Explainability of the Effects of Non-perturbative Data Protection in Supervised Classification

Stefano Locci, Luigi Di Caro
Department of Computer Science
University of Turin
Turin, Italy
0009-0006-9725-2045,
0000-0002-7570-637X

Giovanni Livraga
Department of Computer Science
University of Milan
Milan, Italy
0000-0003-2661-8573

Marco Viviani
Department of Informatics, Systems, and Communication
University of Milano-Bicocca
Milan, Italy
0000-0002-2274-9050

***Abstract*—The increasing availability of online data has meant that data-driven models have been applied to more and more tasks in recent years. In some domains and/or applications, such data must be protected before they are used. Hence, one of the problems only partially addressed in the literature is to determine how the performance of Machine Learning models is affected by data protection. More important, the explainability of the results of such models as a consequence of data protection has been even less investigated to date. In this paper, we refer to this very problem by considering non-perturbative data protection, and by studying the explainability of supervised models applied to the data classification task.**

***Index Terms*—Explainability, Data Protection, Machine Learning, Privacy, Classification**

I. Introduction

In recent years, the availability of both structured and semi-structured/unstructured online data (e.g., open and linked data, Web pages, User-Generated Content) has increased significantly, together with the use of *Machine Learning* (ML) models to perform distinct tasks on such data. In this scenario, the need for more transparent and explainable algorithms has increased, particularly when dealing with personal data, confidential data, or, more in general, data that are not intended for public and indiscriminate disclosure.

In the literature, to the aim of *protecting data*, a possible approach consists of generating a *sanitized* version of a dataset, to be released/shared/processed instead of the original one, with the guarantee that a given *privacy requirement* is satisfied. Such privacy requirement demands that data that should remain private be not disclosed. To this end, data can be sanitized in different ways, and two main families of *data protection techniques* have been proposed. *Perturbative techniques* – as the name says – perturb the original data, for example with the addition of random noise. *Differential privacy* [7] is a well-known privacy model, based on data perturbation, which enforces a privacy requirement limiting the impact of the data of a single individual on the results of a given computation. Regardless of the chosen privacy requirement, the adoption of perturbative data protection techniques inevitably causes a permanent loss of *data truthfulness* (i.e., the truthfulness of the information of each data item) in the sanitized dataset. *Non-perturbative* data protection techniques, instead, protect data while maintaining their truthfulness, for example by *generalizing* them; in this case, data protection is provided by the fact that the sanitized dataset is less detailed/complete. k-*anonymity* [26] and its extensions (e.g., ℓ-*diversity* [17] and t-closeness [13]) are examples of well-known privacy models based on generalization.

Apart from data privacy concerns, there is a growing need for the outcomes of ML models to be *interpretable*, due to several reasons. First, as ML models are being employed in critical domains such as healthcare, finance, and autonomous systems, the ability to explain how a model brings out its outputs becomes increasingly relevant. Second, regulations and legal frameworks, such as the *General Data Protection Regulation* (GDPR) [31], are stressing the right of individuals to understand and challenge decisions made by automated systems. Furthermore, the advent of advanced techniques, such as *Deep Learning* (DL), *Transformer-based architectures*, and *Large Language Models* (LLMs), has introduced a level of complexity that poses challenges to interpretability. These cutting-edge technologies act as "black boxes", i.e., the internal working and decision-making processes are not easily comprehensible or explainable.

For the above-mentioned reasons, the aim of this work is to analyze the impact of data sanitization based on non-perturbative data protection on the performance and, more importantly, the explainability, of commonly used ML models for the task of data classification.

II. Background and Related Work

This section presents basic concepts and state-of-the-art work that independently addresses both data protection and the explainability of ML models. Some works that, like ours, attempt to relate both aspects are also illustrated.

A. Data Protection

Among the various data protection techniques available and briefly illustrated in the Introduction, in the preliminary study presented in this article, we focus just on k-anonymity [26]. This non-perturbative model enforces a privacy requirement demanding that no released data item be related to less than a number k of respondents. It has been designed to operate on

Age	Sex	Disease
70	M	Stroke
50	M	Broken leg
70	F	Stroke
50	F	Asthma

(a)

Age	Sex	Disease
[50-70]	M	Stroke
[50-70]	M	Broken leg
[50-70]	F	Stroke
[50-70]	F	Asthma

(b)

Fig. 1: A microdata table (a) and an example of a 2-anonymous version (b) of it.

microdata tables (i.e., tabular data with one record for each respondent and one column for each attribute of interest). The first step for sanitizing a data collection requires to *remove* (i.e., suppress) all *identifiers*, i.e., attributes that uniquely identify respondents, such as `SSN` or `e-mail address`. This *de-identification* process, however, does not provide any anonymity guarantee, as the table might include other attributes (named *quasi-identifiers*) such as `sex`, `DoB`, and `address` that, in combination, can be linked to external sources of information (e.g., voter lists) to reduce the uncertainty about the identities of the de-identified respondents.[1] k-anonymity then operates by *generalizing* such quasi-identifiers, up to a point where each combination of them appears with at least k occurrences. This means that any linking attack operating on the quasi-identifiers will always find at least k different individuals (i.e., an *equivalence class*) to which each anonymized tuple can correspond, and vice versa.

