DO NOT DISTURB ?
Classifier behavior on perturbed datasets

No Author Given

No Institute Given

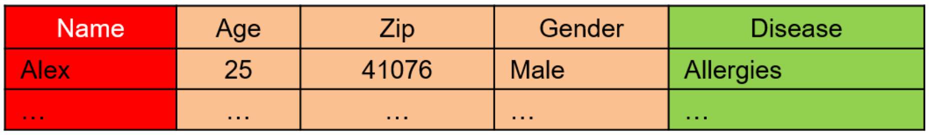
Abstract. Exponential trends in data generation are presenting todays organizations, economies and governments with challenges never encountered before, especially in the field of privacy and data security. One crucial trade-off regulators are facing regards the simultaneous need for publishing personal information for the sake of statistical analysis and Machine Learning in order to increase quality levels in areas like medical services, while at the same time protecting the identity of individuals. A key European measure will be the introduction of the General Data Protection Regulation (GDPR) in 2018, giving customers the 'right to be forgotten', i.e. having their data deleted on request. As this could lead to a competitive disadvantage for European companies, it is important to understand which effects deletion of significant data points has on the performance of ML techniques. In a previous paper we introduced a series of experiments applying different algorithms to a binary classification problem under anonymization as well as perturbation. In this paper we extend those experiments by multi-class classification and introduce outlier-removal as an additional scenario. While the results of our previous work were mostly in-line with our expectations, our current experiments revealed unexpected behavior over a range of different scenarios.

Keywords: Machine learning, knowledge bases, right to be forgotten, perturbation, k-anonymity, SaNGreeA, information loss, cost weighing vector, multi-class classification, outlier analysis, variance-sensitive analysis

1 Introduction and Motivation for Research

2 K-Anonymity and Information loss

While there several data-structures which can contain and convey personal information we might want to protect (free text, audio, images, graph structures etc.) we are focusing our work on tabular data, since most unstructured documents of sensitive nature today can be mapped to tabular data and since delicate information is most easily extracted from those. Figure 1 illustrates the original tabular concept of three different categories of data we will encounter in such tables:



Name	Age	Zip	Gender	Disease
Alex	25	41076	Male	Allergies
…	…	…	…	…

Fig. 1. The three types of data considered in (k-)anonymization

- **Identifiers** directly reveal the identity of a person without having further analysis of the data. Examples are first and last names, email address or social security number (SSN). As personal identifiers are hard to generalized (see Figure 2) in a meaningful way (truncating an email address to 'host' would not yield much usable information), those columns are usually removed. The figure displays this column in a red background color.
- **Sensitive data,** or 'payload', is crucial information for statisticians or researchers and can therefore not be erased or perturbed; such data usually remains untarnished within the released dataset. The table shows one column in green background color representing such data.
- **Quasi identifiers (QI's)**, colored in the table with an orange background, do not directly identify a person (age=35), but can be used in combination to restrict possibilities to such a degree that a specific identity follows logically. For instance, [6] mentioned that 87% of U.S. citizens in 2002 could be re-identified by just using the 3 attributes *zip code*, *gender* and *date of birth*. On the other hand, this information might hold significant information for the purpose of research (e.g. zip code could be of high value in a study on disease spread). Therefore we generalize this kind of information, which means to lower its level of granularity. As an example, one could generalize grades from A+ to B- into A's and B's and then further up to encompass 'all' (also denoted as '*'), as shown in Figure 2.

3 Related Work

A comparison of different Machine Learning algorithms on anonymized datasets was already conducted in 2014 [7] by applying 6 different algorithms on 3 datasets, with very diverse results per algorithm. The main weakness of this paper is its usage of extremely differently-sized datasets which does not easily allow comparison; moreover they only used one very low privacy setting of $k = 2$, preventing the authors from examining more interesting behavior as information content degrades further; this is a main point of our work.

