Privacy Aware Processing

Sharing Uncritical Information to Enable Data Understanding and Preparation on Sensitive Data for Machine Learning

Marian Eleks[1], Jonas Rebstadt[2], Henrik Kortum[3] and Oliver Thomas[4]

Abstract: In many machine learning (ML) applications, the provision of data and the training as well as the analysis of machine learning systems are performed by distinct actors, a data owner and a data consumer. To protect sensitive information in these ML-scenarios, privacy aware machine learning (PAML) methods are often applied to the data before sharing. Based on the type of PAML methods used, data understanding and preparation as defined in the CRISP-DM model become more difficult if not impossible. To enable these steps, we propose a method to share a variety of uncritical information with the data consumer who is then able to define the necessary processing steps on a meta-level. These are then applied to the data in the data owners local trusted environment before the PAML-methods whereupon the prepared and protected data is shared.

Keywords: Privacy Aware Machine Learning, Data Understanding, Data Preparation, CRISP-DM, Artificial Intelligence, Machine Learning

1 Introduction

Privacy aware machine learning (PAML) is becoming increasingly relevant due to the widespread adoption of machine learning (ML) in various application domains, as well as growing concerns about data privacy and security. PAML use cases with distinct data owners and data consumers find growing interest as ML permeates domains where ML knowledge is limited but data is highly sensitive. In a practical example. private households or the housing industry have an urge to reduce their usage of energy but don't want to or are not allowed to share the relevant information for promising ML-services [Ko23]. In these cases, the sensitive information in the patient data would be protected through various PAML-techniques before being sent to the researchers, fulfilling the

[1] Strategion GmbH, Albert-Einstein-Straße 1, 49076 Osnabrück, marian.eleks@strategion.de, https://orcid.org/0000-0002-1516-5129
[2] Deutsches Forschungszentrum für Künstliche Intelligenz GmbH, Smart Enterprise Engineering, Hamburger Straße 24, 49084 Osnabrück, jonas.rebstadt@dfki.de, https://orcid.org/0000-0001-8531-3273
[3] Deutsches Forschungszentrum für Künstliche Intelligenz GmbH, Smart Enterprise Engineering, Hamburger Straße 24, 49084 Osnabrück, henrik.kortum@dfki.de, https://orcid.org/0000-0002-1089-711X
[4] Universität Osnabrück, Informationsmanagement und Wirtschaftsinformatik, Hamburger Straße 24, 49084 Osnabrück, oliver.thomas@uni-osnabrueck.de

privacy goals but preventing meaningful data understanding and preparation. However, a well-founded data understanding and preparation is essential for an optimal modelling outcome and are found in most AI-Lifecycle frameworks [Am19], [AS08], [Er17] including CRISP-DM [WH00]. To enable data understanding and preparation in PAML use cases with distinct actors, a technique of extracting and sharing uncritical information from the data called "privacy aware preview" is developed. Based on this, "privacy aware processing" is constructed with the core idea of the data consumer defining data preparation instructions based on the data understanding achieved in the privacy aware preview, which are then executed in the data owner's local environment to achieve prepared and private data that is then shared with the data consumer. First, related work for privacy aware machine learning and its impact on CRISP-DM is introduced. Then, existing approaches for the previously outlined problem are collected in a literature review followed by the development of this paper's solution, based on the findings in the literature. Lastly, the impact of the solution on CRISP-DM is addressed, followed by a discussion of the results and a conclusion.

2 Privacy Aware Machine Learning and its Consequences for CRISP-DM

The term Privacy Aware Machine Learning (PAML) describes "[…] the discipline of applying Machine Learning techniques in such a way as to protect and retain personal identities during the process." [MKH17]. The goal of the discipline is to share and analyse data while protecting sensitive attributes [El22]. PAML might also be used to describe similar undertakings [AC19]. The literature mainly mentions six techniques: K-Anonymity generalizes and suppresses data, resulting in a dataset where any record has at least k records with identical quasi-identifiers [RMS18], [SD18], [Sw02a]. Differential Privacy's basic idea is to add noise to the data before releasing it, in order to make it difficult to determine whether any particular individual's data is included in the dataset [AC19], [Dw06], [LZ20], [RMS18], [SD18]. Synthetic Data, while mimicking the statistical properties of sensitive data, does not contain any real-world information, making it privacy preserving [Ab19], [HEM19], [Pa18]. Homomorphic Encryption enables computations on encrypted data so that model training may be accomplished on said data without decryption [AC19], [LZ20]. Similarity Preserving Hashing transforms the data to remove any interpretable information while preserving utility for ML [El22]. Lastly, Federated Learning keeps sensitive data on local devices, training individual models locally. A central model is achieved by aggregating the locally trained models without ever combining or seeing local sensitive data [Me21].

