Prifysgol Cymru
Y Drindod Dewi Sant
University of Wales
Trinity Saint David

A Comprehensive Evaluation Approach for Privacy-Preserving Data Publishing Techniques Using Composite Scoring and Decision Support

Student Name: Md. Ariful Islam

Student ID: 2431309

Programme: MSc Cybersecurity and Digital Forensics(Computing)

Supervisor: Dr. Anum Paracha

8th June 2026

Declaration

I, **Md. Ariful Islam**, thus certify that the dissertation entitled: Title: Composite Score and Decision Support for Privacy Preserving Data Publishing My own work and has been developed to the academic standard required for the MSc in Cybersecurity and Digital Forensics (Computing) at the University of Wales Trinity Saint David. The sources used in the preparation of this work have all been acknowledged and cited properly, following the referencing guidelines of the university. This dissertation has not previously been submitted, in whole or in part, for any other academic qualification at this or any other institution. I hereby declare that the research presented here was carried out under the supervision of Dr. Anum Paracha and that all experimental work and analysis were carried out by me as per the university's regulations of academic integrity and the specified word count constraints.

Name: Md. Ariful Islam
Ariful Islam
June 2026

Form of Consent

I, **Md. Ariful Islam**, hereby grant permission for the dissertation entitled: "A Comprehensive Evaluation Approach for Privacy-Preserving Data Publishing Using Composite Scoring and Decision Support" submitted in partial fulfilment of the requirements for the **MSc Cybersecurity and Digital Forensics (Computing)** degree, to be made available for consultation, inter-library loan, and photocopying, in accordance with current copyright legislation and university regulations. I further consent that the title and abstract of this research may be made accessible to external organisations, academic researchers, and educational institutions for the purposes of scholarly inquiry and knowledge dissemination.

Ariful Islam
June 2026

Acknowledgements

First of all I am thankful to the Almighty Allah who made this employment possible by his innumerable mercies. I would also like to sincerely thank my supervisor, **Dr. Anum Paracha**, Lecturer (Computing, London) at the Institute of Inner City Learning (IICL), University of Wales Trinity Saint David, for her expert guidance, invaluable feedback and consistent encouragement throughout each stage of this research. Her dedication to academic integrity and scholarly achievement has been an ever-present inspiration. I am also grateful to the academic and professional support team of the University of Wales Trinity Saint David for the intellectual environment and institutional resources to do this job. I also thank the developers and maintainers of the UCI Machine Learning Repository for providing the Adult Income dataset that serves as the empirical basis for the experimental evaluation in this work. Finally, I would like to extend my deepest gratitude to my family and friends for their patience, encouragement and unwavering support, which have been a constant source of strength during my academic pursuits.

Md. Ariful Islam

Abstract

The proliferation of data-driven systems across healthcare, finance, government, and artificial intelligence domains has intensified the demand for responsible data sharing whilst concurrently safeguarding sensitive personal information. Although a considerable body of research has produced diverse privacy-preserving data publishing (PPDP) techniques, selecting the most appropriate method for any given deployment context remains a complex challenge, largely because privacy protection, data utility, and computational efficiency frequently pull against one another. This dissertation addresses that challenge by proposing a comprehensive multi-dimensional evaluation framework that compares representative PPDP techniques through a composite scoring and decision-support mechanism. Three widely adopted technique categories were implemented and evaluated: k-anonymity, differential privacy, and synthetic data generation, applied to the Adult Income dataset (UCI Machine Learning Repository, 30,162 records). The proposed framework encompasses data preprocessing, feature engineering, privacy-preserving transformation, machine learning utility assessment, privacy analysis, computational performance measurement, composite score calculation, and decision-support ranking. To select the most appropriate downstream classifier, Logistic Regression, Decision Tree, Random Forest, and XGBoost were compared; XGBoost demonstrated the strongest predictive performance, achieving an accuracy of 86.44% and an F1-score of 70.49%, and was accordingly adopted as the evaluation model. Experiments conducted across three train-test split configurations (80:20, 75:25, and 70:30) consistently confirmed that the 80:20 partitioning produced the most favourable outcomes and was therefore designated the final experimental configuration. The composite scoring model integrates privacy, utility, and computational performance through the formula $S = 0.4P + 0.4U + 0.2C$.The final rankings suggest that synthetic data generation achieved the highest composite score (0.7677), followed by k-anonymity (0.6770), differential privacy (0.6157), and the unmodified original dataset (0.4650). Although differential privacy delivered the most robust formal privacy guarantees, synthetic data generation provided the most effective overall balance across all three evaluation dimensions. These findings underscore the insufficiency of privacy-focused optimisation in isolation and demonstrate the practical value of multi-criteria decision-support frameworks for PPDP technique selection in real-world machine learning applications.

Keywords: Privacy-Preserving Data Publishing, k-Anonymity, Differential Privacy, Synthetic Data Generation, Composite Scoring, Decision Support, Machine Learning Utility, XGBoost, Adult Income Dataset.

Contents

1 Introduction **12**

1.1 Background 12
1.2 Problem Statement 14
1.3 Aim and Objectives 15
1.4 Significance of the Research 15
1.5 Scope of the Research 16
1.6 Research Contributions 17
1.7 Dissertation Structure 18
1.8 Summary 18

2 Literature Review **20**

2.1 Introduction 20
2.2 Privacy-Preserving Data Publishing Techniques 21
2.3 k-Anonymity, l-Diversity, and t-Closeness 22
2.4 Differential Privacy 24
2.5 Synthetic Data Generation 25
2.6 Privacy Attacks and Disclosure Risks 26
2.7 Data Utility Evaluation 26
2.8 Computational Performance Evaluation 27
2.9 Existing Evaluation approachs 27
2.10 Research Gap 28
2.11 Summary 29

3 Methodology **31**

3.1 Introduction 31
3.2 Research Design 32
3.3 Dataset Description 34
3.4 Data Preprocessing and Feature Engineering 35
3.4.1 Missing Value Removal 36
3.4.2 Feature Scaling 36
3.4.3 Feature Engineering 36

3.4.4 Train-Test Splitting 37
3.5 Privacy-Preserving Techniques 37
3.5.1 k-Anonymity Through Generalisation 38
3.5.2 Differential Privacy Through Laplace Noise Addition 38
3.5.3 Synthetic Data Generation Through Bootstrap Sampling 39
3.6 Machine Learning Utility Evaluation 39
3.6.1 Train-Test Split Sensitivity 40
3.7 Privacy Evaluation 41
3.8 Computational Performance Evaluation 41
3.9 Composite Scoring Model 41
3.10 Decision-Support Ranking 42
3.11 Ethical Considerations 43
3.12 Summary 43

4 Implementation 44
4.1 Introduction 44
4.2 System Architecture 45
4.3 System Architecture 45
4.4 Implementation Environment 47
4.5 Dataset Loading and Preprocessing Module 48
4.5.1 Missing Value Handling 48
4.5.2 Categorical Encoding 49
4.5.3 Feature Scaling 49
4.6 Feature Engineering Module 49
4.6.1 Capital Net Feature 49
4.6.2 Hours-Age Ratio 50
4.6.3 Education-Hours Feature 50
4.6.4 Grouped Features 50
4.7 k-Anonymity Implementation 50
4.8 Differential Privacy Implementation 51
4.9 Synthetic Data Generation Implementation 52
4.10 Machine Learning Evaluation Module 52
4.11 Utility Evaluation Module 53
4.11.1 Accuracy 53
4.11.2 Precision 53
4.11.3 Recall 53
4.11.4 F1-Score 53
4.12 Privacy Evaluation Module 53
4.13 Computational Performance Module 54

4.14 Composite Scoring Implementation 54
4.15 Decision-Support Ranking Module 55
4.16 Summary 55

5 Experimental Results and Evaluation 56
5.1 Introduction 56
5.2 Experimental Configuration 56
5.3 Machine Learning Model Comparison 57
5.4 Train-Test Split Evaluation 57
5.5 Baseline Performance on Original Data 58
5.6 Utility Evaluation Results 58
5.7 Privacy Evaluation Results 59
5.8 Computational Performance Evaluation 59
5.9 Composite Scoring Results 60
5.10 Critical Discussion 61

6 Discussion 62
6.1 Introduction 62
6.2 Discussion of Model Selection Results 62
6.3 Analysis of Train-Test Split Results 62
6.4 Results Discussion for k-Anonymity 63
6.5 Interpretation of Composite Scoring Results 63
6.6 Answering the Research Questions 63
6.7 Practical Implications 64
6.8 Research Limitations 64
6.9 Future Work 64
6.10 Summary 65

7 Conclusion and Future Work 66
7.1 Introduction 66
7.2 Research Summary 66
7.3 Summary of Experimental Findings 67
7.4 Research Contributions 67
7.4.1 Comprehensive Evaluation Approach 68
7.4.2 Composite Scoring and Decision Support 68
7.4.3 Comparative Evaluation Across Multiple Privacy Techniques . . . 69
7.4.4 Practical Experimental Pipeline 69
7.5 Comparison with Existing Literature 69
7.6 Practical Implications 70
7.7 Research Limitations 70

7.8 Future Work 70
7.8.1 Advanced Synthetic Data Generation 71
7.8.2 Empirical Privacy Attack Evaluation 71
7.8.3 Additional Datasets and Domains 71
7.8.4 Adaptive Weighting Mechanisms 72
7.8.5 Scalability and Explainability 72
7.9 Final Conclusion 72

8 Reflections **73**
8.1 Introduction 73
8.2 Reflections on the research process 73
8.3 Skills Developed 74
8.4 Challenges 74
8.5 Working and Academic Development 75
8.6 Lessons Learned: 75
8.7 Summary 76

List of Figures

3.1 Schematic overview of the proposed privacy-preserving evaluation pipeline 34

4.1 System architecture of the comprehensive evaluation framework 47

5.1 Comparison of composite scores across evaluated techniques 60

5.2 Comparative visualisation of privacy, utility, and computational performance scores 61

List of Tables

2.1 Comparative analysis of existing privacy-preserving evaluation approaches 28
2.2 Research gap analysis 29

3.1 Adult Dataset Attributes 35
3.2 Assigned privacy scores by technique 41

4.1 Summary of implementation environment 48
4.2 Implemented privacy scores by technique 53

5.1 Experimental configuration parameters 57
5.2 Machine Learning Model Comparison Results 57
5.3 Composite Score Comparison Across Train-Test Split Configurations . . 58
5.4 Baseline XGBoost classification performance on the original dataset . . . 58
5.5 Final Utility Evaluation Results Using XGBoost and 80:20 Split 58
5.6 Privacy Evaluation Results 59
5.7 Computational Performance Results for 80:20 Split 59
5.8 Final Composite Scoring Results for 80:20 Split 60

6.1 Final Decision-Support Ranking 63

7.1 Final Composite Ranking Results 67

Chapter 1

Introduction

1.1 Background

We live in a data-driven culture increasingly Organisations across sectors, including healthcare, financial services, public administration, and artificial intelligence research, routinely collect, store, and process large amounts of personal and organisational data to support decision-making, service delivery, and the development of predictive systems [1]. The convergence of artificial intelligence, cloud computing and the Internet of Things has dramatically increased the volume and analytical complexity of data, opening up new avenues for evidence-based innovation, while at the same time raising serious concerns about individual privacy and data governance [2].

The social and economic value of the large scale data analysis is indisputable. Patient records are used by healthcare providers to improve diagnostic accuracy and clinical outcomes . Transaction histories are reviewed by financial organizations to detect fraud and manage credit riskPublic resources and policy development are guided, in part, by population data used by government agencies. In each case, machine learning models trained on rich, high-dimensional datasets have become indispensable instruments for generating actionable insight and automating complex processes. Nevertheless, the utility of data is inextricably bound to the sensitivity of the information it encodes. Datasets drawn from real-world populations frequently contain personal attributes such as age, gender, educational attainment, occupation, income, health status, and geographic origin that, individually or in combination, carry significant re-identification risk. The publication or sharing of such data without appropriate safeguards may expose individuals to a range of privacy harms, including identity disclosure, attribute inference, linkage attacks, and membership inference attacks.

These risks are not merely theoretical. Sweeney [3] demonstrated compellingly that a substantial proportion of individuals within publicly released, ostensibly anonymised datasets could be uniquely re-identified through the intersection of quasi-identifying at-

tributes such as age, gender, and postal code. More recently, Narayanan and Shmatikov [4] amplified this problem by proving that even sparse, high-dimensional datasets might be deanonymised through correlation with publicly available auxiliary information. These contributions, taken together, point to a fundamental contradiction between the social drive to share data for analytical reasons and the ethical imperative to respect the privacy of persons whose information forms the basis of that data. In response to these challenges, the field of privacy-preserving data publishing (PPDP) has developed a diverse portfolio of techniques aimed at reducing disclosure risk whilst preserving analytical utility. The earliest and most widely studied of these is k-anonymity, proposed by Sweeney [3]. which requires each record in a released dataset to be indistinguishable from at least k 1 other records with respect to quasi-identifying attributes, typically achieved through suppression and generalisation. Subsequent refinements, including l-diversity [5]. and t-closeness [6] k-anonymity protects individuals by ensuring that each record is indistinguishable from at least $k - 1$ other records based on quasi-identifiers such as age, gender, and occupation. This is typically achieved through suppression and generalisation of attributes.

Although k-anonymity reduces direct re-identification risks, later studies identified several weaknesses. To avoid homogeneity attacks, machanavajjhala et al. proposed l-diversity [5] and to reduce the risk of attribute disclosure, t-closeness was proposed by Li et al. [7]. These findings illustrate that privacy-preserving data publishing is a developing research domain, with specific weaknesses of the k-anonymity model such as homogeneity assaults and skewed distributions of sensitive attributes being addressed. Dwork [8] [9] introduced a very different and more mathematically rigorous approach. using differential privacy, which provides formal probabilistic guarantees that the inclusion or absence of any individual's record would not appreciably change the output of a computation. Rather than transforming the data directly, differential privacy achieves its guarantees by injecting carefully calibrated random noise into analytical outputs, with the magnitude of the noise governed by a privacy budget parameter . This formal underpinning has made differential privacy highly prominent in both academic research and industrial practice; nonetheless, it poses a well-recognised contradiction between the strength of privacy protection and the preservation of data utility [10].A viable answer to this problem is the privacy-preserving data dissemination using synthetic data synthesis recently emerged. Synthetic data techniques generate artificial records that are statistically representative of the original population instead of offering a modified version of the original dataset that does not directly expose any real individual's information [11]. To this end, Generative Adversarial Networks, Variational Autoencoders and various probabilistic modelling frameworks have been used [12] [13] [14] and dedicated evaluation tools such as Synthcity [10] and SynthEval [15] have been developed to evaluate the quality of synthetic datasets across utility and privacy dimensions.However, synthetic data is not

privacy secure by default since generative models that overfit on their training data can mistakenly repeat sensitive records leading to additional membership inference vulnerabilities [16]. Although the PPDP literature is abundant, the practical problem of picking the most effective technique to be used in a certain application scenario is still a relevant concern. Existing evaluation frameworks tend to see privacy, utility, and computing efficiency as separate, independent concerns, rather than as parts of an integrated, interpretable decision-support system. This fragmentation makes it impossible for practitioners to undertake meaningful cross-technique comparisons, or to find the way that gives the best overall trade-offs for their particular deployment needs. It is this gap that the present research seeks to address.

