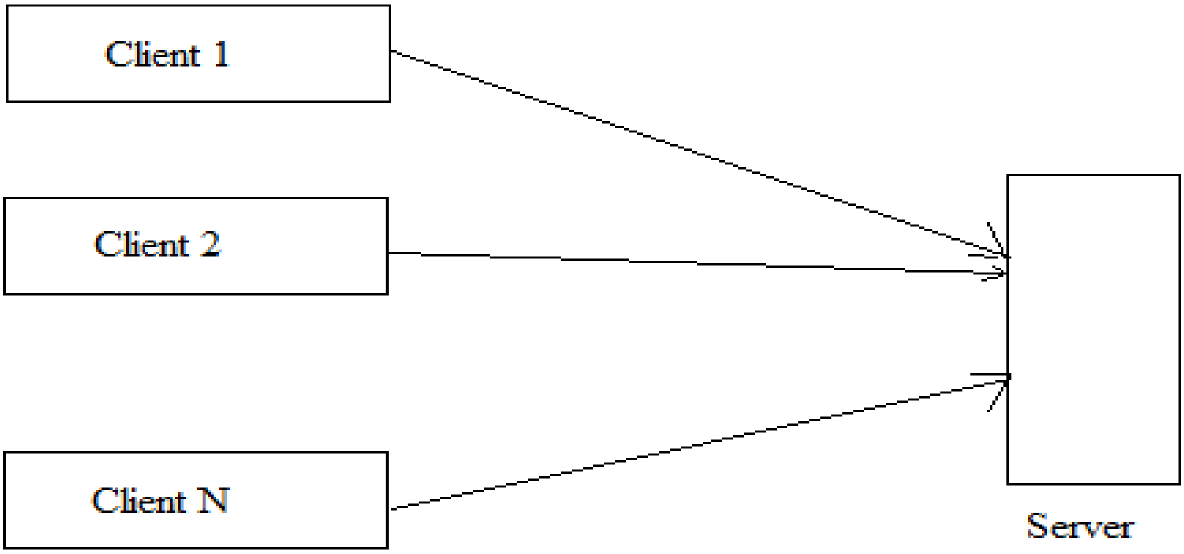

Figure 2.2

When the data appeared on the server, the server reorganizes the data obtained from different parties. Here all the client data is considered in a vertical partitioned format. All the aggregated data is passed on the C4.5 decision tree for generating "IF-THEN-ELSE" rules. After generating the decision rules using the C4.5 algorithm the entire rules are distributed to all the concerning parties. At the client end, the received rules are decrypted using the previously passed cryptographic key. The benefit of this technique is that the client can only view the part of rule which is contributed by own. Additionally, the decision labels are available at all the party. Thus each party can securely obtain the secure decision rules without disclosing any part of data.

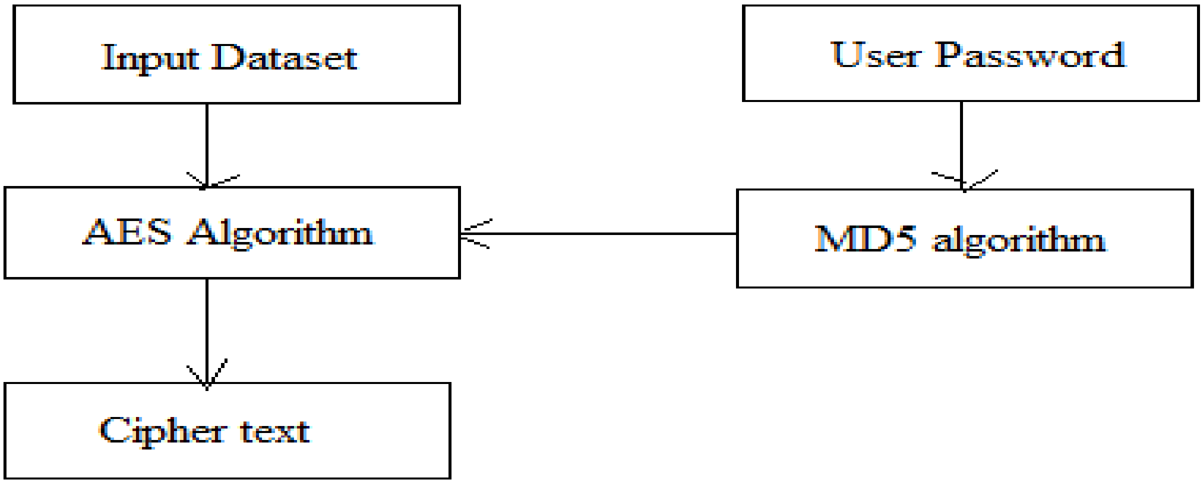

Figure 2.3

In recent years, many PPDM methods have been developed but

there is no standardization in these approaches. To achieve optimized results

while preserving the privacy of the data subjects efficiently, five dimensions

need to be considered and listed below:

(1) The distribution of the basic data

(2) The modification of the basic data

(3) Mining method being used

(4) If basic data or rules are to be hidden and

(5) Additional methods for privacy preservation used.

