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 Regulations (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

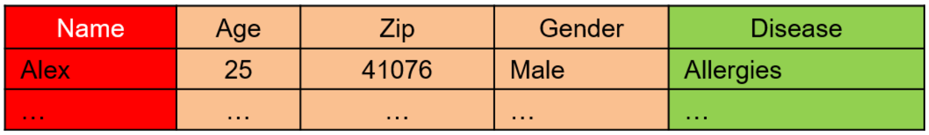
1 Introduction and Motivation for Research

2 Scenarios of incurring information loss in datasets

2.1 Tabular anonymization

3 Experiments

The following sections will describe our series of experiments in detail, encompassing the data source selected, the algorithm used as well as a description of the overall process employed to obtain our results.



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

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

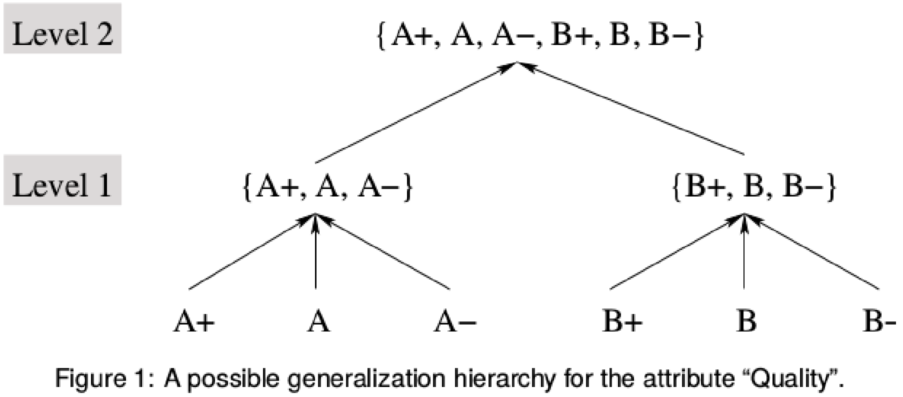
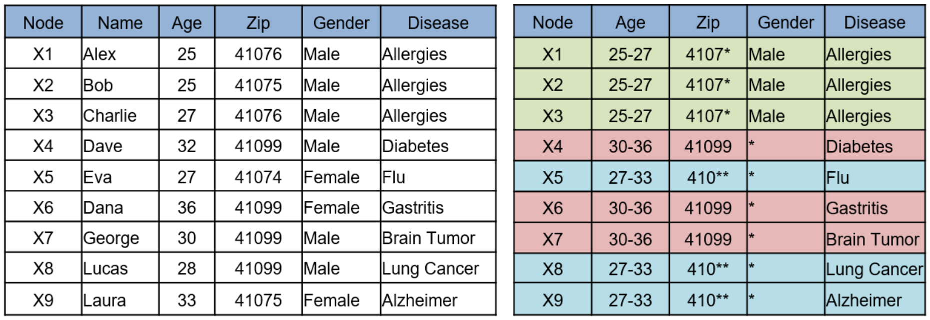


Figure 1: A possible generalization hierarchy for the attribute “Quality”.

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

3.1 Data

As input data we chose the adults dataset from the UCI Machine Learning repository which was generated from US census data of 1994 and contains approximately 32,000 entries; from those 30,162 were selected after preprocessing. Of the attributes (data columns) provided only one was deleted because it was also represented by a column containing its numerical mapping (education =>

