A Blockchain-Based Approach to Data Privacy Using Pseudonymization in Thailand

Woratat Makasiranondh*, and Supanit Angsirikul

College of Digital Innovation Technology, Rangsit University, Pathumthani, Thailand
*Corresponding author, E-mail: mworatat@rsu.ac.th

Abstract

The advent of privacy laws, such as Thailand's PDPA and the GDPR, has created a critical need for robust privacy-enhancing technologies to secure sensitive data in research and online service industries. Data controllers, like RSU-Information Technology Service Center (RSU-ITSC), bear the responsibility to adhere to these regulations. Pseudonymization and anonymization are pivotal in minimizing risks within the online data service industry, each serving different methods and purposes. This research proposes a hybrid framework that integrates hash-based pseudonymization and k-anonymity techniques, utilizing blockchain technology to store direct identifiers separately from general data to ensure non-attribution. The experiment was conducted on the pilot "RSU-Data-Anonymization" simulated dataset across seven machine learning algorithms. The results revealed that Gradient Boosting demonstrates superior performance, while highlighting the inherent trade-off between data utility loss in k-anonymity and re-identification risks in pseudonymization. The experiment provides systematic guidelines for the data industry, offering enhanced protection against breaches and a secure, programmable method for maintaining regulatory compliance while preserving individual privacy.

Keywords: *data anonymization, pseudonymization, k-anonymity, re-identification, cybersecurity*

1. Introduction

Most companies create, store, and process immensely sensitive data. For example, travel information is held on personal mobile devices. Information is embedded in various social media sites. Every day, information about our gender, age, interests, salary, etc. is provided to others via an internet-connected mobile network. Cell phone companies can discover information such as who went where and when because they manage network data; therefore, they take full responsibility for data privacy for their customers. Data privacy directly affects the credibility of the company itself. Moreover, privacy concerns are increasingly emphasized by privacy laws such as GDPR. The data-driven industry needs technological advancements such as privacy-enhancing technologies (PETs) that are specifically designed to safeguard the privacy and confidentiality of personal information. They aim to empower individuals, mitigate surveillance risks, and ensure the security of sensitive data (Information Commissioner's Office [ICO], 2022b).

The implementation of GDPR in 2018 has had a significant impact on data-sharing practices in European countries. Popular data sources like UCI Machine Learning Repository and Kaggle have also been affected, particularly in sharing datasets containing personal information. Typically, these datasets do not include direct identification data, such as full names, to maintain privacy. However, re-identification attacks occurred as early as 2009 (Henriksen-Bulmer and Sheridan, 2016), raising concerns about proper data publication on the internet. This finding has prompted all data publishers to become more vigilant about the risks of re-identification attacks. Recently, most government agencies have asked mobile companies to perform data anonymization before sending personal data to their network. Data anonymization has therefore become very important for mobile companies.

Article 4(5) of the GDPR introduces an innovative approach known as pseudonymization. When applied to independently stored personal data, it adds an extra layer of security. Pseudonymization is becoming increasingly popular in e-commerce due to its reversible nature, making it suitable for historical, scientific, and statistical purposes. Another groundbreaking technology of this era is blockchain, which utilizes a vast distributed network, providing robust protection against data tampering by hackers. Storing

personally identifiable information and sensitive data on a blockchain ensures enhanced security, effectively guarding against online data breaches and preserving the privacy of all personal data.

Pseudonymization, as defined by the GDPR, refers to the process of managing personal data in a way that it is no longer directly linked to a specific individual unless additional information is used (Bourka et al., 2019). In compliance with the GDPR, pseudonymizing a data set requires keeping the "additional information" separate and implementing suitable technical and organizational measures to ensure non-attribution to an identified or identifiable person.

Re-identification (El Emam et al., 2011) is the act of linking previously anonymized data back to its original individuals, unveiling their identities. This process undermines the privacy safeguards intended by data anonymization, often using various techniques to match the data with specific individuals. It underscores the difficulty of balancing data usability with individual privacy protection. As re-identification attacks improve in their ability to de-anonymize data, even information initially considered non-personal can be turned into personally identifiable data. This situation introduces legal uncertainties for researchers conducting re-identification attacks (Lauradoux et al., 2023).

The quasi-identifier is defined by OECD as all other indirect attributes that can re-identify individuals (OECD, 2005). According to the GDPR definition of pseudonymization, "additional information," including unique identifiers and quasi-identifiers, is separated from the original table and separately stored in the blockchain. All other data excluding both identifiers are stored on a server. This scheme preserves original personal data. To ensure non-attribution, the directly identifiable data is securely stored separately from the processed data.