1.2 Problem Statement

The increasing adoption of privacy-preserving data publishing techniques has created a pressing need for evaluation frameworks that are at once comprehensive, interpretable, and practically actionable. Although a substantial body of research has produced diverse PPDP methods each with distinct trade-off profiles across privacy protection, data utility, and computational cost, the tools available to practitioners for comparing and selecting among these methods remain limited. Several specific problems characterise the current state of the field. First, existing evaluation approaches are typically domain-specific or technique-specific, focusing predominantly on a single category of method (most commonly synthetic data generation) whilst providing little basis for comparison across the broader landscape of anonymisation, differential privacy, and generative techniques. This narrow focus makes existing frameworks ill-suited to circumstances in which a practitioner has to select between fundamentally diverse methods. Second, the majority of the evaluation approaches treat privacy, usefulness and performance in isolation, providing many independent measures without an integrating mechanism to provide an overall comparison score. In practice, this means decision makers in the complex and frequently subjective task of balancing competing criteria by hand, a process that is neither transparent nor reproducible. Third, computational performance, which is very important in large scale deployment settings where efficiency of processing directly affects scalability, operational cost and deployment feasibility, is severely under-represented in the existing review work. A technique that provides strong privacy guarantees but imposes prohibitive computational overhead may be practically unsuitable even for applications with stringent privacy requirements. Fourth, many current evaluation approaches lack an explicit decision-support capability — that is, they provide metrics without offering practitioners actionable guidance as to which technique best satisfies the aggregate requirements of their use case.

This research addresses these limitations by proposing a structured multi-dimensional

evaluation approach that compares k-anonymity, differential privacy, and synthetic data generation against a common set of criteria encompassing privacy protection, machine learning utility, and computational efficiency, and that synthesises these assessments into a composite score capable of supporting evidence-based technique selection.

1.3 Aim and Objectives

The aim of this research is to develop a comprehensive evaluation approach for privacy-preserving data publishing that integrates privacy protection, machine learning utility, and computational performance using composite scoring and decision-support mechanisms.

To achieve this aim, the following objectives are defined:

1. To implement and evaluate representative privacy-preserving data publishing techniques including k-anonymity, differential privacy, and synthetic data generation.

2. To assess the effectiveness of privacy-preserving techniques based on privacy protection, machine learning utility, and computational performance.

3. To analyse the trade-offs between privacy preservation and data utility within machine learning-based data publishing systems.

4. To develop a composite scoring and decision-support approach for comparative ranking and selection of the most balanced privacy-preserving technique.

How can composite scoring and decision-support mechanisms improve comparative evaluation and ranking of privacy-preserving techniques?

The research questions steer the investigation, by analyzing trade-offs between various privacy-preserving strategies with respect to privacy protection, machine learning value, and computational efficiency. The questions also help construct a composite grading and decision-support strategy for determining the most effective and practicable privacy-preserving technique.

1.4 Significance of the Research

The present research is of great theoretical and practical significance. From an academic perspective, the paper contributes to the burgeoning subject of multi-dimensional PPDP evaluation by offering an integrated composite scoring methodology that considers privacy, utility and computing efficiency not as distinct aims but as simultaneously optimised criteria. This approach contrasts with the prevailing tendency in the literature to

evaluate these dimensions in isolation and provides a more holistic and reproducible basis for technique comparison. From a practical standpoint, the research addresses a genuine need within organisations that are subject to data protection legislation such as the General Data Protection Regulation (GDPR) and the California Consumer Privacy Act (CCPA), which require the adoption of technically sound privacy-preserving safeguards whilst preserving the analytical value of shared data. The proposed evaluation framework and decision-support mechanism can directly support data governance practitioners, machine learning engineers, and privacy compliance officers in making principled, evidence-based decisions about PPDP technique selection. Additionally, the study provides a demonstration of the importance of computational performance as an evaluation criterion, an aspect frequently neglected in existing research yet essential for deployment-ready privacy-preserving systems.

1.5 Scope of the Research

This research focuses on the evaluation of privacy-preserving data publishing techniques using privacy protection, machine learning utility, and computational performance as the primary assessment dimensions.

The experimental implementation is based on the Adult Income dataset obtained from the UCI Machine Learning Repository [17]. The dataset was selected because it is widely used in privacy-preserving data publishing and machine learning research, allowing meaningful comparison with previous studies.

Three representative privacy-preserving techniques are implemented and evaluated:

- k-anonymity using generalisation and binning techniques;
- differential privacy through Laplace noise addition;
- synthetic data generation using bootstrap-based sampling.

To evaluate machine learning utility, four classification algorithms are initially assessed:

- Logistic Regression;
- Decision Tree;
- Random Forest;
- XGBoost.

The comparative analysis shows that the XGBoost model is the best model in terms of classification accuracy and F1-score. Hence, XGBoost is chosen as the main model for the utility evaluation throughout the rest of the study.

Although the proposed framework provides a complete evaluation technique, the research is limited by some constraints. First, the synthetic data generation method use bootstrap sampling rather than advanced deep learning methods like Conditional Generative Adversarial Networks (CTGANs) or Variational Autoencoders (VAEs). Second, privacy evaluation is based on theoretical privacy scores and established privacy rankings, and not on direct adversarial testing via membership inference or reconstruction assaults. Third, the experiments are only conducted on one benchmark dataset, which may limit the generalisability of the results to other domains.

Nevertheless, despite these constraints, the study provides a practical proof-of-concept implementation indicating the effectiveness of composite scoring and decision-support mechanisms for privacy-preserving data publishing evaluation.

1.6 Research Contributions

This paper contributes to the field of privacy-preserving data publishing in five major ways:

First, it offers a comprehensive assessment framework that can evaluate different kinds of privacy-preserving procedures in a unified experimental setting. Unlike most existing research that focus on only one evaluation dimension, the proposed paradigm unifies privacy protection, utility preservation and computational performance.

Second, A weighted composite scoring model (S=0.4P+0.4U+0.2C) that combines privacy protection, machine learning utility, and computational efficiency into one interpretable statistic.

Third, the research creates a decision support mechanism that evaluates the alternative privacy-preserving strategies in terms of their overall effectiveness. The ranking method helps practitioners to find solutions with a proper combination of privacy protection, analytical utility and implementation efficiency .

Fourth, the work empirically compares k-anonymity, differential privacy and synthetic data generation on the same dataset, pre-processing processes, machine learning model and evaluation criteria. This consistency enables a fair comparison of techniques and avoids methodological bias.

Lastly, the study suggests that privacy assurances alone may not be enough to guide practical decisions when considering privacy-preserving techniques. Instead, an appropriate selection of techniques should be based on a simultaneous evaluation of privacy, utility and computing performance. The suggested approach thus offers a more balanced and application-oriented view on privacy-preserving data publishing research.

1.7 Dissertation Structure

This dissertation is composed of seven chapters.

Chapter 1 describes the research backdrop, problem statement, research aims, research questions, significance of the study, scope and contributions of the study. It sets the motivation for establishing a holistic evaluation methodology for privacy-preserving data dissemination.

Chapter 2 presents a comprehensive overview of the literature on privacy-preserving data publishing, anonymisation strategies, differential privacy, synthetic data generation, privacy assaults, utility evaluation, computational performance and existing evaluation frameworks. This chapter also shows research gap filled by this study.

A research approach as described in Chapter 3 is applied in this examination. It details the research strategy, dataset selection, pre-processing steps, privacy-preserving transformations, machine learning assessment methods, and composite scoring methodology.

Chapter 4 describes the implementation of the suggested evaluation mechanism. This chapter describes the experimental setup, software tools, feature engineering pipeline, privacy-preserving implementations, and evaluation pipeline.

Chapter 5 presents the experimental results and analyzes the implemented strategies with respect to the privacy, utility, computational performance and composite score measures. We provide comparative analyses and visualisations to aid in the interpretation of the results.

In Chapter 6, the suggested framework is evaluated and the results are discussed in the context of the literature. The chapter also discusses the limitations and consequences for future research and practical applications. In chapter 7 the dissertation is concluded by outlining the key findings, highlighting the research contributions, acknowledging constraints and recommending choices for additional work.

1.8 Summary

In this chapter, we introduced the increasing necessity of privacy-preserving data publishing in the modern data-driven environments. The conversation underlined the growing necessity to reconcile the protection of privacy with the utility for analysis and efficiency of computation as businesses continue to depend on big scale data collecting and machine learning systems.

In the chapter, the limits of existing evaluation methodologies were reviewed and the necessity for a comprehensive framework that can integrate privacy protection, machine learning utility and computing performance into a single assessment methodology was found. The study also introduced the research topic, objectives, research questions, significance, breadth and contributions of the study.

The next chapter provides a comprehensive analysis of the literature on privacy-preserving data publishing, covering anonymisation approaches, differential privacy, synthetic data generation, privacy assaults, utility evaluation and evaluation frameworks. The review supplies the theoretical basis for the proposed research and reveals the knowledge gap addressed by this study.

Chapter 2

Literature Review

2.1 Introduction

Due to the rapid rise of data driven systems and the increasing concern of protecting sensitive information, privacy-preserving data publishing has become an essential study subject. Large-scale datasets are a critical resource for modern businesses to support analytics, machine learning, artificial intelligence and automated decision-making systems. However, many datasets contain sensitive personal information, posing major privacy hazards when data is shared, examined or published without sufficient protections.

The rise of machine intelligence has further amplified such privacy issues. High prediction accuracy from machine learning models generally relies on huge datasets, however the direct usage of sensitive datasets might expose individuals to dangers such re-identification, membership inference, linkage attacks, and attribute inference. These concerns are especially pertinent in sectors like healthcare, banking, education and public administration, where data may comprise demographic, financial or medical information.

The goal of privacy-preserving data publishing approaches is to minimize the danger of disclosure while preserving the data's analytical utility. Common ways are anonymization techniques (e.g., k-anonymity, l-diversity, t-closeness), formal privacy mechanisms (e.g., differential privacy), and synthetic data generation (i.e., generation of artificial data sets that resemble the original data).

Such strategies have been widely investigated but evaluation is a challenging problem. Each solution has different trade-offs between privacy protection, data utility and computational performance. For example, one technique may give great privacy protection, but decrease the accuracy of machine learning, while another may maintain utility, but provide lesser privacy guarantees. So evaluation should not be based on a single measure.

In this chapter we critically evaluate the literature on privacy-preserving data publishing, anonymisation, differential privacy, synthetic data generation, privacy assaults, utility evaluation, computational performance, and existing evaluation approaches. This chapter

also brings forth the research gap which this study is trying to fill.

2.2 Privacy-Preserving Data Publishing Techniques

The increasing usage of big data collection and analysis has made privacy-preserving data publication (PPDP) an essential research topic in data science, cybersecurity and machine learning. Modern businesses are continuously obtaining and processing vast amounts of data from individuals, businesses and public bodies to fuel data-driven decision-making, predictive modelling and artificial intelligence applications. High quality data is essential for innovation and improved performance of an organization. However, it also poses significant issues of privacy, confidentiality and data protection. Privacy-preserving data publishing is a technique to allow data sharing and analysis, while minimizing the possibility of releasing sensitive information about individuals.

The main objective of PPDP is to achieve a suitable balance between privacy protection and data value. In practical terms, companies aim to preserve sensitive data without sacrificing too much of its analytical value. This balance is especially important in machine learning applications where the quality and granularity of the base dataset is often the key to the performance of the model. Too much protection of privacy may diminish the useful value of the data. Too little protection of privacy may cause the risk of re-identification, attribute disclosure and membership inference attack to the persons [18, 11].

Historically, data publishers have removed direct identifiers such as names, addresses, telephone numbers or national identity numbers prior to releasing datasets. However, the pioneering work by Sweeney [3] showed that humans can still be re-identified using combinations of quasi-identifiers, such as age, gender and postal code. This research demonstrated that simple de-identification techniques were unable to provide privacy in many real-world contexts. Other studies have since shown that datasets that have been anonymized may often be linked to publicly available information, which allows attackers to recover sensitive personal information [19].

As a consequence, privacy-preserving data publishing has gone beyond standard de-identification methods. Current approaches include anonymization, suppression, generalization, perturbation, noise injection, differential privacy and synthetic data generation. Each of these systems has different privacy guarantees, computational cost and influence on data utility. Anonymisation-based solutions try to reduce the risk of exposure by transforming identifiable traits, whereas differential privacy adds mathematically calibrated noise to constrain the effect of individual data on the analytical results. Recently, the generation of synthetic data has attracted much attention, as artificial datasets may be synthesized without directly exposing original records [15, 16].

The increasing significance of PPDP is significantly associated with the growth of ar-

tificial intelligence and machine learning techniques. Many of today's AI systems require access to massive and diverse information to train and validate themselves. However, sharing sensitive datasets between corporations, research institutes and government agencies could raise major privacy issues. Therefore, privacy-preserving solutions have grown more important in domains such as healthcare, finance, cybersecurity, education and public administration where datasets typically contain sensitive or individually identifiable information [10].

Interest in privacy-preserving technologies has also been prompted by legal and regulatory challenges. Regulations like the General Data Protection Regulation (GDPR) and the California Consumer Privacy Act (CCPA) demand strict restrictions for the collection, storage, processing and sharing of personal information. Therefore, it is crucial for firms to establish technical and organizational measures that are able to show compliance and operational efficiency. Privacy-preserving data publishing is a crucial component of effective data governance and regulatory compliance efforts [55].

The sector has made considerable progress, but the privacy-utility trade-off is still a key problem. In general, the more severe the modification of the original data, the stronger the privacy protection, but at the cost of statistical accuracy and machine learning performance. However, with more usefulness, the risk of leaking privacy will be larger. As such, recent efforts have been made to develop holistic evaluation frameworks in which privacy preservation, utility retention, and computational efficiency are considered together (i.e., not in isolation) [20, 21].

Moreover, the increasing privacy issues expose the shortcomings of many existing techniques. The advent of machine learning and generative artificial intelligence has ushered in a new era of ever more powerful attacks that may deduce sensitive information from the disclosed datasets and trained models. Even synthetic datasets and privacy-preserving machine learning techniques are vulnerable to attacks such as membership inference, reconstruction, and extraction, if not properly evaluated [22, 23]. These improvements call for holistic and multi-faceted evaluation methodologies that may evaluate privacy assurances and practical performance. Privacy-preserving data publishing is evolving with an increased demand for secure data sharing, regulatory compliance and reliable artificial intelligence. The increasing adoption of data-driven technology in enterprises makes the design and assessment of successful PPDP techniques essential to reconcile tradeoffs between innovation, privacy protection and analytical value.