This shows that from a technical viewpoint many different methods and procedures in the perspective of PPDM that can be used.

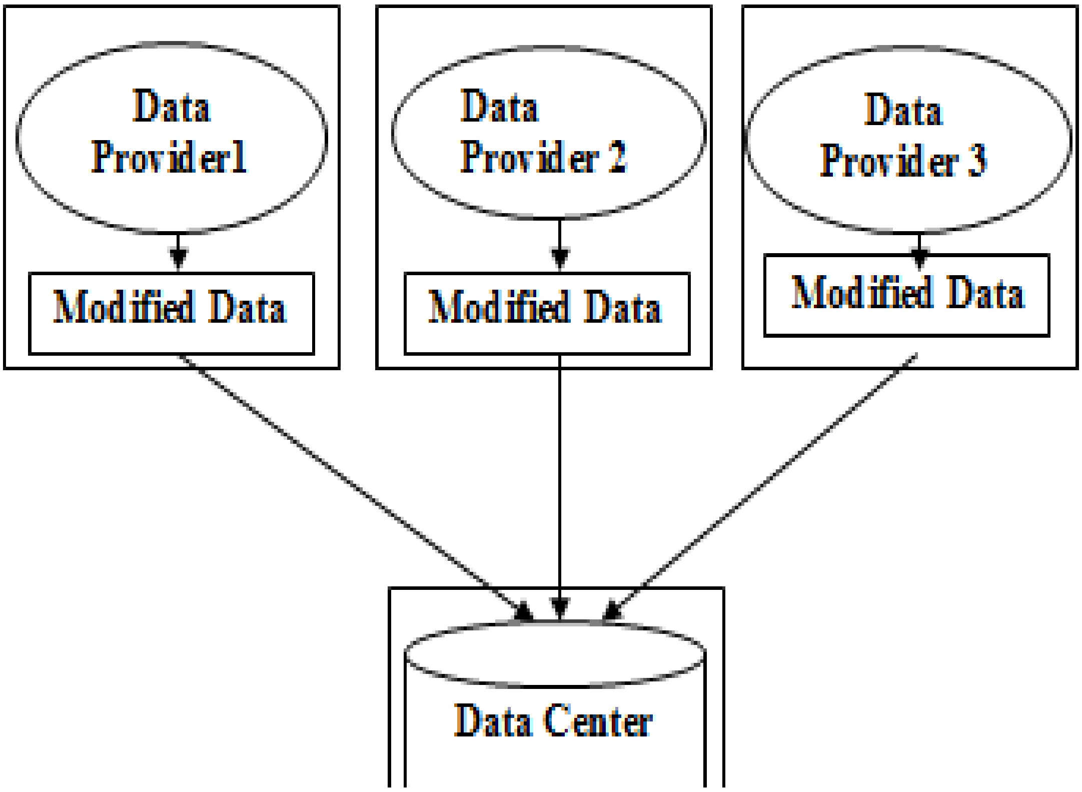
Data Provider1
Modified Data
Data Provider 2
Modified Data
Data Provider 3
Modified Data
Data Center

Figure 2.4

2.2 Procedures Adopted

PPDM is divided into two groups: data hiding and rule hiding. The objective of data hiding is to transform data or design new computation protocols to ensure that private data remains private during or after data mining; when underlying data patterns/models can be discovered. Techniques like multiplicative perturbation, additive perturbation and secure multi-party computation fall in this category. Rule hiding, on the other hand, transforms database so that sensitive rules are masked still allowing underlying patterns to be discovered.

Most privacy computation methods use some form of transformation on data to ensure privacy preservation. Such methods reduce representation granularity to reduce privacy. This results in some data management/mining algorithms loss of effectiveness; a natural trade-off between information loss and privacy. Usually such methods reduce representation granularity to reduce privacy. This reduction leads to some loss in data management/mining algorithms effectiveness - a trade-off between information loss and privacy.The most common techniques are Randomization techniques, group based anonymization and distributed PPDM.

2.2.1 Randomization Techniques

Randomization technique is the process of perturbing the input data to distributed data mining algorithms so that the data values of individual entities are protected from revealing. Several randomization techniques has been identified in PPDM algorithms by including Adding random numbers, Generating random vectors and Random permutation of a sequence.

