DOI: https://doi.org/10.52756/ijerr.2024.v44spl.002 *Int. J. Exp. Res. Rev.*, Vol. 44: 11-19 (2024)

Original Article | *Peer Reviewed* | *Open Access*

International Journal of Experimental Research and Review (IJERR)

ISSN: 2455-4855 (Online)
www.iaph.in
ISSN 2455-4855

A Privacy-Preserving Data Mining Through Comprehensive GNIPP Approach in Sensitive Data Sets

Shailesh Kumar Vyas[1*] and Swapnili Karmore[2]

[1]Department of CSE, G H Raisoni University Saikheda, Chhindwara, 480106, India; [2]Department of CSE, G H Raisoni University Saikheda, Chhindwara, 480106, India

E-mail/Orcid Id:

SV, shailesh.pk.29@gmail.com, https://orcid.org/0009-0003-1262-7353;
SK, swapnilikarmore@gmail.com, https://orcid.org/0000-0003-2068-4043

Article History:

Received: 12th Jan., 2024
Accepted: 16th Oct., 2024
Published: 30th Oct., 2024

Keywords:

Decision tree, privacy preservation, GNIPP, Gaussian noise, random forest, Gini impurity. accuracy

How to cite this Article:

Shailesh Kumar Vyas and Swapnili Karmore (2024). A Privacy-Preserving Data Mining Through Comprehensive GNIPP Approach in Sensitive Data Sets. *International Journal of Experimental Research and Review*, *44*, 11-19.

DOI: https://doi.org/10.52756/ijerr.2024.v44spl.002

Abstract: The quick growth of methods for analyzing data and the availability of easily available datasets have made it possible to build a thorough analytics model that can help with support decision-making. In the meantime, protecting personal privacy is crucial. A popular technique for medical evaluation and prediction, decision trees are easy to comprehend and interpret. However, the decision tree construction procedure may reveal personal information about an individual. By keeping the statistical properties intact and limiting the chance of privacy leaking within a reasonable bound, differential privacy offers a formal mathematical definition of privacy. To construct a boosting random forest that preserves privacy, we propose a Gaussian Noise Integrated Privacy Preservation (GNIPP) in this study. To address the issue of personal information breaches, we have designed a unique Gaussian distribution mechanism in GNIPP that enables the nodes with deeper depth to obtain more privacy during the decision tree construction process. We propose a comprehensive boosting technique based on the decision forest's prediction accuracy for assembling multiple decision trees into a forest. Furthermore, we propose an iterative technique to accelerate the assembly of decision trees. After all, we demonstrate through experimentation that the suggested GNIPP outperforms alternative algorithms on two real-world datasets.

Introduction

The importance of personal information has received increasing emphasis in recent years. Individuals in this data-driven age generate vast amounts of personal data regularly. The revolutionary technique known as data mining can give more personalised and improved services in various industries, including online search, healthcare, and medicine (Vyas and Karmore, 2022). For instance, sophisticated data mining methods can be applied in the healthcare industry to give patients better medical care. However, when external information about patients stored in a medical record system becomes available throughout the process of data mining and evaluation procedures, patient privacy may be compromised. In particular, mining patient electronic medical records may uncover information helpful to medical therapy, such as the underlying relations between different diseases (Vyas et al., 2024). However, this method may potentially expose patients' personal information. Thus, in the field of data mining, an efficient Privacy-Preserving Data Mining (PPDM) technique is crucial, as it can deal with the requirement of exposing database contents while preserving the confidentiality of personal data (Jain et al., 2023; Jain and Thada, 2024).

One crucial data mining technique that is essential to data analysis and prediction is classification. A common example of a tree-like classification model is a decision tree-oriented random forest booster (Karmore and Mahajan, 2016). It performs well in terms of classification accuracy overall and is frequently employed as a classification technique in practical applications. Nevertheless, there is a chance that the decision tree mechanism and the associated counting requirements will reveal private data. Considering that two nearby data sets, with a maximum of one record difference, are utilized to train two trees.

***Corresponding Author:** shailesh.pk.29@gmail.com

A strict concept of privacy that opposes individual privacy disclosure is called differential privacy (Talat et al., 2019). According to this definition, the outcome of the data sets' computation process is unaffected by modifying a single record in the data sets. Differential privacy, which has been extensively utilized in PPDM, was first implemented in the area of statistics database security. It was created to safeguard the privacy of specific database users when publishing statistical data. As we can see in (Chamikara et al., 2021), a variety of data mining techniques, including clustering (Ling et al., 2024), classification (Kumar, 2016), as well as deep learning (Kumar et al., 2023), can accomplish privacy preservation when combined with differential privacy. An architecture for attaining noise-integrated data mining is shown in Figure 1. In this scenario, the data miner submits queries to the Differentially Private Data Set (DPDS) along with the associated privacy parameters, but they are unable to access the original data directly. DPDS calculates the query answer, and it is then modified in a way that respects differential privacy. Since every query complies with differential privacy, data miners are unable to obtain any sensitive information.

