Not a Free Lunch, But a Cheap One: On Classifiers Performance on Anonymized Datasets

Mina Alishahi, Nicola Zannone

Not a Free Lunch, But a Cheap One: On Classifiers Performance on Anonymized Datasets

Mina Alishahi and Nicola Zannone

Eindhoven University of Technology, Eindhoven, The Netherlands
{m.sheikhalishahi, n.zannone}@tue.nl

Abstract. The problem of protecting datasets from the disclosure of confidential information, while published data remains useful for analysis, has recently gained momentum. To solve this problem, anonymization techniques such as k-anonymity, ℓ-diversity, and t-closeness have been used to generate anonymized datasets for training classifiers. While these techniques provide an effective means to generate anonymized datasets, an understanding of how their application affects the performance of classifiers is currently missing. This knowledge enables the data owner and analyst to select the most appropriate classification algorithm and training parameters in order to guarantee high privacy requirements while minimizing the loss of accuracy. In this study, we perform extensive experiments to verify how the classifiers performance changes when trained on an anonymized dataset compared to the original one, and evaluate the impact of classification algorithms, datasets properties, and anonymization parameters on classifiers' performance.

Keywords: Privacy-preserving, k-anonymity, ℓ-diversity, t-closeness, classifiers comparison.

1 Introduction

Classification is the task of identifying to which category (class) a new observation belongs based on a training set of observations whose category membership is known beforehand. Nowadays, data classification algorithms (classifiers) are widely used in many real-world applications, including but not limited to, face and speech recognition, text analysis, fraud and anomaly detection, recommendation system, weather forecasting, and medical image analysis [1, 27].

Classifiers are typically trained over a corpus of training data that is directly accessible by the data analyzer. However, in many real-world scenarios the training data is generated and governed by different entities who are unwilling to share their data with the analyzer. This is because the data might contain privacy-sensitive information, and its disclosure might raises privacy concerns [23]. To solve this issue, a large body of research has investigated how to train a practically useful classifier, while preserving individuals' privacy. Existing solutions can be categorized into two main groups. One category comprises cryptographic-based approaches in which the classifier model is securely computed. The main drawback of these approaches is that they are not scalable, and they are designed only for a specific classification algorithm [10]. The other category of

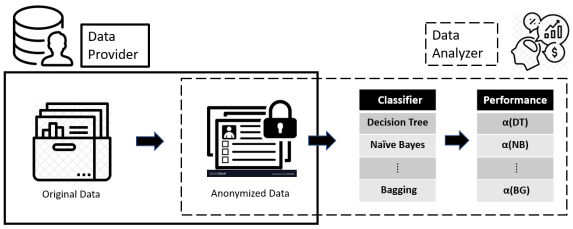


Fig. 1: Reference Architecture.

solutions comprises data anonymization techniques, in which the values in the data are replaced with a more general representation before the dataset is published. This study focuses on the second category, *i.e.*, the application of data anonymization techniques in data classification, where one entity (data provider) owns the data and the other entity (data analyzer) is interested in training a classifier on this data. The data analyzer does not know which classifier outperforms the other classifiers on the shared (anonymized) data. This knowledge would enable the analyzer to decide which classifier to be trained based on the anonymization technique employed, the dataset properties, and the desired performance metric. The problem addressed in this work can be defined as follows.

Problem Statement: A data provider wants to release a dataset T to a data analyzer for modeling a classifier on this data. Each record $\vec{x_i}$ in T is an $(n+1)$-dimensional vector $\vec{x_i} = (v_1, v_2, \ldots, v_n, C_i)$, where the first n elements are the attribute-values and the last element is the class label of that record. The data provider wants to protect the dataset against linking an individual to sensitive information using an anonymization approach. Consider, for instance, a health center wants to share the patients' records with a medical research center for identifying the causes and symptoms of a new disease. However, the shared information might potentially raise the patients' privacy concerns. Thus, the health center only is ready to share the data if no individual record can be linked to the corresponding patient, *i.e.*, dataset can only be revealed in anonymized version.

The data analyzer who has access to the anonymized version of data is interested in training a classifier. In our example, the medical research center wants to model a classifier over the symptoms of a disease to predict whether a new patient suffers from this disease or not. However, the data analyzer has no knowledge which classifier should be selected on the published anonymized data. An overview of this communication model with the following two entities is presented in Figure 1:

- *Data provider* who shares the anonymized table of data respecting his/her privacy requirements, *e.g.*, the published table satisfies 3-anonymity.
- *Data analyzer* who uses the anonymized table of data as training dataset to train a classifier.

We assume that the data analyzer has time limitation and thus is not able to evaluate the performance of different classifiers on the published anonymized data.

Our Contribution: To solve the aforementioned problem, we investigate the classifiers performance on anonymized datasets via answering the following research questions:

RQ1: How the performance of different classification algorithms changes when trained on anonymized datasets?

RQ2: Which classifiers are more affected by the employment of anonymization techniques?

RQ3: Which dataset properties affect the performance of classifiers trained on anonymized datasets?

RQ4: How the classifiers performance is affected by changing the anonymization parameters?

To answer these questions, we compute the performance of eight well-known classification algorithms, namely Decision Tree, Naïve Bayes, k Nearest Neighbors, Support Vector Machine, Random Forest, Logistic Regression, AdaBoost, and Bagging, over 10 benchmark datasets (original and anonymized versions). The performance of classifiers is measured using accuracy, precision, recall, and F1-score metrics. The selected anonymization approaches are k-anonymity [21, 25], ℓ-diversity [15], and t-closeness [12]. The contribution of this work can be summarized as follows:

- We provide insight on the difference between classifiers performance trained over original and anonymized datasets (RQ1). We show that some classifiers significantly outperform the others in this regard.
- We compare the classifiers performance on anonymized datasets and highlight which classifiers outperform the others in terms of the associated performance metric and anonymization approach (RQ2). We show that this outperformance is statistically significant and provide insight on the origin of this difference.
- We investigate the impact of dataset properties, *i.e.*, dataset size, the number of attributes, and the number of class labels on classifiers performance (RQ3). We show which dataset property and to what extent has impact on classifiers performance on anonymized dataset.
- We evaluate the effect of anonymization parameters on classifiers performance through the enforcement of different values of k, ℓ, and t (RQ4). We show that the variation of anonymization parameter has negligible impact on the trend of classifiers performance.
- Based on our experimental results, we draw recommendations to guide data providers and analyzers in the selection of the classification algorithm to be used (cheap lunch).

Outline: The remainder of this paper is organized as follows. The next section presents the background. Section 3 explains our methodology and the setup of the experiments. Section 4 presents the experimental results, whereas Section 5 discusses our findings. Section 6 discusses related work. Finally, Section 7 concludes the paper and provides directions for future work.

2 Preliminaries

This section introduces the anonymization techniques and classification algorithms considered in this work.

