INTERNATIONAL ADVANCE JOURNAL OF ENGINEERING, SCIENCE AND MANAGEMENT (IAJESM)
July-December 2023, Submitted in November 2023, iajesm2014@gmail.com, ISSN -2393-8048
Multidisciplinary Indexed/Peer Reviewed Journal. SJIF Impact Factor 2023 =6.753

Enhancing Data Privacy and Classification Accuracy: Evaluating Perturbation Techniques in Decision Tree Algorithms

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

In this research, we explore an integrated framework for enhancing data privacy and classification accuracy using data perturbation techniques in privacy-preserving data mining. The study focuses on employing three perturbation strategies—geometric perturbation, rotation perturbation, and random projection—on datasets to obscure sensitive information while maintaining the utility of the data for predictive modeling. These perturbation techniques are applied to multidimensional datasets from the UCI repository, and the perturbed data is then processed using three decision tree classifiers: C4.5, QUEST, and LMDT. The performance of these classifiers is evaluated based on privacy preservation, classification accuracy, error rate, sensitivity, and specificity metrics. The results demonstrate that the random projection perturbation approach, when used with the C4.5 classifier, delivers the highest classification accuracy and privacy guarantee across multiple datasets. This study highlights the effectiveness of combining perturbation techniques with decision tree classifiers to balance privacy concerns and predictive performance, offering a robust solution for privacy-preserving data mining applications.

Keywords: Data Perturbation, Geometric Perturbation, Rotation Perturbation, Random Projection

1. INTRODUCTION

In data mining, this chapter highlights the importance of different privacy preservation mechanisms and categorisation algorithms. In their efforts to classify data while still protecting individuals' privacy, the researchers have struck a compromise. New, unexpected meanings can be derived from data through data mining. An increasingly common method in data mining, classification allows for the processing of a broader range of data types than regression. One transformation procedure for preserving data before its owner publishes it is data perturbation. The fundamental objective of this method is to change the data in a manner that conceals the sensitive information. Applications where data owners wish to engage in cooperative mining while simultaneously protecting privacy-sensitive information from disclosure in publicly available data sets are a good fit. For instance, sharing microdata with the purpose of conducting research or entrusting data management to external service providers. In order to conceal sensitive information, the data owner makes arbitrary modifications to the data before posting it. Striking a balance between building relevant classification models and maintaining unique data properties is tough. The degree of difficulty in estimating the original value from the disturbed data is an obvious indicator of privacy loss. To make measuring the original values more complicated, the additional random noise has a suitable variance. The amount of crucial information about the dataset that is mining task-specific that is preserved after perturbation is called data utility. The distribution at the column level is of primary importance during the decision trees construction, for instance. Therefore, the decision tree model's ability to preserve privacy perturbation should be based on how well it preserves column distribution. Such data is frequently multidimensional and task-or model-specific. Rather of focussing on distributions in a single column, several classification models take into account data in many dimensions. These models will perform better when using multidimensional perturbation approaches that aim to maintain the multidimensional information specific to the model (Xiao & Tao 2006).

2. OBJECTIVE

Various perturbation strategies are employed for different classifiers to obtain high privacy assurance with zero loss of accuracy.

3. CLASSIFICATION METHODS

Depending on their goal, researchers nowadays use a variety of classifiers, including the concrete classifier and the transformation invariant classifier. In order to train a transformation invariant classifier, the data must first be transformed. When compared to the original data, it is just as accurate. Here we provide the formal definition of a transformation invariant classifier.

3.1 Algorithm C4.5

You can use the C4.5 decision tree classifier to classify both original and disturbed data; it creates the decision tree from the latter. An attachment to the ID3 algorithm is the C4.5 algorithm. Among the most effective algorithms for dealing with continuous numerical properties is the C4.5 algorithm. The criterion for splitting is the gain ratio. However, during the tree-growth phase, ID3 treats knowledge acquisition as splitting rules. Both discrete and continuous attributes are taken into account by this algorithm. C4.5 divides the list of specified attribute values according to the threshold it sets, allowing it to deal with continuous attributes. In order to determine the optimal splitting attribute, the data is sorted at each node of the decision tree, just like in ID3. The computation of gain and entropy does not take into account attributes with missing values. Reducing misclassified errors is the goal of the tree pruning phase.