Figure 1(b) illustrates a 2-anonymous version of the (de-identified) microdata table of Figure 1(a), reporting medical information for 4 respondents, and considering `Age` and `Sex` as quasi-identifiers. 2-anonymity is enforced by generalizing the original `Age` values to intervals. If a recipient knows that a target female respondent is aged 50, they can easily associate her with the last record in Figure 1(a), while in the 2-anonymous version of Figure 1(b) they cannot pinpoint which one among the last two records pertains to her. This applies to any combination of quasi-identifying values.

B. Explainability in Machine Learning

ML is applied today in an increasingly pervasive manner and with remarkable performance in numerous domains, even critical ones, such as healthcare, transportation, finance, and many more [29]. However, the *black-box* nature of these models can make it difficult or even impossible for people to understand how the system arrived at its decisions, which can be a problem in those contexts where *accountability* and *trust* are becoming more and more important [29]. To date, solutions for explainability mainly follow two paradigms [6]:

- *Transparency by Design*. It entails constructing ML models that are inherently interpretable by their architecture and functionality. *Decision trees*, for example, use a series of decision nodes to create a hierarchical structure of decisions. *Rule-based models* are another example of ML models that achieve explainability by design;
- *Post-hoc Interpretability.* It concerns the extraction of information, a posteriori, from pre-trained ML models. This allows us to interpret such models without compromising their performance. Their internal mechanisms remain not directly observable, because the focus is on understanding the output and the relationship between the input and the output.

In this article, we focus on post-hoc techniques, which in turn can follow two explainability paradigms:

- *Local Interpretability*. It aims at understanding the predictions of specific instances of a model. A widely-used solution is LIME (*Local Interpretable Model-Agnostic Explanations*) [21], which generates explanations for each prediction by locally approximating the model with a simpler, interpretable model around the prediction. Another well-known method employs *Counterfactual Explanations* [28], which proposes a modified instance of the model that would have resulted in a different outcome, providing insights into decision-making by observing the resulting changes;
- *Global Interpretability*. It aims at understanding the overall behavior of a model across its entire input space. This can be achieved with either *feature importance analysis* methods, or explainability frameworks such as SHAP (*SHapley Additive exPlanations*) [16]. While SHAP is fundamentally designed for local interpretation, assigning a Shapley value to each feature to quantify its contribution to the model's prediction for a specific instance, it is also possible to use it for global interpretation by averaging all instances in the data [25]. Finally, feature importance can also be calculated in an ML model based on the decrease in the model's accuracy when a feature is removed.

C. Assessing the Effects of Data Sanitization

When sanitizing a dataset, a certain amount of information loss due to generalization is inevitable [5] and several metrics have been proposed for quantifying it [8], for example assigning penalties based on the size of the equivalence classes (e.g., [1], [12]) or on "how much" generalization is applied (e.g., [30]). Most metrics are general-purpose and do not explicitly consider the impact of data protection on the specific application(s) that will use the sanitized data. Some works have considered the impact of sanitization on data mining tasks and ML models: for example, in [27] the authors evaluate how different k-anonymization approaches affect ML classifiers. We nicely complement this line of work with a specific focus on explainability. The work discussed in [23] is similar to ours in studying the effects of k-anonymity on ML macrotrends, but differs from ours as it focuses on microaggregation-based k-anonymity (a form of sanitization that enforces the k-anonymity requirement through data perturbation). The work in [9] is orthogonal to ours, and proposes a k-anonymiztion approach where generalization is administered so as to limit its impact on the quality of subsequent classification tasks.

[1] A study on the 2000 US Census data showed that a combination of `sex`, `DoB` and `ZIP code` permits to *uniquely* identify 63% of the US population [10].

Another related line of work concerns the impact of data protection on *Explainable Artificial Intelligence* (XAI). In [24], the authors analyze the impact of private learning techniques in combination with *Federated Learning* [11] on generated explanations for Deep Learning models. Other recent works are more domain-specific: for example, the work in [19] focuses on the industrial domain, presenting various approaches for performing perturbative privacy-preserving AI, including *homomorphic encryption* applications on Deep Neural Networks [20], as well as model-agnostic and model-specific methods explainability approaches, and discuss their limitations. Close to our work is [2], which aims to evaluate the effects of data protection on Shapley values for explainability. However, our work differs from these latter in the consideration of non-perturbative (rather than perturbative) sanitization. Other related, but different works, have explored issues of explainability in cybersecurity (e.g., [22]).