Randomization is a technique easily implemented during data collection as noise added to a record is independent of other data records behaviour. This is a weakness as outlier records are difficult to mask. Clearly, where privacy-preservation is not required at data-collection time, a technique where inaccuracy depends on behaviour of the locality of that record is necessary. Another randomization framework weakness is that it does not consider the chances that available records can identify the that record’s owners. It was shown that use of publicly available records leads to privacy

being compromised in high-dimensional cases . This holds good for outlier records easily distinguished from others in their locality. Hence, a broad privacy transformation approach is constructing anonymous records groups that are transformed in group-specifically.

2.2.2 Decision Tree Algorithm

Decision Tree is a Supervised learning technique that can be used for both classification and Regression problems, but mostly it is preferred for solving Classification problems. It is a tree-structured classifier, where internal nodes represent the features of a dataset, branches represent the decision rules and each leaf node represents the outcome.

A decision tree is a flowchart-like tree structure where an internal node represents feature(or attribute), the branch represents a decision rule, and each leaf node represents the outcome. The topmost node in a decision tree is known as the root node. It learns to partition on the basis of the attribute value. It partitions the tree in recursively manner call recursive partitioning. This flowchart-like structure helps you in decision making. It's visualization like a flowchart diagram which easily mimics the human level thinking. That is why decision trees are easy to understand and interpret.

Decision Tree is a white box type of ML algorithm. It shares internal decision-making logic, which is not available in the black box type of algorithms such as Neural Network. Its training time is faster compared to the neural network algorithm. The time complexity of decision trees is a function of the number of records and number of attributes in the given data. The decision tree is a distribution-free or non-parametric method, which does not depend upon probability distribution assumptions. Decision trees can handle high dimensional data with good accuracy.

Types of decision tree is based on the type of target variable we have. It can be of two types:

1. Categorical Variable Decision Tree: Decision Tree which has categorical target variable then it called as categorical variable decision tree. Example:- In above scenario of student problem, where the target variable was “Student will play cricket or not” i.e. YES or NO.
2. Continuous Variable Decision Tree: Decision Tree has continuous target variable then it is called as Continuous Variable Decision Tree.