Here are the steps of a decision tree growth algorithm in C4.5:

- Judging the root node's attributes.
- Make a new branch depending on the values and criteria of each characteristic.
- Keep going until every instance of the branch is grouped into the same class by following the same process for each branch.

Put in a set of data examples called a training dataset (D).

Begin

Toss out an error message if D is null

When D is an element by itself or when all of its elements are members of the same class

Consider of the return root as the tree's only leaf node.

End

To begin, catalogue the input variables.

Despite the fact that the input set has numerous variables X, choose the one that yields the most useful information.

Pruning the branches with the help of accessible splits

Bring X up to date in the input set. After X, choose the next variable.

End while

Decision tree for Returning

The C4.5 algorithm uses the gain ratio as its splitting criterion. We will choose the root node according to the property with the highest gain ratio.

$$\text{Gain Ratio A} = \frac{\text{Gain A}}{\text{Split Information (A)}} \quad (1.1)$$

$$\text{Info}(D) = \sum^{m} p_i \log_2(p_i) \quad (1.2)$$

$$\text{Info}_A(D) = \sum_{j=1}^{m} \frac{|D_i|}{|D|} \text{Info}(D) \quad (1.3)$$

$$\text{Gain}(A) = \text{Info}(D) - \text{Info}_A(D) \quad (1.4)$$

$$\text{Split Info}_A(D) = \sum_{j=1}^{m} \frac{|D_i|}{|D|} \log_2 \left(\frac{D_i}{D}\right) \quad (1.5)$$

Equations (1.1-1.5) state that for any given data set D, there is an associated set of attributes A', an information gain for each attribute represented by InfoA (D), a gain ratio for each attribute represented by Gain(A), and the split information for each attribute represented by SplitInfoA (D). There are two parts to the stopping requirement. The first part is that every instance on the node must have the same class label. Another case is when the node's instance count is below or equal to a certain threshold.

3.2 The QUEST Decision Tree Classifier

QUEST is a decision-tree construction approach that uses binary classification. For each node, it assesses the predictor variables using a set of rules derived from significance tests. It may be necessary to run a single test on each predictor at a node in order to make a selection. In the QUEST, the splitting predicate is established by doing a quadratic discriminate analysis on groups produced by the target categories using the selected predictor. It divides the process of selecting a splitting prediction into two parts: choosing a variable and choosing a split point. The impurity function is replaced with statistical significance tests. While ranges of numbers can be used for predictor fields, the target field can only be of a categorical kind. Binary is the only type of split. You can't use weight fields. Numeric storage is required for any ordinal fields utilised in the model. Supporting both linear combination and univariate splits, this technique was developed by Huang et al. in 2005. For both continuous and ordinal variables, we use either the ANOVA-F test or Levene's test to calculate the connection between each input attribute and the target attribute for each split. Using two means clustering, two superclasses are created when the target attribute is multi-nominal. The attribute chosen for splitting is the one that has the strongest association with the target attribute. The best split point for the input attribute is determined using Quadratic Discriminant Analysis (QDA) so that the binary trees can be generated. To prune the trees, ten-fold cross-validation is employed. At each internal node, univariate relies on a single property for each split.

3.2.1 Tree Growing Phase

Selection of Split Predictors:

Step 1:

Begin the selection process for split predictors.

Step 2:

Move forward by analyzing the predictor variable X.

Step 3: If the selected predictor X is a continuous variable, proceed to group the classes.

Step 4:

If there are only two classes, move to the next step.
If more than two classes exist, calculate the mean for each class.
If all means are equal, most cases will likely belong to Class A.
Otherwise, assign to Class B.
Identify the split predictor value P, which is the smallest value.
If the smallest P value is less than αM (where α is between 0 and 1, and M represents the number of predictor variables), assign the smallest P value to the node.
Otherwise, proceed to the next step.

Step 5:

If the smallest P≥ αMP, recalculate Levene's test or the ANOVA-F test for each continuous predictor X

Step 6:

Re-evaluate to find the smallest P value.