III. An Explainability Study Applied to the Task of Classifying Protected Data

The methodological framework employed to carry out the proposed preliminary study follows the *pipeline* illustrated in Figure 2.[2] It consists of four main steps: (i) *data anonymization* using non-perturbative techniques; (ii) *classification* on both original and anonymized data using four distinct ML models; (iii) *evaluation* of the classification performances; and (iv) *explainability analysis*.

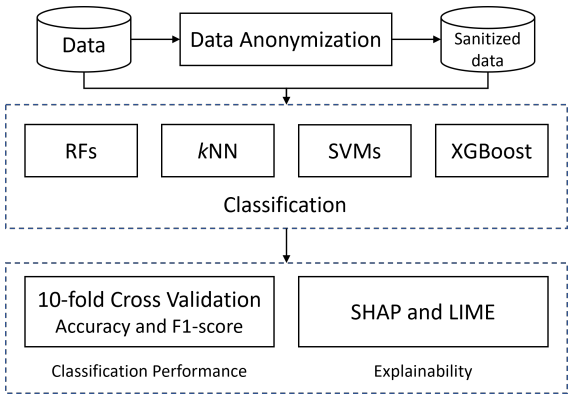

Fig. 2: The experimental pipeline.

Concerning step (i), the selection of two diverse datasets, namely ADULT and *Diabetes* (illustrated in detail in the following), aimed to explore different domains and facilitate a comprehensive exploration of the research objectives. To sanitize these datasets, we employed k-anonymity, for distinct values of k. Subsequently, in step (ii), we proceeded with the training of four distinct ML classifiers, namely *Random Forests*, *k-Nearest Neighbors* (kNNs), *Support Vector Machines* (SVMs), and *XGBoost*, with respect to both original data and each k-anonymized dataset. This allowed us to investigate and compare the classification performance of each ML model, obtained in step (iii), with explainability aspects of the trained models on the original and anonymized data. In particular, to provide the interpretation of each model's decision-making processes, we employed, in step (iv), the SHAP framework as illustrated in Section II-B for global interpretability. Furthermore, for a more localized focus on individual predictions, we utilized the LIME framework.

A. The Data

ADULT[3] This dataset comprises a collection of socio-economic attributes of individuals from the *United States Census Bureau* database.[4] It is commonly used for ML and data mining tasks to predict whether an individual's annual income exceeds a specific threshold (50K US Dollars). It counts around 45K records, each consisting of 14 attributes with associated both categorical and continuous values, i.e., *age*, *workclass*, *education*, *education-num*, *marital-status*, *occupation*, *relationship*, *race*, *sex*, *capital-gain*, *capital-loss*, *hours-per-week*, *native-country*, and the target attribute *income*.

Diabetes[5] This dataset includes around 100K records describing medical and demographic information of patients, with their diabetes (positive/negative) status. This dataset can be utilized to predict whether a patient has diabetes based on the following attributes: *age*, *gender*, *body-mass-index*-BMI, *hypertension*, *heart-disease*, *smoking-history*, *HbA1c-level*, *blood-glucose-level*, and the target attribute *diabetes*.

B. Data Sanitization

The ADULT dataset has been sanitized employing the Python library `Mondrian`,[6] as it supports the generalization of *categorical* attributes (other tools and libraries only support numerical attributes), leveraging a generalization hierarchy. As quasi-identifiers for performing k-anonymity, we chose *age, education, marital-status, occupation, race*.

For the *Diabetes* dataset, anonymization was carried out using the Python library `anonypy`[7] which also employ the Mondrian algorithm [12], and offers enhanced practicality for managing *numerical* attributes. We designated *blood-glucose-level-BMI*, *age*, *HbA1c* as quasi-identifiers.

C. Classification Models

Random Forests (RFs). It is an ensemble learning method that combines multiple decision trees to form a robust predictive model [3]. Each decision tree is trained on a random subset of the data and features, and the final prediction is composed by the aggregation of the predictions of each tree. It can be used for classification and regression tasks. In our study, we employed the `RandomForestClassifier` class

[2]Note that, since we investigate data protection leveraging non-perturbative sanitization within models that guarantee some form of anonymity, in the remainder of this paper we use the terms *anonymization* and *sanitization* interchangeably.

[3]http://archive.ics.uci.edu/dataset/2/adult
[4]https://www.census.gov/data.html
[5]https://archive.ics.uci.edu/dataset/34/diabetes
[6]https://github.com/qiyuangong/Mondrian
[7]https://pypi.org/project/anonypy/

from the `scikit-learn` Python library[8] set the parameter `n_estimators` to 100, and employed the *Gini index*.

k-Nearest Neighbors (kNNs). It is a non-parametric, supervised learning model, used for regression and classification, which predicts the target variable by considering the k nearest data points in the feature space. Specifically, it identifies the k nearest neighbors based on a chosen distance metric, in our case, the *Minkowski metric* [18]. In our study, we employed the `KNeighborsClassifier` class from the `scikit-learn` library, and set $k = 5$.