While PAML techniques make model training on sensitive data possible while preserving privacy, most ML procedure model feature a data understanding as well as a data preparation stage [Am19], [AS08], [Er17], [WH00]. By applying PAML methods to the data in the distinct actor use case, data understanding and preparation are complicated if not impossible since the data consumer receives incomplete (k-Anonymity), perturbated

(Differential Privacy) or incomprehensible (SPH, Homomorphic Encryption) information. The cycle of procedure models like CRISP-DM is broken in PAML use cases, raising a need for techniques that enable data understanding and preparation in a privacy aware context. This need is further justified in the literature [Br21], [Fa20], [JLE14], [LP21].

3 Privacy Aware Preview and Processing

3.1 Existing approaches and current limitations

To find existing solutions and ideas concerning data understanding, preprocessing and feature engineering in a privacy aware environment, a literature review using the search term *("privacy" OR "security" OR "sensitive") AND "machine learning" AND ("preprocessing" OR "feature engineering" OR "data understanding")* is carried out. It yields approaches for privacy aware data understanding and privacy aware data preparation as well as current challenges and needs.

The simplest approach for privacy aware data understanding is to remove all sensitive data before sharing [SBG21], which is not practical since it entails large losses in data utility. Differential Privacy can also be used to share aggregate information about the data while preserving the individuals privacy [JLE14], [Sh23], which is beneficial to assess data distribution, value spaces, imbalances and missing values. Additionally, [JLE14] mention data generation as a possible solution for private data release, raising the possibility to share synthetic plain text data as a proxy for understanding the original data. [Jo19] invert the use case and use feature engineering to hide sensitive information, e.g., transforming GPS-coordinates into a place label like “Walgreens parking lot”. This idea promises great results in use cases where a trusted entity is able to prescribe these kinds of transformations. This entity is not present in the use case of this paper, so it is not applicable. A final approach is identifying similar public datasets to the private data [JLE14], but since these are seldomly available, they will also not be used further.

All approaches for privacy aware data preparation found in the literature implicitly assume that some kind of data understanding has already taken place to be able to clearly define necessary preprocessing and transformation instructions. Thus, privacy aware data understanding is a prerequisite for privacy aware data preparation. The most promising proposition is to have the data consumer define a list of preprocessing and transformation instructions, which are then applied to the data in a trusted environment [HH22], [Jo19]. Another idea is to use homomorphic encryption in a federated learning environment to apply data preparation algorithms to encrypted data [Fa20], [HH22]. Going along with the first approach, these algorithms need to either be applicable to all kinds of raw data or be defined based on a previous data understanding. Finally, [JLE14] mention the use of principal component analysis or forms of feature selection and feature importance calculation to reduce the number of attributes and/or achieve dimensionality reduction in a privacy aware manner.

As mentioned in the data preparation, a need for privacy aware data understanding techniques is present in the literature, either explicitly [Br21], [JLE14] or implicitly [Fa20], [HH22], [Jo19]. Building from that, some papers mention a necessity for clean and complete data, implying a need for privacy aware preprocessing techniques. In some cases, experience in data science is assumed for the data owner [Fa20], [HH22], [Jo19], since they are expected to perform data manipulation. For this paper, the data does not need to have advanced knowledge of data manipulation for the solution to function. Conversely, [HH22], [JLE14] mention that the data needs to be in a relational format for some solutions to function, which is also true for this paper.

