Науковий вісник НЛТУ України
Scientific Bulletin of UNFU
ISSN 1994-7836 (print)
ISSN 2519-2477 (online)
https://doi.org/10.36930/40360310
Article received: 03.04.2026
Article accepted: 26.05.2026
Article published: 29.06.2026
Correspondence author: A. M. Vakhula andrii.m.vakhula@lpnu.ua

UDC 004.056

A. M. Vakhula, O. O. Ivaniuk
Lviv Polytechnic National University, Lviv, Ukraine

IMPACT OF *K*-ANONYMITY ON MACHINE LEARNING MODEL PERFORMANCE: AN EMPIRICAL STUDY

The rapid growth of digital data and the widespread use of machine learning techniques have increased concerns regarding the protection of personal information. Data anonymization is widely used to mitigate privacy risks by transforming datasets before analysis. However, such transformations may reduce data utility and negatively affect the performance of machine learning models. This study investigates the impact of *k*-anonymity-based data anonymization on classification performance using the Adult Income dataset. In the dataset, personal data were marked as quasi-identifiers. Quasi-identifying attributes are generalized and suppressed to achieve different levels of anonymity (k = 5, 10, 20), and machine learning models are trained on both original and anonymized datasets. After training, models are evaluated using a subset of the original dataset. Performance of each *k*-anonymity level in the experiment is based on accuracy, recall, precision, and F1-score. The experimental results demonstrate that increasing the level of anonymization leads to a gradual decrease in predictive performance. Accuracy is only moderately affected: it decreases from 0.828-0.84 on the original dataset to 0.798-0.81 on strongly anonymized data. A more significant reduction is observed in recall and F1-score: from 0.586-0.626 on the original dataset to 0.435-0.555 on strongly anonymized data. Those changes indicate a loss in the ability of models to correctly identify positive instances, which harm performance and reliability of the model. At the same time, moderate anonymization levels, in the range k = 5-10, provide a reasonable balance between privacy protection and analytical utility. The findings confirm the existence of a trade-off between privacy preservation and model performance and highlight the importance of selecting appropriate anonymization parameters in machine learning workflows. Additionally, some machine learning models, like Gradient Boosting, have shown better performance on anonymized data than others, which may help to select an appropriate algorithm for a specific level of anonymized data.

Keywords: data anonymization; performance of machine learning models; data privacy; data usefulness; adult income dataset.

Introduction / Вступ

The rapid growth of digital data and the increasing adoption of machine learning techniques have significantly expanded the use of data-driven decision-making across various domains, including finance, healthcare, and public services. These applications often rely on large datasets containing demographic, socio-economic, and behavioral information. However, such datasets frequently include personal data, which raises serious concerns regarding privacy and the risk of unauthorized disclosure or re-identification of individuals.

Even when explicit identifiers such as names or identification numbers are removed, individuals may still be identified through combinations of indirect attributes, known as quasi-identifiers [5]. This challenge has led to the development of various privacy-preserving data publishing techniques aimed at reducing the risk of re-identification while maintaining the analytical value of the data.

Among these techniques, *k*-anonymity has become one of the most widely adopted approaches due to its conceptual simplicity and practical applicability [4]. The *k*-anonymity model ensures that each record in a dataset is indistinguishable from at least $k - 1$ other records with respect to quasi-identifiers. This is typically achieved through generalization and suppression of attribute values [3]. Despite its effectiveness in reducing re-identification risk, *k*-anonymity introduces information loss, which may negatively affect the performance of machine learning models.

Previous research has demonstrated that anonymization techniques can significantly influence the accuracy and reliability of data analysis and predictive modeling [5]. In particular, transformations applied to quasi-identifiers may distort underlying data distributions and reduce the ability of machine learning algorithms to capture meaningful patterns. As a result, an important research problem is to evaluate how anonymization affects model performance and to determine an appropriate balance between privacy protection and data utility.