There has been progress in building a tree-based model with differential privacy in recent years. The majority of suggested strategies focus on two primary areas. One approach is to reduce unpredictability by creating a novel scoring function with a lower sensitivity or by employing local sensitivity instead of global sensitivity (Kumar et al., 2023). the alternative direction is ensemble (Kulkarni et al., 2022). Nevertheless, the majority of approaches have neglected the influence that privacy allocation brings, meaning that nodes at various levels have varying noise tolerance capacities. One recent article has only proposed a unique technique for privacy allocation that dynamically sets each query's privacy parameter (Chamikara et al., 2021). Nevertheless, figuring out how much privacy is preserved during each inquiry requires the additional value of privacy parameters.

In this research, we offer a booster version of Random Forest algorithm with Gaussian noise integration. Additionally, we carefully combine a number of decision trees into an ensemble to enhance prediction performance. The following clearly describes our principal contributions:

To prevent a large decrease in decision tree performance carried on by the allocation of differential privacy parameters, we create a sensible privacy parameter allocation technique that allows varying amounts of privacy value to nodes at various levels. We often allow a bigger value of privacy parameters to the nodes of the decision tree placed deeper since the true count of such nodes is more vulnerable to noise.

We suggest a selective ensemble technique to increase the ensemble model's capacity for generalization and precision in prediction. Furthermore, we devise an iterative technique to accelerate the aggregating process.

We build a number of simulation experiments using actual data sources. The outcomes of the experiment demonstrate that our GNIPP classification model can outperform other models while still maintaining user privacy.

This is how the remainder of this paper is structured. The relevant work is presented in Section 2. Section 3 presents the preliminary findings. A summary of the suggested GNIPP technique and the entire system threat model are provided in section 4. Section 5 provides an explanation of how decision trees and random forests are created. A theoretical analysis of privacy follows. Section 6 presents the conclusion of the research.

Literature review

There are various methods available today to protect data privacy, including differential privacy and data anonymization (Rafiei et al., 2021). In order to safeguard privacy, data anonymization methods such as k-anonymity (Batista et al., 2022) typically make use of the data generalization operation. However, because it is challenging to model the attackers' prior knowledge, they cannot secure the data's privacy (Jain et al., 2023).

Differential privacy offers a strict and practical definition of privacy. By definition, the calculation results do not provide accurate personal information to attackers. As a result, differential privacy has lately drawn a lot of attention in the PPDM field (Jain and Nandanwar, 2015).

Decision trees are a popular data mining method because of their transparency. However, attackers may be able to obtain personal information by taking advantage of its transparency feature. Several different private decision tree methods have been developed to overcome this issue. The very first decision tree-building technique to incorporate differential privacy was the SuLQ-based ID3 algorithm (Kumar, 2016). Laplacian noise is applied to the query results when determining the information entropy characteristics. On the other hand, the private decision tree's categorization accuracy has decreased dramatically—by a maximum of 30%. The DiffP-ID3and its associated techniques for decision tree classification methods, which use the exponential process to pick the

DOI: https://doi.org/10.52756/ijerr.2024.v44spl.002

splitting characteristics, are proposed by Chamikara et al. (2021) as a solution to the algorithm's shortcoming (Wu et al., 2023).

Using random forest ensemble models is a simple method to lessen the harmful influence of noise on model performance. An effective technique for preserving differential privacy was put out when developing an ID3 classifier (Kiran and Shirisha, 2022). They have demonstrated through experimentation that their suggested approach performs well on both big and small data sets. Vyas et al. (2024) presented an alternative method for building a differentially private random forest. Using their approach, the split attributes are selected at random from the internal nodes instead of according to a predetermined set of criteria. A differentially private ensemble approach was presented (Silva et al., 2019), which can decrease privacy requirements while increasing model accuracy.

Alternative methods concentrate on lessening the randomness brought about by the exponential function. However, the sensitivity of the scoring function and the privacy parameter are related to the randomness generated by the exponential mechanism. When determining the sensitivity of the scoring function, Fletcher and Islam suggested using the local sensitivity as opposed to the global sensitivity (Kumar et al., 2023). Regretfully, a strict definition of differentiated privacy does not exist for local sensitivity. As a result, a method is suggested utilizing the smooth sensitivity to construct a private decision forest (Yu et al., 2023).