3.2 Sharing Uncritical Information for a well-founded Data Understanding

Being a prerequisite for all subsequent steps, a process for privacy aware data understanding based on sharing uncritical information is designed. A preview component generates as much uncritical information as possible from the raw data inside of the data owner's environment. Types of uncritical information from the literature review include synthetic data, which at this point does not need to come from an optimized model to provide an understanding of the data, saving resources and time. Moreover, aggregated statements fulfilling some degree of differential privacy based on requirements from the data owner can be shared to assess data distribution, value spaces, imbalances, and missing values. Using the ideas for dimensionality reduction, the data user can also be informed about the importance of attributes for certain use cases. In addition to the types found in the literature, uncritical meta information can be shared without breaching the privacy of any individual in the data. This includes attribute names, programmatical data types (int, float, string, …) and statistical data types (categorical, ordinal, binary, …). By sharing these types of information about the data with the data user, they gain an understanding of the data without acquiring sensitive information, as is shown in Figure 1.

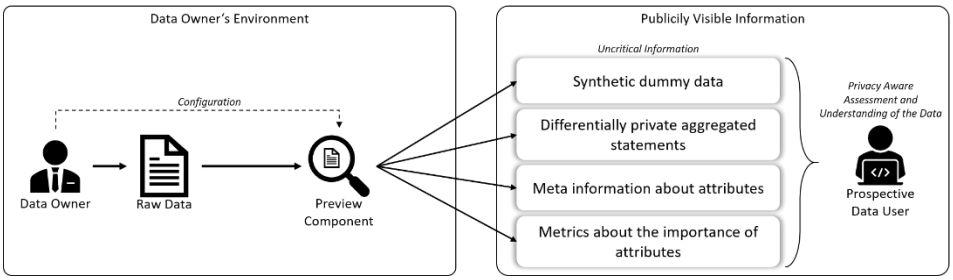

Figure 1: Privacy Aware Preview

Depending on the use case, the data understanding might be sufficient in and of itself, like in data matching or when the data user purchases usage rights to the private data after the preview. In other cases, it is used as a foundation for the following privacy aware processing.

3.3 Processing of Private Data for Data Preparation

Following [HH22], [Jo19], the privacy aware processing component allows the data user to define a list of processing steps that are then executed in the data owners trusted environment on the raw data before applying PAML-methods. This enables data users to perform data preparation on the raw data without having to breach privacy.

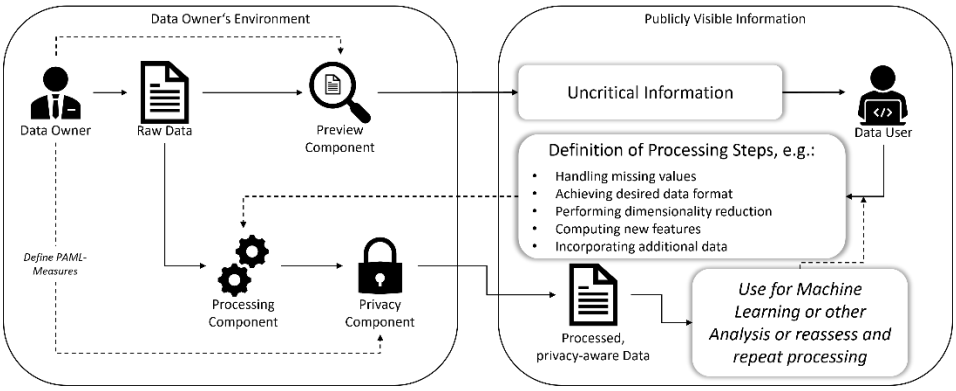

Figure 2: Privacy Aware Processing

In the full process as shown in Figure 2, data owners provide their raw data to the preview component, which they can adjust based on their notion of necessary privacy. The preview component then generates some or all of the possible types of uncritical information defined in 3.2 for the prospective data user to receive. The data user is then able to reach a decision on whether to use the data for their endeavours as well as gain an understanding of the inherent properties and necessary processing steps to undertake for data preparation. Following this, by executing a list of instructions defined by the data user in the environment of the data owner, data processing or data preparation in CRISP-DM terms can be carried out in a privacy preserving manner. Processing instructions can range from dealing with missing values or unclean data through reformatting data all the way to computing new features from the data or incorporating additional attributes from other datasets owned by the data user into the resulting dataset.