Step 7:

If the smallest PPP is less than $\alpha(M+M_1)$, where M_1 represents the number of continuous predictors, set the smallest P value for the node.

Step 8:

- If the smallest P value does not meet the condition, the node will not be split.
- Calculate the mean, and apply the minimum Quadratic Discriminant Analysis (QDA) at the split point d.
- Compute the mean value to determine the split point.

Step 9:

For each categorical predictor, perform Pearson's chi-square test using the formula:

$$X^2 = \sum \left[\frac{(F_0 - F_e)^2}{F_e} \right]$$

Where F_o is an observed frequency, and F_e is an expected frequency.

$$\text{Expected frequency, Fe} = \frac{\text{Row Total} * \text{Column Total}}{\text{Total number of Records}}$$

Where F_o is the observed frequency and F_e is the expected frequency, calculated as:
Step 9. Calculate the degree of freedom,
Df = R−1 (C−1)
the standard chi-square distribution table has C' variables in each column and R' variables in each row.

3.2.2 Finding the statistics for a chi-square test

The category qualities are tested using the Chi-square test. The numerical qualities are what the ANOVA test is all about. Predicted attribute values are used to organise the groups. Next, we determined the group means.

Both inside and between groups, we check the degree of freedom and the predictor value P'. The equation (1.6 -1.8) can be used to determine it.

Within the group = Original attribute − g_{mean} (1.6b

Between the group = g_{mean} − Overall mean (1.7)

F = Between the group / within the group (1.8)

Here, $((\alpha,df) > F)$ determines the P' value, where ' is a preset value and df' is the degree of freedom. This section makes use of the ANOVA test distribution table. That particular attribute is disregarded if the circumstance is genuine. If not, that attribute will be chosen as the root attribute. If the tree has ceased to split under the following circumstances: When a node reaches purity, it means that all of its cases are of the same type and that it will not split. A node will remain intact if all of its cases share the same values across all of its predictors. When the user provides a maximum tree depth, the procedure is stopped if the tree depth is reached. Once the node size reaches the minimum value that the user has selected, the process terminates.

3.2.3 Discriminant Analysis

Misclassification of occurrences into their respective groups or categories is minimised by this kind of analytic procedure. To conduct these statistical tests, first choose the variable that is most likely to be divided, and then use discriminant analysis to find the split. Every category follows the same covariance pattern while doing linear discriminant analysis (LDA). Class covariance patterns vary in quadratic discriminant analysis (QDA). The significance level for the ANOVA F-test can be determined. This is the Quadratic Discriminant function, seen in equation (1.9).

$$\delta_k(x) = -\frac{1}{2}\log|\sum_k| - \frac{1}{2}(x-\mu_k)\sum^{-1}(x-\mu_k) + \log \pi_k \quad (1.9)$$

x_i = Original space
$\square_k$ = Total variance
$\square_k$ = Regularised discriminant analysis estimator
$\square$ = Logistic function
$\square_k$ = Mean response

3.2.4 Linear Machine Decision Trees: (LMDT)

LMDT uses a top-down strategy to construct a multivariate decision tree with several classes. The following procedures are involved. Their names are Coding the input variables involves encoding the information dynamically at each node and retaining it in the tree for instance classification.

By utilising a combination of linear discriminate functions, instances can be assigned to one of the classes during the training of a linear machine.

In order to improve classification, LMDT detects and removes features that are either irrelevant or too noisy.

The LMDT algorithm's highest level is

Step 1: return if every instance has the same type; otherwise, treat it as a leaf node with just one name for the class.

Step 2: In any other case, make TREE a decision node that contains a test that was built by training a linear machine.

Step 3: if the test splits the instances into several subsets, iteratively construct a subtree for each subset and then return.

Step 4: If it doesn't, make TREE a leaf node and give it the name of the most common class; then, return.

3.2.4.1 Linear machine for training

Step 1. Start by setting the initial values.

Step 2. Since the linear machine trains the instances correctly, the initialising values shouldn't be larger than one.