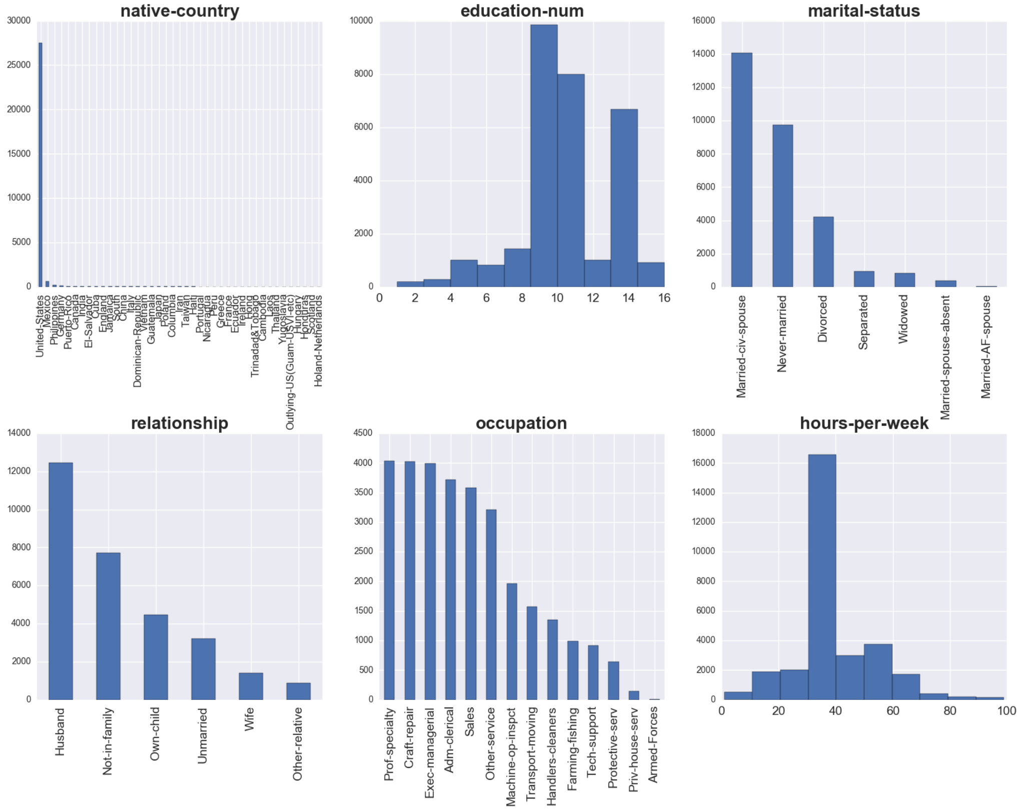


Fig. 4. Initial distribution of six selected data columns of the adult dataset.

education_num). Figure 4 shows the attribute value distribution of the original input dataset with the exception of the sample weights.

3.2 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;