To illustrate the practical use of privacy aware processing, an example based on an insurance cost prediction dataset [Wa23] is established in Figure 3. Instead of receiving e.g. a homomorphically encrypted dataset that is uninterpretable for humans and not prepared to be used in ML, the data user receives the previously described preview. From this, they are able to deduce that with this dataset using the label "charges" and dropping any empty labels (1), a supervised training use case to predict insurance cost is possible. Additionally, they are able to deduce that the attribute region is unnecessary (2) and that the categorical attribute sex needs to be dummy encoded (3). Smoker is a string variable so it needs to be cast to binary/boolean (4) and age contains empty values that can be imputed (5). Listing further examples, the data user might choose to add normalization

instructions to age, children and BMI after seeing their value ranges, correct imbalances identified through the differentially aggregated percentiles or compute additional features (e.g. a special flag for increased COPD risk from the BMI and smoker attributes). All of these instructions are represented in source code, sent to the data owner and executed in their environment.

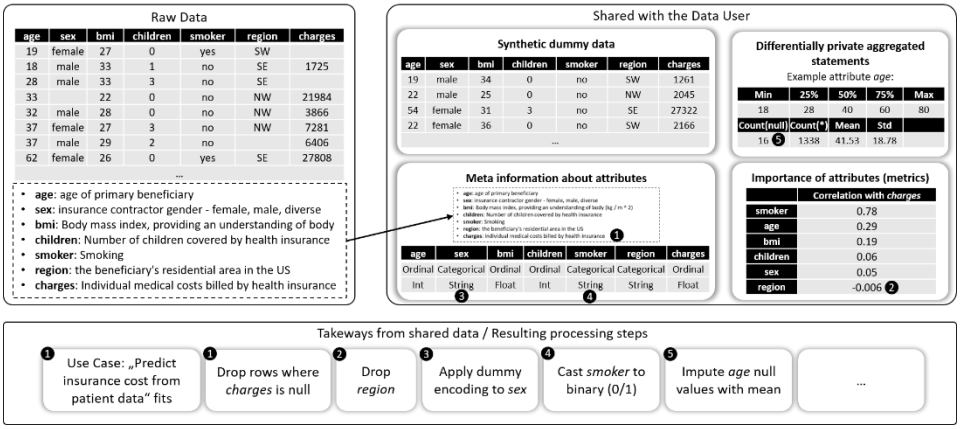

Figure 3: Practical example (insurance cost)

After the data preparation is finished, a privacy component applies PAML-measures to the data depending on the privacy requirements defined by the data owner. These can take the form of achieving k-Anonymity [RMS18], [SD18], [Sw02a] or Differential Privacy [AC19], [Dw06], [LZ20], [RMS18], [SD18] through generalization, suppression, or perturbation . Homomorphically encrypting the dataset is also possible [AC19], [LZ20] as well as training an optimized generative model to synthesize data [Ab19], [HEM19], [Pa18]. A most extreme measure would be the application of Similarity Preserving Hashing Anonymization algorithms to the data to remove any human interpretability [El22]. At this point, the resulting dataset is preprocessed, prepared and void of sensitive information and can be shared with the data user who is then able to use it as they please. They might also choose to revise their processing instructions based on newly gained insights achieved through modelling and evaluation using the data. This creates a cyclic process to achieve increasingly superior results through iterative revision, adhering to the CRISP-DM model.

4 Closing the Loop: Implications for Development Processes with CRISP-DM

In a PAML-context, the CRISP-DM cycle is broken through the harder and sometimes impossible understanding and preparation of privacy aware data. This paper mends the cycle through the ideas of privacy aware data preview and privacy aware data processing.

By providing uncritical information through various privacy aware data sharing techniques, all data understanding activities can be enabled in a privacy aware context. Synthetic initial data is provided and meta information about attributes as well as differentially private aggregated statements enable data description, exploration, and quality verification in a privacy preserving context. Data users are able to define a list of instructions for the processing component based on their data understanding and domain knowledge, enabling all data preparation activities. Additionally, privacy aware data understanding enables an informed modelling technique selection and test design, further facilitating CRISP-DM in a privacy aware environment.

