FOREX Publication
Open Access | Rapid and quality publishing

International Journal of Computer Science Research (IJCSR)
Survey Article | Volume 2, Issue 4 | Pages 130-137 | ISSN XXXX-XXXX

A Privacy Preserving Data Mining through Comprehensive GNIPP Approach in Sensitive Data Sets

Shailesh Kumar Vyas[1*] and Swapnili Karmore[2]
[1] *shailesh.pk.29@gmail.com*
[2] *swapnilikarmore@gmail.com*

***Correspondence:** *shailesh.pk.29@gmail.com*

ABSTRACT- The quick growth of methods for analysing 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 in this study a Gaussian Noise Integrated Privacy Preservation (GNIPP). To address the issue of personal information breach, 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.

Keywords: Decision tree, privacy preservation, GNIPP, Gaussian noise, random forest, Gini impurity, Accuracy

ARTICLE INFORMATION

Author(s): Shailesh Kumar Vyas and Swapnili Karmore;
Received: 17/11/2023; **Accepted**: 15/12/2023; **Published**: 30/12/2023;
e-ISSN: XXXX-XXXX;
Paper Id: IJCSR-020406;
Citation: 10.37391/IJCSR.020406

Publisher's Note: FOREX Publication stays neutral with regard to Jurisdictional claims in Published maps and institutional affiliations.

1. 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 on a regular basis. The revolutionary technique known as data mining is able to give more personalised and improved services in a variety of industries, including online search, healthcare, and medicine [1]. 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 of patient electronic medical records may uncover information that is helpful to medical therapy, such as the underlying relations between different diseases [2]. However, this method may potentially expose patient 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 [5].

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 [6]. 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.

A strict concept of privacy that opposes individual privacy disclosure is called differential privacy [19]. According to this definition, the outcome of the data sets' computation process is unaffected by the modification of a single record in the sets of data. 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 [9], a variety of data mining techniques, including clustering [17], classification [10], as well as deep learning [12], 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. The query answer is calculated by DPDS, 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.

FOREX Publication
Open Access | Rapid and quality publishing
International Journal of Computer Science Research (IJCSR)
Survey Article | Volume 2, Issue 4 | Pages 130-137 | ISSN XXXX-XXXX

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. [15]. the alternative direction is ensemble [16]. Nevertheless, the majority of approaches have neglected the influence that privacy allocation brings, meaning that nodes in various levels have varying noise tolerance capacities. A unique technique to privacy allocation that dynamically sets each query's privacy parameter has only been proposed in one recent article [18]. Nevertheless, figuring out how much privacy preserved during each inquiry requires additional value of privacy parameter.

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 allocation of differential privacy parameter, we create a sensible privacy parameter allocation technique that allots varying amounts of privacy value to nodes in various levels. We often allot bigger value of privacy parameter to the nodes of decision tree placed deeper since the true count of such nodes is more vulnerable to noise.

• To increase the ensemble model's capacity for generalization and precision in prediction, we suggest a selective ensemble technique. 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.

2. LITERATURE REVIEW

There are various methods available today to protect data privacy, including differential privacy and data anonymization [11]. In order to safeguard privacy, data anonymization methods such as k-anonymity [20] typically make use of the data generalization operation. However, because it is challenging to model the attackers' prior knowledge, they are unable to secure the privacy of the data [5].

A strict and practical definition of privacy is offered by differential privacy. By definition, the results of the calculation do not provide accurate personal information to attackers. As a result, differential privacy has lately drawn a lot of attention in the PPDM field [3].

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 differently 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, which was proposed by Blum et al. [10]. When determining the information entropy characteristics, Laplacian noise is applied to the query results. 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 splitting characteristics, are proposed in [9] as a solution to the algorithm's shortcoming in [13].

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 when developing an ID3 classifier was put out by Freidman and Schuster [7]. They have demonstrated through experimentation that their suggested approach performs well on both big and small data sets. In [4], Jagannathan et al. 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 by Rana et al. [8], 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 sensitivity of scoring function, Fletcher and Islam suggested using the local sensitivity as opposed to the global sensitivity [15]. Regretfully, a strict definition of differentiated privacy does not exist for local sensitivity. As a result, Fletcher and Islam moreover suggested utilizing the smooth sensitivity to construct a private decision forest [14].