The Adult Income dataset is widely used as a benchmark for classification tasks and contains multiple attributes that may act as quasi-identifiers. This makes it suitable for evaluating anonymization techniques in a machine learning context. In this study, *k*-anonymity-based anonymization is applied to the dataset, and its impact on classification per-

Вахула Андрій Мирославович, аспірант, кафедра комп'ютеризованих систем автоматики.
Email: andrii.m.vakhula@lpnu.ua; https://orcid.org/0009-0000-6831-5269
Іванюк Олег Олексійович, канд. техн. наук, доцент, кафедра комп'ютеризованих систем автоматики.
Email: oleh.o.ivaniuk@lpnu.ua; https://orcid.org/0000-0003-1908-1907
Цитування за ДСТУ: Вахула А. М., Іванюк О. О. Вплив *k*-анонімності на продуктивність моделей машинного навчання: емпіричне дослідження. Науковий вісник НЛТУ України. 2026, т. 36, № 3. С. 98–103.
Citation APA: Vakhula, A. M., & Ivaniuk, O. O. (2026). Impact of *k*-anonymity on machine learning model performance: an empirical study. *Scientific Bulletin of UNFU*, 36(3), 98–103. https://doi.org/10.36930/40360310

formance is systematically analyzed.

Object of research – applying *k*-anonymity-based data anonymization to structured tabular datasets used in machine learning classification tasks.

Subject of research – methods and means of installation a relationship between the level of *k*-anonymity and the predictive performance of machine learning models, including classification accuracy, precision, recall, and F1-score.

The purpose of the research – evaluate the impact of data anonymization based on the *k*-anonymity model on the performance of machine learning models, the results of which will provide a deeper understanding of how privacy-preserving transformations influence predictive accuracy of the model.

To achieve this purpose, the following main *research objectives are identified*:

- apply *k*-anonymity-based transformations to the Adult Income dataset using multiple values of the parameter $k \in \{5, 10, 20\}$, which will make it possible to evaluate the impact of different anonymization levels on classification results;
- train and evaluate Logistic Regression, Decision Tree, and Gradient Boosting classifiers on both the original and anonymized datasets, which allow to compare models' performance on original and anonymized data;
- quantify the effect of increasing anonymization levels on standard classification metrics: accuracy, precision, recall, and F1-score, which give opportunity to determine changes in classification quality as the anonymization level increases;
- compare the sensitivity of different machine learning model types to information loss introduced by anonymization, which allows to identify anonymization configurations that provide an acceptable practical balance between privacy protection and data utility.

Analysis of recent research and publications. The problem of privacy preservation in data publishing has been extensively studied in the context of preventing re-identification attacks [14] on released datasets. A key challenge arises from the presence of quasi-identifiers, which, when combined, can uniquely identify individuals even in the absence of explicit identifiers. This vulnerability was systematically demonstrated in early work on re-identification attacks, where anonymized datasets were successfully linked with external data sources to reveal personal information.

To address this issue, the *k*-anonymity model was proposed as a practical approach to ensuring privacy in tabular data [4]. Instead of focusing solely on the conceptual definition, *k*-anonymity can be analyzed from the perspective of its implementation and its effect on data transformation. In practical scenarios, achieving *k*-anonymity involves constructing equivalence classes by partitioning the dataset according to quasi-identifiers, followed by applying transformation strategies that ensure each group contains at least *k* records [3].

Different algorithmic approaches have been proposed to achieve this objective while minimizing information loss. For instance, multidimensional partitioning methods, such as the Mondrian algorithm, recursively split the dataset into regions based on attribute ranges, balancing privacy constraints and data utility [15]. Clustering-based approaches, on the other hand, group similar records together before applying generalization, which can reduce distortion compared to global recoding techniques. The choice of anonymization strategy directly affects both the structure of the transformed dataset and the degree of information loss introduced.

An important aspect of *k*-anonymity is the trade-off between privacy guarantees and data granularity [7]. As the value of *k* increases, equivalence classes become larger, which typically requires broader generalization or more aggressive suppression of attribute values. This results in reduced variability of the data and may significantly alter statistical properties, including distributions and correlations between features. To quantify the impact of anonymization, several metrics have been proposed, including information loss measures, discernibility metrics, and classification error-based evaluations. These metrics highlight that the effectiveness of *k*-anonymity should not be assessed solely in terms of privacy guarantees but also in terms of its influence on downstream analytical tasks. Therefore, while *k*-anonymity provides a practical and widely adopted framework for privacy preservation, its real-world applicability depends on carefully selecting anonymization parameters and transformation strategies that balance privacy protection with data utility.