2.3 k-Anonymity, l-Diversity, and t-Closeness

k-anonymity is one of the earliest and most used anonymization approaches. Sweeney proposed it as a paradigm to preserve privacy in the released datasets [3]. The notion of k-anonymity is that each record should be indistinguishable from at least $k - 1$ other

records in terms of quasi-identifiers.

A dataset is k-anonymous if every equivalence class has at least k records:

$$|EC_i| \geq k$$

where EC_i denotes an equivalence class.

k-anonymity is commonly accomplished by generalization and suppression We have categories instead of real values . Generalization . For example, a precise age of 29 can be replaced by an age range of 20-30. Suppression removes values that enhance the re-identification risk totally.

k-anonymity is intuitive and well-known, but has significant shortcomings. The homogeneity attack is one important weakness. If the sensitive attribute values for all the entries in an equivalence class are the same, an attacker can deduce sensitive information even when the individual identities are masked.

Machanavajjhala et al. presented l-diversity [6] to overcome this problem. l-diversity demands the existence of at least l unique values for the sensitive property in each equivalence class:

$$\text{Div}(EC_i) \geq l$$

However, l-diversity has limitations. It might fail if the values of sensitive attributes are semantically identical and/or if the distribution of values overall is significantly skewed. To solve this constraint, Li et al. introduced t-closeness [22]. t-closeness demands that the distribution of sensitive attributes within each equivalence class is near to that of the dataset.

However, even with these advances, data loss may still occur due to the anonymization procedures. Enhanced privacy protection usually necessitates stronger generalisation, which in turn can cause the loss of data granularity and adversely influence machine learning performance. This drawback is crucial in this investigation as the k-anonymity method used in this study uses generalisation and binning which enhances privacy but may lower recall and F1-score.

Recent works tried to improve anonymisation by integrating it with additional privacy methods. Majeed and Hwang introduced a method for privacy-preserving data publishing that merges differential privacy and k-anonymity to minimize the loss of statistical information [9]. Kara et al. also researched hybrid anonymisation with (r, k, ϵ)-anonymisation for combining several privacy principles [23]. The investigations indicate how classical anonymisation is still applicable, but often needs augmenting to address new privacy threats.

2.4 Differential Privacy

Differential privacy is a stringent privacy concept established by Dwork [24]. It provides mathematical assurances that the result of a computation will be mostly unaffected by the inclusion or exclusion of an individual record. We say that a mechanism M is differentially private if:

$$Pr[M(D_1) \in S] \leq e^{\epsilon} Pr[M(D_2) \in S]$$

If D_1 and D_2 are neighboring datasets that differ by one record, and ϵ is the privacy budget.

The value of ϵ controls the level of privacy protection. The smaller the ϵ, the more privacy we get, but that means that we need to add more noise and that may reduce the utility of the data. A bigger value of ϵ retains more benefit at the cost of less privacy protection.

Differential privacy is typically achieved by noise addition procedures like the Laplace mechanism:

$$M(x) = f(x) + Laplace\left(\frac{\Delta f}{\epsilon}\right)$$

where $\Delta f is the sensitivity of the query function.$

Differential privacy has various advantages over regular anonymisation. It gives guarantees that are formal and still hold even when an adversary has access to external information. This provides higher resistance to linking attacks than k-anonymity.

However, differential privacy also poses issues. The main difficulty is the privacy-utility tradeoff. Stronger privacy protection needs larger noise, which may skew the data distribution and degrade the accuracy of machine learning. Kairouz et al. discussed a number of unresolved problems in privacy-preserving machine learning such as privacy budget selection and utility degradation [10]. Pan et al. conducted a survey on differential privacy in deep learning, highlighting the difficulties of maintaining model performance under stringent privacy restrictions [14].

In recent guidelines, NIST emphasizes that differential privacy needs careful interpretation of privacy promises, utility impact, and implementation context [25]. This is important: privacy guarantees alone do not establish if a method is viable for real-world implementation.

The differential privacy in this work is realized by adding Laplace noise. This method can provide high privacy scores. However, the experimental results reveal that the computational performance and utility are worse than synthetic data generation and k-anonymity. This demonstrates that the approach with the most privacy protection is not always the best choice overall, when utility and efficiency are also taken into account.

2.5 Synthetic Data Generation

More attention has been paid to the generation of synthetic data as a way to protect privacy. Synthetic data techniques release artificial records which preserve the statistical structure of the underlying dataset , instead of original records .

Synthetic data offers many advantages. First, it reduces direct access to primary sources. Secondly it can preserve data patterns useful for machine learning. Third, it enables data sharing with limited access to the original sensitive data.

Early methods for synthetic data generation were based on statistical sampling and probabilistic modeling. Current methods are based on machine learning generative models, such as Generative Adversarial Networks, Variational Autoencoders, and copula based models [13].

Generative Adversarial Networks proposed by Goodfellow et al. have a big impact on synthetic data generation [12]. Xu et al. proposed CTGAN, a conditional GAN framework designed for the generation of tabular data [13]. Patki et al. proposed the Synthetic Data Vault for generating and evaluating synthetic tabular data [26]. Qian et al. recently introduced a synthetic data generation approach for several modalities, called [15].

The evaluation of synthetic data has become a major academic topic. SynthEval, proposed by Lautrup et al. [16] enables thorough assessment of utility and privacy of tabular synthetic data. SynthEval is particularly relevant to this research as it highlights the importance of multi-dimensional assessment of synthetic data.

But synthetic data isn't a privacy guarantee in itself. If a generative model memorizes the training data, the synthetic data set could leak critical information. Shokri et al. have demonstrated membership inference attacks against machine learning models [27], and generative models could be vulnerable to similar assaults. Giomi et al. provided a methodology to estimate privacy risk in synthetic data and emphasized the need for empirical privacy assessment [21].

In this study, bootstrap sampling is used to produce synthetic data. This is a basic synthetic data approach compared to advanced models such as CTGAN or DP-CTGAN. Bootstrap sampling can preserve utility by sampling from the original distribution, but it does not have formal privacy guarantees and may have copies of original records. Therefore, results should be interpreted with caution.

While this restriction is there, the experiment demonstrates that synthetic data achieved the highest composite score since it preserved its utility and was computationally resistant. This confirms the claim that fair assessment is necessary.

2.6 Privacy Attacks and Disclosure Risks

Understanding the limitations of privacy-preserving techniques in the face of privacy threats is crucial. A method can be secure, however, disclosed data or model outputs can allow an attacker to deduce sensitive information.

Linkage attacks are those where an attacker uses external information to identify persons in a disclosed dataset. This form of attack is exemplified by Sweeney's re-identification study [3]. Narayanan and Shmatikov also demonstrated successful de-anonymisation attacks on sparse datasets [19].

Membership inference attacks try to recover if a certain individual record was used in the training dataset of a model. Shokri et al. demonstrated that machine learning models are capable of leaking membership information through prediction behaviour [27].

Attribute inference attacks aim to infer hidden sensitive attributes by exploiting available information. Reconstruction attacks are attempts to reconstruct sensitive records from publicly released statistics results.

Recent surveys have demonstrated that privacy threats against machine learning systems are still a substantial danger. Rigaki and Garcia surveyed several privacy threats and demonstrated how contemporary machine learning systems might leak sensitive information [28].

These assaults show that privacy evaluation cannot rely on utility measurement alone. A dataset that works well for machine learning may nonetheless carry privacy hazards. Thus, a holistic approach is needed to address privacy protection, utility and performance in a unified way.

2.7 Data Utility Evaluation

Data utility is the utility of altered or generated data for analysis. In privacy-preserving data publishing, utility is vital, because the privacy transformation should not undermine the analytical usefulness of the data.

Utility can be measured either by statistical similarity or by downstream machine learning performance. Statistical similarity measures the distributions, correlations and descriptive statistics of the original data and the modified data. The machine learning utility examines how good is a model trained on modified data in a prediction challenge.

In this work, the utility is measured by machine learning classification. Comparison of several classifiers including Logistic Regression, Decision Tree, Random Forest and XGBoost. For final utility evaluation, XGBoost is used because it has achieved the strongest prediction performance on the Adult dataset.

Common utility metrics include accuracy, precision, recall, and F1-score:

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN}$$

$$Precision = \frac{TP}{TP + FP}$$

$$Recall = \frac{TP}{TP + FN}$$

$$F1 = 2 \times \frac{Precision \times Recall}{Precision + Recall}$$

F1-score is particularly important when precision and recall must be balanced. In this study, F1-score is used as the utility component in the composite scoring model because it provides a balanced evaluation of classification performance.

Existing approachs such as SynthEval also use utility metrics to assess synthetic data quality [16]. However, utility alone is insufficient because high utility may come at the cost of weak privacy protection.

2.8 Computational Performance Evaluation

However, a key, and often underestimated, aspect of privacy-preserving evaluation is its computational efficiency. In actual systems, privacy-preserving approaches should be practical, scalable and efficient.

The performance can be quantified in terms of execution time, training time, memory utilization, scalability and energy consumption. In this work, the computational performance is measured in terms of model training time. That's improved computational efficiency due to shorter training time.

This dimension is crucial, because a solution with good privacy and high utility can nevertheless be unfeasible if it is computationally expensive. For example, differential privacy may incur overhead from noise creation and data distortion. Generating synthetic data with complex generative models can also be computationally intensive.

Recent studies have progressively stressed the necessity of efficient privacy-preserving technologies. Sanchez-Serrano et al. stressed the need of making practical decisions when selecting methods for privacy-preserving synthetic data [20]. Hence the inclusion of computing performance in the proposed evaluation technique is justified.

2.9 Existing Evaluation approachs

Privacy preservation techniques have been evaluated by a number of approaches proposed recently. These approaches make valuable contributions but have limits.

One of the most relevant current approachs is SynthEval. It gives a thorough assessment of synthetic tabular data with usefulness and privacy metrics [16]. However, SynthEval focuses mostly on synthetic data, and does not provide a clear overall score for different types of privacy-preserving approaches.

Synthcity is a more broad framework for generating and evaluating synthetic data for many data modalities [15]. It supports several synthetic data models. But it is mostly for synthetic data.

Giomi et al. [21] provided an approach for evaluating the privacy risk of synthetic data. The paper is valuable in that it deals with empirical privacy risk. But it does not directly affect computational performance or composite decision support ranking.

Sanchez-Serrano et al. put forward a decision strategy for privacy-preserving synthetic data generation [20]. Their work is very significant as it underlines the need of decision support. However, it concentrates solely on synthetic data production and not on comparison of several privacy-preserving categories.

Current approaches generally give relevant evaluation methods but frequently need a comprehensive composite scoring methodology that bundles privacy, utility and performance across various technique areas.

Approach	**PrimaryFocus**	**Utility Evaluation**	**Privacy Evaluation**	**Decision Support**
SynthEval [16]	Synthetic Data	Yes	Yes	No
Synthcity [15]	Synthetic Data Generation	Yes	Partial	No
Anonymeter [21]	Privacy Risk Assessment	No	Yes	No
Decision Framework [20]	Synthetic Data Selection	Yes	Yes	Yes
Proposed Approach	Multiple Privacy-Preserving Techniques	Yes	Yes	Yes

Table 2.1: Comparative analysis of existing privacy-preserving evaluation approaches

2.10 Research Gap

The literature review indicates some research gaps.

Although there have been significant research efforts in privacy-preserving data pub-

lishing, there are some fundamental shortcomings in existing evaluation approachs. Early privacy-preserving methods such as k-anonymity, l-diversity and t-closeness were mainly aimed at privacy protection by anonymization and generalization techniques without in-depth analysis of machine learning utility and computational efficiency [3, 6, 22]. More recent differential privacy research has introduced mathematically rigorous privacy guarantees, but many studies focus mainly on privacy strength and offer limited multidimensional evaluation of utility preservation and deployment efficiency [29, 10].Similarly, recent synthetic data evaluation methods such as Synthcity and SynthEval provide in-depth analyses of synthetic data utility and privacy risks, but their primary focus is on generating synthetic data rather than carrying out a full comparison of different types of privacy-preserving techniques [15, 16].Often, studies also present numerous independent measures without giving an interpretable ranking or decision-support mechanism to help stakeholders select the most balanced privacy-preserving strategy [20]. Also, the computational performance is generally given less priority yet it is quite important in large-scale real-world deployment situations. Therefore, there is a considerable research gap to develop a holistic and decision-oriented evaluation approach that can integrate privacy protection, machine learning utility, and computing performance into a holistic comparison approach. This reserch closes this gap by presenting a composite scoring and decision-support approach to evaluate and rank privacy-preserving data publishing techniques.

Study/Framework	**Privacy**	**Utility**	**Performance**	**Ranking**
k-Anonymity Studies	✓	×	×	×
Differential Privacy Studies	✓	Partial	×	×
SynthEval	✓	✓	×	×
Synthcity	✓	✓	×	×
Proposed Approach	✓	✓	✓	✓

Table 2.2: Research gap analysis

2.11 Summary

This chapter reviews the literature on privacy preserving data publishing, anonymisation, differential privacy, synthetic data generation, privacy assaults, utility evaluation, computational performance and existing evaluation approaches .

The review found that k-anonymity offers easy and efficient anonymisation but may suffer from information loss and vulnerability to assaults. Differential privacy offers strong theoretical guarantees, but at the cost of reduced utility and increased computational cost. Utility is achievable with synthetic data synthesis but privacy is not guaranteed

automatically.

Existing approaches such as SynthEval, Synthcity, and Anonymeter are providing great contributions but there are constraints for thorough comparison, composite scoring, and decision support across several privacy-preserving technique categories.

Chapter 3

Methodology

3.1 Introduction

In this chapter we describe the research methodology employed to design, execute, and assess the suggested comprehensive solution for privacy-preserving data posting. The approach describes the whole experimental procedure, including dataset pretreatment, feature engineering, application of privacy-preserving algorithms, machine learning evaluation, privacy evaluation, computing performance evaluation, and composite score.

The main purpose of the methodology is to design a systematic and reproducible evaluation framework, which can be used to compare privacy-preserving strategies across three main dimensions:

- Privacy preservation;
- Machine learning utility;
- Computational performance.

The methodology combines these dimensions into a single decision-support system that can score privacy-preserving techniques according to their overall effectiveness. Three privacy-preserving approaches are implemented and empirically evaluated:

- k-anonymity via generalization and quantile binning;
- differential privacy via adding Laplace noise;
- synthetic data production using bootstrap sampling.