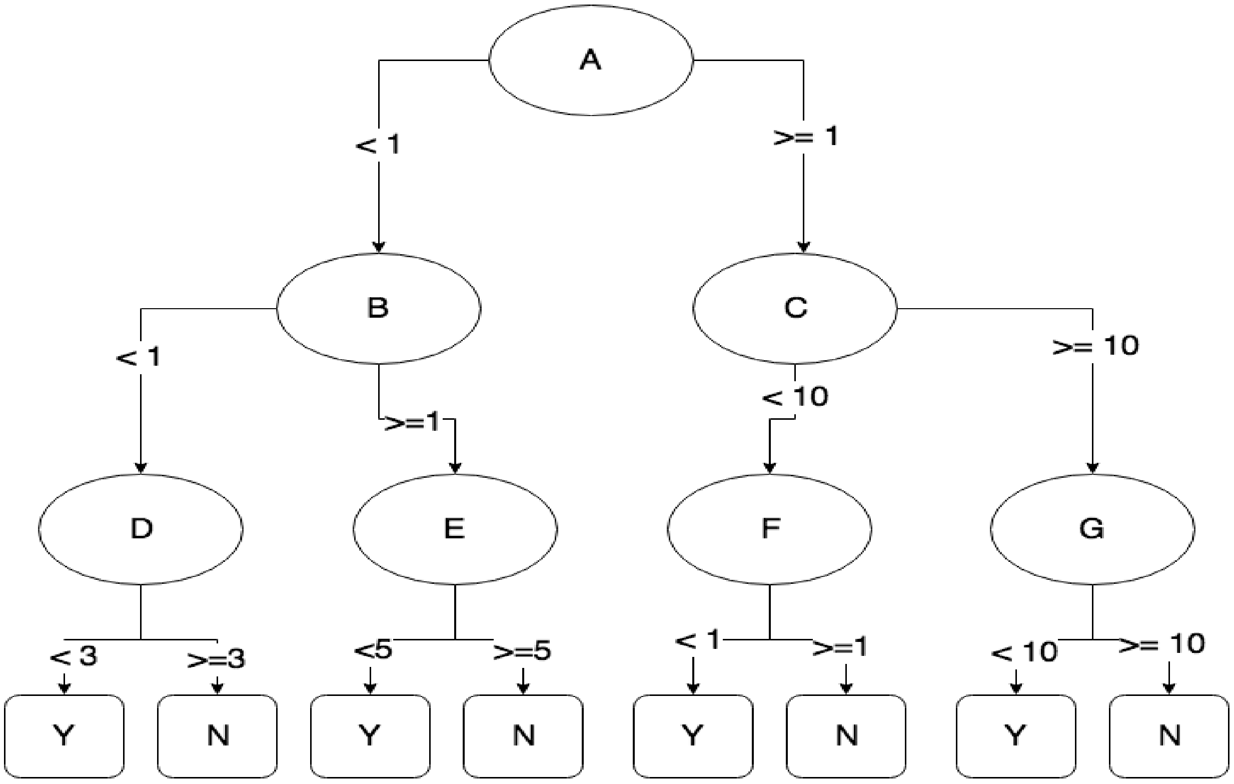

Figure 2.5

2.2.3 Data Preprocessing

Data preprocessing is a data mining technique which is used to transform the raw data in a useful and efficient format.

Steps Involved in Data Preprocessing:

1. Data Cleaning:

The data can have many irrelevant and missing parts. To handle this part, data cleaning is done. It involves handling of missing data, noisy data etc.

(a) Missing Data:

This situation arises when some data is missing in the data. It can be handled in various ways. Some of them are:

1. Ignore the tuples:

This approach is suitable only when the dataset we have is quite large and multiple values are missing within a tuple.

2. Fill the Missing values:

There are various ways to do this task. You can choose to fill the missing values manually, by attribute mean or the most probable value.

(b) Noisy Data:

Noisy data is a meaningless data that can’t be interpreted by machines.It can be generated due to faulty data collection, data entry errors etc. It can be handled in following ways :

3. Binning Method:

This method works on sorted data in order to smooth it. The whole data is divided into segments of equal size and then various methods are performed to complete the task. Each segmented is handled separately. One can replace all data in a segment by its mean or boundary values can be used to complete the task.

4. Regression:

Here data can be made smooth by fitting it to a regression function.The regression used may be linear (having one independent variable) or multiple (having multiple independent variables).

5. Clustering:

This approach groups the similar data in a cluster. The outliers may be undetected or it will fall outside the clusters.

2.Data_Transformation

This step is taken in order to transform the data in appropriate forms suitable for mining process.

This involves following ways:

1. Normalization:

   It is done in order to scale the data values in a specified range (-1.0 to 1.0 or 0.0 to 1.0)

2. Attribute Selection:

   In this strategy, new attributes are constructed from the given set of attributes to help the mining process.

3. Discretization:

   This is done to replace the raw values of numeric attribute by interval levels or conceptual levels.

4. Concept Hierarchy Generation:

   Here attributes are converted from level to higher level in hierarchy. For Example-The attribute “city” can be converted to “country”.

3. Data Reduction:

Since data mining is a technique that is used to handle huge amount of data. While working with huge volume of data, analysis became harder in such cases. In order to get rid of this, we uses data reduction technique. It aims to increase the storage efficiency and reduce data storage and analysis costs.

The various steps to data reduction are:

1. Data Cube Aggregation:

Aggregation operation is applied to data for the construction of the data cube.

2. Attribute Subset Selection:

The highly relevant attributes should be used, rest all can be discarded. For performing attribute selection, one can use level of significance and p- value of the attribute.the attribute having p-value greater than significance level can be discarded.

3. Numerosity Reduction:

This enable to store the model of data instead of whole data, for example: Regression Models.

4. Dimensionality Reduction:

This reduce the size of data by encoding mechanisms.It can be lossy or lossless. If after reconstruction from compressed data, original data can be retrieved, such reduction are called lossless reduction else it is called lossy reduction.