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. N. Borhan et al. suggested an adaptive approach for budget allocation that decides the privacy parameter of every query dynamically, as opposed to assigning a fixed parameter for each query [18]. With this method, we may allocate excess privacy to queries that are susceptible to noise and still obtain reliable accuracy 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

FOREX Publication
Open Access | Rapid and quality publishing
International Journal of Computer Science Research (IJCSR)
Survey Article | Volume 2, Issue 4 | Pages 130-137 | ISSN XXXX-XXXX

privacy that maximizes its utilization. This work's primary goal is to solve this issue.

3. PRELIMINARIES

We first provide a fundamental definition of differential privacy in this section, along with 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 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 neighbouring 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→ $\boldsymbol{V^n}$ n-dimensional vector of real numbers will be produced as $\boldsymbol{V^n}$ 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→ $\boldsymbol{V^n}$, 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, $\boldsymbol{\mu}$ represents the mean value and $\boldsymbol{\gamma}$ represents the value of variance.

Information Entropy (E_s)

In view of data set analysis, the entropy plays an important role to find 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 entropy can be defined by the following formulae-

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

Gini Impurity (G_I)

Gini impurity is a metric that works similar to the entropy but in some situations0 the entropy is not effective to reveal the information gain like if there are 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 the more accurate split operation as compared to the entropy while generating the sub decision trees. The computation of Gini impurity is preferred over the entropy because the computational complexity of 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 more balanced decision tree as compared to entropy. The Gini impurity is computed by the following formulae-

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

Information Gain (I_g)

The information gain is one of the crucial parameter in the construction of 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 decision tree has entropy value as 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, $\boldsymbol{E_p}$ represents the entropy of the parent data set, $\boldsymbol{m_i}$ represents the number of instances in $\boldsymbol{i^{th}}$ child data set, n represents the total number of instances in parent data set and $(\boldsymbol{E_c})_{\boldsymbol{i}}$ represents the entropy of $\boldsymbol{i^{th}}$ child data set.

The information gain can be computed based on either entropy or Gini impurity. The Gini impurity based computation is always more accurate as compared to the entropy based computation.

4. EXPERIMENTAL ANALYSIS

Original Data Set

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

FOREX Publication
Open Access | Rapid and quality publishing
International Journal of Computer Science Research (IJCSR)
Survey Article | Volume 2, Issue 4 | Pages 130-137 | ISSN XXXX-XXXX

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 original data set here one feature is named as 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 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 the sensitive feature like "age" is identified that has to be anonymized by noise integration. Here we have computed a noise value by using 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 anonymized data set.

Noise Integration into Data Set

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

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

Explanation of Proposed Method (GNIPP)

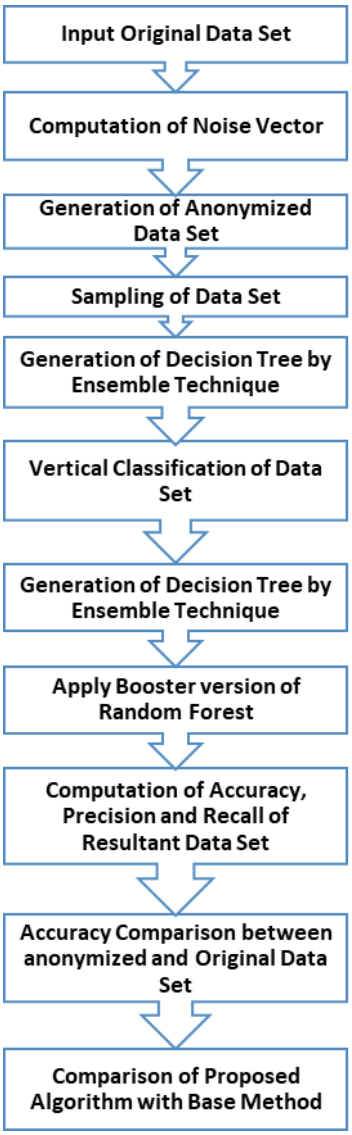

Figure.1 Flow chart of proposed Algorithm

Sampling of Data Set

Three types of sampling has been considered on the noisy data in the experimental analysis that is row wise sampling, column wise sampling and combined sampling. Three samples have been generated for each type of sampling. The purpose of generating samples of noisy data set is to feed each sample data

FOREX Publication
Open Access | Rapid and quality publishing
International Journal of Computer Science Research (IJCSR)
Survey Article | Volume 2, Issue 4 | Pages 130-137 | ISSN XXXX-XXXX

set into decision tree classifier. Decision tree is generated for each sample of data set.