The majority of suggested algorithms neglect to consider the impact of privacy parameter allocation because the noise tolerance capability varies depending on the depth of the generated trees. Scientist suggested an adaptive approach for budget allocation that dynamically decides the privacy parameter of every query instead of assigning a fixed parameter for each query (Chamikara et al., 2021). With this method, we may allocate excess privacy to queries susceptible to noise while obtaining reliable and accurate results. Nevertheless, the computation of the privacy parameter will require additional iterations. Differentially private decision trees perform better when the allocation is optimized by adjusting the privacy parameter before every single query. Nevertheless, none of the currently available works provide a customized parameter for privacy that maximizes its utilization. The primary goal of this work is to solve this issue.

Preliminaries

In this section, we first provide a fundamental definition of differential privacy and two alternative methods. Next, we provide the Gini Index, which is employed in the tree-building process to determine the optimal split attributes.

Differential Privacy

Basically, differential privacy is defined as Whether or not a single record is present in the dataset and has minimal impact on the outcome of the computation. Attackers are, therefore, unable to gather precise personal information by looking at the computation results.

Assume that function R has a randomized calculation and that Range(R) represents all of the potential results. If method R is satisfied for any neighboring data sets S1 and S2 having symmetric difference |S1ΔS2| = 1,

$$P(R(S_1) \in S) \leq e^{\in}.P(R(S_2) \in S) \quad (1)$$

It is stated that function R maintains ∈-differential privacy" for any subset S of Range(R). The parameter that regulates the degree of privacy protection is known as the "privacy parameter," and the degree of privacy protection is inversely correlated with its size.

(Sensitivity). Considering an arbitrary function f: D→ V^n n-dimensional vector of real numbers will be produced as V^n for an arbitrary domain D as input. For f, the sensitivity is-

$$\delta f = \max_{S_1,S_2 \; where \; |S_1 \delta S_2|} |f(S_1) - f(S_2)| \quad (2)$$

Given the sensitivity of function f, we can typically achieve ∈-differential privacy" for numerical inquiries by incorporating noise into the query response that is derived from a measured Gaussian distribution.

(The Gaussian Mechanism), for every domain D, given an arbitrary function f: D→ V^n, the function F offers ∈-differential privacy, the Gaussian noise is given as-

$$P(Y) = \frac{1}{\sqrt{2\pi\gamma}} e^{-(Y-\mu)^2/2\gamma^2} \quad (3)$$

Where, P(Y) is the probability density function of Gaussian random variable Y, μ represents the mean value and γ represents the value of variance.

Information Entropy(E_s)

In view of data set analysis, the entropy plays an important role in finding the amount of uncertain data in our data set. The entropy is inversely proportional to the knowledge about a data set. In other words a data set having maximum number of correct predictions has minimum entropy. The following formulae can define the entropy-

$$E_s = \sum_{k=1}^{n} -P_k \log_2 P_k \quad (4)$$

DOI: https://doi.org/10.52756/ijerr.2024.v44spl.002

Gini Impurity(G_I)

Gini impurity is a metric that works similarly to the entropy, but in some situations, the entropy is not effective in revealing the information gain like if there are the same number of true and false predictions, then entropy is computed as 1 while in the same case, the Gini impurity is computed as 0.5 so the value of Gini impurity is less than the entropy value. The Gini impurity leads to a more accurate split operation than the entropy while generating the sub-decision trees. The computation of Gini impurity is preferred over the entropy because the computational complexity of the Gini approach is less than the entropy-based approach as there is no logarithmic function in Gini computation, but sometimes entropy has its own utility because it generates a more balanced decision tree as compared to entropy. So, we can use either of these two techniques depending on the features and size of the data set. The following formulae compute the Gini impurity-

$$G_I = 1 - (P_T^2 - P_F^2) \quad (5)$$

Information Gain (I_g)

The information gain is one of the crucial parameters in the construction of a decision tree. Information gain is inversely proportional to the entropy. The information gain is recursively computed for each generated decision tree, and this process continues until the leaf node of the decision tree has an entropy value of 0. The zero value of entropy indicates that no more splitting is required for constructing the decision tree. The information gain is computed as follows-

$$I_g = E_p - \frac{m_i}{n}(E_c)_i \quad (6)$$

Where, E_p represents the entropy of the parent data set, m_i represents the number of instances in i^{th} child data set, n represents the total number of instances in the parent data set and $(E_c)_i$ represents the entropy of i^{th} child data set.

Experimental Analysis

Original Data Set

The experimental analysis is carried out by applying a booster version of the random forest classification technique. Firstly, we consider the heart disease data set, and the sensitive features of this data set are identified. The sensitive feature is anonymized by applying Gaussian noise. The following table shows some instances of the data set.

Table 1. Header instances of data set.

Age	Sex	CP	trestbps	chol	fbs	restecg	slope	ca	Target
52	1	0	125	212	0	1	2	2	0
53	1	0	140	203	1	0	0	0	0
70	1	0	145	174	0	1	0	0	0
61	1	0	148	203	0	1	2	1	0
62	0	0	138	294	1	1	1	3	0

The above table shows some features and instances of the original data set. Here, one feature is named as a target that shows whether a patient has heart disease or not. Here, we need to consider a feature or set of features that will be considered for the classification, so the feature “cp” is identified as the most important feature for the classification.