In parallel, the concept of differential privacy has emerged as a formal privacy model that provides strong theoretical guarantees by introducing controlled randomness into data or query outputs [20]. Although differential privacy offers mathematically rigorous protection, its application in machine learning workflows may lead to reduced model performance due to the injected noise, particularly when applied directly to training data.

An important aspect of privacy-preserving data publishing is the impact [11] of anonymization on data utility. In many applications, anonymized datasets are used for downstream analytical tasks, including classification and prediction. Several studies have shown that anonymization techniques, including *k*-anonymity, introduce information loss that may degrade the performance of machine learning models. The extent of this degradation depends on factors such as the number and type of quasi-identifiers, the anonymization strategy, and the chosen privacy parameters.

From a practical perspective, this leads to a fundamental trade-off between privacy protection and analytical utility. Increasing the level of anonymization reduces the risk of re-identification but simultaneously limits the amount of information available for learning patterns in the data. While previous research has acknowledged this trade-off, there remains a need for empirical evaluation of how specific anonymization configurations affect model performance in real-world datasets. The present study addresses this problem by focusing on the impact of *k*-anonymity-based anonymization on machine learning classification tasks using the Adult Income dataset. Unlike purely theoretical analyses, this work provides an experimental evaluation of multiple anonymization levels and their influence on predictive performance. The goal is to quantify the relationship between privacy constraints and model accuracy, as well as to identify configurations that provide a balanced trade-off between privacy preservation and data utility.

Materials and methods of research. The experimental study is conducted using the Adult Income. The dataset contains demographic and socio-economic information and is commonly used for binary classification tasks, where the objective is to predict whether an individual's income exceeds a predefined threshold (50k $). The study considers multiple levels of anonymization and evaluates their effect using several classification algorithms, including Logistic Regression, Decision Tree, and Gradient Boosting.

Research results and their discussion / Результати дослідження та їх обговорення

The scope of this research is centered on the empirical evaluation of data anonymization techniques in the context of machine learning applications. The study focuses specifically on *k*-anonymity as a practical and widely used method for privacy preservation in structured tabular data. The analysis is conducted using the Adult Income dataset, which contains multiple demographic and socio-economic attributes that may function as quasi-identifiers and therefore present a potential risk of re-identification. Within this context, the research examines how the application of *k*-anonymity-based transformations affects the quality of data used for predictive modeling. Particular attention is given to the relationship between the level of anonymization and the resulting performance of machine learning models. The study considers multiple anonymization configurations in order to capture variations in privacy strength and the corresponding impact on analytical outcomes.

The primary objective of this work is to evaluate the influence of *k*-anonymity on classification performance and to analyze the trade-off between privacy preservation and data utility. This objective is addressed through a systematic experimental framework in which the dataset is transformed using different levels of anonymization and subsequently used to train and evaluate several machine learning models. The comparison between original and anonymized datasets enables the assessment of how information loss introduced by anonymization affects predictive accuracy and related evaluation metrics. The expected outcome of the study is a set of insights that support the selection of appropriate anonymization levels for machine learning workflows, ensuring an effective balance between protecting sensitive information and maintaining sufficient data utility for predictive tasks.

The Adult income dataset includes 32,561 records and consists of both numerical and categorical attributes, such as age, education, marital status, occupation, race, sex, and hours worked per week. These attributes provide a rich representation of individuals but also introduce potential privacy risks, as combinations of such attributes may enable re-identification. The target variable is a binary indicator of income level, which is used as the class label in the classification task.

To support the anonymization process, the dataset attributes are categorized based on their role in privacy preservation. Attributes such as age, work class, education, marital status, race, occupation and relationship are treated as quasi-identifiers, as they can be combined to uniquely identify individuals when external information is available. The income attribute is considered sensitive, as it represents confidential information that should not be disclosed through inference. This classification follows common practices in privacy-preserving data analysis, where distinguishing between quasi-identifiers and sensitive attributes is essential for effective anonymization [9].