The authors of [4] propose a scheme for controlling over-generalization of less identity-vulnerable QIs in diverse classes by determining the importance of QIs via Random Forest pre-computations as well as computing sensitive attribute diversity via the Simpson index [5]. Their resulting adaptive anonymization al-

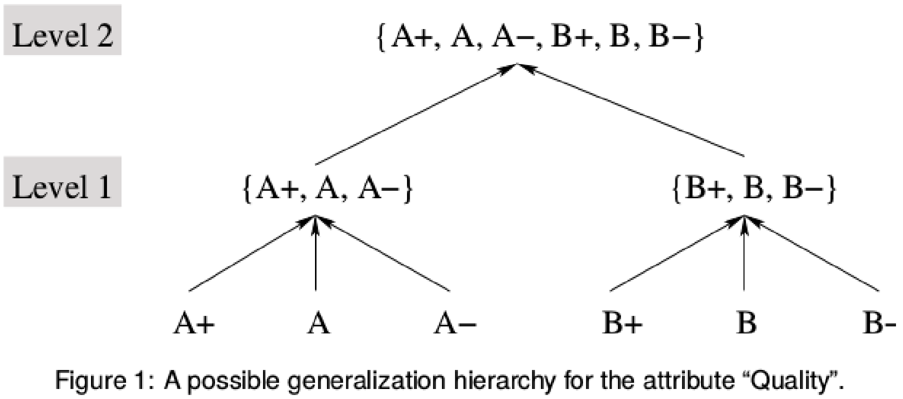
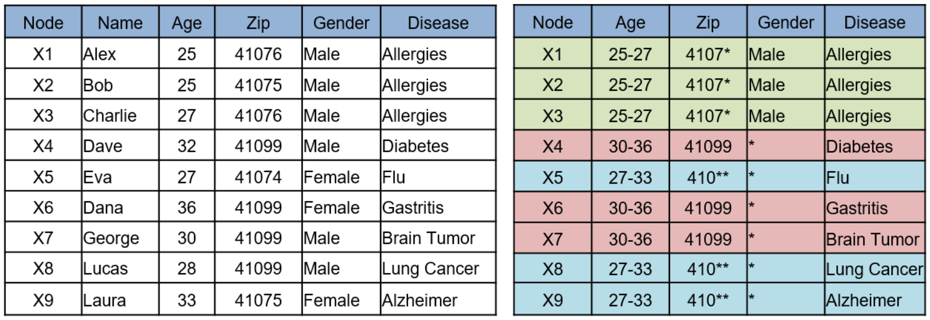

Level 2
{A+, A, A–, B+, B, B–}
Level 1
{A+, A, A–}
{B+, B, B–}
A+
A
A–
B+
B
B-
Figure 1: A possible generalization hierarchy for the attribute "Quality".

Fig. 2. Example of a typical generalization hierarchy taken from [1]



Node	Name	Age	Zip	Gender	Disease
X1	Alex	25	41076	Male	Allergies
X2	Bob	25	41075	Male	Allergies
X3	Charlie	27	41076	Male	Allergies
X4	Dave	32	41099	Male	Diabetes
X5	Eva	27	41074	Female	Flu
X6	Dana	36	41099	Female	Gastritis
X7	George	30	41099	Male	Brain Tumor
X8	Lucas	28	41099	Male	Lung Cancer
X9	Laura	33	41075	Female	Alzheimer

Node	Age	Zip	Gender	Disease
X1	25-27	4107*	Male	Allergies
X2	25-27	4107*	Male	Allergies
X3	25-27	4107*	Male	Allergies
X4	30-36	41099	*	Diabetes
X5	27-33	410**	*	Flu
X6	30-36	41099	*	Gastritis
X7	30-36	41099	*	Brain Tumor
X8	27-33	410**	*	Lung Cancer
X9	27-33	410**	*	Alzheimer

Fig. 3. Tabular anonymization: input table and anonymization result

gorithm was compared to Mondrian [2] as well as IACk [3] and shows improvements w.r.t information loss as well as coverage (the number of descendant leaf nodes of generalized values in the taxonomy). Accuracy measured on classification tree, random forest and SVM shows equal or better performance when applied to a dataset anonymized by their proposed solution; it is interesting to note that their performance on large factors of k not only remains stable, but in some cases increases with k, the same behavior we also observed in some of our experiments.

4 Experiments

The following sections will describe our series of experiments in detail, encompassing the dataset used, the algorithms chosen for classification as well as a description of the overall process employed to obtain our results.