Noise Integrated Data Set

Table 2. Noise integrated data set.

Age	Sex	CP	trestbps	chol	restecg	slope	ca	target
50.18	1	0	125	212	1	2	2	0
41.02	1	0	140	203	0	0	0	0
87.55	1	0	145	174	1	0	0	0
61.86	1	0	148	203	1	2	1	0
52.91	0	0	138	294	1	1	3	0

Now, sensitive features like “age” have been identified and have to be anonymized by noise integration. Here, we compute a noise value using the Gaussian noise formulae. The above table shows the anonymized data set after integrating the value of noise. The following noise vector is added to the feature “age” in the data set. In this way, the resultant data set becomes the noisy data set, or in other words, it is known as an anonymized data set.

Noise Integration into Data Set

The proposed algorithm GNIPP is applied in two phases. The noise is integrated with the sensitive attribute in the proposed algorithm's first phase. The noise vector is computed based on the standard Gaussian noise formulae, and then this noise vector is added as a new feature in the data set it is shown in the following table-

DOI: https://doi.org/10.52756/ijerr.2024.v44spl.002

Explanation of Proposed Method (GNIPP)

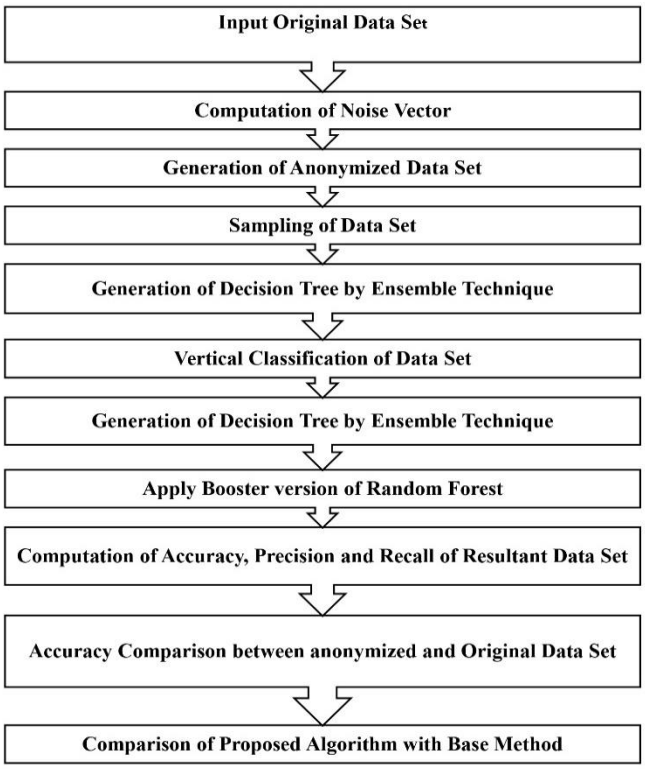

Figure 1. Flow chart of proposed Algorithm.

Table 3. Data Set with Noise Vector.

Age	Sex	CP	trestbps	chol	slope	ca	target	Noise
50.18	1	0	125	212	2	2	0	-0.90
41.02	1	0	140	203	0	0	0	-5.98
87.55	1	0	145	174	0	0	0	8.77
61.86	1	0	148	203	2	1	0	0.43
52.91	0	0	138	294	1	3	0	-4.54

Sampling of Data Set

Table 4. First sample of data set (DS1) after row sampling.

Row No.	Age	Sex	CP	trestbps	chol	Fbs	Rest ecg	Thal ach	Ax ang	Old peak	slope	ca	thal	target
943	64.7396	1	0	120	177	0	1	140	0	0.4	2	0	3	1
60	31.0801	1	1	130	204	0	0	202	0	0.0	2	0	2	1
627	39.1821	1	3	120	231	0	1	182	1	3.8	1	0	3	0
396	72.7102	1	2	180	274	1	0	150	1	1.6	1	0	3	0
373	61.1212	1	1	120	284	0	0	160	0	1.8	1	0	2	0
647	66.3592	0	0	130	303	0	1	122	0	2.0	1	2	2	1
73	55.9917	1	0	140	177	0	1	162	1	0.0	2	1	3	0
364	51.9268	0	1	130	236	0	0	174	0	0.0	1	1	2	0
683	42.2207	1	0	120	177	0	0	120	1	2.5	1	0	3	0

In the experimental analysis, three types of sampling- row-wise sampling, column-wise sampling, and combined sampling- were considered for the noisy data. Three samples have been generated for each type of sampling. The purpose of generating samples from a noisy data set is to feed each sample into a decision tree classifier. A decision tree is generated for each sample of a data set.