Before applying anonymization and machine learning techniques, the dataset undergoes a preprocessing stage to ensure consistency and suitability for analysis. Missing or incomplete records are handled to reduce noise and prevent bias in model training. Categorical attributes are transformed into numerical representations to enable their use in machine learning algorithms, while numerical features are adjusted where necessary to maintain compatibility across models [8, 9].

These preprocessing steps ensure that both the original and anonymized datasets are comparable and suitable for subsequent experimental evaluation.

The anonymization process is based on the *k*-anonymity model and is applied to the quasi-identifiers identified in the dataset. The transformation is performed by constructing equivalence classes through the generalization of attribute values and, where necessary, the suppression of specific entries. Generalization reduces the specificity of attribute values, for example by replacing exact numerical values with intervals or grouping categorical values into broader categories. To evaluate the effect of different privacy levels, multiple anonymized datasets are generated using varying values of the parameter $k \in \{5, 10, 20\}$. Increasing the value of k leads to stronger privacy guarantees, as each record becomes indistinguishable from a larger number of other records. However, this also increases the degree of information loss, which may influence the performance of machine learning models [9].

To assess the impact of anonymization on predictive performance, three machine learning models are selected, representing different classes of learning algorithms. Logistic Regression is used as a linear baseline model that captures global relationships between features and the target variable. Decision Tree represents a rule-based model that relies on hierarchical partitioning of the feature space [10]. Gradient Boosting is employed as an ensemble method that combines multiple weak learners to achieve improved predictive accuracy [6]. The selection of these models allows for the evaluation of how different types of learning algorithms respond to the transformations introduced by anonymization.

Model performance is evaluated using standard classification metrics that capture different aspects of predictive quality. In the formulas below next abbreviations are used: *TP* – true positive, *TN* – true negative, *FP* – false positive, *FN* – false negative [13]:

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN}; \quad (1)$$

$$Precision = \frac{TP}{TP + FP}; \quad (2)$$

$$Recall = \frac{TP}{TP + FN}; \quad (3)$$

$$F1 = \frac{Precision \cdot Recall}{Precision + Recall}. \quad (4)$$

Accuracy measures the overall proportion of correctly classified instances, while precision reflects the correctness of positive predictions. Recall evaluates the ability of the model to identify positive instances, and the F1-score provides a balanced measure that combines precision and recall [10]. These metrics are shown in formulas (1)-(4). They enable a comprehensive assessment of model behavior under different anonymization levels and support the analysis of trade-offs between privacy and utility.

The experimental procedure follows a structured pipeline in which the dataset is first preprocessed and then transformed using *k*-anonymity with different parameter values. Each version of the dataset is subsequently used to train the selected machine learning models, and their performance is evaluated using the described metrics. The results are then compared across different levels of anonymization to assess the impact of privacy-preserving transformations on predictive performance.

The performance of machine learning models was evaluated on both the original dataset and its anonymized versions generated using different values of the parameter *k*. The results are presented in terms of accuracy, precision, recall, and F1-score for three classification models: Logistic Regression, Decision Tree, and Gradient Boosting. The results obtained on the original dataset serve as a baseline for evaluating the impact of anonymization. Among the evaluated models, Gradient Boosting achieved the highest performance, with an accuracy of 0.8399 and an F1-score of 0.626. Logistic Regression and Decision Tree demonstrated comparable accuracy values (approximately 0.83), but slightly lower recall and F1-score. These results indicate that ensemble methods provide improved predictive performance on the given dataset and establish a reference point for further comparison. The results of the experiment and machine learning metrics are listed in Table.