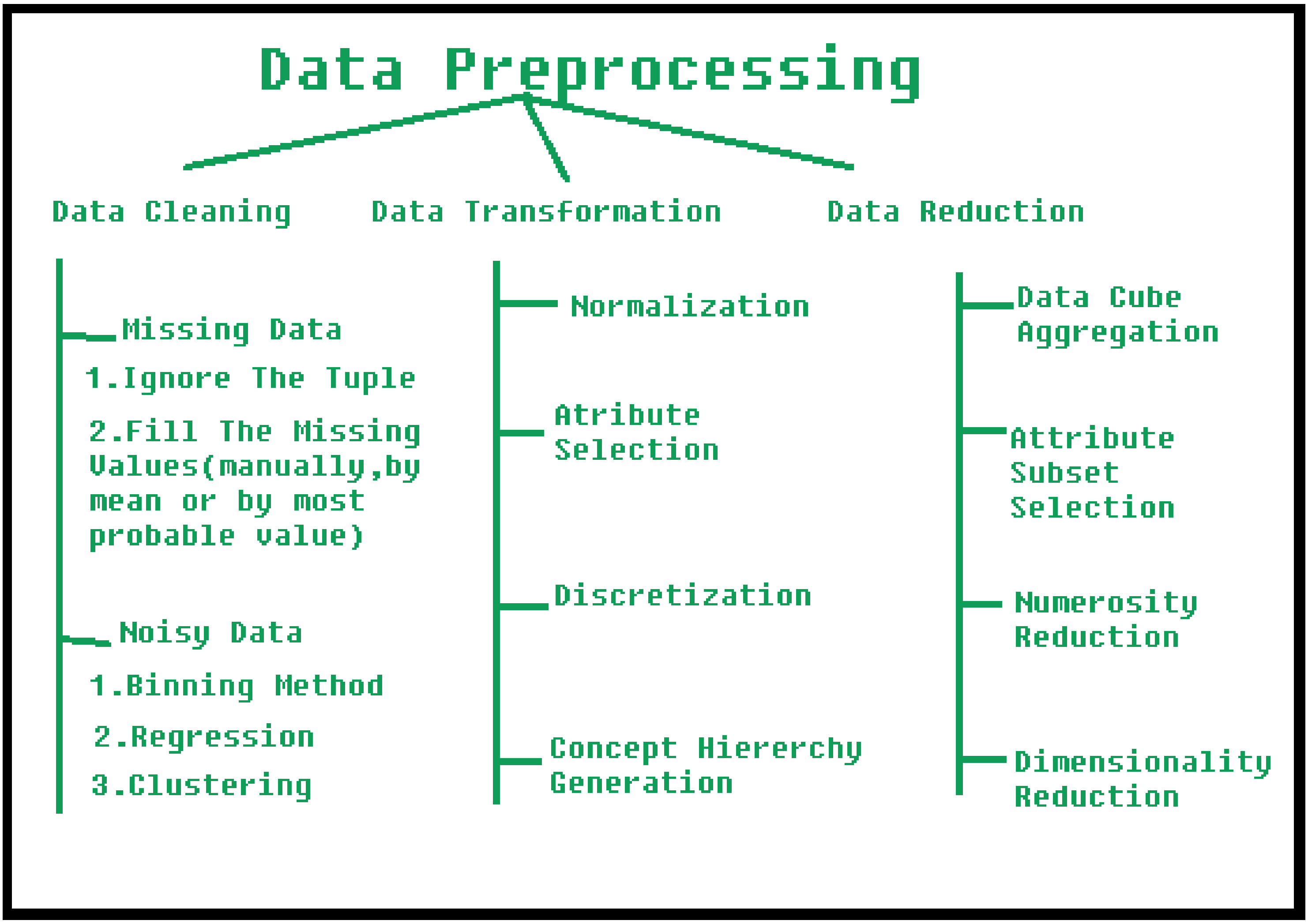

Figure 2.6

2.3 Models of PPDM

i. Trust Third Party Model

The security standard assumes we have a trusted third party to which all data is given. The third party performs computation and delivers results and except for this party, nobody learns anything inferable from own input/ results. Secure protocols aim to reach this privacy preservation level without finding a third party everyone trusts.

ii. Semi-honest Model

In this, all *parties* follow protocol rules using correct input, but when the protocol is free it uses anything it sees during protocol execution to compromise security.

iii. Malicious Model

In malicious model, participants have no restrictions. Any party is free to indulge in any action. Usually, it is difficult to develop efficient protocols valid under a malicious model.

iv. Other Models - Incentive Compatibility

Though semi-honest and malicious models are well researched, other models outside purview of cryptography are also possible. An example is incentive compatibility. A protocol is incentive compatible when a cheating party is either caught/suffers an economic loss. Under the rational economics model, this ensures that parties have no advantage by cheating. Of course, this fails in an irrational model.

Chapter 3

Testing & Results

The proposed privacy-preserving data mining technique is implemented in the previous section. In this section, the experimental analysis of the given model is provided. Thus different performance factors are measured and reported here.

A. Accuracy

The proposed PPDM technique is implemented with the C4.5 decision tree algorithm. Here the performance of the decision tree before and after encryption of the data performance in terms of accuracy of the decision tree is measured and reported in table 3.1 and figure 3.1. The accuracy is an essential factor of performance in data mining. That is measured using the ratio of correctly recognized and total samples to recognize. The following equation can help to understand.

$$accuracy = \frac{total\, correctly\ classified\ X\ 100}{total\, samples} classify$$

The table 3.1 shows the accuracy of both the scenarios in first before encryption the accuracy of the algorithm is computed and then after the accuracy of the algorithm on the same dataset is measured and compared in table 3.1 their graphical representation is given in figure 3.1.