The methodology also considers feature engineering, machine learning model comparison, utility evaluation via XGBoost, privacy scoring, evaluation of computing performance, and weighted composite scoring. We evaluate our methods experimentally on the Adult dataset from UCI Machine Learning Repository [17].

This chapter is structured as follows: The general research design is discussed in Section 3.2. The dataset utilized in the experiment is described in Section 3.3. Preprocessing and feature engineering strategies are described in Section 3.4. The implemented privacy-preserving strategies are described in Section 3.5. The machine learning assessment process is detailed in Section 3.6. Privacy and computational performance evaluation is described in Section 3.7. Section 3.8 introduces composite score and decision support model.

3.2 Research Design

In this work, a quantitative experimental research design was used to examine the effectiveness of several privacy-preserving data publishing approaches. The research focuses on measurable outcomes and comparison using numerical evaluation indicators, hence a quantitative method was judged appropriate. In research in data science and machine learning, experimental procedures are often used when different strategies need to be compared under controlled settings, and evaluated using consistent metrics of performance [30, 31].

The main purpose of this study is to examine the performance of three representative privacy-preserving strategies, i.e., k-anonymity, differential privacy and synthetic data generation, from numerous evaluation dimensions. Such characteristics include privacy protection, utility of machine learning, and computing efficiency. An empirical evaluation procedure was used to give a systematic and reproducible comparison because each technique trades off privacy and analytical utility differently.

The study design is driven by a systematic evaluation pipeline that allows the evaluation of all procedures in the same experimental context. The same pre-processing techniques, machine learning setups, and evaluation criteria are maintained to minimize experimental bias and to allow performance variations to be attributed to the privacy-preserving methods rather than implementation changes. It is believed that such controlled testing is crucial for credible and comparable outcomes in machine learning studies [32].

The total evaluation system comprises eleven successive steps:

1. Dataset collection and preparation;
2. Data preprocessing;
3. Feature engineering;
4. Privacy-preserving transformation;
5. Machine learning model development;
6. Utility assessment;

7. Privacy assessment;
8. Computational performance evaluation;
9. Score normalisation;
10. Composite score calculation;
11. Decision-support ranking.

Different from many prior works that mainly focus on privacy protection or prediction performance, the methodology used in this work integrates many evaluation dimensions in a cohesive framework. This multidimensional view is crucial since privacy-preserving approaches rarely do well on all criteria at the same time. For example, an approach that provides high privacy guarantees may result in information loss that affects the performance of machine learning, while a methodology that preserves predictive accuracy may provide weaker privacy protection. Therefore the evaluation of approaches based on a single statistic may not be the whole picture of their performance.

To tackle this difficulty, the suggested methodology embeds privacy, usefulness and computing performance into an integrated evaluation model. This offers a more fair assessment and facilitates evidence-based decision making when choosing privacy-preserving strategies for real-world deployment circumstances. Moreover, the interpretability is enhanced through the proposed decision-support ranking mechanism, which transforms many performance indicators into a single comparison evaluation procedure.

The entire workflow of the research design and the flow of data through each stage of the assessment framework is presented in Figure 3.1.

The evaluation procedure starts with the acquisition and pre-processing of the experimental dataset. Preprocessing and feature engineering are followed by privacy-preserving transformations executed independently utilizing k-anonymity, differential privacy and synthetic data creation techniques. The generated datasets are evaluated based on machine learning performance metrics, privacy evaluation standards and computational efficiency measurements. Finally, all evaluation outputs are normalized and aggregated in a weighted composite scoring model to produce an overall rating of the examined methodologies.

By integrating multiple performance dimensions into a single evaluation framework, the research design provides a comprehensive and reproducible methodology for assessing privacy-preserving data publishing approaches. This design not only facilitates objective comparison but also contributes to the development of practical decision-support mechanisms for privacy-aware data sharing and machine learning applications.

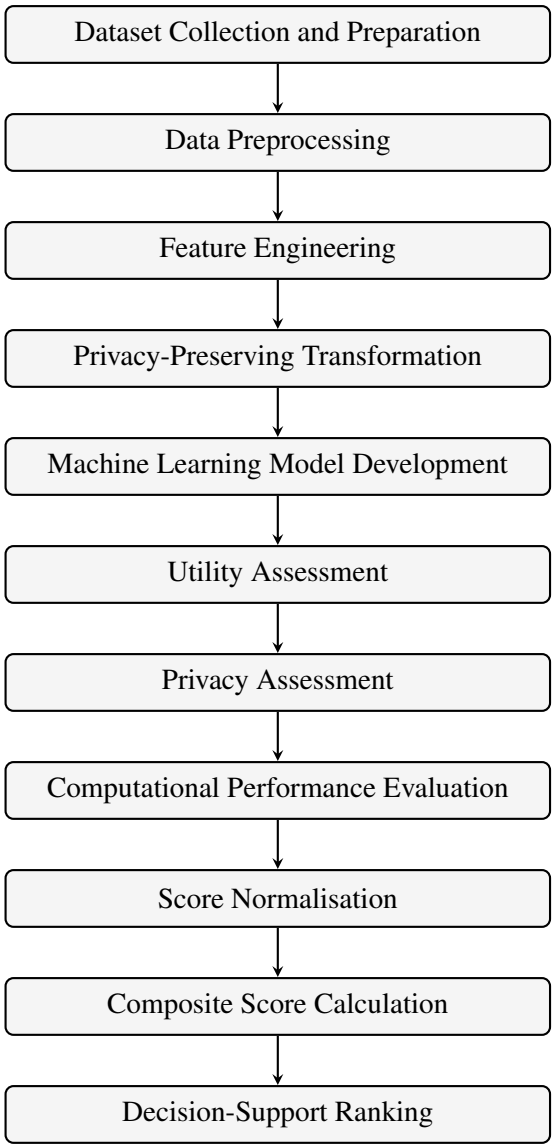

Figure 3.1: Schematic overview of the proposed privacy-preserving evaluation pipeline

3.3 Dataset Description

We use the Adult dataset from the UCI Machine Learning Repository [17] in this experiment. The Adult dataset is one of the most extensively used benchmark datasets in the field of machine learning and privacy-preserving data publishing researches, since it contains demographic and employment-related data with many quasi-identifying features.

The data set is taken from the U.S Census 1994 database and is often used for binary classification tasks for income prediction. The aim variable we will predict here is whether or not a person earns more than $50,000 a year$.

The original dataset has 32,561 records and 15 characteristics. After deleting missing data, there are 30,162 records left for testing.

Table 3.1 summarises the primary dataset attributes.

Attribute	Description
Age	Age of the individual
Workclass	Employment category
Education	Educational qualification
Education-num	Numerical education level
Marital-status	Marital condition
Occupation	Job type
Relationship	Family relationship
Race	Ethnicity category
Sex	Gender
Capital-gain	Capital gain amount
Capital-loss	Capital loss amount
Hours-per-week	Weekly working hours
Native-country	Country of origin
Income	Target class label

Table 3.1: Adult Dataset Attributes

The Adult dataset is an ideal choice for privacy protecting research as it comprises quasi-identifiers such age, education, occupation, marital status, race, and home country. If these traits are combined with external information, persons could be at risk of being re-identified [3].

The dataset consists of both numerical and categorical variables, which makes it suitable for assessing privacy-preserving transformations in machine learning systems.

3.4 Data Preprocessing and Feature Engineering

Data preparation is an important step in machine learning since raw datasets typically contain noisy values, inconsistent formats, missing records and categorical factors that could have adverse impact on prediction performance [33].

The preprocessing pipeline used in this study consists of the following steps:

1. Missing value removal;
2. Categorical encoding;
3. Feature scaling;
4. Feature engineering;
5. Train-test splitting.

3.4.1 Missing Value Removal

In the Adult dataset, missing values are denoted by "?". Some records are dropped to get more uniformity and to lessen the complexity of pre-processing.

The last dataset, after dropping missing information, has 30,162 samples that are acceptable for machine learning and privacy-preserving processing.

In the Adult dataset, many of the parameters are categorical, such as workclass, education, marital status, occupation, relationship, race, sex, and home country.

Machine learning algorithms need the input to be represented numerically. So, label encoding is used to convert the categorical variables into the numerical values.

Label encoding translates the categorical classes into integers:

$$Category_i \rightarrow Integer_i$$

Although one-hot encoding is frequently used in machine learning systems, label encoding was selected to maintain a compact feature representation for the tree-based models used in this study, including Decision Tree, Random Forest, and XGBoost.

3.4.2 Feature Scaling

Numerical attributes are normalised using Min-Max scaling to ensure consistent feature ranges and improve numerical stability [34].

The Min-Max normalisation formula is:

$$X' = \frac{X - X_{min}}{X_{max} - X_{min}}$$

where:

- X is the original feature value;
- X_{min} is the minimum feature value;
- X_{max} is the maximum feature value.

Feature scaling is particularly important for differential privacy implementation because Laplace noise is added to scaled numerical values.

3.4.3 Feature Engineering

Feature engineering is widely recognised as one of the most important stages in predictive modelling because derived attributes can improve machine learning performance and capture hidden relationships within datasets [35].

Several derived features are created in this research:

$$capital_net = capital_gain - capital_loss$$

$$hours_age_ratio = \frac{hours_per_week}{age + 1}$$

$$education_hours = education_num \times hours_per_week$$

Additional binary and grouped features are also introduced:

- age_group;
- high_education;
- has_capital_gain;
- has_capital_loss.

Feature engineering improves downstream machine learning utility by introducing additional semantic relationships between variables.

3.4.4 Train-Test Splitting

The processed dataset is divided into training and testing subsets using an 80:20 ratio.

Stratified splitting is applied to preserve the original class distribution:

$$D = D_{train} \cup D_{test}$$

where:

- D_{train} contains 80% of the dataset;
- D_{test} contains 20% of the dataset.

The testing dataset remains unchanged throughout experimentation to ensure fair comparison across privacy-preserving methods.

3.5 Privacy-Preserving Techniques

This research implements three privacy-preserving techniques:

- k-anonymity through generalisation;
- differential privacy through Laplace noise addition;
- synthetic data generation through bootstrap sampling.

3.5.1 k-Anonymity Through Generalisation

k-anonymity was introduced by Sweeney as a model for reducing re-identification risks in released datasets [3].

A dataset satisfies k-anonymity if every equivalence class contains at least k records:

$$|EC_i| \geq k$$

where EC_i represents an equivalence class.

In this research, k-anonymity is implemented using quantile binning and generalisation.

Continuous quasi-identifying attributes including:

- age;
- fnlwgt;
- hours_per_week;
- capital_net.

are grouped into bins.

For example:

$$29 \rightarrow [20, 30]$$

This transformation reduces attribute granularity and decreases re-identification risk.

However, anonymisation methods may introduce information loss and reduce machine learning performance because detailed attribute relationships are partially removed [6].

3.5.2 Differential Privacy Through Laplace Noise Addition

Differential privacy provides formal mathematical guarantees limiting the influence of individual records on analytical outputs [24].

A mechanism M satisfies differential privacy if:

$$Pr[M(D_1) \in S] \leq e^{\epsilon} Pr[M(D_2) \in S]$$

where D_1 and D_2 are neighbouring datasets differing by one record.

This research implements differential privacy using the Laplace mechanism [29]:

$$X' = X + Laplace\left(0, \frac{1}{\epsilon}\right)$$

where:

- X represents the original feature value;
- ϵ represents the privacy budget.

The experiment uses:

$$\epsilon = 5$$

Smaller values of ϵ provide stronger privacy protection but introduce larger amounts of noise, reducing utility [10].

After noise addition, values are clipped within the interval:

$$0 \leq X' \leq 1$$

This ensures valid feature representations after perturbation.

3.5.3 Synthetic Data Generation Through Bootstrap Sampling

Synthetic data generation aims to preserve statistical structure while reducing direct exposure of original records [26, 15].

This research implements synthetic data generation using bootstrap sampling with replacement.

Bootstrap sampling is widely used in statistical learning because it preserves distributional properties through repeated sampling [36].

The synthetic dataset size satisfies:

$$|D_{synthetic}| = |D_{train}|$$

Although bootstrap sampling is simpler than advanced approaches such as CTGAN and Variational Autoencoders, it provides a useful baseline for evaluating synthetic-data-style privacy-preserving publishing.

3.6 Machine Learning Utility Evaluation

Machine learning utility evaluation measures how effectively original and transformed datasets support predictive modelling. In this research, four supervised classification algorithms are first evaluated to identify the best-performing model for downstream utility assessment:

- Logistic Regression;
- Decision Tree;

- Random Forest;
- XGBoost.

The model comparison stage assures that the final privacy-preserving evaluation is based on the strongest prediction model, not an arbitrary classifier. The experimental comparison showed that XGBoost achieved the best overall performance with 0.8644 accuracy and 0.7049 F1-score. Hence, the XGBoost was used as the final model for utility evaluation.

The final employed XGBoost setting in the privacy preserving evaluation is:

- $n_estimators = 200$;
- $max_depth = 6$;
- $learning_rate = 0.1$;
- $random_state = 42$;
- $eval_metric = logloss$.

The selected model is trained independently using original data, k-anonymised data, differentially private data, and synthetic data. Utility evaluation uses accuracy, precision, recall, and F1-score [37]. F1-score is selected as the primary utility metric because it balances precision and recall.

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN}$$

$$Precision = \frac{TP}{TP + FP}$$

$$Recall = \frac{TP}{TP + FN}$$

$$F1 = 2 \times \frac{Precision \times Recall}{Precision + Recall}$$

3.6.1 Train-Test Split Sensitivity

To assess stability, three train-test split ratios are evaluated: 80:20, 75:25, and 70:30. The final experimental configuration uses the 80:20 split because it achieved the highest composite score for the recommended privacy-preserving method.

3.7 Privacy Evaluation

Privacy evaluation measures the relative privacy protection provided by each method.

This research uses a normalised privacy scoring approach ranging from 0 to 1.

Table-3.2 presents the assigned privacy scores.

Technique	**Privacy Score**
Original Data	0.10
k-Anonymity	0.70
Differential Privacy	0.90
Synthetic Data	0.85

Table 3.2: Assigned privacy scores by technique

Differential privacy receives the highest privacy score because it provides formal mathematical guarantees [24].

Synthetic data receives a high score because artificial records reduce direct exposure of original individuals [16].

k-anonymity provides moderate protection through generalisation and suppression.

3.8 Computational Performance Evaluation

Computational performance is evaluated using model training time because scalability and efficiency are important in practical machine learning deployment [38].

Training time is converted into a normalised computational performance score:

$$C = 1 - \frac{T}{T_{max}}$$

where:

- T represents the training time of a method;
- T_{max} represents the maximum observed training time.

Computational performance is included because privacy-preserving systems must remain practical and scalable for real-world deployment.

