International Advance Journal of Engineering, Science and Management (IAJESM)
ISSN -2393-8048, July-December 2022, Submitted October 2022, iajesm2014@gmail.com

Enhancing Data Privacy in Predictive Modeling: A Comprehensive Approach Using Weight of Evidence and Information Value

Tarun Kumar, Researcher, Department of Computer Science & Engineering, Glocal University, Saharanpur (Uttar Pradesh)
Dr. Bhupendra Kumar, Professor, Department of Computer Science & Engineering, Glocal University, Saharanpur (Uttar Pradesh)

Abstract

This paper presents an innovative approach to Privacy-Preserving Data Mining (PPDM) by integrating Intuitionistic Fuzzy Gaussian Membership Functions with Statistical Transformations. The proposed methodology seeks to balance the imperative need for data privacy with the preservation of data utility for meaningful analysis. Intuitionistic fuzzy sets, characterized by both membership and non-membership values, are employed to effectively manage imprecise and uncertain data. This is complemented by Gaussian membership functions, which facilitate smooth data transformation into a continuous distribution, enhancing its applicability for tasks such as classification and clustering. Additionally, statistical transformations are applied to perturb sensitive data, ensuring privacy while maintaining the data's statistical properties. The result is a novel PPDM technique that addresses the dual challenge of protecting sensitive information and enabling accurate data analysis, making it particularly relevant for sectors where data security and utility must coexist.

Keywords: Privacy-Preserving Data Mining, Intuitionistic Fuzzy Gaussian Membership Functions

1. INTRODUCTION

Geometric change methods have made it possible to use statistical analysis to protect the privacy of data. Statistics like Information Value (IV) and Weight of Evidence (WOE) are used to figure out credit risk. We use WOE and IV as divergence measures to look at credit scores (Guoping Zeng, 2013). WOE is used in supervised learning to change the ways that people show they can do something. It is also often used to bind property values. The author Eftim Zdravevski (2011) says that classification methods, like WOE and IV, are used to hide data. These methods have a big effect on the ability to change original data. This study shows the Statistical Transformation with Intuitionistic Fuzzy (STIF) method for changing data, which can help with the problems of Privacy-Preserving Data Mining (PPDM). In the STIF algorithm, there is an intuitionistic fuzzy Gaussian membership function. This algorithm also uses the statistical methods of WOE and IV. It has been used to look closely at three standard datasets: lung cancer, adult income, and bank marketing. The PPDM method, which is based on reconstructing aggregate-level distributions and changing data, lets you mine data while protecting your privacy. In real life, expectation maximisation methods are used in distribution reconstruction to keep the rate of information loss manageable (Dakshi Agrawal & Charu C. Aggarwal, 2001). In PPDM, data manipulation is done with fuzzy logic, a type of many-valued logic that uses any real number between 0 and 1 to show the degree of truth. Since the fuzzy set was first presented, many additions have been made, and the Intuitionistic Fuzzy Set (IFS) is one of them. The classic fuzzy set (FS) is shown as $\{\langle x, \mu_A(x) \mid x \in E\}$, according to Krussimart T. Atanassov (1986). The improved fuzzy set (IFS) is shown as $\{\langle x, \mu_A(x), 1 - \mu A(x) \mid x \in E\}$. The formula $\pi(x) = 1 - \mu(x) - v(x)$ (Krassimir T. Atanassov, 2017), which is also known as the "hesitating index," shows the confusion or lack of certainty that IFS adds. Instead of using the straight complement rule like in most fuzzy sets, IFS's hesitation index provides a more sophisticated way to deal with uncertainty by taking into account the fact that things may not be clear all the time (Radhika et al., 2016). This index, which is made up of membership and non-membership functions, can help if it's hard to figure out how much an element belongs to a set (Gang Qian et al., 2013). The intuitionistic fuzzy method used in this study makes data perturbation better. WOE and IV have been used for a long time in logistic regression to solve classification problems (Soloshenko, 2015). At first, these models mess up the user's "quasi-identifiers." To improve data privacy, intuitionistic fuzzy methods are used to process the changed values further. This two-layer perturbation guarantees that information extraction before and after the perturbation will be the same. Figure 1.1 shows how the STIF method can be used to change data.

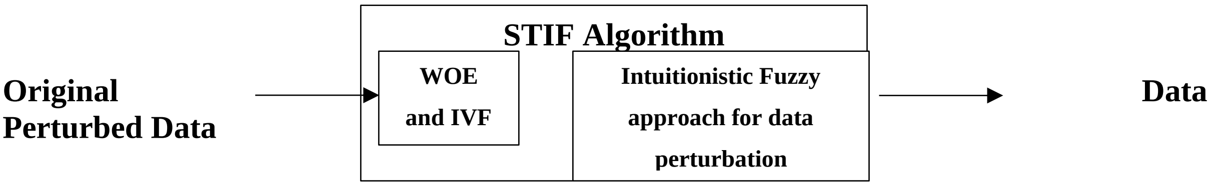


Figure 1.1: Data perturbation using STIF algorithm
https://www.researchgate.net/figure/Data-perturbation-using-STIF-algorithm_fig1_370181638

2. STATISTICAL CONVERSION USING WEIGHT OF EVIDENCE (WOE) AND INFORMATION VALUE (IV)

The formulas for calculating WOE and IV are:

WOE Calculation: $WOE = In(\frac{\% \ of \ non-events}{\% \ of \ events})$ (1.1)

IV Calculation: IV=∑(% of events − % of non-events)×WOE (1.2)

Adjusted WOE $= In(\frac{\text{Number of non-events in a bin+0.5/Number of events}}{\text{Number of events in a bin+0.5/Number of non-events}})$ (1.3)

WOE Calculation for Bin i (1.4): WOE $_{x \in Range} In(\frac{\text{\% of class = 1 where } x \in Range_i}{\text{\% of class = 0 where } x \in Range_i})$
(1.4)

The overall IV is computed by summing up the WOE values across all bins, weighted by the difference between the percentage of events and non-events in each bin.