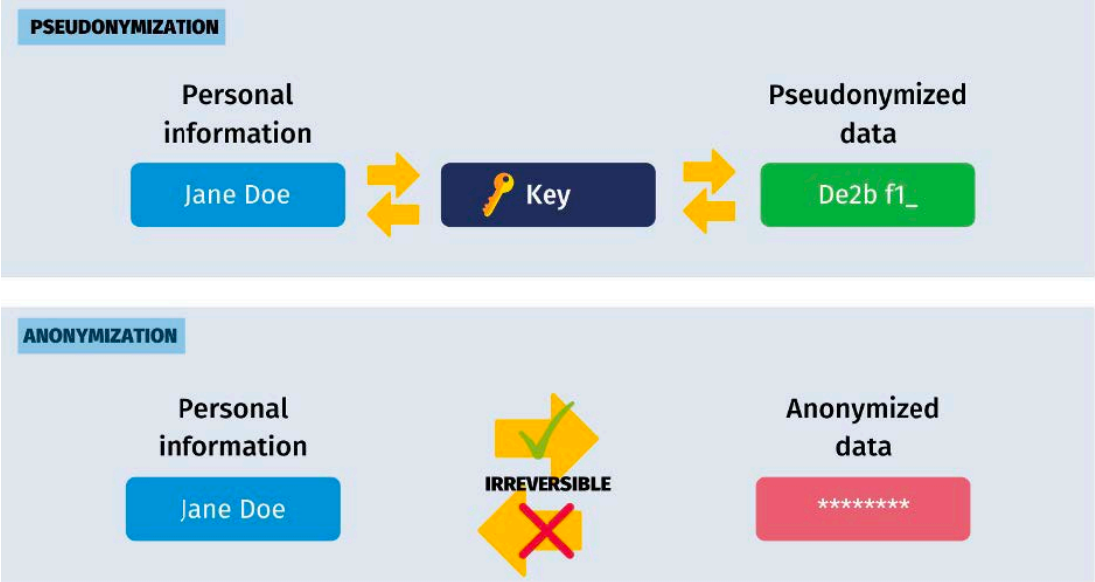


Figure 1 The difference between privacy protection pseudonymization and anonymization
Source: https://www.techyv.com/article/top-ten-privacy-enhancing-technologies/

Figure 1 illustrates the distinction between pseudonymization and anonymization. Anonymization represents a unidirectional and irreversible process. The purpose of this paper is to implement both techniques to leverage the unique advantages they offer.

Anonymization is one of the key elements of privacy-enhancing technologies (PETs). For many years, anonymization has been a widely used process to protect personal data by erasing or encrypting identifiers. Anonymization is based on the principle that breaking the association between specific data items is the primary objective to preserve privacy. The individual data items can be safely stored in cleartext since they do not contain sensitive information and may only consist of non-critical data (e.g., publicly available information). This approach ensures that privacy concerns are addressed by focusing on de-identifying the

data associations while still retaining the utility of the individual data items. However, the de-anonymization method is used by hackers to re-identify original personal data. According to the Information Commissioner's Office (Information Commissioner's Office, 2022a), anonymous data is data that cannot identify individuals; therefore, it does not apply to privacy law.

Pseudonymization and anonymization are closely related techniques, but they have a key difference. In pseudonymization, personally identifiable information is not entirely removed from the data; instead, it is replaced with a pseudonym or unique identifier.

Data breaches are often associated with malicious hacking, but they can stem from human errors or internal mishandling as well (Siteimprove, 2018). The term "personal data breach" encompasses a range of security lapses caused by intentional, negligent, accidental, or unauthorized actions, including computer crimes, cyber threats, mistakes, and more (Tilleke & Gibbins, 2022). This breach classification includes confidentiality, integrity, and availability breaches.

The research paper follows a structured organization. It begins with the introduction of the problem statement in Section 1. Next, Section 2 presents a comprehensive literature review. The methodology employed is described in Section 3. Section 4 is dedicated to presenting the experiment conducted as part of the research. The performance evaluation is discussed in Section 5. Finally, the paper concludes with the findings and conclusions in Section 6. This organized approach allows for a clear and systematic presentation of the research process and results.

2. Literature Review

Al-Zubaidie et al. (2019) proposed a PAX system that incorporates pseudonymization and anonymization to protect patient data and identity in healthcare systems, particularly EHRs (Electronic Health Records). This system addresses both external and internal attacks that could violate privacy.

Abu Attieh et al. (2025) reviewed pseudonymization tools for medical research to help researchers select appropriate tools for various projects. They screened 1,052 articles, narrowed them down to 92 full articles, identified 20 tools, and evaluated 10 based on four dimensions: single/multi-center, short/long-term, small/big data, and integration/standalone. A diverse range of tools was found, such as Software-as-a-Service (SaaS) for centralized use without requiring local infrastructure, capable of providing targeted guidance for research projects.