3.9 Composite Scoring Model

A weighted composite scoring model is introduced to integrate privacy protection, utility preservation, and computational efficiency into a single interpretable score. Composite

scoring models are widely used in multi-criteria decision analysis because they simplify interpretation across conflicting evaluation dimensions [39, 40].

The final composite score is defined as:

$$S = 0.4P + 0.4U + 0.2C$$

where:

- S represents the composite score;
- P represents the privacy score;
- U represents the utility score;
- C represents the computational performance score.

The weighting approach was chosen according to the main purposes of privacy-preserving data publishing systems. The value of privacy protection and utility preservation was to be equal since both dimensions are the core needs of privacy-preserving machine learning applications [41]. The computational performance was also considered as scalability and execution efficiency have a huge impact on the viability of real-world deployment [9]. The two fundamental aims of privacy-preserving data dissemination are privacy and utility, which are considered equally important. Computational performance is also important for deployability and scalability, though less weighted.

The proposed weighted linear scoring technique fits the criteria of MCDA-based decision-support systems, where a full-ranking system reflects the aggregation of multiple competing criteria for practical choice making [42, 20]. Recent research suggest that multi-criteria assessment approaches improve interpretability, transparency and comparative analysis in complicated decision situations [43]. Similarly, decision support systems are more and more based on weighted scoring methodologies to simplify multi-dimensional evaluation and recommendation procedures [44].

Traditional privacy-preserving assessment approaches involve the interpretation of numerous isolated measures manually. The suggested composite scoring model simplifies the ranking process and makes it interpretable, which helps to select the most balanced privacy-preserving methodology.

3.10 Decision-Support Ranking

The decision-support mechanism ranks privacy-preserving techniques according to composite scores.

The ranking process follows:

$$Rank_i = Sort(S_i)$$

where S_i represents the composite score of method i.

This ranking mechanism enables stakeholders to identify the most balanced privacy-preserving technique rather than focusing exclusively on privacy guarantees [45].

3.11 Ethical Considerations

This research uses publicly available secondary data from the UCI Machine Learning Repository [17]. No direct human participants are involved.

The study does not attempt to identify individuals or reconstruct hidden identities. All experiments are conducted solely for academic evaluation purposes.

The research also emphasises responsible use of privacy-preserving technologies and highlights the importance of balancing privacy protection with analytical usefulness.

3.12 Summary

This chapter detailed the study methods adopted for designing and evaluating the proposed comprehensive solution for privacy-preserving data posting.

The chapter details the dataset, preprocessing pipeline, feature engineering process, privacy-preserving transformations, machine learning evaluation processes, privacy scoring system, computational performance evaluation and composite scoring approach.

We propose a holistic assessment pipeline for supporting interpretable decision-making, which jointly considers privacy protection, machine learning utility, and computational efficiency.

The next chapter discusses the implementation of the proposed evaluation approach.

Chapter 4

Implementation

4.1 Introduction

In this chapter, implementation details of the suggested comprehensive approach for privacy-preserving data publishing evaluation are presented. The implementation incorporates preprocessing, feature engineering, privacy-preserving transformation, machine learning evaluation, privacy evaluation, computing performance measurement, and composite scoring into a single modular pipeline.

The technique is implemented in Python in the Google Colab environment. Python was chosen due to its excellent support for machine learning, data pre-processing, statistical analysis and visualisation through libraries such as Pandas, NumPy, Scikit-learn, Matplotlib and Seaborn [46, 47].

- k-anonymity through generalisation;
- differential privacy through Laplace noise addition;
- synthetic data generation through bootstrap sampling.

The approach is based on the methodology of Chapter 3 and develops it into a practical experimental system that provides the means to compare privacy-preserving methods based on privacy, utility and computational performance metrics.

This chapter presents the implementation environment, the preprocessing pipeline, the feature engineering process, privacy-preserving transformations, the machine learning module, the evaluation module, and the composite scoring implementation.

4.2 System Architecture

4.3 System Architecture

The proposed evaluation system is based on a modular design where each module is responsible for a specific step of the privacy-preserving data publishing evaluation process. A modular design allows individual components to be adjusted or extended without affecting the overall workflow, leading to better scalability, maintainability, transparency and reproducibility. Moreover, the architecture guarantees that all privacy-preserving strategies are tested under the same experimental settings, which enables a fair and consistent comparison among techniques.

The framework consists of nine interrelated modules:

1. Dataset Loading Module;
2. Data Preprocessing Module;
3. Feature Engineering Module;
4. Privacy Transformation Module;
5. Utility Evaluation Module;
6. Privacy Assessment Module;
7. Computational Performance Evaluation Module;
8. Composite Scoring Module;
9. Decision-Support Ranking Module.

The evaluation starts with the loading of the Adult Income data set from the UCI Machine Learning Repository. The dataset is then preprocessed by removing missing values, encoding categorizing features, scaling features, and splitting the data into training and testing sets. Then feature engineering is employed to create extra features so as to enhance the predictive power of machine learning models.

The Privacy Transformation Module implements three representative privacy-preserving technique, i.e., k-anonymity, differential privacy and synthetic data creation. Each changed dataset is preprocessed and evaluated using the same preprocessing and evaluation setup. This assures that any differences in performance are due to the privacy-preserving strategies, not due to differences in the experimental setting.

The first four machine learning classifiers were examined to find the best suited utility evaluation model: Logistic Regression, Decision Tree, Random Forest and XGBoost. The experimental findings revealed that XGBoost had the best prediction performance overall

with an accuracy of 86.44% and an F1-score of 70.49%. Thus, XGBoost was used as the final utility evaluation model for all the evaluations of privacy-preserving techniques [**?**].

The Utility Evaluation Module evaluates the analytical utility of every modified dataset using common classification metrics including accuracy, precision, recall, and F1-score. The Privacy Assessment Module utilises predetermined privacy scores to evaluate the relative level of privacy protection of each technique. The Computational Performance Evaluation Module examines the execution efficiency by using training time and resource utilization. The following evaluation dimensions are selected because the realistic privacy-preserving data publishing should trade off privacy protection, analytical benefit and computational expense.

All metrics are assessed, normalised and aggregated in a weighted composite scoring model:

[S = 0.4P + 0.4U + 0.2C]

where (P) is the privacy score, (U) is the utility score and (C) is the computational performance score. The weighting approach gives equal attention to privacy and utility, and considers computational performance as a supporting factor.

Finally, the Decision-Support Ranking Module rates all privacy-preserving strategies according to their composite scores and proposes the most suitable strategy for secure data publishing and machine learning applications. This ranking system provides a realistic framework to enhance decision-making, allowing researchers and organisations to choose privacy-preserving strategies based on many evaluation criteria, instead of relying on a single performance indicator.

Figure 4.1 illustrates the overall architecture of the proposed evaluation framework.

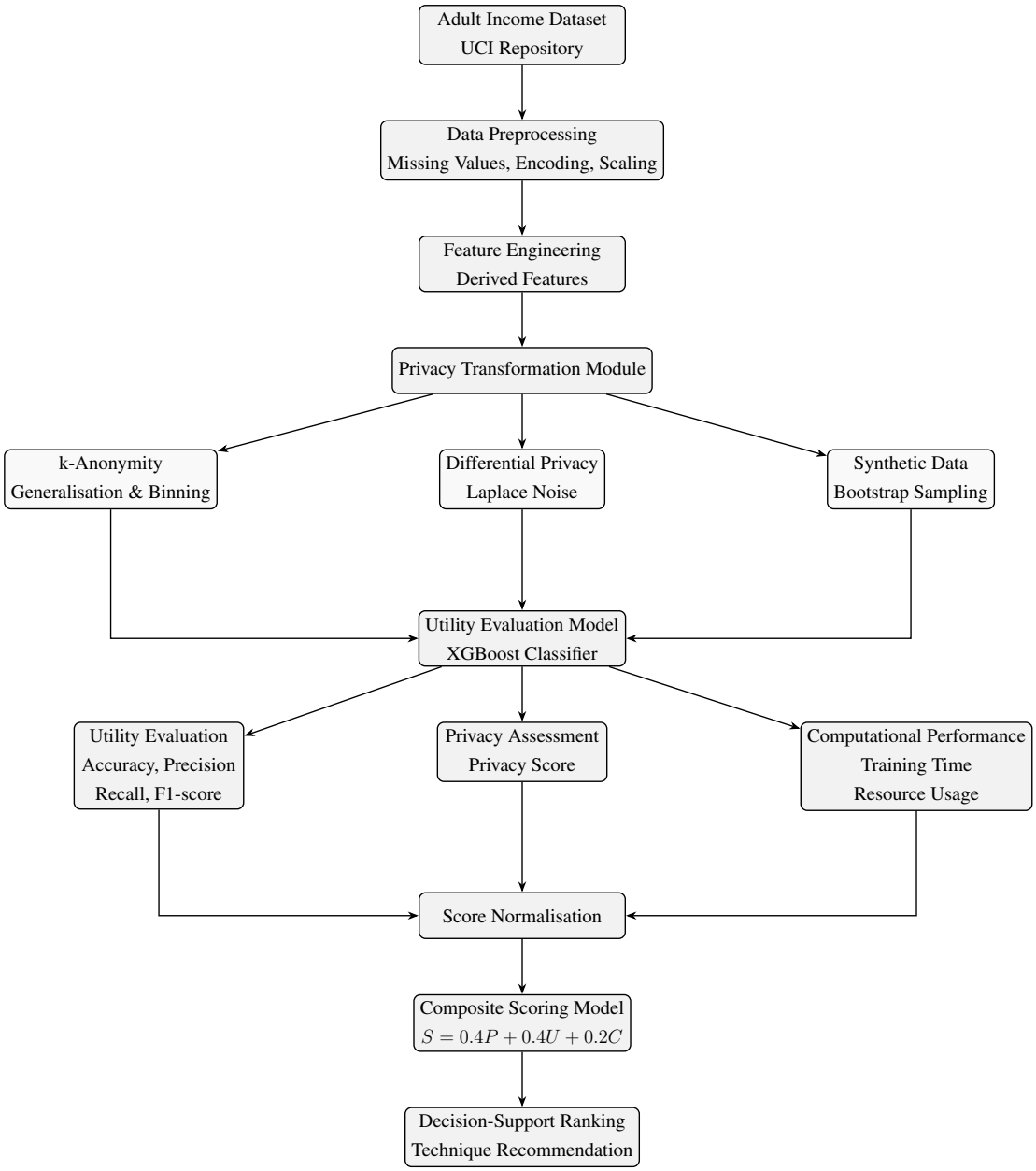

Figure 4.1: System architecture of the comprehensive evaluation framework

The modular architecture allows each privacy-preserving technique to be evaluated independently using identical preprocessing and machine learning configurations. This ensures fair comparison across methods.

4.4 Implementation Environment

The experiment was implemented using Google Colab because it provides cloud-based computational resources and integrated support for Python machine learning libraries.

Table- 4.1 summarises the implementation environment.

Component	Description
Programming Language	Python 3
Development Environment	Google Colab
Data Processing Library	Pandas
Numerical Computing	NumPy
Machine Learning Libraries	Scikit-learn, XGBoost
Visualisation Libraries	Matplotlib, Seaborn
Performance Monitoring	psutil

Table 4.1: Summary of implementation environment

Python was selected because it is widely used in machine learning and privacy-preserving data analysis research [46]. Pandas provides efficient tabular data processing functionality, while NumPy supports numerical computation and matrix operations [47, 48].

Scikit-learn was used for machine learning implementation because it provides robust implementations of classification algorithms, preprocessing utilities, and evaluation metrics.

4.5 Dataset Loading and Preprocessing Module

The dataset loading module imports the Adult dataset into the experimental environment.

The dataset is loaded using the Pandas library:

```
df = pd.read_csv("adult.csv")
```

After loading the dataset, preprocessing operations are applied to improve consistency and machine learning compatibility.

4.5.1 Missing Value Handling

The Adult dataset contains missing values represented using question marks. These records are removed using:

```
df.replace(' ?', np.nan, inplace=True)
df.dropna(inplace=True)
```

Removing missing values improves data consistency and avoids additional complexity associated with imputation methods.

4.5.2 Categorical Encoding

Several attributes contain categorical values that cannot be directly processed by machine learning models.

Label encoding is applied using Scikit-learn:

```
encoder = LabelEncoder()
df[col] = encoder.fit_transform(df[col])
```

Label encoding converts categorical values into integer representations suitable for the machine learning models used in the experiment.

4.5.3 Feature Scaling

Numerical features are normalised using Min-Max scaling [34]:

$$X' = \frac{X - X_{min}}{X_{max} - X_{min}}$$

The implementation uses:

```
scaler = MinMaxScaler()
X_scaled = scaler.fit_transform(X)
```

Feature scaling is particularly important for differential privacy implementation because Laplace noise is added to scaled values.

4.6 Feature Engineering Module

Feature engineering improves machine learning utility by generating additional attributes that capture hidden relationships within the dataset [35].

Several derived features are implemented.

4.6.1 Capital Net Feature

A net capital feature is created using:

$$capital_net = capital_gain - capital_loss$$

Implementation:

```
df['capital_net'] =
df['capital_gain'] - df['capital_loss']
```

This feature provides a more informative representation of financial behaviour.

4.6.2 Hours-Age Ratio

A workload-age relationship feature is introduced:

$$hours_age_ratio = \frac{hours_per_week}{age + 1}$$

Implementation:

```
df['hours_age_ratio'] =
df['hours_per_week'] / (df['age'] + 1)
```

4.6.3 Education-Hours Feature

An interaction feature combining education and workload is introduced:

$$education_hours = education_num \times hours_per_week$$

This feature captures relationships between education level and working intensity.

4.6.4 Grouped Features

Additional grouped and binary features are implemented:

- age_group;
- high_education;
- has_capital_gain;
- has_capital_loss.

Feature engineering increases the number of analytical dimensions and improves downstream utility evaluation.

4.7 k-Anonymity Implementation

k-anonymity is implemented using generalisation and quantile binning [3].

Continuous quasi-identifying attributes are grouped into bins using:

```
pd.qcut(df['age'], q=5)
```

The following attributes are generalised:

- age;

- fnlwgt;
- hours_per_week;
- capital_net.

Example transformation:

$$29 \rightarrow [20, 30]$$

Generalisation reduces attribute granularity and decreases re-identification risk.

However, excessive generalisation may reduce machine learning utility because detailed feature relationships are partially removed [6].

4.8 Differential Privacy Implementation

Differential privacy is implemented using the Laplace mechanism [24, 29].

Noise is sampled from the Laplace distribution:

$$Laplace(0, \frac{1}{\epsilon})$$

where:

$$\epsilon = 5$$

Implementation:

```
noise = np.random.laplace(
0, 1/epsilon, X_train.shape
)
X_noisy = X_train + noise
```

After noise addition, values are clipped within the range:

$$0 \leq X' \leq 1$$

using:

```
X_noisy = np.clip(X_noisy, 0, 1)
```

Differential privacy provides strong theoretical guarantees but may reduce machine learning performance because noise distorts feature distributions [10].