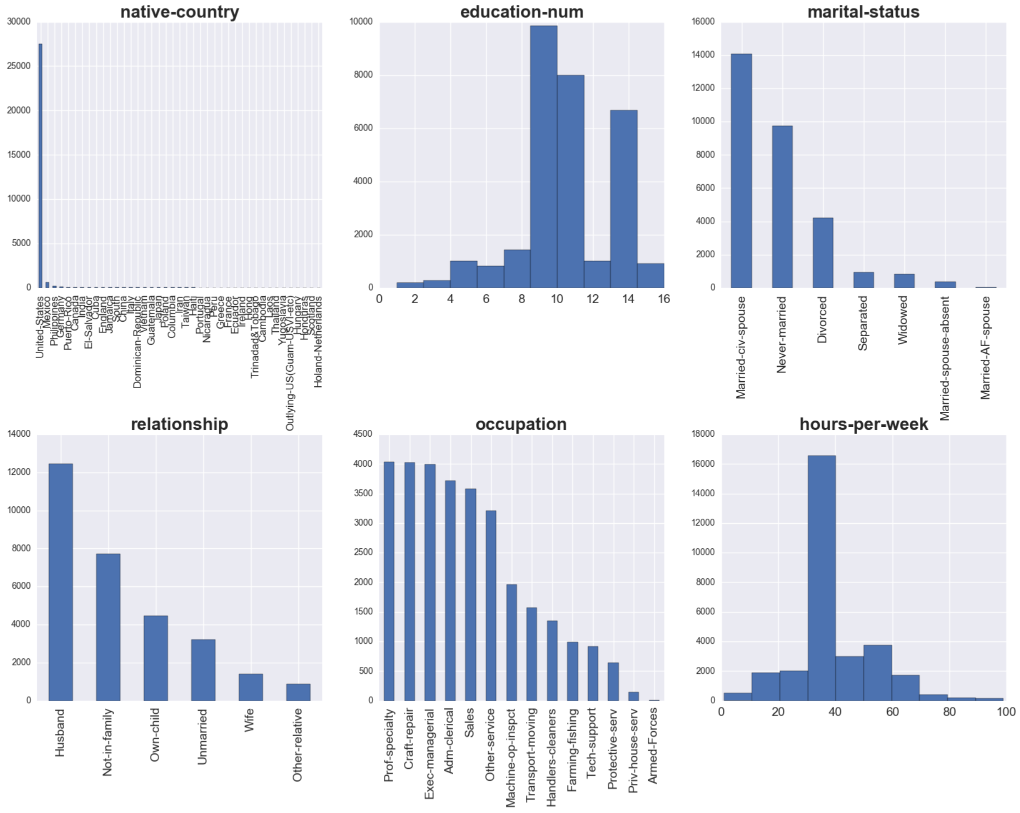


Fig. 4. Initial distribution of six selected data columns of the adult dataset.

4.1 Data

As input data we chose the training set of the adults dataset from the UCI Machine Learning repository which was generated from US census data and contains approximately 32,000 entries (30162 after deleting those with incomplete information). All but one columns were considered for experimentation, the remaining representing duplicate information (education => education_num). Figure 4 shows the attribute value distribution of 6 arbitrarily selected columns of the original dataset.