We have considered 10 percent rows of the data set to generate three samples in row wise sampling while in 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.

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

Ensemble Technique

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

Decision Tree Generation

The generation of decision tree is based on the computation of Gini impurity. The computation of Gini impurity is done by the decision tree classifier and the initial value of Gini impurity for the root node is taken as 0.444. On the basis of this child nodes are generated where the left child reaches to 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 first decision tree is generated where the total number of nodes in the decision tree is 7 and 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 splitted hence each splitted data set corresponds to the generation of 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.

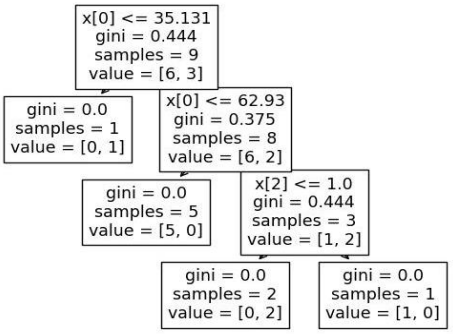

Figure.2 Decision Tree with first row sampling

In the process of random forest, the number of decision trees are 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 an input and this data set has been splitted 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.

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

FOREX Publication
Open Access | Rapid and quality publishing
International Journal of Computer Science Research (IJCSR)
Survey Article | Volume 2, Issue 4 | Pages 130-137 | ISSN XXXX-XXXX

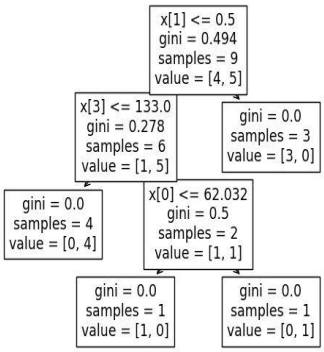

Figure.3 Decision Tree with second row sampling

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

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

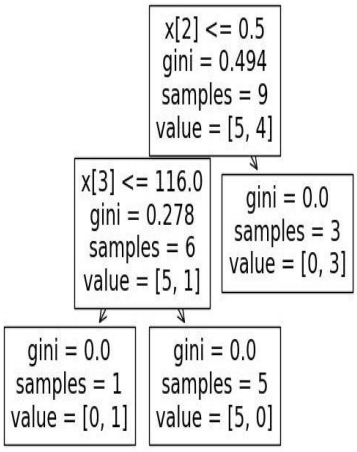

Figure.4 Decision Tree with third row sampling

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

5. 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, tuning of hyper parameters.

Feature Importance Value

During the classification using random forest, various features are assessed based on their importance 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 used to analyse the predictions that have been made by the ensemble technique based random forest model.

FOREX Publication
Open Access | Rapid and quality publishing
International Journal of Computer Science Research (IJCSR)
Survey Article | Volume 2, Issue 4 | Pages 130-137 | ISSN XXXX-XXXX

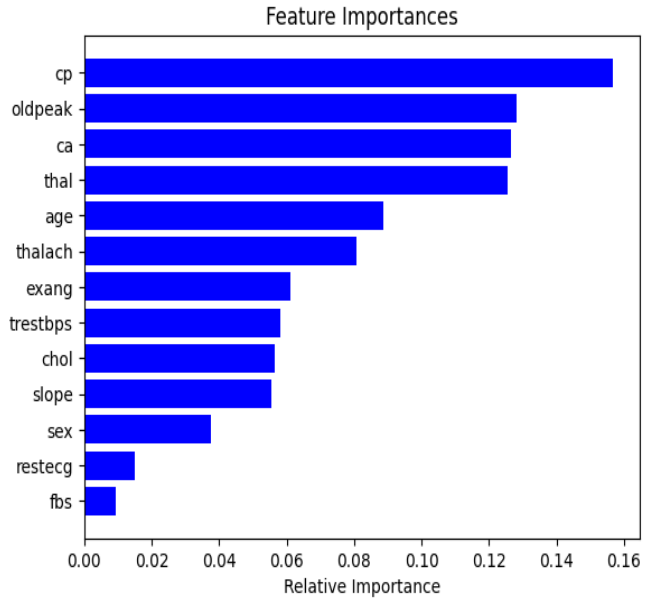

Figure.5 Feature Importance

Accuracy Computation

The accuracy is one of the best metric to analyze the performance of the model. 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.