Dataset Size	Before Encryption(%)	After encryption (%)
100	78	77
300	82	78.3
500	81	77.5
800	82.8	79.6
1000	84.6	81.4
1500	84.9	81.8
2000	86.2	83.2

Table 3.1 accuracy (%)

Figure 3.1 shows a line graph for accuracy, here the X-axis contains the size of datasets that are used for experiments. Additionally, the Y-axis shows the accuracy of the decision tree algorithm in percentage.

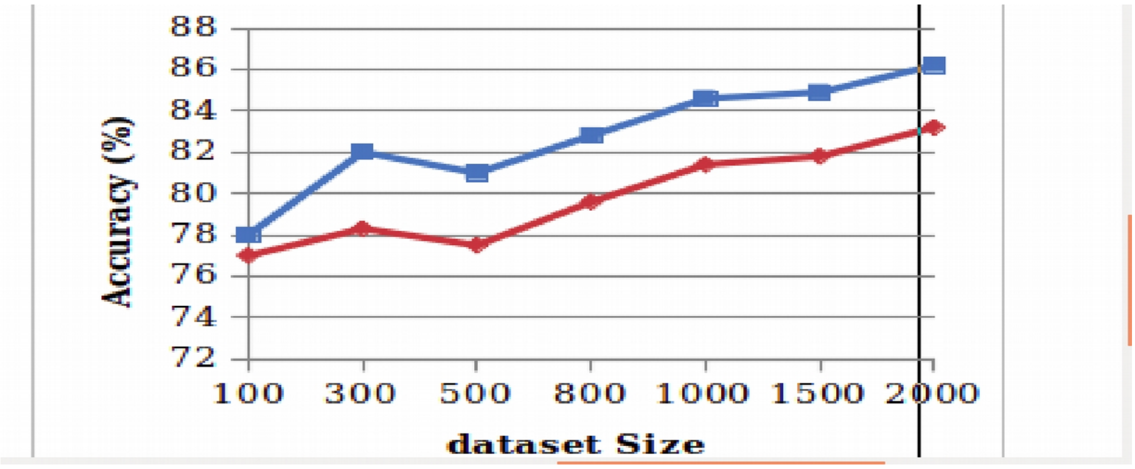

Figure 3.1

The red line shows the performance of the decision tree after encryption and before encryption, the performance is given using the blue line. According to the results, the encoding of data can change the application of a system, additionally, the combination can also impact of data mining application.

B Encryption Time & Decryption Time

The time consumption of an algorithm is also known as time complexity. The time required to execute an algorithm with the application data is termed here as the time consumption. That is estimated using the following formula:

$Time\, consumption = algorithm\, end\, time - start\, time$

Dataset Size	Encryption time(ms)	Decryption time(ms)
100	103	91
300	261	255
500	302	289
800	351	328
1000	389	369
1500	447	428
2000	501	479

Table 3.2 Time consumption

Table 3.2 shows the time consumption of the proposed system during the encryption and decryption of data. The measured values of time in terms of milliseconds (ms) are denoted here.

The line graph representations of table 3.2 values are given in table 3.2. The X-axis of the diagram demonstrates the dataset size and Y-axis shows the time consumption of the algorithm in terms of milliseconds. According to the noticed results, the encryption and decryption approximately require a similar amount of time. But the decryption of data mostly requires less amount of time as compared to encryption.