2.1 Anonymization Techniques

For our study we consider three well-known anonymization techniques: k-anonymity [21, 25], ℓ-diversity [15], and t-closeness [12]. We assume a dataset comprising a set of

records, where each record corresponds to one individual. Each record is described by a number of attributes, which can be divided into three categories: 1) *identifiers* that univocally identify the individuals, *e.g.*, social security number, 2) *quasi-identifier* attributes whose values taken together can be used to potentially identify an individual, *e.g.*, zip-code, birth-date, and gender, 3) *sensitive* attributes that an adversary is not allowed to discover the values of that attribute for any individual, *e.g.*, a patient's disease or an employer's salary.

k-anonymity: A release of data is said to satisfy k-anonymity if each record in the release cannot be distinguished from at least $k-1$ other records in the release with respect to quasi-identifiers [25]. k-anonymity is susceptible to some attacks, *e.g. homogeneity* and *background knowledge* attacks.

ℓ-diversity: The ℓ-diversity model addresses some of the weaknesses of k-anonymity. In particular, k-anonymity does not protect the values of sensitive attributes, specifically when the values in a group are identical. To address this drawback, the ℓ-diversity model introduces constraints on intra-group diversity for sensitive attributes [15]. The ℓ-diversity does not consider the semantic closeness of the distinct values in a sensitive attribute. This problem is addressed by t-closeness.

t-closeness: An equivalence group is said to satisfy t-closeness if the distance between the distribution of a sensitive attribute in this group and the distribution of the attribute in the whole table is not greater than a given threshold t. A dataset is said to satisfy t-closeness if all equivalence classes satisfy t-closeness [12].

2.2 Classification Algorithms

We investigate the performance of eight well-known classification algorithms, namely Decision Tree (DT), Naïve Bayes (NB), k-Nearest Neighbors (kNN), Support Vector Machine (SVM), Random Forest (RF), Logistic Regression (LR), AdaBoost (AB), and Bagging (BG). Next, we briefly introduce these classifiers (for more detail refer to [1]).

Decision Trees (DT) are classification algorithms with a tree-based structure drawn upside down with its root at the top. Each internal node represents a test/condition based on which the tree splits into branches/edges. The end of the branch that does not split anymore (respecting some stopping criteria) is the decision/leaf. The paths from root to leaf represent classification rules. One advantage of DTs is the comprehensibility of the classification structures. This enables the analyzer to verify which attributes determined the final classification. The drawback is that DTs might be non-robust for datasets with a large number of attributes.

Naïve Bayes (NB) algorithms are statistical classifiers based on the Bayes Theorem for calculating probabilities and conditional probabilities. It makes use of all attributes contained in the data, and analyses them individually as though they are equally important and independent (naïve assumption) from each other. Naïve Bayes model is easy to build and particularly useful for very large data sets.

k-Nearest Neighbors (kNN) algorithm is a non-parametric instance-based model that classifies a new instance based on the class of the majority of its k nearest neighbors w.r.t. a given training dataset. To obtain the nearest neighbors for each data point, kNN

uses a measure to compute the distance between pairs of data points (*e.g.*, Euclidean distance). The advantage of kNN lies in its simplicity, while computation time is usually high since all training data has to be revisited for classifying a new instance.

Support Vector Machine (SVM) is a linear modelling with instance-based learning. The algorithm selects a small number of critical boundary instances from each category (class labels) and builds a linear discriminate function that separates them as widely as possible. In the case that no linear separation is possible, the technique of kernel is used to automatically project the training instances into a higherdimensional space and to learn a separator in that space. The SVMs have the advantage of generalization, and also standout for their robustness to high dimensional data. The drawback of the SVMs is the difficulty of model interpretation and the sensibility to parameter tuning.

Random Forest (RF) is an ensemble learning method for classification which operates by constructing a multitude of decision trees at training time and outputting the class that is the mode of the classes of the individual trees. RF corrects for decision trees' habit of overfitting to their training set. The RF has also been recognised to be among the most accurate classifiers.

Logistic Regression (LR) is a statistical model that is largely employed in statistical data analysis to classify binary dependent variables. In regression analysis, logistic regression (or logit regression) is used to estimate the parameters of a logistic model returning the probability of occurrence of a class. To this end, the LR classifier builds a logit variable that contains the natural log of the odds of the class occurring or not. Then, a maximum likelihood estimation algorithm is applied to estimate the probabilities.

AdaBoost (AB) is a technique that builds an ensemble of classifiers (generally Decision Trees) sequentially, one classifier at a time, until the predefined number of classifiers is reached. Each subsequent classifier is trained on a set of samples with weights to emphasize the instances misclassified by the previous classifiers. The ensemble decision is made by weighted voting. The weights are determined by the individual accuracy. AB has been found to be very useful but too sensitive to noise in the data.

Bagging (Bootstrap aggregating) (BG) is an ensemble learning model designed to improve the stability and accuracy of classification algorithms. It considers several homogeneous weak learners, where each weak learner has been trained independently from the others. Then, the bagging model combines them with an averaging process.

3 Experimental Methodology

The experimental analysis aims to compare the classifiers performance when they are trained over anonymized datasets. This comparison evaluates how the employment of anomyzation techniques affect classifiers performance (**RQ1**), and investigate how the choice of classification algorithm (**RQ2**), dataset properties (**RQ3**), and anonymization parameters (**RQ4**) influence the classifiers performance. Next, we present the experimental setup, the datasets used for the experiments and the evaluation approach.

Experimental Setup: We use the implementation of classification algorithms provided by Scikit-learn library[1] with their default parameters. Turning a dataset into an anonymized

[1] `https://scikit-learn.org`

Dataset	# Attributes	# Labels	# Instances
Adult	8	2	48000
Credit	15	239	690
Absent	21	17	740
Derma	33	6	366
Wine	12	2	4898
Network	22	4	1075
Bank	64	2	10,503
Optical	64	10	5600
Diabet	20	2	1151
Heart	13	2	299

Table 1: Datasets information

(k-anonymous, ℓ-diverse, and t-close) dataset is a complex problem in which finding the optimal partition is an NP-hard problem. In this study, we employ the *Mondrian* algorithm, which uses a greedy search algorithm to recursively partition the domain space into regions [11].[2]

Datasets: The classifiers were trained over ten datasets selected from the UCI Repository.[3] Table 1 summarizes the statistics of the selected datasets.

Adult: The dataset contains 48842 instances described by 14 attributes (both numerical and categorical) such as *age*, *occupation*, *education*, and *working class*. The class attribute represents their income, which has two possible values: '$> 50K$' and '$< 50K$'.

Credit: The dataset contains 690 instances of clients' information at a bank described by 15 attributes (both numerical and categorical) such as *age* and *background behavior*, which are used to predict the score of a requester with 239 possibilities (based on this score it is decided whether the credit card application should be accepted, revised, or denied).

Absenteeism at work (Absent): The dataset contains 740 instances described by 21 attributes (both categorical and numerical), *e.g.*, *age*, *education*, *average workload per day*, and *social smoker*. The class label denotes the hours that a new employee might be absent in a month in the future. While the absent hours can vary from one hour to 160 hours in a month, the current dataset only shows 17 distinct values for absent hours (from 20 to 36 hours).

Dermatology (Derma): This dataset contains 366 instances described by 33 numerical attributes such as *age*, *family history*, *knee and elbow involvement*, which are used to predict the type of Eryhemato-Squamous disease as a real problem in dermatology (skin disorders). The majority of attributes take their values from the set $\{0, 1, 2, 3\}$.