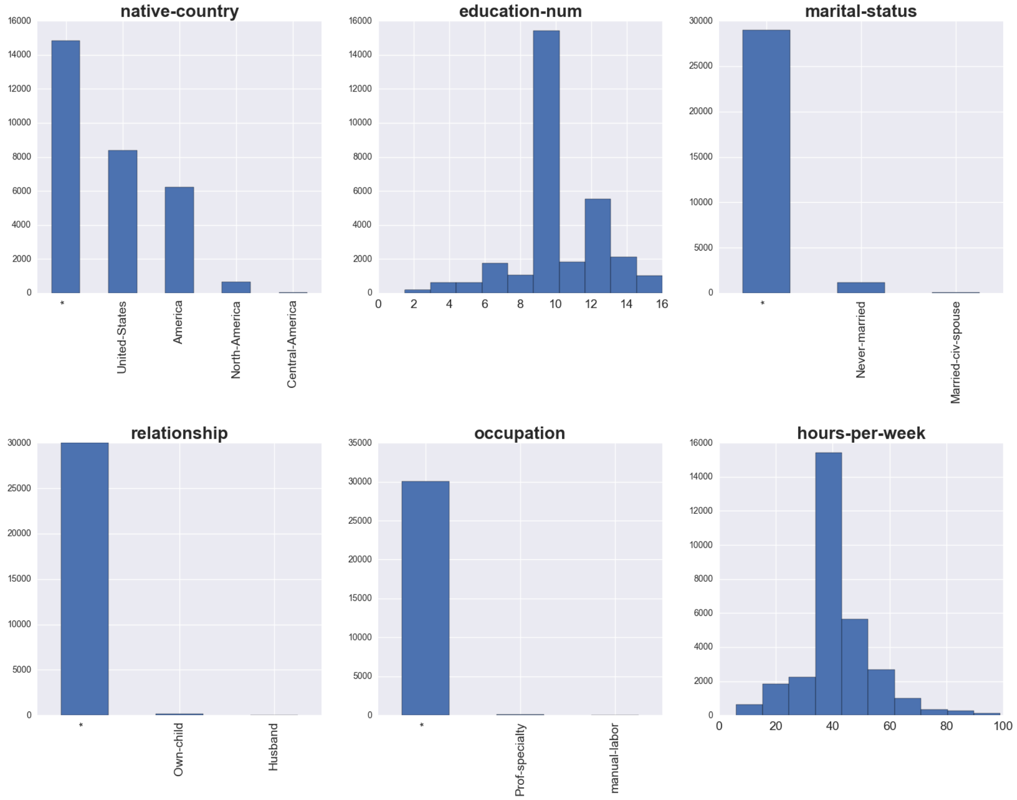


Fig. 5. Anonymized distribution of six selected data columns of the adult dataset, anonymization factor of k=19, equal weight for each attribute.

4.2 Anonymization Algorithm

$$\mathrm{GIL}(cl) = |cl| \cdot (\sum_{j=1}^{s} \frac{size(gen(cl)[N_j])}{size(min_{x\epsilon N}(X[N_j]), max_{x\epsilon N}(X[N_j]))} + \sum_{j=1}^{t} \frac{height(\Lambda(gen(cl)[C_j]))}{height(H_{C_j})})$$

where:
- $|cl|$ denotes the cluster cl's cardinality;
- $size([i1, i2])$ is the size of the interval $[i1, i2]$, i.e., $(i2 - i1)$;
- $\Lambda(w), w\epsilon H_{C_j}$ is the sub-hierarchy of H_{C_j} rooted in w;
- $height(H_{C_j})$ denotes the height of the tree hierarchy H_{C_j};

The total generalization information loss is then given by:

$$\mathrm{GIL}(G, S) = \sum_{j=1}^{v} \mathrm{GIL}(cl_j)$$

And the normalized generalization information loss by:

$$\mathrm{NGIL}(G, S) = \frac{\mathrm{GIL}(G, S)}{n \cdot (s + t)}$$

Distance between two nodes:

$$\mathrm{dist}(X^i, X^j) = \frac{|\{l|l = 1..n \wedge l \neq i, j; b_l^i \neq b_l^j|}{n - 2}$$

Distance between a node and a cluster:

$$\mathrm{dist}(X, cl) = \frac{\sum_{X^j \epsilon cl} \mathrm{dist}(X, X^j)}{|cl|}$$

4.3 Process

To examine the effect of perturbation and anonymization on classification performance, we designed the following processing pipeline:

5 Results & Discussion

5.1 Perturbed Datasets - Selective Deletion

In order to be able to compare the impact of selectively deleting the most / least important attribute values (in fact, the whole data points containing those values) on different classifiers, we chose to select these values via examining the logit coefficients from logistic regression. Although this possibly entails non-erasure of the values specifically significant for each classifier, we chose algorithmic comparison as the more insightful criterion; the implicit assumption that the same attribute values would influence all classifiers approximately equally was largely confirmed by our results.

In contrast to binary classification, determining the 'right' values to delete for a multi-class problem is not always possible: Values contributing highly to the decision boundary for one class might be less significant in the case of another - accordingly one would expect inconclusive behavior in the case of a target for which the highest / lowest log coefficients do not line up over class boundaries.