Table 1.1 WOE and IV Computation for Numerical Attribute

Range	Bins	Count	class = 0	class = 1	% of class = 0	% of class =1	WOE	IV
0-500	1	36260	2542	33718	7.01	92.99	-2.59	222.69
501-1000	2	3972	1532	2440	38.57	61.43	-0.47	10.74
1001-1500	3	750	437	313	58.27	41.73	0.33	5.46
1501-2000	4	146	91	55	62.33	37.67	0.51	12.58
2001-2500	5	36	25	11	69.44	30.56	0.82	31.88
2501-3000	6	9	7	2	77.78	22.22	1.25	69.45
3001-3500	7	9	3	6	33.33	66.67	-0.69	23.01
3501-4000	8	5	3	2	60	40	0.41	8.2
>4000	9	3	2	1	66.66	33.34	0.69	22.91
Total			**4642**	**36548**				

2.1 Procedure for Determining the Values of the Classified Attribute's WoE and IV: Data division into X portions (bins) is unnecessary for categorical attributes. To calculate WoE and IV for categorical values, it is necessary to carry out all the operations from step-2 to step-5 in the exact sequence as described in the pseudo-code. The adult income dataset includes an occupation attribute that should be considered. The occupation attribute has fourteen different types of occupations. The class label was used to derive the value of event and *non*_event. Consequently, the value of WOE and *IV* for categorical values has been determined by computing the percentage of event and *non*_event. For the occupation (categorical) attribute, Table 1.2 shows the statistical transformation value using WOE and *IV*.

Table 1.2 WOE and IV Computation for Categorical Attribute

Occupation	Count	class = 0	class= 1	% of class = 0	% of class =1	WOE	IV
Administrator - clerical	3770	3263	507	86.55	13.45	1.86	135.97
Armed Forces	9	8	1	88.89	11.11	2.08	161.78
Craft-Repair	4099	3170	929	77.34	22.66	1.23	67.25
Executive- Manager	4066	2098	1968	51.60	48.40	0.07	0.22
Farming-Fishing	994	879	115	88.43	11.57	2.03	156.03
Handlers- Cleaners	1370	1284	86	93.72	6.28	2.7	236.10

Machine-opt- inspect	2002	1752	250	87.51	12.49	1.95	146.30
Other Service	3295	3158	137	95.84	4.16	3.14	287.89
Private house servant	149	148	1	99.33	0.67	4.99	492.30
Prof.-specialty	4140	2281	1859	55.10	44.90	0.2	2.04
Protective Service	649	438	211	67.49	32.51	0.73	25.53
Sales	3650	2667	983	73.07	26.93	0.99	45.68
Tech Support	928	645	283	69.50	30.50	0.82	31.99
Transport moving	1597	1277	320	79.96	20.04	1.38	82.70
Total		**23068**	**7650**				

3. INFORMATION DISRUPTION WITH FUZZY SETS AND FUZZY INTUITIONISTIC SETS

3.1 Information Disruption with a Fuzzy Set Approach: One way to think about a Fuzzy Set (FS) is as a mapping from the set of real numbers (Ai) to membership values (xi) that fall between 0 and 1. When expressing the FS as (A,), the set A and the membership function x are both used. The membership function of the fuzzy set B = (A,) is defined as the function x = μB, where $A \to [0,1]$. Usually, the FS can be depicted as = $\{\langle x,\mu A(x)\rangle | x\epsilon X\}$. The degree of membership of $x\epsilon X$ is defined as $\mu(x)$.

3.1.1 TMF for triangle membership: With $a \le b \le c$, the TMF takes three parameters: a lower limit, an upper limit, and a value. The values of a and c indicate the triangle's foot, whereas the argument b indicates the triangle's top. Equation (1.5) provides the TMF.

$$\mu_A(x) = \begin{Bmatrix} 0 & x \le a \\ \left(\frac{x-a}{b-a}\right) & a \le x \le b \\ \left(\frac{c-x}{c-b}\right) & b \le x \le c \\ 0 & x \ge c \end{Bmatrix} \qquad (1.5)$$

3.1.1.1 Trapezoidal membership function (TrMF)

The truncated triangle with a flattened top end is known as the TrMF. In TrMF, there are four arguments: the lower limit a, the upper limit d, the lower support limit b, and the upper support limit c, where $a \le b \le c \le d$. In this case, the trapezium feet are denoted by the arguments and d, whereas the flattened top end of the trapezium is denoted by the arguments b and c. The TrMF can be found in Equation (1.6).