As compared to the base method BDPT that computes the accuracy of the model as around 0.78, hence our suggested technique GNIPP evaluates better accuracy rather than the accuracy of BDPT. The following fig 5 shows the comparison of accuracy among various privacy preservation models.

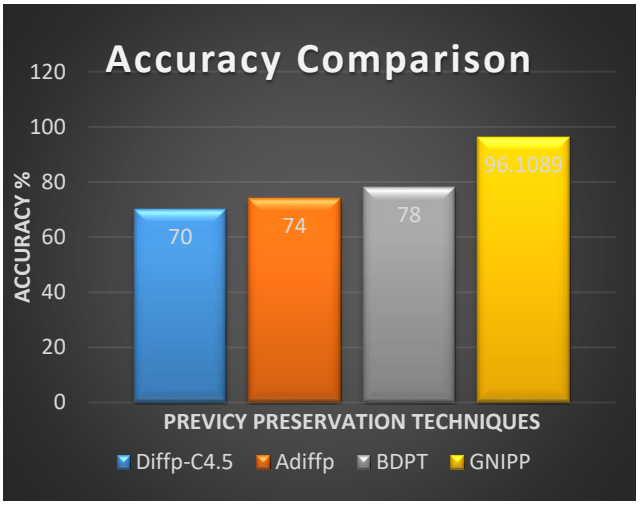

Figure.6 Accuracy Comparison

Accuracy Computation

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

Another metric has been considered for the performance measurement that is the confusion matrix. In 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 approach that is BDPT.

Performance Metrics Comparison

As an evaluation of proposed technique GNIPP, various performance metrics have been computed like F-1 score, Recall, Support and Precision. The following *figure 7* shows the comparison among various performance metrics.

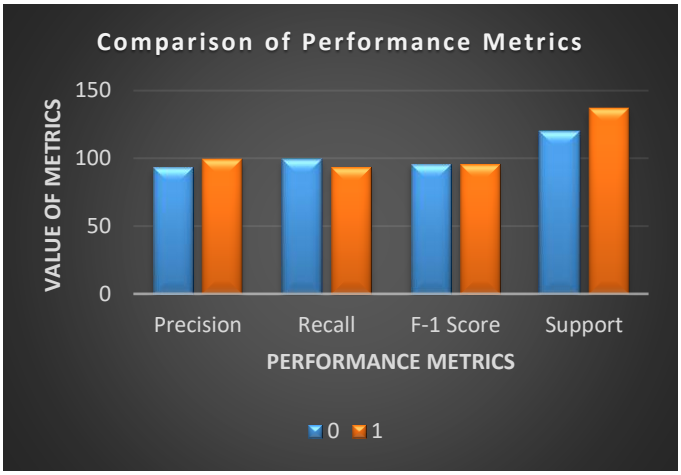

Figure.7 Performance Metrics Comparison

6. 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. For the purpose of integrating the Gaussian noise, less noise is supplied in leaf nodes in order 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

[1] S. Vyas and S. Karmore, "Design and Development of Privacy Preservation Approach in Data Mining: A literature review paper," Social Science Research Network, Jan. 2022, doi: 10.2139/ssrn.4021313.

[2] S. K. Vyas, S. Karmore, and P. Jain, "A Privacy-Preserving Data Mining Approach in Multi-Dimensional Data Set based on the Random and Cumulative Integrated Noise," Feb. 23, 2024. https://www.ijisae.org/index.php/IJISAE/article/view/4892

International Journal of Computer Science Research (IJCSR)
Survey Article | Volume 2, Issue 4 | Pages 130-137 | ISSN XXXX-XXXX

[3] P. Jain and S. Nandanwar, Securing the clustered database using data modification technique. IEEE, 2015. doi: 10.1109/cicn.2015.331.

[4] P. J. V. T. S. K. Vyas, "Achieving highest privacy preservation using efficient machine learning technique," Mar. 16, 2024. https://ijisae.org/index.php/IJISAE/article/view/5434

[5] P. Jain, H. K. Shakya and A. Lala, "Advanced Privacy Preserving Model for Smart Healthcare Using Deep Learning," 2023 6th International Conference on Contemporary Computing and Informatics (IC3I), Gautam Buddha Nagar, India, 2023, pp. 2368-2372, doi: 10.1109/IC3I59117.2023.10397954.