Support Vector Machines (SVMs). They are supervised learning models with associated learning algorithms that build hyperplanes in high-dimensional feature spaces to separate instances into different classes. SVMs can employ different kernel functions to handle both linear and non-linear classification. In our study, we used the `LinearSVC` class from `scikit-learn`, which uses a *linear kernel* and set the regularization parameter C to 1; this controls the trade-off between achieving a larger margin and minimizing classification errors.

XGBoost (XGB). It is an optimized gradient-boosting algorithm known for its high performance [4]. It creates a strong ensemble model combining weak models such as decision trees. It employs boosting techniques to chain model training and makes the subsequent model focus on the enhancement of the previous one. In our study, we used the `XGBClassifier` clas from `scikit-learn` and set the parameter `n_estimators` to 100.

D. Performance Evaluation and Explainability Analysis

The classification performance of ML models on the considered datasets was evaluated by performing 10-*fold cross-validation* using the `cross_val_score` class from `scikit-learn`, selecting as scoring functions `accuracy` and `f1`, and setting `n_jobs` (the number of jobs) to -1, to use all processors available for computation. To assess model explainability, we implemented SHAP and LIME frameworks by means of the official Python classes `shap`[9] and `lime`[10].

IV. Experimental Evaluations

The primary objective of our evaluation is to assess the variations in adaptability and explainability shown by ML models when the considered *features* are sanitized via k-anonymity.[11] To emphasize these variations, each model was trained on distinct sanitized datasets. For each k-anonymous dataset (when $k = 1$, we deal with the original datasets), and for each ML model, we conducted both a *classification performance evaluation* and an *explainability analysis* of features, the latter based on their average importance derived from Shapley values.

[8] https://scikit-learn.org/stable/index.html
[9] https://pypi.org/project/shap/
[10] https://pypi.org/project/lime/
[11] In the remainder of this paper, we refer to dataset *attributes* as *features* when referring to ML tasks.

A. Classification Performance Evaluation

Given that the primary objective of this study is not to optimize the classification performance of ML models, but to examine the impact of non-perturbative sanitization on both performance and explainability (and their interplay), no feature selection/engineering techniques, or model tuning, have been applied to ML classifiers. This choice aims to observe changes in performance outcomes only for the effects of non-perturbative sanitization.

Table I presents performance results on the ADULT dataset in terms of both *accuracy* (Acc) and F1-*score* (F1) (the metrics also used for the other datasets). The results indicate that RFs and XGB models achieve higher performance, outperforming both kNNs and SVMs models, which exhibit comparatively lower performance even on the non-sanitized dataset. As a general observation, it can be noted that, for increasingly higher values of the sanitization parameter k, there is no significant decrease in the F1-score. This suggests that the model can effectively leverage the remaining features to maintain or even enhance its predictive capabilities despite the sanitization process, keeping similar performance to $k = 1$.

In Table II we can observe, on the *Diabetes* dataset, the RFs and XGB models performing slightly better than the other two. In this case, a small decrease in F1-scores is observed across all models when transitioning from the original dataset to the first level of sanitization with $k = 2$. This disparity can be attributed to the presence, in the ADULT dataset, of other influential features (e.g., *capital-gain*) that contribute to maintaining performance levels similar to those of the original dataset, while in the *Diabetes* dataset there are no strong features other than the sanitized ones. This lack of additional influential features leads to a more pronounced, but not heavy, impact on model performance during the sanitization process.

TABLE I: Classification performance on ADULT as k varies.

	RFs		kNNs		SVMs		XGB	
k	Acc	F1	Acc	F1	Acc	F1	Acc	F1
1	0.86	0.85	0.78	0.75	0.80	0.76	0.87	0.86
2	0.85	0.85	0.77	0.75	0.79	0.72	0.87	0.86
5	0.85	0.85	0.77	0.75	0.79	0.73	0.87	0.86
10	0.85	0.84	0.77	0.75	0.79	0.73	0.86	0.86
20	0.84	0.84	0.77	0.75	0.79	0.73	0.87	0.86
50	0.85	0.84	0.77	0.75	0.79	0.74	0.87	0.86
70	0.83	0.83	0.77	0.75	0.80	0.76	0.86	0.85
100	0.83	0.83	0.77	0.75	0.80	0.76	0.86	0.85

TABLE II: Classification performance on *Diabetes* as k varies.