We have considered 10 percent rows of the data set to generate three samples in row-wise sampling, while in the case of column-wise sampling, we have generated three samples in which each sample covers 5 percent of columns. All the samples are generated randomly.

Ensemble Technique

It is a technique where more than one model is trained using the same or different algorithms so in this research, a number of decision trees are used to train the collection of models. In this technique, the final prediction is made on the basis of the aggregation of all the predictions generated by all the decision trees.

Decision Tree Generation

The generation of the decision tree is based on the computation of the Gini impurity. The decision tree classifier computes Gini impurity, and the initial value of Gini impurity for the root node is calculated as 0.444. On the basis of this, child nodes are generated where the left child reaches its leaf level as its Gini value becomes 0. The right child's Gini impurity value is computed as 0.375. This process continues and the first decision tree is generated where the total number of nodes in the decision tree is 7 and the depth of the tree is 3. Gini impurity is a default hyper parameter of the decision tree classifier. Gini impurity is a parameter that is used to recognize a feature along which the data set has to be split; hence, each split data set corresponds to the generation of the decision tree.

The decision trees classifier either works on the computation of entropy or Gini impurity, but in this research, Gini impurity has been considered because the computation of Gini impurity is faster as compared to the computation of entropy as the Gini computation considers squared values of probability.

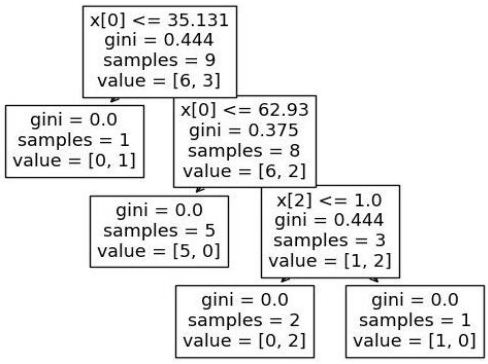

Figure 2. Decision Tree with first row sampling.

DOI: https://doi.org/10.52756/ijerr.2024.v44spl.002

In the process of random forest, the number of decision trees is generated on the basis of row sampling. Figure 2 shows the first decision tree generated as a result of row sampling. The anonymized data set has been taken as input and this data set has been split into 3 sub data sets. The above table 4 shows the first sample of data set which was generated on the basis of 10% of original data set.

The Fig.3 shows the second decision tree generated after row sampling and table 5 shows the randomly generated second subset of the original data set.

The Figure 4 shows the third sample of decision tree that has been generated on the basis of sample subset of the original data set. The subset of the original data set is given in the table 6.

Table 5. Second sample of data set (DS2) after row sampling.

Row No.	Age	Sex	CP	trestbps	chol	Fbs	Rest ecg	Thal ach	Ax ang	Old peak	slope	Ca	thal	target
713	64.7436	0	3	150	226	0	1	114	0	2.6	0	0	3	1
676	57.6434	1	0	130	253	0	1	144	1	1.4	2	1	2	1
911	59.3202	0	1	136	319	1	0	152	0	0.0	2	2	3	0
782	64.4339	0	0	130	303	0	1	122	0	2.0	1	2	3	0
313	75.9737	0	1	120	269	0	0	121	1	1.8	2	1	2	0
635	53.1396	0	0	130	264	0	0	143	0	0.2	1	0	2	1
584	56.7620	1	0	132	353	0	1	132	1	0.4	1	1	3	0
267	68.4241	1	0	120	237	0	1	71	0	1.2	1	1	2	0
359	49.2937	0	2	128	216	0	0	115	0	0.0	2	0	3	0

Table 6. Third sample of data set (DS2) after row sampling.

Row No.	Age	Sex	CP	trestbps	chol	Fbs	Rest ecg	Thal ach	Ax ang	Old peak	slope	ca	thal	target
687	57.8209	1	0	125	300	0	0	171	0	0.0	2	2	3	0
786	64.2504	1	0	125	254	1	1	163	0	0.2	1	2	3	0
318	66.4098	1	0	140	177	0	1	162	1	0.0	2	1	3	0
997	52.1343	1	0	120	188	0	1	113	0	1.4	1	1	3	0
1019	47.6830	1	0	112	204	0	1	143	0	0.1	2	0	2	1
521	59.1519	1	1	125	220	0	1	144	0	0.4	1	4	3	1
215	55.4739	1	1	130	266	0	1	171	0	0.6	2	0	2	1
649	43.9495	0	1	130	234	0	0	175	0	0.6	1	0	2	1
116	63.4143	1	0	130	254	0	0	147	0	1.4	1	1	3	0