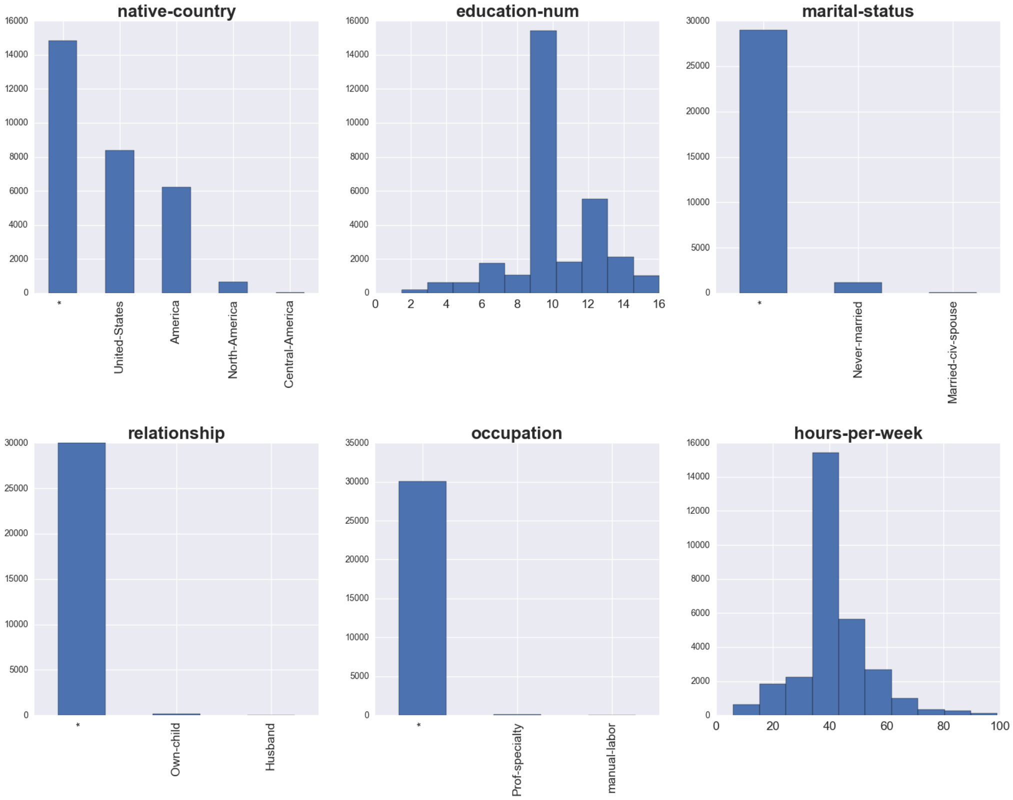


Fig. 5. Anonymized distribution of six selected data columns of the adult dataset, anonymization factor of k=19, equal weight for each attribute.

- $height(H_{C_j})$ 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:

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

3.3 Process

To examine the impact of perturbation and anonymization of datasets on the quality of a classification result, we designed the following processing pipeline:

4 Results & Discussion

5 Open problems Future challenges

6 Conclusion

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. Alina Campan and Traian Marius Truta. Data and structural k-anonymity in social networks. In *Privacy, Security, and Trust in KDD*, pages 33–54. Springer, 2009.
3. Sean Chester, Bruce Kapron, Ganesh Ramesh, Gautam Srivastava, Alex Thomo, and S Venkatesh. k-anonymization of social networks by vertex addition. *ADBIS (2)*, 789:107–116, 2011.
4. Valentina Ciriani, S De Capitani di Vimercati, Sara Foresti, and Pierangela Samarati. κ-anonymity. In *Secure data management in decentralized systems*, pages 323–353. Springer, 2007.
5. A Holzinger, M Plass, K Holzinger, GC Crisan, CM Pintea, and V Palade. Towards interactive machine learning (iml): Applying ant colony algorithms to solve the traveling salesman problem with the human-in-the-loop approach. In *IFIP International Cross Domain Conference and Workshop (CD-ARES)*, page in print. Springer, Heidelberg, Berlin, New York, 2016.
6. Andreas Holzinger. Interactive machine learning for health informatics: When do we need the human-in-the-loop? *Springer Brain Informatics (BRIN)*, 3(2):119–131, 2016.
7. Bruce Kapron, Gautam Srivastava, and S Venkatesh. Social network anonymization via edge addition. In *Advances in Social Networks Analysis and Mining (ASONAM), 2011 International Conference on*, pages 155–162. IEEE, 2011.
8. Peter Kieseberg, Bernd Malle, Peter Frhwirt, Edgar Weippl, and Andreas Holzinger. A tamper-proof audit and control system for the doctor in the loop. *Brain Informatics*, pages 1–11, 2016.
9. Latanya Sweeney. k-anonymity: A model for protecting privacy. *International Journal of Uncertainty, Fuzziness and Knowledge-Based Systems*, 10(05):557–570, 2002.