For each of the targets 'marital-status' and 'education-num' we measured those interesting coefficients in the hope of improving / degrading algorithmic performance; that means deletion of highest logit's is supposed to remove certainty from an algorithm and decreasing performance, while deletion of lowest

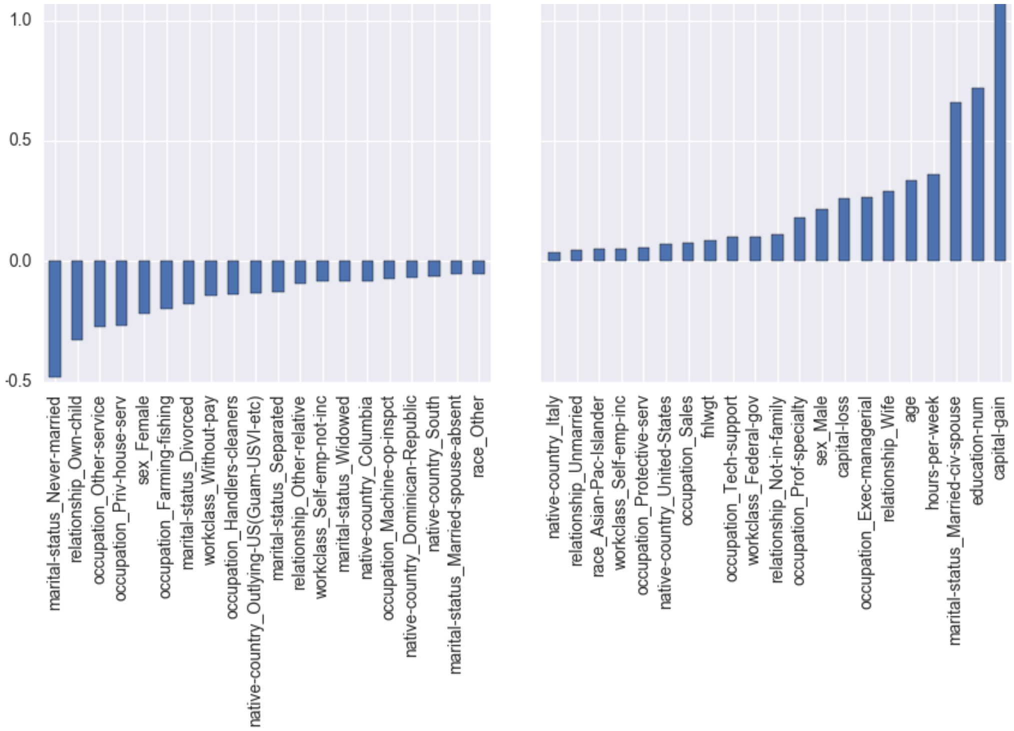


Fig. 6. Attribute values within the adult dataset which contribute highest / lowest certainty to the classification of income (truncated at 1.0). The rightmost columns represent information which enable a classifier to discern most clearly between classes, while the leftmost columns (depending on their actual score) could even confuse the algorithm. We chose this example because income is a binary decision, so the values don't change per category to predict.

logit's is supposed to remove uncertainty, thus improving performance. Our analysis showed that while 'marital-status' had mainly the same most / least significant logit's across all classes, the attribute values for 'education-num' were rather diverse in this area.

In the latter case this lead to erratic behavior of the resulting performance curves, as can be seen in (Figure 7). It is interesting to note that 'income_ >50k' obviously held much larger significance for Logistic Regression than for the other classifiers, as their results showed f1 score improvement with this particular value eviscerating.

In the case of 'marital-status' almost the same attribute values were rated as most / least significant across all classes - this results in very clear outputs with the erasure of highly important values decreasing performance drastically while deletion of confusing values leading to a significant increase in classifier