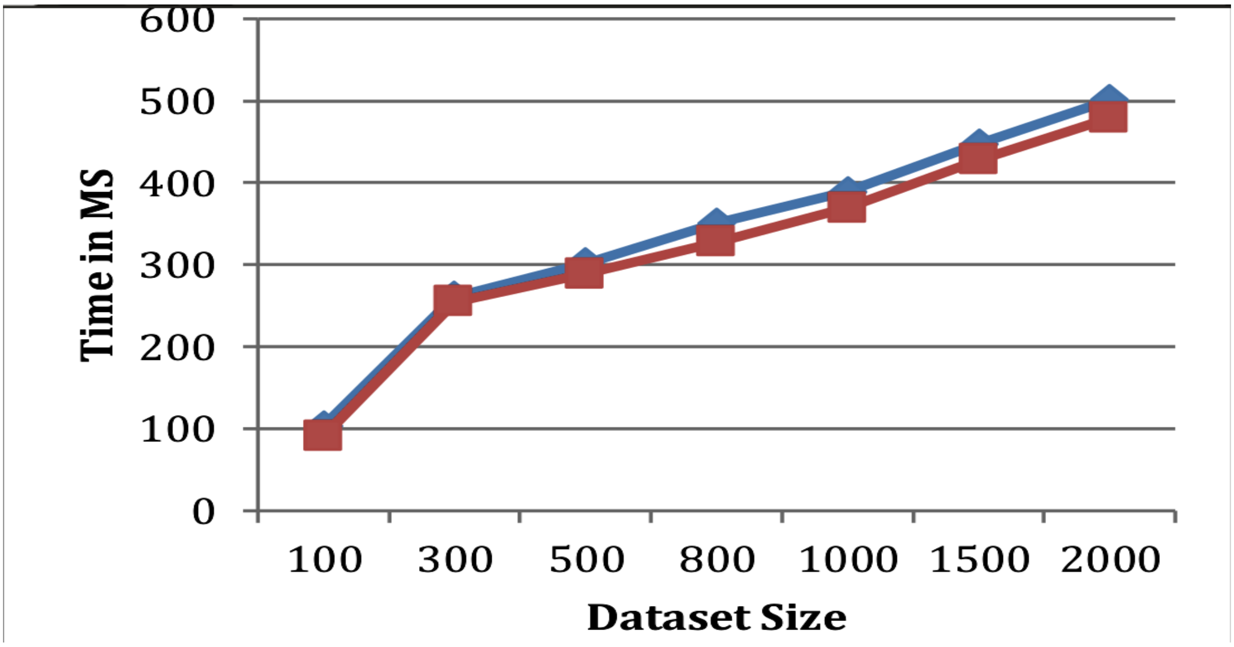

Figure 3.2

C. Encryption and Decryption Memory

Memory usages or consumption is also known as the space complexity of the algorithm. The memory usages of the algorithm during the encryption and decryption of data is measured and reported in this section. JAVA technology enables us to calculate the memory usages of the processes. The following equation can be used to measure it.

Memory usage = Total assigned – Total left

Table 3.3 values are represented in a line graph as given in figure 4.3. In this diagram, the X-axis shows the size of the dataset additionally the Y-axis shows the memory usages in terms of

KB. The red line shows the decryption memory and the blue line shows the memory of encryption. According to the measured memory, the encryption takes a higher amount of memory as compared to the decryption.

Dataset Size	Encryption memory(KB)	Decryption memory(KB)
100	16727	15253
300	16625	15277
500	16829	15529
800	17836	15376
1000	17242	15927
1500	17489	16183
2000	17352	16252