$$\mu_A(x) = \begin{Bmatrix} 0 & x \le a \\ \left(\frac{x-a}{b-a}\right) & a \le x \le b \\ 1 & b \le x \le c \\ \left(\frac{x-d}{c-d}\right) & c \le x \le d \\ 0 & x \ge d \end{Bmatrix} \qquad (1.6)$$

3.1.1.2 Gaussian Membership Function (GMF)

Fuzzy set theorists frequently use the Gaussian Membership Function (GMF) to illustrate the linguistic features and fuzziness of membership functions. The following is the GMF formulation:

$$\mu_A(x)=\exp(-(x-m)^2/2k^2) \qquad (1.7)$$

3.1.2 Data Perturbation using Intuitionistic Fuzzy Set Approach

3.1.2.1 Intuitionistic fuzzy triangular membership function (ITMF): Equations (1.8) and (1.9) provide the membership function and the non-membership function, respectively, which are contained in the ITMF. You may see the ITMF hesitation index in Equation (1.10).

$$\mu_A(x) = \begin{Bmatrix} 0 & x \le a \\ \left(\frac{x-a}{b-a}\right) & a \le x \le b \\ 1 & b \le x \le c \\ \left(\frac{x-d}{c-d}\right) & c \le x \le d \\ 0 & x \ge d \end{Bmatrix} - \pi_A(x) \qquad (1.8)$$

$$\mu_A(x) = \begin{cases} 0 & x \le a \\ \left(\frac{x-a}{b-a}\right) & a \le x \le b \\ 1 & b \le x \le c \\ \left(\frac{x-d}{c-d}\right) & c \le x \le d \\ 0 & x \ge d \end{cases} \quad (1.9)$$

The hesitation index is, $\pi(x)=1-\mu A(x)-\nu A(x)$ (1.10)

Table 1.3 Data perturbation using Fuzzy Set for Age Attribute from Adult Income Dataset after Calculating IV by applying TMF, TrMF and GMF

Operation	**Age Attribute**				
Original Values	39	50	38	53	28
After calculating IV	7.17	7.28	1.90	2.68	7.11
TMF (4.2)	0.991	0.777	0.973	0.719	0.509
TrMF (4.3)	1.00	0.952	1.00	0.880	1.00
GMF (4.4)	0.999	0.704	0.998	0.581	0.740

Table 1.4 Data Perturbation using Intuitionistic Fuzzy Set for Age Attribute from Adult Income Dataset after Calculating IV by Applying ITMF and ITrMF

Operation	**Age Attribute**				
Original Values	39	50	38	53	28
After calculating IV	7.17	7.28	1.90	2.68	7.11
Membership degree for ITMF (3.8)	0.892	0.777	0.899	0.719	0.509
Non membership degree for ITMF (3.9)	0.008	0.122	0.001	0.180	0.390
Membership degree for ITrMF (3.11)	0.889	0.724	0.886	0.880	0.888
Non membership degree for ITrMF (3.12)	0.011	0.176	0.014	0.020	0.012
Hesitation index - Equation (3.10), (3.13)	0.1	0.1	0.1	0.1	0.1
Perturbed data using ITMF & Equation (3.14)	1.00	0.79	0.98	0.74	0.65
Perturbed data using ITrMF & Equation (3.14)	1.00	0.960	1.00	0.89	1.00

3.1.2.2 The Intuitionist Fuzzy Gaussian Membership Function With Respect to Data Perturbation

The Gaussian membership function and non-membership function for intuitionistic fuzzy sets are defined using a Gaussian curve with parameters m (central value) and k (width), where k>0. The narrowness of the Gaussian curve increases as the value of k decreases. The Intuitionistic Gaussian Membership Function (IGMF) is illustrated in Figure 1.2.

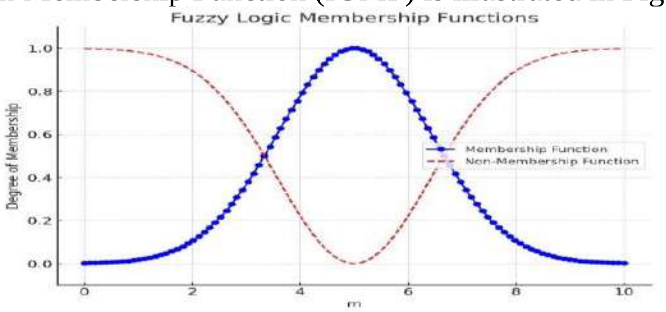


Figure 1.2: Intuitionistic Fuzzy Gaussian Membership Function

Table 1.5: Data Perturbation for Age Attribute from Adult Income Dataset after Calculating IV by Applying IGMF

Operation	**Age Attribute**				
Original Values	39	50	38	53	28
After calculating IV	7.17	7.28	1.90	2.68	7.11
Membership degree for IGMF (3.18)	0.898	0.700	0.895	0.580	0.740
Non membership degree for IGMF (3.19)	0.002	0.200	0.005	0.320	0.160
Hesitation index - equation (3.17)	0.1	0.1	0.1	0.1	0.1
Euclidean 2- norm (3.14)	1.00	0.74	1.00	0.67	0.76

Perturbed data using IGMF	1.00	0.74	1.00	0.67	0.76

Table 1.6 Data Perturbation using IGMF for Adult Income Dataset

Operation	Age	Education	Occupation
Original Value	39	9 th	Sales
After calculating IV	7.17	24.28	46.68
After applying IGMF	1.00	0.77	0.92

4. RESULTS AND ANALYSIS FROM EXPERIMENT

The section showcases the results of implementing the STIF algorithm.