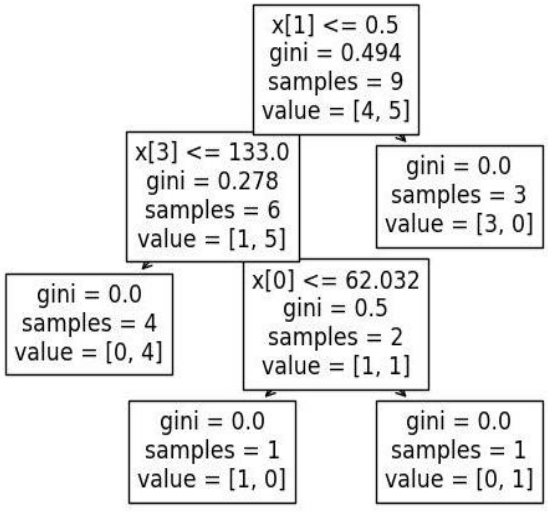

Figure 3. Decision Tree with second-row sampling.

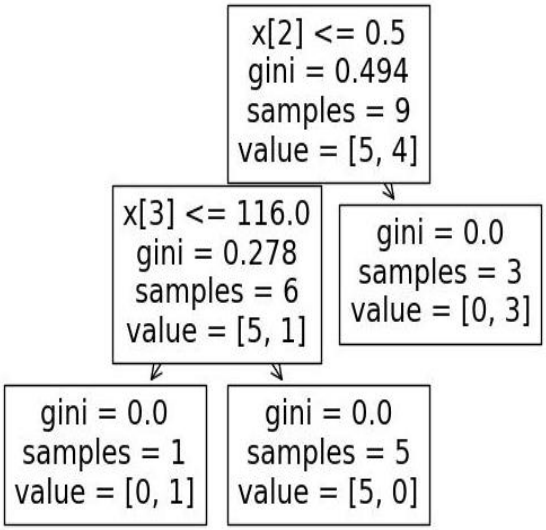

Figure 4. Decision Tree with third-row sampling.

DOI: https://doi.org/10.52756/ijerr.2024.v44spl.002

Result and Discussion

The analysis of proposed algorithm (GNIPP) has been done on various metrics like precision, recall, accuracy and F-1 score. Also the experimental analysis demonstrates the effective computation of various parameters like Gaussian noise computation, information entropy, Gini impurity, information gain and tuning of hyper parameters.

Feature Importance Value

Various features are assessed based on their importance during the classification using random forest. Hence, the feature importance vector is computed as-

Table 7. Feature Importance Value.

Name of Feature	Value representing feature importance
age	0. 08877226
Sex	0. 03775317
CP	0. 1567776
Trestbps	0. 05804852
Chol	0. 05636884
Fbs	0. 00938654
Restecg	0. 01518357
Thalach	0. 08071837
Exang	0. 06114553
Oldpeak	0. 12830501
Slope	0. 05535863
Ca	0.05535863
Thal	0.12655215
target	0.12562981

Performance Metrics

Various performance metrics are used to analyze the predictions made by the ensemble technique-based random forest model.

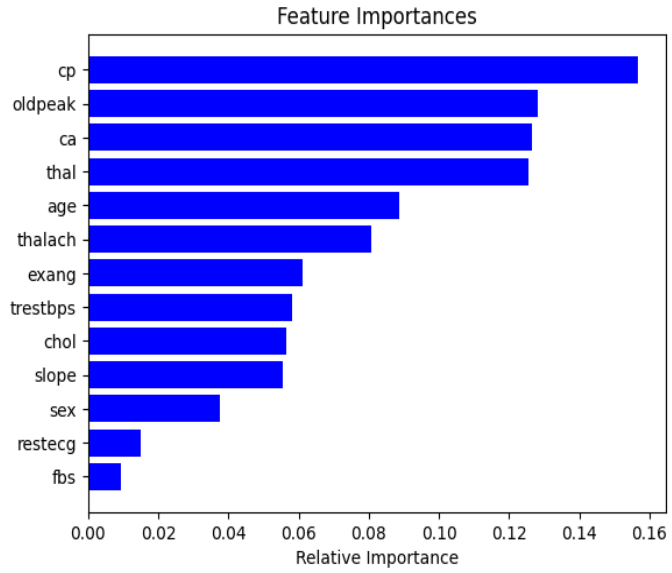

Figure 5. Feature Importance.

Accuracy Computation

Accuracy is one of the best metrics for analyzing a model's performance. Here, the accuracy score represents the total number of correct predictions out of the total prediction. The accuracy score for our suggested algorithm GNIPP is 0.961089.

Compared to the base method BDPT, which computes the model's accuracy at around 0.78, our suggested technique GNIPP evaluates better accuracy than BDPT. Figure 5 shows the accuracy of the comparison of various privacy preservation models.