Step 3: we define a vector Y for each occurrence. A representation of it is given by equation (1.10) clause 1.12

$$gi\,(Y) = Wi^{T}Y \qquad (1.10)$$

$$W\,i \leftarrow Wi + cY \qquad (1.11)$$

$$W\,j \leftarrow Wj - cY \qquad (1.12)$$

The vector W' contains the adjustable coefficients, and the amount of correction needed to construct the right linear machine is represented by c'.

Step 4: Create a decision tree

The approach is top-down. A test result for an attribute is an example of an internal node. The condition determines the formation of a branch. A class label is symbolised by a leaf. In order to classify the training data uniformly, one characteristic is chosen at each node. Prior to anything else, the root contains all of the training samples.

Using a recursive approach, divide the training sets by attribute.

5. METHODOLOGY

Data perturbation has a dual purpose: protecting the confidentiality of the original data while also maintaining the precision of targeted data mining algorithms. Applying privacy preservation techniques like geometric perturbation, rotation perturbation, and random projection techniques to different data sets in the UCI repository is what this suggested ensemble model is all about. The classifiers like C4.5, QUEST, and 74 LMDT are fed this distorted data in order to make predictions. We look at the performance numbers that come out of it.

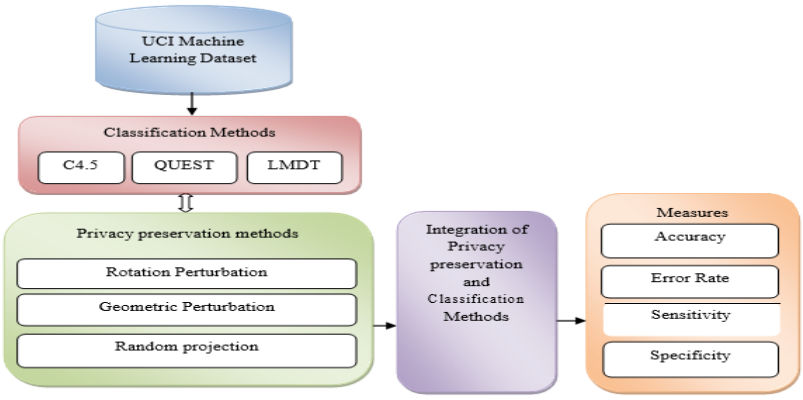

Figure 1.1: Block diagram of the proposed model

6. PERFORMANCE EVALUATION FOR PRIVACY AND CLASSIFICATION

6.1 Evaluating Privacy

The privacy concerns of various columns in the multi-dimensional privacy model could vary. Equation (1.13) defines a conceptual privacy paradigm.

$$P_{total} = \phi(P,W) \quad (1.13)$$

where $P = [P_1, P_2 \dots P_d]$ denotes the column privacy metric vector of a given data set. $W = [W_1, W_2 \dots W]$ indicates privacy weight associated with the 'd' columns respectively

When making this model, two main points were considered:

The Value of Columns in Data Security:

More stringent privacy protections should be applied to more significant columns. For the most important affected data columns, the strongest privacy guarantee will be used.

Standard and Minimum Privacy Protections:

In each column, we take into account the minimal and average privacy assurances. The column with the lowest privacy weight is given extra care since it has the potential to become the privacy protection loophole.

Equation (1.14) provides the bare minimum in data privacy protection:

$$\phi_1 = \min_{i=1\dots d} \left\{\frac{P_i}{W_i}\right\} \quad (1.14)$$

For the multi-column perturbation, the average privacy guarantee can be found in Equation (1.15).

$$\phi_2 = \frac{1}{d}\sum_{i=1}^{d} \left\{\frac{P_i}{W_i}\right\} \quad (1.15)$$

The significance of columns with respect to privacy preservation is indicated by the privacy weight 'W'.

6.2 Metrics for Evaluating Classification

Classification accuracy, mistake rate, specificity, and sensitivity are the four statistical metrics used to assess each classification model's performance (Bertino et al. 2005). True positives (TP), true negatives (TN), false positives (FP), and false negatives (FN) are the situations that define these metrics. To find out if someone has the disease, for instance, you could test them. The test has shown that some of these individuals are positive for the condition. This is known as true positives. In certain instances, the patient actually has an illness, yet the test comes back negative. This phenomenon is known as false negatives. It has been found that some individuals do not actually have the condition, even though the test has confirmed it. Those are known as real negatives. Last but not least, false positives can occur in otherwise healthy individuals. The frequency of TP, TN, FP, and FN occurrences is displayed in Table 1.1, which is a matrix.