Precision

Table 1.7 Accuracy of Adult Income Dataset

Classifier Models	Accuracy in %						
	Original Dataset	TMF	TrMF	GMF	ITMF	ITrMF	STIF
DT	72.25	70.8	72.25	72.25	72.25	72.25	72.25
XGB	78.25	76.25	76.1	76.75	76.35	76.4	76.9
RF	76.35	75.05	75.35	75.7	75.45	75.14	76
SVM	75.05	74.14	74.2	74.23	74.35	73.95	74.8

Table 1.8 Accuracy of Bank Marketing Dataset

Classifier Models	Accuracy in %						
	Original Dataset	TMF	TrMF	GMF	ITMF	ITrMF	STIF
DT	71	71	71	71	71	71	71
XGB	74.25	73.45	73.35	73.75	73.9	73.9	74.2
RF	74.85	74.24	74.25	74.85	74.37	74.35	74.85
SVM	69.2	68.12	68.7	69	68.75	68.25	69.1

Table 1.9 Accuracy of Lung Cancer Dataset

Classifier Models	Accuracy in %						
	Original Dataset	TMF	TrMF	GMF	ITMF	ITrMF	STIF
DT	87.5	87.5	87.5	87.5	87.5	87.5	87.5
XGB	100	87.5	87.5	100	87.5	87.5	100
RF	100	100	100	100	100	100	100
SVM	100	85.7	85.7	92.8	85.7	92.8	100

Ability to Preserve Privacy : Using the algorithms in the attributes from the adult income, bank marketing, and lung cancer datasets, we can compute the privacy preserving capability for TMF, TrMF, GMF, ITMF, ITrMF, and STIF based on Equation (1.3). Maximising the value for Equation (1.3) leads to an increase in PPC. Displayed in Figures 1.3, 1.4, and 1.5, respectively, are the PPC for the adult income dataset, the bank marketing dataset, and the lung cancer dataset.

Ability to Retrieve Data: Data changes as a consequence of data disturbance. It need to be feasible to recover or obtain the original data again as needed after disturbance as well. Rest assured, no data will be lost. Both the privacy and the usefulness of the data should be preserved by the algorithm. We discover the precision and recall by applying Equations (1.4) and (1.5) to the resultant values from the datasets used, which include TMF, TrMF, GMF, ITMF, ITrMF, and STIF. The precision and recall curves for the original dataset and the perturbed dataset treated with STIF are displayed in Figure 1.7. Both datasets display very similar precision and recall charts. This suggests that the information retrieved prior to and subsequent to data perturbation is very similar. The data utility is preserved because STIF has not resulted in any information loss across any datasets. Even after the data perturbation, the original data can be retrieved as needed. The outcome of the mining process will be unaffected.