Table. ML models' performance with anonymized datasets / Продуктивність моделей машинного навчання на анонімізованих наборах даних

k	*model*	*accuracy*	*precision*	*recall*	F1
0	Logistic Regression	0.832	0.696	0.539	0.607
0	Decision Tree	0.828	0.695	0.506	0.586
0	Gradient Boosting	0.840	0.715	0.557	0.626
5	Logistic Regression	0.826	0.677	0.529	0.594
5	Decision Tree	0.821	0.714	0.428	0.535
5	Gradient Boosting	0.821	0.717	0.421	0.531
10	Logistic Regression	0.812	0.639	0.506	0.565
10	Decision Tree	0.801	0.691	0.317	0.435
10	Gradient Boosting	0.809	0.676	0.402	0.504
20	Logistic Regression	0.814	0.654	0.482	0.555
20	Decision Tree	0.798	0.665	0.323	0.435
20	Gradient Boosting	0.808	0.713	0.339	0.459

The application of *k*-anonymity resulted in a gradual decrease in model performance as the value of *k* increased. For $k = 5$, the reduction in accuracy is relatively small. For example, Logistic Regression achieved an accuracy of 0.8257, while Gradient Boosting achieved 0.8207, indicating only a minor degradation compared to the original dataset.

However, for higher values of *k*, the decrease in performance becomes more pronounced. At $k = 10$, the accuracy drops to 0.8122 for Logistic Regression and 0.8096 for Gradient Boosting. Similar results are observed for $k = 20$, where accuracy values remain slightly above 0.8 but are consistently lower than those obtained on the original dataset. These results (Figure 1) demonstrate that increasing the level of anonymization leads to a gradual loss of predictive accuracy due to information loss introduced by generalization and suppression.

Fig. 1. Accuracy of ML models with anonymized datasets / Точність моделей машинного навчання на анонімізованих наборах даних

A more significant effect of anonymization is observed in recall values. For instance, the recall of the Decision Tree model decreases from 0.5064 in the original dataset to 0.3169 for $k = 10$ and remains at a similar level for $k = 20$. Gradient Boosting also exhibits a noticeable reduction in recall, decreasing from 0.5568 in the original dataset to 0.3386 at $k = 20$

This indicates that anonymization significantly affects the model's ability to correctly identify positive instances. As recall decreases, the model becomes less effective in detecting cases belonging to the target class.

The F1-score follows a similar trend, decreasing with higher values of *k*, which reflects the combined effect of reduced precision and recall. Results are shown in Figure 2.

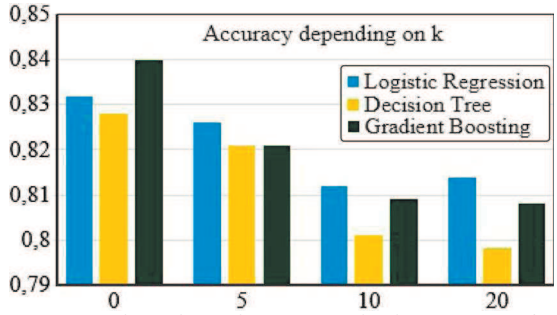

Fig. 2. F1-score dependency from anonymization parameter / Залежність F1-міри від параметра анонімізації

The experimental results reveal that different models exhibit varying sensitivity to anonymization. Decision Tree demonstrates the highest sensitivity, with a substantial decrease in recall and F1-score across all anonymization levels. This suggests that tree-based models are particularly affected by the loss of detailed attribute values caused by generalization. Logistic Regression shows relatively stable performance across different values of *k*, indicating greater robustness to anonymization. Gradient Boosting achieves the best overall performance and demonstrates higher resilience compared to the other models, although it is still affected by increased anonymization levels.

The results clearly demonstrate the existence of a trade-off between privacy preservation and data utility. Increasing the value of *k* enhances privacy by reducing the risk of re-identification but leads to a decrease in model performance. It is important to note that the degradation in accuracy is moderate, while recall is significantly more affected.

The experimental findings indicate that moderate values of *k*, such as $k = 5$, provide a reasonable balance between privacy protection and predictive performance. In contrast, higher values of *k* introduce substantial information loss, which negatively affects model effectiveness.

Discussion of research results. In the work of Abbasi, Mori, and Saracino [1], the trade-off between privacy and utility in the analysis of tabular data using machine learning is examined. The authors emphasize that increasing the level of protection almost inevitably affects the analytical properties of the data and the quality of models. The results of the study conducted within the present paper further specify this conclusion for *k*-anonymity, as the demonstrate that the decrease in accuracy remains relatively moderate, while recall and F1-score decline much more noticeably.