Aamot et al. (2013) address a clinical need for de-anonymizing research data, to benefit patients individually. They proposed a pseudonymization technique that provides both privacy and the ability to safely expose the re-identification data to authorized third parties.

Emerick et al. (2024) aim to solve privacy problems in healthcare data sharing that require high levels of security and utility, using a multi-layered architecture for IoT-enabled environments.

Holmes (2015) stated that the perspective of national security can be divided into two important aspects, namely the first perspective which focuses on the military aspect.

Sweeney (2002) stated that only three attributes can re-identify a personal record by using commercial data or auxiliary information. The re-identification attack is a process that links commercial data with target data. From that linked information, some personal data is revealed. The author proposed a concept of "hiding in the cloud" and named her algorithm k-anonymity. The algorithm generalizes all quasi-identifiers to prevent linkage to an individual personal record.

Riotta (2023) suggested that the Census Bureau should implement a data privacy algorithm to prevent reconstruction attacks. From the Census Bureau attack event, the reconstruction occurred by using five characteristics: Census block, age, sex, race, and ethnicity.

Narayanan and Shmatikov (2008) proposed a de-identification method and demonstrated how to re-identify anonymized data that already removed all directly identifiable records.

Henriksen-Bulmer and Sheridan (2016) revealed that since 2009, several successful re-identification attacks have been conducted using multiple public datasets.

Hamidović et al. (2019) introduced general pseudonymization and anonymization procedures, which are acknowledged as effective protective measures under the GDPR. The emphasis of their research is placed on their application for safeguarding data about patient health. Lapwattanaworakul et al. (2023) discuss how legal protection of personal data, such as GDPR or PDPA, poses challenges to organizations in data sharing,

particularly the risks of re-identification attacks and consent management. The Personal Data Protection Act (PDPA) is Thailand's first data privacy law, which is largely influenced by GDPR (Intellisecure, 2023; Laohapairoj & Sombatsatapornkul, 2023).

Benarous et al. (2020) proposed a privacy-conscious pseudonym management framework using blockchain for vehicular networks (VANETs) to enhance security and privacy.

3. RSU-Data-Anonymization Architecture and Methodology

Under the GDPR, pseudonymization involves processing personal data in a way that prevents its direct association with a specific data subject without using additional information. For effective dataset pseudonymization, the "additional information" required for potential re-identification must be kept separately and safeguarded with appropriate technical and organizational measures to prevent it from being linked to any identified or identifiable individual.

Figure 2 showcases the architectural design of the system. It comprises a single database server and multiple blockchain networks. One of the blockchain networks is specifically dedicated to storing direct identifier segments.

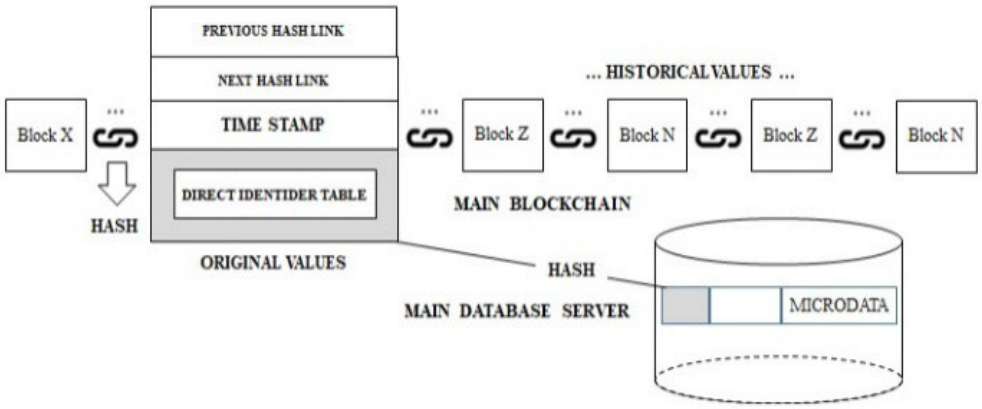


Figure 2 Main architecture diagram

In the system architecture, each node's connection to the database server utilizes a hash function. The adoption of the hash technique is driven by two key properties. Firstly, hashes act as one-way functions, meaning once data is hashed, it cannot be reversed to its original form. This ensures the security and privacy of sensitive information. Secondly, a hash function can serve to verify if any suspicious files have been altered or tampered with, ensuring data integrity. Data stored on the server is associated with the hash of each respective node. When data owners exercise their right to erasure, the data controller simply deletes the actual data from the primary server. This process of applying a hash algorithm to personal data effectively pseudonymizes the data as it indirectly protects the information. According to the Agencia Española de Protección de Datos (2019), hash values used in the blockchain should be considered personal data. As a result, when the data owner exercises their right to erasure, the hash on the blockchain loses its effectiveness and is no longer considered useful or relevant.