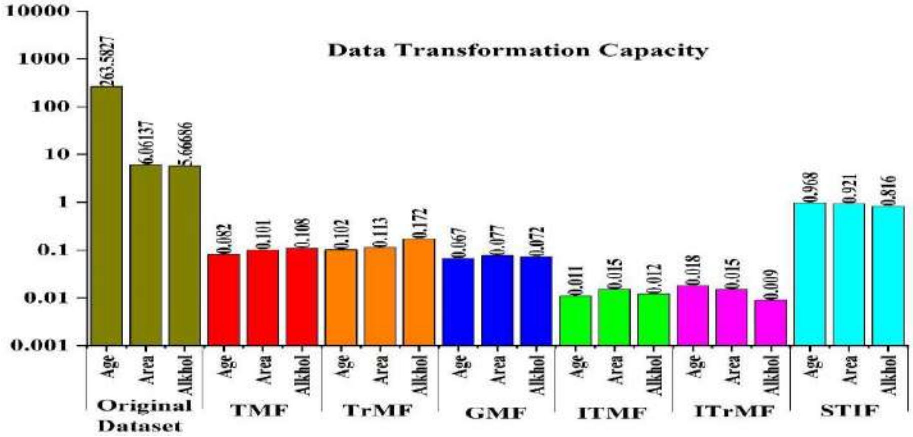


Figure 1.3: DTC of Lung Cancer Dataset

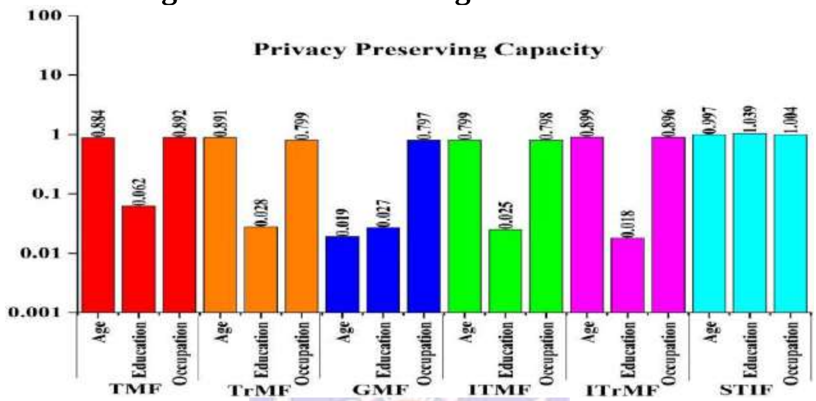


Figure 1.4: PPC of Adult Income Dataset

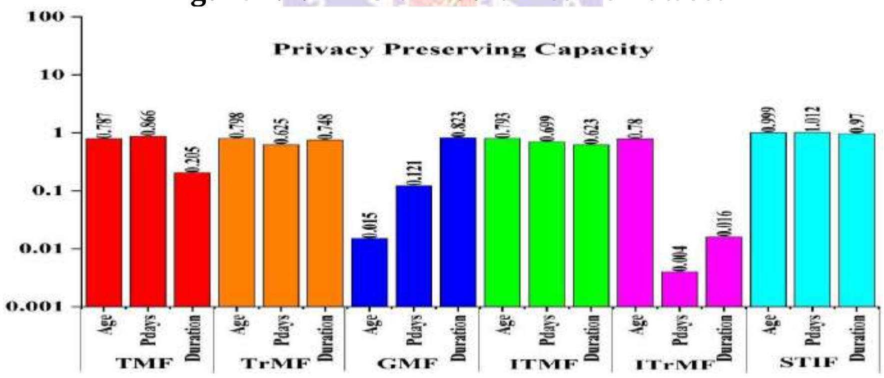


Figure 1.5: PPC of Bank Marketing Dataset

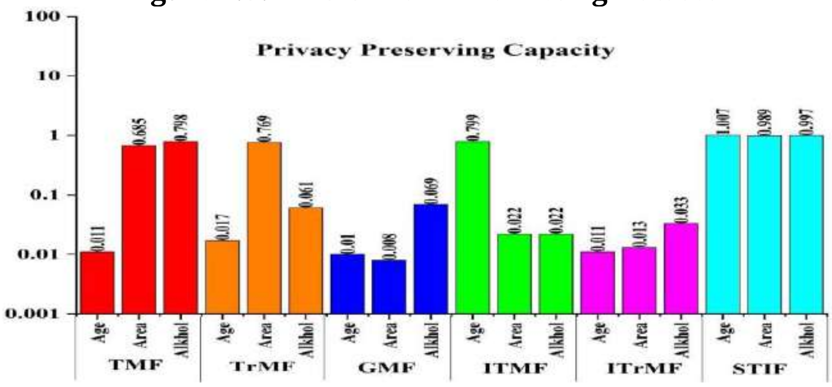
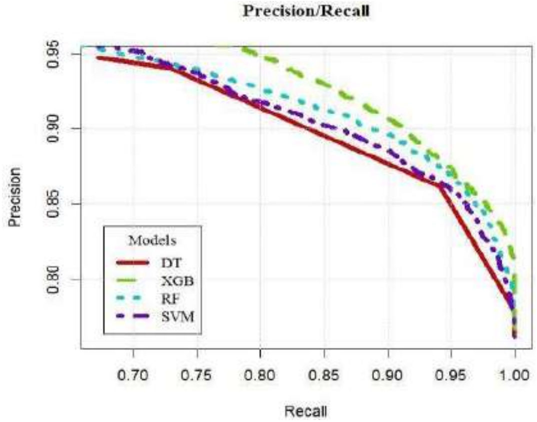
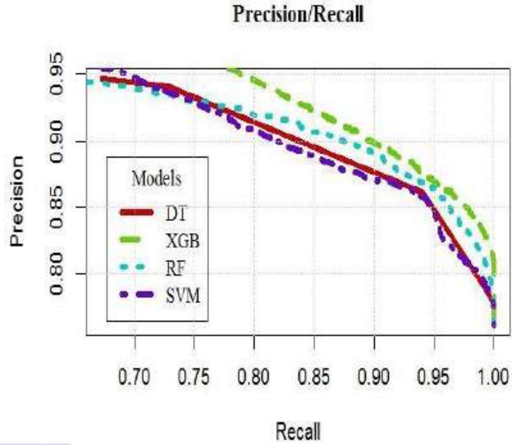


Figure 1.6: PPC of Lung Cancer Dataset

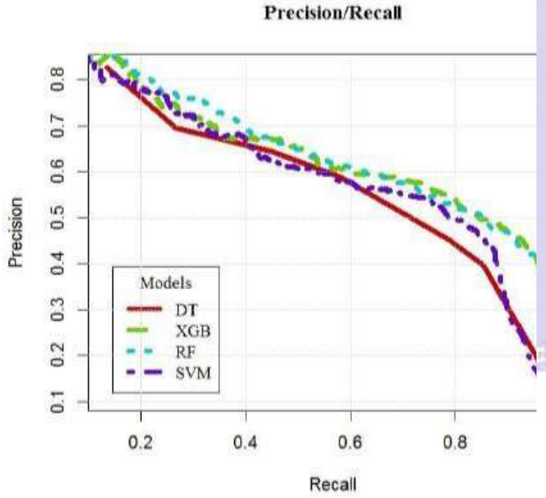
-Measuring Performance: Outputs from algorithms and the accuracy of classification methods are expressed using the words true positives, true negatives, false positives, and false negatives. Classifiers DT, XGB, RF, and SVM have been trained with the suggested STIF algorithm and evaluated on datasets including adult income, bank marketing, and lung cancer. To obtain the sensitivity-specificity curves displayed in Figure 3.11, the values obtained from each dataset using the STIF technique are input into Equations (1.6) and (1.7). The sensitivity-specificity curves for the original datasets are shown in Figure 3.11 (a, c, e), whereas the STIF applied adult income, bank marketing, and lung cancer datasets are portrayed in Figure 3.12 (b, d, f). Figure 3.12 shows that the sensitivity and specificity curves for the original and perturbed datasets are same, indicating that the STIF algorithm performs better. In the analyses of adult income and lung cancer, the sensitivity and specificity scores are both 100%. Also, even after applying STIF to the perturbed dataset, the pattern that was retrieved from the original dataset has not changed. That the STIF algorithm protects both personally identifiable information and sensitive data while still making use of it is demonstrated here. Any third parties can be provided with this disturbed data for analysis. The main achievement of the suggested work is that they are unable to obtain any personal or sensitive information about the individual.