Wine Quality (Wine): The dataset contains 4898 instances of wine samples described by 12 numerical attributes such as *pH value*, *citric acid*, *total sulfur dioxide*, which are used to predict the wine quality (good or bad).

Optical Burst Switching (OBS) Network (Network): The dataset contains 1075 instances of Burst Header Packet (BHP) flood attacks in Optical Burst Switching networks (OBS), described by 22 numerical attributes such as *Average Delay Time per Second*, *Percentage*

[2] The code used for our experiments is available at `https://github.com/minaalishahi/classifiersperformance`.

[3] `https://archive.ics.uci.edu/ml/datasets/`

of Lost Packet rate, to classify the strategy against an attack according to network nodes behavior into four classes as NB-No Block, Block, No Block, NB-Wait (NB= Not Behaving correctly).

Polish Companies Bankruptcy (Bank): The dataset contains 10503 instances of emerging markets around the world described by 64 numerical attributes such as *current assets/short-term liabilities*, *profit on sales/total sales*, to predict whether a Polish company will face bankruptcy or not.

Optical Digits (Optic): The dataset contains 5620 handwritten digits written by 43 persons. Each record is a matrix of 8x8 where each element is an integer in the range $\{0, \ldots, 16\}$ and the class label is one of the integer number in the set $\{0, \ldots, 9\}$.

Diabetic Retinopathy Debrecen (Diabet): The dataset contains 1151 instances of Messidor image set described by 20 attributes such as *the diameter of the optic disc* and *quality assessment*. All features represent a detected lesion, or a descriptive feature of an anatomical part or an image-level descriptor. The class label represents whether an image shows signs of diabetic retinopathy or not.

Heart Failure Clinical Record (Heart): The dataset collected the medical records of 299 patients who had heart failure during a pre-determined number of days (say follow-up period). The dataset is described by 13 attributes such as *age*, to predict whether a patient dies (Boolean) during the follow-up period.

Note that for all datasets, it is assumed that the explicit identifier attributes have been removed from the data, the attribute representing the class label is considered as the sensitive attribute, and the remaining attributes are considered quasi-identifier attributes. In all datasets, categorical attributes either were removed or the categorical values were replaced with numerical values (when the conversion is a valid assumption). This is because in the application of anonymization approaches over categorical attributes, in general, the presence of an expert in the field is required to provide the taxonomy trees for generalization.

Evaluation Approach: We assess the classifiers performance in terms of *Accuracy*, *Precision*, *Recall*, and *F1-score*, which are defined based on True Positives (TP), True Negatives (TN), False Positives (FP), and False Negatives (FN) values. True Positives (TP) are the correctly predicted positive values, *i.e.*, the value of actual class is positive and the value of predicted class is also positive. True Negatives (TN) are the correctly predicted negative values, *i.e.*, the value of actual class is negative and value of predicted class is also negative. False Positives (FP) are the non-correctly predicted positive values, *i.e.*, the value of actual class is negative and the predicted class is positive. False Negatives (FN) are the non-correctly predicted negative values, *i.e.*, the value of actual class is positive but the predicted class is negative. To compute these values for datasets with multiple class labels, we first compute the TP, TN, FP, and FN values for each individual class label against the remaining class labels. Then, the average of these values over all class labels is returned as final TP, TN, FP, and FN. Accuracy is the ratio of correctly predicted observations over the total number of observations. Precision is the ratio of correctly predicted positive observations to the total number of predicted positive observations. Recall is the ratio of correctly predicted positive observations to all observations in actual class positive. F1-Score is the weighted average of Precision and Recall. Therefore, this

score takes both false positives and false negatives into account. Formally:

$$\text{Accuracy} = \frac{\text{TP + TN}}{\text{TP + FP+ FN + TN}} \qquad \text{Precision} = \frac{\text{TP}}{\text{TP + FP}}$$
$$\text{Recall} \quad = \frac{\text{TP}}{\text{TP + FN}} \qquad \text{F1-Score} = 2 \cdot \frac{\text{Recall} \cdot \text{Precision}}{\text{Recall + Precision}}$$

To answer research questions RQ1, RQ2, and RQ3, we compute classifier performance in terms of accuracy, precision, recall, and F1-scores on the original, 3-anonymity, 2-diversity, and 0.2-closeness datasets. To answer RQ4, we compute classifier accuracy when anonymization parameters, *i.e.*, k, ℓ, and t, vary.

Criteria for RQ1: To investigate to what extent the use of anonymization techniques affect a classifier performance, we compare the performance of classifiers trained on anonymized datasets with the one of classifiers trained on the corresponding original datasets by computing the *performance ratio*. For each considered performance metric, the performance ratio is computed by dividing the performance of a classifier trained over an anonymized dataset by the performance of the classifier trained over the corresponding original datasets. As we are interested on the average performance of classification algorithms, we aggregate performance metrics over the 10 datasets.

To verify whether the differences between classification algorithms are statistically significant, we use a non-parametric statistical test, named *Wilcoxon* test [26]. The Wilcoxon test can be adapted to our problem as follows.

Definition 1 (Wilcoxon test). *Given two classification algorithms, let d_i be the signed difference between the performance scores of the classifiers obtained by applying each algorithm on a given dataset for a given privacy level. The differences d_i ($1 \leq i \leq N$ where N is the number of anonymized datasets to which the classification algorithms are applied) are ranked based on the absolute values (average rank is assigned for equal performances). Let R^+ denote the sum of the ranks for datasets and privacy level on which $d_i > 0$, and let R^- be the sum of the ranks for datasets and privacy level on which $d_i < 0$ (dividing the sum of the ranks for which $d_i = 0$ evenly),* i.e.*,*

$$R^+ = \sum_{d_i>0} rank(d_i) + \frac{1}{2}\sum_{d_i=0} rank(d_i) \ , \ R^- = \sum_{d_i<0} rank(d_i) + \frac{1}{2}\sum_{d_i=0} rank(d_i)$$

Let $T = \min(R^+, R^-)$, then

$$z = \frac{T - \frac{1}{4}N(N+1)}{\sqrt{\frac{1}{4}N(N+1)(2N+1)}}$$

*is approximately distributed normally. Under this condition, the difference between the accuracy distribution of the two classification algorithms is statistically significant (*i.e.*, the null hypothesis is rejected) if the p-value is less than or equal to a given significance level σ. In our experiments, we require a 95% confidence interval, which corresponds to $\sigma = 0.05$ (*i.e.*, the null-hypothesis can be rejected if z is smaller than -1.96).*

Criteria for RQ2: To investigate the impact of the adoption of anonymization techniques on classifiers performance, we first compute the average of classifiers' performance for each

performance metric over the 10 datasets. Then, we employ a non-parametric statistical test, named the *Friedman* test [5], on these results. This test was designed to compare classification algorithms over multiple datasets, and the outcome determines whether the algorithms are equal in terms of performance or not. If the classification algorithms exhibit different performance, p-values (in *Holm* methodology) are used to order them based on their performance. The Friedman test can be adapted to our problem as follows.