4.9 Synthetic Data Generation Implementation

Synthetic data generation is implemented using bootstrap sampling with replacement [36].

Implementation:

```
synthetic_data = resample(
X_train,
replace=True,
n_samples=len(X_train)
)
```

Bootstrap sampling preserves statistical distributions and machine learning utility while reducing direct exposure of original records.

Although bootstrap sampling is simpler than advanced generative models such as CT-GAN and Variational Autoencoders, it provides an interpretable baseline for evaluating synthetic-data-style privacy-preserving publishing.

4.10 Machine Learning Evaluation Module

The implementation includes a model comparison module followed by a final utility evaluation module. Four classifiers were implemented and evaluated on the preprocessed Adult dataset:

- Logistic Regression;
- Decision Tree;
- Random Forest;
- XGBoost.

The XGBoost classifier was implemented using the following configuration:

```
model = XGBClassifier(
    n_estimators=200,
    max_depth=6,
    learning_rate=0.1,
    random_state=42,
    eval_metric="logloss",
    n_jobs=-1
)
```

4.11 Utility Evaluation Module

Utility evaluation uses four classification metrics [37].

4.11.1 Accuracy

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN}$$

4.11.2 Precision

$$Precision = \frac{TP}{TP + FP}$$

4.11.3 Recall

$$Recall = \frac{TP}{TP + FN}$$

4.11.4 F1-Score

$$F1 = 2 \times \frac{Precision \times Recall}{Precision + Recall}$$

Scikit-learn evaluation functions are used:

```
accuracy_score()
precision_score()
recall_score()
f1_score()
```

F1-score is selected as the primary utility metric because it balances precision and recall.

4.12 Privacy Evaluation Module

Privacy evaluation assigns normalised privacy scores ranging from 0 to 1.

Table- 4.2 presents the implemented scores.

Technique	**Privacy Score**
Original Data	0.10
k-Anonymity	0.70
Differential Privacy	0.90
Synthetic Data	0.85

Table 4.2: Implemented privacy scores by technique

Differential privacy receives the highest score because it provides formal mathematical guarantees.

Synthetic data receives a high score because it reduces direct exposure of original records.

4.13 Computational Performance Module

Computational performance is evaluated using model training time.

Execution time is measured using:

```
start = time.time()
...
end = time.time()
```

Performance scores are normalised using:

$$C = 1 - \frac{T}{T_{max}}$$

where:

- T is the training time;
- T_{max} is the maximum observed training time.

Computational performance is included because privacy-preserving systems must remain scalable and deployable.

4.14 Composite Scoring Implementation

The approach integrates privacy, utility, and performance into a weighted composite score.

The composite score is defined as:

$$S = 0.4P + 0.4U + 0.2C$$

where:

- P represents privacy score;
- U represents utility score;
- C represents computational performance score.

Implementation:

```
score = (
0.4 * privacy +
0.4 * utility +
0.2 * performance
)
```

Composite scoring improves interpretability and supports multi-criteria decision-making [39, 40].

4.15 Decision-Support Ranking Module

The decision-support module ranks methods according to composite scores.

Implementation:

```
results.sort_values(
by='Composite Score',
ascending=False
)
```

The ranking system enables stakeholders to identify the most balanced privacy-preserving method.

4.16 Summary

In this chapter, the specifics of the implementation of the proposed privacy-preserving assessment approach were provided.

In this chapter, we discussed the preprocessing pipeline, feature engineering process, privacy-preserving transformations, machine learning implementation, utility evaluation module, privacy evaluation system, computational performance analysis, and composite scoring mechanism.

This technique implements the theoretical techniques discussed in Chapter 3 in a realistic and reproducible experimental setup.

Next chapter gives the experimental findings and comparative evaluation of the implemented privacy preservation approaches.

Chapter 5

Experimental Results and Evaluation

5.1 Introduction

This chapter reports on the experimental results and evaluation of the suggested privacy-preserving data publishing method. The review targets comparing privacy-preserving strategies on the grounds of privacy protection, machine learning value, and computational performance. The evaluation was done in the Google Colab environment, using the Adult dataset from the UCI Machine Learning Repository. Three privacy-preserving strategies were assessed, including k-anonymity via generalisation and binning, differential privacy via Laplace noise addition, and synthetic data generation by bootstrap sampling.

The evaluation also involves a train-test split sensitivity analysis and model selection. Four machine learning models were tested and XGBoost was selected as a final utility evaluation model because of its highest prediction performance. Moreover, three split ratios were considered, 80:20, 75:25 and 70:30.

5.2 Experimental Configuration

The experiment was conducted using Python in Google Colab. The implementation used Pandas, NumPy, Scikit-learn, XGBoost, Matplotlib, Seaborn, and psutil for data processing, modelling, evaluation, visualisation, and performance monitoring [46, 47, 48].

Parameter	Value
Dataset	Adult Dataset
Records After Cleaning	30,162
Final Train-Test Split	80:20
Final Utility Model	XGBoost
Number of Estimators	200
Maximum Tree Depth	6
Learning Rate	0.1
Differential Privacy Budget	$\epsilon = 5$
Evaluation Metrics	Accuracy, Precision, Recall, F1-score
Composite Score Weights	Privacy 0.4, Utility 0.4, Performance 0.2

Table 5.1: Experimental configuration parameters

5.3 Machine Learning Model Comparison

Before evaluating privacy-preserving techniques, four machine learning models were compared on the preprocessed dataset.

Model	Accuracy	Precision	Recall	F1-Score
Logistic Regression	0.8226	0.7218	0.4680	0.5679
Decision Tree	0.8389	0.6912	0.6378	0.6634
Random Forest	0.8576	**0.7736**	0.6052	0.6791
XGBoost	**0.8644**	0.7693	**0.6505**	**0.7049**

Table 5.2: Machine Learning Model Comparison Results

XGBoost achieved the highest accuracy, recall, and F1-score. Although Random Forest achieved slightly higher precision, XGBoost produced the strongest balance between precision and recall. Since F1-score is the primary utility metric in this research, XGBoost was selected as the final model for privacy-preserving utility evaluation.

5.4 Train-Test Split Evaluation

To evaluate the stability of the proposed approach, three train-test split ratios were tested: 80:20, 75:25, and 70:30. The recommended method remained Synthetic Data across all split configurations.

Split	Synthetic Data	k-Anonymity	Differential Privacy	Original Data
80:20	**0.7677**	0.6770	0.6157	0.4650
75:25	0.7430	**0.6802**	0.6118	**0.4828**
70:30	0.7399	0.6295	0.6144	0.4668

Table 5.3: Composite Score Comparison Across Train-Test Split Configurations

The 80:20 split produced the highest composite score for Synthetic Data, with a score of 0.7677. Therefore, the 80:20 split was selected as the final experimental configuration. The consistency of the ranking across all splits also indicates that the decision-support result is stable.

5.5 Baseline Performance on Original Data

The baseline experiment evaluated XGBoost using the original dataset without privacy-preserving transformation.

Metric	Score
Accuracy	0.8644
Precision	0.7693
Recall	0.6505
F1-Score	0.7049

Table 5.4: Baseline XGBoost classification performance on the original dataset

The baseline XGBoost model achieved an accuracy of 86.44% and an F1-score of 0.7049.

5.6 Utility Evaluation Results

Machine learning utility evaluation measures how effectively transformed datasets preserve predictive performance after privacy-preserving transformation. F1-score was used as the utility score because it balances precision and recall [37].

Technique	Utility Score / F1-Score
Original Data	0.7049
k-Anonymity	0.6093
Differential Privacy	0.6392
Synthetic Data	**0.7035**

Table 5.5: Final Utility Evaluation Results Using XGBoost and 80:20 Split

Synthetic Data achieved the highest utility among the privacy-preserving techniques, with an F1-score of 0.7035. This value is very close to the original data baseline of 0.7049, indicating that bootstrap-based synthetic generation preserved important predictive patterns.

5.7 Privacy Evaluation Results

Technique	**Privacy Score**
Original Data	0.10
k-Anonymity	0.70
Differential Privacy	0.90
Synthetic Data	0.85

Table 5.6: Privacy Evaluation Results

Differential Privacy achieved the highest privacy score because it provides formal mathematical privacy guarantees [24]. Synthetic Data achieved the second-highest privacy score because artificial records reduce direct exposure of original records.

5.8 Computational Performance Evaluation

Computational performance was evaluated using model training time. Training times were converted into normalised performance scores using:

$$C = 1 - \frac{T}{T_{max}}$$

Technique	**Performance Score**
Original Data	0.7150
k-Anonymity	0.7665
Differential Privacy	0.0000
Synthetic Data	0.7315

Table 5.7: Computational Performance Results for 80:20 Split

k-Anonymity achieved the highest computational performance score, while Differential Privacy achieved the lowest performance score.

5.9 Composite Scoring Results

The proposed evaluation approach integrates privacy protection, utility preservation, and computational performance into a weighted composite score:

$$S = 0.4P + 0.4U + 0.2C$$

Rank	Technique	Privacy	Utility	Performance	Composite Score
1	Synthetic Data	0.8500	0.7035	0.7315	**0.7677**
2	k-Anonymity	0.7000	0.6093	0.7665	0.6770
3	Differential Privacy	0.9000	0.6392	0.0000	0.6157
4	Original Data	0.1000	0.7049	0.7150	0.4650

Table 5.8: Final Composite Scoring Results for 80:20 Split

Synthetic Data received the greatest scores for the overall composite score of 0.7677. Differential Privacy has the highest privacy score but its lower computational performance caused it to be ranked lower overall. K-Anonymity was ranked second since it provides moderate privacy and good computational performance but its lower utility score lowered its final outcome. Data had the lowest composite score as it did not offer any substantial privacy protection.

Figure 5.1 illustrates the final ranking results.

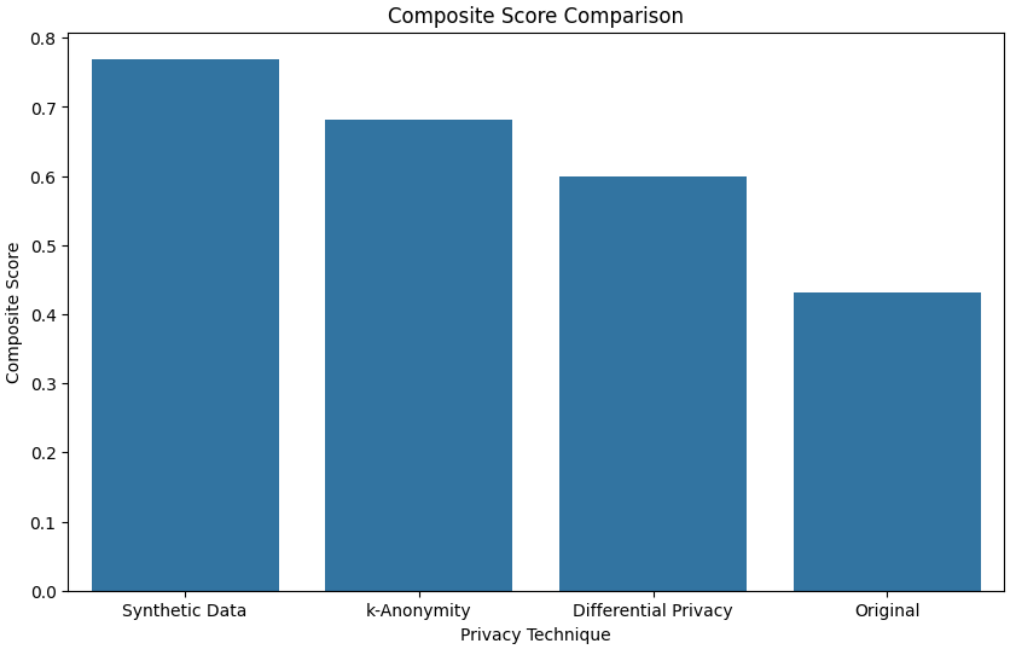

Figure 5.1: Comparison of composite scores across evaluated techniques

Figure 5.2 Privacy, Utility, Performance Comparison.

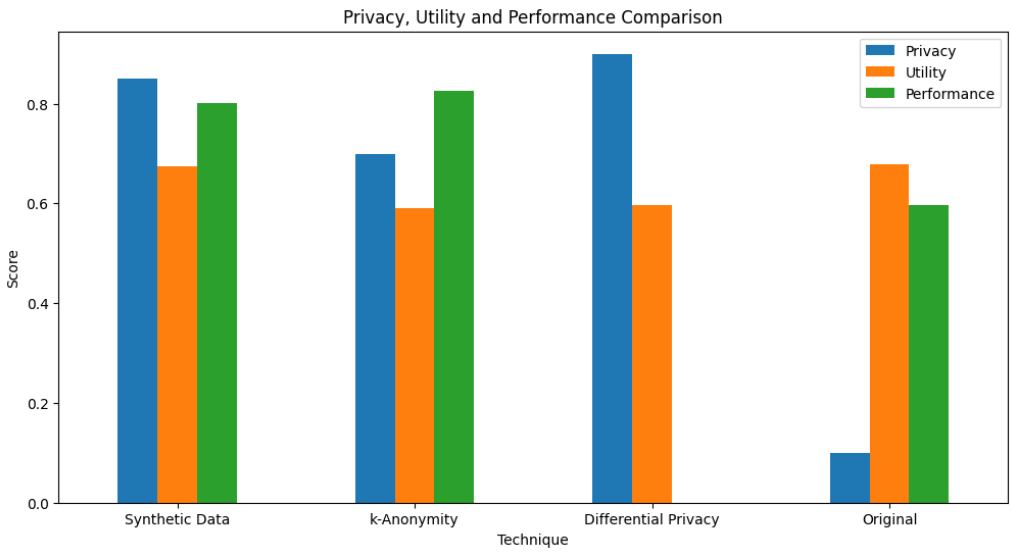

Figure 5.2: Comparative visualisation of privacy, utility, and computational performance scores

5.10 Critical Discussion

The model comparison showed that for this data set XGBoost was the best performing classifier. The best accuracy and F1 score was found for XGBoost and is therefore suitable for the last utility assessment. The split analysis also showed that the 80:20 split had the greatest composite score for Synthetic Data. Importantly, Synthetic Data was placed first across all three split configurations, demonstrating that the proposed ranking is stable.

Our results suggest that there is no single privacy enhancing strategy that outperforms others for all evaluation metrics. Differential Privacy performed best in terms of privacy but worse in terms of computational performance. k-Anonymity generalization gained efficiency at the cost of predictive utility. Synthetic Data provided the optimum trade-off, preserving predictive utility, while delivering robust privacy and practical computing efficiency.