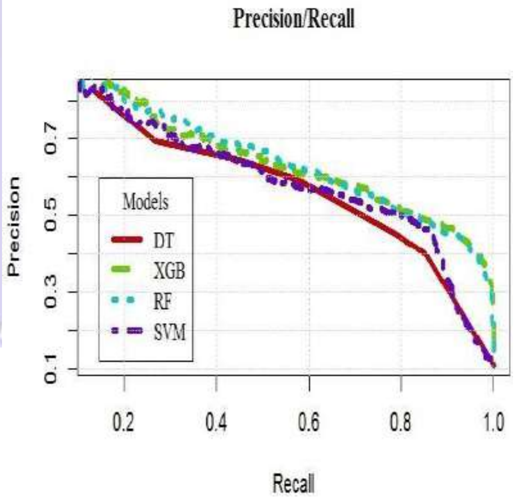
a. Actual adult income dataset



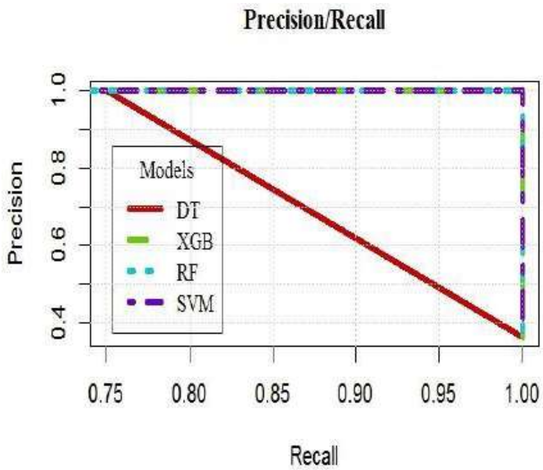
b. STIF applied adult income dataset



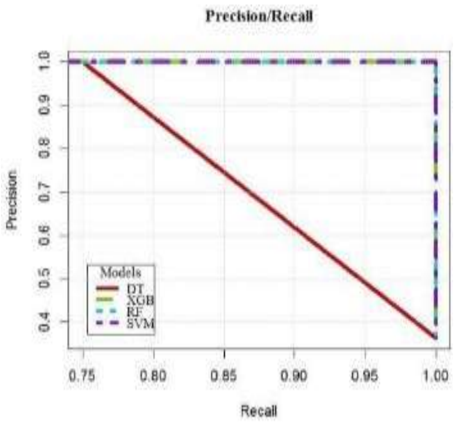
c. Actual bank marketing dataset



d. STIF applied bank marketing dataset

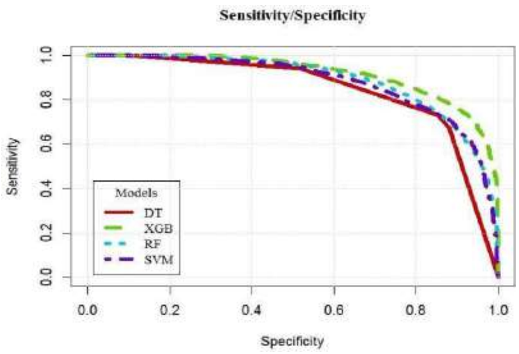


e. Actual lung cancer dataset

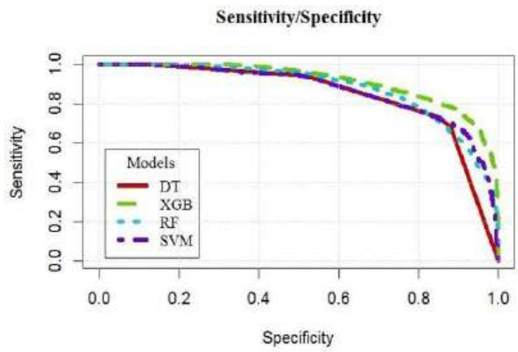


f. STIF applied lung cancer dataset

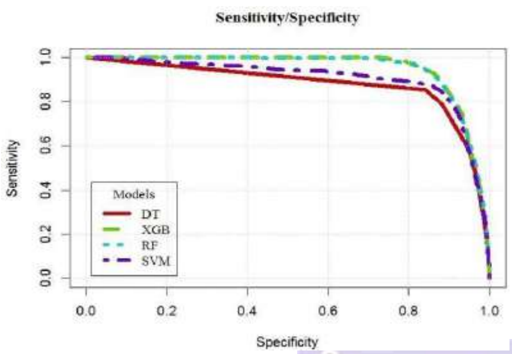
Figure 1.7 Precision/Recall Plot for Actual and Perturbed Dataset Using STIF Algorithm