Definition 2 (Friedman Test). *Given n classification algorithms and m datasets, let r_{ij} denote the rank of j-th algorithm on the i-th dataset. The Friedman test compares the average ranks of algorithms,* i.e.*, $R_j = \frac{1}{m}\sum_i r_{ij}$. Under the null-hypothesis stating that all algorithms are equivalent (their average ranks R_j are close), the Friedman statistic*

$$\chi_F^2 = \frac{12m}{n(n+1)}\Big(\sum_j R_j^2 - \frac{n(n+1)^2}{4}\Big)$$

is distributed according to the well-known χ_F^2 distribution with $n-1$ degrees of freedom, for n and m big enough ($n \geq 5$, $m \geq 10$). To decide if the classifiers' performance is significantly different, the Friedman test is used as:

$$F_F = \frac{(m-1)\chi_F^2}{m(n-1)-\chi_F^2}$$

*where the probability distribution can be approximated by a F-distribution with $n-1$ and $(n-1)(m-1)$ degrees of freedom. The table of critical values can be found in statistical books [9]. The difference between the performance distribution of all classification algorithms is statistically significant (*i.e.*, the null hypothesis is rejected) if the p-value is less than or equal to a given significance level σ. In our experiments, we require a 95% confidence interval, which corresponds to $\sigma = 0.05$.*

If the null-hypothesis is rejected, we need to determine where the differences truly came from. To answer this question, generally, a post-hoc statistical test named the *Nemenyi* test is used. However, in some cases, the Nemeneyi test is not able to detect why the Friedman test has rejected the null-hypothesis. In this regard, the tests like *Bonferroni* or *Holm* are more powerful in detecting where the difference comes from using a control variable [5, 8]. In this study, we use the average performance of classifiers as the required control variable. This test assigns a score to classifiers by comparing the respective p-values of a classifier compared to the others such that the classifier with higher score has the better performance for that specific assessment.

Criteria for RQ3: Research question RQ3 aims to understand the effect of dataset properties on the performance of classifiers when trained on anonymized datasets. To this end, we investigate how the classifier accuracy, precision, recall, and F1-scores vary over anonymized datasets with different sizes, the number of attributes, and the number of class labels. To assess this variation, for each dataset and each performance metric we compute the average performance over all classification algorithms. To determine whether the impact of a specific dataset property is significant, we compare the distribution of classifiers' performance based on aggregated results for that specific property.

Anonymity	Metric	Classifier							
		DT	NB	kNN	SVM	RF	LR	AB	BG
3-anonymity	Accuracy	0.93	0.90	1.00	1.03	0.85	0.91	**1.10**	0.84
	Precision	0.88	0.89	1.04	1.09	0.90	0.90	**1.20**	0.86
	Recall	0.87	0.80	1.00	1.03	0.85	0.91	**1.10**	0.84
	F1 score	0.87	0.81	1.03	1.11	0.87	0.90	**1.22**	0.85
2-diversity	Accuracy	0.83	0.86	0.93	0.93	0.75	0.86	**1.05**	0.75
	Precision	0.77	0.86	0.95	0.96	0.77	0.85	**1.08**	0.75
	Recall	0.78	0.77	0.93	0.93	0.75	0.86	**1.05**	0.75
	F1 score	0.77	0.79	0.95	0.99	0.76	0.85	**1.15**	0.75
0.2-closeness	Accuracy	0.71	0.61	0.70	0.71	0.62	0.68	**0.77**	0.62
	Precision	0.55	0.54	0.66	0.68	0.54	0.61	**0.71**	0.54
	Recall	0.65	0.57	0.70	0.71	0.62	0.68	**0.77**	0.62
	F1 score	0.59	0.57	0.69	0.72	0.58	0.63	**0.76**	0.58

Table 2: Classifier performance ratio.

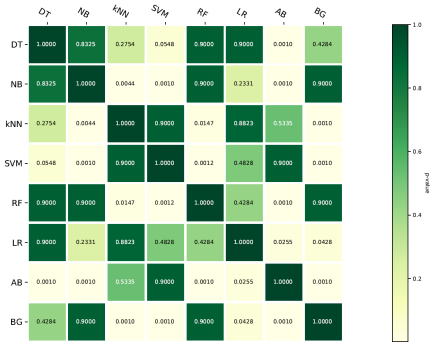


Fig. 2: Heatmap of Wilcoxon test.

Criteria for RQ4: To investigate the effect of anonymization parameters, *i.e.*, k, ℓ, and t in k-anonymity, ℓ-diversity, and t-closeness on classifiers performance, we compute classifier accuracy, precision, recall, and F1-scores on anonymized datasets for $k \in \{3, 4, 5, 6\}$, $\ell \in \{2, 3, 4, 5\}$, and $t \in \{0.2, 0.3, 0.4, 0.5\}$.

It is worth noting that the selection of value ℓ from the set $\{2, 3, 4, 5\}$ requires that the dataset under analysis contains at least 5 distinct class labels. Out of the 10 selected datasets, the datasets satisfying this requirement (*i.e.*, with more than five class labels) are Credit, Absent, Derma, and Optical datasets. Due to the lack of space, we only present the results for the three datasets with the greatest number of class labels.

4 Experimental Results

This section presents the results of our experiments.

RQ1: How the performance of different classification algorithms changes when trained on anonymized datasets? The performance ratio for the considered classification algorithms (aggregated over the 10 datasets) is reported in Table 2. We can observe that the AB classifier shows the highest ratio for all performance metrics and anonymization techniques. The performance is even higher when AB is trained on 3-anonymous and 2-diverse datasets compared to when it is trained on the original dataset (the ratio is greater than 1). This performance improvement can also be observed for the kNN and SVM classifiers. The worst classifiers in terms of performance ratio are NB and BG.

We used the Wilcoxon test to verify the statistical significance of these differences. Fig. 2 depicts the heatmap of mutual comparison of classifier performance ratios in terms of p-values over accuracy results. The lower p-value (lighter color) shows more confidence in rejecting the null-hypothesis (*i.e.*, more different performance). It can be observed that the null-hypothesis of the Wilcoxon test is rejected in the mutual comparison of AB with DT, NB, RF, BG, and LR with high confidence (small p-value), *i.e.*, the AB classifier shows a different behavior compared to the other algorithms (except from SVM and kNN).

RQ2: Which classifiers are more affected by the employment of anonymization techniques? The average performance of classifiers is reported in Table 3. From the

Anonymization	Metric	Classifier							
		DT	**NB**	**kNN**	**SVM**	**RF**	**LR**	**AB**	**BG**
Original	Accuracy	82.61	65.73	72.06	71.87	**85.73**	80.03	65.40	85.47
	Precision	82.60	81.45	68.39	71.72	83.74	77.25	58.89	**84.14**
	Recall	84.03	73.65	72.06	71.87	**85.73**	80.03	65.40	85.47
	F1 score	82.95	71.76	68.38	67.00	84.25	77.21	59.67	**84.41**
3-anonymity	Accuracy	73.03	59.23	71.10	**73.30**	73.28	72.10	67.61	72.04
	Precision	71.48	70.94	68.03	71.94	**72.65**	68.76	64.08	71.56
	Recall	73.03	59.23	71.10	**73.30**	73.28	72.10	67.61	72.04
	F1 score	71.38	58.80	68.01	71.79	**72.39**	68.77	62.96	71.27
2-diversity	Accuracy	64.27	56.71	65.28	66.04	64.75	**68.11**	64.36	63.80
	Precision	62.34	**68.63**	62.04	64.34	63.81	65.15	58.70	62.76
	Recall	64.27	56.71	65.28	66.04	64.75	**68.11**	64.36	63.80
	F1 score	62.47	56.30	62.18	64.21	63.82	**64.85**	60.12	62.81
0.2-closeness	Accuracy	50.97	37.99	49.74	51.44	50.41	**52.93**	51.76	50.41
	Precision	41.96	**46.12**	43.94	41.44	41.95	44.85	42.97	42.71
	Recall	50.97	37.99	49.74	51.44	50.41	**52.93**	51.76	50.41
	F1 score	45.72	36.20	45.89	45.60	45.57	46.14	**46.24**	45.74