	RFs		kNNs		SVMs		XGB	
k	Acc	F1	Acc	F1	Acc	F1	Acc	F1
1	0.97	0.97	0.95	0.95	0.96	0.96	0.97	0.97
2	0.95	0.95	0.95	0.94	0.94	0.93	0.96	0.95
5	0.95	0.95	0.95	0.94	0.94	0.93	0.96	0.95
10	0.95	0.95	0.95	0.94	0.94	0.93	0.95	0.95
20	0.95	0.94	0.95	0.94	0.94	0.93	0.95	0.95
50	0.94	0.94	0.95	0.94	0.93	0.91	0.95	0.95
70	0.94	0.94	0.95	0.94	0.93	0.92	0.95	0.95
100	0.94	0.94	0.95	0.94	0.93	0.90	0.95	0.94

B. Explainability Analysis with SHAP

We employed SHAP to determine *feature importance ranks*. Following the generation of *beeswarm plots*, the observed fluctuation in rank as k increases was systematically recorded and subsequently represented in the line graphs presented herein. Visual trends of feature importance ranks are shown by means of the `matplotlib` Python class,[12] for distinct values of k on the considered datasets and classifiers.

ADULT. The impact of the anonymization of the quasi-identifiers of ADULT (see Section III-B) using the four ML models is illustrated in Figure 3.

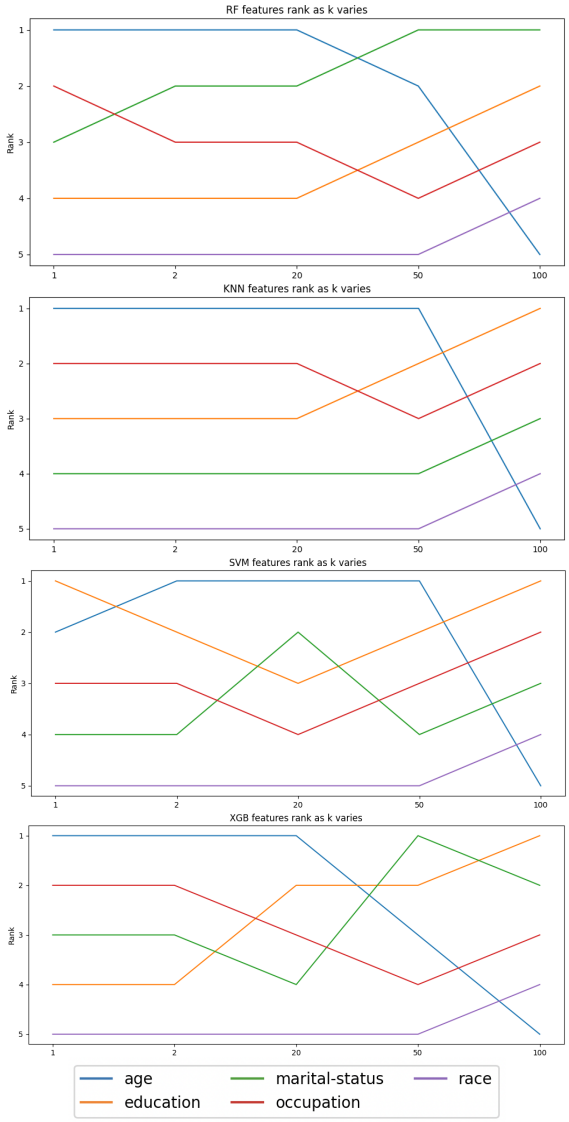

Fig. 3: Feature importance rank trends associated with RFs, kNNs, SVMs, and XGB, as k varies on the ADULT dataset.

[12] https://matplotlib.org/

Observing the results, it is evident that *age* holds a dominant position in the feature importance ranking for the non-anonymized dataset with $k = 1$ (except for SVMs). However, as k increases, its importance undergoes a gradual decline until reaching a significant drop at $k = 20$ (RFs and XGB) and at $k = 50$ (kNNs and SVMs), experiencing a drastic decline with further anonymization ($k \in \{50, 70, 100\}$) for all models. The effects of this decline appear to be offset by the growth in the importance of alternative features, indicating the models' ability to adapt when some features are obscured or less discernible. Greater variability can be observed for XGB; this is likely because of the model's adaptive boosting and its capacity to adjust its group of weak learners to capture and utilize changing patterns and data relationships.

Diabetes. The anonymization of the quasi-identifiers of the *Diabetes* dataset (see Section III-B) yielded distinctly different outcomes in terms of feature importance ranks compared to the results observed in the ADULT dataset. In *Diabetes*, the process of anonymization exhibited a minimal impact, with the ranks of the features displaying greater stability across various levels of anonymization. Since no significant changes in feature importance ranks are observed for RFs and XGB models, for the sake of conciseness, Figure 4 illustrates the feature importance rank trends for kNNs and SVMs models only; indeed, they exhibit just initial rank fluctuations among feature importance.