5 Discussion and Conclusion

In this paper, the challenge of performing data understanding and preparation analogue to CRISP-DM in privacy aware use cases is uncovered and established through relevant literature. Based on current approaches, a solution consisting of a privacy aware data preview based on sharing uncritical information and a privacy aware data processing based on executing predefined preparation instructions in a trusted environment is designed. The solution promises to close the loop opened by applying CRISP-DM to privacy aware machine learning use cases. Its applicability is currently limited to CRISP-DM. While other AI process models seem promising, the solutions utility remains to be assessed. Additionally, it is limited to relational or tabular data and not suitable for e.g. images. As a main limitation, this paper stops at the design stage, leaving implementation and evaluation open for future research. Thus, utility assessments and acceptance by data consumers and owners respectively remain to be explored with special attention given to the most used and most useful processing steps. The designed solution does not reinvent data preprocessing or preparation but rather shifts the surrounding factors and underlying conditions to achieve better security for the data owner as well as the possibility for the data user to process raw data in their own image without having to breach privacy. As for the practical implications, this paper provides a groundwork on which domains with limited ML knowledge and highly sensitive data can build their training processes. It creates a starting point for new business models based on fully privacy aware machine learning as a service, providing solutions in domains like healthcare or smart living that fulfil the previously mentioned criteria. By extending an existing framework, service providers are able to build upon existing infrastructure and established pipelines, reducing redundant work and lowering the barrier to entry for these novel business cases. Overall, this paper solves a problem present in the literature and in established process models and provides a solution applicable to practical use cases in specific domains. Future research should focus on implementing and evaluating the privacy aware preview and processing designs in relevant applications domains. Based on the evaluation results and requirements for data consumers and owners, business models can be derived and realised in practice.

Bibliography

[Ab19] Abay, N. C.; Zhou, Y.; Kantarcioglu, M.; Thuraisingham, B.; Sweeney, L.: Privacy preserving synthetic data release using deep learning. Machine Learning and Knowledge Discovery in Databases, pp. 510–526, 2019.

[AC19] Al-Rubaie, M.; Chang, J. M.: Privacy-Preserving Machine Learning: Threats and Solutions. In: IEEE Security and Privacy 2/19, pp. 49–58, 2019.

[Am19] Amershi, S.; Begel, A.; Bird, C.; DeLine, R.; Gall, H.; Kamar, E.; Nagappan, N.; Nushi, B.; Zimmermann, T.: Software Engineering for Machine Learning: A Case Study. Proceedings - 2019 IEEE/ACM 41st International Conference on Software Engineering: Software Engineering in Practice, ICSE-SEIP 2019, Montréal, pp. 291–300, 2019.

[AS08] Azevedo, A.; Santos, M. F.: KDD, SEMMA and CRISP-DM: a parallel overview. In: IADS-DM, 2008.

[Br21] Briguglio, W.; Moghaddam, P.; Yousef, W. A.; Traoré, I.; Mamun, M.: Machine learning in precision medicine to preserve privacy via encryption. In: Pattern Recognition Letters 1/21, pp. 148–154, 2021.

[Dw06] Dwork, C.: Differential Privacy. International Colloquium on Automata, Languages, and Programming, Bd. 4052 LNCS, pp. 1–12, 2006.

[El22] Eleks, M.; Rebstadt, J.; Fukas, P.; Thomas, O.: Learning without Looking: Similarity Preserving Hashing and Its Potential for Machine Learning in Privacy Critical Domains. INFORMATIK 2022, 2022.

[Er17] Ericson, G.; Rohm, W.; Martens, J.; Harvey, B.; Schonning, N.: Team data science process documentation. Microsoft, 2017.

[Fa20] Fang, P.; Cai, Z.; Chen, H.; Shi, Q.: FLFE: A Communication-Efficient and Privacy-Preserving Federated Feature Engineering Framework. arXiv preprint arXiv:2009.02557 [cs.LG], 2020.

[HEM19] Hittmeir, M.; Ekelhart, A.; Mayer, R.: Utility and Privacy Assessments of Synthetic Data for Regression Tasks. Proceedings - 2019 IEEE International Conference on Big Data, Big Data 2019, Los Angeles, CA, pp. 5763–5772, 2019.