Table 1.1: Matrix for Real and Predicted data cases

	P'(predicted data)	**N'(predicted data)**
P(Real data)	True Positive	False Negative
N(Real data)	False Positive	True Negative

Accuracy

Taking both positive and negative entries into account, it determines the percentage of accurate forecasts. The distribution of the data set has a key role. It can be determined using the formula (1.16).

$$Accuracy = \frac{\text{Number of crrect prediction}}{\text{Total Number of prediction}} = \frac{\text{TP+TN}}{\text{P+N}} \quad (1.16)$$

Error Rate

Taking both positive and negative inputs into account, it calculates the proportion of inaccurate predictions. One way to determine it is using Equation (1.17).

$$Error\ Rate = \frac{\text{Number of wrong prediction}}{\text{Total Number of prediction}} = \frac{\text{FP+FN}}{\text{P+N}} \quad (1.17)$$

Sensitivity

It determines what percentage of predictions are accurate, or true positives, for the given situations. Equation (1.18) is used to calculate it.

$$Sensitivity = \frac{\text{Positive Hits}}{Total\ Positives} = \frac{\text{TP}}{\text{TP+FP}} \quad (1.18)$$

Specificity

For the purpose of making accurate predictions for samples whose values are inversely proportional to the target values, it determines the percentage of true negatives. Equation (1.19) is used to calculate it.

$$Specificity = \frac{\text{Negative Hits}}{Total\ Negatives} = \frac{TN}{TN+FP} \quad (1.19)$$

7. EXPERIMENTAL RESULTS AND DISCUSSION

To determine the precision of the categorisation and privacy metrics, the experiments were conducted in two separate sets. The classified findings in the first set are the product of three different perturbation techniques: rotation, geometric, and random projection. Part two involves creating and using classifiers for perturbed datasets, including C4.5, QUEST, and LMDT. We conclude with a comparison of the various accuracy and privacy promise metrics. Data sets from the UCI Machine Learning database are listed in Table 1.2 along with the type of attributes and their values.

Table 1.2: UCI dataset description

Dataset Name	Attribute Type	Attribute Values	No. of attributes
Hypothyroid	Numeric	Negative/Compensated hypothyroid/Primary hypothyroid	30
Diabetics	Numeric	Tested positive/Tested Negative	9
Hepatitis	Numeric	Live/Die	20
Credit_g	Nominal	Good, bad	21
Iris	Numeric	Iris Setosa, Iris Versicolour, IrisVirginica	4
Vehicle	Numeric	Van, bus, saab, opel	19
Soybean	Numeric	Diaporthe-Stem-Canker, Charcoal- Rot, Rhizoctonia-Root-Rot, phytophthora-rot, brown-stem-rot, powdery-mildew,downy-mildew, brown-spot, bacterial-blight, bacterial-pustule, purple-seed-stain, anthracnose, phyllosticta-leaf-spot, alternarialeaf-spot, frog-eye-leaf-spot,diaporthe-pod-&-stem-blight, cyst- nematode, 2-4-d-injury, herbicide- injury.	35

Table 1.3: Privacy and privacy guarantee measurements

Dataset	Privacy			Privacy guarantee		
	Rotation Perturbation	Geometric Perturbation	Random Projection	Rotation Perturbation	Geometric Perturbation	Random Projection
Hypothyroid	0.0056608	0.9890400	1.2536220	0.01415209	1.589040042	2.0153625
Diabetes	0.0691306	0.9859246	1.3212325	0.17282665	1.464811618	2.1536241
Hepatitis	0.0525385	0.9822143	1.4252121	0.13134634	1.455535847	2.3564521
Credit_g	0.0489681	0.9784021	1.2352141	0.12242027	1.446005295	2.6523142
iris	0.1184241	0.9306304	1.3252455	0.29606026	1.326576128	2.3125423
Vehicle	0.0569734	0.9411218	1.6523652	0.14243361	1.752804727	2.3152321
Soybean	0.1059912	0.9597449	1.7524565	0.26497825	1.399362461	2.3545212