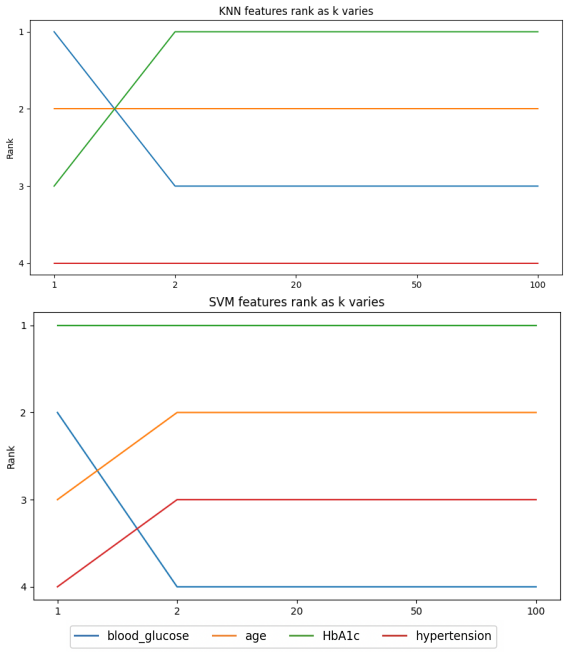

Fig. 4: Feature importance rank trends associated with kNNs and SVMs as k varies on the *Diabetes* dataset.

C. Explainability Analysis with LIME

Through LIME, we analyzed the importance of each feature and determined if the same subset of features that maintained

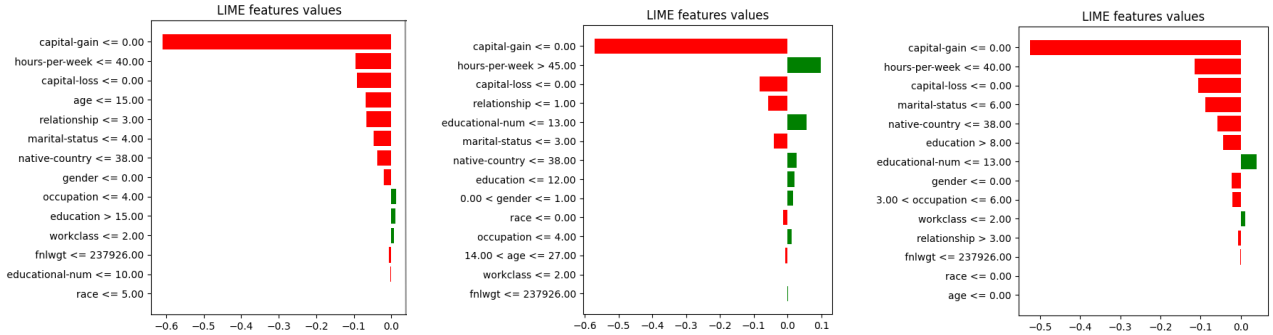

Fig. 5: Feature importance obtained with LIME for three anonymized ADULT instances using the RFs model. Three key values of k are tested, i.e., $k = 2$ (a), $k = 20$ (b), $k = 100$ (c).

high ranks in the feature importance rank trends obtained with SHAP align with the features identified as influential by LIME.

In Figure 5, three examples of feature interpretability using LIME are presented for the ADULT dataset using the RFs model, for $k = 2$ (Figure 5(a)), $k = 20$ (Figure 5(b)), and $k = 100$ (Figure 5(c)). A prominence of the *capital-gain* feature as a highly influential factor can be observed, confirming the assertion made in the previous section that this feature played a crucial role in mitigating significant performance drops during the anonymization process. These findings align with the observed trends in feature importance ranks from the previous section on SHAP analysis. They also confirm that feature *age*, which initially held a top rank for $k = 2$, gradually decreased for $k = 20$, and ultimately reached zero for $k = 100$, thereby providing further support to the previous findings. Due to space constraints, we report in the paper only this example, with the note that further evidence is available at the link that contains supporting material.

D. Discussion of the Results

The analysis of both model classification performance and explainability analysis provides useful insights into the impact of non-perturbative anonymization on ML models. From the feature importance rank trends, we observed consistent patterns across different datasets and models.

In the ADULT dataset, the anonymization brought some light changes, where the most notable was on the *age* feature which, from a global interpretability point of view, started in the top-rank position and ended in the last one for high k values. This suggests that the models adapted to use alternative features to maintain their performance in the presence of anonymization.

In the *Diabetes* dataset, the impact of anonymization on feature importance rank was less pronounced. In this case, ML models exhibited a stronger dependency on the selected quasi-identifiers for generating predictions, even post-anonymization. Such stability in feature importance ranks could be indicative of the inherent characteristics of the Diabetes dataset, where the quasi-identifiers may hold substantial predictive power that is less susceptible to the effects of anonymization.