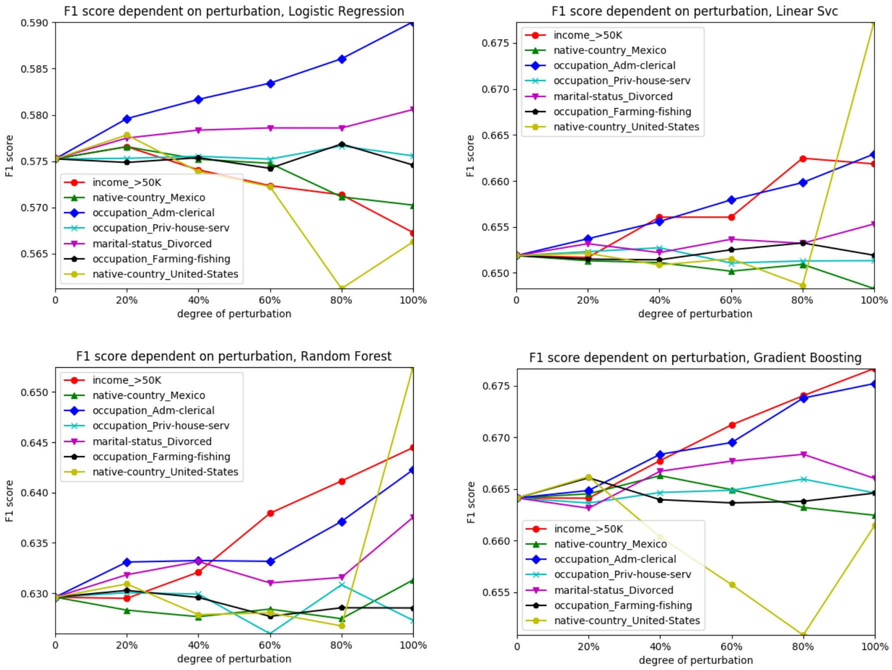


Fig. 7. Multi-class classification of 'education-num' under perturbation by selective deletion of the most / least contributing attribute values. Since different values are significant for deciding on different classes of education level, progressive deletion of this data results in indeterminate behavior.

performance (Figure 8). While it is not surprising that relationship information shows high correlation with marital status, the opposite effects of *sex_Female* and *sex_Male* stand out as a slight curiosity - being a woman in this dataset seems to point less distinctly to a specific marital status than being a man.

5.2 Anonymized Datasets

5.3 Outliers removed

5.4 Anonymization on Outliers removed

6 Open problems Future challenges

7 Conclusion

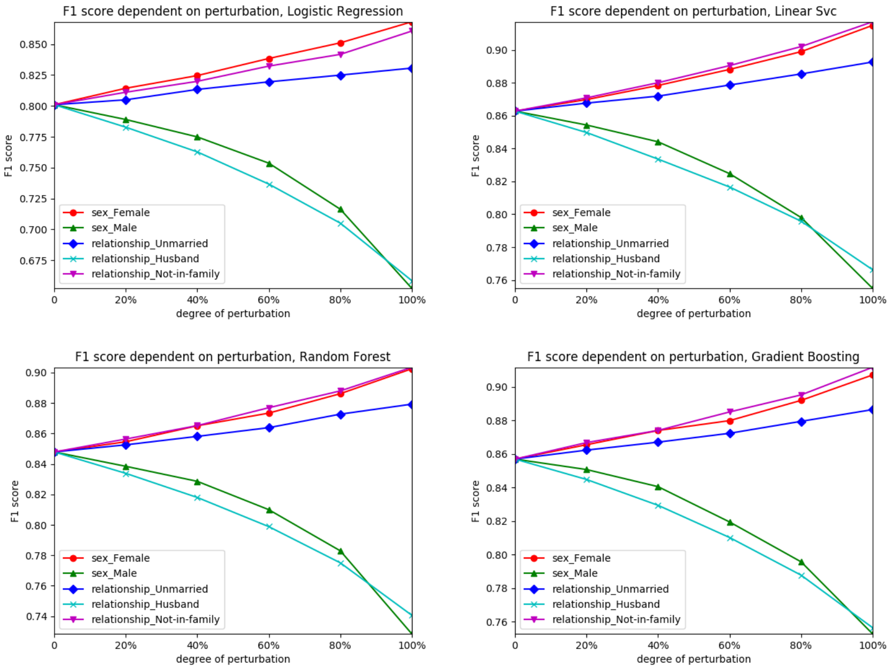


Fig. 8. Multi-class classification of 'marital-status' under perturbation by selective deletion of the most / least contributing attribute values. Since the same values are significant for deciding different classes of marital status, progressive deletion leads to orderly increase / decrease of ML performance.