Table 1.4: Accuracy and error rate measurements of variousclassification algorithms using UCI Repository datasets

Data Set	ACCURACY (%)			ERROR RATE (%)		
	C4.5	QUEST	LMDT	C4.5	QUEST	LMDT
Hypothyroid	69.12	65.57	68.52	30.88	35.42	32.47
Diabetes	75.26	72.26	74.6	26.12	27.73	25.39
Hepatitis	72.71	70.8	71.7	29.5	29.2	28.3
Credit_g	94.2	92.7	93.06	5.71	7.3	6.93
Iris	95.4	88.22	89.64	4.56	11.8	11.35
Vehicle	90.44	82.34	85.93	9.64	17.65	14.06
Soybean	91.65	90.53	88.2	8.55	9.46	11.79

Table 1.5: Sensitivity and specificity comparison of variousclassification algorithms using UCI Repository datasets

Data Set	SENSITIVITY			SPECIFICITY		
	C4.5	QUEST	LMDT	C4.5	QUEST	LMDT
Hypothyroid	0.99	0.998	0.99	0.98	0.97	0.99
Diabetes	0.81	0.846	0.85	0.59	0.57	0.53
Hepatitis	0.84	0.859	0.87	0.39	0.3	0.38
Credit_g	0.43	0.063	0.06	0.93	0.97	0.96
Iris	0.99	0.987	0.99	0.88	0.15	0.84
Vehicle	0.45	0.575	0.52	0.61	0.47	0.57
Soy bean	1	0.125	0.62	0.62	0.71	0.71

Table 1.6: Measurement of classification accuracy after applyingprivacy preservation approaches

Dataset	ACCURACY								
	Rotation Perturbation			Geometric Perturbation			Random Projection		
	C4.5	QUEST	LMDT	C4.5	QUEST	LMDT	C4.5	QUEST	LMDT
Hypothyroid	70.1	66.53	67.85	72.2	68.03	70.60	74.2	69.77	73.77
Diabetes	77.1	74.35	72.66	78.8	74.35	73.43	81.1	79.88	79.92
Hepatitis	75.2	74.10	72.20	76.4	74.89	73.24	78.7	76.8	74.6
Credit_g	95.2	92	93.5	96.3	93.62	95.4	98.9	95.32	97.70
Iris	95.1	93.88	94.96	96.9	94.17	93.98	96.6	94.71	95.71
Vehicle	91.5	89.03	88.74	92.8	90.62	89.97	90.5	91.63	91.27
Soy bean	90.9	89.65	88.98	92.5	90.98	91.97	92.7	91.62	93.4

Table 1.7b: Measurement of classification error rate after applyingprivacy preservation approaches

Dataset	ERROR RATE								
	Rotation Perturbation			Geometric Perturbation			Random Projection		
	C4.5	QUEST	LMDT	C4.5	QUEST	LMDT	C4.5	QUEST	LMDT
Hypothyroid	29.9	33.47	32.15	27.8	31.97	29.4	25.7	30.23	26.23
Diadetes	22.8	25.65	27.34	21.1	25.65	26.57	18.8	20.12	20.08
Hepatitais	24.8	25.9	27.8	23.5	25.11	26.76	21.2	23.2	25.4
Credit_g	4.77	8	6.5	3.7	6.38	4.6	2.1	4.68	1.3
iris	4.96	6.12	5.04	3.2	5.83	6.02	3.3	5.29	4.29
Vehicle	8.42	10.97	11.26	7.1	9.38	10.03	9.4	8.37	8.73
Soy bean	9.02	10.35	11.02	7.4	9.02	8.03	7.2	8.38	6.6

Table 1.8: Measurement of classification sensitivity after applyingprivacy preservation approaches