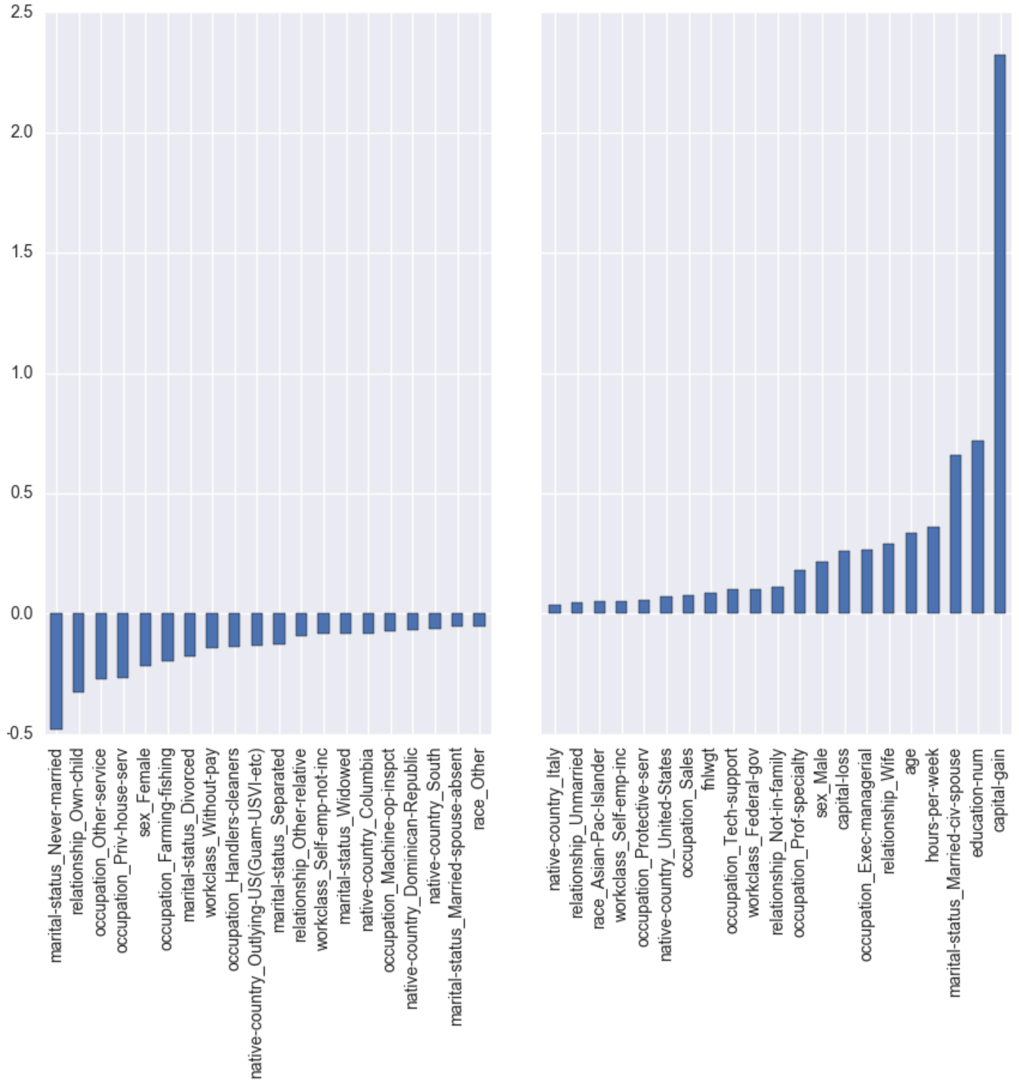


Fig. 6. The attribute values of the adult dataset which contribute most positively / negatively to the classification result. The columns to the right strongly indicate a yearly income of above 50k, whereas the columns to the outer left indicate a yearly income of below 50k. The least significant columns in the middle part were cut out.

10. Wei-Shi Zheng, Shaogang Gong, and Tao Xiang. Reidentification by relative distance comparison. *IEEE transactions on pattern analysis and machine intelligence*, 35(3):653–668, 2013.
11. Bin Zhou, Jian Pei, and WoShun Luk. A brief survey on anonymization techniques for privacy preserving publishing of social network data. *ACM Sigkdd Explorations Newsletter*, 10(2):12–22, 2008.