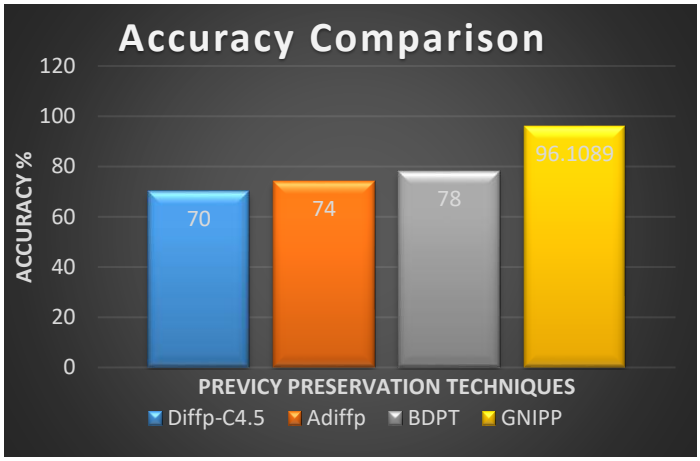

Figure 6. Accuracy Comparison.

Accuracy Computation

$$\begin{bmatrix} 119 & 9 \\ 1 & 128 \end{bmatrix}$$

Another metric considered for the performance measurement is the confusion matrix. In the above confusion matrix, the true positive predictions are 119 out of 127 predictions, while true negative predictions are 128 out of 129, so the overall performance is much better than the existing BDPT approach.

Performance Metrics Comparison

Various performance metrics, such as F-1 score, recall, support, and precision, have been computed to evaluate the proposed technique GNIPP. Figure 7 shows the comparison among various performance metrics.

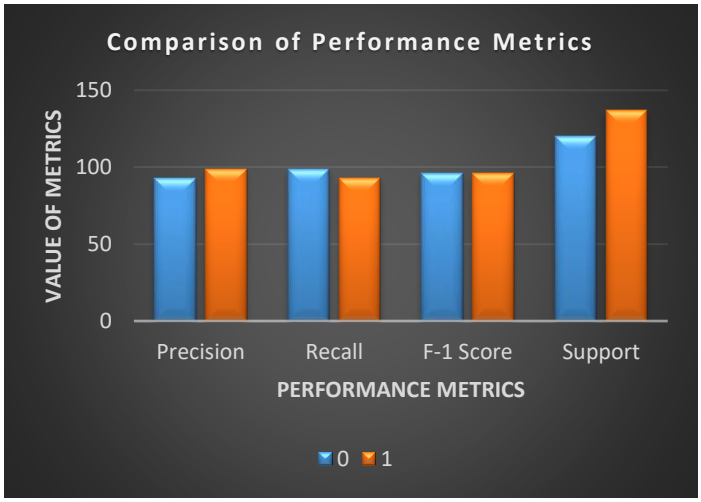

Figure 7. Performance Metrics Comparison.

DOI: https://doi.org/10.52756/ijerr.2024.v44spl.002

Conclusion and Future Work

In this study, we introduce the GNIPP technique, which enables data miners to build a highly useful decision forest while maintaining privacy. Our privacy parameter allocation technique gives the nodes more privacy when compared to previous works, which is important for leaf nodes.

As the tree gets deeper, there are fewer and fewer sample instances of the nodes, indicating that the leaf nodes are more susceptible to the noise added to preserve privacy when building the decision tree. To integrate the Gaussian noise, leaf nodes supply less noise to balance the noise and true counts. Furthermore, our selective aggregation approach enables us to choose trees that can support the ultimate performance for aggregation while aggregating them into a forest. Lastly, we carry out comprehensive experiments to demonstrate that the suggested GNIPP approach can accomplish a more favorable trade-off between privacy and utility.

References

Batista, E., Martínez-Ballesté, A., & Solanas, A. (2022). Privacy-preserving process mining: A microaggregation-based approach. *Journal of Information Security and Applications*, *68*, 103235. https://doi.org/10.1016/j.jisa.2022.103235.

Chamikara, M., Bertok, P., Khalil, I., Liu, D., & Camtepe, S. (2021). PPAAS: Privacy Preservation as a service. *Computer Communications*, *173*, 192–205. https://doi.org/10.1016/j.comcom.2021.04.006.

Chamikara, M., Bertok, P., Liu, D., Camtepe, S., & Khalil, I. (2019). Efficient privacy preservation of big data for accurate data mining. *Information Sciences*, *527*, 420–443. https://doi.org/10.1016/j.ins.2019.05.053

Jain, P., & Nandanwar, S. (2015). Securing the clustered database using data modification Technique. https://doi.org/10.1109/cicn.2015.331.