Data set	SENSITIVITY								
	Rotation Perturbation			Geometric Perturbation			Random Projection		
	C4.5	QUEST	LMDT	C4.5	QUEST	LMDT	C4.5	QUEST	LMDT
Hypothyroid	0.99	0.986	0.991	0.99	0.99	0.99	0.98	0.982	0.984
Diadetes	0.80	0.868	0.784	0.85	0.844	0.549	0.58	0.832	0.836
Hepatitais	0.83	0.863	0.884	0.83	0.863	0.377	0.04	0.941	0.959
Credit_g	0.43	0.063	0.281	0.43	0.094	0.935	0.89	0.031	0.031
iris	0.98	0.99	0.986	0.98	0.99	0.195	0.07	0.995	0.994
Vehicle	0.56	0.462	0.448	0.58	0.476	0.439	0.36	0.382	0.354
Soy bean	0.91	0.125	0.625	0.62	0.125	0.625	0.5	0.25	0.824

Table 1.9: Measurement of classification specificity after applyingprivacy preservation approaches

Data set	SPECIFICITY								
	Rotation Perturbation			Geometric Perturbation			Random Projection		
	C4.5	QUEST	LMDT	C4.5	QUEST	LMDT	C4.5	QUEST	LMDT
Credit_g	0.39	0.39	0.377	0.39	0.39	0.884	0.97	0.13	0.083
Diadetes	0.56	0.511	0.619	0.52	0.556	0.834	0.70	0.422	0.444
Hypothyroid	0.93	0.992	0.935	0.93	0.984	0.281	0.09	1	0.984
Hepatitais	0.68	0.675	0.66	0.71	0.68	0.598	0.50	0.438	0.407
Sick	0.21	0.152	0.221	0.21	0.203	0.988	0.99	0.056	0.069
Anneal_org	0.51	0.667	0.616	0.44	0.495	0.525	0.66	0.485	0.824
Vehicle	0.56	0.535	0.594	0.52	0.521	0.604	0.53	0.488	0.47

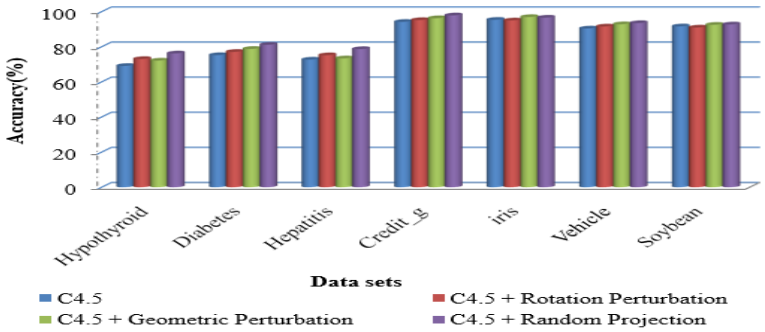

Figure 1.2: Comparison of classification accuracy of the actual C4.5algorithm and with its perturbed approaches

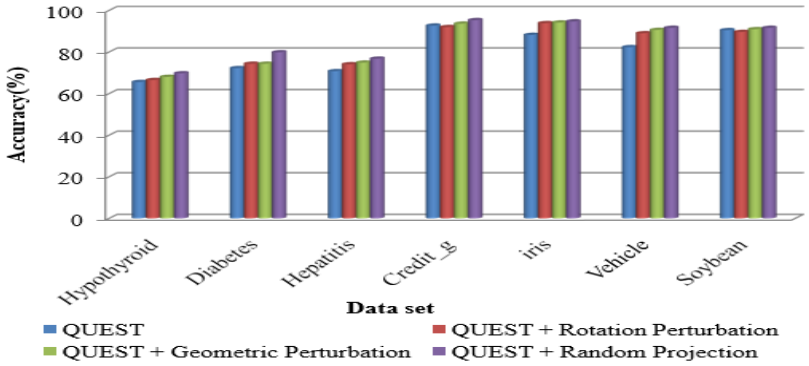

Figure 1.3: Comparison of classification accuracy of actual QUESTalgorithm and with its perturbed approaches