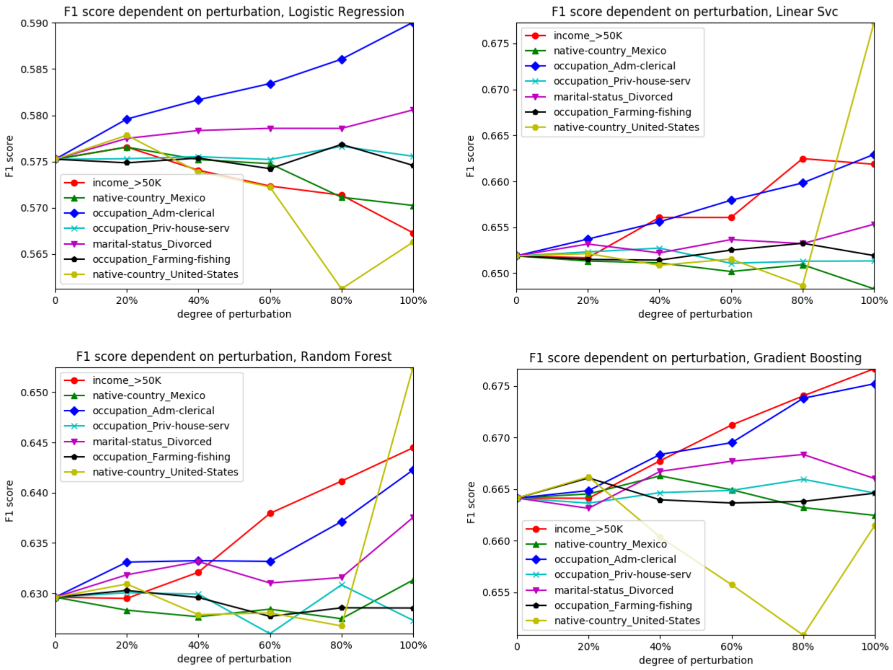


Fig. 7. Multi-class classification of education level under perturbation by selective deletion of important data attributes.

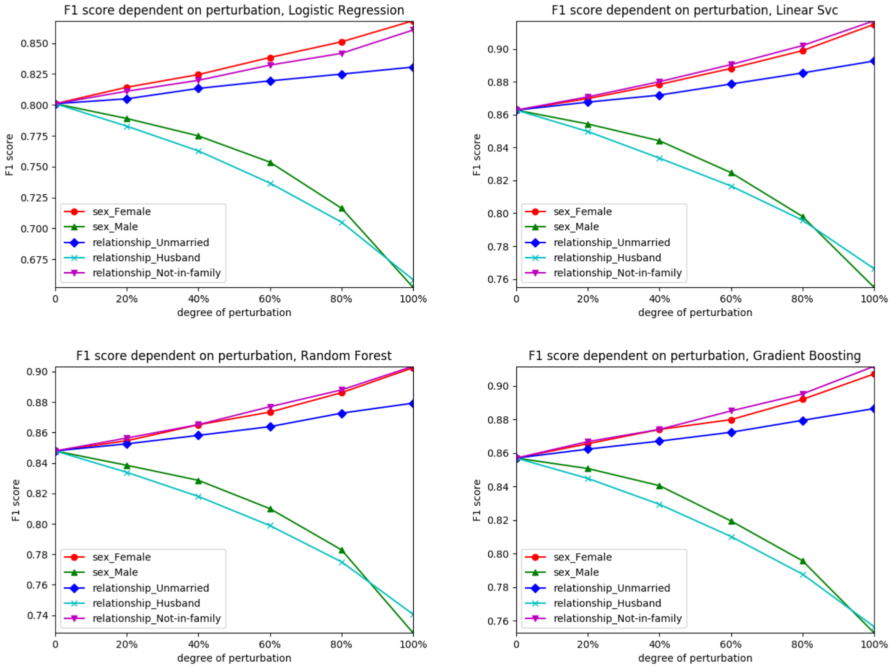


Fig. 8. Multi-class classification of marital status under perturbation by selective deletion of important data attributes.