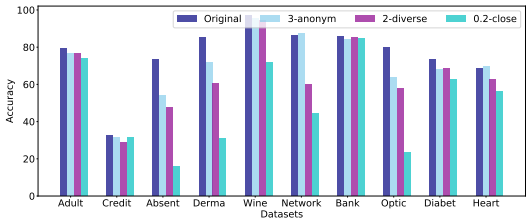
Table 3: Average performance for the considered classifiers. The highest value in each row (representing the best performance) is highlighted in bold.

	Accuracy	**Precision**	**Recall**	**F1-score**
Original	BG	BG	BG	LR
3-anonymity	LR	RF	LR	LR
2-diversity	LR	NB	LR	LR
0.2-closeness	LR	NB	LR	LR

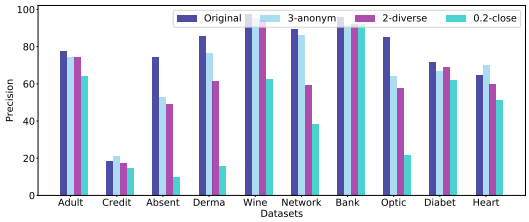
Table 4: The best scored classifiers using the Holm methodology.

table, we can observe that 1) there is not a single classifier that outperforms all other classifiers for all performance metrics; 2) while for specific metrics and anonymization techniques, some classifier outperforms the others (highlighted values), in some cases the results are very close (*e.g.*, the accuracy of the SVM and RF on 3-anonymous datasets).

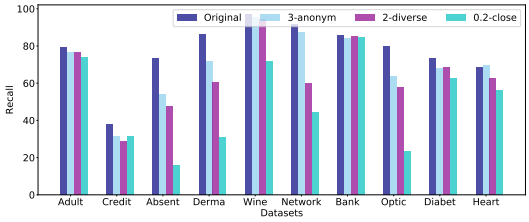
To get a better insight on these observations, we performed the Friedman test on average results. The null-hypothesis of this test was rejected in our experiments meaning that all classifiers are not equal in terms of performance and there is a significant difference. To determine where this difference comes from, in Tables 5, 6, 7, and 8 (in Appendix) we report respectively the Holm scores (simply score from now on) of classifiers with respect to accuracy, precision, recall, and F1-score. The higher scores represent higher performance for the associated classification algorithm and metric. Table 4 summarizes the best classifiers for each performance metric and anonymization technique according to the Holm scores. We can observe that the LR classifier outperforms the other classifiers in terms of accuracy, recall, and F1-score over anonymized datasets. However, in terms of precision, the RF and NB classifiers are the best scored ones over 3-anonymous and 2-diverse (and 0.2-close) datasets.



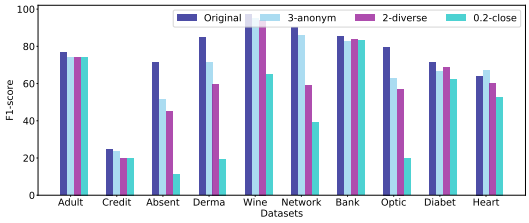
(a) Accuracy



(b) Precision



(c) Recall



(d) F1-score

Fig. 3: Average performance of the selected classifiers for each dataset.

RQ3: Which dataset properties affect the performance of classifiers on anonymized datasets? To evaluate the impact of datasets on classifiers performance, we computed the accuracy, precision, recall, and F1-scores for each individual dataset averaged over all classifiers performance when trained on original and anonymized datasets. Figures 3a, 3b, 3c, and 3d show respectively the average accuracy, precision, recall, and F1-scores on all classifiers results. From these results, we can infer that the number of attributes has no di-

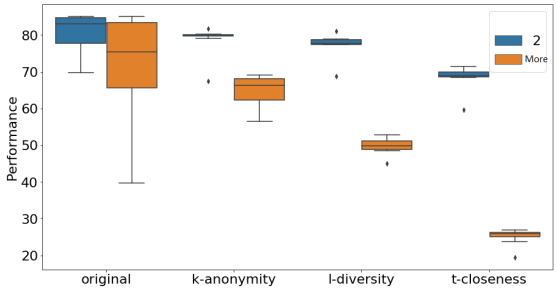


Fig. 4: Performance of the selected classifiers for datasets with two class labels vs. multi-class labels.

rect impact on classifiers performance. The Bank and Optic datasets with an equal number of attributes show completely different impact on the classifiers' performance. The size of datasets also shows no direct impact on classifiers performance. It can be observed that the Credit and Absent, which have a comparable number of records, show different impact on classifiers performance. This can be seen for Network and Diabet datasets as well.

The number of class labels, differently from the previous properties, shows a considerable impact on classifiers performance on anonymized datasets. The datasets with two class labels, *i.e.*, Adult, Wine, Bank, Diabet, and Heart datasets, show better and more stable performance in both original and anonymized versions. On the other hand, the Credit and Absent datasets, which have multi class labels (239 and 17, respectively) result in poor performance in all scenarios.

To gain better insight on the impact of the number of class labels on classifiers performance, we compare the performance of binary classifiers (*i.e.*, classifiers trained on datasets which have two class values, namely Adult, Wine, Bank, Diabet, and Heart), and multi-class classifiers (*i.e.*, classifiers trained on datasets which have multiple class values, namely Credit, Absent, Derma, Network, and Optimal). Fig. 4 shows the performance distribution of binary (blue) and multi-class (orange) classifiers when trained on the original data as well as when 3-anonymity, 2-diversity and 0.2-closeness are applied. Each box represents the distribution over five datasets (the blue ones over 2-labeled and the orange ones over more-labeled datasets) and over classifiers' accuracy (*i.e.*, each box shows the average value computed over 5 accuracy values). We can observe that on the original datasets, the blue and orange boxes have close median values (*i.e.*, small difference). The difference in the distribution between the boxes (and median values) increases when anonymyzation is applied.

RQ4: How the classifiers's performance is affected by changing the anonymization parameters? To evaluate whether a classifier' performance is affected by the values of k, ℓ, and t, we computed the performance metrics for all considered classifiers for different values of k, ℓ, and t on Credit, Absent, and Optic datasets. The results are reported in Figures 5 to 16 (in Appendix). We observed that except from the kNN, NB, and RF classifiers showing a negligible difference for some values of k, ℓ, and t, the other classifiers preserve the performance trend when the values of k, ℓ, and t vary.