In general, we can observe that when strong features in both datasets are sanitized, the ML models demonstrate the ability to adapt and leverage alternative features to maintain pretty satisfactory performance levels, which are almost comparable to those of pre-sanitization.

V. Conclusions and Further Research

In the context of online data dissemination, some of which need to be protected, we studied the impact of non-perturbative sanitization on the performance and explainability of four ML classifiers considering different publicly available datasets. In particular, with regard to explainability, we used Shapley values to build a feature importance rank for each anonymized dataset at distinct anonymization levels, and, by analyzing the resulting changes, we showed insights into the effects of anonymization on the interpretability of the selected ML models. This study further reveals discernible patterns in feature rank trends, thereby illuminating the adaptability or challenges encountered by models, as exemplified by the linear SVMs model.

Future research can explore additional anonymization techniques (e.g., ℓ-diversity and t-closeness) and their impact on additional datasets and ML models. Indeed, in this preliminary study, we used only basic classification models; furthermore, they were performed on minimally processed data, without possibly studying parameter optimization versus data anonymization. Therefore, it becomes important to investigate the performance of more advanced (and optimized) models combined with anonymization techniques and explainability paradigms, as well as advanced data pre-processing methods. For example, studying the effects of anonymization on Deep Learning models. We must also emphasize that we considered structured data that can be disseminated online. However, we are well aware of data protection issues involving semi-structured or unstructured data, which largely characterize the Web ecosystem [14], [15]. For this reason, it will be equally important to study the impact of innovative data protection

solutions in this context, including with respect to the use of cutting-edge technologies like Large Language Models (LLMs).

Acknowledgements

This work was supported in part by project SERICS (PE00000014) under the NRRP MUR program funded by the EU - NGEU, by project KURAMi: Knowledge-based, explainable User empowerment in Releasing private data and Assessing Misinformation in online environments (Italian Ministry of University and Research – PRIN 2022: 20225WTRFN) https://kurami.disco.unimib.it/ and by the EC under grants MARSAL (101017171) and GLACIATION (101070141).

Code and Data Availability

The code developed for this study is available, along with additional material, at the URL: https://github.com/stefanolocci/WIC-WI-IAT-23. The datasets employed in this work are publicly accessible and their URLs have been provided within the paper itself.

References

[1] R. J. Bayardo and R. Agrawal, "Data privacy through optimal k-anonymization," in *Proc. of ICDE 2005*, Tokoyo, Japan, April 2005.

[2] A. Bozorgpanah, V. Torra, and L. Aliahmadipour, "Privacy and explainability: The effects of data protection on Shapley values," *Technologies*, vol. 10, no. 6, p. 125, 2022.

[3] L. Breiman, "Random forests," *Machine Learning*, vol. 45, no. 1, pp. 5–32, Oct 2001. [Online]. Available: https://doi.org/10.1023/A:1010933404324

[4] T. Chen and C. Guestrin, "XGBoost: A scalable tree boosting system," in *Proc. of KDD 2016*, San Francisco, California, USA, 2016, pp. 785–794. [Online]. Available: https://doi.org/10.1145/2939672.2939785

[5] S. De Capitani di Vimercati, S. Foresti, G. Livraga, and P. Samarati, "Data privacy: Definitions and techniques," *IJUFKBS*, vol. 20, no. 6, pp. 793–817, 2012.

[6] F. K. Došilović, M. Brčić, and N. Hlupić, "Explainable artificial intelligence: A survey," in *Proc. of MIPRO 2018*, Opatija, Croatia, 2018.

[7] C. Dwork, "Differential privacy," in *Proc. of ICALP 2006*, Venice, Italy, 2006.

[8] B. C. Fung, K. Wang, R. Chen, and P. S. Yu, "Privacy-preserving data publishing: A survey of recent developments," *ACM CSUR*, vol. 42, no. 4, pp. 1–53, 2010.

[9] B. C. Fung, K. Wang, and S. Y. Philip, "Anonymizing classification data for privacy preservation," *IEEE TKDE*, vol. 19, no. 5, pp. 711–725, 2007.

[10] P. Golle, "Revisiting the uniqueness of simple demographics in the US population," in *Proc. of WPES 2006*, Alexandria, VA, USA, 2006.

[11] J. Konečný, H. B. McMahan, F. X. Yu, P. Richtarik, A. T. Suresh, and D. Bacon, "Federated learning: Strategies for improving communication efficiency," in *Proc. of PMPML 2016*, Barcelona, Spain, 2016. [Online]. Available: https://arxiv.org/abs/1610.05492

[12] K. LeFevre, D. DeWitt, and R. Ramakrishnan, "Mondrian multidimensional k-anonymity," in *Proc. of ICDE 2006*, Atlanta, GA, USA, 2006.