Jain, P., Shakya, H. K., & Lala, A. (2023). Advanced privacy Preserving model for smart healthcare using deep learning. *2023 6th International Conference on Contemporary Computing and Informatics (IC3I)*, Gautam Buddha Nagar, India. *39*, 2368–2372. https://doi.org/10.1109/ic3i59117.2023.10397954

Jain, P., & Thada, V. (2024). Securing the Data Using an Efficient Machine Learning Technique. *International Journal of Experimental Research and Review*, *40*(Spl Volume), 217-226. https://doi.org/10.52756/ijerr.2024.v40spl.018

Karmore, S. P., & Mahajan, A. R. (2016). New Approach for Testing and Providing Security Mechanism for Embedded Systems. *Procedia Computer Science*, *78*, 851–858. https://doi.org/10.1016/j.procs.2016.02.073.

Kiran, A., & Shirisha, N. (2022). K-Anonymization approach for privacy preservation using data perturbation techniques in data mining. *Materials Today Proceedings*, *64*, 578–584. https://doi.org/10.1016/j.matpr.2022.05.117.

Kulkarni, Y. R., Jagdale, B., & Sugave, S. R. (2022). Optimized key generation-based privacy preserving data mining model for secure data publishing. *Advances in Engineering Software*, *175*, 103332. https://doi.org/10.1016/j.advengsoft.2022.103332.

Kumar, G. S., Premalatha, K., Maheshwari, G. U., & Kanna, P. R. (2023). No more privacy Concern: A privacy-chain based homomorphic encryption scheme and statistical method for privacy preservation of user's private and sensitive data. *Expert Systems with Applications*, *234*, 121071. https://doi.org/10.1016/j.eswa.2023.121071.

Kumar, G. S., Premalatha, K., Maheshwari, G. U., Kanna, P. R., Vijaya, G., & Nivaashini, M. (2023). Differential privacy scheme using Laplace mechanism and statistical method computation in deep neural network for privacy preservation. *Engineering Applications of Artificial Intelligence*, *128*, 107399. https://doi.org/10.1016/j.engappai.2023.107399.

KumarTripathi, K. (2016). Discrimination Prevention with Classification and Privacy Preservation in Data mining. *Procedia Computer Science*, *79*, 244–253. https://doi.org/10.1016/j.procs.2016.03.032.

Ling, J., Zheng, J., & Chen, J. (2024). Efficient Federated Learning Privacy Preservation Method with Heterogeneous Differential Privacy. *Computers & Security*, *139*, 103715. https://doi.org/10.1016/j.cose.2024.103715.

Rafiei, M., & Van Der Aalst, W. M. (2021). Group-based privacy preservation techniques for process mining. *Data & Knowledge Engineering*, *134*, 101908. https://doi.org/10.1016/j.datak.2021.101908.

Silva, J., Cubillos, J., Villa, J. V., Romero, L., Solano, D., & Fernández, C. (2019). Preservation of confidential information privacy and association

rule hiding for data mining: a bibliometric review. *Procedia Computer Science*, *151*, 1219–1224. https://doi.org/10.1016/j.procs.2019.04.175.

Talat, R., Obaidat, M. S., Muzammal, M., Sodhro, A. H., Luo, Z., & Pirbhulal, S. (2019). A decentralised approach to privacy preserving trajectory mining. *Future Generation Computer Systems*, *102*, 382–392. https://doi.org/10.1016/j.future.2019.07.068.

Vyas, P. J. V. T. S. K. (2024). *Achieving highest privacy preservation using efficient Machine Learning Technique*. https://ijisae.org/index.php/IJISAE/article/view/5434.

Vyas, S. K., Karmore, S., & Jain, P. (2024). *A Privacy-Preserving Data Mining Approach in Multi-Dimensional Data Set based on the Random and Cumulative Integrated Noise*. https://www.ijisae.org/index.php/IJISAE/article/view/4892.

Vyas, S., & Karmore, S. (2022). Design and Development of Privacy Preservation Approach in Data Mining: A literature review paper. *Proceedings of the International Conference on Innovative Computing & Communication* (ICICC), 2022. https://doi.org/10.2139/ssrn.4021313.

Wu, H., Ran, R., Peng, S., Yang, M., & Guo, T. (2023). Mining frequent items from high-dimensional set-valued data under local differential privacy protection. *Expert Systems with Applications*, *234*, 121105. https://doi.org/10.1016/j.eswa.2023.121105.

Yu, S., Wei, Z., Sun, G., Zhou, Y., & Zang, H. (2023). A double auction mechanism for virtual power plants based on blockchain sharding consensus and privacy preservation. *Journal of Cleaner Production*, *436*, 140285. https://doi.org/10.1016/j.jclepro.2023.140285.

How to cite this Article:

Shailesh Kumar Vyas and Swapnili Karmore (2024). A Privacy-Preserving Data Mining Through Comprehensive GNIPP Approach in Sensitive Data Sets. *International Journal of Experimental Research and Review*, *44*, 11-19.

DOI : https://doi.org/10.52756/ ijerr.2024.v44spl.002