5 Discussion

In this work, we selected eight well-known classification algorithms and trained them over 10 datasets manipulated using three anonymization techniques, *i.e.*, k-anonymity, ℓ-diversity and t-closeness. We assessed how the employment of these anonymization techniques affect classifiers accuracy, precision, recall and F1-score, and investigated whether these effects depends on the chosen classification algorithm, dataset properties and anonymization parameters. We now discuss some interesting findings and report the threats to validity.

Findings: The performance of the considered classifiers when trained on anonymized datasets in comparison to when trained on the original datasets show that the AB classifier returns the most similar performance results between the original and anonymized datasets. The AB (outperformance) difference is significant compared to DT, NB, RF, LR, and BG, but it is not significant compared to the kNN and SVM classifiers. This result suggests the employment of AB, kNN, and SVM on anonymized data when these classifiers perform accurately when trained on the original data.

Our results also show that there is not a single classifier that significantly outperforms the other classifiers for all performance metrics. Nonetheless, we observe that LR is the best classifier in terms of accuracy, recall, and F1-score on anonymized datasets, whereas NB and RF are superior in terms of precision.

Among dataset properties, the number of class labels considerably affects the classifiers performance on anonymized dataset. The other properties, *i.e.*, dataset size and number of attributes, have negligible (or no) impact on classifiers' performance. This outcome is independent from the performance metric considered.

The variation of anonymization parameters, apart from some exceptions with a negligible difference, does not affect the *trend* of classifiers' performance. This outcome allows us to generalize (to some extent) our findings on 3-anonymity, 2-diversity, and 0.2-closeness to other k-anonymous, ℓ-diverse, and t-close datasets.

Threats to validity: Several variations have been proposed in the literature for some classification algorithms. For instance, the polynomial and RBF kernel-based SVM are two types of SVM classifiers, and the Bernoulli and Gaussian are two types of Naïve Bayes classifiers. Moreover, each classification algorithm has one or more configuration parameters, *e.g.*, the value of k in kNN classifier or the number of trees in the AB classifier. The selection of other types of a classification algorithm or tuning its configuration parameters might affect the performance. Beside the selection of classification algorithms, the selection of alternative datasets (*e.g.*, datasets with million records), the selection of other anonymization algorithm (*e.g.*, Incognito instead of Mondrian) and anonymization parameters (*e.g.*, $k, \ell > 10$) might provide different results.

To mitigate the effect of aforementioned validity threats on our findings, we have selected the classification algorithms and their configuration parameters as suggested by the Scikit-learn library. These parameters have been tuned to their highest performance (for the majority of cases) to provide a fair comparison among classifiers. The datasets have been selected to meet the diverse requirements in terms of the dataset size, the number of attributes, and the number of class labels. The selected anonymization algo-

rithm (*i.e.*, Mondrian algorithm) has shown higher performance performance compared to other algorithms [2].

While the control variables of this study were carefully chosen, we expect that the selection of the alternative classification algorithm types (and parameter tuning), dataset, and anonymization technique, will not considerably affect (some of) our findings. For instance, we expect that the anonymization process makes the datasets linearly separable resulting in (generally) higher performance for the LR classifier compared to other classifiers. Also, the anonymization approaches tend to perform poorly when several records with multi-class labels are grouped together. These claims and the other aforementioned findings of this study need more work to investigate the results on a wider range of datasets, different types of classifiers, other adjustments of the classification algorithms parameters (with the use of cross-validation or greedy search for parameter adjustment), and different anonymization algorithms and approaches.

6 Related Work

Data anonymization has become a widely investigated research direction in an effort to protect individuals' privacy when data is supposed to be released publicly. k-anonymity, which was proposed as the initial definition of anonymity [21, 25], has been extended to new additional constraints such as ℓ-diversity [15] and t-closeness [12]. The proposed approaches have been optimized in terms of a generic measurement with no emphasis on the utility of anonymized data for classification. For instance, in [13], a novel anonymization technique, named *slicing* is proposed, which handles high-dimensional data and improves the data structure compared to preliminary generalization technique. Nergiz et al. [20] suggest a hybrid generalization technique which by data relocation provides a trade-off between utility and privacy. The design and application of anonymization techniques when data utility is critical for data classification has been investigated in several studies. Ye et al. [28] propose a new anonymization approach based on rough set theory, which measures data quality for accurate classifiers construction and guides the anonymization process through combining rough set theory and attribute value taxonomies. In [22] and [18], a trade-off between privacy and data utility is achieved by appropriate feature suppression to publish the anonymized dataset to be used for classification. The focus of these approaches is to create anonymized datasets in which the features effectively discriminate the class labels. However, the performance of a specific classifier over the anonymized datasets has not been sufficiently investigated. In [6], anonymization (k-anonymity) is embedded within the decision tree induction process, which provides better accuracy compared to the scenario in which data is first anonymized and then used for inducing the tree. A similar methodology has been proposed in [4] for embedding anonymization within the association rule mining algorithm. Mostly, the data perturbation for building a specific classification algorithm is performed with the use of differential privacy approach [7]. In this set of approaches, the perturbation process is task-specific dependent on the classification algorithm, *e.g.*, for constructing Naïve Bayes classifier.

Another research steam focuses on the effect of dataset features on data anonymization. In [19], for instance, novel methods are proposed to identify which features of documents need to change and how they must be changed to accomplish document anonymization.

Brikke and Shmatikov [3] investigate whether generalization and suppression of quasi-identifier features offer any benefit over trivial sanitization which simply separates quasi-identifier features from sensitive ones. In [17] and [16], a series of experiments is conducted to study the effect of anonymization on classifiers accuracy by applying four different classifiers to the Adult dataset (before and after being perturbed).

While some work in the literature compares the impact of privacy in the context of classifier training, *e.g.*, over encrypted data [24] and under differential privacy [14], to the best of our knowledge, no prior work has provided a comparison of the performance achieved by different classifiers when trained on different datasets before and after being anonymized.

7 Conclusion

This paper investigate the performance achieved by classifiers when they are trained over anonymized datasets. Accordingly, ten benchmark datasets have been anonymized using k-anonymity, ℓ-diversity, and t-closeness approaches. Then, eight well-known classifiers have been trained on these datasets, and their performance in terms of accuracy, precision, recall, and F1-score has been compared. Our experimental results show that depending on performance metric and dataset properties, one classifier might outperform the others.

In future work, we plan to provide a thorough comparison among a wider range of classifiers on a broader range of benchmark datasets with mixed types of attributes (*e.g.*, categorical). Moreover, we plan to evaluate the impact of classifiers' parameter configuration on classifiers performance trained on anonymized datasets.

Acknowledgement

This work has been supported by H2020 EU funded project SECREDAS [GA #783119].