In the study by Yuan, Yu, Yang, and Chen [19], the Unitanony framework is proposed for more fine-grained anonymization aimed at reducing information loss. The authors show that improving the utility of anonymized data constitutes a separate scientific problem, and the quality of the outcome depends not only on the value of *k* but also on the method used to generalize quasi-identifiers.

In the work of Verdonck, De Boeck, Willocx, Lapon, and Naessens [17], a hybrid anonymization pipeline is proposed to improve the balance between privacy and data utility in ML scenarios. The authors emphasize that the practical value of an anonymized dataset should be assessed not only by formally satisfying privacy requirements but also by its suitability for subsequent machine learning tasks. This aligns with the logic of the conducted study, in which the quality of anonymization was evaluated not abstractly, but through changes in performance metrics across several classification models.

In the study by Arbelaez and Climent [2], a local search approach for *k*-anonymity with minimized information loss is proposed. The authors demonstrate that even within the same privacy model, it is possible to significantly reduce data loss compared to existing heuristics. The conducted research shows that the observed degradation of metrics with increasing *k* does not negate the feasibility of *k*-anonymity, but rather indicates the need to select more effective anonymization algorithms and generalization configurations that enable a better balance between privacy and classification quality.

In the article by Manocchio, Layeghy, Gwynne, and Portmann [12], a configurable anonymization approach is proposed that is explicitly oriented toward balancing privacy levels and data utility for machine learning. The authors demonstrate that anonymization should be treated as a tunable tool rather than a single-step procedure with a fixed outcome.

Thus, the discussion of the research results confirms the relevance of conducting this study. The analyzed contemporary works are primarily focused either on improving anonymization algorithms or on the general trade-off between privacy and data utility, whereas the presented study provides a quantitative evaluation of this trade-off within a unified experimental framework for three widely used machine learning models and multiple levels of *k*-anonymity. This made it possible not only to confirm the general trend of decreasing model performance with increasing *k*, but also to demonstrate the differential sensitivity of Logistic Regression, Decision Tree, and Gradient Boosting to anonymization, which constitutes a distinct contribution compared to existing results.

So, based on the results of the work performed, it is possible to formulate the following scientific novelty and practical significance of the research results.

Scientific novelty of the research results – for the first time, a method for evaluating the differential sensitivity of Logistic Regression, Decision Tree, and Gradient Boosting classifiers to *k*-anonymity-based data transformations on the Adult Income dataset has been developed by constructing model-specific performance degradation curves for $k \in \{5, 10, 20\}$.

Practical significance of the research results – can be applied when selecting anonymization parameters for preparing datasets intended for machine learning tasks in privacy-sensitive environments such as healthcare, finance, and public services, enabling practitioners to choose a *k* value that satisfies regulatory requirements while maintaining acceptable predictive model performance.

Conclusions / Висновок

An assessment of the impact of data anonymisation has been carried out based on the *k*-anonymity model on the performance of machine learning models, the results of which will provide a deeper understanding of how privacy-preserving transformations influence predictive accuracy of the model. Based on the results of the research the following main conclusions can be drawn.

1. Investigated the impact of *k*-anonymity-based data anonymization on the performance of machine learning models using the Adult Income dataset, which confirmed that increasing the anonymization parameter leads to information loss affecting classification performance metrics.
2. Established that as the value of the parameter *k* increases, a gradual decrease in predictive accuracy is observed, while a more significant reduction occurs in recall and F1-score, indicating a stronger influence of anonymization on the ability of models to correctly identify positive instances.
3. Analyzed the sensitivity of different machine learning models to anonymization, which showed that Decision Tree demonstrates the highest sensitivity, Logistic Regression exhibits more stable behavior, and Gradient Boosting provides higher robustness compared to other models.
4. Identified the trade-off between privacy protection and analytical utility, where increasing the level of anonymization enhances privacy by reducing the risk of re-identification but leads to reduced predictive performance.
5. Determined that moderate levels of anonymization, particularly $k = 5$, provide a reasonable balance between privacy preservation and model effectiveness for practical machine learning applications.
6. Established the practical applicability of the obtained results for supporting decision-making when preparing datasets for machine learning tasks in privacy-sensitive environments, ensuring compliance with data protection requirements while maintaining acceptable model performance.
7. Future research may focus on extending the analysis to other anonymization techniques, such as differential privacy, as well as evaluating the impact of anonymization on different types of datasets and machine learning models. Additionally, further investigation into methods that minimize information loss while maintaining strong privacy guarantees remains an important direction for future work.