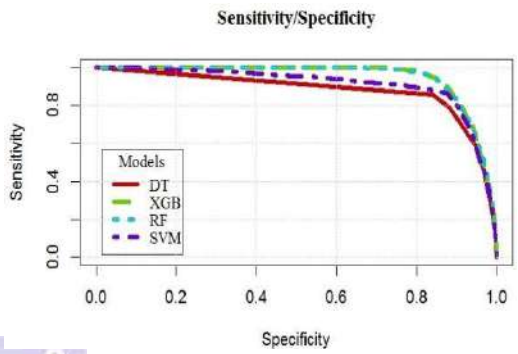
a. Actual adult income dataset



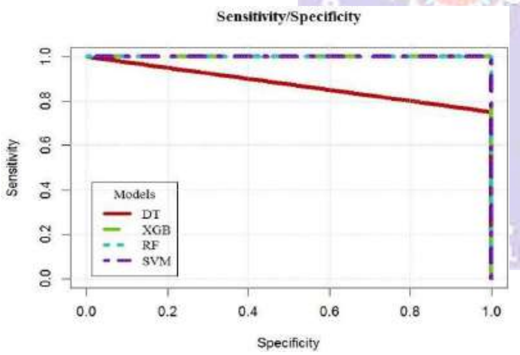
b. STIF applied adult income dataset



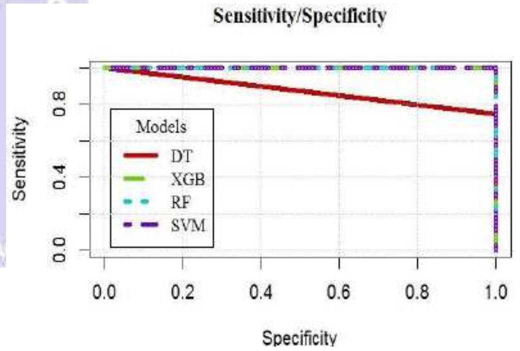
c. Actual bank marketing dataset



d. STIF applied bank marketing dataset



e. Actual lung cancer dataset



f. STIF applied lung cancer dataset

Figure 1.8: Sensitivity/Specificity Plot for Actual and Perturbed Dataset Using STIF algorithm

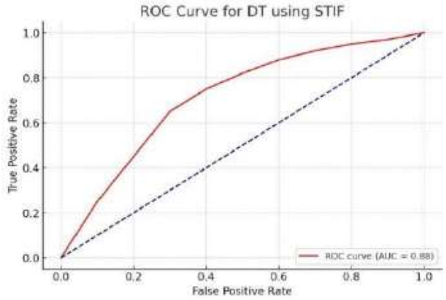
Performance at Various Thresholds

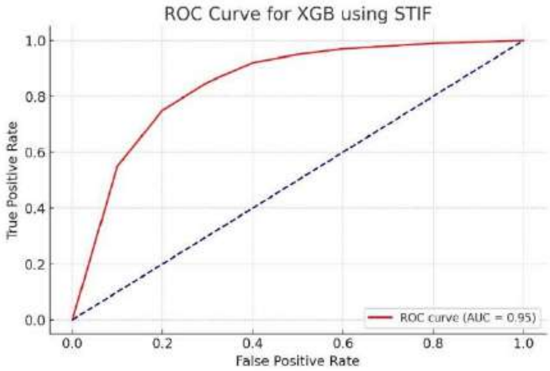
Table 1.10 AUC for Adult Income Dataset

Classifier	AUC for Actual Dataset	AUC for Perturbed Dataset using STIF algorithm
DT	0.85	0.84
XGB	0.92	0.91
RF	0.9	0.88
SVM	0.88	0.87

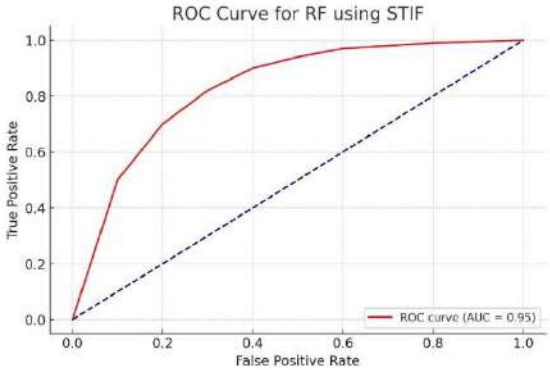
Table 1.11 AUC for Bank Marketing Dataset