References

1. Aggarwal, C.C.: Data Classification: Algorithms and Applications. Chapman and HallCRC (2014)
2. Ayala-Rivera, V., McDonagh, P., Cerqueus, T., Murphy, L.: A systematic comparison and evaluation of k-anonymization algorithms for practitioners. Trans. Data Privacy 7(3), 337–370 (2014)
3. Brickell, J., Shmatikov, V.: The cost of privacy: Destruction of data-mining utility in anonymized data publishing. In: International Conference on Knowledge Discovery and Data Mining. p. 70–78. ACM (2008)
4. Ciriani, V., di Vimercati, S.D.C., Foresti, S., Samarati, P.: k-anonymous data mining: A survey. In: Privacy-Preserving Data Mining: Models and Algorithms, pp. 105–136 (2008)
5. Demšar, J.: Statistical comparisons of classifiers over multiple data sets. J. Mach. Learn. Res. 7, 1– (2006)
6. Friedman, A., Schuster, A., Wolff, R.: k-anonymous decision tree induction. In: Knowledge Discovery in Databases. pp. 151–162 (2006)
7. Gong, M., Xie, Y., Pan, K., Feng, K., Qin, A.: A survey on differentially private machine learning. IEEE Comp. Intell. Mag. 15(2), 49–64 (2020)

8. Holm, S.: A simple sequentially rejective multiple test procedure. Scandinavian Journal of Statistics 6(2), 65–70 (1979)
9. James, G., Witten, D., Hastie, T., Tibshirani, R.: An Introduction to Statistical Learning: With Applications in R. Springer Publishing Company, Incorporated (2014)
10. Khodaparast, F., Sheikhalishahi, M., Haghighi, H., Martinelli, F.: Privacy preserving random decision tree classification over horizontally and vertically partitioned data. In: Conference on Dependable, Autonomic and Secure Computing. pp. 600–607 (2018)
11. LeFevre, K., DeWitt, D.J., Ramakrishnan, R.: Mondrian multidimensional k-anonymity. In: International Conference on Data Engineering. pp. 25–25 (2006)
12. Li, N., Li, T., Venkatasubramanian, S.: t-closeness: Privacy beyond k-anonymity and l-diversity. In: 23rd International Conference on Data Engineering. pp. 106–115. IEEE (2007)
13. Li, T., Li, N., Zhang, J., Molloy, I.: Slicing: A new approach for privacy preserving data publishing. IEEE Transactions on Knowledge and Data Engineering 24(3), 561–574 (2012)
14. Lopuhaä-Zwakenberg, M., Alishahi, M., Kivits, J., Klarenbeek, J., n van der Velde, G.J., Zannone, N.: Comparing classifiers' performance under differential privacy. In: International Conference on Security and Cryptography (SECRYPT) (2021)
15. Machanavajjhala, A., Kifer, D., Gehrke, J., Venkitasubramaniam, M.: l-diversity: Privacy beyond k-anonymity. ACM Transactions on Knowledge Discovery from Data 1(1), 3–es (2007)
16. Malle, B., Kieseberg, P., Holzinger, A.: DO NOT disturb? classifier behavior on perturbed datasets. In: Machine Learning and Knowledge Extraction. pp. 155–173 (2017)
17. Malle, B., Kieseberg, P., Weippl, E., Holzinger, A.: The right to be forgotten: Towards machine learning on perturbed knowledge bases. In: Availability, Reliability, and Security in Information Systems. pp. 251–266 (2016)
18. Martinelli, F., Alishahi, M.S.: Distributed data anonymization. In: Conference on Dependable, Autonomic and Secure Computing (DASC). pp. 580–586 (2019)
19. McDonald, A.W.E., Afroz, S., Caliskan, A., Stolerman, A., Greenstadt, R.: Use fewer instances of the letter "i": Toward writing style anonymization. In: Privacy Enhancing Technologies. pp. 299–318 (2012)
20. Nergiz, M.E., Gök, M.Z.: Hybrid k-anonymity. Computers & Security 44, 51 – 63 (2014)
21. Samarati, P.: Protecting respondents' identities in microdata release. IEEE Trans. on Knowl. and Data Eng. 13(6), 1010–1027 (2001)
22. Sheikhalishahi, M., Martinelli, F.: Privacy-utility feature selection as a privacy mechanism in collaborative data classification. In: Enabling Technologies: Infrastructure for Collaborative Enterprises. pp. 244–249 (2017)
23. Sheikhalishahi, M., Saracino, A., Martinelli, F., Marra, A.L.: Privacy preserving data sharing and analysis for edge-based architectures. International Journal of Information Security 1(2), 1–23 (2021)
24. Sheikhalishahi, M., Zannone, N.: On the comparison of classifiers' construction over private inputs. In: International Conference on Trust, Security and Privacy in Computing and Communications. pp. 691–698 (2020)
25. Sweeney, L.: k-anonymity: A model for protecting privacy. International Journal of Uncertainty, Fuzziness and Knowledge-Based Systems 10(05), 557–570 (2002)
26. Wilcoxon, F.: Individual comparisons by ranking methods. Biometrics Bulletin 1, 60–83 (1945)
27. Yang, Q., Liu, Y., Chen, T., Tong, Y.: Federated machine learning: Concept and applications. ACM Trans. Intell. Syst. Technol. 10(2) (2019)
28. Ye, M., Wu, X., Hu, X., Hu, D.: Anonymizing classification data using rough set theory. Knowledge-Based Systems 43 (2013)

Appendix

Tables 5, 6, 7, and 8 report respectively the Holm scores of classifiers with respect to accuracy, precision, recall, and F1-score. The higher scores show better performance results for the associated classification algorithm and associated metric.

Anonymity	**Classifier**							
	DT	**NB**	**kNN**	**SVM**	**RF**	**LR**	**AB**	**BG**
Original	4.15	2.83	3.51	2.92	5.66	4.02	3.97	**5.80**
3-anonymity	4.06	3.33	3.79	3.97	4.34	**5.34**	4.47	3.56
2-diversity	3.10	4.70	4.79	4.15	3.38	**5.20**	4.11	3.42
0.2-closeness	3.61	3.33	3.79	4.93	3.38	**5.25**	4.93	3.65

Table 5: Classifier accuracy scores.

Anonymity	**Classifier**							
	DT	**NB**	**kNN**	**SVM**	**RF**	**LR**	**AB**	**BG**
Original	4.66	3.93	3.29	2.15	5.75	3.74	3.15	**6.21**
3-anonymity	3.83	5.11	2.88	3.65	**5.16**	4.11	3.65	4.47
2-diversity	2.92	**5.66**	4.43	3.61	4.02	4.75	4.02	3.47
0.2-closeness	3.79	**5.75**	3.97	3.01	3.01	5.39	3.93	4.02

Table 6: Classifier precision scores.

Anonymity	**Classifier**							
	DT	**NB**	**kNN**	**SVM**	**RF**	**LR**	**AB**	**BG**
Original	4.38	3.63	3.38	3.29	5.07	3.70	4.02	**5.39**
3-anonymity	4.52	3.33	3.79	3.97	4.34	**5.34**	4.47	3.56
2-diversity	3.10	4.70	4.79	4.15	3.38	**5.20**	4.11	3.42
0.2-closeness	3.61	3.33	3.79	4.93	3.38	**5.25**	4.93	3.65

Table 7: Classifier recall scores.

Anonymity	**Classifier**							
	DT	**NB**	**kNN**	**SVM**	**RF**	**LR**	**AB**	**BG**
Original	3.83	4.38	4.43	3.38	4.29	**5.11**	3.70	3.74
3-anonymity	4.02	3.88	3.97	3.56	4.70	**4.84**	4.02	3.88
2-diversity	2.83	4.47	4.83	3.65	4.11	**4.84**	4.20	3.93
0.2-closeness	4.02	4.47	4.56	3.51	3.83	**4.66**	4.52	3.29

Table 8: Classifier F1-score scores.