References

1. Abbasi, W., Mori, P., & Saracino, A. (2024). Further Insights: Balancing Privacy, Explainability, and Utility in Machine Learning-based Tabular Data Analysis. *Proceedings of the 19th International Conference on Availability, Reliability and Security* (*ARES 2024*), pp. 1–10. https://doi.org/10.1145/3664476.3670901
2. Arbelaez, A., & Climent, L. (2025). Iterative local search for preserving data privacy. *Applied Intelligence*, article number 189. https://doi.org/10.1007/s10489-024-05909-w
3. Carvalho, T. M., Moniz, N., Faria, P., & Antunes, L. (2023). Survey on Privacy-Preserving Techniques for Microdata Publication. *ACM Computing Surveys*, 55(14s), 1–37. https://doi.org/10.1145/3588765
4. De Capitani di Vimercati, S., & Samarati, P. (2021). *k*-Anonymity. In: Jajodia, S., Samarati, P., Yung, M. (Eds). *Encyclopedia of Cryptography, Security and Privacy*. Springer, pp. 1–5. https://doi.org/10.1007/978-3-642-27739-9_754-2
5. De Capitani di Vimercati, S., Foresti, S., Livraga, G., & Samarati, P. (2023). *k*-Anonymity: From Theory to Applications. *Transactions on Data Privacy, 16*(1), 25–49. URL: https://www.tdp.cat/issues21/tdp.a460a22.pdf
6. Friedman, J. H. (2001). Greedy function approximation: A gradient boosting machine. *Annals of Statistics, 29*(*5*), 1189–1232. https://doi.org/10.1214/aos/1013203451
7. Gadotti, A., Rocher, L., Houssiau, F., Creţu, A.-M., & de Montjoye, Y.-A. (2024). Anonymization: The imperfect science of using data while preserving privacy. *Science Advances, 10*(*29*), article ID eadn7053. https://doi.org/10.1126/sciadv.adn7053
8. Géron, A. (2019). Hands-on machine learning with scikit-learn, keras, and tensorflow (2end edition). *OReilly Media*, pp. 19-30.

URL: https://www.oreilly.com/library/view/hands-on-machine-learning/9781492032632/
9. Han, J., Pei, J., & Tong, H. (2022). *Data Mining: Concepts and Techniques* (4th ed.). Elsevier, pp. 23–26. https://doi.org/10.1016/C2013-0-18660-6
10. James, G., Witten, D., Hastie, T., Tibshirani, R., & Taylor, J. (2023). *An Introduction to Statistical Learning: with Applications in Python*. Springer, pp. 15–155. https://doi.org/10.1007/978-3-031-38747-0
11. Karagiannis, S., Ntantogian, C., Magkos, E., Tsohou, A., & Ribeiro, L. L. (2024). Mastering data privacy: leveraging *k*-anonymity for robust health data sharing. *International Journal of Information Security, 23*, 2189–2201. https://doi.org/10.1007/s10207-024-00838-8
12. Manocchio, L. D., Layeghy, S., Gwynne, D., & Portmann, M. (2024). A configurable anonymisation approach for network flow data: Balancing utility and privacy. *Computers & Security / Computers and Electrical Engineering*, vol. 118. https://doi.org/10.1016/j.compeleceng.2024.109465
13. Rainio, O., Teuho, J., & Klén, R. (2024). Evaluation metrics and statistical tests for machine learning. *Scientific Reports 14,* article ID 6086. https://doi.org/10.1038/s41598-024-56706-x
14. Sweeney, L. (2000). Simple demographics often identify people uniquely. *Carnegie Mellon University, Data Privacy Working Paper,* vol 3. URL: https://dataprivacylab.org/projects/identifiability/
15. Torra, V., & Navarro-Arribas, G. (2023). Attribute disclosure risk for *k*-anonymity: the case of numerical data. *International Journal of Information Security, 22*, 2015–2024. https://doi.org/10.1007/s10207-023-00730-x
16. Torskyi, O. I., & Hrytsiuk, Y. I. (2025). Application of machine learning to enhance the efficiency of automated software testing. *Scientific Bulletin of UNFU*, 35(4), 142-149. https://doi.org/10.36930/40350416
17. Verdonck, J., De Boeck, K., Willocx, M., Lapon, J., & Naessens, V. (2023). A hybrid anonymization pipeline to improve the privacy-utility balance in sensitive datasets for ML purposes. *Proceedings of the 18th International Conference on Availability, Reliability and Security* (*ARES 2023*), pp. 1–11. https://doi.org/10.1145/3600160.3600168
18. Veseliak, V. V., & Hrytsiuk, Y. I. (2024). Machine learning methods in epidemiological research. *Scientific Bulletin of UNFU*, 34(4), 59–67. https://doi.org/10.36930/40340408
19. Yuan, S., Yu, J., Yang, T., & Chen, C. (2025). Unitanony: a fine-grained and practical anonymization framework for better data utility. *Cybersecurity, 8,* article number *43*. https://doi.org/10.1186/s42400-024-00345-2
20. Zhao, Y., & Chen, J. (2022). A Survey on Differential Privacy for Unstructured Data Content. *ACM Computing Surveys, 54*(10s), article ID 207, 1–28. https://doi.org/10.1145/3490237