The hash, which previously indirectly identified the data, is now regarded as "personal data" and must be treated as such to comply with privacy regulations. This ensures that the privacy rights of individuals are respected, and their data is adequately protected by the applicable data protection laws.

A hash function is a procedure that converts any arbitrary dataset into a fixed-length character series, regardless of the input data size. The resulting output can be referred to as the hash value, code, digest, or simply hash. The term "hash" is commonly used to describe both the hash function itself and its output when applied to a specific message. The data processed through the hash function is known as the message

preimage, while the collection of all possible messages or preimages constitutes the message domain or message space. The primary goal of the hash function is to be irreversible, ensuring that applying a hash function to a direct identifier prevents its practical re-identification. However, the very "order" inherent in processing activities, which makes the hash function effective as a unique identifier, also increases the chances of potentially revealing the original message from the hash.

Before implementing hash techniques for pseudonymization or anonymization of personal data, it is crucial to conduct a thorough re-identification risk analysis specific to the chosen hash technique. This analysis should encompass the entire hash process and all other elements of the hash system, with careful consideration of any potentially linked information. The ultimate objective is to provide an unbiased and objective assessment of the long-term probability of re-identification.

4. Experiment

The well-known dataset named "Adult" is used as a dataset prototype. Income is intuitively influenced by the individual education level, age, gender, occupation, etc. The "Adult" dataset is available in the UCI machine learning repository (Lapwattanaworakul et al., 2022). To demonstrate this experiment, we combine social security numbers, names, last names, and addresses with the adult dataset. The combined dataset named "Full-Adult-dataset.csv" is used as a prototype for the demonstration. Python and scikit-learn are widely used machine-learning tools. This experiment adopts both tools for data manipulation, especially formatting the dataset into three segments: a Direct identifier segment, a Quasi-identifiers segment, and an Anonymized segment.

A. Step 1: Segmentation

In the initial step, "Full-Adult-dataset.csv" is partitioned into three distinct segments: a Direct identifier segment, a Quasi-identifiers segment, and an Anonymized segment, as shown in Figure 3.

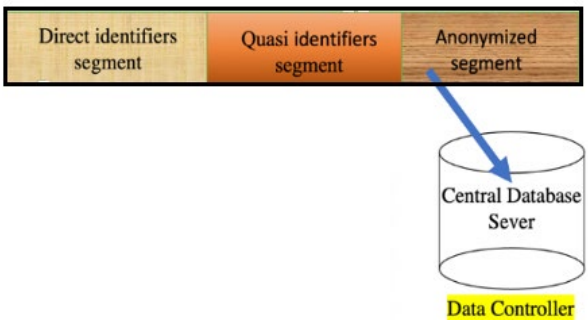


Figure 3 Data segmentation

In Figure 3, the data preparation process is depicted as being divided into three segments, each undergoing transformation.

B. Step 2: Quasi-Identifier Attributes Generalization

The purpose of this section is to demonstrate the generalization of quasi-identifier attributes. The quasi-identifier is the smallest set of attributes that has the potential to re-identify a person. The fact of re-identification makes the dataset anonymized, referred to as "k-anonymization". Some quasi-attributes such as age are generalized. A simple group-by command is used to reduce the unique identifiers to a minimal set.

The k-anonymity algorithm introduces the idea of "hiding in the crowd". Within this concept, a quasi-identifier cannot be used to re-identify an individual. Figure 4 shows a subclass distribution of attributes showing the possibility of duplication.

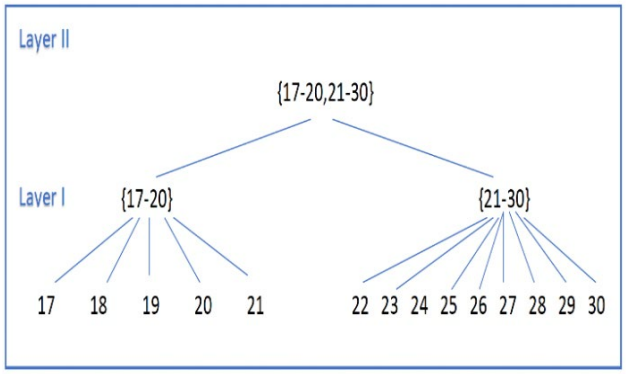


Figure 4 A part of the generalization hierarchy of the attribute "age"

In Figure 4, row 10 shows that only one young white female executive in the United States is 17 years old and has never been married. Her educational background is high school graduation, and she works for a big private company. This record can identify a specific person by combining it with another dataset that contains the full name. This process is called "re-identification risk". This set of attributes is a quasi-identifier. Therefore, the quasi-identifier, in this case, is {'age', 'sex', 'native-country', 'marital-status', 'occupation', 'race', 'work-class', 'education'}, k-anonymity guarantees that each released record will relate to at least k individuals.