Table 3.3 Memory Consumption

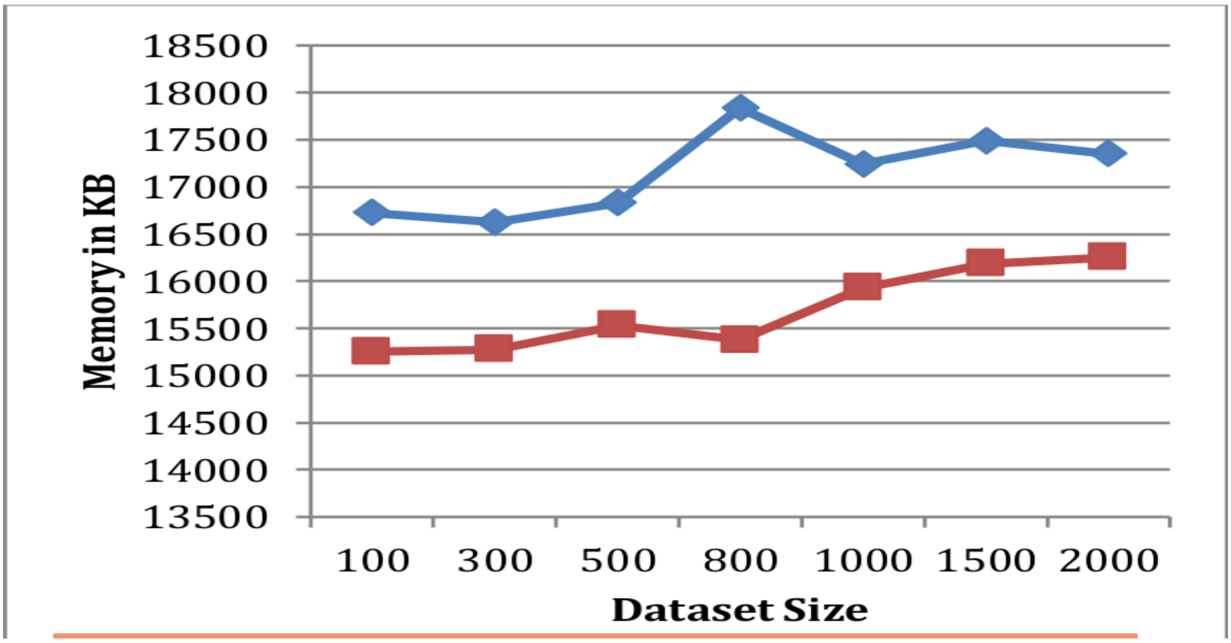

Figure 3.3 Memory Consumption

Chapter 4

Advantage of PPDM

PPDM has emerged to protect the privacy of sensitive data and also give valid data mining results. Figure 1.2 shows a distributed PPDM scenario which can achieve reasonable privacy and good accuracy. Often a trade-off between privacy and accuracy are needs to be made. On the one hand, privacy requires that the original data records must be fully obfuscated before data mining analysis. On the other hand, accuracy needs that the “patterns” in the original data should be mined out in spite of the perturbation (Likun Liu et al 2012)

There are two major methods in PPDM First by using cryptographic representation and The other is by using heuristic algorithms which ensures that sensitive data is not revealed.

Most of the current industry requires that these data can be secured during transmission and also when the data is present in the data warehouse (Kumar et al 2013). Originally PPDM extended the traditional data mining techniques to work with data hiding sensitive information, but the major issue was how to modify data and how to recover data mining results from that

modified data. The goals of a PPDM algorithm include:

i. Prevent the discovery of sensible information.

ii. Being uncompromised to access and to use the non-sensitive data.

iii. Being usable on large amounts of data.

iv. Must have less exponential computational complexity.

Privacy is an important issue in many data mining applications and it deals with some application fields such as

- Health care,
- Security,
- Financial and
- Other types of sensitive applications (Kamakshi et al 2010).

Chapter 5

Conclusions and Future Work

The proposed work is motivated to explore and investigate the privacy-preserving data mining technique. During the study, it is recognized that data mining can be used to manage the secure data modeling. In this context, a cryptographic technique is applied to the client-side on which the data owner first encrypts the data and then submits their data to the server for processing or decision rule mining. At the end of the server, only data is combined and then processed using decision tree C4.5. Finally, the recovered rule is distributed to all the clients and they can use these rules by deciphering the rules by their own private keys. On the basis of design observations and obtained outcomes of the experiments, the following facts are concluded.

(a) The encryption of data can impact on the performance of the supervised classifiers

(b) The decryption of data requires less amount of time and memory to process the data

(c) Accuracy of the decision tree also depends on other data owner's part of the information

By these concluding facts, the proposed model is extended for the following task.

A. Improve the data model to manage the actual data utility of the different party inputs

B. Need to reduce the cryptographic time and memory consumption using some other alternative techniques

Chapter 6

Bibliography and References

- K. S. Deepashri, A. Kamath, “Survey on Techniques of Data Mining and its Applications”, International Journal of Emerging Research in Management &Technology ISSN: 2278-9359 (Volume-6, Issue-2)

- D. Bhamare, P. Suryawanshi, “Review on Reliable Pattern Recognition with Machine Learning Techniques”, Fuzzy Information and Engineering, 2019, Vol. 10, No. 3, 362–377

- M. S. Mahdavinejad, M. Rezvan, M. Barekatain, P. Adibi, P. Barnaghi, A. P. Sheth, “Machine learning for internet of things data analysis: a survey”, Digital Communications and Networks, 2018, 4, 161-175

- V. Manikandan, V. Porkodi, A. S. Mohammed, M. Sivaram, “Privacy Preserving Data Mining Using Threshold Based Fuzzy C-means Clustering”, ICTACT Journal on Soft Computing, Oct. 2018, Volume: 09, Issue: 01

- S. Upadhyay, C. Sharma, P. Sharma, P. Bharadwaj, K. R. Seeja, “Privacy preserving data mining with 3-D rotation transformation”, King Saud University Journal of King Saud University – Computer and Information Sciences, 2018, 30, 524-530

- N. Rajesh, K. A. Sujatha, A. L. Selvakumar, “Survey on Privacy Preserving Data Mining Techniques using Recent Algorithms”, International Journal of Computer Applications (0975 – 8887), Volume 133 – No.7, January 2016

- R. Lu, K. Heung, A. H. Lashkari, A. A. Ghorvani, “A Lightweight Privacy-Preserving Data Aggregation Scheme for Fog Computing-Enhanced IoT”, 2169-3536 2017 IEEE, Vol. 5, 2017

- S. Scardapane, R. Altilio, V. Ciccarelli, A. Uncini, M. Panella, “Chapter 12: Privacy-Preserving Data Mining for Distributed Medical Scenarios”, Smart Innovation, Systems and Technologies 69, Springer International Publishing AG 2018