А. М. Вахула, О. О. Іванюк

Національний університет "Львівська політехніка", м. Львів, Україна

ВПЛИВ *K*-АНОНІМНОСТІ НА ПРОДУКТИВНІСТЬ МОДЕЛЕЙ МАШИННОГО НАВЧАННЯ: ЕМПІРИЧНЕ ДОСЛІДЖЕННЯ

Швидке зростання цифрових даних і широке використання методів машинного навчання моделей посилили занепокоєння щодо захисту персональної інформації. Анонімізація даних широко використовується для зменшення ризиків їх конфіденційності шляхом перетворення наборів даних перед аналізом. Однак, такі перетворення можуть зменшити корисність даних і негативно вплинути на продуктивність моделей машинного навчання. Проаналізовано вплив анонімізації даних на основі *k*-анонімності на продуктивність процесу класифікації даних з використанням набору даних Adult Income. У цьому наборі персональні дані було позначено як квазіідентифікатори, атрибути яких узагальнюються та пригнічуються для досягнення різних рівнів анонімності $k = \{5, 10, 20\}$, а моделі машинного навчання навчаються як на оригінальних, так і на анонімізованих наборах даних. Після навчання моделі оцінюються з використанням підмножини оригінального набору даних. Продуктивність кожного рівня *k*-анонімності в експерименті базується на accuracy, recall, precision та F1-score. Експериментальні результати демонструють, що збільшення рівня анонімізації призводить до поступового зниження прогностичної продуктивності моделей машинного навчання. Метрика точності прогнозування Accuracy зазнає тільки помірного впливу: воно знижується від 0,828-0,84 на оригінальному наборі даних і до 0,798-0,81 – на сильно анонімізованих даних. Істотніші зміни спостерігаються для метрик recall та F1-score: з 0,586-0,626 на оригінальному наборі даних до 0,435-0,555 – на сильно анонімізованих даних. Ці зміни вказують на втрату здатності моделей правильно ідентифікувати позитивні приклади, що погіршує продуктивність і надійність моделі. Водночас, помірні рівні анонімізації, у діапазоні $k = 5\text{-}10$, забезпечують розумний баланс між конфіденційністю даних і аналітичною їх корисністю. Отримані результати підтверджують існування компромісу між збереженням конфіденційності даних та продуктивністю моделей і вказують на важливість вибору відповідних параметрів їх анонімізації у процесах машинного навчання. З'ясовано, деякі моделі машинного навчання, такі як Gradient Boosting, продемонстрували кращу продуктивність на анонімізованих даних порівняно з іншими, що може допомогти обрати відповідний алгоритм для конкретного рівня анонімізації даних.

Ключові слова: анонімізація даних; продуктивність моделей машинного навчання; збереження конфіденційності даних; корисність даних; набір даних про дохід дорослого населення.