A minimum set of factors includes age, sex, race, native country, education, work class, and occupation can act as quasi-identifier attributes that can be linked to a person. The problem is called "re-identification risk".

After generalizing all the quasi-identifiers, Figure 5 shows that none of the rows in the dataset remain unique. The generalization process has led to the grouping of records, and there are now at least two rows that contain the same information.

```
df=df.groupby(['age','sex','native-country','race','occupation',
               'education']).size().reset_index(name='count')
df.head(15)
```

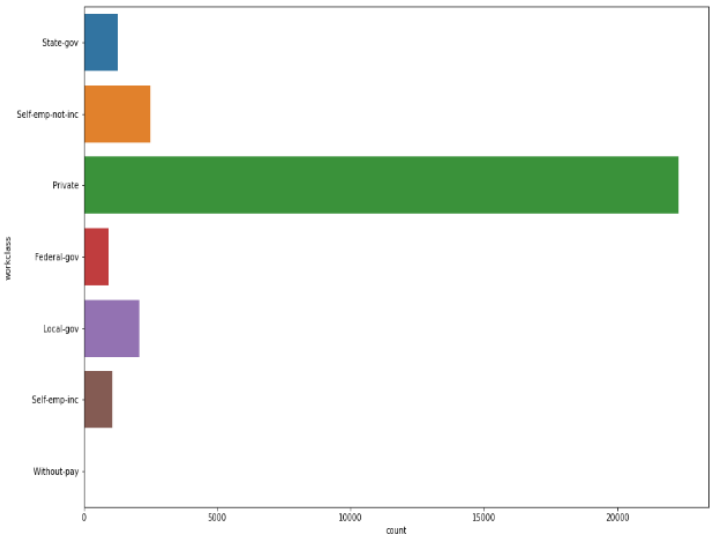
	age	sex	native-country	race	occupation	education	count
0	17-30	Female	United-States	White	Industry	HS-College	270
1	17-30	Female	United-States	White	Industry	Higher Education	208
2	17-30	Female	United-States	White	Service	HS-College	1011
3	17-30	Female	United-States	White	Service	Higher Education	236
4	17-30	Female	United-States	Not White	Industry	HS-College	54
5	17-30	Female	United-States	Not White	Industry	Higher Education	31
6	17-30	Female	United-States	Not White	Service	HS-College	229
7	17-30	Female	United-States	Not White	Service	Higher Education	34
8	17-30	Female	Non United-States	White	Industry	HS-College	32
9	17-30	Female	Non United-States	White	Industry	Higher Education	4
10	17-30	Female	Non United-States	White	Service	HS-College	88
11	17-30	Female	Non United-States	White	Service	Higher Education	12
12	17-30	Female	Non United-States	Not White	Industry	HS-College	12
13	17-30	Female	Non United-States	Not White	Industry	Higher Education	12
14	17-30	Female	Non United-States	Not White	Service	HS-College	46

Figure 5 Generalized all quasi-identifiers

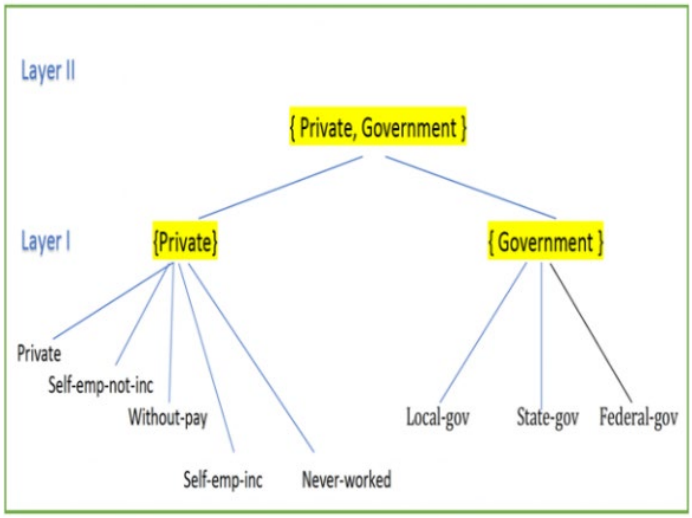
C. Step 3: Subclass Distribution of Attributes

This section aims to present the generalization hierarchy for each of the quasi-identifiers: age, sex, race, native country, education, work class, and occupation, as they collectively constitute the quasi-identifier.