References

1. Gagan Aggarwal, Tomas Feder, Krishnaram Kenthapadi, Rajeev Motwani, Rina Panigrahy, Dilys Thomas, and An Zhu. Approximation algorithms for k-anonymity. *Journal of Privacy Technology (JOPT)*, 2005.
2. Kristen LeFevre, David J DeWitt, and Raghu Ramakrishnan. Mondrian multidimensional k-anonymity. In *Data Engineering, 2006. ICDE'06. Proceedings of the 22nd International Conference on*, pages 25–25. IEEE, 2006.
3. Jiuyong Li, Jixue Liu, Muzammil Baig, and Raymond Chi-Wing Wong. Information based data anonymization for classification utility. *Data & Knowledge Engineering*, 70(12):1030–1045, 2011.
4. A Majeed, F Ullah, and S Lee. Vulnerability-and Diversity-Aware Anonymization of Personally Identifiable Information for Improving User Privacy and Utility of Publishing Data. *Sensors*, pages 1–23, 2017.
5. Edward H Simpson. Measurement of diversity. *Nature*, 1949.
6. Latanya Sweeney. k-anonymity: A model for protecting privacy. *International Journal of Uncertainty, Fuzziness and Knowledge-Based Systems*, 10(05):557–570, 2002.
7. Hayden Wimmer and Loreen Powell. A Comparison of the Effects of K-Anonymity on Machine Learning Algorithms. pages 1–9, 2014.

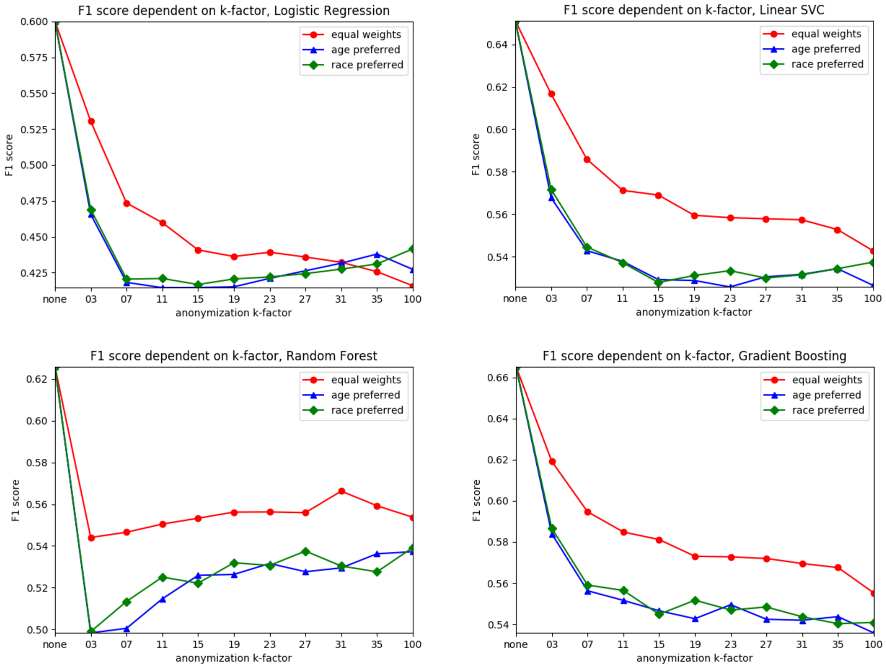


Fig. 9. Multi-class classification of education num on the adult dataset under several degrees of k-anonymization.