In this chapter, we have shown the experimental findings and evaluation of the suggested privacy-preserving evaluation approach. XGBoost was chosen as the final model due to its highest F1-score among the tested classifiers. The 80:20 split resulted in the greatest composite score and was adopted in the final configuration. Finally, the best composite score is 0.7677 Synthetic Data in the final ranking. Then k-Anonymity, Differential Privacy and Original Data.

Chapter 6

Discussion

6.1 Introduction

In Chapter 6, we critically evaluate the experimental results shown in Chapter 5 and analyse the consequences of the suggested privacy-preserving evaluation approach. In this paper we address the interplay between privacy protection, machine-learning utility, computational performance, model selection, train-test split configuration and composite decision-support ranking.

6.2 Discussion of Model Selection Results

In the model comparison phase, Logistic Regression, Decision Tree, Random Forest and XGBoost were investigated. XGBoost had the best overall classification performance with an accuracy of 0.8644 and F1-score of 0.7049. Since the key utility measure was F1-score, XGBoost was selected as the final model for privacy-preserving utility evaluation.

Logistic Regression is a linear model that may not be able to capture complex non-linear interactions in the Adult dataset and hence lowest F1-score. Decision Tree performed better than Logistic Regression but still was not as excellent as ensemble techniques. Random Forest had a good precision but worse recall than XGBoost. XGBoost presented the best trade-off between recall and precision and was the most suitable model for the final evaluation.

6.3 Analysis of Train-Test Split Results

The study employed three train-test splits, 80:20, 75:25 and 70:30. The best Synthetic Data composite score was 0.7677 for the 80:20 split. The Synthetic Data composite scores for 75:25 and 70:30 splits were lower at 0.7430 and 0.7399 respectively.

This outcome indicates that the 80:20 split offered the optimum compromise between the availability of training data and the reliability of evaluation. Importantly, Synthetic Data was ranked first in all three split settings, indicating the advice was stable and not dependent on a particular split setting.

6.4 Results Discussion for k-Anonymity

k-Anonymity was the runner-up in the final composite score with a score of 0.6770. It provided moderate privacy protection and good computing performance. However, the utility score was lower than Synthetic Data and Original Data due to the reduced feature granularity due to generalisation.

6.5 Interpretation of Composite Scoring Results

The composite scoring model was designed to combine privacy, utility, and computational performance:

$$S = 0.4P + 0.4U + 0.2C$$

Rank	Technique	Composite Score
1	Synthetic Data	**0.7677**
2	k-Anonymity	0.6770
3	Differential Privacy	0.6157
4	Original Data	0.4650

Table 6.1: Final Decision-Support Ranking

This ranking shows that the highest privacy score alone does not guarantee the best overall method. Differential Privacy had the strongest privacy score but ranked third because its computational performance was weak. Synthetic Data ranked first because it achieved the strongest balance across the three dimensions.

6.6 Answering the Research Questions

The first study topic was to understand the differences between k-Anonymity, Differential Privacy and Synthetic Data, in terms of the privacy protection and machine learning utility. We find that Differential Privacy offers the strongest privacy protection, and Synthetic Data the strongest utility among privacy-preserving approaches.

The second study topic examined the differences in computing performance between the privacy-preserving approaches. The greatest score for computational performance was given to k-Anonymity, followed by Synthetic Data and Original Data. Differential Privacy obtained the lowest performance score since adding noise increases computing cost and reduces feature stability.

The third study topic explored the role of composite scoring and decision-support systems in improving comparative evaluation. Results indicate that composite scoring offers a single interpretable ranking that balances privacy, utility and performance. This gives a simpler way to choose the most balanced strategy instead of manually analyzing individual indicators.

6.7 Practical Implications

The results have various practical consequences. First, privacy-preserving solutions should not be selected solely on privacy pledges by organisations. Real world deployment is also influenced by utility and computational efficiency. Second , Synthetic Data can be a good practical solution when the goal is to keep machine learning utility while reducing direct exposure of original data. Third, Differential Privacy is desirable when formal privacy protections are necessary, especially in sensitive areas like healthcare, finance, and government systems.

6.8 Research Limitations

Several limitations should be acknowledged. First, k-Anonymity was developed utilizing a quantile-based generalization instead of complex equivalence-class optimisation. Second, we implemented Differential Privacy by adding Laplace noise, rather than using more complex differentially private learning methods such as DP-SGD. Third, Synthetic Data was generated by bootstrap sampling rather than sophisticated generative models. Fourth, privacy scores were determined by theoretical and practical interpretation, not by empirical assault simulations. Fifth, the experiment only employed the Adult dataset, limiting generalisability. Finally the composite score weights were fixed and may need to be adjusted for domain particular applications.

6.9 Future Work

Future study should enhance the approach provided here by deploying advanced synthetic data generation models such as CTGAN, DP-CTGAN, PATE-GAN, Variational Autoencoders, and diffusion models. Future work should include consider empirical privacy

attack evaluation, such as membership inference, attribute inference, reconstruction and linkage attacks. The technique should also be tested on other datasets from healthcare, banking, cyber security, education, and public administration.

6.10 Summary

This chapter critically evaluates the experimental results. Utility evaluation found the best model to be XGBoost and the strongest experimental setup to be 80:20 split. The last results show that the Synthetic Data had the best overall score and was recommended as the most balanced technique for privacy preservation.

Chapter 7

Conclusion and Future Work

7.1 Introduction

This chapter summarizes the main results, contributions, limits and directions for future research of the dissertation, and provides a brief reflective synthesis of the work. It revisits the study aim and objectives set out in Chapter 1, evaluates the extent to which they have been realized and discusses the contribution of the findings to theory and practice. We developed and experimentally evaluated three privacy-preserving strategies using the Adult dataset from the UCI Machine Learning Repository:

- k-anonymity via generalization and binning;
- differential privacy by adding Laplace noise;
- synthetic data generation using bootstrap sampling.

Our evaluation methodology proposes a weighted composite score scheme that might help decision-oriented comparison and ranking of privacy-preserving strategies.

This chapter summarizes the main study findings, highlights the key academic and practical contributions, evaluates limits, and identifies potential for further research.

7.2 Research Summary

The problem of evaluating privacy-preserving data publishing techniques is studied in this dissertation by means of a multi-dimensional framework that takes into account privacy protection, machine learning utility and computational performance. The motivation for the research was the fragmentation of existing evaluation methodologies that usually evaluate these characteristics in isolation and hence do not give practitioners with the integrated and interpretable recommendations needed for evidence-based technique selection. The proposed framework bridges this gap with a composite scoring model

(S=0.4P+0.4U+0.2C) and a decision-support ranking mechanism applied on a controlled empirical evaluation of three representative PPDP technique categories — k-anonymity, differential privacy, and synthetic data generation — using the Adult Income dataset and XGBoost as the utility evaluation model.

7.3 Summary of Experimental Findings

The experimental results show large differences between the privacy-preserving techniques that were investigated.

Since the initial dataset was not privacy treated, it was most valuable for machine learning. But it also ended up with the worst privacy protection. k-anonymity gave a fair degree of privacy protection and was very computationally efficient. Generalization downscaled feature complexity and sped up training. The loss of information substantially affected memory and the overall predictive usefulness.

Differential privacy scored highest on privacy because it offers explicit mathematical guarantees to the reduction of the effect of individual records over analytical outputs [24]. However, the addition of Laplace noise lowered the value of machine learning and created a significant processing burden.

Synthetic data generation performed the best overall composite score because of its preservation of machine learning utility, good computational performance, and reasonably excellent privacy protection.

The final standings were as follows:

Rank	Technique	Composite Score
1	Synthetic Data	0.7677
2	k-Anonymity	0.6770
3	Differential Privacy	0.6157
4	Original Data	0.4650

Table 7.1: Final Composite Ranking Results

These findings demonstrate that the strongest privacy guarantee does not necessarily produce the best overall solution when utility and computational efficiency are also considered.

7.4 Research Contributions

This study makes several important academic and practical contributions.

7.4.1 Comprehensive Evaluation Approach

The research proposed a comprehensive evaluation approach integrating privacy protection, machine learning utility, and computational performance into a single evaluation pipeline.

Unlike many existing approaches that evaluate privacy and utility independently, the proposed evaluation approach provides integrated multidimensional evaluation.

7.4.2 Composite Scoring and Decision Support

The proposed composite scoring model was developed to support multidimensional evaluation of privacy-preserving data publishing techniques. Privacy-preserving methods often exhibit conflicting characteristics across privacy protection, machine learning utility, and computational performance. Consequently, evaluating a technique using a single metric may produce misleading conclusions. To address this challenge, a weighted composite scoring model was adopted based on principles commonly used in Multi-Criteria Decision Analysis (MCDA) and decision-support systems [39, 40].

A major contribution of this research is the introduction of a weighted composite scoring mechanism:

$$S = 0.4P + 0.4U + 0.2C$$

where:

- P represents privacy score;
- U represents utility score;
- C represents computational performance score.

The weighting strategy reflects the primary objectives of privacy-preserving data publishing. Privacy protection and utility preservation were assigned equal weights of 0.4 because these dimensions represent the two fundamental and often competing requirements of privacy-preserving machine learning systems. A technique that provides strong privacy but poor utility may have limited practical value, while a technique that preserves utility but fails to protect sensitive information cannot satisfy privacy requirements. Therefore, privacy and utility were considered equally important evaluation criteria.

Computational performance was assigned a lower weight of 0.2 because execution efficiency primarily affects deployability rather than the core objectives of privacy-preserving data publishing. Although computational performance remains important in real-world environments, it is generally considered secondary to privacy protection and analytical

usefulness. Similar weighting strategies have been adopted in multi-criteria evaluation research where primary criteria receive greater emphasis than supporting criteria [49, 50].

Therefore, the specified weights must be understood as research-driven priorities, not as mathematically ideal values. Depending on the application needs, alternative schemes of weighting could be used. For instance, healthcare applications may focus more on privacy protection, whereas corporate analytics applications may focus more on utility preservation. Future study may explore adaptive weighting strategies, expert-based weighting approaches like Analytic Hierarchy Process (AHP), or data-driven optimization procedures for the automatic determination of criterion weights.

The suggested weighting strategy aims to offer a visible, interpretable and practically useful decision support framework for evaluating privacy-preserving strategies inside a single evaluation framework.

The composite score approach enhances interpretability and aids practical decision making by combining many assessment dimensions into a single ranking outcome [51] .

7.4.3 Comparative Evaluation Across Multiple Privacy Techniques

The approach evaluated anonymisation, differential privacy, and synthetic data generation within the same experimental environment.

This comparative evaluation provides broader insights than approachs focusing only on individual privacy-preserving categories.

7.4.4 Practical Experimental Pipeline

The dissertation also provides a practical and reproducible implementation pipeline using Python and Google Colab. The approach can be extended to additional datasets and privacy-preserving methods.

7.5 Comparison with Existing Literature

The results obtained in this study are consistent with other major findings reported in the literature.

Previous research showed that differential privacy gives strong formal guarantees, but may diminish utility by adding noise [29, 10]. The experimental results validated this observation as differential privacy led to a decrease in classification performance and computing efficiency.

Similarly, the information loss due to generalization of k-anonymity has been demonstrated in prior work [6]. This finding is supported by the reduced recall in this dissertation.

Recent studies on synthetic data showed that synthetic datasets could preserve analytical utility while maximizing the non-direct exposure of original records [15, 16]. The good performance of synthetic data generation in our work is in line with these results.

However, the suggested evaluation approach adds to the current research by incorporating privacy, utility and computing performance in a whole composite scoring and decision-support system.

7.6 Practical Implications

The proposed evaluation approach has certain practical consequences for companies adopting privacy-preserving machine learning system.

First of all the results indicate that the choice of the privacy-preserving strategy should not be based just on the privacy promises. Besides, the utility preservation and computing efficiency are of great importance in practical deployment situations.

Second, synthetic data production can provide a helpful privacy-utility trade-off in situations where direct data sharing is not allowed.

Third, differential privacy is still crucial in highly sensitive sectors like as healthcare and government systems, where formal guarantees are needed at the expense of diminished utility.

The suggested composite score methodology may assist in decision-making regarding data governance, regulatory compliance, and the use of privacy-preserving AI.

7.7 Research Limitations

The principal limitations of this research are as follows: the implementation of each privacy-preserving technique is representative rather than optimal (simplified quantile binning for k-anonymity, direct Laplace noise addition for differential privacy, and bootstrap sampling for synthetic data generation); privacy evaluation relies on theoretical scores rather than empirically simulated attacks; the experiment is conducted on a single dataset, limiting cross-domain generalisability; and the composite scoring weights are fixed rather than adaptive. These limitations are acknowledged throughout the dissertation and are addressed in the future research agenda below.

7.8 Future Work

Several opportunities exist for extending this research.

7.8.1 Advanced Synthetic Data Generation

Future research should implement advanced generative models such as:

- CTGAN;
- DP-CTGAN;
- PATE-GAN;
- Variational Autoencoders;
- Diffusion models.

These approaches may provide stronger privacy protection while maintaining high utility.

7.8.2 Empirical Privacy Attack Evaluation

Future work should integrate empirical attack simulation including:

- Membership inference attacks;
- Attribute inference attacks;
- Reconstruction attacks;
- Linkage attacks.

This would provide stronger empirical privacy evaluation beyond assigned privacy scores.

7.8.3 Additional Datasets and Domains

The approach should be evaluated using additional datasets from:

- Healthcare;
- Finance;
- Cybersecurity;
- Education;
- Social media analytics.

This would improve generalisability and practical relevance.

7.8.4 Adaptive Weighting Mechanisms

Future research should explore adaptive and domain-specific weighting strategies for composite scoring.

Different application domains may prioritise privacy, utility, or computational efficiency differently.

7.8.5 Scalability and Explainability

Future work should also evaluate:

- Scalability;
- Memory consumption;
- Energy efficiency;
- Explainable AI integration.

These extensions would improve practical deployment capabilities.

7.9 Final Conclusion

Here in, we present a complete assessment system for privacy-preserving data publishing, which we have created, implemented and verified. This work fills a significant vacuum in the existing literature. The framework combines privacy protection, machine learning utility, and computational performance into a weighted composite scoring and decision-support system, providing a principled, interpretable and actionable basis for selecting PPDP techniques that exceeds the piecemeal, single-dimensional assessments that characterize the state of the art. The empirical evaluation clearly reveals that synthetic data generation via bootstrap sampling yields the best overall trade-off among the three evaluation dimensions for the Adult Income dataset, with a composite score of 0.7677 in the 80:20 train-test setting. Notably, the study proves that differential privacy, even with its strongest formal privacy guarantees, is not the best solution in practice when utility and computational efficiency are considered together – a finding with broad implications for data governance practices. The research highlights that selecting a suitable privacy-preserving data publishing technique is essentially a multi-objective optimization issue. If a framework does not encompass all relevant elements in an integrated evaluation approach, it could provide practitioners with inaccurate advice. The composite scoring methodology proposed in this dissertation makes an important progress towards solving that challenge, and its refinement and development constitutes a fruitful agenda for future research in privacy-preserving artificial intelligence.