In Figure 6(a), the original distribution of the 'work-class' attribute is displayed. In Figure 6(b), the generalization hierarchy of the 'work-class' attribute is presented.



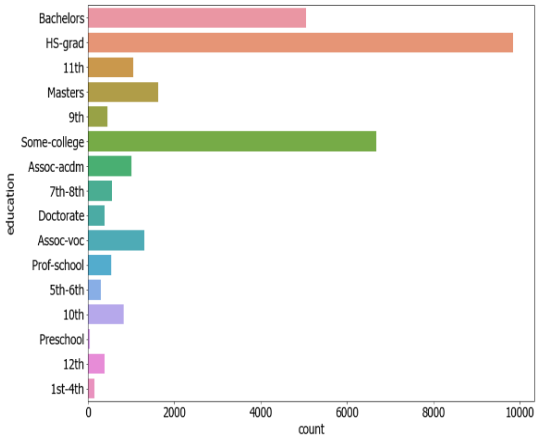
(a)



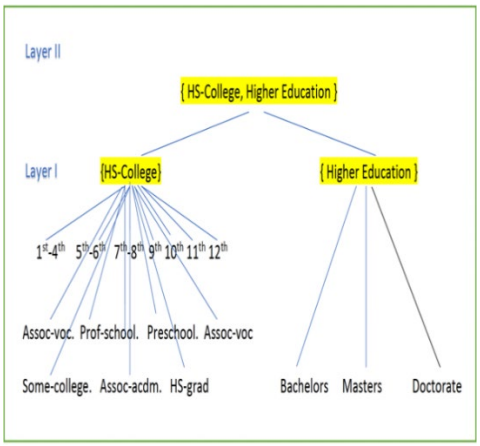
(b)

Figure 6 'work-class' distribution and generalization hierarchy to the attribute 'work-class'

In Figure 6(a), the original distribution of the 'education' attribute is depicted. In Figure 6(b), the generalization hierarchy of the 'education' attribute is demonstrated.



(a)



(b)

Figure 7 'education' distribution and generalization hierarchy to the attribute 'education'

When statisticians request microdata that has undergone anonymization methods like k-anonymity, they can download the anonymized microdata from the main database server.

D. Step 4: Show Distribution of Quasi-Attributes

Figure 8 illustrates the distribution of quasi-attributes visually after applying k-anonymity techniques. This representation shows how the data in a dataset is grouped based on similar generalized values for the quasi-attributes. Each group comprises records that share these generalized values, and the distribution illustrates the formation of these groups and the number of records in each group. The goal is to showcase the effectiveness of anonymization in protecting individual privacy while still enabling meaningful analysis of the grouped data.

```
dfn.sort_values('age')
```

	age	education-num	income	count
0	[17.0]	[7.200598802395209]	<=50k	334
120	[17.0]	[9.0]	<=50k	14
43	[17.0]	[10.0]	<=50k	5
469	[17.0]	[3.0]	<=50k	3
615	[17.0]	[4.0]	<=50k	5
...	...	...	...	...
746	[90.0]	[9.0]	>50k	4
818	[90.0]	[14.0]	<=50k	2
819	[90.0]	[14.0]	>50k	3
804	[90.0]	[6.5]	<=50k	4
750	[90.0]	[10.545454545454545]	<=50k	9

822 rows × 4 columns

Figure 8 Generalization by partitioning quasi-attributes

Figure 8 displays the result of executing the source code. Notably, all quasi-attributes have been regrouped into ranges of values.

5. Performance Evaluation

In this section, we compare performance before and after running seven algorithms including Gradient Boosting (GB), Logistic Regression (LR), Support Vector Classifier (SVC), Gaussian NB (GNB), Random Forest (RF), Decision Tree (DT), and K-Nearest Neighbors (KNN). The drawback of perturbation such as k-anonymity is loss of data utility. Figure 9 shows the percentage of data loss by comparing before and after. Pseudonymization and k-anonymity are chosen to compensate for its drawback. The drawback of pseudonymization is a high chance of a re-identification attack but no data loss.