[13] N. Li, T. Li, and S. Venkatasubramanian, "t-closeness: Privacy beyond k-anonymity and l-diversity," in *2007 IEEE 23rd international conference on data engineering*. IEEE, 2006, pp. 106–115.

[14] G. Livraga, A. Motta, and M. Viviani, "Assessing user privacy on social media: The Twitter case study," in *Proc. of OASIS@HT 2022*, virtual, 2022.

[15] G. Livraga and M. Viviani, "Data confidentiality and information credibility in on-line ecosystems," in *Proc. of MEDES 2019*, Limassol, Cyprus, 2019.

[16] S. M. Lundberg and S.-I. Lee, "A unified approach to interpreting model predictions," *Advances in neural information processing systems*, vol. 30, 2017.

[17] A. Machanavajjhala, D. Kifer, J. Gehrke, and M. Venkitasubramaniam, "ℓ-diversity: Privacy beyond k-anonymity," *ACM TKDD*, vol. 1, no. 1, pp. 3:1–3:52, 2007.

[18] M. Mailagaha Kumbure and P. Luukka, "A generalized fuzzy k-nearest neighbor regression model based on Minkowski distance," *Granular Computing*, vol. 7, no. 3, pp. 657–671, 2022.

[19] I. Ogrezeanu, A. Vizitiu, C. Ciușdel, A. Puiu, S. Coman, C. Boldișor, A. Itu, R. Demeter, F. Moldoveanu, C. Suciu, and L. Itu, "Privacy-preserving and explainable AI in industrial applications," *Applied Sciences*, vol. 12, no. 13, 2022. [Online]. Available: https://www.mdpi.com/2076-3417/12/13/6395

[20] C. Orlandi, A. Piva, and M. Barni, "Oblivious neural network computing via homomorphic encryption," *EURASIP Journal on Information Security*, vol. 2007, 2007.

[21] M. T. Ribeiro, S. Singh, and C. Guestrin, "Why should I trust you?: Explaining the predictions of any classifier," in *Proc. of KDD 2016*, San Francisco, CA, USA, 2016, pp. 1135–1144. [Online]. Available: https://doi.org/10.1145/2939672.2939778

[22] G. Rjoub, J. Bentahar, O. A. Wahab, R. Mizouni, A. Song, R. Cohen, H. Otrok, and A. Mourad, "A survey on explainable artificial intelligence for cybersecurity," *IEEE TNSM*, 2023, pre-print.

[23] A. Rodríguez-Hoyos, J. Estrada-Jiménez, D. Rebollo-Monedero, J. Parra-Arnau, and J. Forné, "Does k -anonymous microaggregation affect machine-learned macrotrends?" *IEEE Access*, vol. 6, pp. 28 258–28 277, 2018.

[24] S. Saifullah, D. Mercier, A. Lucieri, A. R. Dengel, and S. Ahmed, "Privacy meets explainability: A comprehensive impact benchmark," *ArXiv*, vol. abs/2211.04110, 2022.

[25] R. Saleem, B. Yuan, F. Kurugollu, A. Anjum, and L. Liu, "Explaining deep neural networks: A survey on the global interpretation methods," *Neurocomputing*, vol. 513, pp. 165–180, 2022. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0925231222012218

[26] P. Samarati, "Protecting respondents' identities in microdata release," *IEEE TKDE*, vol. 13, no. 6, pp. 1010–1027, 2001.

[27] D. Slijepčević, M. Henzl, L. D. Klausner, T. Dam, P. Kieseberg, and M. Zeppelzauer, "k-Anonymity in practice: How generalisation and suppression affect machine learning classifiers," *COSE*, vol. 111, 2021.

[28] I. Stepin, J. M. Alonso, A. Catalá, and M. Pereira-Fariña, "A survey of contrastive and counterfactual explanation generation methods for explainable artificial intelligence," *IEEE Access*, vol. 9, pp. 11 974–12 001, 2021.

[29] F. Xu, H. Uszkoreit, Y. Du, W. Fan, D. Zhao, and J. Zhu, "Explainable AI: A brief survey on history, research areas, approaches and challenges," in *Proc. of NLPCC 2019*, Dunhuang, China, 2019.

[30] J. Xu, W. Wang, J. Pei, X. Wang, B. Shi, and A.-C. Fu, "Utility-based anonymization for privacy preservation with less information loss," *ACM SIGKDD Explorations Newsletter*, vol. 8, no. 2, pp. 21–30, 2006.

[31] R. N. Zaeem and K. S. Barber, "The effect of the GDPR on privacy policies: Recent progress and future promise," *ACM TMIS*, vol. 12, no. 1, pp. 1–20, 2020.