Figures 5 to 16 show the classifiers performance trained on anonymized Credit, Absent, and Optic datasets for different values of k, ℓ, and t.

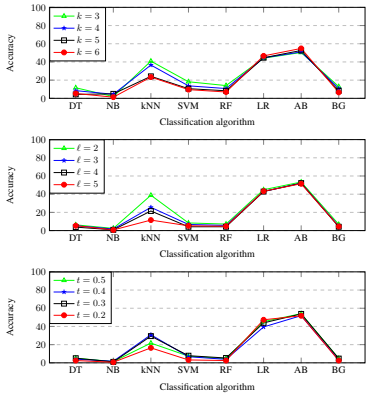


Fig. 5: Accuracy on Credit.

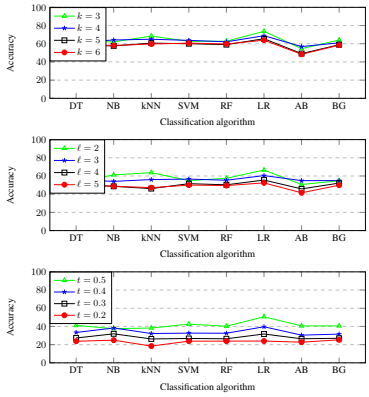


Fig. 6: Precision on Credit.

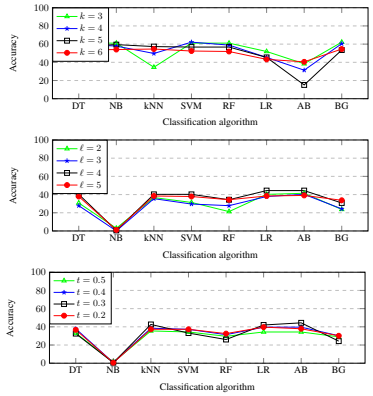


Fig. 7: Recall on Credit.

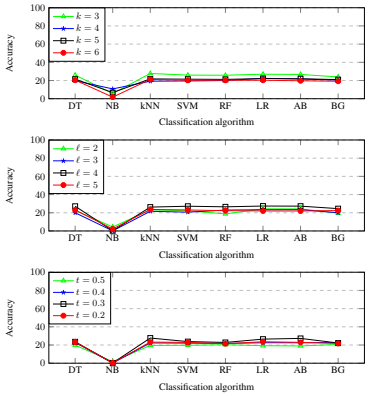


Fig. 8: F1-score on Credit.

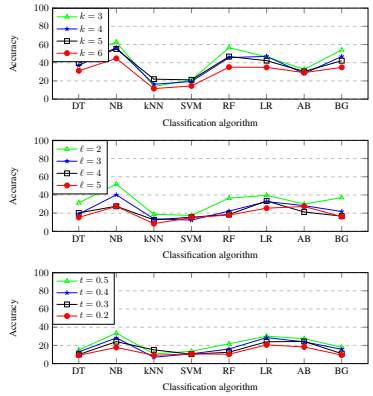


Fig. 9: Accuracy on Absent.

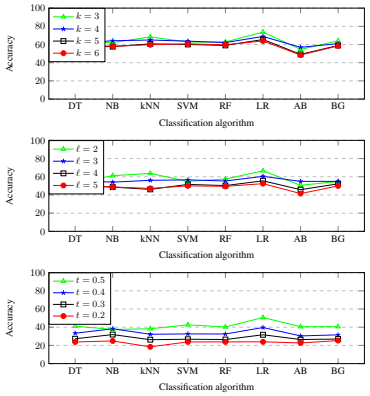


Fig. 10: Precision on Absent.

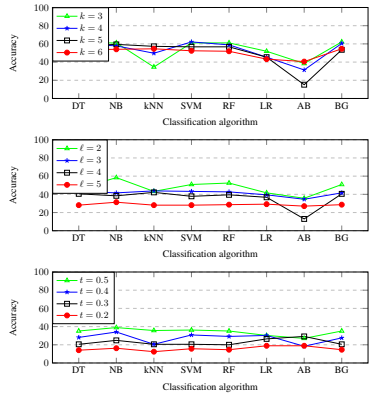


Fig. 11: Recall on Absent.

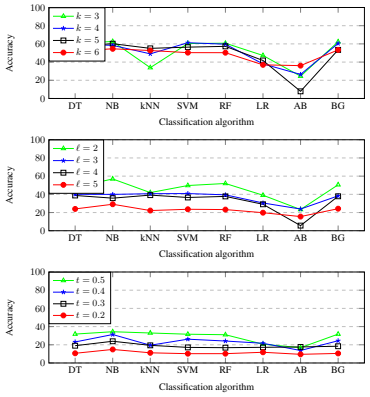


Fig. 12: F1-score on Absent.

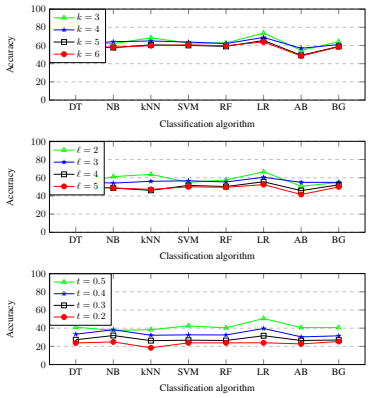


Fig. 13: Accuracy on Optic.

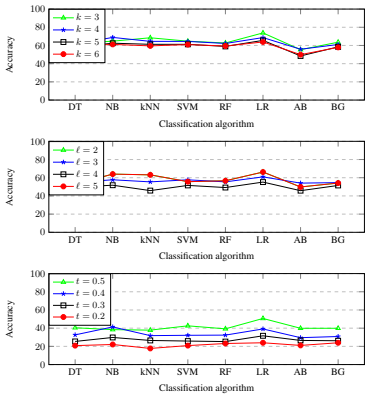


Fig. 14: Precision on Optic.

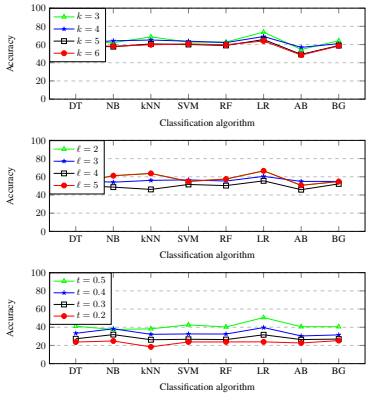


Fig. 15: Recall on Optic.

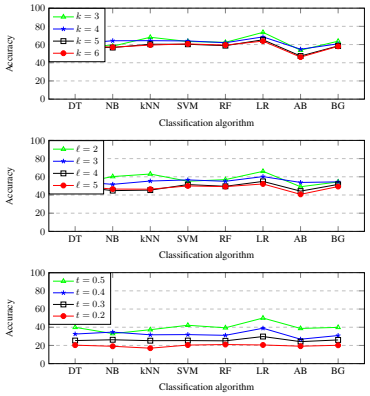


Fig. 16: F1-score on Optic.