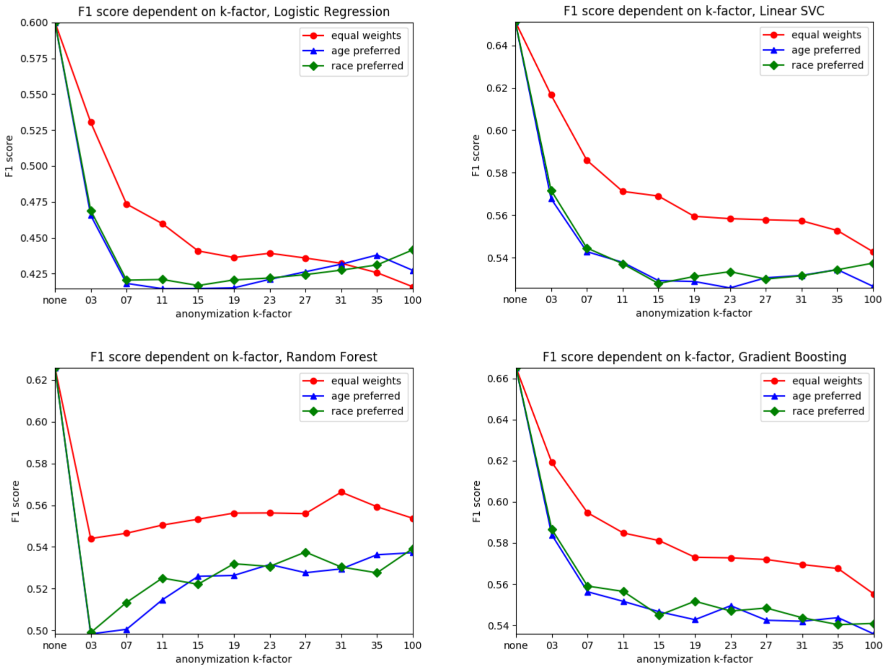


Fig. 9. Multi-class classification of education num on the adult dataset under several degrees of k-anonymization.

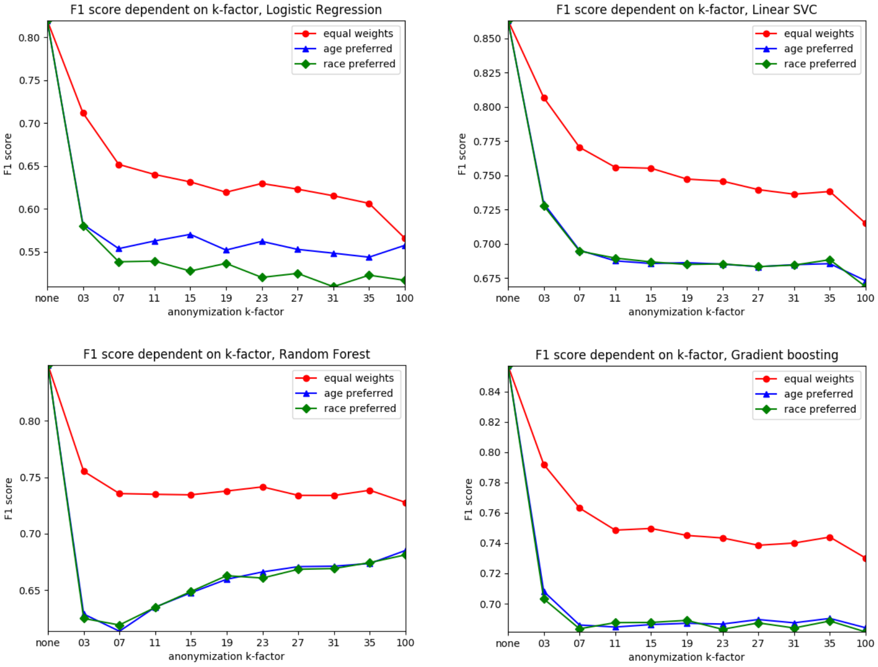


Fig. 10. Multi-class classification of marital status on the adult dataset under several degrees of k-anonymization.