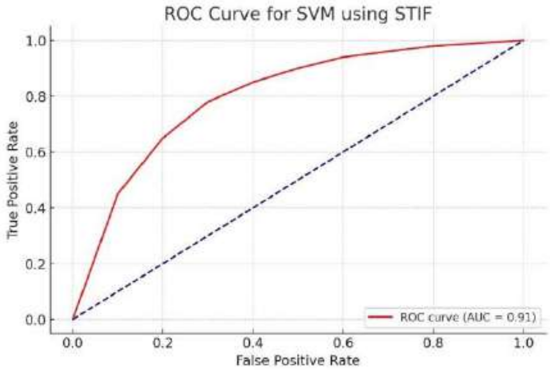
Classifier	AUC for Actual Dataset	AUC for Perturbed Dataset using STIF algorithm
DT	0.88	0.88
XGB	0.95	0.95
RF	0.95	0.95
SVM	0.9	0.91





a. ROC-AUC for DT using STIF algorithm b. ROC-AUC for XGB using STIF algorithm





c. ROC-AUC for RF using STIF algorithm d. ROC-AUC for SVM using STIF algorithm

Figure 1.9: ROC and AUC for the Perturbed Bank Marketing Dataset using STIF Algorithm

Table 1.12 AUC for Lung Cancer Dataset

Classifier	AUC for Actual Dataset	AUC for Perturbed Dataset using STIF algorithm
DT	0.88	0.88
XGB	1	1
RF	1	1
SVM	1	1

5. CONCLUSION

Modern businesses can benefit from data mining since it reduces costs while improving service and quality. It gives third parties access to personally identifiable information. The security of sensitive information is jeopardised. Using the STIF algorithm, this work tackles the problem of PPDM. Intuitionistic fuzzy Gaussian membership function for data perturbation and statistical transformation methods WOE and IV make up the STIF algorithm. Datasets pertaining to adult income, bank marketing, and lung cancer are processed using the STIF algorithm. Examining the STIF algorithm is done with the help of the classifier models DT, XGB, RF, and SVM. In the adult income dataset, the STIF algorithm achieves a perfect score of 100%; in the bank marketing dataset, it achieves a perfect score of 100%; and in the lung cancer datasets, it achieves a perfect score of 100% for DT, XGB, RF, and SVM. When compared to state-of-the-art algorithms, STIF performs better in both data transformation capacity and privacy maintaining capacity as determined by the variance metric. On the adult income dataset, the STIF algorithm has a data retrieval capacity of over 95%; on the bank marketing dataset, it is over 100%; and on the lung cancer dataset, it is over 100%. As a measure of STIF's performance, we have used sensitivity-specificity and ROC-AUC. For all datasets, the resultant sensitivity-specificity plot is virtually identical for both the original and STIF applied perturbed datasets. The adult income dataset has an AUC of 0.91, the bank marketing dataset yields an AUC of 0.95, and the lung cancer dataset yields an AUC of 1. Algorithmic performance is enhanced with a higher AUC value. Also, there is no difference between the patterns obtained before and after the data perturbation that STIF did. As a result, the data utility is maintained and no information is lost during data disturbance. In addition to data utility, the results show that the STIF algorithm better protects an individual's private and sensitive data.

REFERENCES

1. Good, I. J. (1985). "Weight of Evidence: A Brief Survey." *Bayesian Statistics 2*, 249-270.
2. Siddiqi, N. (2006). *Credit Risk Scorecards: Developing and Implementing Intelligent Credit Scoring*. Wiley Finance.
3. Hand, D. J., & Henley, W. E. (1997). "Statistical Classification Methods in Consumer Credit Scoring: A Review." *Journal of the Royal Statistical Society: Series A (Statistics in Society)*, 160(3), 523-541.
4. Anderson, R. (2007). *The Credit Scoring Toolkit: Theory and Practice for Retail Credit Risk Management and Decision Automation*. Oxford University Press.
5. Jiang, W., Li, X., & Zhao, L. (2009). "Application of Fuzzy Logic in Data Mining." *International Journal of Data Mining and Knowledge Management Process*, 1(1), 1-12.
6. Kandel, A. (1986). *Fuzzy Mathematical Techniques with Applications*. Addison-Wesley.
7. Zadeh, L. A. (1999). "Fuzzy Sets as a Basis for a Theory of Possibility." *Fuzzy Sets and Systems*, 100(1-3), 9-34.
8. Atanassov, K. T. (1986). "Intuitionistic Fuzzy Sets." *Fuzzy Sets and Systems*, 20(1), 87-96.
9. Li, X., & Liu, Y. (2013). "Privacy-Preserving Data Mining Based on Fuzzy Sets." *Journal of Applied Mathematics*, 2013, 1-10.
10. Sivagaminathan, R. (2014). "Privacy Preserving Data Mining using Fuzzy Logic Approach." *International Journal of Computer Applications*, 98(17), 1-6.
11. Khemchandani, V., & Chandra, S. (2016). "Fuzzy Logic-Based Classification Techniques for Credit Scoring." *Soft Computing*, 20(2), 1-12.
12. Van Gestel, T., Baesens, B., & Martens, D. (2005). "Credit Scoring for Good and Bad Credit Using Fuzzy Rule-Based Classifiers." *European Journal of Operational Research*, 160(3), 791-806.
13. Sun, H., Liu, J., & Li, F. (2017). "Application of Weight of Evidence and Information Value in Risk Prediction Models." *Risk Analysis Journal*, 37(9), 1728-1743.
14. Mohamed, F. M., & Rasheed, H. (2019). "Enhancing Predictive Modeling with Intuitionistic Fuzzy Systems." *Journal of Fuzzy Systems*, 27(4), 1056-1068.