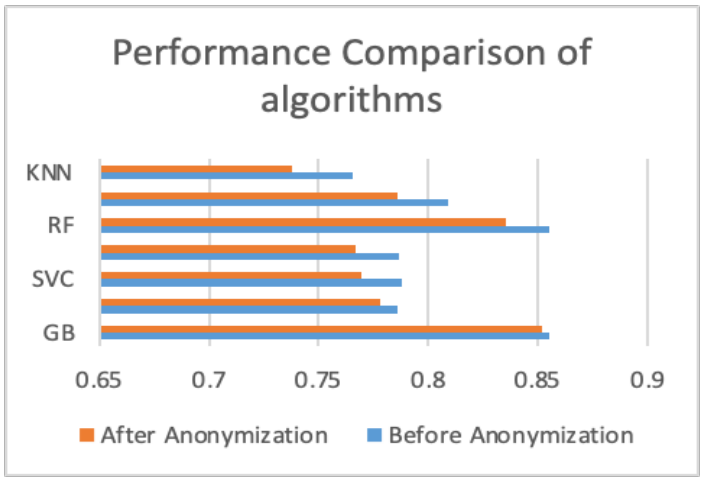


Figure 9 Performance comparison of the algorithm

According to Figure 9, Gradient Boosting demonstrates superior performance compared to other methods. While the pseudonymization technique alone does not lead to data loss, it still carries a risk of potential re-identification attacks. Therefore, this research could serve as a valuable guideline for data centers aiming to implement data privacy projects. However, it is important to note that Python is suitable for quick programmability, but it may not provide robust data security. The language itself may have certain pitfalls that could pose security risks. As an author, it is challenging to guarantee complete security solely due to language limitations and vulnerabilities. Caution and additional security measures are necessary when implementing data privacy projects using Python or any other programming language.

6. Discussion

Gradient Boosting (GB) emerged as the most effective algorithm among the seven tested. This superior performance can be attributed to the iterative nature of GB, which builds an ensemble of weak learners, where each new model attempts to correct the errors made by previous models. Unlike simpler models such as Logistic Regression (LR) or Gaussian NB (GNB), which may struggle with complex, non-linear relationships in tabular datasets as observed in this work. GB optimizes a loss function through gradient descent, allowing it to capture intricate patterns in the quasi-identifier data. Furthermore, while the Decision Tree (DT) and Random Forest (RF) models also showed strong performance, the boosting approach in GB often provides a more refined predictive boundary than the bagging approach used in RF, leading to the observed performance gains in this specific privacy-utility context. This finding is also aligned with previous comparison presented in Caruana and Niculescu-Mizil (2006), where the foundation of GB is among the most consistently high-performing algorithms across various datasets, often outperforming RF, SVC, GNB, and LR.

7. Conclusion

This research aims to implement a data privacy project in a programming style using Python, and k-anonymity was chosen because it is easily programmable. It is inevitable to incur data loss from data privacy while gaining data privacy. It is a trade-off between them. From Section 5, seven algorithms are implemented to show the percentage of data loss. The Privacy-Enhancing Technologies (PETs) Framework is a standard defined by ICO. The “RSU-Data-Anonymization” pilot demonstrates show how to comply with this standard in a programmable manner. Blockchain technology is suggested as an auxiliary system in this study. All mapping tables are stored in the blockchain. The hashing address is unique enough to serve as a primary key that can link back to microdata. Each mapping table is in a box. Each chain stores a history of each mapping table. The last box of the chain is the current version of the dataset

8. References

Aamot, H., Kohl, C. D., Richter, D., & Knaup-Gregori, P. (2013). Pseudonymization of patient identifiers for translational research. *BMC Medical Informatics and Decision Making*, *13*(1), Article 75. https://doi.org/10.1186/1472-6947-13-75

Abu Attieh, H., Müller, A., Wirth, F. N., & Prasser, F. (2025). Pseudonymization tools for medical research: A systematic review. *BMC Medical Informatics and Decision Making*, *25*(1), Article 128. https://doi.org/10.1186/s12911-025-02958-0

Agencia Española de Protección de Datos. (2019). Introducción al hash como técnica de seudonimización de datos personales. Retrieved from https://www.aepd.es/guias/estudio-hash-anonimidad.pdf

Al-Zubaidie, M., Zhang, Z., & Zhang, J. (2019). PAX: Using pseudonymization and anonymization to protect patients’ identities and data in the healthcare system. *International Journal of Environmental Research and Public Health*, *16*(9), Article 1490. https://doi.org/10.3390/ijerph16091490

Benarous, L., Kadri, B., & Bouridane, A. (2020). Blockchain-based privacy-aware pseudonym management framework for vehicular networks. *Arabian Journal for Science and Engineering*, *45*(8), 6033-6049. https://doi.org/10.1007/s13369-020-04448-z

Bourka, A., Drogkaris, P., & Agrafiotis, I. (2019). *Pseudonymization techniques and best practices: Recommendations on shaping technology according to data protection and privacy provisions*. European Union Agency for Cybersecurity (ENISA). Retrieved from https://www.enisa.europa.eu/publications/pseudonymisation-techniques-and-best-practices

Caruana, R., & Niculescu-Mizil, A. (2006). An empirical comparison of supervised learning algorithms. In *Proceedings of the 23rd International Conference on Machine Learning (ICML 2006)*. pp. 161-168, Pittsburgh, PA. https://doi.org/10.1145/1143844.1143865