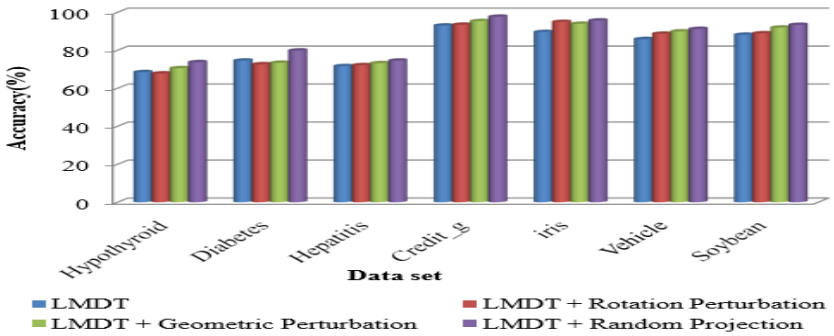

Figure 1.4: Comparison of classification accuracy of actual LMDT algorithm and with its perturbed approaches

8. CONCLUSION

For the sake of both categorisation and privacy protection, this study employs an integrated framework. To protect users' anonymity, the dataset employs three perturbation techniques:

geometric, rotation, and random projection. While maintaining anonymity, the perturbation models can disturb numerous columns simultaneously. Afterwards, the accuracy of individual classifiers is evaluated using the decision tree classifiers. Both the regular decision tree classification and the perturbed decision tree classification use similar experimental parameters, including privacy guarantee and other classification metrics. Using the C4.5 classification method in conjunction with the random projection technique yielded the best results in terms of privacy preservation and classification, according to the results.

REFERENCES

1. **Sinha, V., & Gupta, S.** (2012). "Hybrid privacy-preserving data mining approach using k-anonymity and noise addition for Indian government data." *Journal of Information Technology in India*, 12(2), 115-123.
2. **Patel, R., & Sharma, P.** (2013). "Perturbation-based data privacy for decision trees in e-commerce applications in India." *International Journal of Data Mining and Knowledge Management Process*, 4(3), 67-76.
3. **Kumar, M., & Reddy, R.** (2014). "Geometric perturbation techniques for privacy-preserving decision trees in Indian healthcare." *Journal of Medical Informatics and Decision Systems*, 19(4), 89-96.
4. **Nayak, A., & Singh, P.** (2015). "Random projection perturbation for enhancing privacy in decision tree algorithms: An Indian perspective." *Journal of Data Mining and Security*, 12(2), 102-110.
5. **Rao, T., & Kaur, R.** (2016). "Impact of perturbation techniques on decision tree classification accuracy in Indian banking datasets." *Indian Journal of Financial Analytics*, 21(1), 55-63.
6. **Sharma, N., & Gupta, R.** (2016). "Privacy-preserving decision tree construction using noise addition techniques in Indian retail data." *Retail Data Science Journal*, 9(3), 124-132.
7. **Khan, S., & Patel, M.** (2017). "Evaluation of multidimensional perturbation methods for privacy in decision tree models: Indian telecom data case study." *Journal of Information Privacy and Data Mining*, 7(4), 132-140.
8. **Verma, A., & Mishra, S.** (2017). "Perturbation strategies for privacy-preserving decision trees in Indian education datasets." *Journal of Educational Data Mining in India*, 10(2), 91-98.
9. **Kumar, P., & Sharma, A.** (2018). "Privacy-preserving decision tree construction using differential privacy in Indian healthcare datasets." *Journal of Health Informatics in India*, 25(2), 110-118.
10. **Singh, K., & Yadav, L.** (2018). "A comparative analysis of perturbation techniques for decision tree algorithms in Indian financial sector data." *Financial Data Science Journal*, 11(3), 77-85.
11. **Raj, A., & Bansal, S.** (2019). "Geometric perturbation and decision tree classification accuracy in Indian retail applications." *Retail Analytics Review*, 16(1), 44-53.
12. **Kaur, P., & Jain, R.** (2019). "Decision tree algorithms with rotation perturbation for privacy preservation in Indian telecom data." *Telecom Data Science Journal*, 14(2), 65-73.
13. **Agarwal, V., & Singh, D.** (2020). "Privacy-preserving decision trees using perturbation techniques in Indian healthcare systems." *Healthcare Data Analytics in India*, 29(3), 112-120.
14. **Sharma, V., & Patel, H.** (2020). "Enhancing privacy and classification accuracy using random projection perturbation in decision tree models for Indian financial data." *Journal of Data Privacy and Analytics*, 13(4), 101-110.