[6] S. Karmore and A. Mahajan, "New Approach for Testing and providing Security Mechanism for Embedded Systems," Procedia Computer Science, vol. 78, pp. 851–858, Jan. 2016, doi: 10.1016/j.procs.2016.02.073.

[7] A. Kiran and N. Shirisha, "K-Anonymization approach for privacy preservation using data perturbation techniques in data mining," Materials Today: Proceedings, vol. 64, pp. 578–584, Jan. 2022, doi: 10.1016/j.matpr.2022.05.117.

[8] J. Silva, J. Cubillos, J. V. Villa, L. Romero, D. Solano, and C. N. Fernández, "Preservation of confidential information privacy and association rule hiding for data mining: a bibliometric review," Procedia Computer Science, vol. 151, pp. 1219–1224, Jan. 2019, doi: 10.1016/j.procs.2019.04.175.

[9] M. a. P. Chamikara, P. Bertók, D. Liu, S. Camtepe, and I. Khalil, "Efficient privacy preservation of big data for accurate data mining," Information Sciences, vol. 527, pp. 420–443, Jul. 2020, doi: 10.1016/j.ins.2019.05.053.

[10] K. KumarTripathi, "Discrimination Prevention with Classification and Privacy Preservation in Data mining," Procedia Computer Science, vol. 79, pp. 244–253, Jan. 2016, doi: 10.1016/j.procs.2016.03.032.

[11] M. Rafiei and W. M. P. Van Der Aalst, "Group-based privacy preservation techniques for process mining," Data and Knowledge Engineering, vol. 134, p. 101908, Jul. 2021, doi: 10.1016/j.datak.2021.101908.

[12] G. S. Kumar, K. Premalatha, G. Maheshwari, and P. R. Kanna, "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, vol. 234, p. 121071, Dec. 2023, doi: 10.1016/j.eswa.2023.121071.

[13] H. Wu, R. Ran, S. Peng, M. Yang, and T. Guo, "Mining frequent items from high-dimensional set-valued data under local differential privacy protection," Expert Systems With Applications, vol. 234, p. 121105, Dec. 2023, doi: 10.1016/j.eswa.2023.121105.

[14] S. Yu, Z. Wei, G. Sun, Y. Zhou, and H. Zang, "A double auction mechanism for virtual power plants based on blockchain sharding consensus and privacy preservation," Journal of Cleaner Production, vol. 436, p. 140285, Jan. 2024, doi: 10.1016/j.jclepro.2023.140285.

[15] G. S. Kumar, K. Premalatha, G. Maheshwari, P. R. Kanna, G. Vijaya, and M. Nivaashini, "Differential privacy scheme using Laplace mechanism and statistical method computation in deep neural network for privacy preservation," Engineering Applications of Artificial Intelligence, vol. 128, p. 107399, Feb. 2024, doi: 10.1016/j.engappai.2023.107399.

[16] Y. R. Kulkarni, B. N. Jagdale, and S. R. Sugave, "Optimized key generation-based privacy preserving data mining model for secure data publishing," Advances in Engineering Software, vol. 175, p. 103332, Jan. 2023, doi: 10.1016/j.advengsoft.2022.103332.

[17] J. Ling, J. Zheng, and J. Chen, "Efficient Federated Learning Privacy Preservation Method with Heterogeneous Differential Privacy," Computers & Security, vol. 139, p. 103715, Apr. 2024, doi: 10.1016/j.cose.2024.103715.

[18] M. a. P. Chamikara, P. Bertók, I. Khalil, D. Liu, and S. Camtepe, "PPAAS: Privacy Preservation as a service," Computer Communications, vol. 173, pp. 192–205, May 2021, doi: 10.1016/j.comcom.2021.04.006.

[19] R. Talat, M. S. Obaidat, M. Muzammal, A. H. Sodhro, Z. Luo, and S. Pirbhulal, "A decentralised approach to privacy preserving trajectory mining," Future Generation Computer Systems, vol. 102, pp. 382–392, Jan. 2020, doi: 10.1016/j.future.2019.07.068.

[20] E. Batista, A. Martínez-Ballesté, and A. Solanas, "Privacy-preserving process mining: A microaggregation-based approach," Journal of Information Security and Applications, vol. 68, p. 103235, Aug. 2022, doi: 10.1016/j.jisa.2022.103235.