El Emam, K., Jonker, E., Arbuckle, L., & Malin, B. (2011). A systematic review of re-identification attacks on health data. *PloS One*, *6*(12), Article e28071. https://doi.org/10.1371/journal.pone.0028071

Emerick, P. H. R., Sampaio, S. C., Dalmazo, B. L., Riker, A., Neto, A. V., & Immich, R. (2024). Enhancing privacy in healthcare: A multilevel approach to (pseudo) anonymization. In *Proceedings of the 2024 International Wireless Communications and Mobile Computing (IWCMC)*. pp. 1814-1819. Ayia Napa, Cyprus. https://doi.org/10.1109/IWCMC61514.2024.10592397

Hamidović, H., Kabil, J., & Šehić, E. (2019). EU General Data Protection Regulation (GDPR) – Anonymisation and pseudonimisation in function of data protection. In *Proceedings of International Scientific Conference on Digital Economy (DIEC 2019)*, pp. 219–230.

Henriksen-Bulmer, J., & Sheridan, J. (2016). Re-identification attacks – A systematic literature review. *International Journal of Information Management*, *36*(6), 1184–1192. https://doi.org/10.1016/j.ijinformgt.2016.08.002

Holmes, K. R. (2015). *What is national security?* The Heritage Foundation. Retrieved from https://www.heritage.org/military-strength-essays/2015-essays/what-national-security

Information Commissioner's Office. (2022a). *Chapter 3: Pseudonymisation, anonymisation, and privacy-enhancing technologies guidance*. Retrieved from https://ico.org.uk/media/about-the-ico/consultations/4019579/chapter-3-anonymisation-guidance.pdf

Information Commissioner's Office. (2022b). *Chapter 5: Privacy-enhancing technologies (PETs). Anonymisation, pseudonymisation, and privacy-enhancing technologies guidance*. Retrieved from https://ico.org.uk/media/about-the-ico/consultations/4021464/chapter-5-anonymisation-pets.pdf

Intellisecure Co., Ltd. (2023). *Overview of the Data Protection Act in Thailand.* Retrieved from https://www.intellisecure.co/assets/downloads/whitepapers/Thai%20PDPA%20Whitepaper%20v1.0.pdf

Laohapairoj, P., & Sombatsatapornkul, T. (2023). *Thailand data protection & cybersecurity. The Legal 500 Country Comparative Guides 2023*. Retrieved from https://www.chandlermhm.com/content/files/pdf/publications/The%20Legal%20500%20Country%20Comparative%20Guides%202023%20Thailand%20-%20Data%20Protection%20and%20Cyber%20Security.pdf

Lapwattanaworakul, J., Srisa-An, C., & Angsirikul, S. (2022). Guideline for data anonymization for data privacy in Thailand. In *The 6th International Conference on Information Technology (InCIT)*, pp. 211–215. Nonthaburi, Thailand. https://doi.org/10.1109/InCIT56086.2022.10067859

Lapwattanaworakul, J., Srisa-An, C., & Aribarg, T. (2023). Blockchain-based auxiliary systems for pseudonymization and consent management. *TEM Journal*, *12*(4), 2470–2480. https://doi.org/10.18421/TEM124-59

Lauradoux, C., Curelariu, T., & Lodie, A. (2023). *Re-identification attacks and data protection law*. AI-Regulation. Retrieved from https://ai-regulation.com/re-identification-attacks-and-data-protection-law/

Narayanan, A., & Shmatikov, V. (2008). Robust de-anonymization of large sparse datasets. *2008 IEEE Symposium on Security and Privacy*, pp. 111–125. Oakland, CA, USA.https://doi.org/10.1109/SP.2008.33

OECD. (2005). *Glossary of statistic terms: Quasi-identifier*. Retrieved from https://stats.oecd.org/glossary/detail.asp?ID=6961

Riotta, C. (2023). *Census Bureau data susceptible to 'reconstruction attacks' exposing individual data, report claims*. NEXTGOV/FCW. Retrieved from

https://www.nextgov.com/cybersecurity/2023/02/census-bureau-data-susceptible-reconstruction-attacks-exposing-individual-data/383170/

Siteimprove. (2018). *What is a data breach, exactly?*. Retrieved from https://www.siteimprove.com/glossary/gdpr-data-breaches/

Sweeney, L. (2002). k-Anonymity: A model for protecting privacy. *International Journal of Uncertainty, Fuzziness and Knowledge-Based Systems*, *10*(05), 557–570.

Tilleke & Gibbins. (2023). *Responding to a data breach in Southeast Asia*. Retrieved from https://www.tilleke.com/wp-content/uploads/2023/02/Tilleke-Responding-to-a-Data-Breach-in-Southeast-Asia.pdf