Chapter 8

Reflections

8.1 Introduction

In this chapter I offer a reflexive assessment of the research process in the making of this dissertation. It is about the knowledge obtained, the problems experienced, the abilities acquired and the lessons learned throughout the project. The reflection also examines the strengths and limitations of the research and how the experience has helped in academic and professional development.

8.2 Reflections on the research process

The work of this dissertation gave me excellent experience to perform independent research in the topic of privacy-preserving data publishing. The main aim at the beginning of the research was to examine the evaluation of various privacy-preserving methods by adopting a balanced framework in terms of privacy protection, machine learning utility, and computational performance.

A major lesson acquired from the study process is that it is hard to balance opposing agendas. At first it seemed that enhanced privacy protection could lead to a better solution. Yet, literature research and experimental results suggested that privacy preservation often brings tradeoffs that impact data value and predictive performance. This knowledge has a strong influence on the design of the suggested composite scoring framework.

The research approach also emphasized the need for systematic experimentation. We spent a lot of work on literature review, selection of evaluation criteria, pre-processing of the data, implementation of privacy-preserving transformations and validation of the machine learning models. The efforts showed the importance of evidence-based decision making and the usefulness of rigorous methodological planning.

8.3 Skills Developed

The project was a good opportunity to gain technology skills related to cybersecurity, data science and machine learning.

In the implementation phase, we found practical expertise in:
beginitemize

Feature engineering and data preprocessing

approaches for privacy-preserving data transformation

creation and evaluation of machine learning models

Performance analysis and experimental design • Statistical analysis of outcomes • Data visualization and reporting

This also increased competence in Python and associated data science libraries. The design and use of several privacy-preserving algorithms have improved the knowledge of the real machine learning processes and privacy engineering ideas.

In addition, the development of the composite scoring and decision-support framework boosted analytical thinking and problem-solving skills by necessitating the integration of various evaluation variables into a single assessment model.

8.4 Challenges

Several obstacles were experienced during the project.

One of the challenges was how to identify appropriate privacy-preserving strategies for comparison. The field is composed of many approaches, each with different assumptions, advantages and disadvantages. An intensive literature review and critical evaluation were needed to find strategies that were representative and actually implementable.

"Another challenge was to choose appropriate evaluation metrics. Privacy, utility, and computational performance are all fundamentally different notions, thus direct comparison is problematic. A composite scoring framework was designed with particular attention to weighting strategies and normalisation approaches for meaningful comparisons.

There were additional technological obstacles in the installation process. We have to iterate through data pre-processing, feature encoding, privacy translation, and machine learning assessment to explore and improve. It was very crucial to keep the experimental workflow similar, and to ensure a fair comparison across the different approaches.

Another additional problem was time management because the experimentation was iterative. The project involved repeated testing and modification in several stages before good findings were attained.

section titleEvaluation of the Research Results

The results of the research show that the proposed evaluation framework has fulfilled the objectives set at the beginning of the investigation.

The methodology permitted the comparison of k-anonymity, differential privacy and synthetic data generation under a single set of evaluation criteria. This composite scoring approach offered a visible, interpretable method to rank privacy-preserving strategies by overall efficacy.

Experimental results showed that synthetic data generation achieved the highest overall composite score and that differential privacy provided the strongest privacy guarantees. These results validate the basic premise of the dissertation that privacy-preserving systems should be evaluated along different dimensions, not just on privacy protection.

The framework met the targeted objectives but the research also identified areas for improvement. Future implementations can include more datasets, sophisticated synthetic data creation techniques and empirical privacy attack simulations to further fortify the evaluation process.

8.5 Working and Academic Development

This dissertation has been a great contribution to my academic and professional growth.

At an academic level, the project contributed to the knowledge of privacy-preserving technologies, evaluation of machine learning models, data governance, and principles of cybersecurity. It also boosted skills of academic writing, critical analysis, literature review and research methodologies.

The experience enhanced skills immediately applicable to professions in cybersecurity, data analytics, artificial intelligence, and privacy engineering. The capacity to assess technical solutions, analyze experimental data, and convey research findings proficiently is a crucial skill set for future employment and career progression.

The research heightened awareness regarding the ethical obligations associated with managing sensitive information and creating data-driven solutions. In the current digital landscape, where data sharing and analytics are prevalent, understanding the equilibrium between privacy protection and data utility is becoming increasingly crucial.

8.6 Lessons Learned:

The study highlighted several crucial lessons.

First, privacy preservation should not be seen in isolation of usability and performance. Effective solutions require balance, not optimisation of one purpose.

Second, evaluation frameworks need to remain interpretable if they are to be useful for real decision-making. The utility of a large number of measures to practitioners could

be reduced if not presented with a clear mechanism for assessment.

Third, the reliability of the machine learning assessment outcomes is critically dependent on the experimental design and data pretreatment. Thus, careful planning and validation is important throughout the research process.

Finally, working in fast-moving domains such as privacy-preserving machine learning and data protection requires ongoing learning and adaptation.

8.7 Summary

In this chapter we reflected on the research journey undertaken throughout the dissertation. It described the abilities gained, obstacles met, lessons learned and the overall contribution made by the project to academic and professional development.

The reflection pointed out the importance of a trade-off between privacy protection, machine learning value, and computing performance when assessing privacy-preserving data publishing approaches. It also illustrated how information and experience accumulated during the project has contributed to a greater understanding of privacy engineering, cybersecurity and data-driven decision-making.

Bibliography

[1] M. N. M. Khan, "Artificial intelligence driven big data and business analytics: A comprehensive review of multi-sectoral applications in healthcare, finance, supply chain, and organizational innovation," *Pacific Journal of Business Innovation and Strategy*, vol. 2, no. 4, pp. 122–137, 2025.

[2] S. Kunungo, S. Ramabhotla, and M. Bhoyar, "The integration of data engineering and cloud computing in the age of machine learning and artificial intelligence," *Iconic Research And Engineering Journals*, vol. 1, no. 12, pp. 79–84, 2018.

[3] L. Sweeney, "k-anonymity: A model for protecting privacy," *International journal of uncertainty, fuzziness and knowledge-based systems*, vol. 10, no. 05, pp. 557–570, 2002.

[4] S. Shaham, A. Hajisafi, M. K. Quan, D. C. Nguyen, B. Krishnamachari, C. Peris, G. Ghinita, C. Shahabi, and P. N. Pathirana, "Privacy and fairness in machine learning: A survey," *arXiv preprint*, 2025.

[5] A. Machanavajjhala, D. Kifer, J. Gehrke, and M. Venkitasubramaniam, "l-diversity: Privacy beyond k-anonymity," *Acm transactions on knowledge discovery from data (tkdd)*, vol. 1, no. 1, pp. 3–es, 2007.

[6] ——, "l-diversity: Privacy beyond k-anonymity," *ACM Transactions on Knowledge Discovery from Data*, vol. 1, no. 1, pp. 1–52, 2007.

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

[8] C. Dwork and J. Lei, "Differential privacy and robust statistics," in *Proceedings of the forty-first annual ACM symposium on Theory of computing*, 2009, pp. 371–380.

[9] A. Majeed and S. O. Hwang, "Differential privacy and k-anonymity-based privacy preserving data publishing scheme with minimal loss of statistical information," *IEEE Transactions on Computational Social Systems*, vol. 11, no. 3, pp. 3753–3767, 2024.

[10] P. Kairouz *et al.*, "Advances and open problems in federated learning," *Foundations and Trends in Machine Learning*, vol. 14, no. 1–2, pp. 1–210, 2021.

[11] J. Jordon, L. Szpruch, F. Houssiau, M. Bottarelli, G. Cherubin, C. Maple, S. N. Cohen, and A. Weller, "Synthetic data – what, why and how?" *arXiv preprint arXiv:2205.03257*, 2022.

[12] I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio, "Generative adversarial networks," *Communications of the ACM*, vol. 63, no. 11, pp. 139–144, 2020.

[13] L. Xu, M. Skoularidou, A. Cuesta-Infante, and K. Veeramachaneni, "Modeling tabular data using conditional gan," in *Advances in Neural Information Processing Systems*, 2019.

[14] K. Pan *et al.*, "Differential privacy in deep learning: A literature survey," *Neurocomputing*, 2024.

[15] Z. Qian *et al.*, "Synthcity: Facilitating innovative synthetic data benchmarks," *Nature Communications*, vol. 14, no. 1, 2023.

[16] S. Lautrup *et al.*, "Syntheval: A comprehensive framework for synthetic data evaluation," *Data Mining and Knowledge Discovery*, 2025.

[17] B. Becker and R. Kohavi, "Adult," UCI Machine Learning Repository, 1996, DOI: https://doi.org/10.24432/C5XW20.

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

[19] A. Narayanan and V. Shmatikov, "Robust de-anonymization of large sparse datasets," *IEEE Symposium on Security and Privacy*, pp. 111–125, 2008.

[20] M. Sanchez-Serrano *et al.*, "Decision frameworks for privacy-preserving synthetic data selection," *Journal of Information Security and Applications*, 2025.

[21] M. Giomi *et al.*, "Anonymeter: Quantifying privacy risks in synthetic data," *Proceedings on Privacy Enhancing Technologies*, 2023.

[22] N. Li, T. Li, and S. Venkatasubramanian, "t-closeness: Privacy beyond k-anonymity and l-diversity," in *IEEE International Conference on Data Engineering*, 2007, pp. 106–115.

[23] B. C. Kara *et al.*, "(r, k, epsilon)-anonymization: Privacy-preserving data publishing using k-anonymity and differential privacy," *Expert Systems with Applications*, 2025.

[24] C. Dwork, "Differential privacy," in *International Colloquium on Automata, Languages and Programming*, 2006, pp. 1–12.

[25] J. P. Near *et al.*, "Guidelines for evaluating differential privacy guarantees," National Institute of Standards and Technology, Tech. Rep. NIST Special Publication 800-226, 2025.

[26] N. Patki, R. Wedge, and K. Veeramachaneni, "The synthetic data vault," in *IEEE International Conference on Data Science and Advanced Analytics*, 2016, pp. 399–410.

[27] R. Shokri, M. Stronati, C. Song, and V. Shmatikov, "Membership inference attacks against machine learning models," in *IEEE Symposium on Security and Privacy*, 2017, pp. 3–18.

[28] M. Rigaki and S. Garcia, "A survey of privacy attacks in machine learning," *ACM Computing Surveys*, vol. 56, no. 4, pp. 1–34, 2024.

[29] C. Dwork and A. Roth, *The Algorithmic Foundations of Differential Privacy*. Now Publishers, 2014.

[30] J. W. Creswell and J. D. Creswell, *Research Design: Qualitative, Quantitative, and Mixed Methods Approaches*, 5th ed. SAGE Publications, 2018.

[31] M. Saunders, P. Lewis, and A. Thornhill, *Research Methods for Business Students*, 8th ed. Pearson, 2019.

[32] I. Goodfellow, Y. Bengio, and A. Courville, *Deep Learning*. MIT Press, 2016.

[33] J. Han, M. Kamber, and J. Pei, *Data Mining: Concepts and Techniques*, 3rd ed. Morgan Kaufmann, 2011.

[34] A. K. Jain, K. Nandakumar, and A. Ross, "Score normalization in multimodal biometric systems," *Pattern Recognition*, vol. 38, no. 12, pp. 2270–2285, 2005.

[35] A. Zheng and A. Casari, *Feature Engineering for Machine Learning*. O'Reilly Media, 2018.

[36] B. Efron and R. Tibshirani, *An Introduction to the Bootstrap*. CRC Press, 1994.

[37] D. M. W. Powers, "Evaluation: From precision, recall and f-measure to roc, informedness, markedness and correlation," *Journal of Machine Learning Technologies*, vol. 2, no. 1, pp. 37–63, 2011.

[38] A. J. Tyagi *et al.*, "Scaling deep learning models: Challenges and solutions for large-scale deployments," *WORLD JOURNAL OF ADVANCED ENGINEERING TECHNOLOGY AND SCIENCES*, vol. 16, no. 2, pp. 010–020, 2025.

[39] E. Triantaphyllou, *Multi-Criteria Decision Making Methods: A Comparative Study*. Springer, 2000.

[40] V. Belton and T. Stewart, *Multiple Criteria Decision Analysis: An Integrated Approach*. Springer, 2002.

[41] N. B. Njungle, E. Jahns, Z. Wu, L. Mastromauro, M. Stojkov, and M. A. Kinsy, "Guardianml: anatomy of privacy-preserving machine learning techniques and frameworks," *IEEE Access*, vol. 13, pp. 61 483–61 510, 2025.

[42] J. Wątróbski, J. Jankowski, P. Ziemba *et al.*, "Generalised framework for multi-criteria method selection," *Omega*, vol. 86, pp. 107–124, 2019.

[43] W. Sałabun *et al.*, "Comparison of multi-criteria decision analysis methods under comprehensive sensitivity analysis," *IEEE Access*, 2026.

[44] M. Z. D. la Cal *et al.*, "A multi-criteria decision support system for data-driven strategic planning," *Sustainability*, vol. 18, no. 3, p. 1412, 2026.

[45] X.-k. Wang, H.-y. Zhang, Y.-n. Wang, W.-h. Hou, J.-q. Wang, and H.-g. Peng, "A decision support system for comments-adjusted ranking of hotels," *Journal of the Operational Research Society*, vol. 76, no. 9, pp. 1880–1897, 2025.

[46] F. Pedregosa, G. Varoquaux, A. Gramfort *et al.*, "Scikit-learn: Machine learning in python," in *Journal of Machine Learning Research*, vol. 12, 2011, pp. 2825–2830.

[47] W. McKinney, "Data structures for statistical computing in python," in *Proceedings of the 9th Python in Science Conference*, 2010, pp. 51–56.

[48] C. R. Harris, K. J. Millman, S. J. van der Walt *et al.*, "Array programming with numpy," *Nature*, vol. 585, no. 7825, pp. 357–362, 2020.

[49] J. Wątróbski *et al.*, "Generalised framework for multi-criteria method selection," *Omega*, vol. 86, pp. 107–124, 2019.

[50] W. Sałabun, "The characteristic objects method: A new distance-based approach to mcda problems," *Expert Systems with Applications*, vol. 159, p. 113575, 2020.

[51] D. V. Carvalho, E. M. Pereira, and J. S. Cardoso, "Machine learning interpretability: A survey on methods and metrics," *Electronics*, vol. 8, no. 8, p. 832, 2019.