[HH22] Hsu, R.-H.; Huang, T.-Y.: Private Data Preprocessing for Privacy-preserving Federated Learning. 2022 IEEE 5th International Conference on Knowledge Innovation and Invention (ICKII), pp. 173–178, 2022.

[JLE14] Ji, Z.; Lipton, Z. C.; Elkan, C.: Differential Privacy and Machine Learning: a Survey and Review. arXiv preprint arXiv:1412.7584 [cs.LG], 2014.

[Jo19] Jones, M.; Johnson, M.; Shervey, M.; Dudley, J. T.; Zimmerman, N.: Privacy-Preserving Methods for Feature Engineering Using Blockchain: Review, Evaluation, and Proof of Concept. In: Journal of Medical Internet Research 8/19, pp. e13600, 2019.

[Ko23] Kortum, H.; Hagen, S.; Eleks, M.; Rebstadt, J.; Remark, F.; Lowin, M.; Mihale Wilson, C.; Eberhardt, B.; Roß, A.; Maihöfner, D.; Hinz, O.; Thomas, O.: SECAI – Sustainable Heating through Edge-Cloud-based AI Systems. In: HMD Praxis der Wirtschaftsinformatik 4/23, pp. 1-22, 2023.

[LP21] Lau, A.; Passerat-Palmbach, J.: Statistical Privacy Guarantees of Machine Learning Preprocessing Techniques. arXiv preprint arXiv:2109.02496 [cs.LG], 2021.

[LZ20] Lisin, N.; Zapechnikov, S.: Methods and Approaches for Privacy-Preserving Machine Learning. Advanced Technologies in Robotics and Intelligent Systems. Bd. 80Springer, pp. 141–148, 2020.

[Me21] Mercier, D.; Lucieri, A.; Munir, M.; Dengel, A.; Ahmed, S.: Evaluating Privacy-Preserving Machine Learning in Critical Infrastructures: A Case Study on Time-Series Classification. In: IEEE TRANSACTIONS ON INDUSTRIAL INFORMATICS X, 2021.

[MKH17] Malle, B.; Kieseberg, P.; Holzinger, A.: Interactive Anonymization for Privacy aware Machine Learning. Institute for Medical Informatics, Statistics & Documentation, Medical University Graz, Austria, Graz, 2017.

[Pa18] Park, N.; Mohammadi, M.; Gorde, K.; Jajodia, S.; Park, H.; Kim, Y.: Data synthesis based on generative adversarial networks. In: Proceedings of the VLDB Endowment 10/18, pp. 1071–1083, 2018.

[RMS18] Ram Mohan Rao, P.; Murali Krishna, S.; Siva Kumar, A. P.: Privacy preservation techniques in big data analytics: a survey. In: Journal of Big Data 1/18, pp. 1–12, 2018.

[SBG21] Soni, M.; Barot, Y.; Gomathi, S.: A review on Privacy-Preserving Data Preprocessing. In: Journal of Cybersecurity and Information Management 2/21: Special Issue-RIDAPPH, pp. 16–30, 2021.

[SD18] Salas, J.; Domingo-Ferrer, J.: Some Basics on Privacy Techniques, Anonymization and their Big Data Challenges. In: Mathematics in Computer Science 3/18, pp. 263–274.

[Sh23] Shen, H.; Li, J.; Wu, G.; Zhang, M.: Data release for machine learning via correlated differential privacy. In: Information Processing & Management 3/23, pp. 103349, 2023.

[Sw02a] Sweeney, L.: k-Anonymity: A Model for Protecting Privacy, International Journal on Uncertainty, Fuzziness and Knowledge-based Systems 5/02, pp. 557–570, 2002.

[Wa23] Wakefield, B.: Prediction of Insurance Charges, Kaggle, 2023. URL: https://www.kaggle.com/datasets/thedevastator/prediction-of-insurance-charges-using-age-gender, accessed 02/06/2023.

[WH00] Wirth, R.; Hipp, J.: CRISP-DM: Towards a Standard Process Model for Data Mining. Proceedings of the 4th international conference on the practical applications of knowledge discovery and data mining, London, pp. 29–39, 2000.