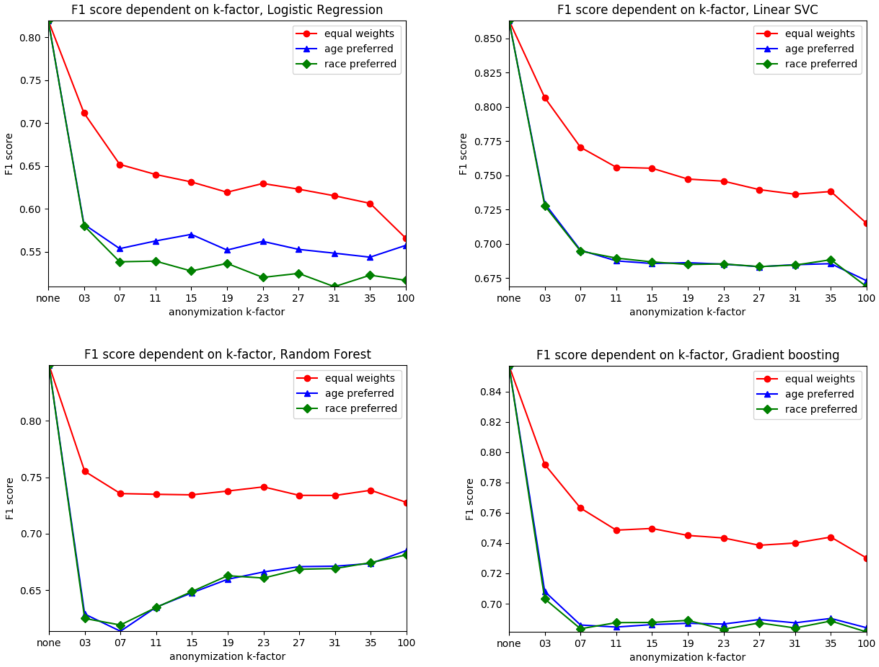


Fig. 10. Multi-class classification of marital status on the adult dataset under several degrees of k-anonymization.

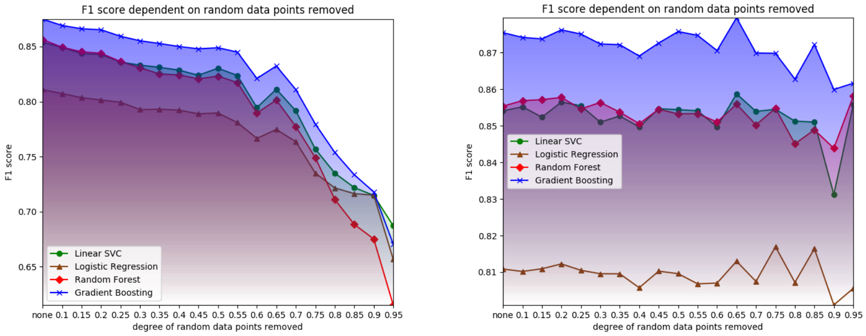
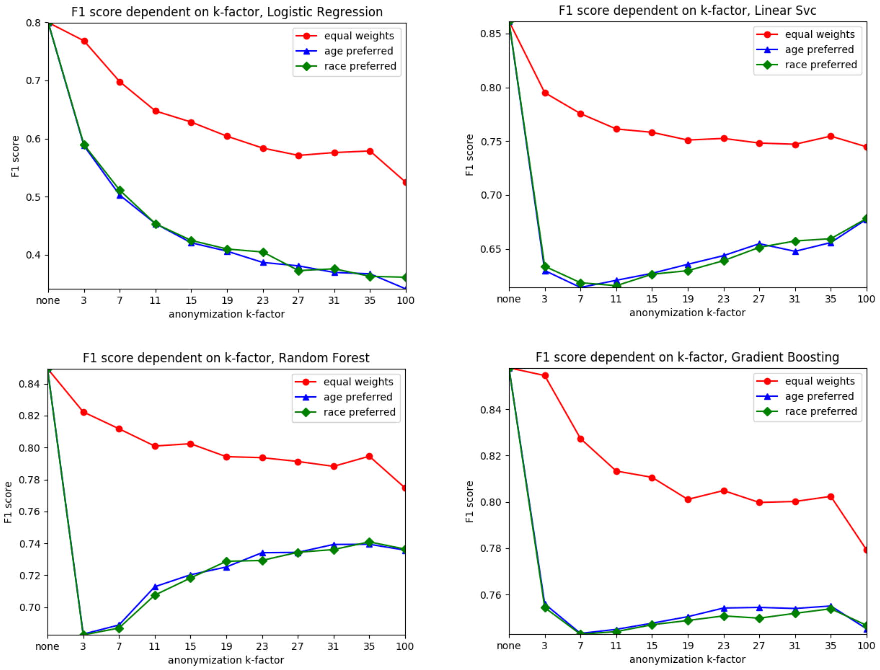


Fig. 11. Binary classification on target income based on a dataset with different degrees of outliers removed (= variance loss) vs. the same degree of data randomly deleted.


F1 score dependent on k-factor, Logistic Regression

F1 score dependent on k-factor, Linear Svc

F1 score dependent on k-factor, Random Forest

F1 score dependent on k-factor, Gradient Boosting

Fig. 12. Multi-class classification on target marital status based on a dataset with 30% outliers removed AND different degrees of k-anonymization.